RAPORT KOŃCOWY PROJEKTU ANALITYCZNEGO W SAS

Autor: Paulina Przybyłek

24 STYCZNIA 2022

RAPORT KOŃCOWY PROJEKTU ANALITYCZNEGO W SAS

Spis treści

RAPORT KOŃCOWY PROJEKTU ANALITYCZNEGO W SAS

1. Wstęp.

1.1. Cel projektu.

Celem zadania jest zapoznanie się z narzędziami SAS podczas realizacji projektu dotyczącego metod przetwarzania danych, eksploracji oraz modelowania predykcyjnego z wykorzystaniem technik uczenia maszynowego. Projekt zakłada postawienie problemu badawczego oraz rozwiązanie go przy użyciu wybranego zbioru danych oraz środowiska firmy SAS.

1.2. Problem badawczy.

Metody uczenia maszynowego stają się coraz popularniejsze w użyciu i zastępują zadania niegdyś wykonywane przez ludzi. Dotyczy to wielu różnych dziedzin życia, czy to rekomendacji produktów czy bankowości. Uczenie maszynowe pozwala wykorzystać zawarte informacje do celów pomagających różnym osobom czy firmom w poprawie jakości usługi. Zaczyna się również wprowadzać te metody do medycyny.

Ludzie, póki co nadal chorują i chorować będą. Różne technologie są wykorzystywane do przeprowadzania badań i uzyskiwania skuteczniejszych wyników czy leczenia. Metody uczenia maszynowego mogą wspomóc lekarzy podczas stawiania diagnozy pacjentom. Wówczas lekarz szybciej mógłby wprowadzić leczenie czy skierować danego pacjenta na odpowiednie badania, wzorując się na predykcji uzyskanej za pomocą tych metod. Problem polega na tym czy można wykorzystać uczenie maszynowe do tego celu? Jak jest to skuteczne? Czy może ono zastąpić diagnozę specjalisty i wspomóc pacjentów? Decyzje dotyczące życia drugiej osoby są naprawdę trudne i ważne. W tym projekcie została przeprowadzona analiza zbioru pacjentów, gdzie występują osoby zdrowe oraz posiadające chorobę tarczycy. Sprawdzono czy metody modelowania predykcyjnego mogą być używane do diagnozy choroby tarczycy u pacjentów.

2. Opis rozwiązania.

2.1. Zbiór danych.

Wybrany zbiór danych pochodzi ze strony UCI Machine Learning Repository [1] i zawiera dane dotyczące chorób tarczycy dostarczone przez Instytut Garavan i J. Ross Quinlana z Instytutu Nowej Południowej Walii w Sydney w Australii. Rekordy z badaniami pacjentów zostały zebrane w roku 1987.

Na stronie UCI dotyczącej wybranych danych znajduje się kilka różnych zbiorów o różnych chorobach tarczycy. Zbiory są jednak stare i według autora w większości zostały zanieczyszczone, więc wykorzystanie ich do postawionego problemu badawczego mogłoby mijać się z celem. Dlatego też wykorzystano jedyny zbiór, który został oczyszczony i powinien zawierać wiarygodne informacje. Jest to zbiór *sick*. W projekcie pobrano ten zbiór ze strony DataHub [2], gdzie zbiór ten występuje w formacie .csv i było łatwiej go pobrać oraz zaimportować do środowiska firmy SAS.

Zbiór *sick* zawiera 29 atrybutów, w tym 23 dyskretne i 7 numerycznych, oraz binarną zmienną celu, określającą chorobę tarczycy lub jej brak u pacjenta. Atrybuty zawierają informacje o pacjentach - wywiad z nim oraz wyniki badań. Liczba rekordów wynosi 3772, przy czym wśród nich jest 5.4% brakujących wartości.

2.2. Plan projektu oraz wykorzystane technologie.

W Tabeli 1 przedstawiono plan postępowania wraz z wykorzystanymi narzędziami SAS przy realizacji projektu.

Tab. 1 Plan postępowania przy realizacji projektu.

Nr	Zadanie	Wykorzystane technologie
1	Import danych	SAS Enterprise Guide
2	Zapoznanie się z danymi i sprawdzenie ich poprawności	SAS Enterprise Guide
3	Wstępne przetwarzanie danych	SAS Enterprise Guide
4	Wstępna eksploracja danych	SAS Enterprise Guide
5	Podział zbioru na treningowy, testowy i walidacyjny	SAS Enterprise Miner
6	Zaawansowana eksploracja danych na zbiorze treningowym	SAS Enterprise Miner
7	Analiza zbioru i jego atrybutów	SAS Enterprise Guide, SAS Enterprise Miner oraz dostępne źródła internetowe dotyczące informacji o atrybutach zbioru

8	Przetwarzanie danych do modelowania	SAS Enterprise Miner
9	Modelowanie i ocena jakości modeli	SAS Enterprise Miner

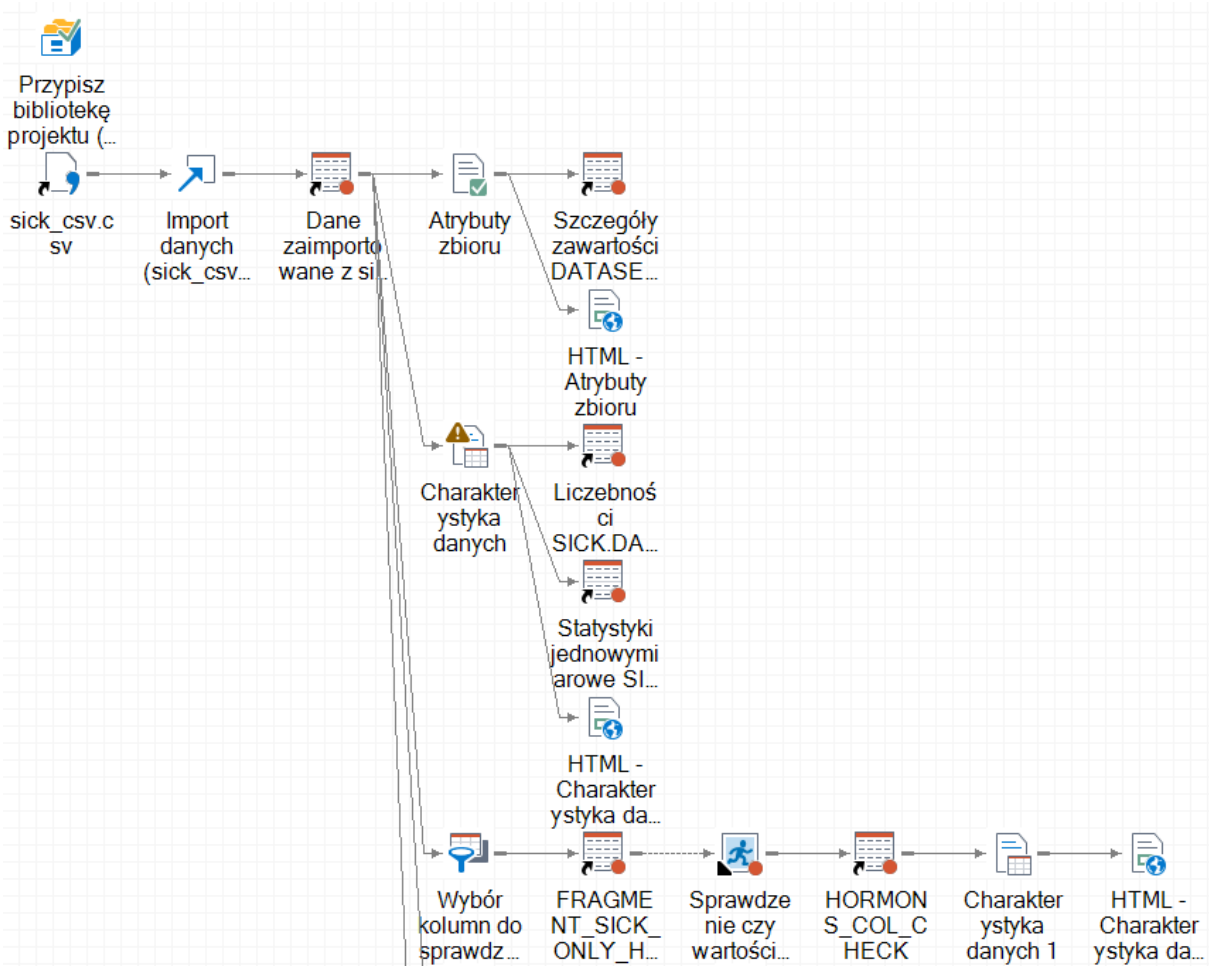
Szczegółowy opis przeprowadzonych zadań oraz wyciągnięte z nich wnioski zawarto w kolejnych rozdziałach raportu.

3. Przygotowanie oraz EDA zbioru danych.

3.1. Import i zapoznanie się z danymi.

Pobrany zbiór danych był w formacie .csv, jednak do pracy ze środowiskiem SAS należało go przekształcić w zbiór sasowy. SAS Enterprise Guide umożliwia wczytanie pliku .csv i zaimportowanie danych do odpowiedniej formy.

Na Rysunku 1 przedstawiono proces importu danych po uprzednim stworzeniu biblioteki dla tego projektu oraz wykonaną krótką analizę powstałego zbioru. Dane zaimportowane zostały przeanalizowane w celu sprawdzenia poprawności ich importu i dokonania ewentualnej korekty np. typów kolumn.



Rys. 1 Import i analiza wczytanego zbioru danych w programie SAS Enterprise Guide.

Na zaimportowanym zbiorze dokonano trzech analiz: sprawdzenia atrybutów zbioru, charakterystyki danych oraz czy kolumny zawierające pomiary liczbowe a kolumny o takich samych nazwach z dopiskiem “measured” pokrywają się.

Atrybuty zbioru

Wczytany zbiór z pliku zawierał taką samą liczbę kolumn i atrybutów jak podana na stronie, z której została pobrana .csv, więc żaden rekord czy wiersz nie został pominięty. Jednakże okazało się, że jedna z kolumn jest całkowicie pusta (posiada same braki danych) przez co jej format trzeba było ustawić ręcznie, gdyż SAS nie potrafił się go domyślić. Dzięki wykonaniu zadania *"Atrybuty zbioru"* można było poznać informacje o atrybutach, zbiorze jak i o mechanizmach zależnych od środowiska.

W Tabeli 2 pokazano wszystkie automatyczne jak i ustawione informacje o wczytanych atrybutach.

Tab. 2 Informacje o atrybutach zbioru *sick*.

Liczba zmiennych	Nazwa	Typ	Format	Etykieta	Długość
1	age	Numeryczny	BEST		8
2	sex	Znakowy	$CHAR		1
3	on_thyroxine	Znakowy	$CHAR		1
4	query_on_thyroxine	Znakowy	$CHAR		1
5	on_antithyroid_medication	Znakowy	$CHAR		1
6	sick	Znakowy	$CHAR		1
7	pregnant	Znakowy	$CHAR		1
8	thyroid_surgery	Znakowy	$CHAR		1
9	I131_treatment	Znakowy	$CHAR		1
10	query_hypothyroid	Znakowy	$CHAR		1
11	query_hyperthyroid	Znakowy	$CHAR		1
12	lithium	Znakowy	$CHAR		1
13	goitre	Znakowy	$CHAR		1
14	tumor	Znakowy	$CHAR		1
15	hypopituitary	Znakowy	$CHAR		1
16	psych	Znakowy	$CHAR		1
17	TSH_measured	Znakowy	$CHAR		1
18	TSH	Numeryczny	BEST		8
19	T3_measured	Znakowy	$CHAR		1
20	T3	Numeryczny	BEST		8
21	TT4_measured	Znakowy	$CHAR		1
22	TT4	Numeryczny	BEST		8
23	T4U_measured	Znakowy	$CHAR		1
24	T4U	Numeryczny	BEST		8
25	FTI_measured	Znakowy	$CHAR		1
26	FTI	Numeryczny	BEST		8
27	TBG_measured	Znakowy	$CHAR		1
28	TBG	Numeryczny	BEST		8
29	referral_source	Znakowy	$CHAR		5
30	thyroid_disease	Znakowy	$CHAR	target	8

Charakterystyka danych

Na surowo wczytanych danych charakterystyka zbioru nie powiodła się bezbłędnie - pojawiło się ostrzeżenie, ponieważ jedna kolumna jest cała pusta i nie da się dla niej uzyskać żadnych informacji.

Charakterystyka zmiennych kategorycznych zliczyła wystąpienia każdej wartości w zmiennej i przedstawiła wyniki w postaci liczby oraz w procentach. Większość zmiennych kategorycznych zawierała jedynie “f” i “t”, oznaczające odpowiednio false i true. Jedynie zmienna celu, zmienna *sex* oraz zmienna *referal_source* zawierały inne wartości.

W Tabelach 3, 4 oraz 5 przedstawiono charakterystyki tych zmiennych.

Tab. 3 Charakterystyka zmiennej celu.

Variable	Label	Value	Frequency Count	Percent of Total Frequency
thyroid_disease	target	negative	3541	93.876
		sick	231	6.124

Okazuje się, że tylko nieco ponad 6% obserwacji zawiera dane o chorych pacjentach. Zbiór jest więc mocno niezbalansowany.

Tab. 4 Charakterystyka zmiennej sex.

Variable	Label	Value	Frequency Count	Percent of Total Frequency
sex		F	2480	65.748
		M	1142	30.276
		Brak	150	3.977

W przypadku zmiennej sex, czyli dotyczącej płci, obserwujemy braki danych w prawie 4% obserwacji oraz sporą przewagę kobiet nad mężczyznami wśród pacjentów z podejrzeniem choroby tarczycy.

Tab. 5 Charakterystyka zmiennej referal_source.

Variable	Label	Value	Frequency Count	Percent of Total Frequency
referral_source		other	2201	58.351
		SVI	1034	27.413
		SVHC	386	10.233
		STMW	112	2.969
		SVHD	39	1.034

Zmienna referal_source zawiera aż 5 różnych wartości, zmienna oznacza źródło skierowania pacjenta.

Charakterystyka zmiennych numerycznych zawiera podstawowe informacje z funkcji agregujących jak suma, minimum, maksimum, średnia, mediana, odchylenie standardowe, a także liczbę wartości jak i liczbę braków danych. Okazuje się, że wszystkie zmienne zawierają jakieś braki danych. Dodatkowo po dokładnym przyjrzeniu się tym informacjom agregującym można zauważyć, że maksymalny wiek wynosi 455 lat.

W Tabeli 6 przedstawiono statystyki dla zmiennej wieku (*age*).

Tab. 6 Charakterystyka zmiennej age.

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
age		3771	1	195096.00	1.000	51.736	54.00	455.00	0.32707

Zadanie “Charakterystyka danych” pokazuje również liczebność wartości w atrybutach na wykresach. Nie zostały one załączone w tym rozdziale.

Sprawdzenie kolumn numerycznych i kategorycznych o tych samych nazwach z dopiskiem “measured”

Rozważane atrybuty numeryczne zawierają liczbę bądź brak danych. Jest to 6 zmiennych, mianowicie:

- TSH,
- T3,
- TT4,
- T4U,
- FTI,
- TBG.

Zmienne kategoryczne z odpowiadającymi nazwami zawierają jedynie wartość “f” lub “t”.

Okazuje się, że zmienne te pokrywają się w 100%, gdy jest jakaś wartość liczbowa w zmiennej numerycznej to mamy wartość “t” w odpowiadającej jej zmiennej, a gdy brak danych to “f”.

3.2. Wstępne przetwarzanie danych.

Po wstępnej analizie zbioru postanowiono usunąć pustą kolumnę TBG, ponieważ nie wnosi ona żadnej informacji oraz dwa wiersze ze zmiennej age - z brakiem danych i wartością 455 (najstarszy człowiek miał 123 lata, więc jest to ewidentnie zła wartość). Wówczas po wywołaniu “Charakterystyki danych” nie pojawia się ostrzeżenie, a zadanie wykonuje się bez żadnych problemów.

Uznano jednak, że zamiast usuwać wiek równy 455, co najprawdopodobniej jest błędem podczas wpisywania do bazy wieku 45 bądź 55, można go zamienić na wartość brakującą, żeby nie tracić potrzebnych informacji, gdzie zbiór i tak jest mały.

Ostatecznie wybrano zbiór z usuniętą kolumną TBG i zamienieniem wartości 455 w kolumnie wieku na wartość brakującą.

3.3. Wstępna eksploracja danych w SAS Enterprise Guide.

Po przekształceniach i otrzymaniu oczekiwanej formy zbioru dokonano jego analizy. W tym celu w SAS Enterprise Guide zastosowano na zbiorze:

- charakterystykę danych tylko dla zmiennych kategorycznych i wyłącznie biorąc pod uwagę pacjentów chorych na tarczycę,
- charakterystykę wszystkich danych,
- statystyki agregujące ze względu na płeć i zmienną celu,
- statystyki agregujące ze względu na zmienną celu,
- analizę rozkładu,
- korelacje zmiennych numerycznych ze zmienną celu.

Charakterystyki całościowych danych dokonano, aby sprawdzić poprawność przekształconego zbioru oraz jak przejawiają się teraz charakterystyki poszczególnych zmiennych. Natomiast charakterystyki dla zmiennych kategorycznych i chorych pacjentów uzyskano dla większej ilości informacji o objawach choroby tarczycy.

Statystyki agregujące, analizę rozkładu oraz korelacje zastosowano, żeby poznać lepiej zmienne i ich oddziaływanie na siebie nawzajem oraz na zmienną celu.

Poniżej przedstawiono krótki opis wykonanej analizy (z wyłączeniem analizy rozkładu, która nie wniosła nowych interesujących informacji).

Charakterystyka danych dla zmiennych kategorycznych i pacjentów chorych na tarczycę

Z interesujących informacji można wypisać, że osoby chore na tarczycę w posiadanych danych:

- nie mają historii operacji tarczycy (100% pacjentów),
- w 58.874% to kobiety,
- nie są w ciąży (100% pacjentów),
- tylko dwie osoby mają guza (ciekawostka!).

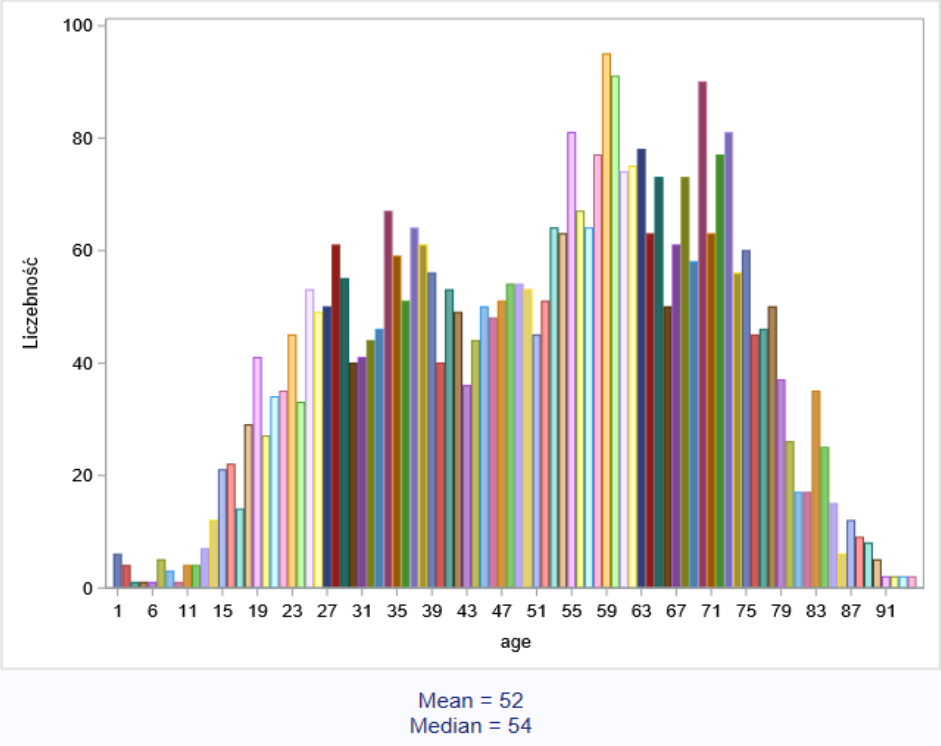
Można wnioskować, że to kobiety częściej chorują a po operacji tarczycy choroba (jeśli była) to nie wraca. Możliwe też, że choroba tarczycy powoduje problemy z zajściem w ciążę. Jednak są to na ten moment spekulacje na podstawie jedynie uzyskanych wyników.

Charakterystyka danych

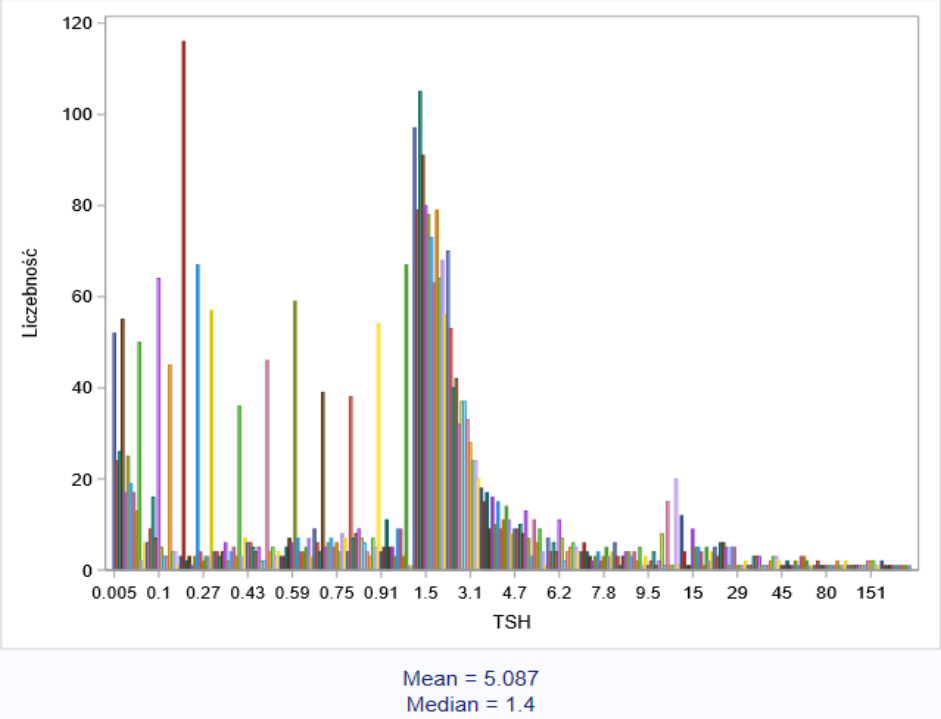
Po tym etapie można zauważyć, że zmienne kategoryczne mają dużo rzadziej wartości “t” niż “f”. Dodatkowo ciekawe są informacje z rozkładów zmiennych numerycznych. Większość z nich ma rozkłady zbliżone do normalnego, jedynie odbiegają rozkłady zmiennej age oraz zmiennej TSH.

Na Rysunku 2 przedstawiono rozkład zmiennej age. Średnia z wieku wynosi 52 lata a mediana 54 lata. Można zauważyć, że pacjenci są z każdego przedziału wiekowego, przy czym największe ich skupisko oscyluje przy 60 latach, gdzie jest największy pik. Drugi pik jest również koło 35, może to właśnie kobiety, które nie mogą zajść w ciążę przychodzą się zbadać na tarczycę?

Na Rysunku 3 pokazano rozkład zmiennej TSH. Widać, że przyjmowane wartości są z bardzo szerokiego zakresu, największe wartości są między 1-3. TSH, czyli tyreotropina, przyjmuje wartości między 0 a nawet 400, więc taki duży zakres jest normalny. Oczywiście norma wynosi od 0.32 do 4 albo 5 mlU/l (nie wliczając dzieci i kobiet w ciąży).



Rys. 2 Rozkład zmiennej age (wieku).

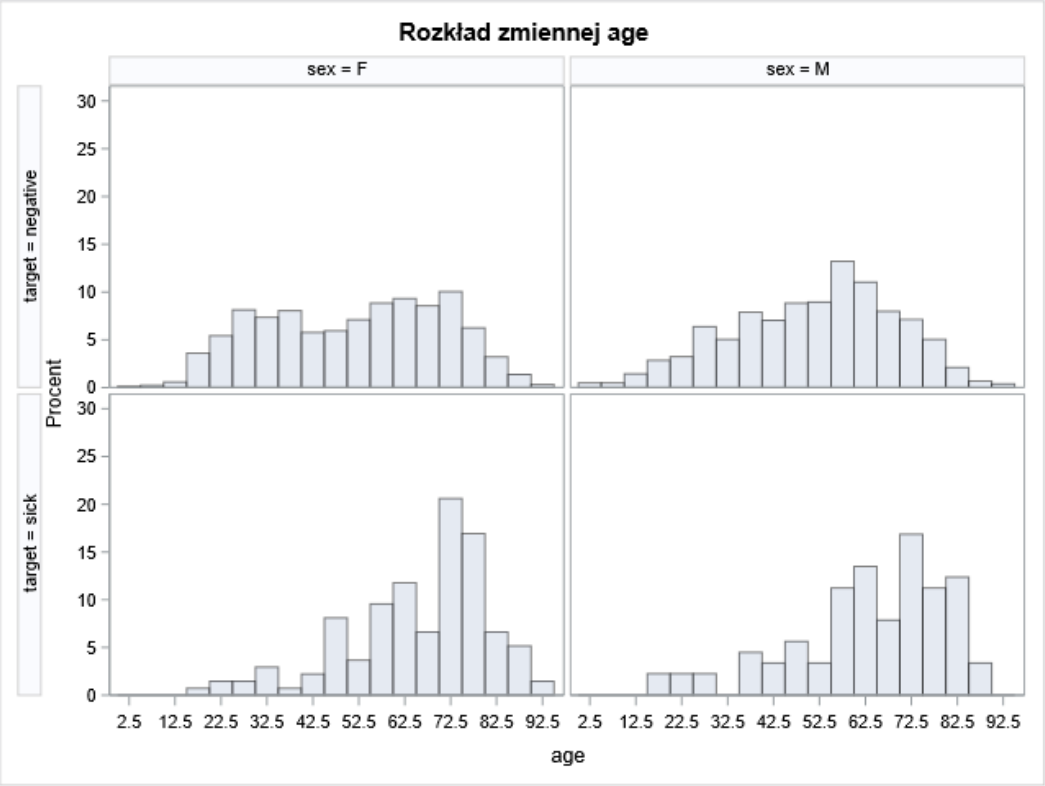


Rys. 3 Rozkład zmiennej TSH.

Statystyki agregujące ze względu na płeć i zmienną celu

Statystyki te pokazały różnice między kobietami a mężczyznami względem różnych badań czy cech w zależności od tego czy są chorzy czy zdrowi. Z interesujących informacji można wypisać, że:

- z chorych pacjentów kobiety miały trochę większe wartości FTI czy TT4,
- mediana wieku chorych kobiet jest większa,
- rozkłady zmiennych numerycznych (pomijając wiek) są do siebie bardzo zbliżone niezależnie od płci czy choroby
- rozkład wieku posiada różnice w zależności od choroby - osoby około 70 r.ż. chorują częściej (Rysunek 4).



Rys. 4 Rozkład zmiennej age z podziałem na płeć i wartość zmiennej celu.

Statystyki agregujące ze względu na zmienną celu

W Tabeli 7 pokazano statystyki agregujące dla zmiennych numerycznych. Można z niej wyczytać różnice statystyk w zależności od choroby bądź jej braku.

Ważną obserwacją jest to, że braki danych pojawiają się głównie wśród osób zdrowych, co znaczy, że nie wykonanie badania hormonów sugeruje, że lekarz zna diagnozę i pacjent jest zdrowy, dlatego nie powinno się tych danych imputować w późniejszych krokach.

Tab. 7 Statystyki agregujące dla zmiennych numerycznych.

target	N obs.	Zmienna	Średnia	Odch. std.	Minimum	Maksimum	Moda	N	N braków
negative	3541	age	50.8265047	18.8697708	1.0000000	94.0000000	59.0000000	3539	2
		TSH	5.1462713	25.1420812	0.0050000	530.0000000	0.2000000	3174	367
		T3	2.1056180	0.7830901	0.0500000	10.6000000	2.0000000	2775	766
		TT4	109.4171601	35.6262480	2.0000000	430.0000000	93.0000000	3310	231
		T4U	1.0062758	0.1941236	0.2500000	2.3200000	0.9900000	3162	379
		FTI	110.4385904	33.0104169	2.0000000	395.0000000	100.0000000	3164	377
sick	231	age	63.9220779	16.3143838	16.0000000	93.0000000	73.0000000	231	0
		TSH	4.2620087	13.2011923	0.0050000	143.0000000	1.3000000	229	2
		T3	0.8923246	0.4405597	0.0500000	2.6000000	1.0000000	228	3
		TT4	92.5887446	31.3808257	19.0000000	192.0000000	73.0000000	231	0
		T4U	0.8351121	0.1352171	0.4600000	1.1400000	0.8200000	223	8
		FTI	110.9103139	34.2682980	17.0000000	219.0000000	93.0000000	223	8

Korelacje zmiennych numerycznych ze zmienną celu

Zmienną najbardziej skorelowaną ze zmienną celu jest T3, czyli trijodotyronina, jeden z hormonów wpływających na poprawne działanie tarczycy.

W Tabelach 8 i 9 przedstawiono wynik korelacji Pearsona i Spearmana.

Tab. 8 Współczynniki korelacji Pearsona.

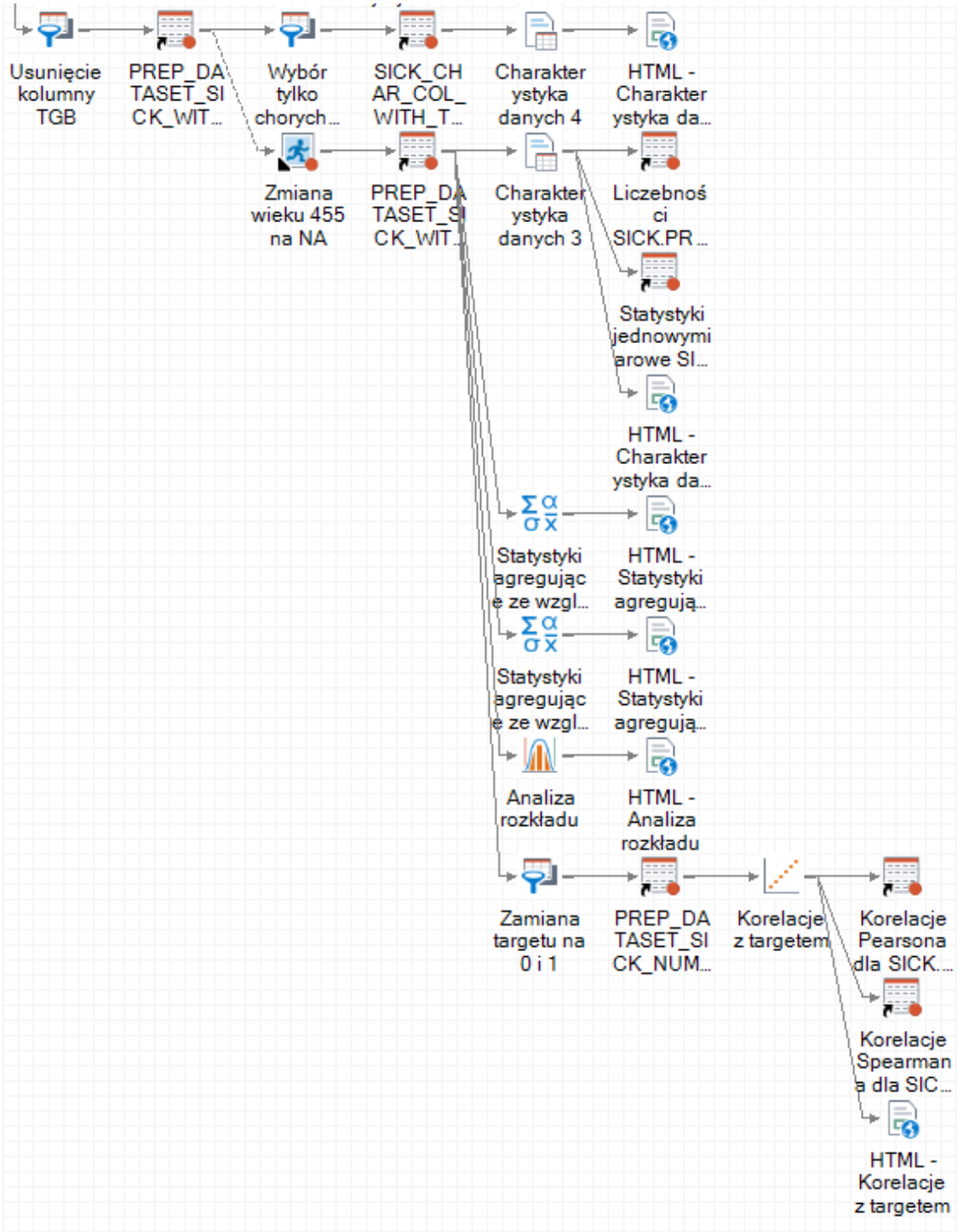
Współczynniki korelacji Pearsona Prawd. > |r| przy H0: rho=0 Liczba obserwacji						
	TSH	T3	TT4	T4U	FTI	age
target diagnostic class	-0.00904 0.5983 3403	-0.38846 <.0001 3003	-0.11673 <.0001 3541	-0.21727 <.0001 3385	0.00354 0.8370 3387	0.16547 <.0001 3770

Tab. 9 Współczynniki korelacji Spearmana.

Współczynniki korelacji Spearmana Prawd. > |r| przy H0: rho=0 Liczba obserwacji						
	TSH	T3	TT4	T4U	FTI	age
target diagnostic class	0.02418 0.1585 3403	-0.40239 <.0001 3003	-0.13938 <.0001 3541	-0.24108 <.0001 3385	0.00690 0.6880 3387	0.16881 <.0001 3770

Podsumowanie eksploracji

Na Rysunku 5 przedstawiono całościową przeprowadzoną analizę przekształconego zbioru danych w programie SAS Enterprise Guide.



Rys. 5 Proces eksploracji danych na przekształconym zbiorze w SAS Enterprise Guide.

3.4. Podział zbioru.

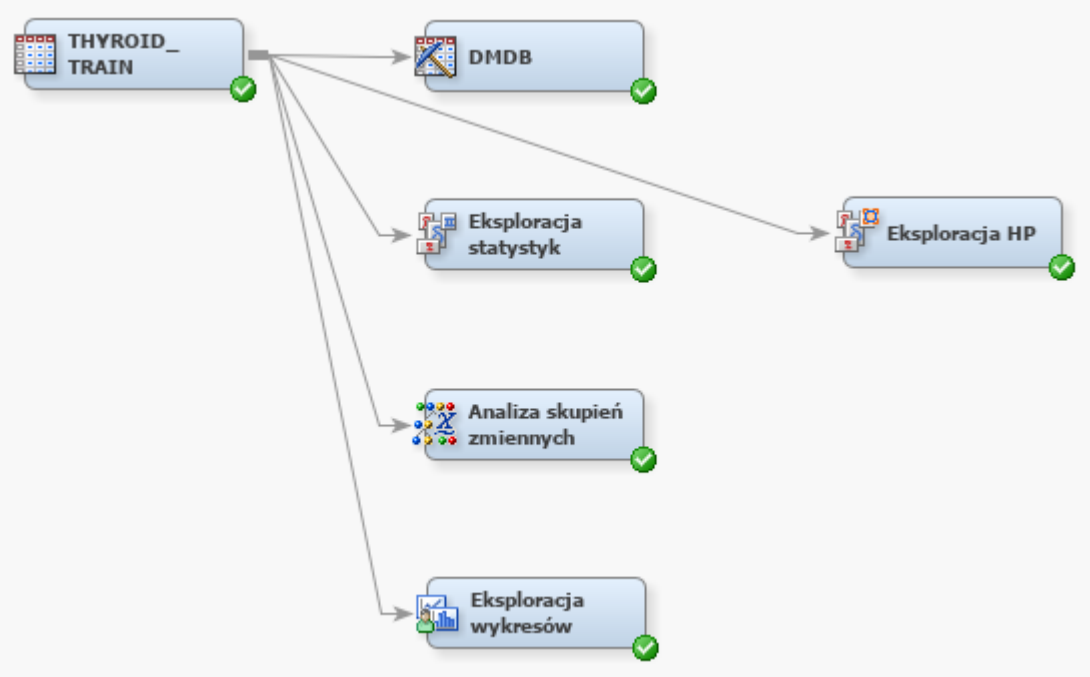
Zbiór przetworzony w SAS Enterprise Guide został wczytany do SAS Enterprise Miner i od razu podzielony na treningowy, walidacyjny i testowy w proporcjach odpowiednio po 60%, 20% i 20%. Podział odbył się z zachowaniem proporcji zmiennej celu. Ustawione ziarno losowości to 123.

3.5. Eksploracja danych w SAS Enterprise Miner na zbiorze treningowym.

Podczas wczytywania zbioru do SAS Enterprise Miner porzucono kolumnę referal_source, ponieważ skierowanie pacjenta nie powinno mieć wpływu na modelowanie, gdy nie mamy żadnych informacji o tym miejscu. Dodatkowo zmienne kategoryczne dostały typ BINARY, ponieważ zawierały jedynie informację “t” lub “f”.

Po podziale na zbiory treningowy, walidacyjny i testowy przeprowadzono eksplorację danych treningowych, aby zapoznać się z aktualnymi danymi, aby później móc zdecydować co zrobić w przypadku modelowania.

Na Rysunku 6 przedstawiono schemat przeprowadzonej eksploracji.



Rys. 6 Schemat eksploracji zbioru treningowego.

Większość analiz powtarza się z tymi przeprowadzonymi wcześniej, jednak należało sprawdzić poprawność wczytanych danych do SAS Enterprise Miner oraz jak zbiór został podzielony.

Poniżej przedstawiono krótko niektóre analizy i wnioski.

DMDB

Z interesujących informacji można wypisać, że:

- dwie kolumny (atrybuty) posiadają tylko jedną wartość - są to hypopituitary oraz TBG_measured, to znaczy, że nie wniosą nam one żadnych informacji podczas uczenia modelu,
- 90 braków danych w zmiennej sex trafiło do zbioru treningowego, czyli 60 braków jest w innych zbiorach,
- statystyki opisowe (agregujące) zmiennych trochę się zmieniły względem poprzednich wartości (Tabela 10).

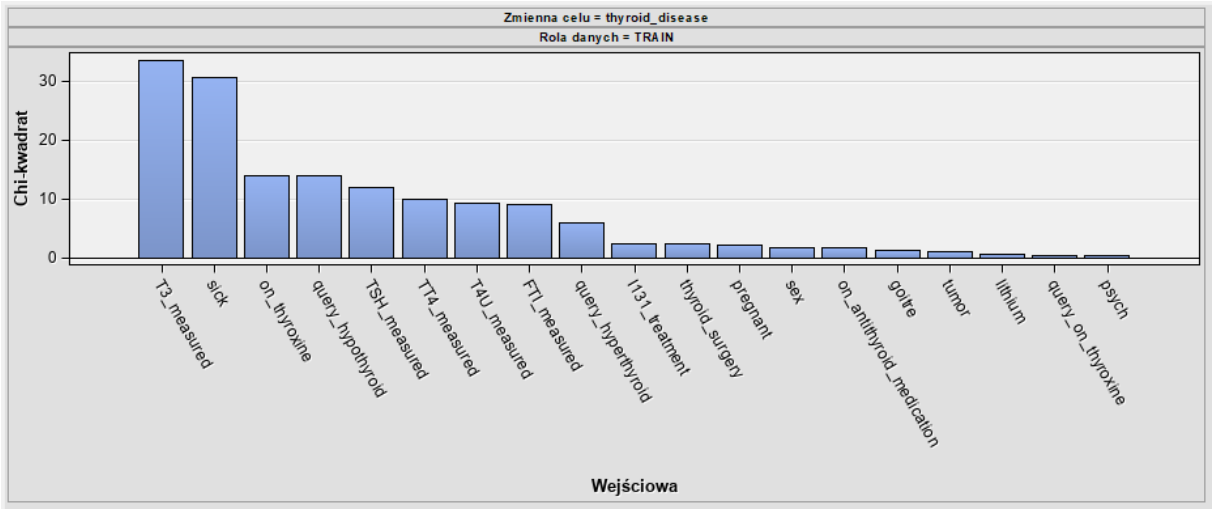
Tab. 10 Statystyki agregujące zmiennych numerycznych w zbiorze treningowym.

Zmienna	Etykieta	Braki danych	N	Minimum	Maksimum	Średnia	Odchylenie standardowe	Skośność	Kurtoza
FTI		237	2026	2.000	362.00	109.740	32.8887	1.1968	6.531
T3		470	1793	0.050	10.60	2.002	0.8181	1.7515	11.307
T4U		238	2025	0.310	2.32	0.995	0.1937	1.2509	4.131
TSH		223	2040	0.005	478.00	5.649	27.3356	12.4343	185.516
TT4		143	2120	2.000	372.00	107.611	34.9828	0.9969	4.108
age		0	2263	1.000	93.00	51.911	18.9950	-0.2291	-0.844

Eksploracja statystyk i eksploracja HP

Podczas eksploracji z podziałem na chorych i zdrowych pacjentów okazało się, że część zmiennych przyjmuje tylko jedną wartość dla chorych pacjentów.

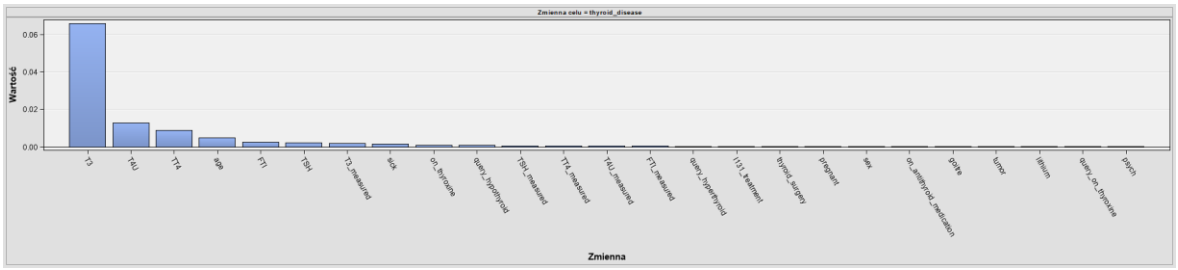
Na Rysunku 7 można zobaczyć wartości Chi-kwadrat dla zmiennych kategorycznych. Pomiar bądź jego brak hormonu T3 ma największy wpływ na zmienną celu. Inne kolumny z wykonaniem pomiarów również są o wysokich wartościach, jednak na miejscach 2-4 znajdują się sick (złe samopoczucie), on_thyroxine (pacjent zażywa tyroksynę) oraz query_hypothyroid (być może niedoczynność tarczycy).



Rys. 7 Wartość Chi-kwadrat dla zmiennych binarnych (kategorycznych).

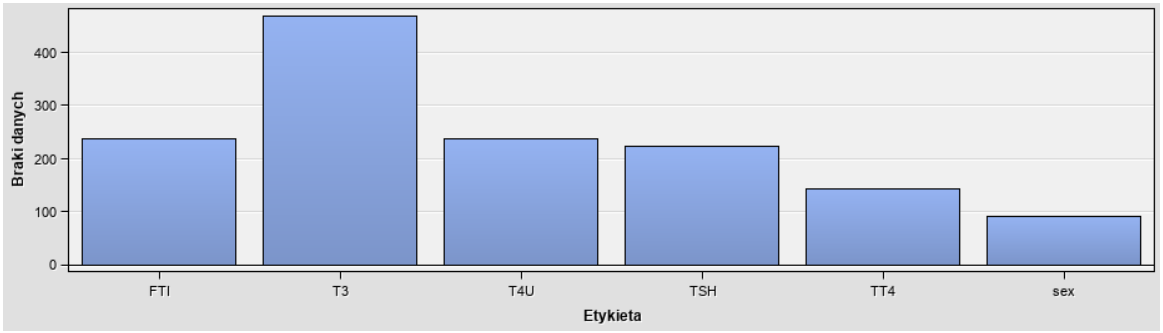
Rysunek 8 przedstawia wartości wszystkich zmiennych w zbiorze. Zmienne przedziałowe (numeryczne) znalazły się wszystkie przed kategorycznymi, których ułożenie jest takie samo jak na wykresie Chi-kwadrat.

Kolejność zmiennych numerycznych: T3, T4U, TT4, age, FTI, TSH.



Rys. 8 Wartość zmiennych w zbiorze.

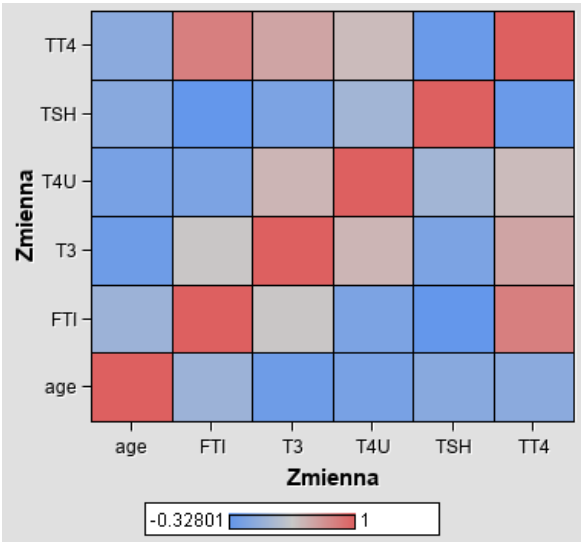
Rysunek 9 pokazuje liczbę braków danych z podziałem na zmienne ze zbioru treningowego. Największą liczbę braków zabiera atrybut T3, co znaczy, że testy na ten hormon robiono najrzadziej wśród pacjentów.



Rys. 9 Liczba braków danych.

Analiza skupień zmiennych

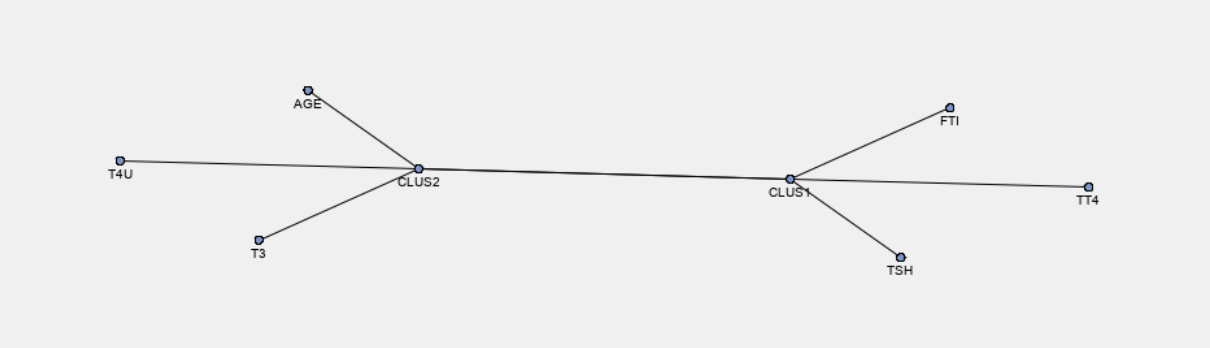
Na Rysunku 10 pokazano korelacje zmiennych numerycznych. Największą korelację ma TT4 z FTI (0.81). Również wysokie są korelacje TT4 z T3 (0.57). Korelacje T4U z T3 oraz TT4 są na poziomie około 0.4. Należy więc wykluczyć z uczenia zmienne skorelowane.



Rys. 10 Korelacje zmiennych numerycznych.

Analiza skupień podzieliła zmienne numeryczne na dwa klastry. Na Rysunku 11 pokazano, jak wygląda ten podział. W klastrze pierwszym występuje TT4, FTI oraz TSH. W drugim T3, wiek oraz T4U. Można zauważyć, że są to po dwie zmienne skorelowane i jedna nieskorelowana.

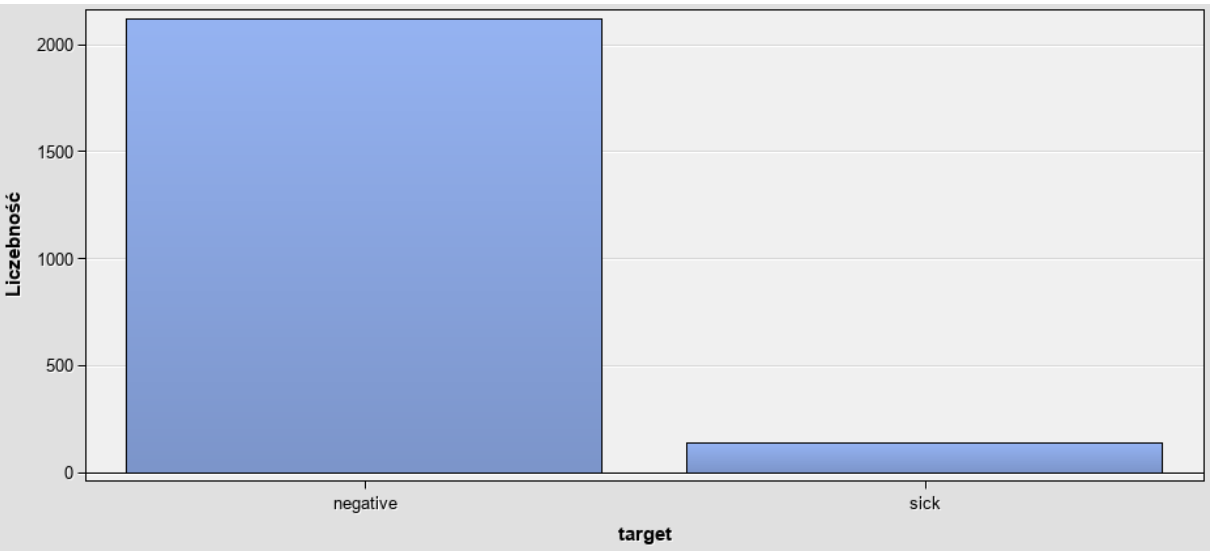
Według wiedzy medycznej o hormonach, dzięki wolnemu T4 oraz TSH można często określić chorobę tarczycy. Wolne T4 występuje we wzorach na TT4 oraz FTI.



Rys. 11 Wykres skupień zmiennych.

Eksploracja wykresów

Przypominając, że zbiór jest niezbalansowany również po podziale (zachował takie same proporcje) poniżej załączono wykres liczebności wartości w zmiennej celu (Rysunek 12). Dane są zebrane ze szpitali, więc jest to rzeczywisty stosunek chorych do zdrowych w społeczeństwie z okresu, w którym dane były zbierane.



Rys.12 Liczebność wartości występujących w zmiennej celu.

4. Analiza zbioru.

4.1. Opis zmiennych w zbiorze.

Podczas przekształcania i eksploracji zbioru potrzebna była wiedza o znaczeniu tych zmiennych, aby następnie wyniki EDA oraz tę wiedzę połączyć i przejść do wykonania działań w kierunku przygotowania zbioru do modelowania. Ten rozdział zawiera opis zmiennych ze zbioru zebranych w jedno miejsce. Informacje o znaczeniu zmiennych zebrano z różnych źródeł internetowych ([3],[4]) oraz korzystając z tych nazw zmiennych, które były oczywiste.

Krótki opis zmiennych zawarto w Tabeli 11.

Tab. 11 Lista atrybutów zbioru.

Nr	Nazwa atrybutu	Znaczenie / Opis
1	age	wiek w latach
2	sex	płeć
3	on_thyroxine	pacjent zażywający tyroksynę
4	query_on_thyroxine	pacjent być może zażywający tyroksynę
5	on_antithyroid_medication	pacjent na lekach przeciwtarczycowych
6	sick	pacjent zgłasza złe samopoczucie
7	pregnant	pacjent w ciąży
8	thyroid_surgery	historia operacji tarczycy
9	I131_treatment	pacjent na leczeniu I131
10	query_hypothyroid	być może niedoczynność tarczycy
11	query_hyperthyroid	być może nadczynność tarczycy
12	lithium	pacjent przyjmujący lit
13	goitre	pacjent ma wole (powiększenie gruczołu tarczycy)
14	tumor	pacjent ma guza
15	hypopituitary	pacjent z niedoczynnością przysadki
16	psych	objawy psychologiczne
17	TSH_measured	czy zmierzono TSH
18	TSH	wartość TSH (tyreotropiny) - zwiększenie masy tarczycy, zwiększenie przepływu krwi przez ten narząd oraz nasilenie

		produkcji i wydzielania hormonów tarczycy: tyroksyny i trójjodotyroniny
19	T3_measured	czy zmierzono T3
20	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 chodzi o wolne T3
21	TT4_measured	czy zmierzono TT4
22	TT4	wartość TT4 (tyroksyna) - T4 całkowita, norma 58–154 nmol/l
23	T4U_measured	czy zmierzono T4U
24	T4U	wartość T4U - wykorzystanie tyroksyny przez organizm (eksploatacja, pobór)
25	FTI_measured	czy zmierzono FTI
26	FTI	wartość FTI - wolny testosteron, wiarygodny indeks do oceniania stanu tarczycy, FTI = Thyroxine (T4)/Thyroid Binding Capacity
27	TBG_measured	czy zmierzono TBG
28	TBG	wartość TBG - stężenie globuliny wiążącej tyroksynę
29	referal_source	źródło skierowania
30	Class	klasa diagnostyczna, określa chorobę tarczycy lub jej brak

4.2. Ocena jakości danych.

Wybrany zbiór danych zawierał braki danych, błędne wartości (wiek wynoszący 455) a nawet całe puste kolumny (TBG). Dodatkowo na stronie UCI nie dało się zrozumieć opisu zawartych zbiorów, większość była uszkodzona oraz nie istniały opisy zmiennych. Na szczęście kolumny są nazwane w miarę oczywiście oraz istnieją artykuły zawierające wyjaśnienia do tego zbioru.

Zbiór jest niezbalansowany, więc trudno się z nim pracuje przy modelowaniu. Niektóre zmienne są całkowicie nieistotne (jedna wartość dla wszystkich obserwacji) a inne skorelowane i tworzą szum w danych.

Biorąc pod uwagę wszystkie czynniki to zbiór nie jest dobrej jakości, należy się mu dokładnie przyjrzeć, aby go oczyścić i przetransformować do dalszej pracy. Oczywiście najpierw trzeba znaleźć nieuszkodzony plik z danymi w internecie, ponieważ jeśli dane byłyby uszkodzone bądź zaszumione to w przypadku medycznych wartości trudno byłoby bez wiedzy eksperckiej to naprawić.

5. Modelowanie.

5.1. Zastosowane rozwiązania.

Po przeprowadzonej dokładnej eksploracji i zrozumieniu znaczenia nazw kolumn, przyszedł czas na wprowadzenie przekształceń danych pod etap modelowania.

Dane treningowe, walidacyjne i testowe przeszły szereg zmian doprowadzając ostatecznie dane do czterech różnych postaci, na których przeprowadzono modelowanie. Wybrane modele predykcyjne były zbudowane na danych treningowych i walidacyjnych, a ich ocena została wykonana na podstawie wyników na zbiorze testowym.

Wszelkie zmiany zastosowane na danych:

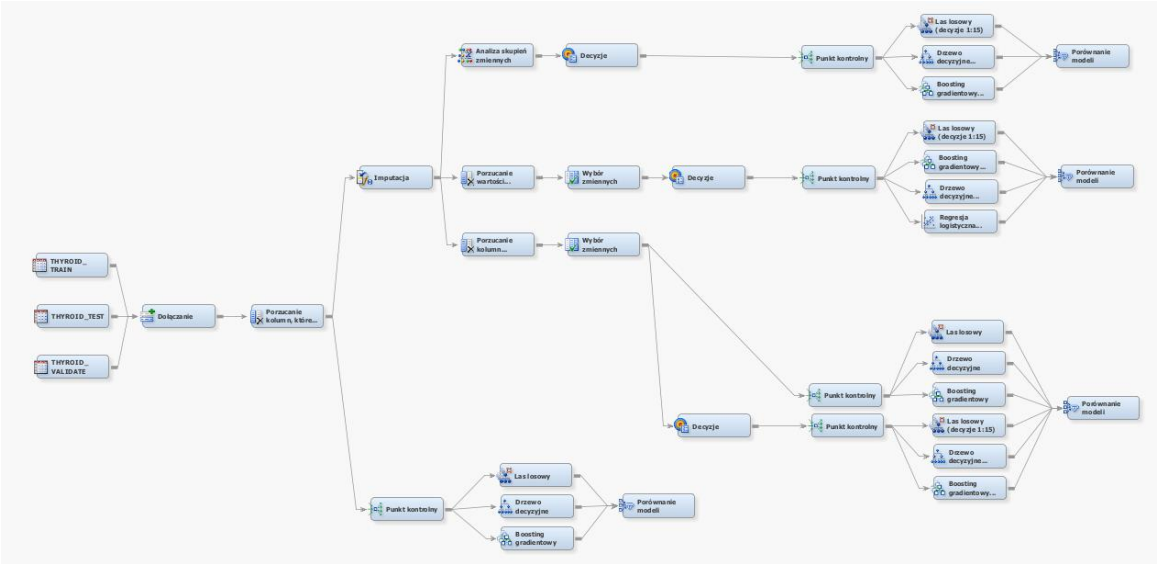
- porzucenie kolumn z jedną wartością dla wszystkich wierszy,
- imputacja braków danych - dotyczy tylko zmiennej wieku i płci, ponieważ jak opisano wyżej uznano, że informacja o brakach dotycząca hormonów jest potrzebna do diagnozy:
  - wiek zaimputowano medianą,
  - płeć zaimputowano wartością o przeważającej liczności (modą),
- analiza skupień zmiennych,
- porzucenie kolumn z pomiarami hormonów bądź porzucenie kolumn “measured”, które określały, czy pomiar jest wykonany,
- wybór zmiennych na podstawie ich wartości Chi-kwadrat oraz ręczne usunięcie możliwego szumu na podstawie własnej wiedzy po eksploracji i analizie danych,
- określenie wag decyzji błędów I i II rodzaju:
  - domyślne - odpowiednio 0 i 0,
  - na podstawie poziomu niezbalansowania danych - odpowiednio 1 i 15 na minusie.

Cztery końcowe postaci, na których przeprowadzono modelowanie nie przeszły wszystkich wyżej wymienionych zmian. Lista wprowadzonych zmian do każdej z czterech końcowych postaci danych poddanych modelowaniu:

- surowe dane - porzucenie kolumn z jedną wartością dla wszystkich wierszy,
- wykorzystanie skupień zmiennych - porzucenie kolumn z jedną wartością dla wszystkich wierszy oraz imputacja,
- dane po odrzuceniu pomiarów hormonów - porzucenie kolumn z jedną wartością dla wszystkich wierszy, imputacja, wybór zmiennych,
- dane po odrzuceniu kolumn “measured” - analogicznie jak wyżej.

Na wszystkich czterech przygotowanych zbiorach przeprowadzono etap modelowania dla obu wersji ustawionych decyzji.

Na Rysunku 13 pokazano uproszczony schemat modelowania, to znaczy taki po usunięciu niektórych rozwiązań np. testowania dla decyzji 0 i 0, która przyniosła gorsze wyniki predykcji (większa liczba wystąpień False Negative) czy zastosowania punktu odcięcia, który nie przyniósł poprawy jakości modelu. Zrobiono to ze względu na uproszczenie schematu i pozostawienie jedynie wartościowych informacji o modelowaniu.



Rys.13 Uproszczony schemat przeprowadzonego modelowania.

5.2. Wybrane modele do klasyfikacji.

Uczenie maszynowe w medycynie ma wspierać lekarzy i zastępować ich w diagnozie. Lekarz musi rozumieć działanie modelu, aby móc ewentualnie skorygować diagnozę, jeśli model się myli. Żeby to jednak się działo model uczenia zbioru medycznego musi być wyjaśnialny.

W tym celu w projekcie zastosowano modele drzewiaste oraz regresję logistyczną. Zbudowano cztery różne modele:

- regresję logistyczną dla zbioru z danymi kategorycznymi (bez pomiarów hormonów, które mają braki danych i mogą źle wpływać na model regresyjny),
- drzewo decyzyjne,
- las losowy,
- oraz boosting gradientowy.

Za najlepsze przygotowanie zbioru do modelowania uznano opcję z pozostawieniem wartości pomiarów (porzucenie kolumn “measured”), gdzie odrzucono zmienne skorelowane czy zmienne nieistotne oraz zaimputowano braki danych. Wybrano ten zbiór danych ze względu na to, że pozbyto się zbędnego szumu mogącego wpływać na modele a także według wiedzy medycznej wartości pomiarów niektórych hormonów porównane ze sobą powinny same wystarczyć do określenia diagnozy. Stąd wybór najlepszych parametrów modeli był sprawdzany na tym zbiorze, z pominięciem regresji logistycznej nauczonej tylko dla jednego zbioru.

Kryterium wyboru to liczba błędnych klasyfikacji, którą chcemy minimalizować.

Regresja logistyczna

Najważniejsze wybrane parametry to:

- model wyboru - krokowa,
- kryterium wyboru - błędne klasyfikacje w walidacji krzyżowej,
- domyślne wartości dla technik uczenia modelu.

Drzewo decyzyjne

Najważniejsze wybrane parametry to:

- kryterium nominalnej zmiennej celu - Współczynnik Gini,
- poziom istotności - 0.1,
- użycie braków danych w wyszukiwaniu,
- maksymalna głębokość - 7,
- walidacja krzyżowa.

Las losowy

Najważniejsze wybrane parametry to:

- maksymalna liczba drzew - 200,
- ziarno losowości - 123,
- udział obserwacji w każdej z prób - 0.8,
- maksymalna głębia - 4,
- użycie braków danych w wyszukiwaniu,
- poziom istotności - 0.1,
- metoda istotności zmiennych - redukcja strat.

Boosting gradientowy

Najważniejsze wybrane parametry to:

- N iteracji - 100,
- ziarno losowości - 123,
- udział uczących - 0.8,
- maksymalna głębia - 4,
- użycie braków danych w wyszukiwaniu,
- miara oceny - błędne klasyfikacje.

5.3. Surowe dane.

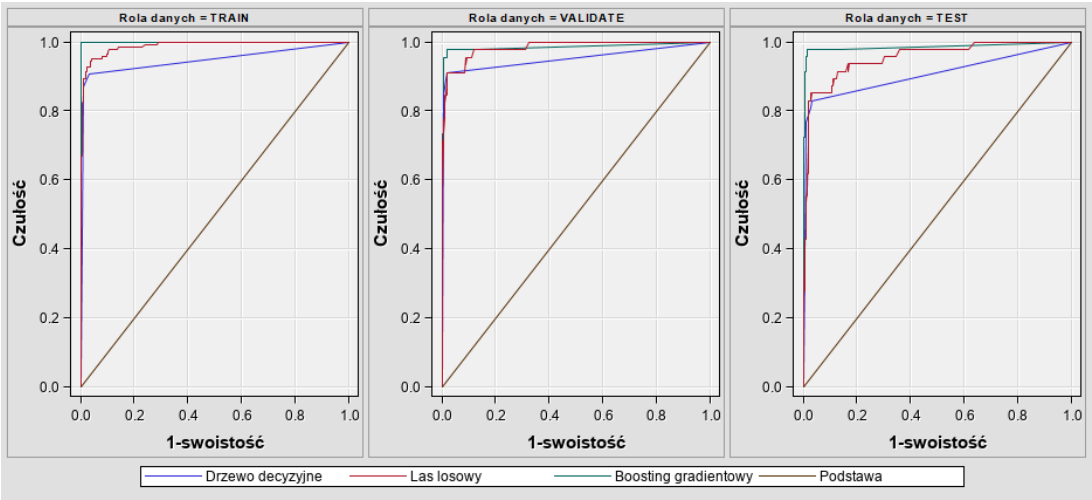
Są to dane jedynie po porzuceniu kolumn TBG_measured oraz hypopituitary, które mają tylko jedną wartość dla wszystkich wierszy. Zastosowano na nich modele drzewa decyzyjnego, lasu losowego oraz boostingu gradientowego. Wybierano modele na podstawie kryterium błędnych klasyfikacji. W Tabeli 12 przedstawiono wyniki dla zbioru testowego.

Tab. 12 Odsetek błędnych klasyfikacji na zbiorze testowym.

Boosting gradientowy	0.011905
Drzewo decyzyjne	0.02381
Las losowy	0.062169

Po wynikach na zbiorze testowym nie widać tego, ale model nadmiernie dopasował się do zbioru treningowego (przeuczenie), dodatkowo mamy tu dużo szumu, zmiennych skorelowanych itp., przez co nie jest to najlepszy model do oceny chorób.

Na Rysunku 14 przedstawiono porównanie modeli za pomocą krzywej ROC. Widać, że dla zbioru treningowego jest to 1.0 dla boostingu gradientowego.



Rys.14 Porównanie wyników modeli dla surowych danych.

5.4. Wykorzystanie skupień zmiennych.

Po odrzuceniu kolumn z jedną wartością wszędzie i imputacji wieku oraz płci wykonano analizę skupień zmiennych. Tutaj również wykorzystano modele drzewa decyzyjnego, lasu losowego oraz boostingu gradientowego oraz wybierano je na podstawie kryterium błędnych klasyfikacji.

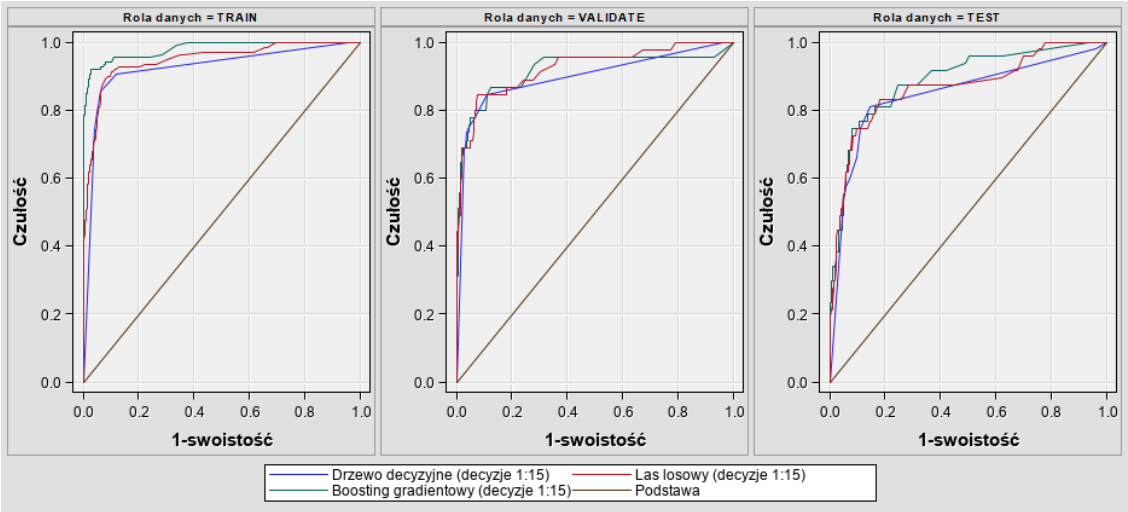
Wyjaśnienie czym są skupienia zmiennych i jak działa uczenie może być nieintuicyjne i za trudne dla medyków/lekarzy, więc nie jest to najlepszy pomysł uczenia. Dodatkowo wyniki modelowania również nie wyszły najlepiej jak byśmy tego chcieli.

Przedstawione poniżej wyniki są dla ustalonych wag decyzji -1 i -15, ponieważ były one lepsze. Wyniki liczbowe zastosowanego kryterium przedstawiono w Tabeli 13.

Tab. 13 Odsetek błędnych klasyfikacji na zbiorze testowym.

Boosting gradientowy (decyzje 1:15)	0.05291
Las losowy (decyzje 1:15)	0.063492
Drzewo decyzyjne (decyzje 1:15)	0.075397

Na Rysunku 15 przedstawiono porównanie modeli za pomocą krzywej ROC. Tutaj również najlepiej poradził sobie boosting gradientowy.



Rys.15 Porównanie wyników modeli dla danych po analizie skupień zmiennych.

5.5. Porzucenie kolumn z pomiarami poziomu hormonów.

Postanowiono sprawdzić, czy potrzebne są pomiary wartości hormonów czy może wystarczy wiedzieć, czy takie badanie zostało zlecone bądź nie. Usunięto zatem nie tylko kolumny z jedną wartością, ale też wszystkie zmienne numeryczne dotyczące hormonów. Oczywiście zmienne wieku i płci zostały zaimputowane. Po selekcji zmiennych za pomocą wartości Chi-kwadrat do uczenia wybrano następujące atrybuty zbioru:

- sex,
- T3_measured,
- age,
- psych,
- query_hypothyroid
- oraz sick.

Zastosowano modele drzewa decyzyjnego, lasu losowego, boostingu gradientowego oraz regresji logistycznej. Wybierano modele na podstawie kryterium błędnych klasyfikacji. Są to jedyne dane, gdzie było możliwe sprawdzenie modelu regresji, co tym bardziej skłaniało do testowania czy jest to dobra opcja do diagnozy.

Przedstawione poniżej wyniki są dla ustalonych wag decyzji -1 i -15, ponieważ były one lepsze. Wyniki liczbowe zastosowanego kryterium przedstawiono w Tabeli 14.

Tab. 14 Odsetek błędnych klasyfikacji na zbiorze testowym.

Regresja logistyczna (decyzje 1:15)	0.062169
Drzewo decyzyjne (decyzje 1:15)	0.062169
Las losowy (decyzje 1:15)	0.062169
Boosting gradientowy (decyzje 1:15)	0.063492

Wyniki są wysokie oraz dla trzech modeli identyczne. Tym razem boosting gradientowy był najgorszym modelem. Jednak wagi decyzji zostały przypisane inne dla błędu I i II rodzaju, skoro

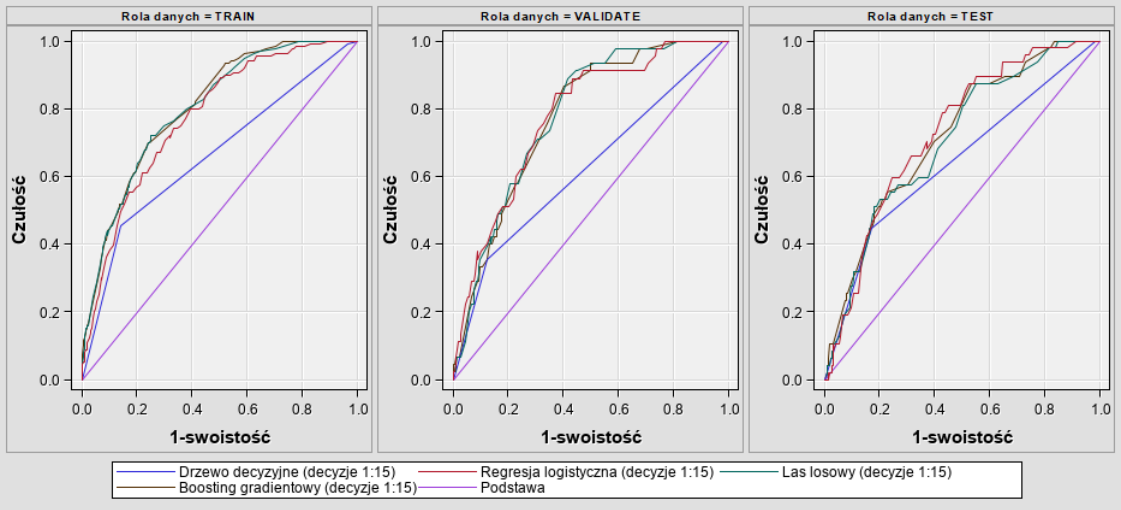
mamy takie same odsetki błędów to, aby udowodnić, że to nie jest błąd a przypadkowa zbieżność poniżej w Tabeli 15 przedstawiono zysk dla tego modelowania.

Tab. 15 Średni zysk na zbiorze testowym.

Regresja logistyczna (decyzje 1:15)	-0.87037
Drzewo decyzyjne (decyzje 1:15)	-0.87037
Las losowy (decyzje 1:15)	-7.82979
Boosting gradientowy (decyzje 1:15)	-0.89153

W tym przypadku regresja i drzewo mają największy zysk. Indeks ROC ma największa regresja, więc to ona jest najlepsza dla takich danych.

Na Rysunku 16 przedstawiono porównanie modeli za pomocą krzywej ROC.



Rys.16 Porównanie wyników modeli dla danych po usunięciu pomiarów hormonów.

5.6. Porzucenie kolumn zawierających w nazwie “measured”.

Zastosowane tutaj przekształcenia i wyniki na zbiorze danych uznano za najlepsze.

Do przygotowania zbioru do modelowania usunięto kolumny z jedną wartością oraz z nazwą z dopiskiem “measured”, a także zmienne wieku i płci zostały zaimputowane. Po selekcji zmiennych za pomocą wartości Chi-kwadrat do uczenia wybrano następujące atrybuty zbioru:

- FTI
- T3
- T4U
- TSH
- age
- on_thyroxine
- oraz TT4 (usunięto ją ręcznie ze względu na korelacje oraz na to, że całkowita tyroksyna może pogarszać diagnozę przez nieznany jej wpływ w niektórych przypadkach).

Przedstawione poniżej wyniki są dla ustalonych wag decyzji -1 i -15 oraz dla 0 i 0. Dla obu opcji przetestowano trzy modele (drzewa decyzyjnego, lasu losowego, boostingu gradientowego), przy czym ważne jest, że te modele na tych danych były tuningowane tak, aby wyniki były jak najlepsze. Wybierano modele na podstawie kryterium błędnych klasyfikacji.

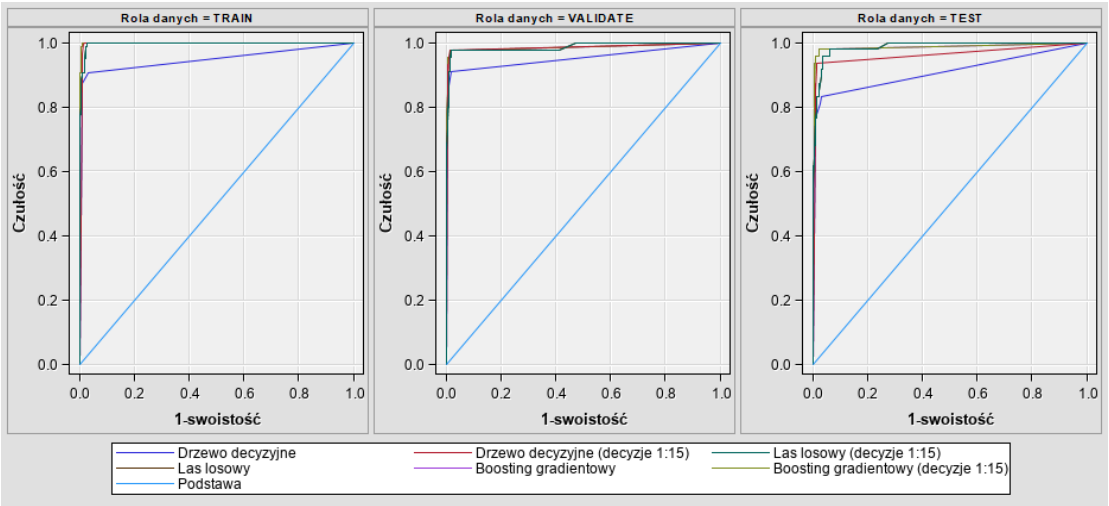
Wyniki liczbowe zastosowanego kryterium oraz średniego zysku (tylko dla wag decyzji -1 i -15) przedstawiono w Tabeli 16.

Tab. 16 Odsetek błędnych klasyfikacji (druga kolumna) oraz średni zysk (trzecia kolumna) na zbiorze testowym.

Boosting gradientowy	0.018519	.
Boosting gradientowy (decyzje 1:15)	0.018519	-0.02651
Drzewo decyzyjne (decyzje 1:15)	0.019841	-0.03624
Drzewo decyzyjne	0.02381	.
Las losowy	0.025132	.
Las losowy (decyzje 1:15)	0.025132	-0.23022

Odsetek błędnych klasyfikacji jest taki sam dla modeli: boosting gradientowy oraz las losowy. Z jakiegoś powodu te modele nie zmieniły wyników mimo większej kary za błąd II rodzaju. Natomiast jak widać las losowy poprawił swoje wyniki. Średni zysk jest na minusie. Patrząc na te wyniki liczbowe to boosting gradientowy jest najlepszy.

Na Rysunku 17 przedstawiono porównanie modeli za pomocą krzywej ROC. Największy indeks ROC posiada las losowy. Mimo to wszystkie modele osiągnęły indeks rok powyżej 0.9.



Rys.17 Porównanie wyników modeli dla danych po usunięciu kolumn skorelowanych, "measured" oraz z jedną wartością; zaimputowanych i po selekcji zmiennych.

W predykcji nie interesuje nas największy indeks ROC czy najwięcej idealnych predykcji. Chcemy minimalizować błąd II rodzaju, czyli pojawienie się False Negative. Teraz skupimy się na najlepszym modelu, aby pokazać różnice przy zastosowaniu wag decyzji -1 i -15 a ich braku.

Na Rysunkach 18 i 19 przedstawiono fragmenty wyników z SASa dotyczące predykcji zmiennych ze względu na decyzje bądź ich brak. W zbiorze treningowym, gdy ustawiono decyzje nie było żadnego FN (False Negative). Przy czym dla braku decyzji było to 9 osób (0.4%).

Tabela klasyfikacji

Rola danych=TRAIN Zmienna celu=thyroid_disease Etykieta zmiennej celu=target

Zmienna celu	Wynik	Procent docelowy	Procent wynikowy	Liczba wystąpień	Całkowity procent
NEGA	NEGA	99.5753	99.3409	2110	93.2391
SICK	NEGA	0.4247	6.4748	9	0.3977
NEGA	SICK	9.7222	0.6591	14	0.6186
SICK	SICK	90.2778	93.5252	130	5.7446

Rola danych=VALIDATE Zmienna celu=thyroid_disease Etykieta zmiennej celu=target

Zmienna celu	Wynik	Procent docelowy	Procent wynikowy	Liczba wystąpień	Całkowity procent
NEGA	NEGA	99.5751	99.2938	703	93.3599
SICK	NEGA	0.4249	6.6667	3	0.3984
NEGA	SICK	10.6383	0.7062	5	0.6640
SICK	SICK	89.3617	93.3333	42	5.5777

Rys.18 Wynik klasyfikacji dla zbioru bez ustawionych decyzji.

Tabela decyzji

Rola danych=TRAIN Zmienna celu=thyroid_disease Etykieta zmiennej celu=target

Zmienna celu	Wynik	Procent docelowy	Procent wynikowy	Liczba wystąpień	Całkowity procent
NEGA	NEGATIVE	100.000	98.588	2094	92.5320
NEGA	SICK	17.751	1.412	30	1.3257
SICK	SICK	82.249	100.000	139	6.1423

Rola danych=VALIDATE Zmienna celu=thyroid_disease Etykieta zmiennej celu=target

Zmienna celu	Wynik	Procent docelowy	Procent wynikowy	Liczba wystąpień	Całkowity procent
NEGA	NEGATIVE	99.7143	98.5876	698	92.6959
SICK	NEGATIVE	0.2857	4.4444	2	0.2656
NEGA	SICK	18.8679	1.4124	10	1.3280
SICK	SICK	81.1321	95.5556	43	5.7105

Rys.18 Wynik klasyfikacji dla zbioru z ustawionymi wagami decyzji na -1 (błąd I rodzaju) i -15 (błąd II rodzaju).

Po dokładnym przyjrzeniu się powyższym wynikom można zauważyć, że pule predykcji innych typów zmieniły się, powodując na przykład wzrost błędu I rodzaju. Dlatego w przypadku danych medycznych jest to mniejszy problem i należy tak robić to do innych problemów wzrost błędu I rodzaju może nie być dobrym rozwiązaniem.

Podsumowując, trochę zmieniła się jakość predykcji po wprowadzeniu wag decyzji, nawet jeśli mowa o 1 czy 2 pacjentach to po pierwsze jest to mały zbiór a po drugie nie można ocenić czy to

mało czy nie bo chodzi o ludzkie życia. Są to koszty niepoliczalne (ewentualnie przez ekspertów można to określić w jakiś sposób).

Dodatkowo w Tabeli 17 przedstawiono ciekawe statystyki dla wykorzystywanych zmiennych. Można zauważyć na przykład istotność T3, TSH i FTI - co wskazuje na to, że medyczne zależności mają potwierdzenie w informatycznych rezultatach (chodzi o możliwość postawienia diagnozy tylko na podstawie TSH oraz wolnego T4 i wolnego T3).

Tab. 17 Informacje o zmiennych wykorzystanych do uczenia zbioru.

Nazwa zmiennej	Liczba reguł podziału	Istotność ▼	Istotność walidacji	Iloraz istotności walidacji i istotności uczenia
T3	25	1	1	1
TSH	38	0.55004	0.260758	0.47407
FTI	71	0.398526	0.323831	0.812572
T4U	10	0.175737	0.138587	0.788607
age	16	0.164715	0.126431	0.767571
on_thyroxine	7	0.158908	0.179246	1.127985

6. Podsumowanie.

W projekcie okazało się, że tworzenie modeli uczenia maszynowego, które mają wesprzeć lekarzy w stawianiu diagnozy, jest bardzo ciężkie, kiedy nie ma się wiedzy eksperckiej z medycyny. Bardzo łatwo przeoczyć pewne korelacje, usunąć ważne zmienne, jeśli się nie zna ich znaczenia czy nawet jeśli chodzi o kwestię imputacji - może się okazać, że braki danych są mile widziane i pomogą przy szukaniu najlepszego rozwiązania. Co prowadzi co przeuczenia, wycieku bądź błędnej interpretacji danych przez model.

Wykorzystane modele nie są idealne, więc mogą jedynie wesprzeć w diagnozie a nie zastąpić lekarza. Jednak modele drzewiaste dla postawionego problemu przyniosły bardzo dobre rezultaty. Są wyjaśnialne, przy czym w prostych przypadkach radzą sobie poprawnie.

Dobrym podejściem jest ustanowienie wag decyzji czy kosztów dla uczenia. Chcemy unikać błędów II rodzaju nawet kosztem zwiększenia błędu I rodzaju. Wagi ustawiono jako jeden z domyślnie stosowanych sposobów, jednak powinien je ustalić ekspert i podać jakie są koszty wystąpienia tych błędów.

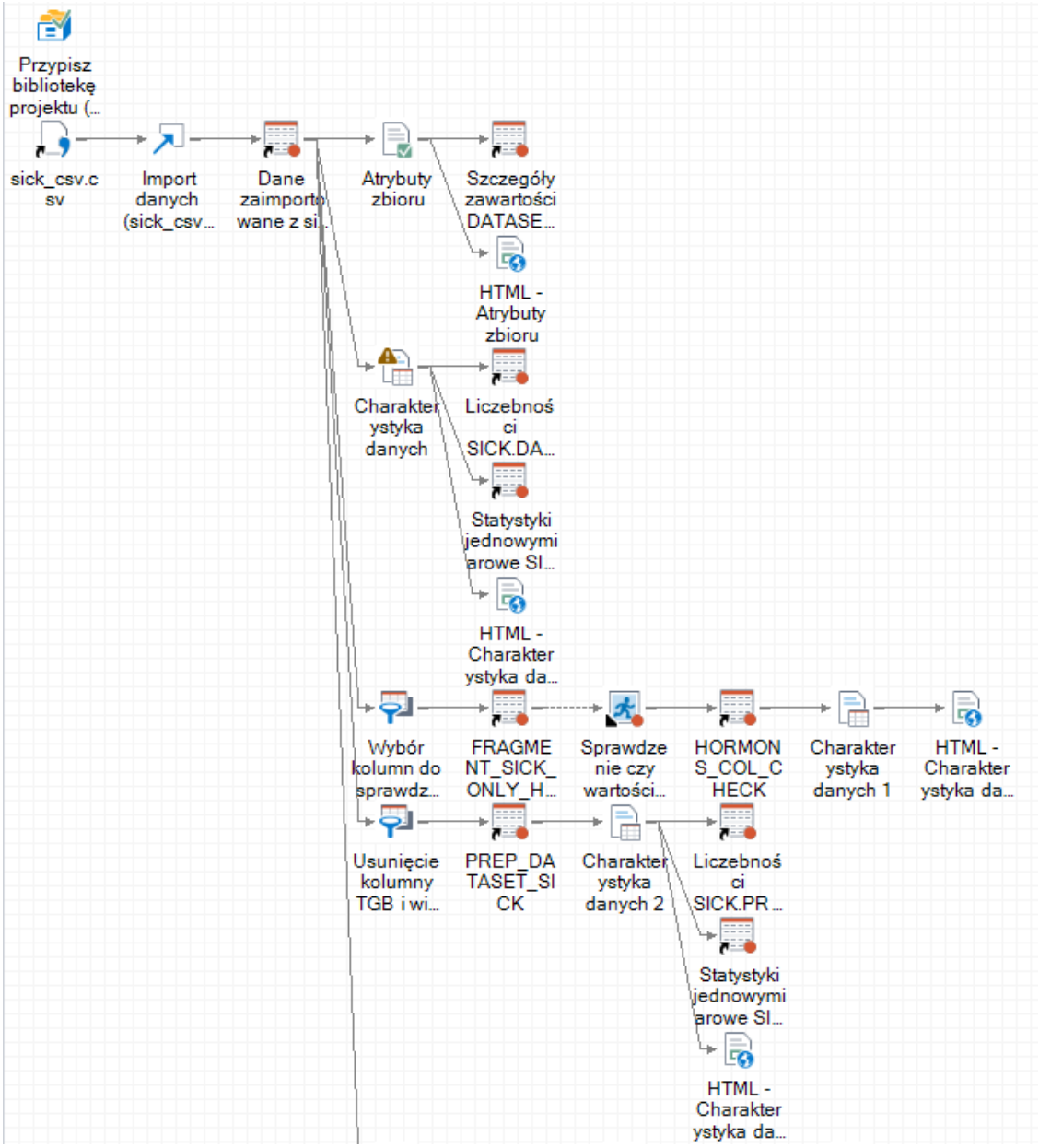
Klasyfikacja sprawdziła się całkiem dobrze na tym zbiorze potwierdzając, że można wykorzystać uczenie maszynowe do medycyny. Innym podejściem, wartym sprawdzenia, byłoby zastosowanie Cost Sensitive Learning (zamiast klasyfikacji), który może się do medycyny okazać jednak lepszy, ponieważ bierze pod uwagę koszt każdej zmiennej (np. koszt badania).

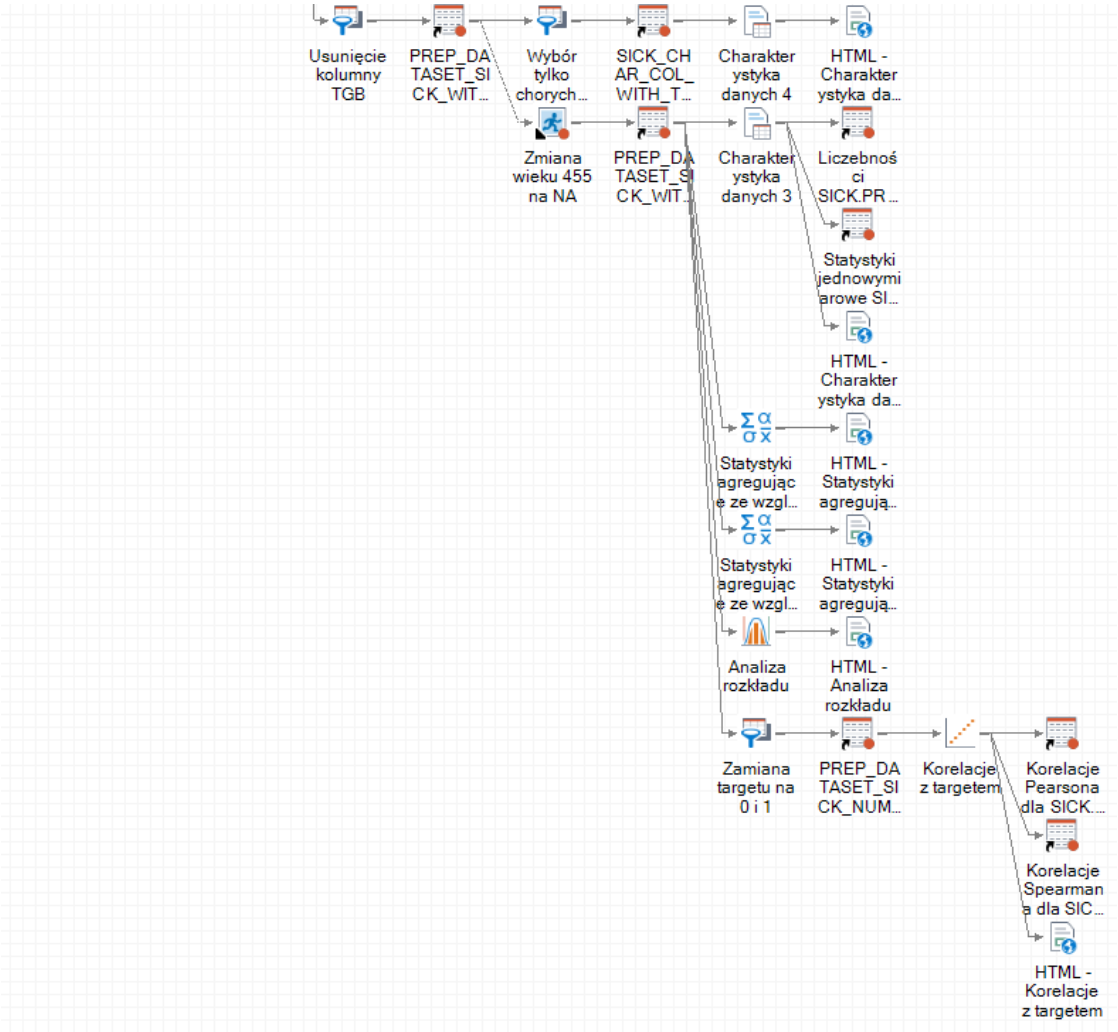
7. Bibliografia.

1) Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
2) Zbiór sick, DataHub [https://datahub.io/machine-learning/sick].
3) Wikipedia[https://pl.wikipedia.org/wiki/Tyreotropina].
4) Turney, Peter. (1995). Cost-Sensitive Classification: Empirical Evaluation of a Hybrid Genetic Decision Tree Induction Algorithm. J Artif Intell Res. 2. 10.1613/jair.120.

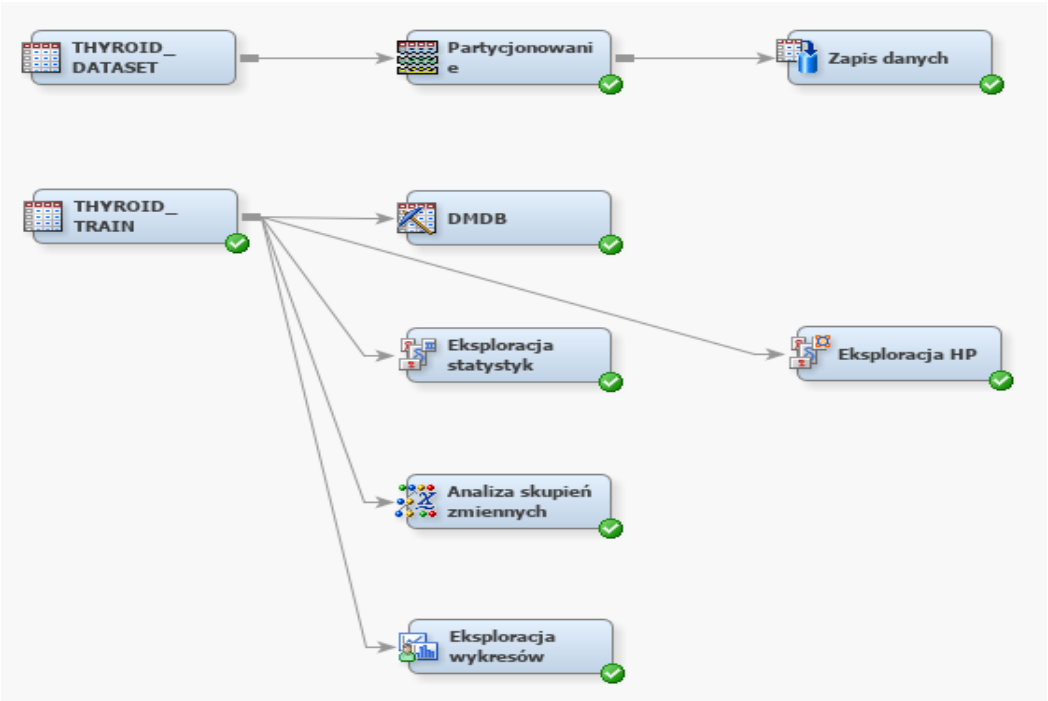
8. Załączniki.

Schemat SAS Enterprise Guide





SAS Enterprise Miner



RAPORT KOŃCOWY PROJEKTU ANALITYCZNEGO W SAS