<https://www.kaggle.com/datasets/aadarshvelu/aids-virus-infection-prediction/data>

# Wstęp i opis problemu

HIV (Human Immunodeficiency Virus) jest wirusem, który atakuje układ odpornościowy człowieka, a dokładniej komórki CD4, kluczowe dla prawidłowego funkcjonowania układu immunologicznego. Bez odpowiedniego leczenia HIV może prowadzić do AIDS (Acquired Immunodeficiency Syndrome). AIDS to zaawansowane stadium zakażenia HIV, charakteryzujące się znacznym osłabieniem układu odpornościowego, co zwiększa podatność na zakażenia oportunistyczne i niektóre nowotwory.

Obecnie leczenie HIV/AIDS opiera się na terapii antyretrowirusowej (ART), która polega na stosowaniu kombinacji leków antyretrowirusowych. ART nie eliminuje HIV, ale skutecznie kontroluje replikację wirusa, pozwalając pacjentom żyć dłużej i zdrowiej. Terapia ta zmniejsza ilość wirusa we krwi do niewykrywalnego poziomu, co również obniża ryzyko przeniesienia wirusa na inne osoby.

Zydowudyna, zatwierdzona w 1987 roku jako pierwszy lek antyretrowirusowy, należy do klasy nukleozydowych inhibitorów odwrotnej transkryptazy (NRTI). Działa poprzez hamowanie enzymu odwrotnej transkryptazy, co blokuje replikację wirusa. Jej wprowadzenie było przełomem w leczeniu HIV/AIDS, dając nadzieję na poprawę jakości życia pacjentów i zmniejszenie śmiertelności związanej z chorobą.

Odpowiedź na terapię antyretrowirusową może się różnić w zależności od pacjenta, co wynika z różnych czynników, takich jak stadium choroby, czynniki biologiczne czy historia leczenia. Wprowadzenie nowych leków antyretrowirusowych może prowadzić do lepszej skuteczności leczenia niż zydowudyna. Aby zbadać te zagadnienia, przeprowadzono badanie ACTG 175.

Celem niniejszej pracy jest analiza danych z badania ACTG 175, aby zidentyfikować zmienne wpływające na progresję infekcji HIV oraz ocenić efektywność różnych terapii stosowanych w leczeniu pacjentów. Wykorzystane modele będą oceniać, które czynniki mają kluczowe znaczenie dla przewidywania ryzyka progresji do zaawansowanego stadium choroby.

# Opis danych

Projekt został zrealizowany na podstawie danych "AIDS Virus Infection Prediction" pobranych ze strony Kaggle.com[[1]](#footnote-1). Dane pochodzą z badania ACTG 175 przeprowadzonego przez National Institute of Allergy and Infectious Diseases w 1995 roku. Badanie miało na celu porównanie różnych form leczenia osób zakażonych HIV-1 z liczbą komórek CD4 między 200 a 500 na milimetr sześcienny. Pod obserwację brani byli pacjenci z 43 ośrodków badań klinicznych AIDS i 9 ośrodków National Hemophilia Foundation w Stanach Zjednoczonych i Portoryko, którzy między innymi spełniali następujące kryteria: wiek co najmniej 12 lat, laboratoryjnie potwierdzona infekcja HIV-1, liczba komórek CD4 między 200 a 500 na milimetr sześcienny w ciągu 30 dni przed randomizacją, brak historii choroby definiującej AIDS (poza minimalnym śluzówkowo-skórnym mięsakem Kaposiego), wskaźnik sprawności Karnofsky'ego na poziomie co najmniej 70 oraz akceptowalne wyniki laboratoryjne.

Pacjenci zostali losowo przydzieleni do jednej z czterech grup leczenia:

* 200 mg zydowudyny trzy razy dziennie,
* 200 mg zydowudyny trzy razy dziennie plus 0,75 mg zalsytabiny trzy razy dziennie,
* 200 mg zydowudyny trzy razy dziennie plus 200 mg dydanozyny dwa razy dziennie lub
* 200 mg dydanozyny dwa razy dziennie.

Główny punkt końcowy został zdefiniowany jako spadek liczby komórek CD4 o ponad 50% w stosunku do wartości wyjściowej, co zostało zinterpretowane jako następująca progresja zespołu nabytego niedoboru odporności (AIDS). Takim punktem była również śmierć lub wystąpienie ciężkich zdarzeń niepożądanych (zgodnie ze schematem oceny ACTG), które obejmowały ciężkie objawy lub nieprawidłowości laboratoryjne.

W ramach badania zebrano też dane dotyczące opieki zdrowotnej i kategoryczne informacje na temat pacjentów, dostarczając w ten sposób podstawowych informacji na ich temat. Zbiór zawiera obserwacji i 23 zmienne, których opis przedstawiono w tabeli 2.1, znajdującej się na następnej stronie.

|  |  |  |  |
| --- | --- | --- | --- |
| NAZWA ZMIENNEJ | RODZAJ ZMIENNEJ | OPIS ZMIENNEJ | PRZYJMOWANE WARTOŚCI |
| infected | Dychotomiczna | Progresja zespołu nabytego niedoboru odporności (AIDS) | * 0 – nie rozwinięcie się infekcji do zaawansowanego stadium * 1 – osiągnięcie zaawansowanego stadium infekcji |
| drugs | Dychotomiczna | Informacja o tym czy pacjent w przeszłości zażywał dożylnie narkotyki | * 0 – nie * 1 - tak |
| gender | Dychotomiczna | Płeć pacjenta | * 0 - Kobieta * 1 - Mężczyzna |
| hemo | Dychotomiczna | Informacja o tym czy pacjent choruje na hemofilię | * 0 - nie * 1 - tak |
| homo | Dychotomiczna | Informacja o tym czy pacjent uprawia aktywność homoseksualną | * 0 - nie * 1 - tak |
| offtrt | Dychotomiczna | Informacja o tym czy pacjent w trakcie obserwacji zakończył przyjmowanie leków przed 96 (± 5) tygodniami | * 0: nie * 1: tak |
| oprior | Dychotomiczna | Informacja o tym czy pacjent przed zakwalifikowaniem do badania był poddawany innej terapii przeciwretrowirusowej niż ZDV | * 0 - nie był poddawany innej terapii przeciwretrowirusowej * 1 - był wcześniej poddawany innej terapii przeciwretrowirusowej |
| race | Dychotomiczna | Informacja o tym czy pacjent był innej rasy niż biała | * 0 - rasa biała * 1 - inna rasa niż biała |
| str2 | Dychotomiczna | Informacja o tym czy pacjent przed obserwacją był poddany leczeniu antyretrowirusowemu | * 0 - nie * 1 - tak |
| symptom | Dychotomiczna | Informacja o tym czy u pacjenta wystąpiły objawy choroby | * 0 - nie * 1 - tak |
| Treat | Dychotomiczna | Informacja o tym czy w czasie leczenia pacjent przyjmował inne leki niż zydowudynę | * 0: pacjent przyjmował tylko zydowudynę * 1: pacjent przyjmował też / lub inne leki |
| z30 | Dychotomiczna | Informacja o tym czy pacjent 30 dni przed rozpoczęciem obserwacji przyjmował zydowudynę | * 0: nie przyjmował * 1: przyjmował |
| strat | Nominalna porządkowa | Pogrupowanie ze względu na historię antyretrowirusową - podziału pacjentów na trzy grupy na podstawie ich wcześniejszej historii leczenia antyretrowirusowego | * 1 – pacjent przed przystąpieniem do badania nie miał wcześniejszego doświadczenia z terapią antyretrowirusową, * 2 – pacjent przed przystąpieniem do badania przebywał na terapii antyretrowirusowej, trwającej więcej niż 1 tydzień, ale nie dłużej niż 52 tygodnie, * 3 – pacjent przed przystąpieniem do badania miał przebywał na terapii antyretrowirusowej, trwającej ponad 52 tygodnie |
| trt | Nominalna | Sposób leczenia pacjenta w czasie badania:  - 200 mg zydowudyny trzy razy na dobę,  - 200 mg zydowudyny trzy razy na dobę plus 0,75 mg zalcytabiny trzy razy na dobę,  - 200 mg zydowudyny trzy razy na dobę plus 200 mg dydanozyny dwa razy na dobę  - 200 mg dydanozyny dwa razy na dobę | * 0 - tylko ZDV * 1 - ZDV + ddI * 2 - ZDV + Zal * 3 - tylko ddl |
| age | Przedziałowa | Wiek pacjenta (warunek ukończenia 12 lat) w momencie rozpoczęcia obserwacji | <12, 70> |
| cd40 | Przedziałowa | Poziom CD4 na początku badania | <0, 1199> |
| cd420 | Przedziałowa | CD4 po 20+/-5 tygodniach | <49, 1119> |
| cd80 | Przedziałowa | Poziom CD8 na początku badania | <40, 5011> |
| cd820 | Przedziałowa | CD8 po 20+/-5 tygodniach | <124, 6035> |
| karnof | Przedziałowa | Wynik Karnofsky'ego | <70, 100> |
| preanti | Przedziałowa | Czas od zarażenia wirusem HIV do momentu rozpoczęcia obserwacji w ramach badania (w dniach) | <0, 2851> |
| time | Przedziałowa | Czas trwania obserwacji w ramach badania (w dniach) | <14, 1231> |
| wtkg | Przedziałowa | Waga pacjenta w momencie rozpoczecia obserwacji | <31, 159,94> |

Do budowy modeli sieci neuronowych zastosowano algorytm Sequential z pakietu Keras. Sieci neuronowe opierają się na strukturze wielowarstwowego perceptronu (MLP), która polega na przepływie informacji przez kolejne warstwy neuronów w celu nauczenia się złożonych wzorców w danych. Podstawowym elementem sieci neuronowej jest neuron, który przekształca sygnał wejściowy za pomocą funkcji aktywacji, a następnie przesyła sygnał do kolejnych warstw.

W celu przewidywania wystąpienia AIDS u pacjentów z HIV zastosowano logistyczną funkcję straty. Utworzono trzy modele sieci neuronowych z wykorzystaniem różnych technik i parametrów, oznaczonych kolejnymi liczbami od 1 do 3.

W każdej z trzech sieci neuronowych wykorzystano różne parametry i hiperparametry, które wpływają na proces uczenia się modelu. Poniżej znajduje się lista wszystkich używanych parametrów i hiperparametrów wraz z ich opisami.

# Porównianie sieci i analiza wyników

Wszystkie wytrenowane sieci uzyskały bardzo dobre wyniki i nadają się do aplikacji w celu uzyskania predykcji. Analizując pole pod krzywą ROC, możemy zauważyć że sieć 1 oraz sieć 2 uzyskały zbliżone wyniki (następująco 0.94, 0.92), sieć 3 zwraca trochę gorsze wyniki, które mimo wszystko są bardzo dobre. Większe różnice występują natomiast na macierzy pomyłek. Ze względu na charakter badania chcemy zminimalizować błąd, kiedy wynik jest pozytywny, a predykcja wskazuje wynik negatywny. Ze względu na znaczenie predykcji dla ludzkiego życia, kluczowe jest minimalizowanie tego rodzaju błędów.

A graph of a function

Description automatically generated with medium confidence

Porównanie macierzy pomyłek dla poszczególnych sieci

Sieć 1

A blue squares with white text

Description automatically generatedA graph of a line

Description automatically generated with medium confidence

Sieć 2

A blue squares with white text

Description automatically generatedA graph of a function

Description automatically generated with medium confidenceSieć 3

A diagram of a confusion matrix

Description automatically generatedA graph of a function

Description automatically generated with medium confidence

1. [↑](#footnote-ref-1)