# Analiza zbioru danych "Patient-level dataset to study the effect of COVID-19 in people with Multiple Sclerosis" z wykorzystaniem nadzorowanego algorytmu uczenia maszynowego One-Class SVM(Support Vector Machines)

Anna Mrozek, Bartosz Panek i Przemysław Jura

ARTICLE INFO

*Keywords*:
One-Class SVM
COVID-19

STRESZCZENIE

Artykuł analizuje zbiór danych pacjentów ze stwardnieniem rozsianym (SM), którzy przeszli COVID-19, przy użyciu nadzorowanego algorytmu One-Class SVM. Celem jest identyfikacja przypadków o zwiększonym ryzyku ciężkiego przebiegu infekcji. Umożliwi to zidentyfikowanie charakterystycznych cech pacjentów bardziej narażonych na powikłania.

## 1. Wprowadzenie

Pandemia COVID-19 miała znaczący wpływ na osoby z chorobami przewlekłymi, w tym na pacjentów ze stwardnieniem rozsianym (SM). SM jest przewlekłą chorobą autoimmunologiczną, która wpływa na układ nerwowy i może prowadzić do trwałego uszkodzenia neuronów oraz ograniczenia funkcji motorycznych i poznawczych. Ze względu na charakter choroby i stosowane terapie immunosupresyjne, pacjenci z SM mogą być bardziej narażeni na cięższy przebieg infekcji wirusowych, w tym COVID-19. Analiza danych od pacjentów z SM, którzy przeszli COVID-19, może dostarczyć ważnych informacji na temat ryzyka powikłań i zidentyfikować czynniki przyczyniające się do poważniejszych objawów, co ostatecznie może pomóc w lepszym monitorowaniu i opiece nad tą grupą pacjentów.

## 2. Cel

Celem badania jest wykorzystanie nadzorowanego algorytmu One-Class SVM do analizy zbioru danych pacjentów ze stwardnieniem rozsianym, którzy przeszli COVID-19, w celu zidentyfikowania cech pacjentów o zwiększonym ryzyku ciężkiej infekcji. Oczekuje się, że wyniki analizy dostarczą wskazówek do opracowania bardziej ukierunkowanych strategii opieki i środków zapobiegawczych dla pacjentów ze stwardnieniem rozsianym w kontekście potencjalnych przyszłych pandemii lub innych zagrożeń wirusowych.

## 3. Przegląd literatry

## 4. Metodologia

W badaniu zastosowano algorytm nadzorowanego uczenia maszynowego One-Class SVM w celu identyfikacji pacjentów z podwyższonym ryzykiem ciężkiego przebiegu COVID-19 w grupie osób ze stwardnieniem rozsianym. Analiza obejmowała cechy kliniczne pacjentów, takie jak wiek, płeć, typ leczenia oraz choroby współistniejące, aby zidentyfikować wzorce związane z większą podatnością na powikłania.

### 4.1. Dataset

Stwardnienie rozsiane (MS) to przewlekła choroba autoimmunologiczna, która wywołuje stan zapalny w obrębie ośrodkowego układu nerwowego. Choroba prowadzi do różnych stopni utraty funkcji przez uszkodzenia mieliny oraz włókien nerwowych [1]. Osoby z MS są bardziej podatne na infekcje z powodu złożonego działania samej choroby, jej leczenia i naturalnego przebiegu [2]. Z inicjatywy COVID-19 and MS Global Data Sharing Initiative (GDSI) zbadano, jak leki immunosupresyjne lub immunomodulujące wpływają na COVID-19 i jego przebieg u osób z MS. GDSI miała na celu zwiększenie skali zbierania danych dotyczących COVID-19 i dostarczenie społeczności związanej z MS informacji opartych na danych podczas pandemii [3]. W ramach GDSI wybrano kluczowe zmienne obejmujące informacje o COVID-19, stopniu jego ciężkości, leczeniu, dane demograficzne, historię i nasilenie MS, stosowanie leków modyfikujących przebieg choroby (DMT), choroby współistniejące i wybrane zachowania związane ze stylem życia, takie jak palenie tytoniu. Globalna społeczność MS współpracowała, przekazując dokumentację statusu COVID-19 u osób z MS za pośrednictwem centralnej platformy udostępnionej przez QMENTA [4].

Ten zbiór danych został zebrany za pomocą narzędzia do szybkiego wprowadzania danych, które umożliwiało klinicystom, osobom ze stwardnieniem rozsianym (PwMS) lub ich przedstawicielom bezpośrednie wprowadzanie informacji do centralnej platformy GDSI. Narzędzie to zawierało kwestionariusz oparty na wcześniej ustalonych zmiennych i nie gromadziło bezpośrednich danych osobowych, aby chronić prywatność użytkowników. Narzędzie zostało wyłączone 3 lutego 2022 roku.

Zbiór danych obejmuje informacje o 1141 osobach ze stwardnieniem rozsianym (PwMS). Aby zapewnić zgodność danych z wytycznymi HIPAA, przeprowadzono proces deidentyfikacji. Po zebraniu danych dokonano oceny ryzyka

związanego z małymi komórkami (SCRA), klasyfikując zmienne na trzy kategorie: bezpośrednie identyfikatory, zmienne wrażliwe i identyfikatory pośrednie. Bezpośrednie identyfikatory to zmienne, które mogą jednoznacznie zidentyfikować osobę, zmienne wrażliwe to te, które respondent może chcieć zachować w tajemnicy, natomiast identyfikatory pośrednie mogą zidentyfikować osobę, jeśli są połączone z danymi z innych zbiorów.

Ponieważ w danych nie zbierano imion pacjentów, deidentyfikacja skupiła się na datach i wieku pacjentów. Wiek pacjentów sklasyfikowano w cztery grupy: 0 dla osób w wieku 0–17 lat, 1 dla osób między 18 a 50 lat, 2 dla osób między 51 a 70 lat, oraz 3 dla osób powyżej 71 lat. Dzięki temu żadne dokładne wartości wieku powyżej 90 lat nie zostały ujawnione. Po klasyfikacji zmiennych i wdrożeniu odpowiednich środków ostrożności dane zostały zdeidentyfikowane i spełniają standardy HIPAA, zachowując jednocześnie wartość badawczą. Ponadto, aby zapewnić ochronę prywatności, zastosowano techniki takie jak K-anonimizacja oraz różnorodność.

Zbiór danych obejmuje zestaw z góry określonych zmiennych (n=47), takich jak płeć, kategoria wiekowa, typ MS, wynik EDSS, status palenia oraz kategoria BMI. Te zmienne dostarczają informacji o demografii pacjentów, ich stanie klinicznym oraz symptomach związanych z COVID-19. Szczegółowy opis typów zmiennych i ich statystyki znajduje się w sekcji „Opis Danych".

Zbiór zawiera dane zebrane od 1141 osób z stwardnieniem rozsianym (PwMS) w ramach inicjatywy COVID-19 i MS Global Data Sharing Initiative. W tym zbiorze danych znajdują się różnorodne zmienne kategoryczne i numeryczne, takie jak płeć, typ SM, status palenia, kategoria wiekowa, wynik EDSS, objawy COVID-19 oraz kategoria BMI. Dane te zostały zebrane w celu lepszego zrozumienia wpływu COVID-19 na osoby z SM oraz wsparcia badań i analiz dotyczących tej choroby. Dla ochrony prywatności pacjentów, dane zostały dodatkowo zanonimizowane przy użyciu technik K-anonimowości oraz l-różnorodności. Obecnie zbiór ten jest dostępny dla społeczności naukowej do celów badawczych.[5]

### 4.2. Opis metody
### 4.3. Opis przeprowadzonych obliczeń
### 4.4. Wykorzystane metryki oceny
## 5. Wyniki
## 6. Opis wyników
## 7. Podsumowanie
## 8. Bibliografia

[1] Calabresi PA. Diagnosis and management of multiple sclerosis. Am Fam Physician. 2004 Nov

[2] Montgomery S, Hillert J, Bahmanyar S. Hospital admission due to infections in multiple sclerosis patients. Eur J Neurol. 2013 Aug

[3] Peeters LM, Parciak T, Walton C, Geys L, Moreau Y, De Brouwer E, et al. COVID-19 in people with multiple sclerosis: A global data sharing initiative. Mult Scler Houndmills Basingstoke Engl. 2020 Sep

[4] Simpson-Yap S, De Brouwer E, Kalincik T, Rijke N, Hillert JA, Walton C, et al. Associations of Disease-Modifying Therapies With COVID-19 Severity in Multiple Sclerosis. Neurology. 2021 Nov

[5] https://physionet.org/content/patient-level-data-covid-ms/1.0.1/ Published: Jan. 2, 2024. Version: 1.0.1

https://github.com/przemyslawJura00/RozpoznawanieWzorcowGrup

https://physionet.org/content/patient-level-data-covid-ms/1.0.1/

https://www.geeksforgeeks.org/understanding-one-class-support-vector-machines/