Weryfikacja płci na podstawie danych dotyczących głosu przy wykorzystaniu modeli:

- **regresji logistycznej,**
- **drzewa decyzyjnego,**
- **sieci neuronowej**

Przy wykorzystaniu oprogramowania SAS Enterprise Miner oraz metodologii CRISP-DM

Spis treści

Wstęp

Identyfikacja płci na podstawie głosu umożliwia przykładowo predykcję cech rozmówcy, między innymi wiek lub też emocje oraz potencjalne zainteresowanie produktem danego przedsiębiorstwa. Omawiana identyfikacja to dzisiaj nieodłączny element systemu weryfikacji głosowej.

1. Cel projektu

Celem projektu jest eksploracja posiadanych zasobów danych dotyczących głosu grupy osób oraz zbudowanie, porównanie i wybranie najlepszego spośród modeli: regresji logistycznej, drzewa decyzyjnego oraz sieci neuronowych, zbudowanych na podstawie tych danych służącego do różnicowania płci badanych osób.

2. Metodologia pracy – CRISP-DM

Projekt opiera się na wykorzystaniu metodologii CRISP-DM (Cross Industry Standard Process for Data Mining). Jest to metodologia obejmujący cały proces eksploracji danych, od zrozumienia problemu biznesowego po wdrożenie modelu. CRISP-DM składa się z sześciu następujących etapów projektu:

- ocena wymagań biznesowych
- ocena danych
- przygotowanie danych
- modelowanie
- weryfikacja
- wdrożenie

3. Problem biznesowy

Wyniki przeprowadzonych badań mogą posłużyć do rozwoju technik identyfikacji klienta. W dobie systemów komunikacji z klientem przy pomocy maszyn i sztucznej inteligencji automatyczne rozpoznawanie płci rozmówcy jest istotnym czynnikiem w segmentacji klienta czy personalizacji interakcji utworzonej na linii rozmówca-komputer. Rozpoznanie płci umożliwia poprawę predykcji cech klienta czy emocji którymi się kieruje. Rozpoznanie płci to dzisiaj nie odłączna część każdego systemu weryfikacji głosu.

4. Zrozumienie i przygotowanie danych

4.1 Poznanie danych

Zbiór danych pochodzi zasobów strony kaggle.com, https://www.kaggle.com/primaryobjects/voicegender(dostęp: 20.11.2020) i dotyczy analizy akustycznej głosu. W zbiorze znajduje się 3 168 obserwacji oraz 20 cech głosu każdej z przebadanych jednostek. Zmienną celu stanowi zmienna „label”.

Tabela 1. Zmienne wraz z objaśnieniami

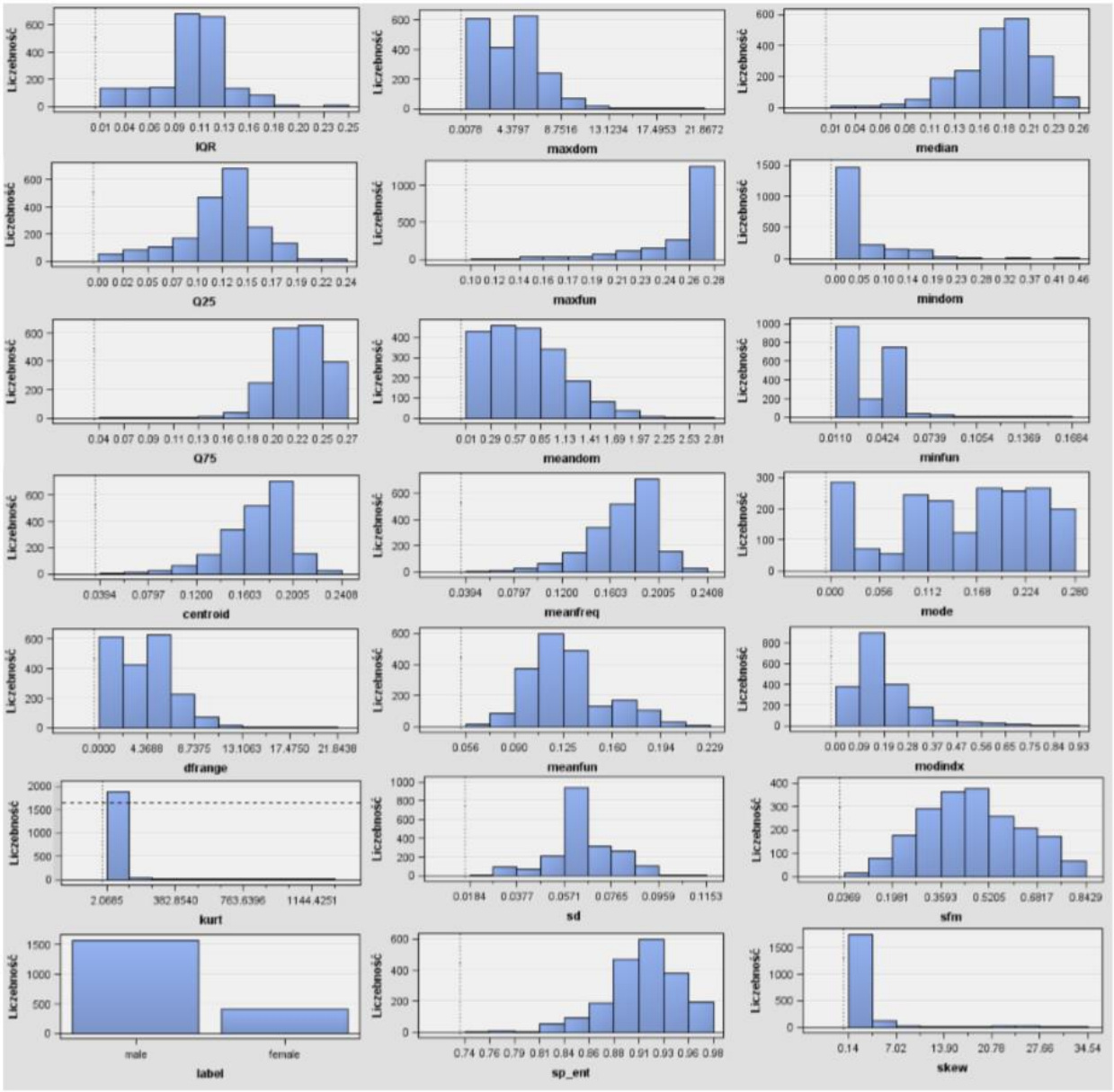
Lp.	Zmienna	Objaśnienie
1.	Label	Płeć (zmienna celu)
2.	duration	Długość badanego sygnału
3.	meanfreq	Średnia częstotliwość (w kHz)
4.	sd	Odchylenie standardowe częstotliwości
5.	median	Mediana częstotliwości (w kHz)
6.	Q25	Pierwszy kwartyl (w kHz)
7.	Q75	Trzeci kwartyl (w kHz)
8.	IQR	Zakres ćwiartkowy (w kHz)
9.	skew	skośność
10.	kurt	kurtoza
11.	sp_ent	Entropia spektralna
12.	sfm	Płaskość spektralna
13.	mode	Dominanta częstotliwości
14.	centroid	Centroid częstotliwości
15.	peakf	Najwyższa częstotliwość
16.	meanfun	Średnia podstawowej częstotliwości mierzonej w trakcie sygnału akustycznego
17.	minfun	Minimalna podstawowa częstotliwości mierzonej w trakcie sygnału akustycznego
18.	maxfun	Maksymalna podstawowa częstotliwości mierzonej w trakcie sygnału akustycznego
19.	meandom	Średnia podstawowa częstotliwości mierzonej w trakcie sygnału akustycznego
20.	mindom	Minimum podstawowej częstotliwości mierzonej w trakcie sygnału akustycznego
21.	maxdom	Maksimum podstawowej częstotliwości mierzonej w trakcie sygnału akustycznego
22.	dfrange	Zakres dominującej częstotliwości mierzonej w trakcie sygnału akustycznego
23.	modindx	modulacja

4.2 Podstawowe statystyki i wykresy dotyczące danych

Celem poznania badanego zbioru danych, poniżej zaprezentowane zostały liczebności poszczególnych zmiennych przy użyciu podstawowych węzłów eksploracji danych oraz statystyki poszczególnych zmiennych wywołane przez węzeł eksploracji statystyk w SAS Enterprise Miner.

W badanym zbiorze nie występują braki danych.

Rysunek 1. Liczebność poszczególnych wartości badanych zmiennych



Źródło: Opracowanie własne przy użyciu oprogramowania SAS Enterprise Miner

Tabela 2. Podstawowe statystyki dla zmiennych numerycznych

Obs.	NAME	NMISS	N	MIN	MAX	MEAN	STD	SKEWNESS	KURTOSIS
1	IQR	0	3168	0.01456	0.25	0.0843	0.043	0.29543	-0.4482
2	Q25	0	3168	0.00023	0.25	0.1405	0.049	-0.49088	0.0183
3	Q75	0	3168	0.04295	0.27	0.2248	0.024	-0.90031	2.9818
4	centroid	0	3168	0.03936	0.25	0.1809	0.030	-0.61750	0.8052
5	dfrange	0	3168	0.00000	21.84	4.9946	3.520	0.72826	1.3180
6	kurt	0	3168	2.06846	1309.61	36.5685	134.929	5.87259	35.9321
7	maxdom	0	3168	0.00781	21.87	5.0473	3.521	0.72619	1.3147
8	maxfun	0	3168	0.10309	0.28	0.2588	0.030	-2.23853	5.2039
9	meandom	0	3168	0.00781	2.96	0.8292	0.525	0.61102	-0.0548
10	meanfreq	0	3168	0.03936	0.25	0.1809	0.030	-0.61750	0.8052
11	meanfun	0	3168	0.05557	0.24	0.1428	0.032	0.03914	-0.8600
12	median	0	3168	0.01097	0.26	0.1856	0.036	-1.01278	1.6295
13	mindom	0	3168	0.00488	0.46	0.0526	0.063	1.66111	2.1876
14	minfun	0	3168	0.00978	0.20	0.0368	0.019	1.87800	10.7581
15	mode	0	3168	0.00000	0.28	0.1653	0.077	-0.83724	-0.2559
16	modindx	0	3168	0.00000	0.93	0.1738	0.119	2.06433	5.9249
17	sd	0	3168	0.01836	0.12	0.0571	0.017	0.13692	-0.5218
18	sfm	0	3168	0.03688	0.84	0.4082	0.178	0.33996	-0.8359
19	skew	0	3168	0.14174	34.73	3.1402	4.241	4.93331	25.3634
20	sp_ent	0	3168	0.73865	0.98	0.8951	0.045	-0.43093	-0.4239

Źródło: Opracowanie własne przy użyciu oprogramowania SAS Enterprise Miner

4.3 Przygotowanie danych

Pierwotny zbiór posiadający 3 168 obserwacji pomniejszam o 30% celem stworzenia odrębnego zbioru testowego.

Pomniejszony o 30% pierwotny zbiór danych zawiera obecnie 2 217 obserwacji. W następnym kroku wykonuję partycjonowanie z wykorzystaniem metody partycjonowania – losowanie warstwowe i dzielę zbiór w 70% na zbiór uczący i w 30% na zbiór walidacyjny.

Tabela 3 oraz Tabela 4 przedstawiają podstawowe charakterystyki poszczególnych, wyodrębnionych zbiorów danych.

Tabela 3. Podstawowe charakterystyki zbioru uczącego

Zmienna	Wartość liczbowa	wartość sformatowana	Liczba wystąpień	Procent
label	.	fema	777	50.0967
label	.	male	774	49.9033

Źródło: Opracowanie własne przy użyciu oprogramowania SAS Enterprise Miner

Tabela 4. Podstawowe charakterystyki zbioru walidacyjnego

Zmienna	Wartość liczbowa	wartość sformatowana	Liczba wystąpień	Procent
label	.	fema	334	50.1502
label	.	male	332	49.8498

Źródło: Opracowanie własne przy użyciu oprogramowania SAS Enterprise Miner

5. Budowa modeli

5.1 Model Regresji Logistycznej

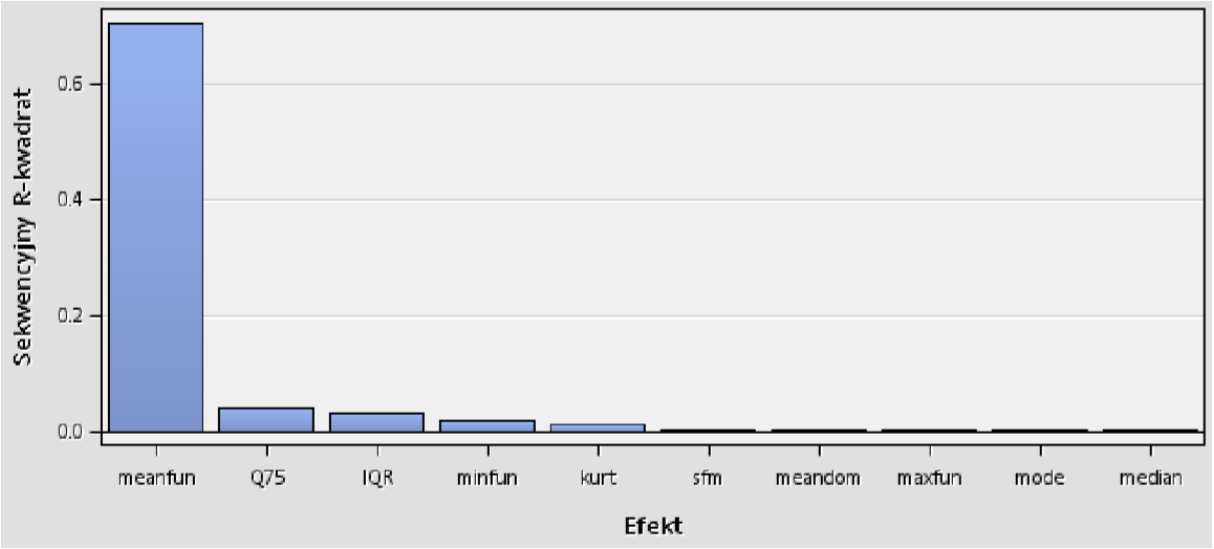
Podczas tworzenia projektu utworzono kilka modeli regresji logistycznej, poniżej zostaną przedstawione rezultaty najlepszego z nich. Najlepsze model uzyskano przy następującej konfiguracji opcji dostępnych w SAS Enterprise Miner:

- Uwzględnienie efektów głównych,
- Uwzględnienie interakcji dwuczynnikowych,
- Uwzględnienie potęg zmiennych, gdzie najwyższy stopień wielomiany wyniósł: 2,
- Funkcją łączącą był Logit

W kolejnym etapie projektu zbiór danych poddano selekcji zmiennych celem ujęcia w modelu jedynie istotnych statystycznie zmiennych, które znaczącym stopniu wpływają na zmienną zależną label, oznaczającą płeć badanego.

Wszystkie badane zmienne niezależne są zmiennymi przedziałowymi. Kryterium wybrane do ich selekcji to $R2$. Rezultaty selekcji zmiennych obrazuje Rysunek 2.

Rysunek 2. Zmienne wykazujące największy wpływ na zmienną celu



Źródło: Opracowanie własne przy użyciu oprogramowania SAS Enterprise Miner

Wykres 2 prezentuje efekt wykonania procesu selekcji zmiennych, gdzie wybrano ostatecznie 10 zmiennych z badanego zbioru danych, które mają największy wpływ na badaną zmienną celu oraz wnoszą tym samym do modelu najwięcej istotnych informacji.

Do grona 10 wyselekcjonowanych zmiennych objaśniających należy:

- meanfun,
- Q75,
- IQR,
- minfun,
- kurt,
- sfm,

- meandom,
- maxfun,
- mode,
- median

Pozostałe zmienne nie zostały uwzględnione w modelu.

W kolejnym kroku dla wyselekcjonowanych zmiennych wykonana została selekcja krokowa (stepwise). Każda iteracja została poddana ocenie poprzez kryterium informacyjne AIC. Modelem z najlepszymi wynikami okazał się ten z włączonymi interakcjami dwuczynnikowymi oraz wyrazami wielomianu stopnia drugiego. W tabeli 4 przedstawione zostały efekty, które zostały uwzględnione w modelu.

Tabela 5. Oszacowania parametrów modelu regresji logistycznej

Parametr	DF	Estimate	Standard Error	Wald-Chi-Square	Pr.>chi-kw.	Exp(est)
Intercept	1	-10.2136	3.4506	8.76	0.0031	0.00
IQR	1	0.3280	0.0531	38.20	<.0001	1.388
minfun	1	0.6041	0.2675	5.10	0.0240	1.830
IQR*IQR	1	-0.00073	0.000160	21.00	<.0001	0.999
IQR*meanfun	1	-0.00082	0.000343	5.67	0.0173	0.999
maxfun*meanfun	1	-0.00051	0.000142	12.88	0.0003	0.999
maxfun*minfun	1	0.00274	0.000678	16.37	<.0001	1.003
meanfun*meanfun	1	0.000895	0.000204	19.32	<.0001	1.001
meanfun*minfun	1	-0.00975	0.00210	21.62	<.0001	0.990
median*median	1	0.000099	0.000027	13.29	0.0003	1.000
median*sfm	1	-0.0746	0.0158	22.25	<.0001	0.928
minfun*sfm	1	0.1968	0.1039	3.59	0.0580	1.218

Źródło: Opracowanie własne na podstawie oprogramowania SAS Enterprise Miner

W Tabeli 5 znajdują się ponadto oceny parametrów, a wartości statystyki p-value potwierdzają statystyczną istotność zmiennych przy poziomie istotności wynoszącym 0.05. Jedynie zmienna minfun*fm jest nieistotna statystycznie, ponieważ jej wartość p-value przekracza poziom 0.05, dlatego ta zmienna nie będzie podlegać interpretacji.

Przykładowa interpretacja ilorazu szans w otrzymanym modelu dla zmiennej minfun jest następująca: wzrost minimalnej częstotliwości sygnału o 1 Hz zwiększa szanse na to, że badana osoba jest mężczyzną o 83 %. W tym miejscu należy nadmienić, iż zmienna minfun*sfm nie jest istotna statystycznie, tym samym nie podlega interpretacji ilorazu szans.

Przy wyborze najlepszego modelu istotną rolę odgrywają statystyki dopasowania. Tabela 5 prezentuje statystyki dopasowania wybranego, najlepszego modelu, z kolei tabela 6 obrazuje statystyki dopasowania jednego z odrzuconych modeli.

Tabela 6. Statystyki dopasowania najlepszego modelu regresji logistycznej

Zmienna celu	Statystyka dopasowania	Etykieta statystyk	Uczenie	Walidacja
label	_AIC_	Kryterium informacyjne Akaikego	183.4737	.
label	_ASE_	Przeciętny błąd kwadratowy	0.013257	0.021245
label	_MISC_	Odsetek błędnych klasyfikacji	0.019342	0.024024

Źródło: Opracowanie własne przy użyciu oprogramowania SAS Enterprise Miner

Tabela 7. Statystyki dopasowania jednego z odrzuconych modeli regresji logistycznej

Zmienna celu	Statystyka dopasowania	Etykieta statystyk	Uczenie	Walidacja
label	_AIC_	Kryterium informacyjne Akaikego	280.5527	.
label	_ASE_	Przeciętny błąd kwadratowy	0.020204	0.0201
label	_MISC_	Odsetek błędnych klasyfikacji	0.0245	0.0241

Źródło: Opracowanie własne przy użyciu oprogramowania SAS Enterprise Miner

Jak wynika z tabeli 5. wybrany model uzyskał wartość kryterium informacyjnego Akaike na poziomie 183.4737. Wartość ta była najniższa spośród wszystkich zbudowanych modeli regresji logistycznej. Statystyki dopasowania drugiego (odrzuconego) modelu, niezawierającego interakcji zmiennych i wyrazów wielomianu, zaprezentowane w tabeli 6. Gdzie widać, iż kryterium informacyjne Akaike było wyższe i wynosiło 280.5527. Wartości przeciętnego błędu kwadratowego oraz odsetek błędnych klasyfikacji w obu porównywanych modelach są na niskim poziomie, nie przekraczającym 3% (tym samym model nie jest nie douczony). Ważnym czynnikiem jest również to, iż wartości w zbiorach treningowym (nie zaprezentowanym w tabeli) oraz walidacyjnym nie różnią się znacząco. Dowodzi to faktu, iż wybrany model nie jest przeuczony.

Dla otrzymanego modelu utworzona została również macierz pomyłek (confusion matrix), która zobrazowana została w tabeli 7.

Tabela 8. Macierz pomyłek najlepszego modelu regresji logistycznej

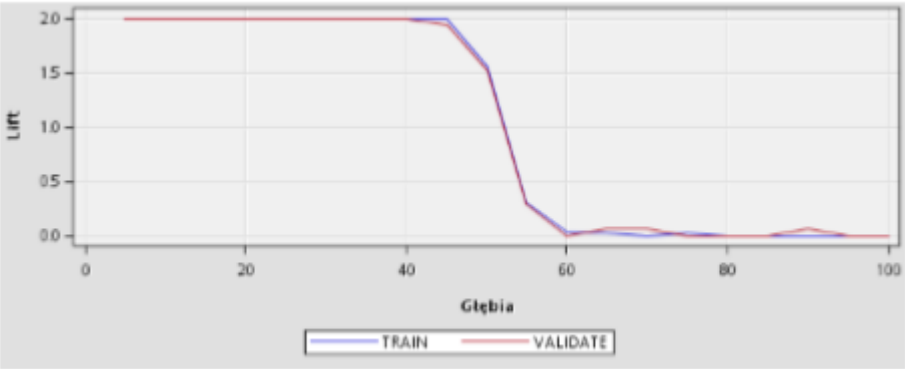
	0	1
0	327	9
1	7	323

Źródło: Opracowanie własne przy użyciu oprogramowania SAS Enterprise Miner

Zgodnie z tabelą 7. na zbiorze walidacyjnym model poprawnie zaklasyfikował 650 z 666 obserwacji. Ponadto odnotowano 7 klasyfikacji fałszywie dodatnich oraz 9 klasyfikacji fałszywie ujemnych. Taki rezultat macierzy pomyłek obrazuje, iż model poprawnie klasyfikuje jednostki.

Podczas analizy wybranego modelu regresji logistycznej dokonano również graficznej oceny jego jakości. Takim przykładem jest wykres krzywej LIFT zaprezentowany na rysunku 3.

Rysunek 3. Krzywa LIFT dla najlepszego modelu regresji logistycznej



Źródło: Opracowanie własne przy użyciu oprogramowania SAS Enterprise Miner

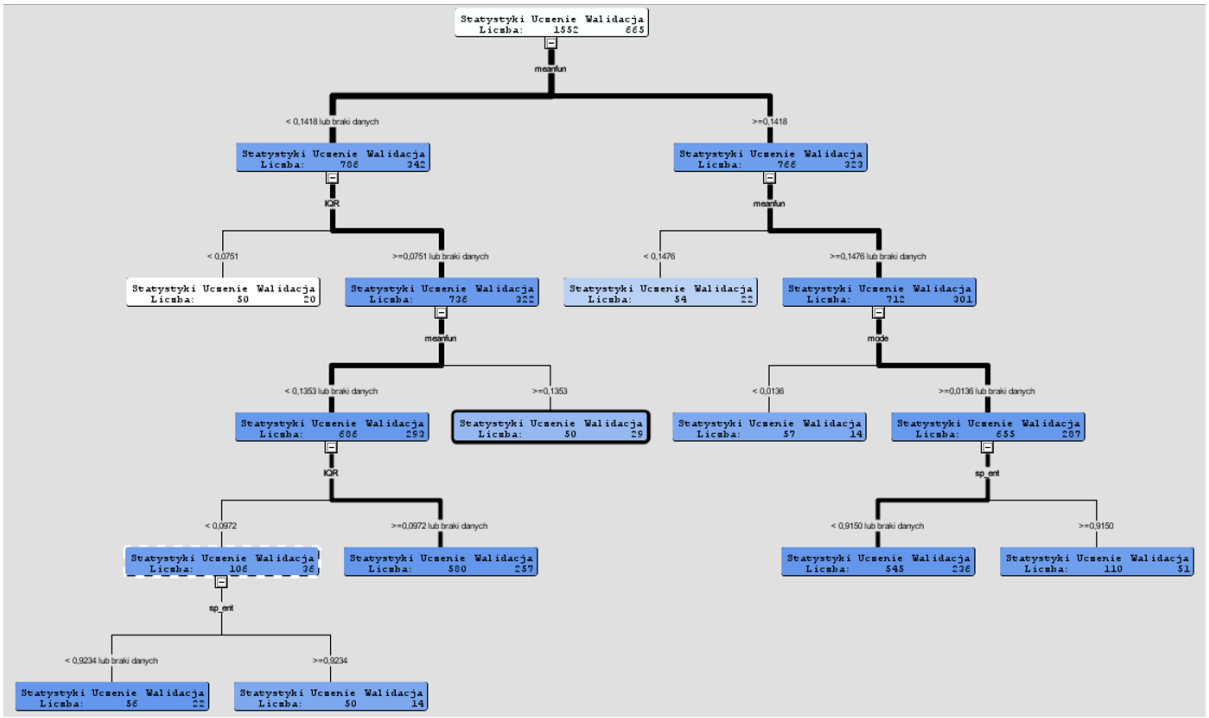
Zgodnie z rysunkiem 1. krzywe LIFT dla zbioru treningowego oraz walidacyjnego dla najlepszego modelu regresji logistycznej kształtują się podobnie, co pozytywnie świadczy o zbudowanym modelu, nie jest on przetrenowany. Kształt krzywych LIFT informuje, iż zysk z zastosowania wybranego, najlepszego modelu regresji logistycznej względem nie zastosowania go jest znaczący.

5.2 Model drzewa decyzyjnego

Procedurami odpowiedzialnymi za budowę drzewa decyzyjnego w oprogramowaniu SAS są: SPLIT oraz HPSPLIT. W SAS Enterprise Miner dostępnych jest wiele opcji konfiguracji budowy modelu drzewa decyzyjnego. Ostatecznie, spośród kilku zbudowanych modeli drzewa decyzyjnego, najlepsze okazało się drzewo decyzyjne binarne z maksymalną głębokością równą 5 i minimalną wielkością liścia równą 50. Ponadto okazało się, iż zmiany pozostałych dostępnych opcji jak na przykład kryterium podziału nie wpłynęło znacząco na różnicę w wynikach zbudowanych modeli.

Dla wybranego drzewa wartość wskaźnika testowego wyniosła 0.031276, a drzewo posiada 9 liści. Obraz uzyskanego drzewa decyzyjnego prezentuje rysunek 2.

Rysunek 4. Schemat najlepszego modelu drzewa decyzyjnego



Źródło: Opracowanie własne przy użyciu oprogramowania SAS Enterprise Miner

Zgodnie z uzyskanym schematem drzewa decyzyjnego kluczową zmienną dla odróżnienia mężczyzn od kobiet jest zmienna meanfun. W sytuacji, gdy średnia częstotliwość głosu w kHz jest mniejsza od 0.1418 lub nie posiadamy informacji, to osoba z wysokim prawdopodobieństwem będzie mężczyzną (w tym węźle znalazło się 749 mężczyzn i jedynie 37 kobiet). Po drugiej stronie stosunek odnośnie płci był podobny, jednakże na korzyść kobiet których w omawianym węźle znalazło się 742, natomiast mężczyzn zaledwie 24.

Kolejne rozgałęzienia dążą do jeszcze lepszego rozróżnienia płci danej osoby. Okazuje się, iż poza meanfun, ważnymi zmiennymi okazały się zmienne: IQR (rozstęp częstotliwości w kHz), mode (dominanta częstotliwości) oraz sp_ent(entropia epktralna). Na wizualizacji drzewa uwagę przykuwa liść wypełniony białym kolorem. Są to osoby z przeciętną częstotliwością mniejszą od 0.1418 oraz rozstępem częstotliwości mniejszym od 0.0751. Do omawianego liścia trafiło 25 mężczyzn oraz 25 kobiet, co wyjaśnia brak kolejnego podziału, ponieważ jak już wcześniej wspomniano, w opcjach ustawiona została minimalna wielkość liścia równa 50.

Podobnie jak w modelu regresji logistycznej tak i w modelu drzewa decyzyjnego należy poddać analizie statystyki dopasowania modelu drzewa decyzyjnego. Obrazuje je tabela 8.

Tabela 9. Statystyki dopasowania najlepszego modelu drzewa decyzyjnego

Statystyka	Próba	
	Ucząca	Walidacyjna
Odsetek błędnych klasyfikacji	0.0393	0.0466
Maksymalny błąd bezwzględny	0.9636	1

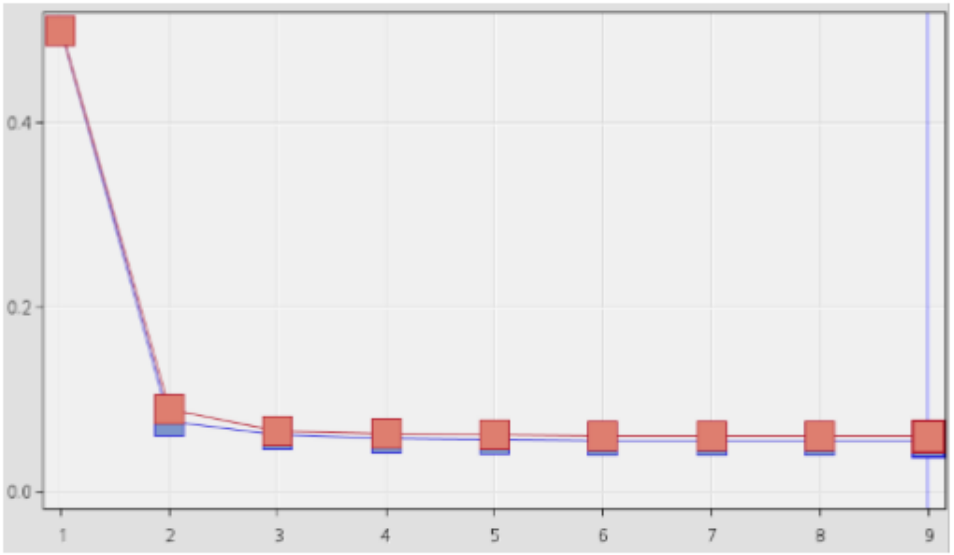
Średni błąd kwadratowy	0.0272	0.0297
Pierwiastek ze średniego błędu kwadratowego	0.1648	0.1724

Źródło: Opracowanie własne na podstawie oprogramowania SAS Enterprise Miner

Statystyki dopasowania najlepszego zbudowanego modelu drzewa decyzyjnego zobrazowane w tabeli 8 wskazują na dobrą wartość predykcyjną omawianego modelu. Odsetek błędnych klasyfikacji dla próby uczącej wynosi 0.0393, natomiast w walidacyjnej 0.0466. Spadek jakości pomiędzy próbami jest na akceptowalnym poziomie. Istotne są również niskie wartości średniego błędu kwadratowego, który w żadnej z prób nie przekroczył 3 %.

Kolejny istotny wykres zaprezentowany na rysunku 3. obrazuje wartości błędu standardowego w zależności od liczby liści w drzewie. Kolorem czerwonym oznaczone zostały wartości dla próby walidacyjnej, natomiast niebieskim dla próby uczącej.

Rysunek 5. Wartości błędu standardowego w zależności od liczby liści w drzewie

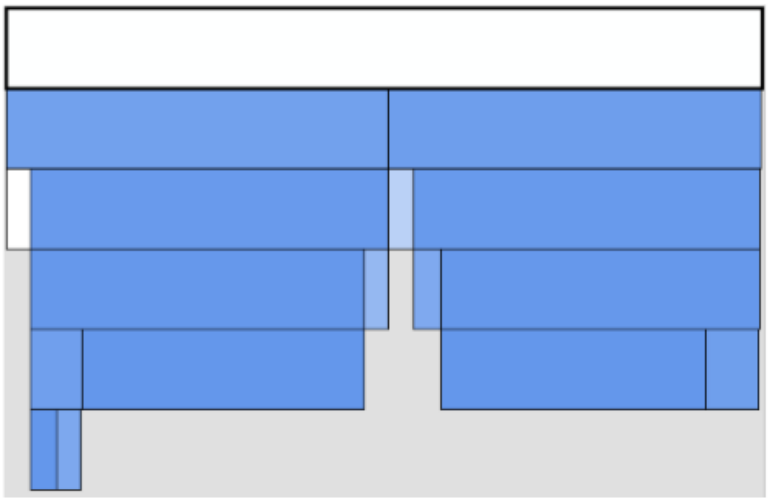


Źródło: Opracowanie własne przy użyciu oprogramowania SAS Enterprise Miner

Przedstawiony na rysunku 3. wykres prezentuje jaka korzyść płynie z kolejnych podziałów. Zgodnie z omawianym rysunkiem, po pierwszym, znaczącym spadku wartości, kolejne są już niewielkie. Jednakże, warto zauważyć, iż drzewo z 2 liśćmi było dość nie stabilne, ponieważ zauważalna była stosunkowo duża różnica między wartością średniego błędu kwadratowego dla próby uczącej oraz walidacyjnej. Należy również zaobserwować, iż kolejne podziały są tworzone na liściach, do których trafia coraz mniejsza liczba obserwacji, tym samym kolejne spadki są coraz mniejsze.

Kolejnym istotnym rysunkiem jest ten oznaczony numerem 3. Obrazujący wykres kafelkowy omawianego modelu drzewa decyzyjnego.

Rysunek 6. Wykres kafelkowy najlepszego modelu drzewa decyzyjnego

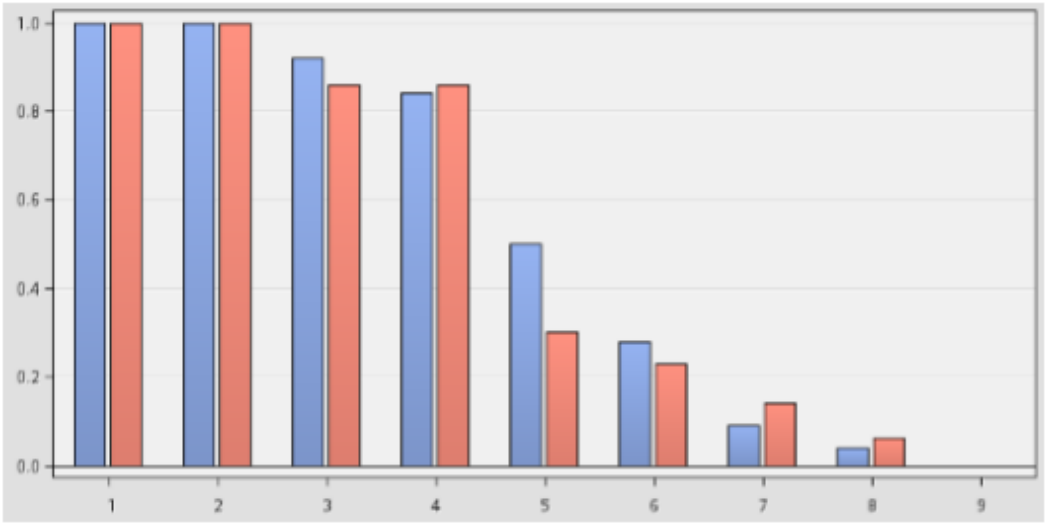


Źródło: Opracowanie własne przy użyciu oprogramowania SAS Enterprise Miner

Rysunek 3. prezentuje w jaki sposób pomiędzy poszczególne gałęzie została podzielona próba ucząca. Zgodnie z tym co prezentuje wspomniany rysunek, końcowo została ona podzielona na kilka mniejszych i większych części. W tym miejscu można się zastanowić na przycięciem drzewa i przykładowo usunąć jedyny podział na 5 poziomie, lecz finalnie podział nie został usunięty.

Następnym analizowanym rysunkiem będzie rysunek 4 obrazujące statystyki ilościowe omawianego modelu drzewa decyzyjnego.

Rysunek 7. Statystyki ilościowe najlepszego modelu drzewa decyzyjnego

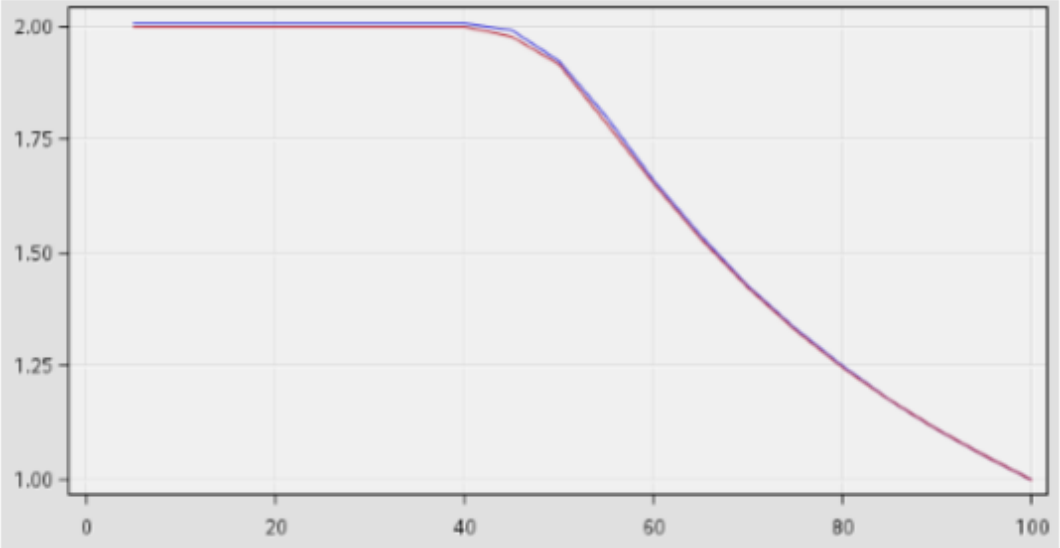


Źródło: Opracowanie własne przy użyciu oprogramowania SAS Enterprise Miner

Na rysunku 4. można zaobserwować ile procent obserwacji z danego węzła zostało zaliczonych do jednej kategorii. Wartości bliskie 1 lub 0 są oczekiwane w tej analizie, ponieważ mówią, iż dany liść dobrze różnicuje kobiety i mężczyzn. W większości liści w drzewie występują właśnie takie liście. Najbardziej problematyczny jest liść numer 5, który poza słabym odróżnianiem płci cechuje się niestabilnością między próbami, gdzie kolorem niebieskim oznaczona jest próba ucząca, a czerwonym próba walidacyjna.

Jak w przypadku modelu regresji logistycznej, tak samo w przypadku wybranego modelu drzewa decyzyjnego nastąpiła graficzna ocena jakości modelu przy użyciu wykresu krzywej LIFT zwizualizowanej na rysunku 5.

Rysunek 8. Krzywa LIFT dla najlepszego modelu drzewa decyzyjnego



Źródło: Opracowanie własne przy użyciu oprogramowania SAS Enterprise Miner

Kształt omawianej krzywej jest zgodny z oczekiwaniami i pokrywa się z dotychczasową analizą modelu. Dla ponad 40 % obserwacji z najwyższymi prawdopodobieństwami model jest dwukrotnie lepszy od podejścia losowego.

5.3 Model sieci neuronowej

Podobnie jak w poprzednich modelach, także tutaj spośród kilku utworzonych modeli sieci neuronowych wybrany został najlepszy z nich. Wybrany model sieci neuronowej posiadał 1 warstwę ukrytą, 5 neuronów w warstwie ukrytej, funkcją aktywacji COS, maksymalną liczbą iteracji wynoszącą 1 000 oraz dziewięcioma przeprowadzonymi procesami uczenia sieci dla różnych startowych wag. Poprzez wybór znacznej liczby iteracji, algorytm osiągnął zbieżność. Ponadto ustawiono ziarno inicjalizacji na poziomie 5 000 oraz kryterium wyboru modelu na podstawie minimalnej liczby błędnych klasyfikacji na zbiorze walidacyjnym. Wykorzystana została również opcja uczenia wstępnego, by wagi losowe nie były brane jako wartości początkowe w procesie uczenia sieci. Pozostałe parametry pozostały domyślne, sugerowane przez oprogramowanie SAS Enterprise Miner.

Wspomniany na początku projektu zbiór składający się z 2 217 obserwacji, przeznaczony do uczenia oraz walidacji został poddany partycjonowaniu w identyczny sposób jak uprzednio opisany. Wykorzystane zostało ziarno losowe równe 5 000, otrzymano zbiór uczący (70% próby) oraz walidacyjny (30% próby). Z powodu braku występowania braków danych nie wystąpiła potrzeba imputacji.

W pierwszej kolejności otrzymane rezultaty ocenione zostały na podstawie statystyk dopasowania modelu umieszczonych w tabeli 9.

Tabela 10. Statystyki dopasowania najlepszego modelu sieci neuronowej

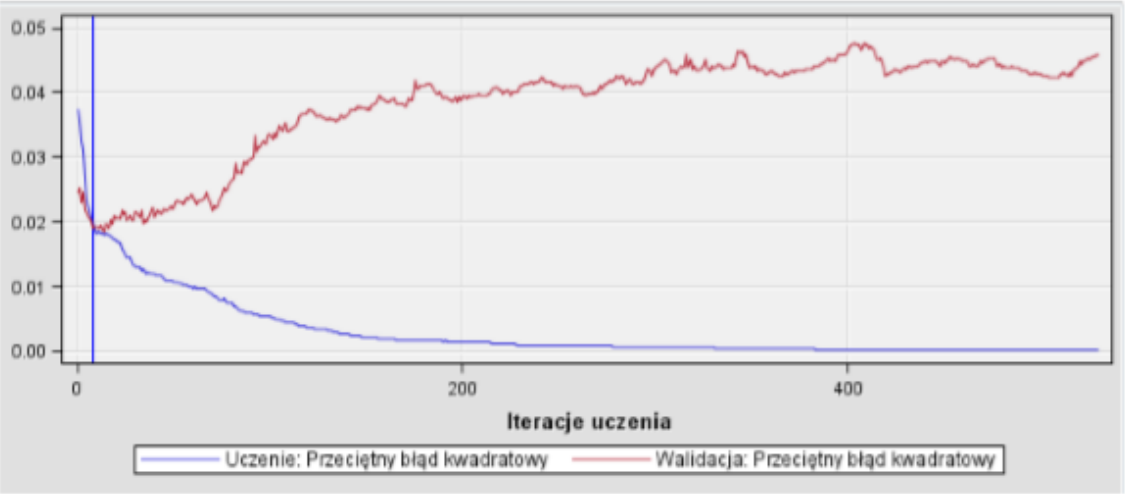
Statystyka	Próba	
	Ucząca	Walidacyjna
Odsetek błędnych klasyfikacji	0.0252	0.0180
Maksymalny błąd bezwzględny	0.9889	0.9950
Średni błąd kwadratowy	0.0220	0.0201
Pierwiastek ze średniego błędu kwadratowego	0.1484	0.0180

Źródło: Opracowanie własne na podstawie oprogramowania SAS Enterprise Miner

Zgodnie ze statystykami zaprezentowanymi w tabeli 9. przeciętny błąd kwadratowy na zbiorze walidacyjnym wyniósł 0.0201, co stanowi bardzo niski wynik i jednocześnie świadczy o dobrym dopasowaniu modelu do danych. Ponadto, omawiany błąd kształtował się na zbliżonym poziomie również na próbie uczącej, co potwierdza o braku przeuczenia modelu. Brak nadmiernego dopasowania do zbioru uczącego pozwala na uogólnianie jego wyników i przyjmowanie ich jako wiarygodnych. Podobnie odsetek błędnych klasyfikacji jest niski i nie przekraczający 2 % na zbiorze walidacyjnym.

Kolejny etap analizy wybranego modelu sieci neuronowych dotyczy wykresów średniego błędu kwadratowego oraz odsetka błędnych klasyfikacji, które zostały zobrazowane na rysunkach 6 oraz 7.

Rysunek 9. Średni błąd kwadratowy



Źródło: Opracowanie własne przy użyciu oprogramowania SAS Enterprise Miner

Z rysunku 6. można wywnioskować, iż proces optymalizacji potrzebował 9 iteracji do osiągnięcia zbieżności. Pionowa, niebieska linia oznacza moment, gdzie wartość średniego błędu kwadratowego dla zbioru walidacyjnego zaczyna rosnąć, zwiększając tym samym różnicę wartości pomiędzy próbami uczącą oraz walidacyjną. Obrazuje to problem przeuczenia, który wymusza w tym momencie zaprzestanie uczenia sieci.

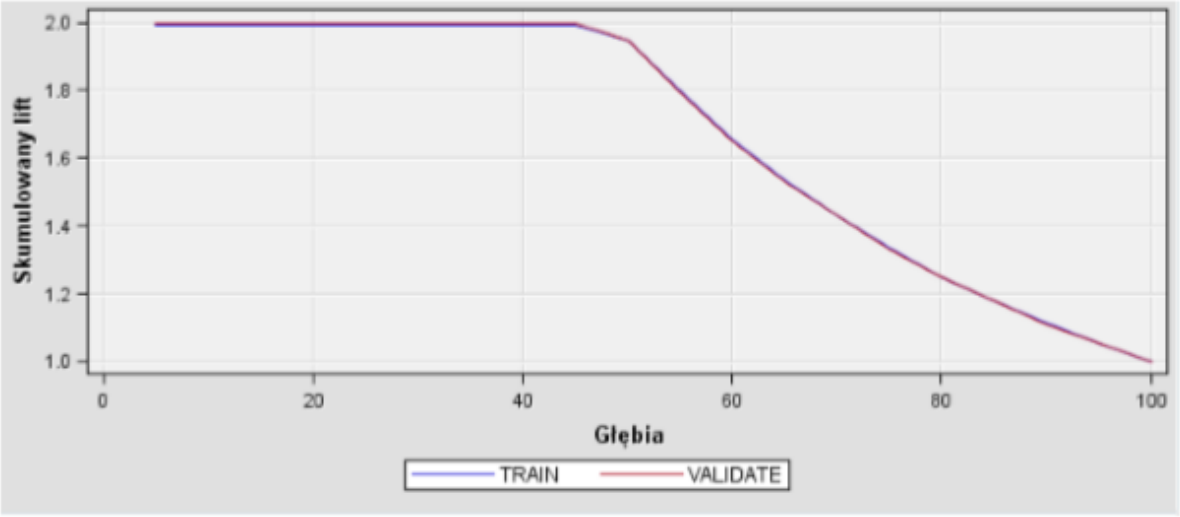
Rysunek 10. Odsetek błędnych klasyfikacji

Źródło: Opracowanie własne przy użyciu oprogramowania SAS Enterprise Miner

Podobnie sytuacja wygląda w przypadku odsetka błędnych klasyfikacji. Proces optymalizacji potrzebował również 9 iteracji do osiągnięcia zbieżności. W dalszych krokach po przekroczeniu pionowej linii, wartość odsetka błędnych klasyfikacji dla zbioru walidacyjnego zaczyna rosnąć, tym samym konieczne jest przerwanie uczenia siec neuronowej.

Podobnie jak w przypadku dwóch wcześniej przeanalizowanych modeli (regresja logistyczna, drzewo decyzyjne), tak i w kwestii sieci neuronowej w analizie modelu wykorzystana została metoda graficzna w postaci krzywej LIFT, co zostało zaprezentowane na rysunku 8.

Rysunek 11. Krzywa LIFT najlepszego modelu sieci neuronowych



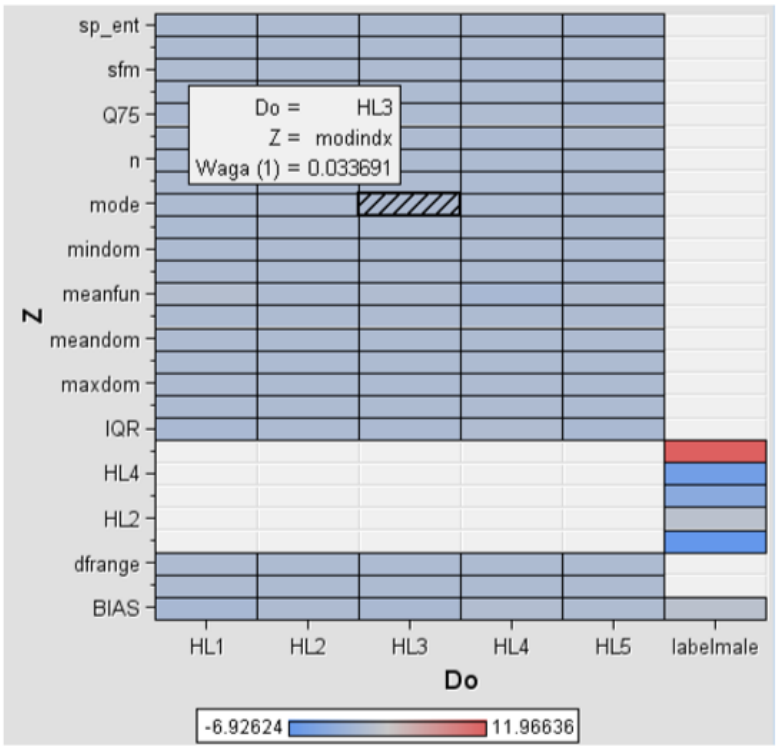
Źródło: Opracowanie własne przy użyciu oprogramowania SAS Enterprise Miner

Zgodnie z tym co obrazuje rysunek 8. Krzywe LIFT dla zbioru treningowego oraz walidacyjnego przebiegają w bardzo zbliżony sposób, co pozytywnie świadczy o analizowanym modelu. Zarówno kształt jak i położenie obu krzywych ukazuje, iż zysk z zastosowania modelu, w porównaniu do stosowania modelu losowego jest znaczący, tym samym skuteczność predykcji modelu można ocenić jako bardzo dobrą. Jak można zauważyć,

dla 45 % obserwacji z oszacowanym najwyższym prawdopodobieństwem, omawiany model jest dwukrotnie lepszym klasyfikatorem niż model losowy.

Na kolejnym rysunku (rysunek 9.) zwizualizowany został schemat wag końcowych dla analizowanego, najlepszego modelu sieci neuronowych. Na wejściu neuron otrzymuje zestaw sygnałów, czyli dane wejściowe. Dane z warstwy wejściowej sumowane są przez funkcję kombinacji, która wymnaża je przez odpowiednie wartości (wagi), a otrzymana suma jest traktowana jako argument funkcji aktywacji.

Rysunek 12. Wagi końcowe najlepszego modelu sieci neuronowej

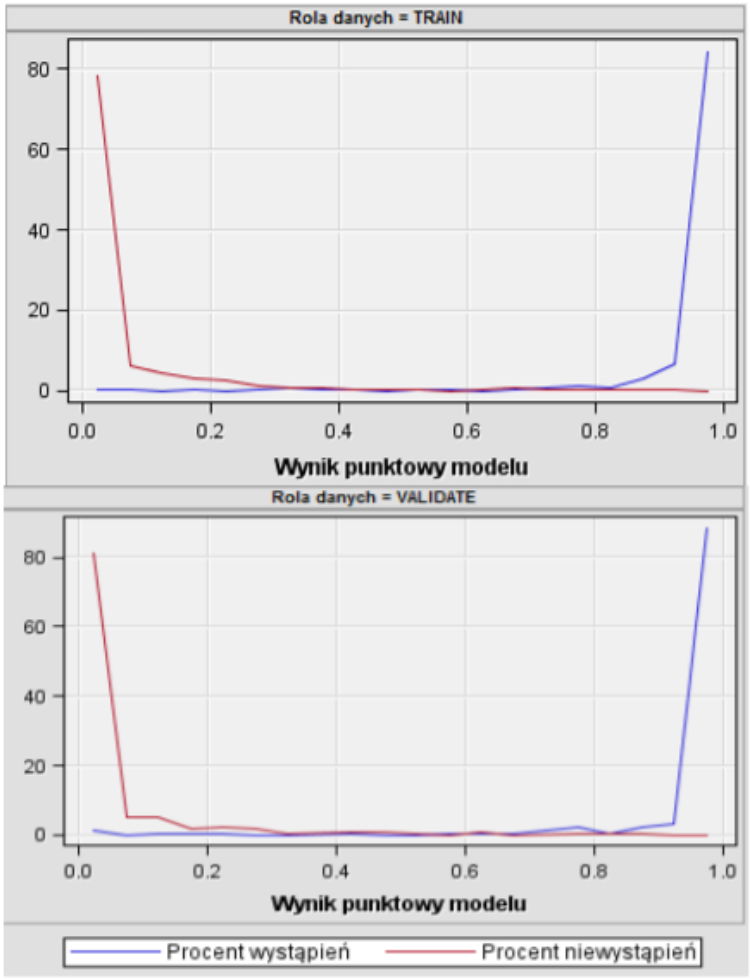


Źródło: Opracowanie własne przy użyciu oprogramowania SAS Enterprise Miner

Zgodnie z rysunkiem 9. w przypadku omawianego modelu sieci neuronowej wykorzystana została jedna warstwa ukryta, złożona z pięciu neuronów. Sieć jest pełna, gdyż każdy neuron w danej warstwie jest połączony z wszystkimi neuronami z kolejnej warstwy. Otrzymane wagi odpowiadają konkretnym połączeniom, a ich wartości zostały przedstawione przy użyciu odcieni. Poziomy ujemne zostały zobrazowane za pomocą odcieni koloru niebieskiego, dodatnie za pomocą koloru czerwonego. Najwyższy poziom wagi dla połączenia został oszacowany na poziomie blisko 11.97. Dla każdego połączenia możliwe jest sprawdzenie skąd i dokąd prowadzi oraz jaka waga została mu przypisana.

Kolejna wykorzystana w analizie wizualizacja odnosi się do rozkładu ocen omawianego modelu sieci neuronowej. Wspomniany rozkład został zaprezentowany na rysunku 10.

Rysunek 13. Rozkład ocen najlepszego modelu sieci neuronowej



Źródło: Opracowanie własne przy użyciu oprogramowania SAS Enterprise Miner

Wizualizacja przedstawiona na rysunku 10. ukazuje sposób w jaki kształtuje się rozkład wystąpień oraz niewystąpień badanego zjawiska w oparciu o predykcje wygenerowane przez analizowany model sieci neuronowej. Kształty wykresów zarówno dla zbioru treningowego jak i walidacyjnego są podobne, co pozytywnie świadczy o stabilności modelu. Ocena punktowa w przypadku niewystąpień skoncentrowana jest wokół 0, natomiast dla wystąpień wokół 1, pozwala to na wyciągnięcie wniosku, iż w analizowanym modelu sieci neuronowej występuje dobre zróżnicowanie obu populacji.

Ostatnim krokiem w analizie najlepszego modelu sieci neuronowej jest odniesienie się do tabel klasyfikacji zdarzeń, confusion matrix, dla zbiorów: treningowego oraz walidacyjnego, co prezentuje tabela 10.

Tabela 11. Tabela klasyfikacji zdarzeń

Tabela klasyfikacji zdarzeń

Rola danych=TRAIN Zmienna celu=label Etykieta zmiennej celu=' '

Fałszywie ujemne	Prawdziwie ujemne	Fałszywie dodatnie	Prawdziwie dodatnie
18	752	21	759

Rola danych=VALIDATE Zmienna celu=label Etykieta zmiennej celu=' '

Fałszywie ujemne	Prawdziwie ujemne	Fałszywie dodatnie	Prawdziwie dodatnie
8	327	6	326

Źródło: Opracowanie własne przy użyciu oprogramowania SAS Enterprise Miner

Znajdujące się w tabeli 10. wyniki obrazują, że na zbiorze walidacyjnym model sieci neuronowej przyporządkowywał błędnie jedynie pojedyncze obserwacje. Mając na uwadze, iż klasyfikacje fałszywie ujemne oraz fałszywie dodatnie kształtują się na zbliżonym poziomie, można wyciągnąć wniosek, że analizowany model nie ma problemu z gorszym rozpoznawaniem którejś z kategorii zmiennej zależnej.

6. Ocena i ewaluacja zbudowanych modeli

Każdy z przeanalizowanych modeli: regresji logistycznej, drzewa decyzyjnego oraz sieci neuronowej został zbudowany pod kątem jak najwyższego dopasowania oraz minimalizacji błędów, celem jak najlepszego odróżnienia płci badanej osoby. Wymienione modele zostały poddane ocenie i porównane między sobą pod kątem poziomu dopasowania jak również efektywności.

Celem porównania trzech analizowanych modeli użyty został węzeł porównania modeli dostępny w oprogramowaniu SAS Enterprise Miner. W tym miejscu głównie zostaną wymienione statystyki na próbie walidacyjnej.

Tabela 12. Porównanie odsetków błędnych klasyfikacji w modelach

Model	Odsetek błędnych klasyfikacji
Regresja logistyczna	0.024
Drzewo decyzyjne	0.047
Sieć neuronowa	0.018

Źródło: Opracowanie własne na podstawie oprogramowania SAS Enterprise Miner

Zaprezentowane w tabeli 11. porównanie odsetków błędnych klasyfikacji analizowanych modeli przemawia za wyborem modelu sieci neuronowej. Omawiana tabela i statystyka w niej ujęta szczególnie obnaża słabość drzewa decyzyjnego.

Do porównania modeli wykorzystana została również statystyka przeciętnego błędu kwadratowego zaprezentowana w tabeli 12.

Tabela 13. Porównanie przeciętnych błędów kwadratowych w modelach

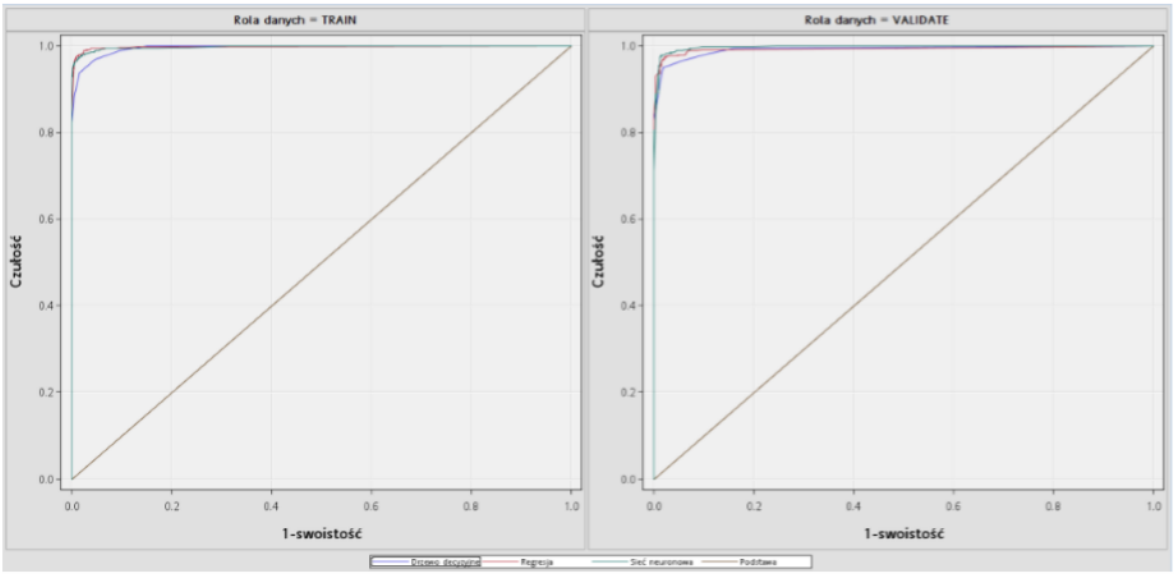
Model	Przeciętny błąd kwadratowy
Regresja logistyczna	0.021
Drzewo decyzyjne	0.029
Sieć neuronowa	0.02

Źródło: Opracowanie własne na podstawie oprogramowania SAS Enterprise Miner

Podobnie jak przy porównaniu modeli za pomocą odsetków błędnych klasyfikacji, tak samo w przypadku porównania przeciętnych błędów kwadratowych modeli drzewo decyzyjne prezentuje najsłabszy wynik. Wartości przytoczonych przeciętnych błędów kwadratowych sugerują, że przynajmniej dla sieci neuronowej oraz regresji logistycznej wyniki wskazują na dobrą jakość wymienionych modeli i potwierdzają istnienie reguł zachodzących między zmiennymi objaśniającymi, a zmienną objaśnianą.

Do porównania modeli wykorzystana została również krzywa ROC, dla każdego z trzech modeli, zobrazowana na rysunku 11.

Rysunek 14. Porównanie krzywych ROC dla analizowanych modeli



Źródło: Opracowanie własne przy użyciu oprogramowania SAS Enterprise Miner

Zgodnie z rysunkiem 11. obrazującym krzywe ROC dla zbioru treningowego oraz walidacyjnego wszystkie trzy modele są wynikami zbliżone do siebie, pole pod krzywymi ROC mają podobną wielkość.

Decyzja, który z wyżej omówionych modeli w tym wypadku zależy od kilku czynników. Łatwość w interpretacji oraz przejrzystą i atrakcyjną wizualizację oferuje drzewo decyzyjne. Jednakże w omawianym przypadku wadą tego modelu są przeciętne statystyki dopasowania. Model sieci neuronowej z kolei nie jest najprostszy w interpretacji oraz nie posiada atrakcyjnej wizualizacji, jednakże jest w stanie nauczyć się postępowania z danymi, które nie niosą żadnej informacji lub są błędne. Ostatni z omawianych modeli, model regresji logistycznej, cechuje się łatwością w interpretacji wyników i opisie zależności, jednakże w

odróżnieniu od modeli drzewa decyzyjnego czy sieci neuronowych, nie wychwytuje bardzo złożonych i nieliniowych zależności w danych oraz ma słabsze moce predykcyjne.

Ostatecznie, najlepszym modelem będzie model sieci neuronowych. Pomimo, iż interpretacja wyników sieci neuronowych jest trudniejsza niż w przypadku drzewa decyzyjnego czy regresji logistycznej, to model ten uzyskał bardzo dobre statystyki dopasowania oraz jest stabilny.

Podsumowanie

Nadrzędnym celem projektu było porównanie modeli regresji logistycznej, sieci neuronowych oraz drzewa decyzyjnego i wybranie tego, który będzie w stanie najlepiej różnicować płeć badanych osób na podstawie danych o ich głosie. Na potrzeby projektu zbudowanych zostało kilka różnych modeli, jednak do porównania wybrane zostały 3 najlepsze modele. Celem budowy modeli było osiągniecie jak najlepszego dopasowania oraz minimalizacja błędów w procesie rozpoznawania płci badanego.

Wszystkie trzy modele dla zbiorów: uczącego oraz walidacyjnego, na podstawie wykresów krzywej ROC oraz pola pod krzywą ROC , zaprezentowały podobne rezultaty. Ostatecznie jednak wybrany został model sieci neuronowej jako ten prezentujący minimalnie najlepsze rezultaty.

Spis rysunków

Spis tabel