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 oraz poznawczych.[1][2] 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.[3][4] 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

W literaturze medycznej i naukowej, od początku pandemii COVID-19, wiele badań skupiało się na analizie wpływu wirusa SARS-CoV-2 na osoby cierpiące na choroby przewlekłe i autoimmunologiczne, takie jak stwardnienie rozsiane (SM). Wyniki tych badań sugerują, że pacjenci z SM, zwłaszcza ci stosujący terapie immunosupresyjne, mogą być bardziej narażeni na cięższy przebieg COVID-19 oraz związane z nim powikłania[5][6]. U pacjentów z SM często obserwuje się obniżoną odporność oraz większą podatność na infekcje, co wynika zarówno z samej choroby, jak i z efektów leków tłumiących układ odpornościowy[7]. Dodatkowo w badaniach zwraca się uwagę na takie czynniki ryzyka jak wiek, płeć, rodzaj stosowanego leczenia oraz inne choroby towarzyszące, które mogą zwiększać ryzyko ciężkiego przebiegu infekcji u tej grupy pacjentów.

W ostatnich latach coraz większą popularność zyskuje zastosowanie algorytmów uczenia maszynowego, takich jak Support Vector Machines (SVM), do analizy danych medycznych. Algorytmy te pozwalają na wykrywanie wzorców i anomalii w dużych, zróżnicowanych zbiorach danych[8]. W szczególności algorytm One-Class SVM, używany do wykrywania anomalii, jest przydatny do identyfikacji przypadków wysokiego ryzyka w danych medycznych, gdzie przypadki nietypowe (np. pacjenci bardziej narażeni na ciężki przebieg COVID-19) występują rzadko. Badania pokazują, że One-Class SVM dobrze sprawdza się w analizie populacji pacjentów, gdy dostęp do dużych zbiorów danych o przypadkach zdrowych jest ograniczony, co często stanowi wyzwanie w analizie medycznej[9].

W kontekście badań nad COVID-19 i SM, algorytmy nadzorowanego uczenia maszynowego, takie jak SVM, umożliwiają dokładniejszą analizę czynników ryzyka oraz lepsze prognozowanie ciężkiego przebiegu choroby. Wyniki takich analiz mogą być przydatne nie tylko dla lekarzy, ale także dla decydentów zajmujących się zdrowiem publicznym, ponieważ pozwalają na opracowanie bardziej ukierunkowanych strategii opieki dla osób z SM, które są bardziej narażone na zagrożenia związane z wirusami, takimi jak COVID-19.

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 [10]. Osoby z MS są bardziej podatne na infekcje z powodu złożonego działania samej choroby, jej leczenia i naturalnego przebiegu [11]. 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 [12]. 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 [13].

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. Daty w kolumnie „stop-or-end-date-combined" zostały przesunięte o losową liczbę dni (między -15 a 15), aby uniemożliwić identyfikację na podstawie dat. 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".

4.2. Opis metody

Jednoklasowy SVM (One-Class Support Vector Machine, OCSVM) to algorytm przeznaczony do wykrywania anomalii w zbiorze danych. Zamiast klasyfikować dane do dwóch lub więcej klas, jak w klasycznym SVM, OCSVM koncentruje się wyłącznie na danych normalnych.[14][15] Jego celem jest zbudowanie granicy wokół większości przypadków normalnych, tworząc "strefę normalności", która odróżnia standardowe przypadki od anomalii.

Główna zasada działania OCSVM opiera się na maksymalizacji marginesu między przypadkami normalnymi a granicą, która oddziela normę od anomalii. Dzięki temu model lepiej rozpoznaje dane odstające, które znajdują się poza wyznaczoną strefą. W algorytmie znajduje się hiperparametr „nu", który pozwala kontrolować czułość modelu – decyduje on o maksymalnym odsetku błędów marginesowych oraz liczbie wektorów nośnych, wpływając na balans między surowością a tolerancją modelu.[16][17]

Kluczowym elementem działania OCSVM jest funkcja decyzyjna oparta na hiperpłaszczyźnie, która oddziela dane od początku układu współrzędnych. Wzór na hiperpłaszczyznę wyrażany jest jako:

$$f(x) = \mathbf{w} \cdot x - \rho \tag{1}$$

gdzie:

- w – wektor normalny do hiperpłaszczyzny, wyznaczony podczas procesu uczenia,
- x – próbka danych,
- ρ – jest wartością progową (bias), która definiuje margines.

OCSVM posiada szereg parametrów, dzięki którym możliwe jest dostosowywanie modelu do specyficznych potrzeb:

- **Parametr** ν: Kontroluje liczbę obserwacji uznawanych za anomalie. Parametr ten przyjmuje wartości w przedziale od 0 do 1 i determinuje udział anomalii, jakie model może tolerować w zbiorze treningowym (np. jeśli $\nu = 0.05$, oznacza to, że około 5% próbek zostanie sklasyfikowanych jako anomalie).

- **Parametr** γ: Wpływa na kształt granicy decyzyjnej poprzez dopasowywanie modelu do danych. Mniejsza wartość γ oznacza „szerszą" granicę, obejmującą więcej punktów, natomiast większa wartość γ powoduje, że granica jest bardziej dopasowana do danych, co może zwiększać ryzyko nadmiernego dopasowania (overfitting).

- **Jądro (*ang. kernel*)**: One-Class SVM często wykorzystuje funkcje jądrowe, aby modelować nieliniowe granice decyzyjne i lepiej dopasować się do danych, które nie są liniowo rozdzielne w przestrzeni cech. Wybór jądra, zwany również „sztuczką jądra", wpływa na charakter dopasowania modelu i jego wydajność.

Podczas treningu OCSVM analizuje wyłącznie dane normalne, co sprawia, że jest szczególnie użyteczny w sytuacjach, gdy anomalie są rzadkie lub trudne do zidentyfikowania. Model staje się wtedy bardziej niezawodny w rzeczywistych zastosowaniach, takich jak wykrywanie oszustw, monitorowanie awarii czy zabezpieczanie sieci komputerowych. OCSVM może wykorzystać różne funkcje jądra, co umożliwia mu wykrywanie zarówno prostych, jak i bardziej złożonych odchyleń.[18]

4.3. Opis przeprowadzonych obliczeń

1. Przygotowanie i czyszczenie danych - rozpoczęliśmy od wczytania danych oraz wybrania interesujących nas cech, które następnie zostały przekształcone w sposób umożliwiający przeprowadzenie obliczeń:

- Dla kolumn binarnych (yes/no) wartości zostały zamienione na liczby 0 i 1.

- Dla kolumn kategorycznych i ordinalnych (np. age-in-cat, covid19-outcome-levels-2, report-source) zostały przypisane wartości liczbowe, zgodnie z przyjętym mapowaniem.

- Brakujące wartości w danych zostały wypełnione zerami, co pozwoliło na uniknięcie problemów podczas analizy i treningu modelu.

2. Tworzenie nowych zmiennych: symptom-score oraz comorbidity-score, aby skonsolidować informacje o objawach i chorobach współistniejących.

- Stworzyliśmy dwie kolumny symptom-score oraz comorbidity-score, które zliczają ilość objawów wirusa COVID-19 oraz liczby chorób współistniejących dla każdego pacjenta.

- Kolumny te były następnie skalowane przy użyciu StandardScaler, co pozwala na lepszą interpretację wyników oraz standaryzację w zakresie modelowania.

Efektem było, uzyskanie wartości numerycznych, które reprezentują intensywność symptomów oraz liczbę chorób współistniejących dla każdego pacjenta.

3. Przygotowanie danych do treningu - wykorzystaliśmy pacjentów bez chorób współistniejących jako grupę uczącą (zbiór X-train). Wszystkie próbki (zarówno te z chorobami współistniejącymi, jak i bez) tworzyły zestaw testowy (zbiór X-test). Wybór pacjentów bez chorób współistniejących pozwolił modelowi nauczyć się cech charakterystycznych dla „normalnych" próbek, które posłużyły do wykrywania potencjalnych anomalii.

4. Trening modelu One-Class SVM - zastosowaliśmy algorytm One-Class Support Vector Machine (SVM) z jądrem rbf, który dobrze radzi sobie z modelowaniem nieliniowych granic decyzyjnych. Model trenowano, aby rozpoznawał typowe cechy pacjentów bez chorób współistniejących, a następnie przewidywał, które próbki są podobne do tego wzorca, a które mogą być anomaliami. Po treningu model przypisywał każdej próbce etykietę 1 (normalna) lub -1 (anomalia). Próbki oznaczone jako anomalia zawarto w nowej ramce anomalies w celu dalszej analizy.

5. Analiza wyników - na podstawie wykrytych anomalii przeprowadziliśmy szereg analiz, tworząc wizualizacje, które pomagają zrozumieć charakterystykę anomalii[19]

4.4. Wykorzystane metryki oceny

Classification Report: Zawiera dokładność (precision), czułość (recall) i f1-score dla obu klas (1 - normalna, -1 - anomalia). Daje to pogląd na efektywność wykrywania anomalii oraz odsetek poprawnie klasyfikowanych normalnych próbek.

Confusion Matrix (Macierz Pomyłek): Macierz błędów wyświetla liczby klasyfikacji poprawnych i błędnych dla klas normalnych i anomalii, co pozwala zidentyfikować potencjalne problemy z fałszywie pozytywnymi lub fałszywie negatywnymi klasyfikacjami.

ROC AUC (Area Under the Curve of Receiver Operating Characteristic) — pomimo braku klasy binarnej, można skonstruować ROC, badając, jak dobrze model oddziela „normę" od „odchyleń".

Procent wykrytych anomalii — określa, jaki odsetek faktycznych anomalii został prawidłowo wykryty.

False Positive Rate (FPR) — aby zminimalizować fałszywe alarmy.

5. Wyniki

5.1. Ocena modelu One-Class SVM

Precision: Jest to miara, która określa, jaki odsetek przypadków oznaczonych przez model jako „Anomaly" rzeczywiście jest anomaliami. W tym przypadku precyzja dla obu klas (Normal i Anomaly) wynosi 1.00 (100%), co oznacza, że wszystkie przypadki zaklasyfikowane jako „Normal" i „Anomaly" przez model są poprawne.

Recall: Jest to miara, która określa, jaki odsetek rzeczywistych przypadków „Anomaly" został poprawnie wykryty przez model. Wartość 1.00 dla obu klas oznacza, że model poprawnie rozpoznał 100% przypadków normalnych i anomalnych.

```
Ocena modelu One-Class SVM:
              precision    recall  f1-score   support

      Normal       1.00      1.00      1.00        37
     Anomaly       1.00      1.00      1.00        47

    accuracy                           1.00        84
   macro avg       1.00      1.00      1.00        84
weighted avg       1.00      1.00      1.00        84

Macierz pomyłek:
[[37  0]
 [ 0 47]]
ROC AUC: 1.0
Procent wykrytych anomalii (TPR): 100.00%
False Positive Rate (FPR): 0.00%
```

Figure 1: Ocena modelu One-class SVM

F1-Score: To średnia harmoniczna precyzji i czułości, która jest przydatna w sytuacjach, gdzie istnieje nierównowaga między klasami (np. liczba przypadków normalnych i anomalnych). Wartość F1-Score 1.00 oznacza, że model jest bardzo dobrze zbalansowany i radzi sobie perfekcyjnie w wykrywaniu obu klas.

Support: To liczba przypadków w każdej klasie (37 przypadków normalnych i 47 anomalii).

Accuracy: Całkowita dokładność modelu (accuracy) to 100%, co oznacza, że model poprawnie sklasyfikował wszystkie 84 przypadki testowe (37 normalnych i 47 anomalnych).

Macierz pomyłek: [[37, 0], [0, 47]] Macierz ta pokazuje, że wszystkie 37 przypadków normalnych zostały poprawnie sklasyfikowane jako normalne, a wszystkie 47 anomalii zostały poprawnie oznaczone jako anomalie. Brak błędnych klasyfikacji (nie ma wartości poza główną przekątną), co wskazuje na perfekcyjne działanie modelu.

ROC AUC (Area Under the Curve): wynosi 1.0, co oznacza idealną zdolność modelu do rozróżniania między klasami „Normal" i „Anomaly". Wartość 1.0 w AUC oznacza, że model jest perfekcyjny, a jego wyniki są w 100% poprawne.

Procent wykrytych anomalii (True Positive Rate, TPR): Procent wykrytych anomalii wynosi 100%. TPR, czyli odsetek rzeczywistych anomalii, które zostały poprawnie oznaczone jako anomalie, wynosi 100%. Model wykrył wszystkie anomalia bez pominięcia żadnego przypadku. False Positive Rate (FPR): FPR wynosi 0%, co oznacza, że model nie popełnił żadnego błędu, oznaczając przypadek normalny jako anomalię. Wartość FPR 0% oznacza, że model idealnie oddzielił przypadki normalne od anomalii bez żadnych fałszywych alarmów.

5.2. Wynik analizy

Wykresy znajdują się na końcu opracowania.

6. Opis wyników

Opis wynkresów znajdujących się na końcu opracowania:

Figure 2: Na wykresie widzimy histogram z liczbą próbek oznaczonych jako normalne i anomalia przez model One-Class SVM. Z wykresu wynika, że większość próbek została sklasyfikowana jako normalne (oznaczone jako 1), podczas gdy mniejsza liczba próbek została uznana za anomalie. Wysoka liczba normalnych próbek w stosunku do anomalii wskazuje, że model wykrywa anomalia rzadziej, co jest zgodne z celem detekcji anomalii, ponieważ anomalie w zdrowiu są zwykle mniej częste.

Figure 3: Histogram pokazuje, że większość przypadków anomalii ma niski symptom_score, głównie na poziomie 0, z kilkoma przypadkami rozproszonymi na wyższych wartościach aż do 7. Niskie wartości mogą sugerować, że większość anomalii wykazuje niewiele lub żadnych symptomów, choć istnieją wyjątki, gdzie symptom score jest wyższy.

Figure 4: Histogram przedstawia, że większość anomalii ma comorbidity_score równy zero, co oznacza brak chorób współistniejących, chociaż kilka przypadków posiada wyższe wartości aż do 5. To sugeruje, że choć wiele przypadków anomalii nie ma dodatkowych schorzeń, niektóre z nich charakteryzują się złożonymi chorobami współistniejącymi.

Figure 5: Histogram pokazuje bardziej równomierny rozkład w porównaniu do anomalii. Najwięcej przypadków ma symptom_score wynoszący około 3, ale przypadki są bardziej równomiernie rozproszone w przedziale od 0 do 6. Sugeruje to, że normalne przypadki mogą mieć różny poziom symptomów, ale nie wyróżniają się skrajnie niskimi ani wysokimi wartościami.

Figure 6: Tutaj widzimy, że większość przypadków normalnych ma comorbidity_score bliski zero, choć niektóre przypadki osiągają wartość 1. Wskazuje to, że większość przypadków normalnych nie ma chorób współistniejących, ale mogą występować łagodne współistniejące schorzenia.

Figure 7: Na wykresie przedstawiono średnie wartości cech binarnych dla przypadków zaklasyfikowanych jako anomalie. Są to cechy, które wskazują obecność lub brak pewnych objawów lub stanów zdrowotnych. Wysokie średnie wartości wskazują na częstsze występowanie danej cechy wśród anomalii: Najwyższe średnie wartości dotyczą cech takich jak current_dmt (obecnie przyjmowana terapia modyfikująca chorobę), covid19_self_isolation (izolacja w związku z COVID-19), oraz covid19_has_symptoms (obecność symptomów COVID-19). Oznacza to, że osoby klasyfikowane jako anomalie często są poddane leczeniu, mają objawy lub były w izolacji. Kolejne wysokie cechy to np. covid19_confirmed_case (potwierdzone przypadki COVID-19) oraz różne objawy związane z COVID-19, takie jak dry_cough (suchy kaszel), fever (gorączka) i fatigue (zmęczenie), co sugeruje, że anomalie mogą być powiązane z przypadkami COVID-19 lub innymi poważnymi objawami chorobowymi.

Figure 8: Na tym wykresie poziomym przedstawiono średnie wartości dla różnych cech binarnych w normalnych przypadkach (bez anomalii). Cecha o najwyższej średniej wartości to current_dmt, co wskazuje, że pacjenci normalni często używają terapii modyfikujących chorobę (DMT). Inne często występujące cechy to covid19_has_symptoms, covid19_self_isolation, oraz covid19_sympt_dry_cough, co sugeruje, że objawy COVID-19, takie jak kaszel i izolacja, są częste w tej grupie. Natomiast mniej powszechne cechy, jak comorbidities_other, covid19_icu_stay, czy covid19_sympt-_pneumonia, występują rzadko, co oznacza, że przypadki normalne rzadko dotyczą ciężkich objawów COVID-19 lub pobytu na oddziale intensywnej terapii.

Figure 9: Większość anomalii dotyczy osób młodszych (kategoria 1), co może sugerować, że młodsze osoby mają większe ryzyko występowania anomalii w danych.

Figure 10: Zdecydowana większość przypadków anomalii dotyczy osób z niższym BMI (kategoria 0), co może wskazywać na związek między BMI a nietypowymi przypadkami.

Figure 11: Wśród anomalii przeważają kobiety (kategoria 1), co sugeruje, że w danych kobiet częściej występują anomalie.

Figure 12: Większość anomalii dotyczy osób bez przypisania do ms_type2 (kategoria 0), ale niektóre przypadki należą do kategorii 1 lub 2.

Figure 13: Rozkład wskazuje, że wśród anomalii dominują łagodniejsze wyniki COVID-19 (kategoria 0), choć są też przypadki umiarkowane (kategoria 1) i ciężkie (kategoria 2).

7. Podsumowanie

Zgromadzony zbiór danych liczył jedynie około 1100 przypadków, co w kontekście analiz statystycznych i uczenia maszynowego jest liczbą stosunkowo niewielką, szczególnie przy analizie złożonych zjawisk, takich jak przebieg COVID-19 w określonych populacjach. Dodatkowo dane te charakteryzowały się dużą ilością brakujących wartości – aż około 70% przypadków zawierało znaczne luki w informacjach, co ograniczało ich użyteczność. Tylko 7% przypadków dotyczyło osób hospitalizowanych, co dodatkowo zmniejszało wartość predykcyjną modelu, ponieważ ciężkie przypadki stanowiły bardzo niewielką część danych i były słabo reprezentowane.

Problemem była również nierównomierność rozkładu danych w kluczowych grupach, takich jak płeć i wiek, co prowadziło do stronniczości wyników. Na przykład, nierównomierna liczba kobiet i mężczyzn w próbie lub różnice w liczbie obserwacji pomiędzy grupami wiekowymi powodowały, że model mógł błędnie wyciągać wnioski na podstawie nadreprezentowanych cech. Brak równowagi w danych utrudniał także identyfikację rzeczywistych wzorców w populacji i prowadził do fałszywego postrzegania wiarygodności wyników.

W tych okolicznościach zastosowanie modelu one-class SVM nie było właściwym podejściem. Model ten zakłada, że dostępne dane normalne są wystarczająco reprezentatywne, by umożliwić wykrycie anomalii. Jednakże w tym przypadku, z powodu małej liczby przypadków, braku wystarczających przykładów rzeczywistych anomalii (np. ciężkich przypadków COVID-19), a także braków danych i ich nierównomiernego rozkładu, model nie był w stanie skutecznie odróżnić anomalii od danych normalnych. W konsekwencji one-class SVM wykazał fałszywie wyniki.

Podsumowując, niedostateczna liczba przypadków, liczne braki danych, nierównomierność ich rozkładu oraz brak reprezentatywności ciężkich przypadków sprawiły, że analiza oparta na one-class SVM nie była miarodajna. Wyniki uzyskane przez model były w dużej mierze fałszywie pozytywne i nie miały praktycznego znaczenia. Aby przeprowadzić skuteczniejszą analizę, konieczne byłoby zgromadzenie większej, bardziej zrównoważonej i kompletnej próby danych, uwzględniającej większą liczbę przypadków klinicznych i ciężkich przypadków COVID-19.

8. Bibliografia

[1] https://diag.pl/pacjent/artykuly/jakie-sa-pierwsze-objawy-i-sposoby-leczenia-stwardnienia-rozsianego/

[2] https://www.medicover.pl/o-zdrowiu/stwardnienie-rozsiane-objawy-przyczyny-i-leczenie,6619,n,192

[3] https://pl.wikipedia.org/wiki/Stwardnienie_rozsiane

[4] https://ptsr.org.pl/strona/133,covid-19-a-sm

[5] Sormani MP, De Rossi N, Schiavetti I, Carmisciano L, Cordioli C, Radaelli M, et al. Disease modifying therapies and COVID-19 severity in multiple sclerosis. Ann Neurol. 2021 Apr

[6] Louapre C, Collongues N, Stankoff B, Giannesini C, Papeix C, Bensa C, et al. Clinical characteristics and outcomes in patients with coronavirus disease 2019 and multiple sclerosis. JAMA Neurol. 2020 Sep

[7] 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

[8] Schiff MA, Rae-Grant A, Gilden D, Franklin GM. Practice guideline: Disease-modifying therapies for adults with multiple sclerosis: Report of the guideline development, dissemination, and implementation subcommittee of the American Academy of Neurology. Neurology. 2019 Jan

[9] Erfani P, Mitchell AJ, Hameed S, Heydarpour P, Ghaffaripour R, Sahraian MA. Systematic review of health-related quality of life in multiple sclerosis patients: The impact of pharmacological treatments and lifestyle. J Neurol Sci. 2016 Dec

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

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

[12] 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

[13] 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

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

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

[16] https://scikit-learn.org/dev/modules/generated/sklearn.svm.OneClassSVM.html

[17] https://www.baeldung.com/cs/one-class-svm

[18] https://medium.com/@roshmitadey/anomaly-detection-using-support-vectors-2c1b842213ed

[19] https://github.com/przemyslawJura00/RozpoznawanieWzorcowGrupa4

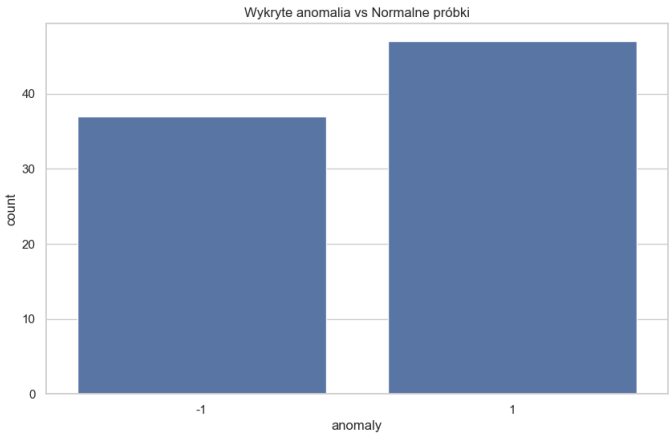

Figure 2: Wykryte anomalia vs Normalne próbki: -1: anomalia 1: normalny przypadek

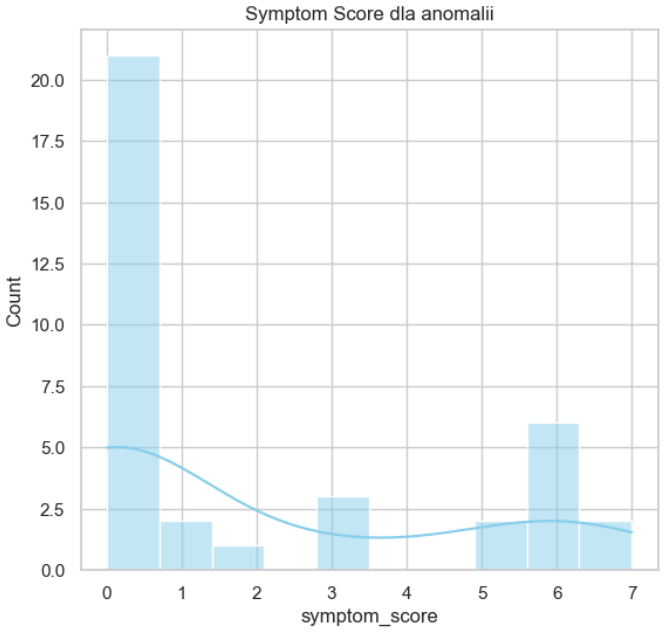

Figure 3: Rozkład ilości symptomów dla anomalii

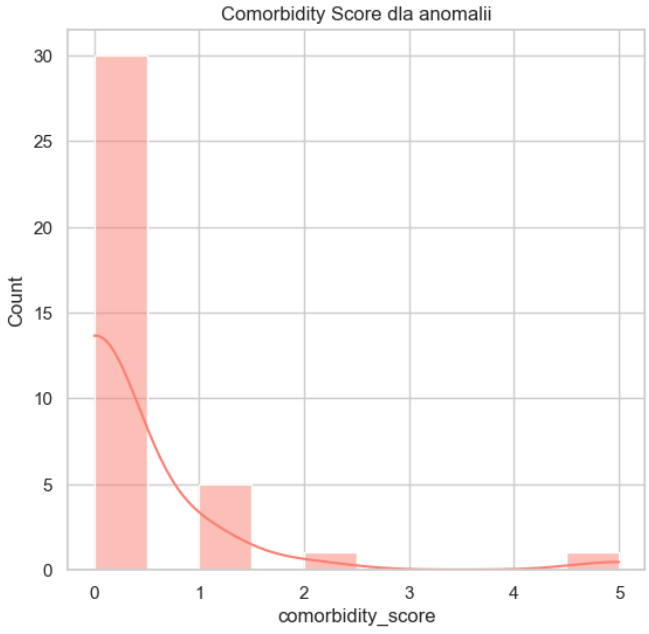

Figure 4: Rozkład ilości chorób wspóistniejących dla anomalii

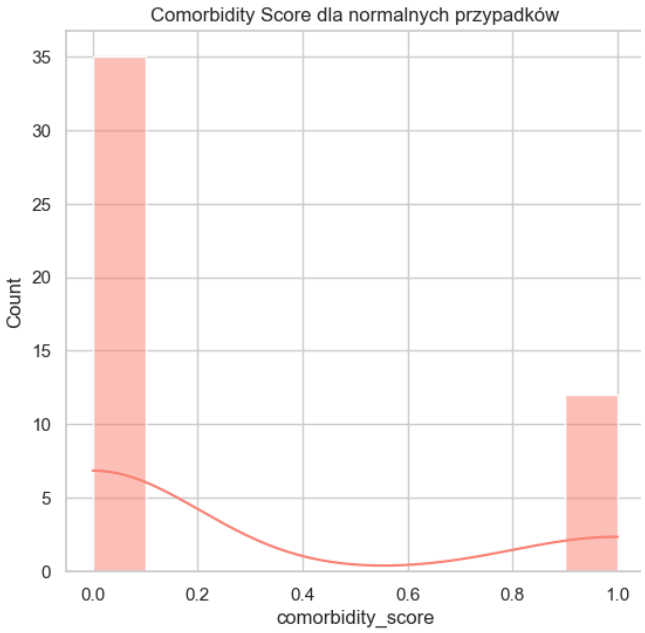

Figure 6: Rozkład ilości chorób wspóistniejących dla przypadków normalnych

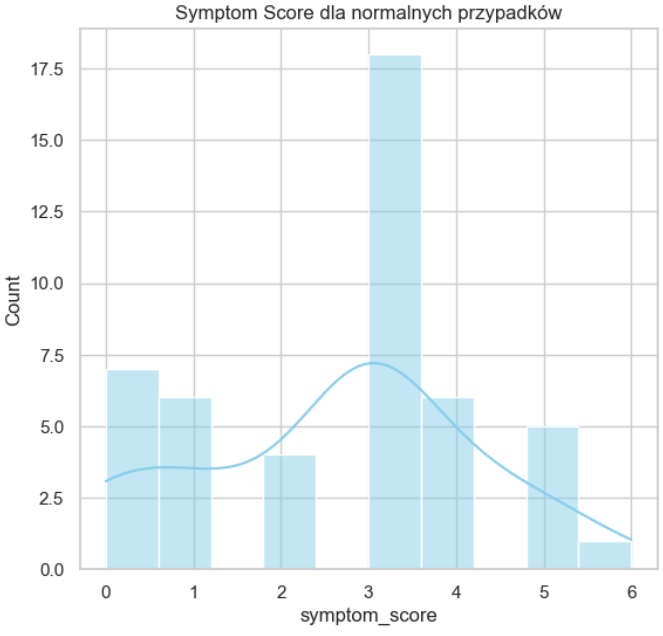

Figure 5: Rozkład ilości symptomów dla przypadków normalnych

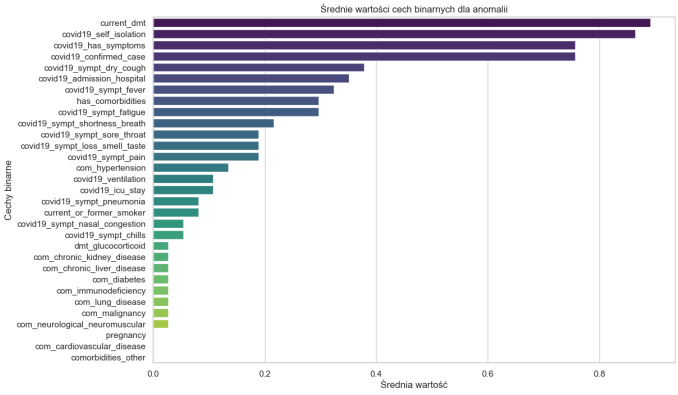

Figure 7: Średnie wartości charakterystycznych cech dla anomalii

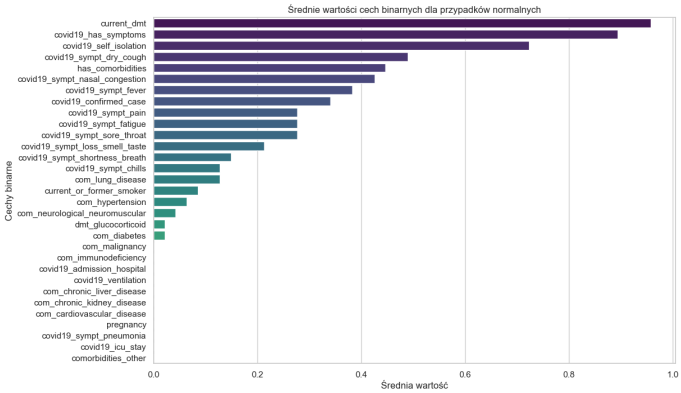

Figure 8: Średnie wartości charakterystycznych cech dla przypadków normalych

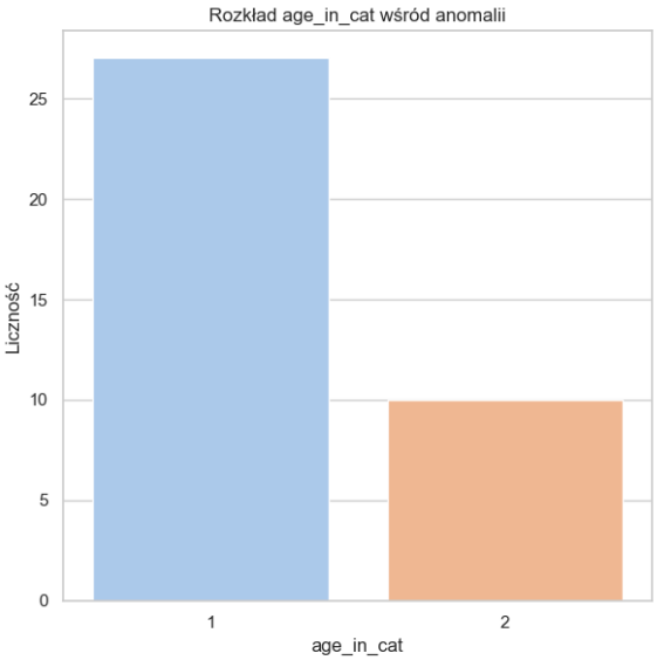

Figure 9: Rozkład wieku wśród anomalii: 0: jeśli zakres wieku mieści się w przedziale od 0 do <18. 1: jeżeli przedział wiekowy mieści się w przedziale od 18 do <=50 lat. 2: jeżeli przedział wiekowy mieści się w przedziale od 51 do <=70 lat. 3: jeśli przedział wiekowy wynosi 71 lat lub więcej..

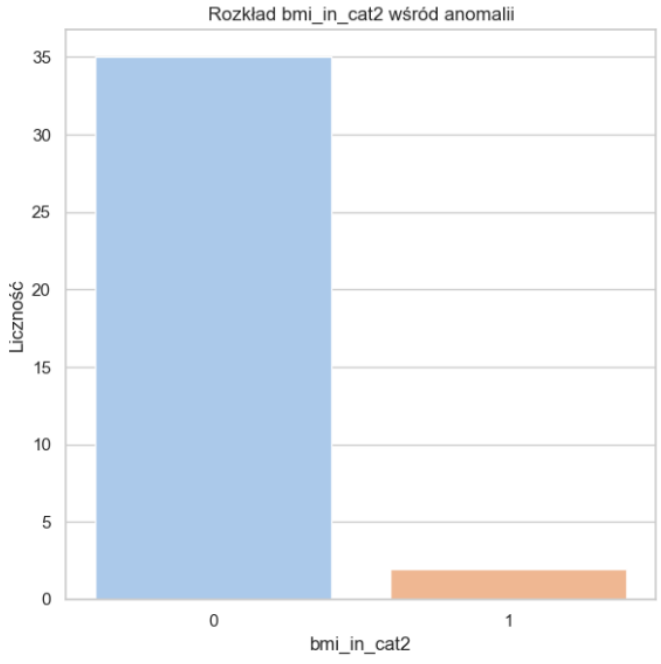

Figure 10: Rozkład bmi wśród anomalii: 0: not_overweight: if BMI <= 30 kg/m2 1: overweight: if BMI > 30 kg/m2.

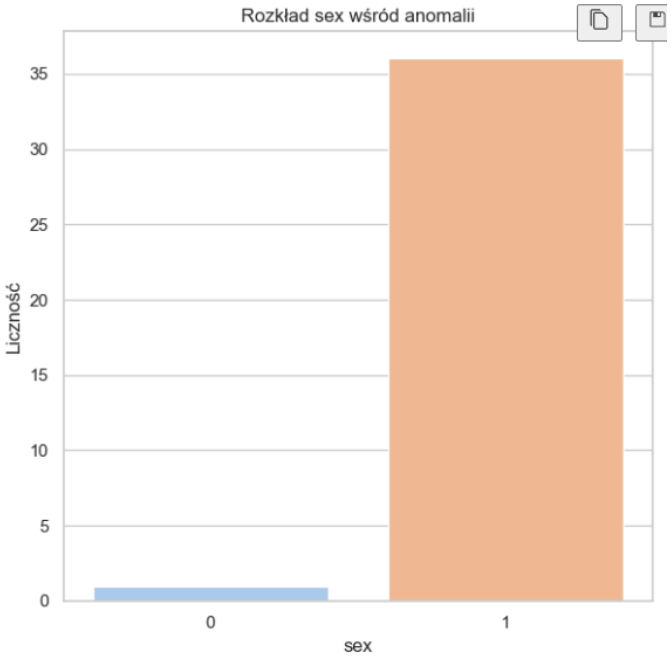

Figure 11: Rozkład płci wśród anomalii: 0: mężczyźni 1: kobiety

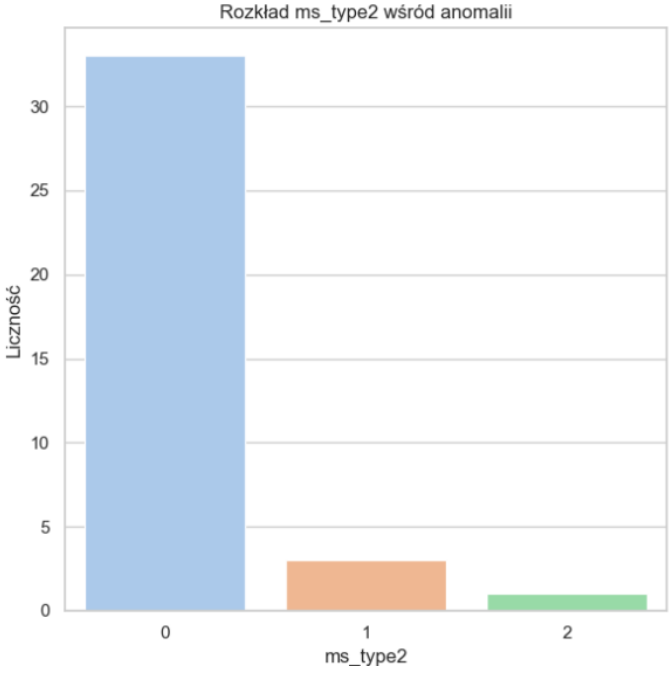

Figure 12: Rozkład typu stwardnienia rozsianego wśród anomalii: 0: relapsing_remitting: jeśli typ stwardnienia rozsianego to stwardnienie rozsiane rzutowo-remisyjne (RRMS) 1: progresywny_MS: jeśli typ stwardnienia rozsianego to wtórnie postępujące stwardnienie rozsiane (SPMS) lub pierwotnie postępujące stwardnienie rozsiane (PPMS) 2: inny: jeśli typ stwardnienia rozsianego to zespół izolowany klinicznie (CIS) lub pusty lub „niepewny", w przypadku gdy pacjent lub lekarz nie był pewien.

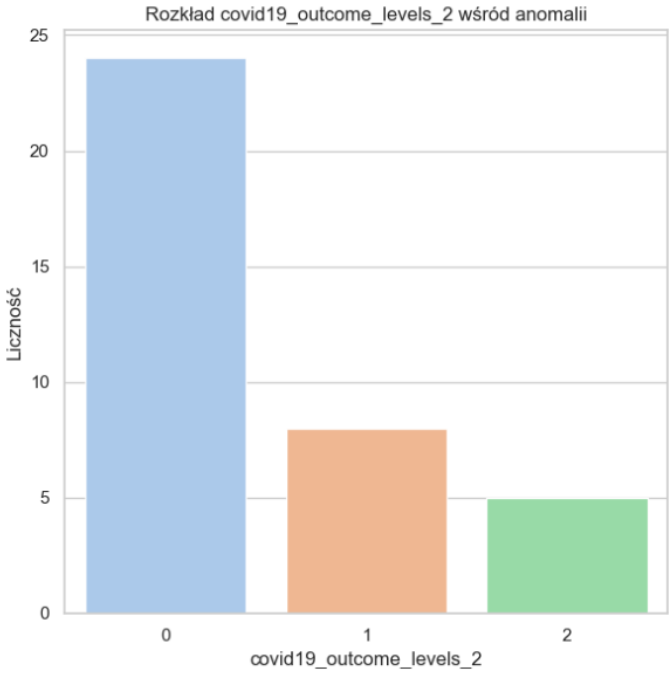

Figure 13: Rozkład hospitalizowanych przypadków: 0: Jeśli dana osoba ma Covid-19, ale nie była hospitalizowana. 1: Osoba ma Covid-19 i została hospitalizowana. 2: Osoba ma Covid-19, była hospitalizowana, przebywała na oddziale intensywnej terapii i/lub przebywała w ośrodku wentylacyjnym. 3: Osoba zmarła z powodu Covid-19 (nieobecna w tym zbiorze danych).