Raport Alzheimer

Natan Warelich

Spis treści

1 Wprowadzenie

Raport alzheimera. Wykorzystany model to MLP

2 Eksploracyjna analiza danych

2.1 Wczytanie danych

Dane wczytujemy wykorzystując bibliotekę pandas.

```
import pandas as pd

df = pd.read_csv("alzheimer_wersja1.csv", sep=";", header=0)
```

Na wstępie napotkano problem niekompatybilności formatu danych z narzędziami analitycznymi – wartości zmiennoprzecinkowe były zapisane z użyciem przecinka jako separatora dziesiętnego. W celu zapewnienia poprawnej interpretacji danych przez oprogramowanie, zastosowano konwersję formatu, jak przedstawiono w poniższym fragmencie kodu:

```
for col in df.select_dtypes(include='object').columns:
    df[col] = (df[col].str.replace(',', '.', regex=False)
                .astype(float))
```

Teraz dane są gotowe do przetwarzania.

2.2 Zbadanie zgodności danych z opisem

Wszystkie dane były sprawdzone pod względem zawartości, oraz zgodności z założeniami. Schemat sprawdzania wyglądał w następujący sposób:

```
age_check = dane['Age'].between(60, 90).all()
print(f"age check: {age_check}")

gender_check = dane['Gender'].unique()
print(f"gender check: {gender_check}")
```

Wyniki analizy wskazują, że dane są kompletne i zgodne z opisem, co umożliwia przeprowadzenie dalszej analizy bez potrzeby imputacji czy odrzucania rekordów.

2.3 Rozkład wartości, korelacje danych

Istotnym aspektem przygotowania danych do modelowania jest analiza rozkładu wartości zmiennych jakościowych. Jak wskazuje poniższy wykres, większość tych zmiennych zawiera przewagę wartości zerowych, co może wpływać na wyniki modelowania

```
column Gender contains 49.37% 0
column Smoking contains 71.15% 0
column FamilyHistoryAlzheimers contains 74.78% 0
column MemoryComplaints contains 79.20% 0
column BehavioralProblems contains 84.32% 0
column DifficultyCompletingTasks contains 84.13% 0
column Forgetfulness contains 69.85% 0
column Diagnosis contains 64.63% 0
```

Kolejnym krokiem była analiza korelacji zmiennych z cechą docelową – diagnozą. Spośród analizowanych zmiennych tylko trzy wykazały istotny poziom korelacji. Zmienne związane z paleniem tytoniu oraz zapominalstwem nie wykazują korelacji z diagnozą i zostały wykluczone z dalszej analizy w modelach klasyfikacyjnych.

```
Correlation between Diagnosis and Age: -0.01
Correlation between Diagnosis and Gender: -0.02
Correlation between Diagnosis and BMI: 0.03
Correlation between Diagnosis and Smoking: -0.00
Correlation between Diagnosis and AlcoholConsumption: -0.01
Correlation between Diagnosis and PhysicalActivity: 0.01
Correlation between Diagnosis and FamilyHistoryAlzheimers: -0.03
Correlation between Diagnosis and CholesterolTotal: 0.01
Correlation between Diagnosis and MemoryComplaints: 0.31
Correlation between Diagnosis and BehavioralProblems: 0.22
Correlation between Diagnosis and ADL: -0.33
Correlation between Diagnosis and DifficultyCompletingTasks: 0.01
Correlation between Diagnosis and Forgetfulness: -0.00
```

3 Model MLP

3.1 Omówienie ustawień modelu

Ze względu na znikomy wpływ na diagnozę większości zmiennych rozsądnym będzie przeprowadzenie dwóch eksperymentów:

- Uwzględniający tylko dane wpływowe
- Omawiający wszystkie dane, które wstępnie nie zostały wykluczone

Badania wykonano przy następujących hiperparametrach danych:

```
random_state = 318153
test_size = 0.25
```

Model MLP został zdefiniowany zgodnie z poniższymi parametrami

```
mlp = MLPClassifier(hidden_layer_sizes=(32, 16),
                    activation='relu',
                    solver='adam',
                    max_iter=1000,
                    random_state=318153)
```

3.2 Wstępne sprawdzenie modelu

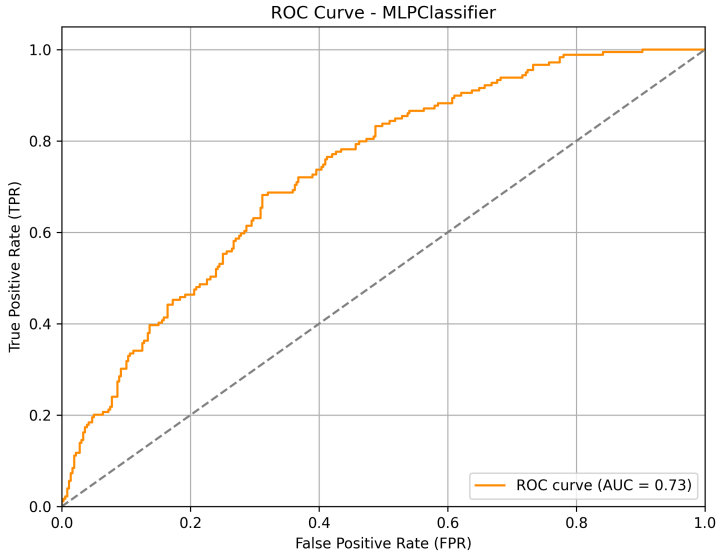
Dla zdefiniowanego modelu wstępne statystystyki na danych prezentują się następująco

Tabela 1: Uzyskane wyniki na zbiorze uczącym dla modelu z istotnymi cechami

Klasa	**Precyzja**	**Czułość (Recall)**	**F1-score**	**Próbka (Support)**
0	76%	86%	81%	1030
1	68%	52%	59%	581
Trafność (Accuracy)			74%	1611

Tabela 2: Uzyskane wyniki na zbiorze testowym dla modelu z istotnymi cechami

Klasa	**Precyzja**	**Czułość**	**F1**	**Próbka**
0	77%	84%	81%	359
1	62%	51%	56%	179
Trafność (accuracy)			73%	538



Rysunek 1: Wykres 1: Krzywa ROC modelu MLP na zbiorze testowym

Jak widać model ten, nadaje się do wykrywania osób zdrowych, lecz w kwestii chorych 51% czułości wskazuje raczej na to, że model zgaduję, aniżeli widzi jakiś patent. Model ten ma praktycznie identyczne trafność na obydwu zbiorach

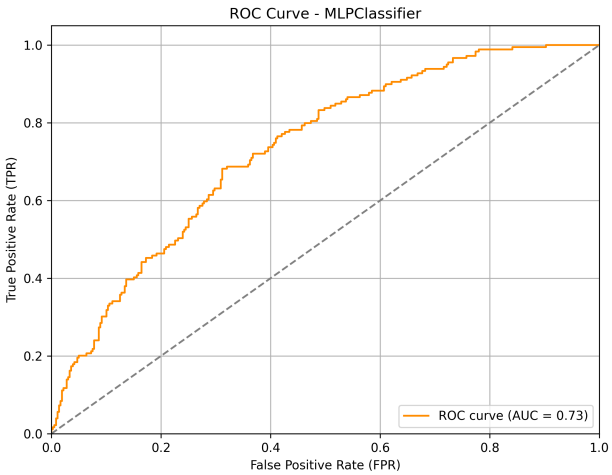
W przypadku uwzględnienia wszystkich cech, w tym mniej istotnych, model wykazuje zwiększone przeuczenie — trafność na zbiorze uczącym wynosi 88,33%, a na testowym spada do 68,03%. W związku z tym dalsze badanie bedzie odbywać się tylko dookoła zmiennych mających wysokie korelacje z diagnozą. Szczegółowe metryki to:

Tabela 3: Wyniki na zbiorze uczącym dla modelu MLP (wszystkie cechy)

Klasa	**Precyzja**	**Czułość**	**F1**	**Próbka**
0	89%	94%	91%	1030
1	88%	79%	83%	581
Trafność (accuracy)			88%	1611

Tabela 4: Wyniki na zbiorze testowym dla modelu MLP (wszystkie cechy)

Klasa	**Precyzja**	**Czułość**	**F1**	**Próbka**
0	75%	77%	76%	359
1	52%	49%	51%	179
Trafność (accuracy)			68%	538



Rysunek 2: Wykres 2: Krzywa ROC modelu MLP na zbiorze testowym(wszystkie cechy)

Wyniki te potwierdzają, że mniej istotne dane wprowadzają szum, obniżając jakość modelu, co jest także widoczne na krzywej ROC — pole pod krzywą jest mniejsze i mniej korzystne.

Z czego może wynikać nasz stosunkowo niesatysfakcjonujący wynik? Powodów może być wiele od słabej jakości modelu, aż po nieodpowiednie dane. Jednak najprawdopodobniej w tym przypadku przyczyna leży w wspomnianej wcześniej dysproporcji w ilości zer w zmiennych jakościowych. Powoduje to, że model zamiast faktycznie nauczyć się rozpoznawać wszystkie przypadki równomiernie, uczy się głównie wzorców charakterystycznych dla osób zdrowych.

3.3 Próby poprawy jakości modelu

W celu poprawy jakości naszego modelu możemy zastosować techniki przedstawione poniżej.

3.3.1 Zmiana funkcji aktywacyjnej/algorytmu optymalizacji

Założenie tego jest banalne. Uczymy nasz algorytm przy użyciu zupełnie innych metod, aby sprawdzić czy osiągniemy lepsze wyniki. Niestety, każdy z parametrów funkcji aktywacyjnej (logistic, identity, tanh) oraz algorytmów optymalizacji (lbfgs, sgd) powodowały zwiększanie się przeuczenia, oraz obniżanie ogólnej jakości modelu.

3.3.2 Edycja warstwy ukrytej

Przy tym rozwiązaniu edytujemy nasze warstwy ukryte poprzez np. zmianę liczby neuronów, zwiększenie liczby warstw itp. Analogicznie jak w podpunkcie 1. rozwiązanie te sprawiało, że zbiór uczący coraz bardziej dążył do przeuczenia.

3.3.3 Edycja danych

Istnieje szansa, że jeżeli przeszkolimy model na wyrównanej liczbie danych, poprzez oversampling/downsampling to nasza maszyna dzięki temu nie będzie faworyzować jednego typu danych. Jako, iż nadmiarowo posiadamy zera, adekwatnym do sytuacji będzie redukcja danych osób zdrowych.

Tutaj warto zaznaczyć, iż dla iteracji równej tysiąc python wskazuje błąd

```
ConvergenceWarning: Stochastic Optimizer: Maximum iterations
1000 reached and the optimization hasn't converged yet.
```

W związku z czym w przypadku downsamplingu ilość iteracji będzie wynosić 1300.

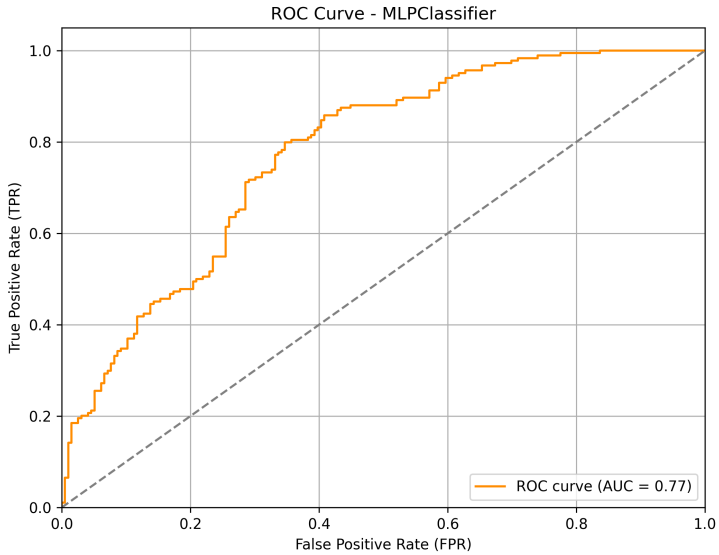
Natomiast statystyki prezentują się następująco

Tabela 5: Wyniki modelu po downsamplingu – zbiór uczący

Klasa	**Precyzja**	**Czułość**	**F1**	**Próbka**
0	100%	100%	100%	564
1	100%	100%	100%	576
Trafność (accuracy)			100%	1140

Tabela 6: Wyniki dla modelu po downsamplingu zbioru

Klasa	Precyzja	Czułość	F1	Próbka
0	66%	65%	66%	196
1	63%	65%	64%	184
Trafność (accuracy)			65%	380



Rysunek 3: Wykres 3: Krzywa ROC modelu MLP na zbiorze testowym dla downsamplingu

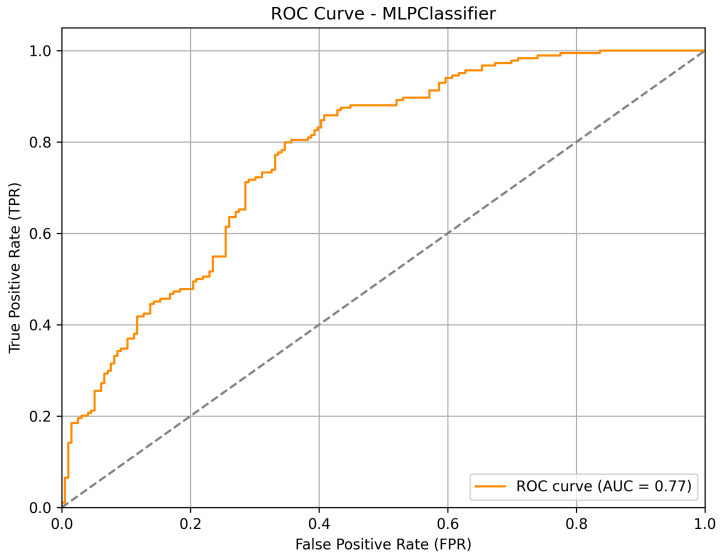
W tym przypadku model wyuczył się wszystkich danych na pamięć. Przy zmniejszeniu znacznym iteracji do np 100. Model faktycznie ma optymalne wyniki,

Tabela 7: Wyniki modelu na zbiorze treningowym po downsamplingu (wersja z zredukowanymi iteracjami)

Klasa	Precyzja	Czułość	F1	Próbka
0	81%	75%	78%	564
1	77%	83%	80%	576
Trafność (accuracy)			79%	1140

Tabela 8: Wyniki modelu na zbiorze testowym po downsamplingu (wersja z zredukowanymi iteracjami)

Klasa	Precyzja	Czułość	F1	Próbka
0	73%	69%	71%	196
1	69%	72%	71%	184
Trafność (accuracy)			71%	380



Rysunek 4: KRzywa ROC MLP downsampling min

Jednakże przeuczenie dalej jest widoczne, więc warto rozważyć jeszcze inne rozwiązania.

3.3.4 Dane walidacyjne

Kolejnym sposobem na rozwiązanie problemu jest wprowadzenie danych walidacyjnych. Dane te mają na celu skorygować model w trakcie szkolenia, w celu uzyskania większej dokładności w ostatecznym rozrachunku. W przypadku zastosowania tej techniki, należy spojrzeć również na wyniki walidacji.

Tabela 9: Wyniki modelu na zbiorze treningowym

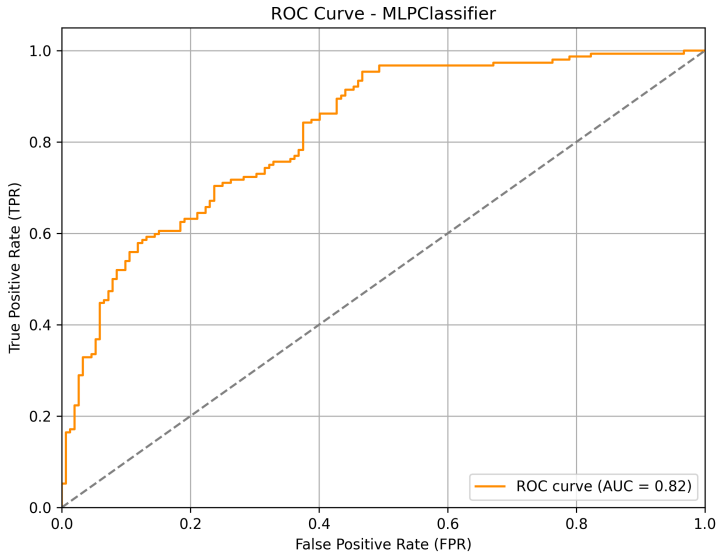
Klasa	Precyzja	Czułość	F1	Próbka
0	76%	90%	83%	831
1	73%	50%	59%	458
Trafność (accuracy)			76%	1289

Tabela 10: Wyniki modelu na zbiorze walidacyjnym

Klasa	Precyzja	Czułość	F1	Próbka
0	73%	87%	79%	277
1	64%	43%	52%	153
Trafność (accuracy)			71%	430

Tabela 11: Wyniki modelu na zbiorze testowym

Klasa	Precyzja	Czułość	F1	Próbka
0	75%	88%	81%	281
1	66%	45%	53%	149
Trafność (accuracy)			73%	430



Rysunek 5: Wykres 4: Krzywa ROC modelu MLP na zbiorze testowym dla danych walidacyjnych

Jak widać jest to model odrobinę lepszy od pierwotnego, jednakże dalej występuje problem z diagnozą chorych.
Warto jeszcze tutaj przyjrzeć się krzywej, wyglądowo coraz bliżej jej do "perfekcji".
Skoro rezultaty dla szkolenia zarówno stosując dane walidacyjne, jak i undersampling w pewnym sensie były w stanie coś ulepszyć, to intuicyjnie powinniśmy jeszcze sprawdzić, co się stanie, gdy zastosujemy te dwie metody naraz.

3.3.5 Użycie walidacji + downsampling'u

Tutaj od razu warto pochwalić się rezultatami gdyż one najwięcej nam powiedzą o skuteczności tego rozwiązania.

Tabela 12: Wyniki modelu na zbiorze treningowym (walidacja + downsampling)

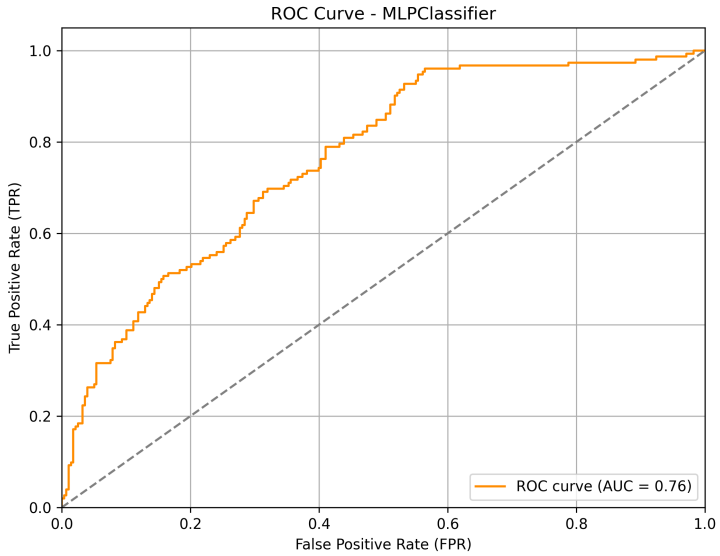
Klasa	**Precyzja**	**Czułość**	**F1**	**Próbka**
0	84%	65%	73%	456
1	72%	88%	79%	456
Trafność (accuracy)			76%	912

Tabela 13: Wyniki modelu na zbiorze walidacyjnym (walidacja + downsampling)

Klasa	**Precyzja**	**Czułość**	**F1**	**Próbka**
0	79%	61%	69%	152
1	68%	84%	75%	152
Trafność (accuracy)			72%	304

Tabela 14: Wyniki modelu na zbiorze testowym (walidacja + downsampling)

Klasa	**Precyzja**	**Czułość**	**F1**	**Próbka**
0	77%	62%	69%	152
1	68%	81%	74%	152
Trafność (accuracy)			72%	304



Rysunek 6: Wykres 5: Krzywa ROC modelu MLP na zbiorze testowym dla walidacji + downsampling

Uzyskane wyniki wskazują, że model, który wcześniej wykazywał się skutecznością głównie w identyfikacji osób zdrowych, został znacząco usprawniony. Pomimo utrzymania się ogólnej trafności na podobnym poziomie, model ten charakteryzuje się wyraźną poprawą w rozpoznawaniu przypadków choroby. Statystyki dotyczące wykrywania pacjentów chorych, a zwłaszcza czułość, są na tyle wysokie, że wykluczają przypadkowe trafienia.
Warto również zauważyć, że pole pod krzywą ROC nie jest najwyższe w porównaniu do wcześniejszych wersji modelu, jednakże jego znaczenie jest relatywnie niewielkie. W tej wersji model osiąga bowiem najlepsze wyniki w zakresie wykrywania przypadków chorobowych, co czyni go najbardziej efektywnym spośród analizowanych wariantów.

3.4 Podsumowanie

Reasumując, w celu uzyskania modelu będącego w stanie wykrywać chorobę Alzheimera na podstawie otrzymanych danych podjęliśmy następujące kroki:

1. Zdefiniowanie modelu MLP:
    - określenie architektury warstw ukrytych,
    - ustalenie podziału danych na zbiory uczący i testowy.
2. Wstępna analiza:
    - eliminacja modeli uwzględniających dane o niskiej istotności,
    - identyfikacja problemu z wykrywaniem przypadków choroby.
3. Nieefektywna próba rozwiązania poprzez zmiany parametrów:
    - modyfikacja struktury warstw ukrytych,
    - eksperymenty z różnymi funkcjami aktywacji,
    - zmiana algorytmów optymalizacji.
4. Downsampling:
    - wybór metody redukcji liczby próbek dominującej klasy,
    - określenie zmienianych danych,
    - zmniejszenie liczby przykładów klasy negatywnej,
    - zwiększenie liczby iteracji treningowych.
5. Walidacja:
    - wydzielenie zbioru walidacyjnego,
    - obserwacja pozytywnych rezultatów zarówno dla wersji modelu wykorzystującej walidację, jak i w podejściu z downsamplingiem.
6. Walidacja + Downsampling:
    - integracja obu metod w celu poprawy jakości modelu,
    - uzyskanie wyraźnej poprawy czułości, co potwierdza, że wykrywanie przypadków choroby nie było efektem losowości.