

Analiza danych medycznych dotyczących chorób tarczycy

★ ★ ★

Raport końcowy

Paulina Przybyłek

23.01.2023

Spis treści

1. Wstęp	3
1.1. Cel Projektu	3
1.2. Problem badawczy	3
1.3. Zbiór danych	4
2. Charakterystyka i analiza danych	5
2.1. Opis zmiennych w zbiorze	5
2.2. Krótkie zapoznanie się z danymi	6
2.3. Charakterystyka danych	7
2.4. Eksploracja danych	9
3. Przetworzenie danych	13
3.1. Wstępne przetworzenie danych	13
3.2. Podział na zbiór treningowy i testowy	13
3.2.1. Eksploracja danych treningowych	14
3.3. Wybór najlepszego zbioru do modelowania	15
3.3.1. Surowe dane	16
3.3.2. Porzucenie kolumn z pomiarami hormonów	16
3.3.3. Porzucenie kolumn zawierających w nazwie „measured”	17
3.3.4. PCA	17
3.3.5. Najlepszy wybrany zbiór	18
4. Ostateczne modele wraz z uzyskaną predykcją	20
4.1. Wpływ transformacji	20
4.2. Klasyfikatory	20
4.3. Wybór najlepszego modelu	21
4.4. XAI	23
5. Podsumowanie	25

1. Wstęp

Projekt realizowany w ramach przedmiotu *Podstawy Przetwarzania Danych*.

1.1. Cel Projektu

Celem projektu jest klasyfikacja pacjenta jako zdrowego bądź chorego na podstawie danych medycznych pacjentów, w tym wyników testów morfologicznych. Jednak sam temat można podzielić na analizę tych danych oraz na predykcję przetworzonych danych zgodnie z przeprowadzoną analizą.

Podobny projekt został zrealizowany uprzednio w programie SAS, stąd dodatkowym założeniem jest odtworzenie tych analiz w Pythonie oraz porównanie pewnych wyników między sobą.

1.2. Problem badawczy

Uczenie maszynowe wraz z modelami do predykcji różnych wartości są już stosowane w wielu dziedzinach życia. Próbuje się też wykorzystać je w medycynie, jednak oddanie życia pacjenta predykcji modelu nie należy do łatwych decyzji. Jeśli w biznesie nastąpi pomyłka przewidzianej wartości to możliwą stratą są pieniądze bądź inne dobra materialne. Niestety w przypadku danych medycznych sytuacja staje się gorsza, ponieważ stawką jest ludzkie życie. Nikt nie jest w stanie i nawet nie próbuje wycenić ludzkiego życia, mimo to jest to w pewnym sensie potrzebne do uczenia takich modeli. Jeśli model zaklasyfikuje osobę zdrową jako chorego pacjenta to po wykonaniu kolejnych badań lekarz zauważa, że to jednak była pomyłka. Jednak, gdy chory pacjent zostanie uznany za zdrowego i wypuszczony do domu to potem może umrzeć. Kogo wówczas obwinia się za taką sytuację? Możliwe, że lekarz był w stanie zauważać chorobę nawet jeśli model nie potrafił. Taki błąd jest nazywany błędem II rodzaju i w przypadku modeli stosowanych do danych medycznych należy go jak najbardziej zminimalizować. Jeśli uda się znaleźć model, który nie popełnia takich błędów, tzw. „model idealny”, to i tak nie powinno się powierzać całkowitej decyzji modelowi, ponieważ w przypadku wyniki badań pacjentów mają czasem anomalie, które wcześniej się nie pojawiły, i to mogłyby zostać źle ocenione przez model. Stąd najlepiej, aby modele były stosowane jako pomoc dla lekarzy w decyzji co zrobić z pacjentem - na zasadzie dodatkowego narzędzia. Nawet jako narzędzia a nie samodzielna jednostka

decydującą za lekarza, to uczenie maszynowe mogłoby przydać się lekarzom ułatwiając ich pracę bądź przyspieszając ją. W projekcie podjęto się próbie znalezienia takiego „idealnego modelu”, który będzie minimalizował błąd II rodzaju, dla danych medycznych pacjentów, którzy byli badani pod względem posiadania jakiejś choroby tarczycy.

1.3. Zbiór danych

Dane pacjentów badanych pod względem choroby tarczycy pochodzą ze strony UCI Machine Learning Repository [Dua i Graff 2017, Quinlan 1987] i zostały dostarczone w 1987 roku przez Garavan Institute i J. Rossa Quinlana z Instytutu Nowej Południowej Walii w Sydney w Australii. Na powyższej stronie znajduje się kilka zbiorów danych, jednak można znaleźć informacje o tym, że możliwe jest, że zbiory te są uszkodzone poprzez zaszumienie danych i jedynym poprawnie przygotowanym zbiorem jest ten o nazwie *sick*. Stąd to ten konkretny zbiór rozważano w projekcie.

Zbiór danych *sick* posiada 29 zmiennych objaśniających oraz 1 zmienną celu binarną, gdzie pacjenci są podzieli na zdrowych i na tych co mają jakąś chorobę tarczycy. Zmienne objaśniające są kategoryczne oraz numeryczne, przy czym jest ich odpowiednio 22 i 7. Atrybuty zawierają wyniki badań pacjenta bądź informacje o nim z wywiadu lekarskiego. Liczba wszystkich rekordów (to jest pacjentów) wynosi 3772.

2. Charakterystyka i analiza danych

Niniejszy rozdział zawiera EDA (exploratory data analysis) zbioru w postaci wybranych informacji z przeprowadzonych analiz.

2.1. Opis zmiennych w zbiorze

Podczas analizy danych przydatna jest wiedza o tym czego dotyczy dana kolumna. Z tego powodu Tabele 2.1 i 2.2 przedstawiają wszystkie atrybuty zbioru danych wraz z krótkim opisem ich znaczenia. Informacje zebrane zostały ze strony zbioru a także z wykorzystaniem innych źródeł mogących pomóc w medycznym zrozumieniu niektórych rzeczy.

Nazwa atrybutu w zbiorze	Znaczenie
age	wiek
sex	płeć
on_thyroxine	pacjent zażywający tyroksynę
query_on_thyroxine	pacjent może być zażywający tyroksynę
on_antithyroid_medication	pacjent na lekach przeciwtarczycowych
sick	pacjent zgłasza złe samopoczucie
pregnant	pacjent w ciąży
thyroid_surgery	historia operacji tarczycy (po operacji tarczycy)
I131_treatment	pacjent na leczeniu I131
query_hypothyroid	być może niedoczynność tarczycy
query_hyperthyroid	być może nadczynność tarczycy
lithium	pacjent przyjmujący lit
goitre	pacjent ma wole (powiększenie gruczołu tarczycy)
tumor	pacjent ma guza
hypopituitary	pacjent z niedoczynnością przysadki
psych	objawy psychologiczne

Tabela 2.1: Lista atrybutów (zmiennych objaśniających wraz ze zmienną celu) zbioru *sick*, cz. 1/2.

Nazwa atrybutu w zbiorze	Znaczenie
TSH_measured	czy zmierzono TSH
TSH	wartość TSH (tyreotropiny) - zwiększenie masy tarczycy, zwiększenie przepływu krwi przez ten narząd oraz nasilenie produkcji i wydzielania hormonów tarczycy: tyroksyny i trójjodotyroniny
T3_measured	czy zmierzono T3
T3	wartość T3 (trijodotyronina) - wyróżniamy dwie: całkowitą i wolną. T3 całkowita (TT3) ma normę 1,3–3,1 nmol/l, a T3 wolna (FT3) ma normę 4,0–7,8 pmol/l, założono, że w projekcie najpewniej chodzi o wolne T3
TT4_measured	czy zmierzono TT4
TT4	wartość TT4 (tyroksyna) - T4 całkowita, norma 58–154 nmol/l
T4U_measured	czy zmierzono T4U
T4U	wartość T4U - wykorzystanie tyroksyny przez organizm (eksplatacja, pobór)
FTI_measured	czy zmierzono FTI
FTI	wartość FTI - wolny testosteron, wiarygodny indeks do oceniania stanu tarczycy, FTI = Thyroxine (T4)/Thyroid Binding Capacity
TBG_measured	czy zmierzono TBG
TBG	wartość TBG - stężenie globuliny wiążącej tyroksynę
referral_source	źródło skierowania
Class	target (zmienna objaśniana, zmienna celu) określa chorobę tarczycy lub jej brak

Tabela 2.2: Lista atrybutów (zmiennych objasniających wraz ze zmienną celu) zbioru *sick*, cz. 2/2.

2.2. Krótkie zapoznanie się z danymi

Surowy zbiór danych posiada 30 kolumn, których wartości są typu numerycznego bądź kategorycznego. Poniżej wymieniono podstawowe bądź ciekawe informacje o analizowanych danych.

- Zmienna celu jest binarna, ale w postaci zmiennej kategorycznej o wartościach „negative” (zdrowy pacjent) i „sick” (pacjent z chorobą tarczycy).
- Zbiór posiada **braki danych**, przy czym jedna z kolumn jest całkowicie pozbawiona wartości - jest pusta. Brakujące wartości występują dla zmiennych numerycznych (wiek oraz wartości hormonów) oraz dla zmiennej płeć. Procent braków przedstawia Tabela 2.3.

2.3. CHARAKTERYSTYKA DANYCH

Nazwa atrybutu w zbiorze	Procent brakujących wartości [%]
age	0.03
sex	3.98
TSH	9.78
T3	20.39
TT4	6.12
FTI	10.21
T4U	10.26
TBG	100

Tabela 2.3: Procent brakujących wartości dla danej zmiennej w zbiorze.

- Większość zmiennych kategorycznych jest binarna, na zasadzie „Tak”/„Nie”.
- Kolumny „_measured” odpowiadają tym bez tego przyrostka, czyli nie wnoszą nowych informacji mimo to, czasem dla modelu tak opisana informacja może być bardziej znacząca niż sam pomiar. Założymy, że dane badanie jest specjalistyczne i kosztowne - wówczas, jeśli nie mamy podejrzeń do choroby u pacjenta to badanie nie zostanie wykonane, stąd informacja o braku pomiaru może sugerować zdrowego pacjenta.
- W zbiorze występuje więcej kobiet niż mężczyzn. Może wynikać z faktu, iż w rzeczywistości to częściej kobiety chorują na tarczycę, więc częściej powinny się badać.
- Zbiór jest **niezbalsowanym**, przy czym współczynnik wynosi ponad 15.

2.3. Charakterystyka danych

Dla zmiennych kategorycznych wykonano charakterystykę w postaci zliczenia liczby rekordów z podziałem na wartości występujące w zmiennej. Wynik przedstawiono w postaci liczbowej oraz procentowej. Rysunek 2.1 przedstawia owe wyniki dla wybranych zmiennych. Charakterystyki te potwierdzają zauważone w sekcji 2.2 informacje o zbiorze.

Kobiet wśród pacjentów jest prawie 66%, przy czym mężczyzn nieco ponad 30%. Dodatkowo nie wszyscy pacjenci mają określoną płeć.

Można zauważyć, że tylko jeden pacjent wśród 3772 posiadał niedoczynność przysadki oraz żaden z pacjentów nie miał badanego hormonu TBG, z nieznanych nam przyczyn.

Zmienna określająca źródło skierowania posiada pięć różnych wartości, niestety nigdzie nie jest wyjaśnione do czego dokładnie się one odnoszą. Większość pacjentów pochodzi z innych placówek niż wymienione z nazwy. Informacja ta może jedynie wskazywać na to, że w danym miejscu (położenie

geograficzne) ludzie chorują na tarczycę częściej bądź dane miejsce jest znane z dobrego leczenia tych chorób.

Niebalansowanie zbioru o współczynniku 15 odnosi się do tego, że jedynie ok. 6% rekordów dotyczy chorych pacjentów, reszta to zdrowe osoby. Taki podział etykiet w zadaniu klasyfikacji jest utrudnieniem, stąd podczas uczenia i ocenie modeli trzeba brać to pod uwagę.

sex			
	Frequency	Count	Percent of Total Frequency
F	2480		0.657476
M	1142		0.302757
NaN	150		0.039767

hypopituitary			
	Frequency	Count	Percent of Total Frequency
f	3771		0.999735
t	1		0.000265

TBG_measured			
	Frequency	Count	Percent of Total Frequency
f	3772		1.0

referral_source			
	Frequency	Count	Percent of Total Frequency
other	2201		0.583510
SVI	1034		0.274125
SVHC	386		0.102333
STMW	112		0.029692
SVHD	39		0.010339

Class			
	Frequency	Count	Percent of Total Frequency
negative	3541		0.938759
sick	231		0.061241

Rysunek 2.1: Charakterystyka zmiennych: sex, hipopituitary, TBG_measured, referral_source oraz Class (zmienna celu).

Dodatkowo przyglądając się tym wynikom dla reszty zmiennych oraz wykresom rozkładu to zmienne, które mają dwie wartości w większości rekordów mają wartość „Nie”, co w sumie jest oczywiste biorąc pod uwagę, że tylko 6% rekordów to chore osoby, aby zmienne te określają powody do wystąpienia choroby.

W przypadku zmiennych numerycznych zdecydowano się na wywołanie podstawowych statystyk, które oferuje funkcja *describe()* w Pythonie. Rysunek 2.2 przedstawia tabelę z wynikami.

Wartości uzyskiwane przez hormony nie mówią nam bez odpowiedniej wiedzy medycznej, natomiast warto zwrócić uwagę na inną zmienną - wiek. W zbiorze znajduje się pacjent, którego wiek wynosi 455, co jest niemożliwe, aby się stało, stąd jest to błąd w danych. Prawdopodobnie jest to pomyłka przy wpisywaniu wieku - może być to wartość 45 bądź 55. Nie jesteśmy w stanie tego stwierdzić, więc najlepiej będzie usunąć tę wartość.

2.4. EKSPLORACJA DANYCH

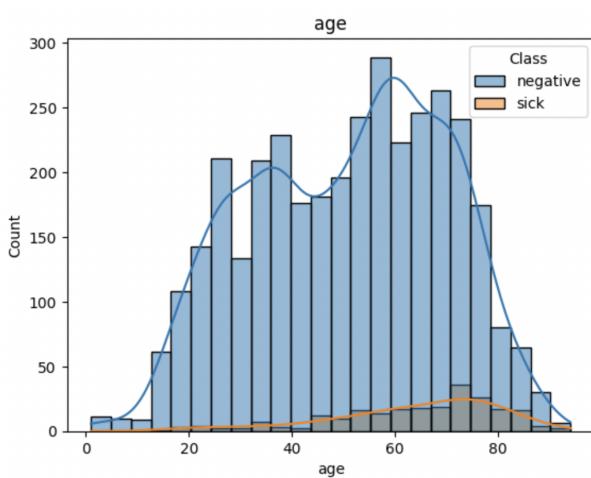
	age	TSH	T3	TT4	T4U	FTI
count	3771.000000	3403.000000	3003.000000	3541.000000	3385.000000	3387.000000
mean	51.735879	5.086766	2.013500	108.319345	0.995000	110.469649
std	20.084958	24.521470	0.827434	35.604248	0.195457	33.089698
min	1.000000	0.005000	0.050000	2.000000	0.250000	2.000000
25%	36.000000	0.500000	1.600000	88.000000	0.880000	93.000000
50%	54.000000	1.400000	2.000000	103.000000	0.980000	107.000000
75%	67.000000	2.700000	2.400000	124.000000	1.080000	124.000000
max	455.000000	530.000000	10.600000	430.000000	2.320000	395.000000

Rysunek 2.2: Podstawowe statystyki dla zmiennych numerycznych z wykluczeniem kolumny TBG.

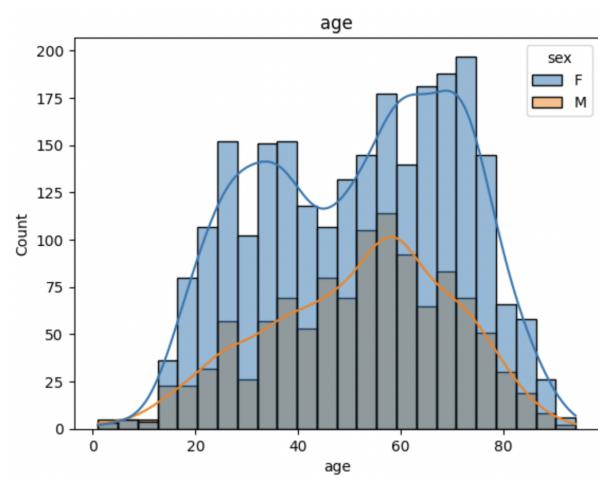
2.4. Eksploracja danych

Rozkłady zmiennych

Zdecydowano się sprawdzić rozkłady zmiennych numerycznych. W raporcie zaprezentowano wykresy dla wybranych zmiennych. Rysunek 2.3 dotyczy rozkładu dla wieku pacjenta. Wraz z wiekiem wzrasta liczba chorych na tarczycę. Dodatkowo warto zauważać dwa piki dla osób zdrowych. Można wywnioskować (choć jest to bardziej zgadywanie), że najczęściej w wieku bliżej 30 r.ż. ludzie decydują się zbadać, aby sprawdzić zdrowie a następnie już po 60 r.ż., kiedy szanse na zachorowanie wzrastają. Na Rysunku 2.4 widać, że te piki dotyczą kobiet głównie, więc wiedząc, że kobiety częściej chorują na tarczycę oczywistym jest zakładanie, że w młodości jeszcze sprawdzają stan zdrowia.



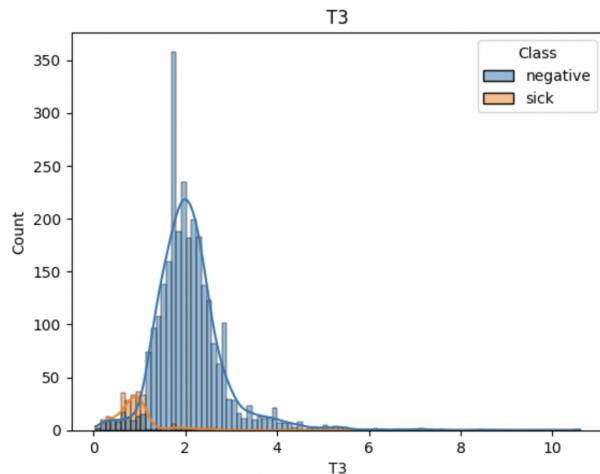
Rysunek 2.3: Rozkład zmiennej dotyczącej wieku pacjenta w zależności od wartości zmiennej celu.



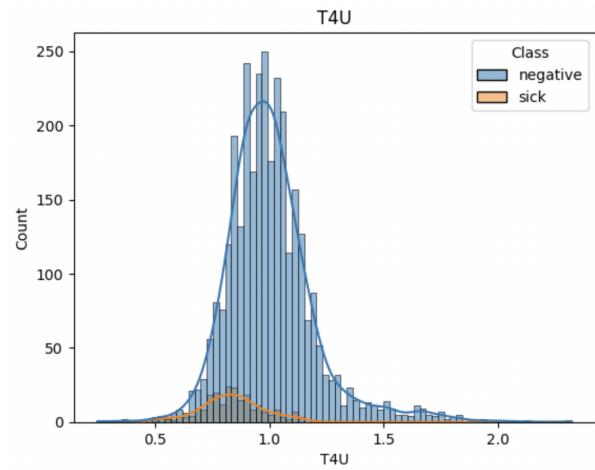
Rysunek 2.4: Rozkład zmiennej dotyczącej wieku pacjenta w zależności od płci.

Najciekawszym wykresem jest rozkład zmiennej T3. Rysunek 2.5 pokazuje, że wartości dla chorych pacjentów skupią się na niższych wartościach hormonu T3 niż w przypadku zdrowej osoby. W przypadku hormonu TT4 czy T4U również wartości są mniejsze w przypadku chorego pacjenta, ale nie jest

to już tak widoczne (Rysunek 2.6).



Rysunek 2.5: Rozkład zmiennej T3 (hormonu) w zależności od wartości zmiennej celu.



Rysunek 2.6: Rozkład zmiennej T4U (hormonu) w zależności od wartości zmiennej celu.

Charakterystyka danych dla zmiennych kategorycznych i pacjentów chorych na tarczycę

Powtórzono charakterystykę dla atrybutów kategorycznych, tyle że z podziałem na osoby chore i zdrowe. Z interesujących informacji można wskazać, że wśród chorych pacjentów:

- 58.87% osób to kobiety,
- osoby chore nie mają wartości true dla kolumny on_antithyroid_medication, czyli nie są na lekach przeciwtarczycowych,
- brakuje osób w ciąży,
- nikt nie miał operacji tarczycy,
- tylko 2 osoby mają powiększony gruczoł tarczycy,
- tylko 2 osoby mają guza,
- TT4 zmierzono wszystkim.

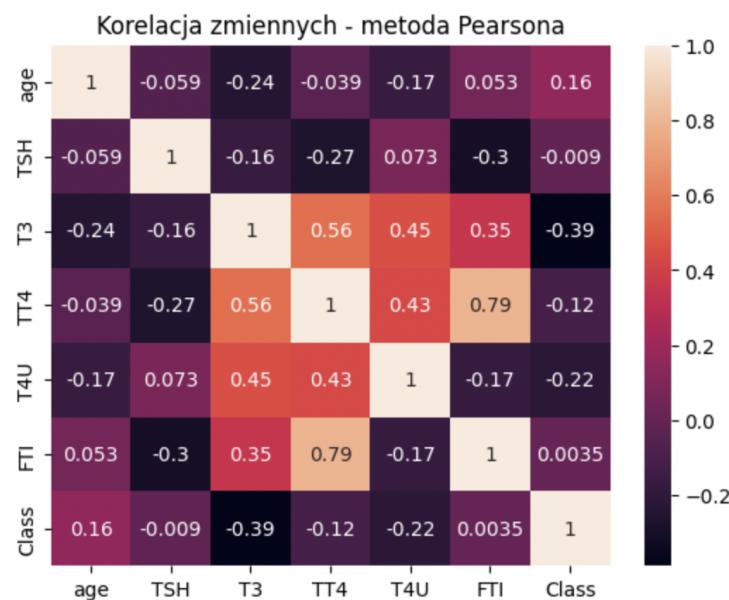
Korelacja zmiennych

Sprawdzono dla zmiennych numerycznych korelację między nimi i zmienną celu. Na Rysunku 3.2 przedstawiono korelację metodą Pearsona. Przykładowe wnioski z wykresu korelacji:

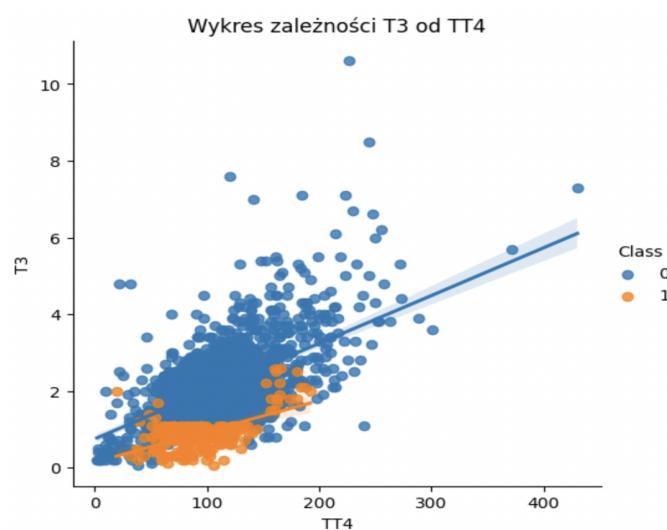
- wysoka korelacja zachodzi między TT4 a FTI i jest to spowodowane tym, że wskaźnik FTI wylicza się z wartości hormonu TT4;

2.4. EKSPLORACJA DANYCH

- spadek T3 wpływa na niedoczynność tarczycy, w sumie to wszystkie hormony mają ujemną korelację, a wzrost wieku sprzyja chorobom tarczycy;
- TT4 i T3 mają dodatnią korelację i to całkiem dużą - występuje między nimi zależność z punktu medycznego (zależność widać na Rysunku 3.1);
- T4U ma dodatnią korelację o dość sporej wartości z TT4 i T3;
- powyższe dwa punkty to hormony wytwarzane przez tarczyce, więc ich zależności mogły być do przewidzenia.



Rysunek 2.7: Korelacje między zmiennymi numerycznymi i zmienną celu wykonane metodą Pearsona.



Rysunek 2.8: Wykres zależności T3 od TT4 z podziałem na zmienną celu.

W przypadku zmiany metody korelacji na Spearmana, jedyną dużą różnicą jest spadek korelacji między FTI a T3 do 0.11.

Obserwacje odstające

Zdecydowano się też za pomocą kryterium Chauvenet'a znaleźć obserwacje odstające. Dla wszystkich zmiennych numerycznych metoda zwróciła jakieś wartości uznane za odstające. Oczywiście, zastosowane kryterium odrzuciło anomalię w postaci wieku równego 455. Przy innych wartościach hormonów również pojawiły się wartości uznane za odstające. Z punktu medycznego, osiągnięcie takich danych jest wykonalne, więc można założyć, że usunięcie wartości odstających dla hormonów nie będzie najlepszym rozwiązaniem, gdyż te anomalie mogą być powodem choroby - w medycynie wszystko jest możliwe, a skoro są to osiągalne wyniki to lepiej je zostawić.

3. Przetworzenie danych

Po przeprowadzeniu analizy zbioru przystąpiono do jego oczyszczenia i przetworzenia w ramach przygotowania do modelowania. Rozdział ten zawiera opis przeprowadzonych w tym celu działań.

3.1. Wstępne przetworzenie danych

Wstępne przetworzenie polegało na następujących zmianach:

1. Usunięcie kolumny referral_source - źródło skierowania pacjenta (szpital) nie powinno mieć wpływu na uczenie. Możliwe, że jakiś szpital skupia się na leczeniu danej choroby i tam będzie więcej chorych, jednak planowane modelowanie ma klasyfikować na podstawie informacji o samych pacjencie.
2. Usunięcie kolumny TBG ze względu na to, iż jest to cała pusta kolumna, pozbawiona wyników badań.
3. Zmiana wartości „f” i „t” na „0” i „1”, aby zamienić zmienne kategoryczne na numeryczne.
4. Zmiana wartości zmiennej celu na „0” i „1”, gdzie oznacza to odpowiednio „negative” (zdrowego pacjenta) i „sick” (pacjenta z chorobą tarczycy).
5. Zmiana wartości zmiennej płeć (sex) na „0”, „1” oraz „NAN”, gdzie oznacza to odpowiednio kobietę, mężczyznę oraz brakującą wartość.
6. Pozbycie się wartości odstających - uznano, że jedynie wartość wieku równa 455 jest wartością odstającą, stąd zdecydowano się zmienić ją na brak danych, żeby nie usuwać całego wiersza.

3.2. Podział na zbiór treningowy i testowy

Tak przygotowany zbiór jak opisano to w poprzedniej sekcji (3.1) podzielono na zbiór treningowy oraz testowy zachowując proporcje rozkładu klas zmiennej celu. Ustalono podział 75% rekordów dla zbioru treningowego a pozostałe 25% to zbiór testowy. Zrezygnowano z tworzenia zbioru walidacyjnego ze względu na jego małe wykorzystanie w Pythonie jako oddzielnego zbioru. Stąd ustalone zbiory

treningowy i testowy różnią się od podziału w projekcie z SASa, więc od tej pory należało mieć to na uwadze przy porównywaniu wyników.

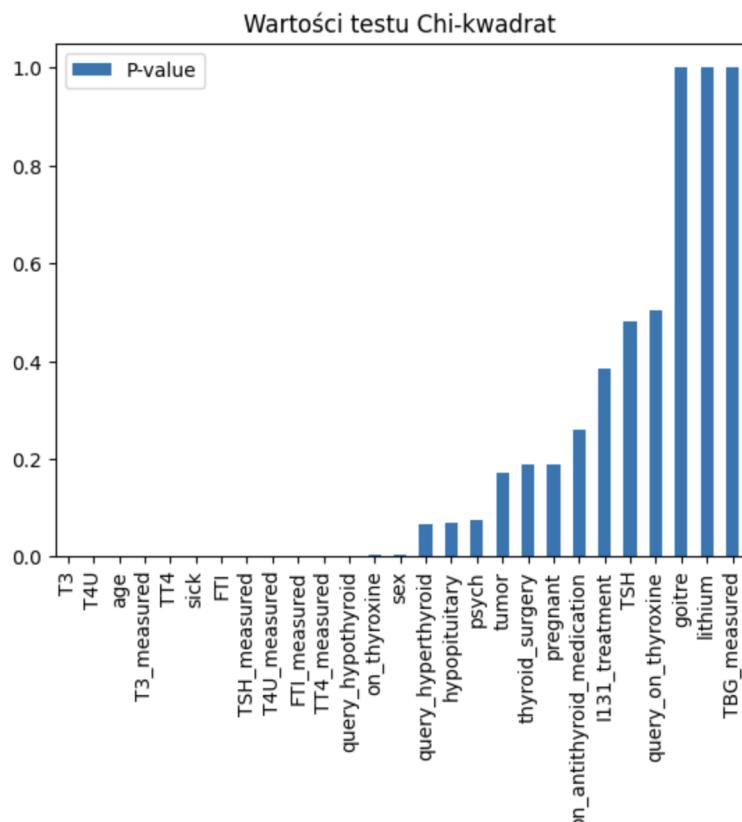
3.2.1. Eksploracja danych treningowych

Powtórzono część EDA dla zbioru treningowego, aby zdecydować o sposobie przetworzenia danych do uczenia modeli. Jak wiadomo w realnej sytuacji dostęp ma się tylko do danych treningowych, więc o zbiorze testowym nie wiemy nic. Stąd niektóre wyniki i analizy zostały ponownie przeprowadzone, aby zobaczyć czy może coś się nie zmieniło.

Rozkłady zmiennych czy korelacje pozostały bez zmian jak dla całego zbioru. Zmieniła się liczba braków danych, ale procentowo jest nadal zbliżone do poprzedniego udziału braków w danych. Stąd wnioski o danych można pozostawić takie same jak poprzednio.

Test Chi-kwadrat

Postanowiono, jak w uprzednio w projekcie z SAS, sprawdzić ważność zmiennych poprzez wykonanie testu Chi-kwadrat. Na Rysunku 3.1 przedstawiono wyniki testu.



Rysunek 3.1: Wykres p-value dla zmiennych z testu Chi-kwadrat.

Im mniejsza wartość p-value tym zmienna jest ważniejsza. W ten sposób można powiedzieć, że zmienne, które mają znaczenie w zbiorze przy poziomie istotności 0.05 to: Po odrzuceniu warto-

3.3. WYBÓR NAJLEPSZEGO ZBIORU DO MODELOWANIA

ści hormonów, w SASie dokonano selekcji zmiennych za pomocą wartości Chi-kwadrat i do uczenia wybrano następujące atrybuty zbioru: age, T3_measured, sick, TSH_measured, T4U_measured, FTI_measured, TT4_measured, query_hypothyroid, on_thyroxine oraz sex. Różnią się one od wyboru w projekcie z SASa, ponieważ zbiór treningowy został wybrany w inny sposób. Analiza z SASa wybrała kolumny: sex, T3_measured, age, psych, query_hypothyroid oraz sick. W aktualnym projekcie wybrano więcej zmiennych, przy czym wszystkie poza jedną się powtórzyły.

Analiza skupień

Analiza skupień na zmiennych numerycznych zwróciła dwa klastry. Jeden zawiera TSH, TT4 i FTI, gdzie FTI jest współczynnikiem liczymy z TT4, a TSH jest dość powiązany z TT4, więc taki klaster ma sens. Drugi z klastrów zawiera resztę zmiennych, do nich nie da się powiedzieć już czegoś o powiązaniach.

3.3. Wybór najlepszego zbioru do modelowania

W SAS rozważano:

1. surowe dane - porzucenie kolumn z jedną wartością dla wszystkich wierszy;
2. analiza skupień;
3. dane po odrzuceniu pomiarów hormonów + porzucenie kolumn z jedną wartością dla wszystkich wierszy, imputacja, wybór zmiennych;
4. dane po odrzuceniu kolumn "measured" + porzucenie kolumn z jedną wartością dla wszystkich wierszy, imputacja, wybór zmiennych.

Postanowiono to jak najlepiej odwzorować i na podstawie wybranych modeli oraz wyników ich predykcji wybrać najlepszy zbiór danych do uczenia. Zrezygnowano z metody analizy skupień, gdyż Python nie oferował takiego samego sposobu jaki był użyty w SAS. Dopiero potem skupić się na nim i wytrenować najlepszy model. W przypadku modeli do znalezienia zbioru zastosowano:

- regresję logistyczną dla zbioru z danymi kategorycznymi (bez pomiarów hormonów, które mają braki danych i mogą źle wpływać na model regresyjny) - ustalone parametry: random_state=123, class_weight = 'balanced', max_iter = 1000;
- drzewo decyzyjne - ustalone parametry: criterion='gini', random_state = 123, max_depth = 7, class_weight = 'balanced';
- las losowy - ustalone parametry: random_state = 123, n_estimators = 200, max_depth = 4, max_samples = 0.8, criterion = 'log_loss', class_weight = 'balanced';

- oraz XGBoost - ustalone parametry: random_state = 123, learning_rate = 0.01, booster = 'gbtree', n_estimators = 100, max_depth = 4, subsample = 0.8, scale_pos_weight = 15.33.

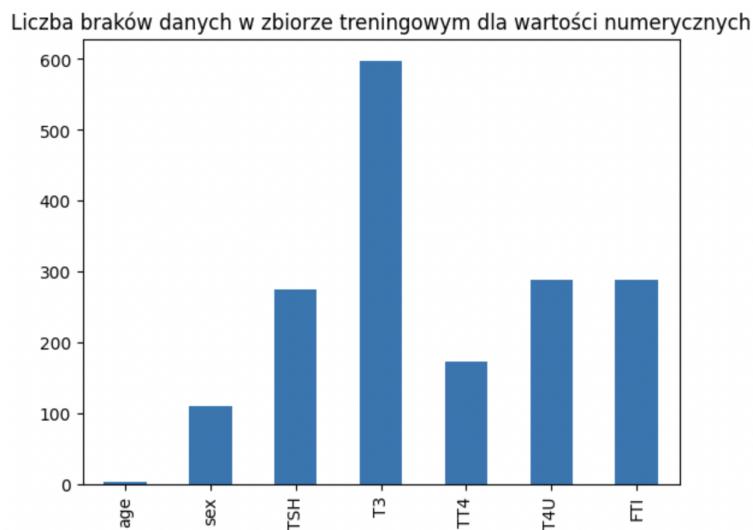
Modele były najbardziej zbliżone do tych wykorzystywanych w SAS. Wszystkie mają parametr, który uwzględnia niezbalansowanie danych, aby wyniki predykcji były jak najbardziej poprawne.

3.3.1. Surowe dane

Pierwszym sposobem było modelowanie na surowych danych, jednak dotyczy to zbioru po zmianach opisanych w sekcji 3.1. Usunięto tylko kolumny, które mają jedną wartość dla wszystkich danych, czyli zmienne TBG_measured i hypopituitary. Przez to, że braki danych nie są imputowane to jedynie model XGBoost zadziałał na takim zbiorze. Wynik modelu na zbiorze testowym wynosił 97.14%, jednak jest możliwe, że jest niepoprawny bądź przeuczony.

3.3.2. Porzucenie kolumn z pomiarami hormonów

Jak poprzednio usunięto kolumny, które mają jedną wartość dla wszystkich danych, czyli zmienne TBG_measured i hypopituitary. Następnie przystąpiono do imputacji braków danych. Rysunek 3.2 przedstawia liczbę braków danych z podziałem na zmienne. Jest to jedna zmienna kategoryczna - płeć oraz zmienne numeryczne - wiek i wartości hormonów. Aby móc stosować różne modele braki danych trzeba było uzupełnić. Wykorzystano w tym celu imputację medianą i średnią jak w SAS.



Rysunek 3.2: Liczba braków danych w zbiorze treningowym z podziałem na zmienne.

Po zastosowaniu imputacji wybrano jedynie kolumny uznane za ważne w teście Chi-kwadrat (opisane w sekcji 3.2.1). Wyniki predykcji dla takiego zbioru były w przedziale 55-65%, co pokazuje jak bardzo się pogorszyły względem „surowych danych”.

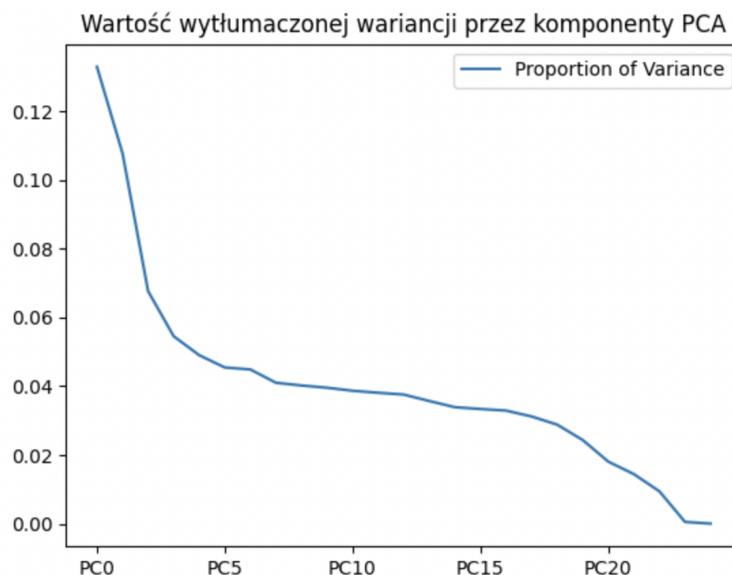
3.3.3. Porzucenie kolumn zawierających w nazwie „measured”

W tym przypadku zbiór, opisany w poprzedniej podsekcji zmniejszono o kolumny z przyrostkiem „measured”. Dodatkowo w analizie w SASie odrzucono kolumnę TT4 ręcznie ze względu na korelacje oraz na to, że całkowita tyroksyna może pogarszać diagnozę przez nieznany jej wpływ w niektórych przypadkach. W aktualnym projekcie sprawdzono oba przypadki - z i bez zmiennej TT4. Różnice w wynikach modeli są minimalne i dotyczą jedynie drzewa oraz regresji. Mimo to odrzucenie TT4 dało tam wyższe wyniki, więc możliwe, że przynosi to korzyści.

Wyniki modelowania były powyżej 90%, przy czym najgorzej wypadła regresja logistyczna a najlepiej las losowy. Mając dobre wyniki predykcji postanowiono sprawdzić czy zmiana imputacji poprawi predykcję. Stąd sprawdzono dodatkowe dwie metody: k najbliższych sąsiadów (k-Nearest Neighbors) oraz imputację funkcją *IterativeImputer* (zmienne numeryczne) + moda (zmienne kategoryczne). Przy porównywaniu imputacji najlepsze okazało się podejście z funkcją *IterativeImputer* dla zmiennych numerycznych i moda dla zmiennych kategorycznych. Dla trzech modeli wynik predykcji przekroczył wówczas 97%.

3.3.4. PCA

Na zbiorze po odrzuceniu kolumn TBG_measured i hypopituitary i po imputacji najlepszą metodą (*IterativeImputer* + moda) dokonano standaryzacji i następnie wykonano PCA. Sprawdzono dwa kryteria wyboru liczby komponentów do PCA. Metoda "łokcia" do wykresu nie wydała się łatwa, gdyż nie ma dobrego miejsca zgięcia (dowód na Rysunku 3.3), dopiero po 23/24 komponentach, dlatego wybrano 18 komponentów, zgodnie z drugim kryterium (Kryterium części wyjaśnionej wariancji).



Rysunek 3.3: Wykres wytłumaczonej wariancji przez odpowiednie komponenty PCA.

Wyniki predykcji dla wszystkich modeli okazały się dobre - powyżej 80%, jednak pogorszyły się

względem uzyskanych na zbiorze z porzuconymi kolumnami „measured”.

3.3.5. Najlepszy wybrany zbiór

Na podstawie wyników predykcji zdecydowano się na zbiór:

- po usunięciu kolumn z jedną wartością,
- z imputacją IterativeImputer + moda,
- z wybraniem najważniejszych zmiennych na podstawie testu Chi-kwadrat,
- i po odrzuceniu kolumn „measured” wraz z kolumną TT4.

Wyniki na zbiorze testowym przedstawiono w Tabeli 3.1.

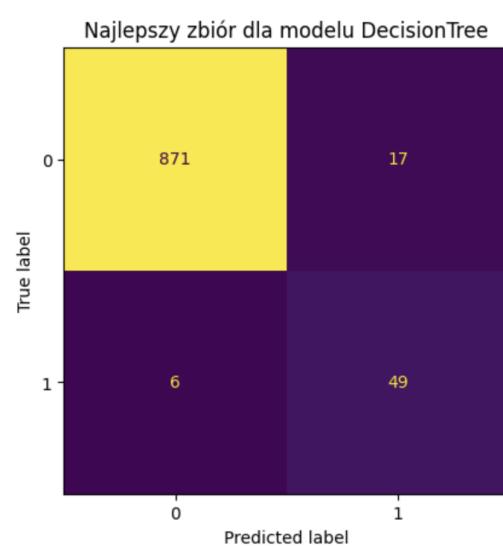
Model	Wynik predykcji (score) [%]
XGBoost	97.14
Las losowy	97.35
Regresja logistyczna	89.82
Drzewo decyzyjne	97.56

Tabela 3.1: Wyniki predykcji modeli na zbiorze testowym.

Drzewo decyzyjne ma najlepszy wynik w tym przypadku. Sprawdzono też dla niego macierz pomyłek, aby zobaczyć, jak przejawia się taki wynik. Rysunek 3.4 przedstawia macierz dla drzewa decyzyjnego i wybranego zbioru. Po porównaniu go z innymi modelami, rzeczywiście ten model popełnił najmniej błędów. W przypadku danych medycznych interesuje nas błąd II rodzaju, więc przypadki, gdzie chory pacjent jest zaklasyfikowany jako zdrowy. Model drzewa zaklasyfikował błędnie tak 6 pacjentów.

Wykorzystując średnią geometryczną (GM) predykcja modelu drzewa wynosi 93.5%, przy czym drugi najlepszy model - las losowy - miał $GM = 92.47\%$. Nawet model automatyczny nie osiągnął takiego wyniku (miał on 90.68%). Można powiedzieć, że jest to dobrze wytrenowany model dla tych danych.

3.3. WYBÓR NAJLEPSZEGO ZBIORU DO MODELOWANIA



Rysunek 3.4: Macierz pomyłek dla modelu drzewa decyzyjnego i wybranego najlepszego zbioru.

4. Ostateczne modele wraz z uzyskaną predykcją

Po przetworzeniu odpowiednio danych przystąpiono do znalezienia najlepszego klasyfikatora. Rzeczywiście wybrany model, który do tej pory był najlepszy to drzewo decyzyjne. Jednak, aby znaleźć najlepszy model wypadałoby wybrać odpowiednie hiperparametry w jakiś sensowny sposób a nie na ślepo jak do tej pory. Stąd rozważano teraz jak poprawić modele, aby były jeszcze lepsze niż do tej pory.

Ten rozdział opisuje podjęte kroki i wyniki podczas szukania najlepszego modelu dla przetworzonych danych o chorobach tarczycy.

4.1. Wpływ transformacji

Model regresji do tej pory miał najgorsze wyniki predykcji, więc postanowiono zobaczyć, czy transformacja zmiennych - standaryzacja czy normalizacja poprawi jakość predykcji. Dla modeli drzewiastych standaryzacja i inne opcje transformacji nie mają wpływu i nie powinny, co udowodnił test na modelu lasu losowego. Natomiast w przypadku regresji logistycznej standaryzacja daje najlepsze wyniki a normalizacja gorsze. Stąd uznano, że by ją stosować do zbioru dla tego modelu.

4.2. Klasyfikatory

Najlepszy model dla przygotowanego zbioru zdecydowano wybrać spośród klasyfikatorów takich jak:

- las losowy,
- k najbliższych sąsiadów,
- regresji logistycznej,
- XGBoost,
- oraz gradient boosting.

Dla każdego modelu wykonano strojenie hiperparametrów zarówno za pomocą 'siatki poszukiwań' jak i losowo, biorąc pod uwagę wskazane wartości. Dodatkowo dokonano transformacji danych (standaryzacja, normalizacja), aby zdecydować, czy poprawi to wyniki klasyfikacji.

4.3. WYBÓR NAJLEPSZEGO MODELU

Wszystkie modele poza k najbliższych sąsiadów i gradient boostingiem miały daną opcję brania pod uwagę niezbalansowania zbioru. Dodatkowo ustawiono ziarno losowości na 123, w tych modelach, które przyjmowały taką możliwość. Strojenie parametrów wykorzystywało 5-krotną kroswalidację, a ocena modeli była dokonywana za pomocą średniej geometrycznej. Uczenie wykonano na zbiorze treningowym. Standaryzacja poprawiała wynik dla k najbliższych sąsiadów oraz regresji logistycznej. Inne klasyfikatory nie potrzebowaly transformacji danych. Gradient boosting i regresja logistyczna miały błędy podczas obliczeń - pierwszy z nich nie uzyskał dobrego wyniku predykcji (zwrócono zera) dla losowego przeszukiwania, a drugi osiągał maksymalną liczbę iteracji (ostrzeżenia się o tym pojawiły). Pomijając te problemy wybrano najlepsze kombinacje hiperparametrów dla każdego klasyfikatora i wśród nich najlepszy okazał się odpowiednio wytrenowany model XGBoost.

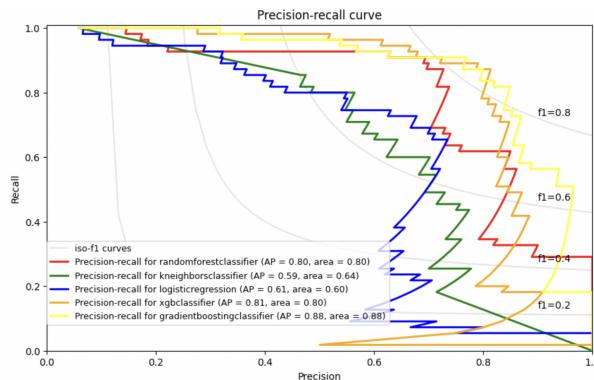
4.3. Wybór najlepszego modelu

Po wykonanym strojeniu hiperparametrów sprawdzono jakość predykcji na zbiorze testowym dla wytrenowanych modeli o wybranych parametrach i transformacji danych. Tabela 4.1 przedstawia wyniki przy użyciu średniej geometrycznej. Dwa modele przekroczyły wartość 90%. Są to modele XGBoost i lasu losowego, gdzie lepszym z nich jest XGBoost.

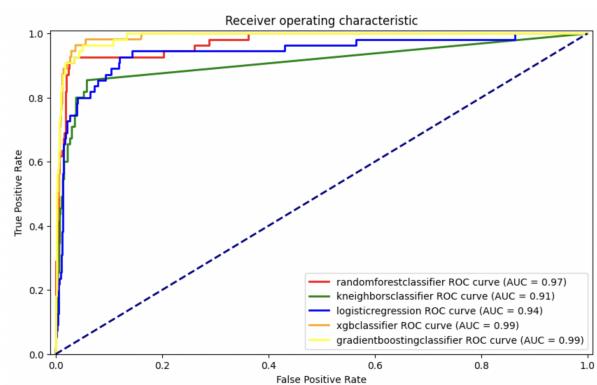
Model	Wynik predykcji (GM) [%]
XGBoost	93.53
Las losowy	92.42
Regresja logistyczna	88.67
Gradient boosting	88.02
k najbliższych sąsiadów	70.87

Tabela 4.1: Wyniki predykcji, przy użyciu średniej geometrycznej (GM), dla wytrenowanych modeli na zbiorze testowym.

Wykonano też wykresy krzywej ROC i precision-recall dla tych modeli, aby nie opierać się na jednej metryce. Rysunek 4.1 przedstawia wykres krzywych precision-recall. Najlepiej wygląda predykcja dla modelu gradient boostingu, a następnie dla XGBoost i lasu losowego. Na Rysunku 4.2 można zobaczyć wykres krzywej ROC. Tutaj najlepiej wypadł XGBoost i gradient boosting. Mają one nawet taką samą wartość AUC. Modele niedrzewiaste, wypadły gorzej niż wspomniane trzy modele oparte na drzewach. Gradient boosting mimo, że dla średniej geometrycznej nie miał najlepszego wyniku tutaj pokazuje się jako lepszy bądź porównywalnie dobry model niż XGBoost.



Rysunek 4.1: Wykres krzywej precision-recall dla wytrenowanych modeli i ich wyników na zbiorze testowym. W legendzie można zobaczyć wartość miary pola pod krzywą i AP (average precision).



Rysunek 4.2: Wykres krzywej ROC dla wytrenowanych modeli i ich wyników na zbiorze testowym. W legendzie można zobaczyć wartość miary AUC (będącej polem pod krzywą ROC).

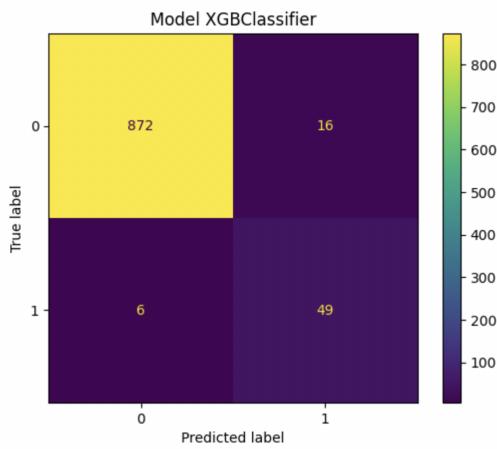
Zdecydowano się porównać ostatecznie trzy modele, które mają szansę o wybranie jako najlepszy - XGBoost, gradient boosting oraz las losowy - wraz z modelem drzewa decyzyjnego wytrenowanego ręcznie przy ustalaniu najlepszego przetworzonego zbioru do uczenia. Wykorzystano do tego metryki takie jak: średnia geometryczna (GM), F1, precision, recall, zbalansowaną skuteczność (balanced accuracy, BACC), pole pod krzywą ROC (AUC). W Tabeli 4.2 przedstawiono wyniki jakości predykcji dla wybranych modeli. Pogrubiono najlepszy wynik dla danej metryki. Wówczas można zauważyć, że najlepiej wypadł model XGBoost.

Model	GM	F1	Precision	Recall	BACC	AUC
XGBoost	93.53	81.67	75.38	89.09	93.64	93.64
Gradient boosting	88.02	81.13	84.31	78.18	88.64	88.64
Las losowy	92.42	78.69	71.64	87.27	92.57	92.57
Drzewo decyzyjne	93.48	80.99	74.24	89.09	93.59	93.59

Tabela 4.2: Jakość predykcji w procentach dla wybranych modeli na zbiorze testowym. Przedstawione metryki to: średnia geometryczna (GM), F1, precision, recall, zbalansowana skuteczność (balanced accuracy, BACC) oraz pole pod krzywą ROC (AUC).

Sprawdzono też dla najlepszego modelu (XGBoost) macierz pomyłek, aby zobaczyć jak przejawia się taki wynik. Rysunek 4.3 przedstawia tę macierz. Porównując ją z macierzą dla drzewa decyzyjnego (Rysunek 3.4), można zauważyć, że błąd II rodzaju się nie zmienił, ale błąd I rodzaju jest mniejszy. Stąd widać, że chociaż nie zmniejszył się interesujący nas błąd, to model XGBoost jest lepszy.

4.4. XAI



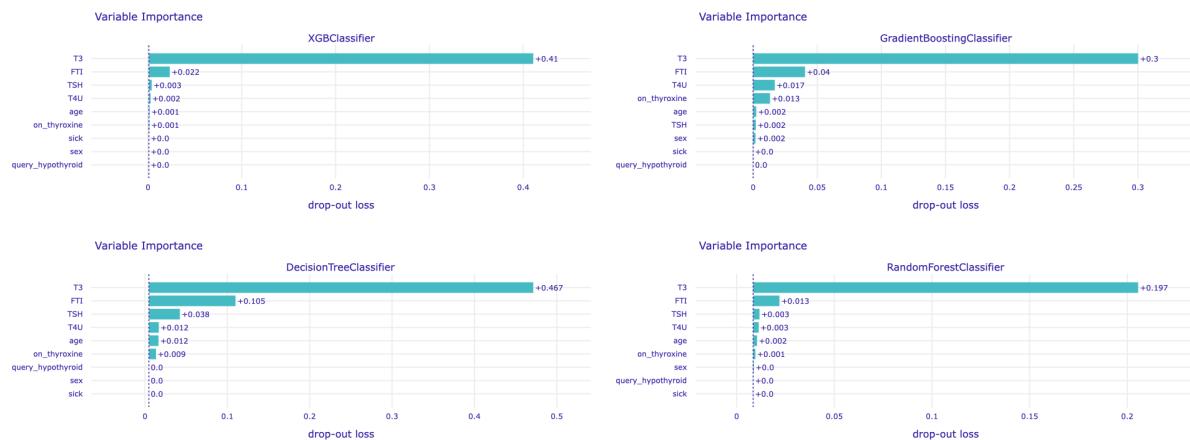
Rysunek 4.3: Macierz pomyłek dla modelu XGBoost i wybranego najlepszego zbioru (zbiór testowy).

4.4. XAI

Wyuczzone modele poddano wyjaśnialnemu uczeniu maszynowemu (XAI), aby zobaczyć jaki wpływ mają zmienne na wynik predykcji.

Dla najlepszego modelu - XGBoost - zbadano poprawnie i błędne przewidziane obserwacje. Zauważono duży wpływ zmiennych T3 oraz FTI, stąd, jeśli tylko ta zmienna jest anomalią dla pacjenta to możliwa jest błędna predykcja, gdyż modele bardzo opierają swoją klasyfikację na tych zmiennych.

Sprawdzono ważność zmiennych w modelu poprzez XAI dla różnych modeli. Wykresy ważności dla czterech najlepszych modeli przedstawia Rysunek 4.4. W przypadku innych modelu ważność zmiennych się różni. Jednak pierwsze dwie najważniejsze zmienne dla wszystkich sprawdzanych modeli były takie same i są to już wspomniane - T3 i FTI. W projekcie w SAS również zmienna T3 była najważniejszą w predykcji.



Rysunek 4.4: Wykresy ważności zmiennych dla wyuczonych modeli XGBoost, Gradient boostingu, drzewa decyzyjnego i lasu losowego.

Ważność zmiennej T3 wyjaśnia jej rozkład, który pokazano na Rysunku 2.5. Rozkład dla zdrowych

4. OSTATECZNE MODELE WRAZ Z UZYSKANĄ PREDYKCJĄ

i chorych pacjentów różni się od siebie, stąd wartość T3 może ułatwiać rozpoznanie choroby tarczycy. W przypadku zmiennej FTI nie udało się znaleźć prostego wytłumaczenia na jej ważność.

5. Podsumowanie

Projekt został zrealizowany bez większych problemów. Udało się przeprowadzić analizę zbioru wraz z EDA oraz modelowanie. Ostateczny zbiór do klasyfikacji nie był poddany standaryzacji czy redukcji wymiarowości, bo nie poprawiało to jakości wyników. Jednak zdecydowano się na selekcję zmiennych i imputację, a także pozbycie się zmiennych odstających. Wytrenowane modele osiągają na zbiorze testowym bardzo dobre wyniki, gdzie najlepszy z nich ma średnią geometryczną predykcji równą 93.53% i jedynie 6 pacjentów na 55 chorych uznał za zdrowych (błąd II rodzaju).

Porównując analizę w SAS to EDA i charakterystyka danych była tam prostsza w przypadku wykresów i tabel, jednak skomplikowane rzeczy łatwiej było wykonać w Python. Najlepszy zbiór do modelowania wybrano w taki sam sposób, jednak wybrane zmienne różnią się ze względu na inny podział danych. W przypadku uczenia modeli SAS nie sprzyja użytkownikom - brakuje tam metryk do porównania wyników oraz wybór modeli nie jest zbyt szeroki. Stąd Python miał większe możliwości do tego zadania.

Bibliografia

Dua, Dheeru i Casey Graff (2017). *UCI Machine Learning Repository*. URL: <http://archive.ics.uci.edu/ml>.

Quinlan, J. Ross (1987). *Thyroid disease records supplied*. URL: <https://archive.ics.uci.edu/ml/datasets/thyroid+disease>.