Sprawozdanie:

1) **Wprowadzenie:**

Celem mojego projektu jest opracowanie modelu, który klasyfikuje osoby za względu na ich stan psychiczny na podstawie ulubionego gatunku muzycznego, częstotliwości słuchania i wieku. Dokonam także analizy wpływu muzyki na zdrowie psychiczne człowieka. Skorzystałam z zestawu danych pobranego z platformy Kaggle. Jest to ankieta, w której badane są m.in. ulubione rodzaje muzyki, częstotliwość słuchania różnych gatunków muzycznych oraz samoocena stanu zdrowia psychicznego respondentów (depresja, bezsenność, lęki, zaburzenia obsesyjno-kompulsywne). Zbiór danych obejmuje około 750 osób.

Aby uzyskać większą liczbę odpowiedzi i skonfrontować się z bardziej zaawansowanym oczyszczaniem danych, postanowiłam zebrać także własne odpowiedzi. Korzystając z ankiet Google opracowałam ankietę, z takimi samymi pytaniami jak ta z Kaggle, w której zapytania zostały przetłumaczone na język polski. Zebrałam ponad 100 odpowiedzi.

W sprawozdaniu przedstawię proces opracowania modelu klasyfikacyjnego na podstawie tych danych. Opiszę kroki podejmowane w celu oczyszczenia i przygotowania danych oraz eksploracyjną analizę danych. Następnie przedstawię wybór i implementację odpowiedniego modelu uczenia maszynowego, ocenę jakość modeli oraz wysunę wnioski.

Wykonując ten projekt powinnam nauczyć się przetwarzania i analizy danych, budowy modelu oraz interpretacji wyników.

Wykorzystam kilka klasyfikatorów, a następnie ocenię ich jakość przy użyciu odpowiednich metryk. Dowiem się jakie metody wykorzystać, aby poprawić jakość przeprowadzonej klasyfikacji, i które z wypróbowanych metod wpłyną na moje wyniki.

2) **Opis przeprowadzonych badań**

Do wykonania projektu używałam języka programowania Python.

Do przygotowania danych wykorzystałam bibliotekę Pandas. Dzięki niej załadowałam dane do ramki, oczyściłam dane tzn. usunęłam zbędne kolumny, pozbyłam się pustych wartości i zakodowałam wartości tekstowe numerycznie.

Pracowałam na dwóch osobnych ramkach danych (dane z platformy Kaggle i moja ankieta), więc wartości numeryczne, które kodowałam w obu ramkach musiały być takie same. Dodatkowo dzięki Pandas dokonałam konkatenacji tych obiektów, tworząc w ten sposób jedną ramkę danych.

Do uczenia maszynowego skorzystałam z biblioteki Sklearn.

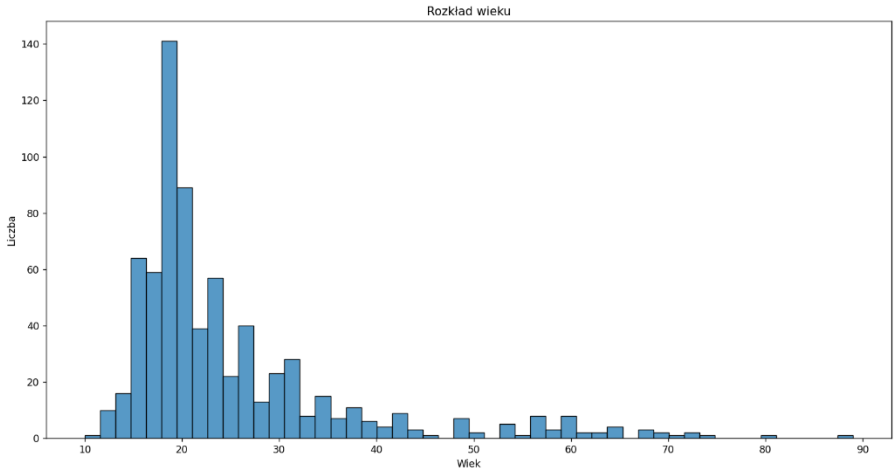
Jako, że nie miałam klasy docelowej w moim zbiorze danych, ponieważ w ankiecie mam 4 cechy określające stan psychiczny, stworzyłam nową cechę na bazie tych czterech, która będzie klasą docelową. Dokonałam podziału na osoby będące w: dobrym stanie psychicznym, średnim i słabym.

Następnie podzieliłam dane na zbiór testowy i treningowy, dokonałam implementacji wybranych modeli uczenia maszynowego takich jak: regresja logistyczna, k-najbliżsi sąsiedzi, las losowy, maszyna wektorów nośnych (SVM), wzmocnienie gradientowe

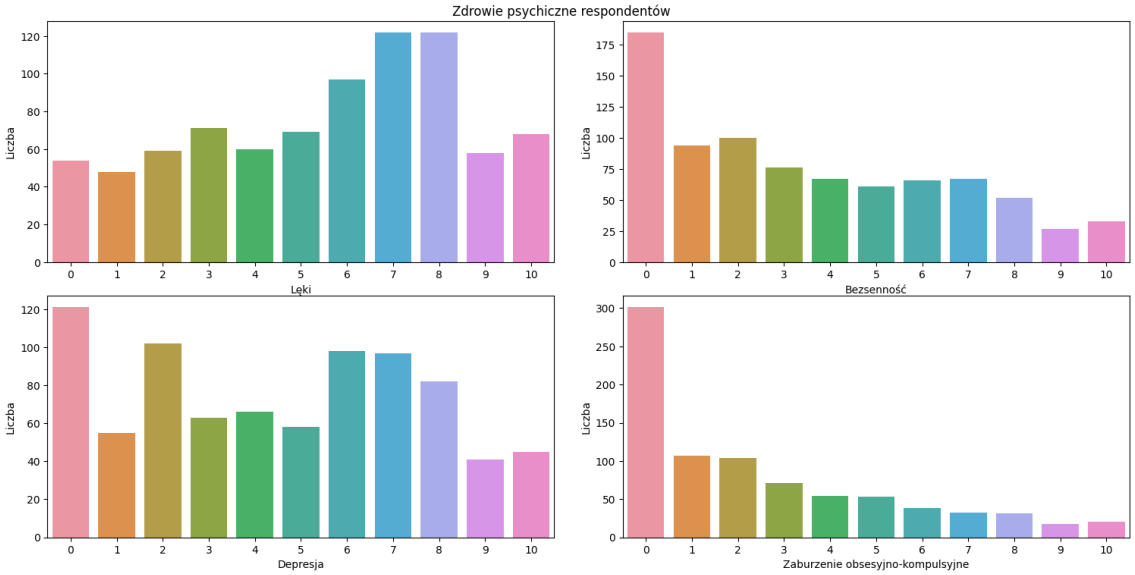
(gradient boosting). Dzięki tej bibliotece dokonałam oceny tych klasyfikatorów przy użyciu metryk: precyzja, czułość, dokładność, F-1 score, macierz pomyłek, krzywa ROC.
Badałam też wpływ doboru cech na jakość klasyfikacji przez analizę wariancji.
W trakcie pracy z danymi zauważyłam, że etykiety klasy docelowej są niezrównoważone. Najwięcej osób było w ciężkim stanie psychicznym. Nierównowaga klas może wpływać na proces uczenia maszynowego, szczególnie w przypadku algorytmów klasyfikacyjnych, które mogą być skłonne do skupiania się na dominującej klasie i osiągania niższej dokładności dla mniej licznych klas, zatem w celu poprawy jakości klasyfikacji przeprowadziłam nadpróbkowanie (oversampling) za pomocą pakietu Imblearn. Wykonałam również bootstrapping dla każdego klasyfikatora, tworząc losowe podzbiory danych treningowych, by przekonać się czy ta metoda polepszy jakość mojej klasyfikacji. Do wizualizacji danych zostały wykorzystane biblioteki Matplotlib oraz Seaborn.

3) Wyniki

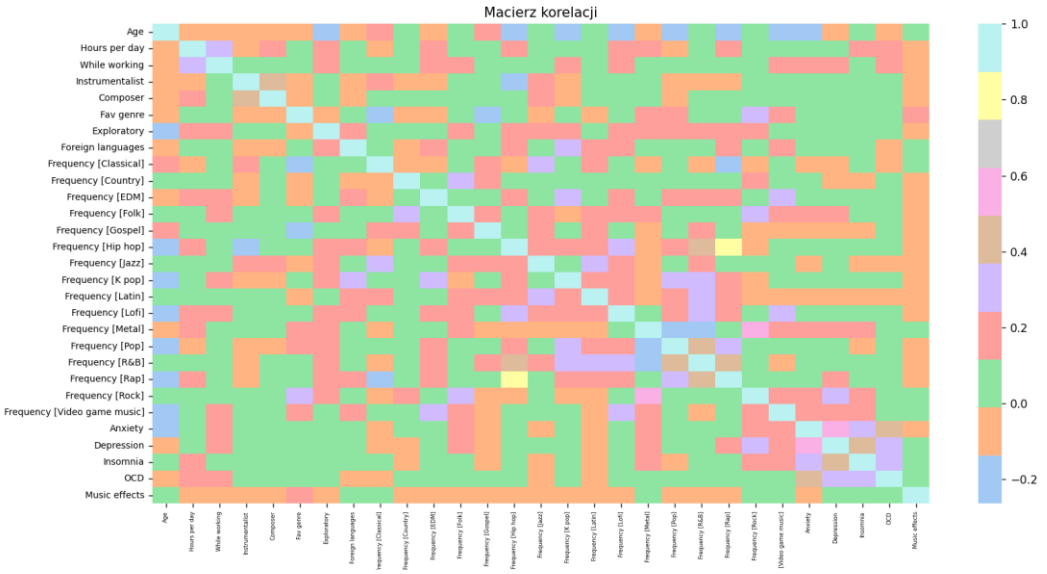
a) Rozkład wieku respondentów



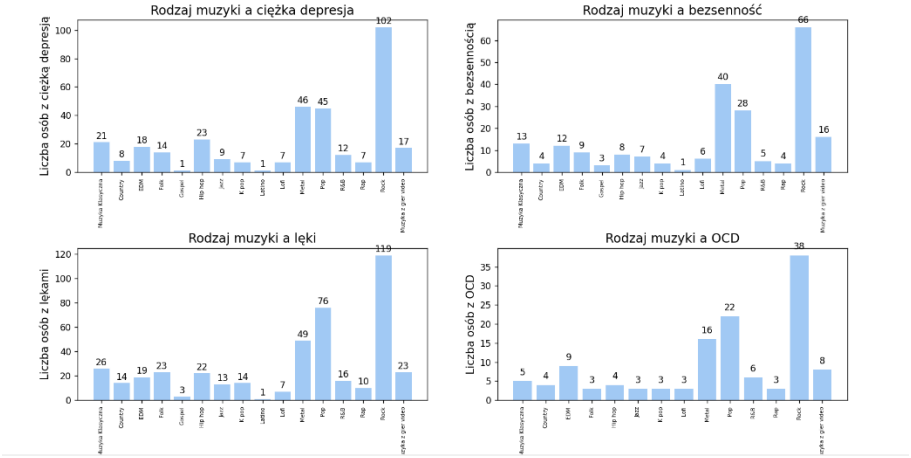
b) W ankiecie odpowiadający oceniali stan swojego zdrowia psychicznego w skali od 0 do 10, pytania były o ocenę odczuwanych lęków, bezsenności, depresji i zaburzeń obsesyjno-kompulsywnych.

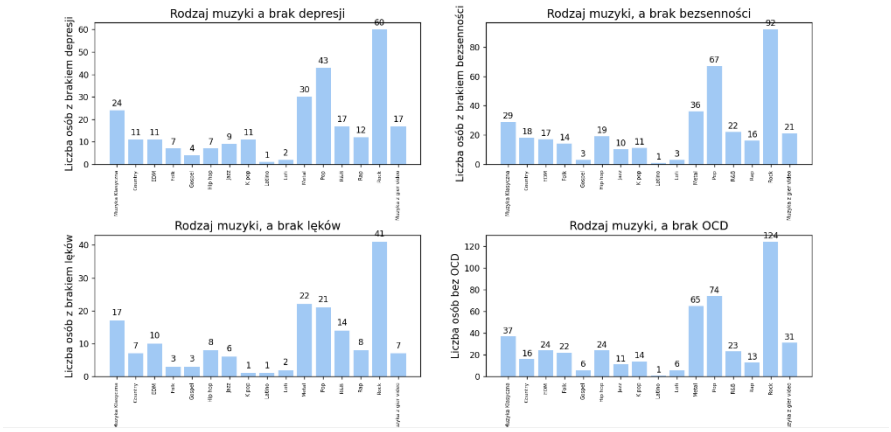


c) Macierz korelacji

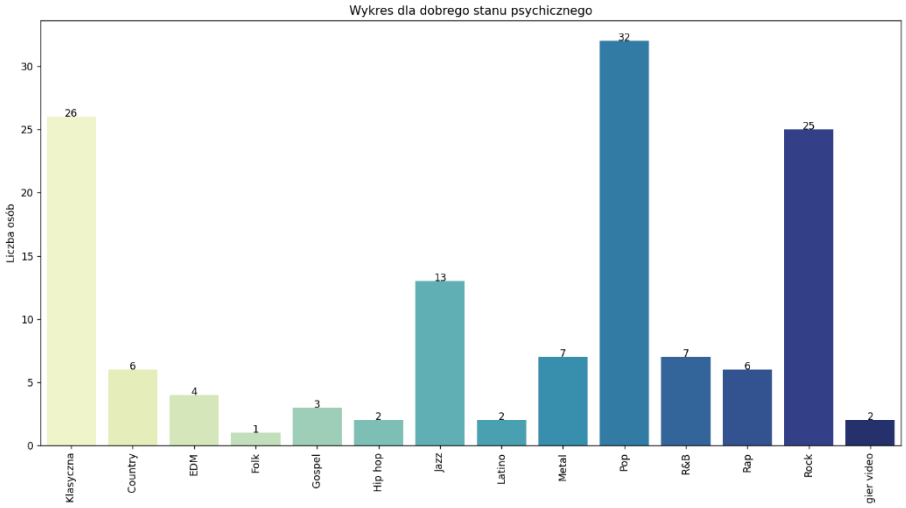


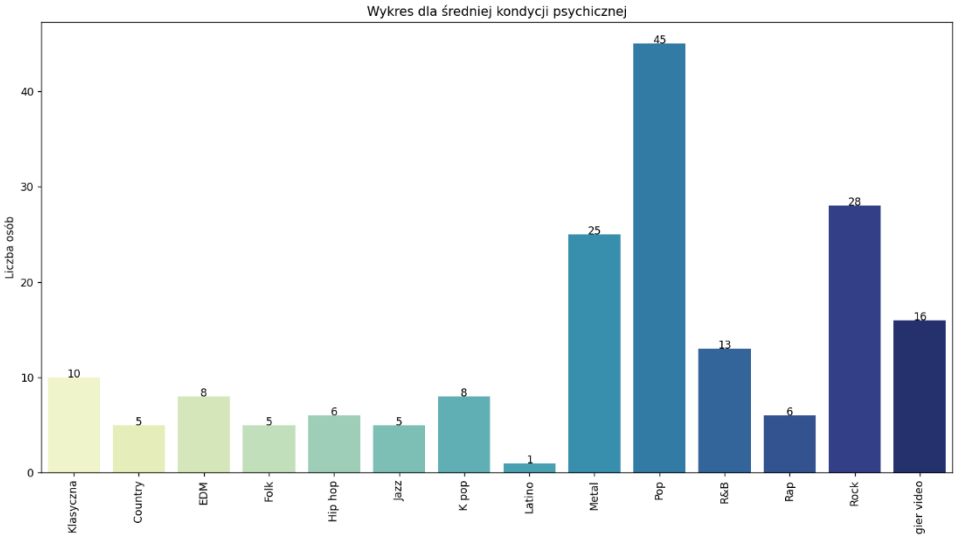
d) Ulubiony rodzaj muzyki, a stan depresja, bezsenność, lęki, zaburzenia obsesyjno-kompulsywne

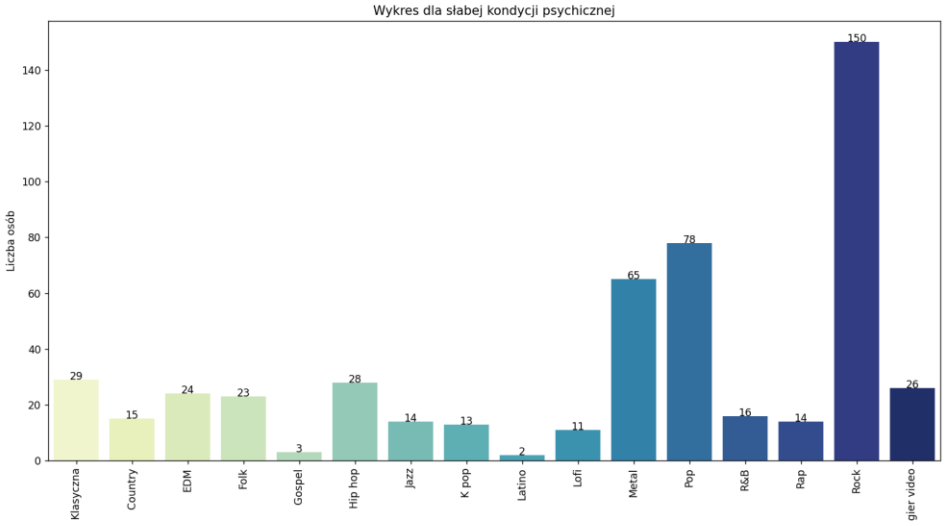




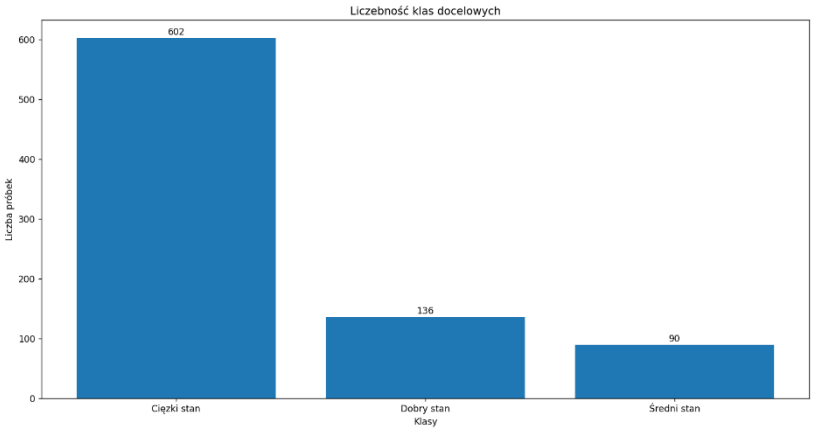
e) Zły, średni, dobry stan psychiczny, a ulubiony gatunek muzyczny







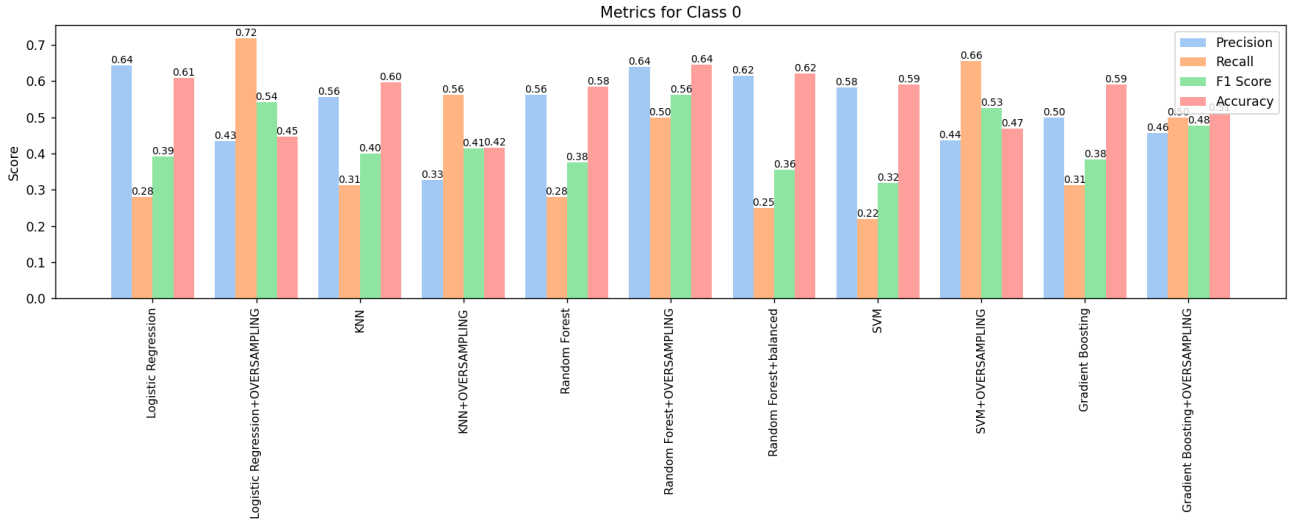
f) Liczebność klas docelowych

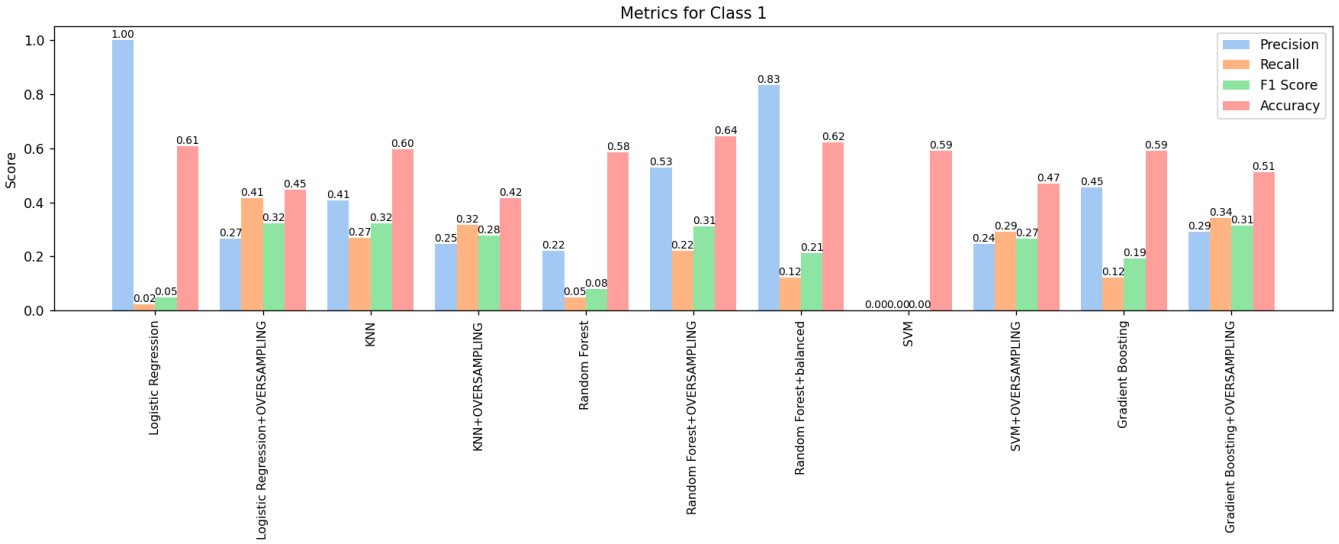


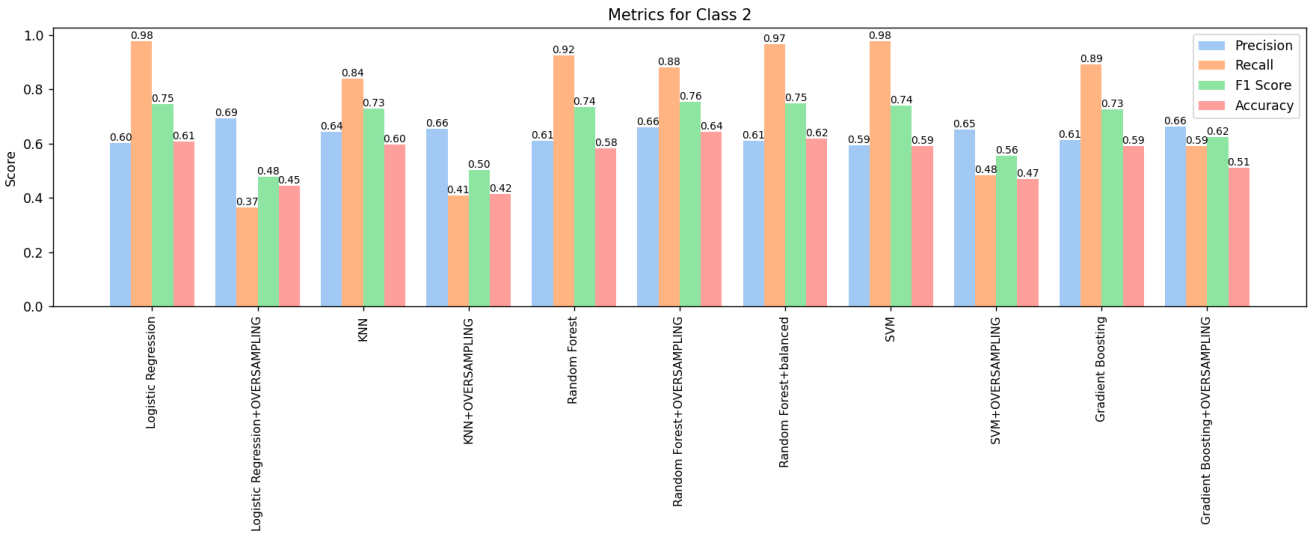
g) Analiza wariancji

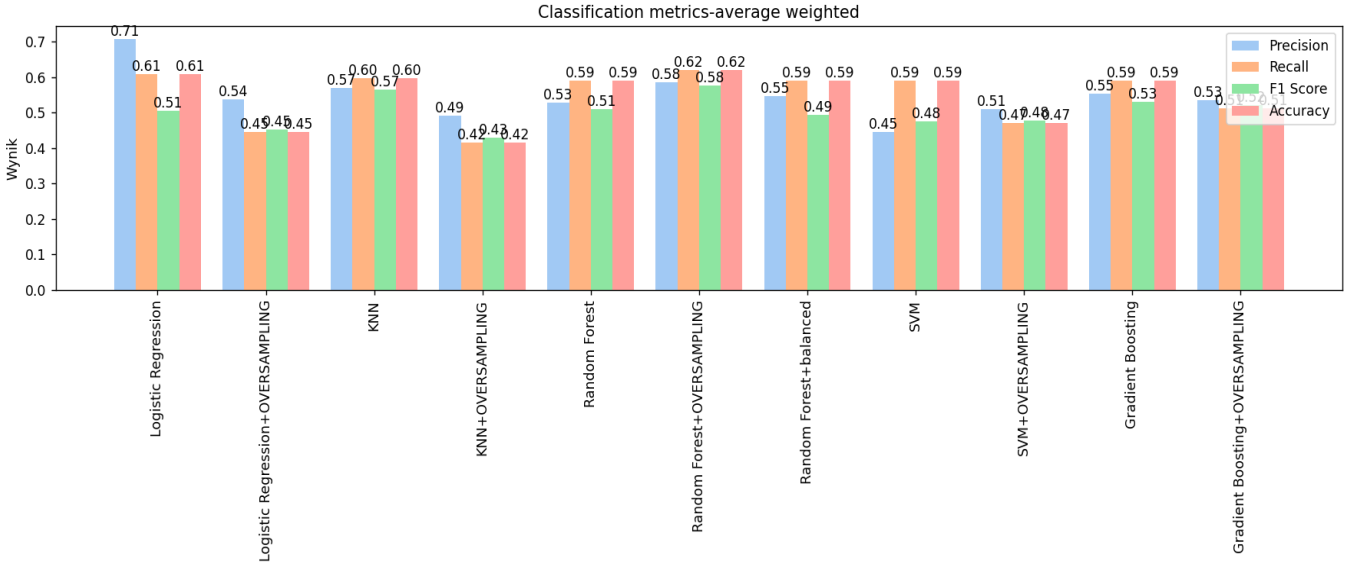
Age	140.239311
Hours per day	8.330096
While working	0.189298
Instrumentalist	0.212660
Composer	0.137273
Fav genre	22.702578
Exploratory	0.216300
Foreign languages	0.245588
Frequency [Classical]	0.981791
Frequency [Country]	0.822287
Frequency [EDM]	1.104826
Frequency [Folk]	1.012159
Frequency [Gospel]	0.511657
Frequency [Hip hop]	1.061593
Frequency [Jazz]	0.929532
Frequency [K pop]	0.936724
Frequency [Latin]	0.817224
Frequency [Lofi]	1.045684
Frequency [Metal]	1.303572
Frequency [Pop]	0.894121
Frequency [R&B]	1.140113
Frequency [Rap]	1.110987
Frequency [Rock]	1.169082
Frequency [Video game music]	1.166158
Anxiety	8.541430
Depression	9.609431
Insomnia	9.354902
OCD	7.865849
Music effects	0.241562
Mental Health Status	0.574558

h) Analiza metryk oceniających jakość klasyfikacji bez i z oversamplingiem; precyzja, czułość, dokładność, F-1 score

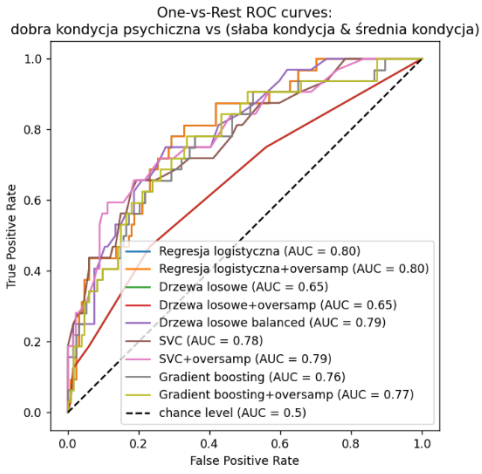


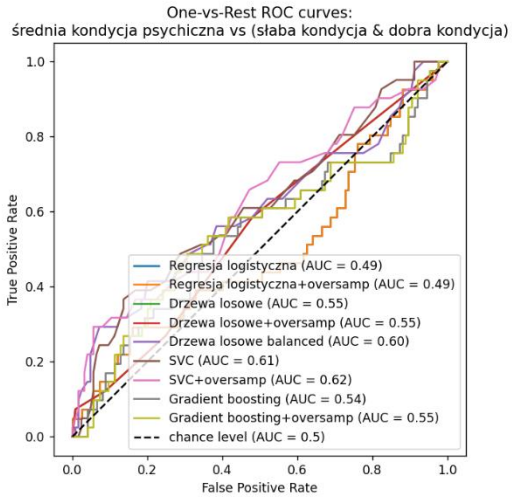


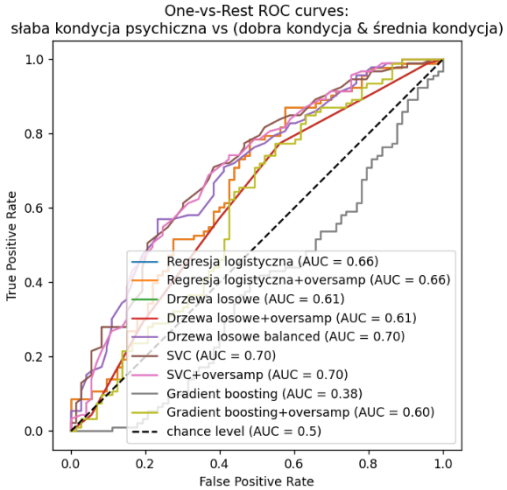




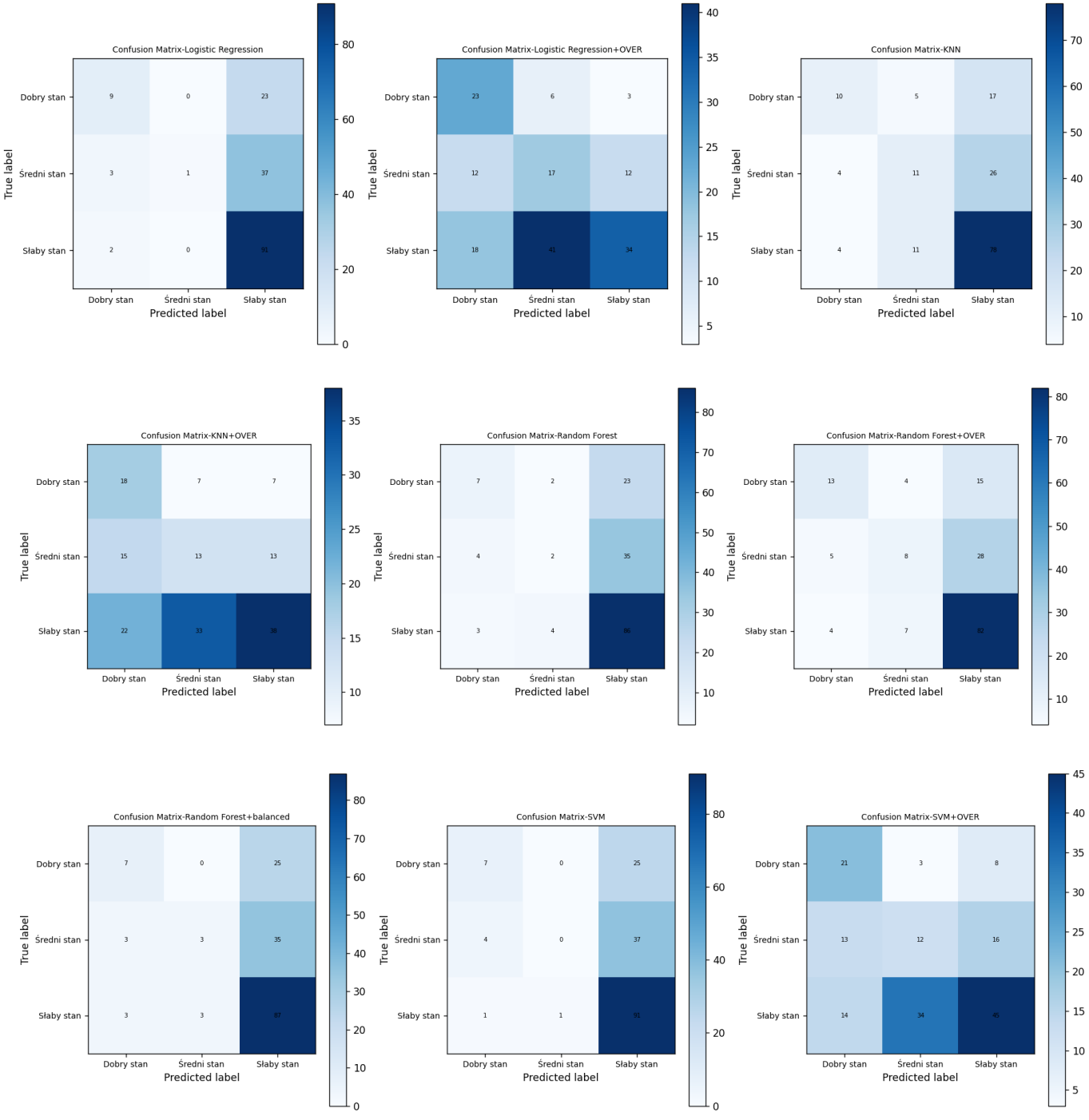
i) Krzywe ROC, klasyfikatory bez i z oversamplingiem

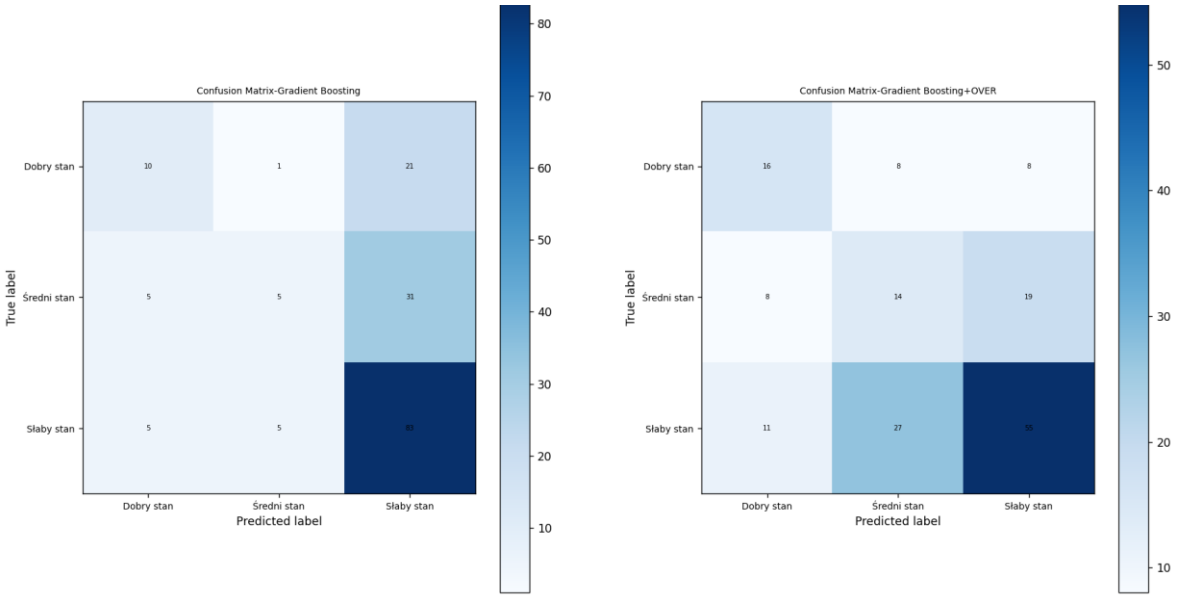




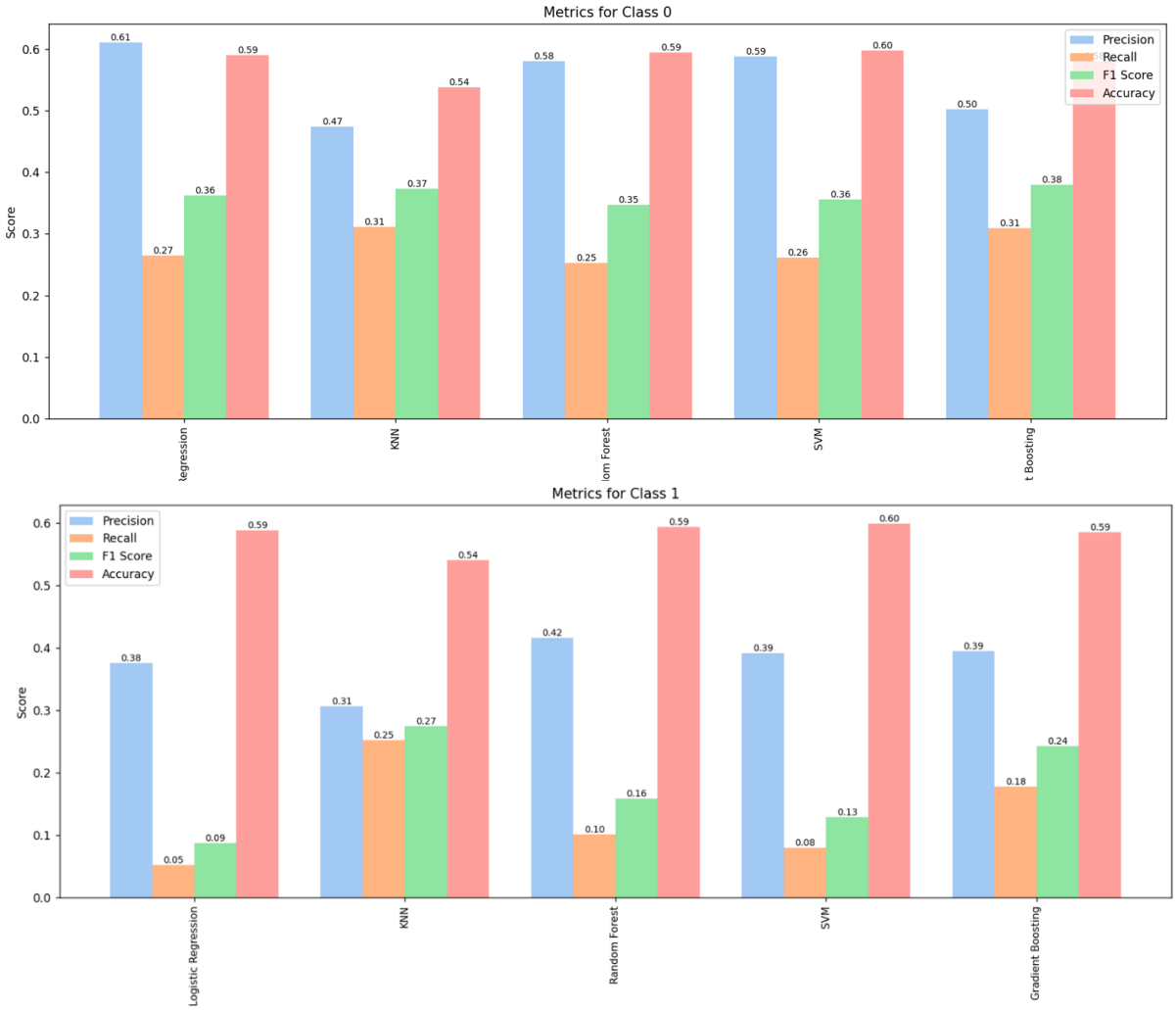


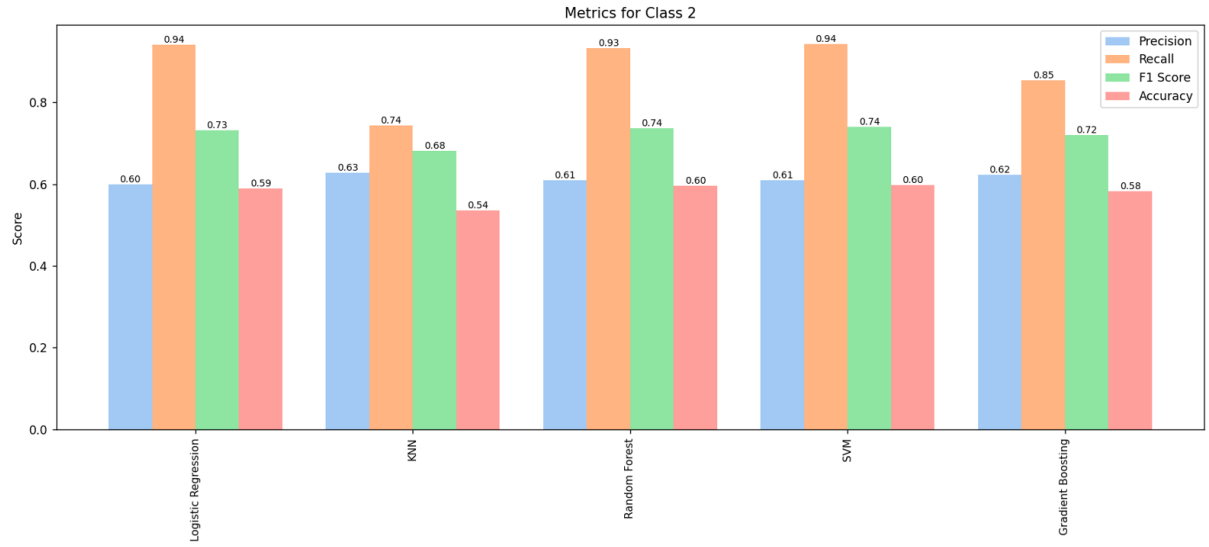
j) Macierze pomyłek

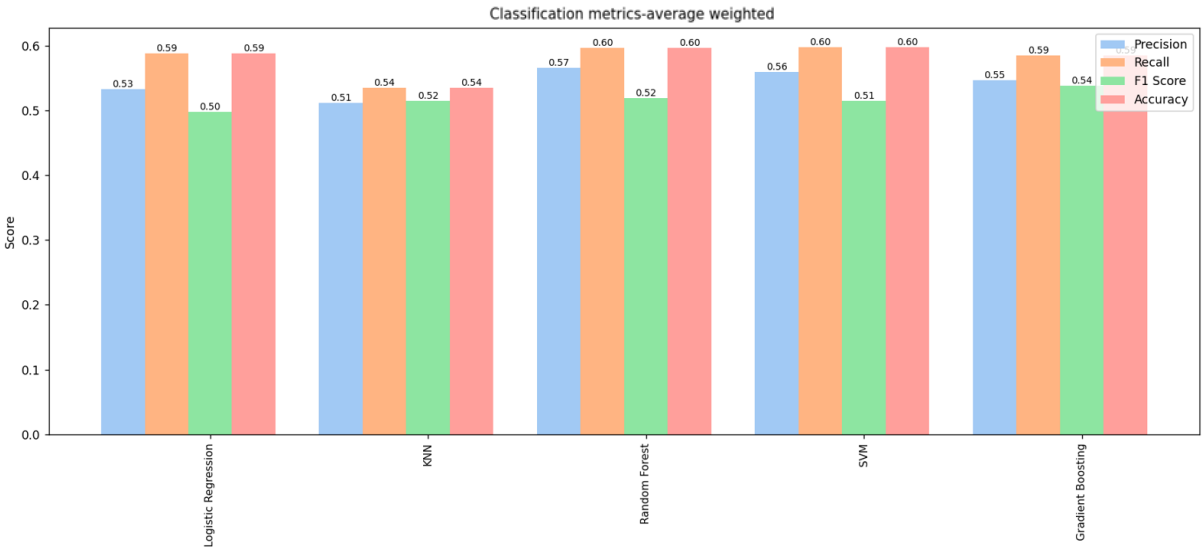




k) Metryki jakości klasyfikacji po użyciu metody bootstrap







4) Analiza wyników

Ad a) Wykres rozkładu wieku przedstawiający liczbę osób w zależności od wieku wykazał, że największa liczba osób ankietowanych koncentruje się w okolicach wieku 19 i 20 lat.

Ad b) W ankiecie odpowiadający oceniali stan swojego zdrowia psychicznego w skali od 0 do 10, pytania były o ocenę odczuwanych lęków, bezsenności, depresji i zaburzeń obsesyjno-kompulsywnych. Największa liczba ankietowanych zgłosiła brak objawów bezsenności, depresji czy zaburzeń obsesyjno-kompulsywnych. Jeśli chodzi o lęki, blisko 120 osób oceniło je na poziome 7-8. Jeśli chodzi o depresję ok. 100 ankietowanych oceniło jej stan na 2, co interpretuję jako stan gorszego samopoczucia, niekoniecznie musi to być depresja. Dokonując analizy zaburzeń obsesyjno-kompulsywnych w tym przypadku jest najmniej odpowiedzi innych niż 0, może to wynikać z małej świadomości czym się objawiaja ta choroba.

Ad c) Macierz korelacji nie dała mi oczekiwanych efektów tzn. nie zaobserwowałam silnych zależności między cechami określającymi stan zdrowia psychicznego, a

ulubionymi rodzajami i częstotliwością słuchanej muzyki. Współczynnik korelacji między ulubionym rodzajem słuchanej muzyki, a stanem psychicznym wynosi 0.13.

Zaobserwowałam silniejsze korelacje między czterema cechami określającymi stan psychiczny, jest to dość oczywiste, szczególnie jeśli chodzi o związek depresji i lęków. Depresja i lęki są dwoma odrębnymi zaburzeniami, ale są często wzajemnie powiązane. Choć mają różne objawy i mechanizmy, mogą się nawzajem nasilać i współistnieć u jednej osoby. Osoby z depresją często doświadczają również objawów lękowych.

Ponadto ciekawym jest fakt, że współczynnik korelacji między częstotliwością słuchania rapu i hip-hopu jest wysoki-0.76. Częstotliwość słuchania rocka i metalu również ma ze sobą związek, wartość współczynnika wynosi 0.52.

Ad d) Jako, że analiza macierzy korelacji nie wpłynęła znacznie na rozwiązywany przeze mnie problem, kolejnym krokiem było badanie związku między preferencjami muzycznymi, a stanem psychicznym.

Określiłam, że osoby dające ocenę depresji, bezsenności, zaburzeniom obsesyjno-kompulsywnych w zakresie od 0 do 3, jako stan dobry (ich ocena ewentualnie wynika z obniżenia nastroju), te oceniające go od 7 do 10, jako stan ciężki.

Zarówno w przypadku ciężkich stanów psychicznych jak i dobrych przeważającym ulubionym rodzajem muzyki był rock, co nie daje nam żadnych podstaw do stwierdzenia, że konkretny gatunek muzyczny jest ulubionym wśród osób w złym bądź dobrym stanie psychicznym.

Ad e) Jako, że nie udało mi się wysunąć żadnych wniosków dotyczących rodzaju słuchanej muzyki, a stanem zdrowia psychicznego analizując bezsenność, lęki, zaburzenia obsesyjno-kompulsywne z osobna, dodałam nową kolumnę określającą stan: zły, średni, dobry na podstawie tych czterech cech. Sklasyfikowałam osoby, które wszystkie cechy oceniły w zakresie 0-3 jako stan dobry. Za osoby w stanie ciężkim uznałam te, które oceniły chociaż jedną z cech w przedziale 7-10, pozostałe osoby są określone jako będące w stanie średnim.

Na tych wizualizacjach mogę zaobserwować, że najwięcej osób, które są w:

- dobrym kondycji psychicznej wskazuje jako swój ulubiony: pop(32), muzyka klasyczna(26), rock (25).
- średniej kondycji psychicznym wskazuje jako swój ulubiony: pop(45)
- słabej kondycji psychicznej wskazuje jako swój ulubiony: rock(150).

Ad f) Mój zbiór danych jest niezbalansowany. Klasą dominującą są osoby będące w złym stanie psychicznym.

Ad g) Następnie zajęłam się selekcją cech na podstawie wariancji. Po analizie wyników wykluczyłam cechy, które mają wartość wariancji poniżej 0.9. Testowałam różne wartości tego hiperparametru; dla 0.9 metryki oceniające klasyfikację miały najwyższe wartości.

Ad h) Jako, że zmagam się z problemem klasyfikacji wieloklasowej (mam 3 etykiety: zły, średni, dobry stan psychiczny) najpierw przedstawiłam wyniki metryk dla każdej

klasy z osobna. Te wykresy przedstawiają porównanie jakości działania różnych klasyfikatorów na oryginalnym zbiorze danych, a także na modelach trenowanych na danych nadpróbkowanych.

Jeśli chodzi o klasę 0 (dobry stan psychiczny), najwyższą czułość (recall)(0.72) uzyskał model regresji logistycznej trenowany na danych nadpróbkowanych, co sugeruje, że ma on dobrą zdolność do wykrywania przypadków tej klasy, jednak precyzja i dokładność mają wartość jedynie ok. 0.40. Moją uwagę zwrócił także klasyfikator lasu losowego z oversamplingiem, tu wszystkie metryki mają przybliżone do siebie wartości (zakres 0.50-0.64).

Metryki klasy 1 (średni stan psychiczny) mają bardzo rozbieżne wartości, np. w przypadku regresji logistycznej wysoka precyzja (równa 1) przy jednocześnie niskiej czułości i niskim F1 może sugerować, że model jest przetrenowany.

Metryki klasy 2(zły stan psychiczny) mają najlepsze wyniki, w porównaniu do poprzednich klas. Spośród wszystkich klasyfikatorów w których trenowałam model zbalansowanym zbiorze danych dobrze jedynie sprawdził się las losowy, reszta modeli daje słabsze wyniki przy użyciu tej metody. Wszystkie klasyfikatory trenowane na oryginalnym zbiorze dały całkiem zadowalające wyniki. Wartości czułości są bardzo wysokie, nawet na poziome 0.98.

Jako, że ciężko było wyłonić klasyfikator, który dawał najlepsze wyniki metryk przeszłam do stworzenia wykresu średniej metryki. Użyłam uśrednienia ważonego, które uwzględnia równowagę klas. Tutaj najlepszym modelem okazał się model regresji logistycznej oraz model lasu losowego z metodą oversamplingu.

Ad i) Dalszym etapem oceny wydajności klasyfikatorów było użycie krzywej ROC. Krzywa ROC jest mniej podatna na wpływ niezbalansowanych danych, w których jedna klasa dominuje nad drugą, zatem postanowiłam ją wykorzystać. Jako, że mój problem dotyczy klasyfikacji wieloklasowej użyłam metody One-vs-Rest, która porównuje każdą klasę ze wszystkimi pozostałymi. Najwyższe wartości AUC wskazują na lepszą zdolność klasyfikatora do rozróżniania między klasami.

Najwyższe AUC (0.80) dostałam dla porównania: dobry stan vs średni stan, zły stan, dla regresji logistycznej i regresji logistycznej z oversamplingiem. Najgorszy wynik uzyskałam dla drzew losowych z oversamplingiem (0.65).

Jeśli chodzi o średni stan psychiczny vs dobry, zły wszystkie wyniki AUC dla klasyfikatorów wahają się na poziome ok. 0.50, oznacza to, że klasyfikatory mają losową wydajność.

Słaby stan vs zły, dobry najgorszy wynik AUC jest dla gradientowego wzmocnienia (gradient boosting) (0.38), a drzewa losowe zbalansowane i SVC z oversamplingiem dają wyniki na poziome 0.70.

Ad j) Przeanalizowałam macierze pomyłek, zauważyłam, że w większości klasyfikatorów ma tendencję do klasyfikowania błędnie średniego stanu jako stanu słabego. Klasyfikatorami, które dały dość satysfakcjonujące wyniki to gradientowe wzmocnienie (gradient boosting) i las losowy (random forest).

Ad k) Jako, że nadpróbkowanie z pakietu Imblearn nie spowodowało znacznej poprawy klasyfikacji, postanowiłam spróbkować także metody bootstrap. W metryce ważonej 3 klas wyniki dla KNN, gradientowego wzmocnienia i lasu losowego nie różnią się znacznie od tych z wykorzystanym oversamplingiem czy też bez niego. Jednak w regresji logistycznej zaobserwowałam spadek precyzji.

Analizując każdą z klas z osobna, metryki miały nieco niższe wartości niż w wcześniej prezentowanych metrykach, aczkolwiek zauważyłam zmiany w klasie 1 (średni stan zdrowia psychicznego).

5) Wnioski

Analiza wpływu muzyki na stan zdrowia psychicznego okazała się trudnym zadaniem. Dane były niezbalansowane oraz było ich dość mało.
Uważam, że zgromadzenie większej liczby ankietowanych w badaniu mogłoby pomóc w uzyskaniu bardziej reprezentatywnego obrazu. Próbka powinna być bardziej różnorodna pod względem wieku i innych istotnych czynników, aby wyniki były bardziej ogólne i wiarygodne. W dodatku 4 cechy oceniające swój stan zdrowia psychicznego to mała liczba, być może stworzenie bardziej szczegółowej ankiety dostarczyło by więcej informacji na temat powiązania zdrowia psychicznego z muzyką.
Po oczyszczeniu danych, dokonałam analizy wizualizacji. Okazało się, że osoby oceniające swój stan psychiczny jako zły zaznaczyły, że ich ulubionym gatunkiem jest rock.
W średniej kondycji psychicznej ulubionym rodzajem muzyki jest pop, może to wynikać z faktu, że największa liczba ankietowanych była w wieku ok. 20 lat, a jest to jeden z popularniejszych gatunków wśród młodych osób. Nie udało się jednoznacznie określić przeważających ulubionych gatunków dla osób w dobrej kondycji psychicznej, z tego względu, że różnice w ilościach osób między rodzajami muzyki są niewielkie.
Te obserwacje są interesujące i sugerują pewne korelacje, ale nie można jednoznacznie stwierdzić, czy istnieje związek przyczynowo - skutkowy między preferowanymi gatunkami muzycznymi, a stanem zdrowia psychicznego.
Jeśli chodzi o problem klasyfikacji z jakim się mierzyłam spodziewałam się gorszych wyników jakości klasyfikatorów, jednak wyniki okazały się trudne dla mnie do interpretacji. Nie jestem w stanie określić jednego klasyfikatora, który sprawdził się na moim zbiorze danych najlepiej, wszystko zależy od naszych oczekiwań względem modelu np. czy model ma dobrze rozróżniać stan dobry od złego i średniego czy też np. zły od dobrego i średniego, ponieważ wyniki klasyfikatorów różniły się w tym przypadku.
Niewątpliwie najlepiej klasyfikatory radziły sobie z klasyfikacją osób w złym stanie psychicznym, ze względu na to, że była to najczęściej występująca klasa w zbiorze danych.
Przypuszczam, że na wyniki może też mieć wpływ moje własne kategoryzowanie osób według ciężkiej, średniej i dobrej kondycji psychicznej, ponieważ z 4 cech określających stan zdrowia psychicznego (depresja, lęki, zaburzenia obsesyjno-kompulsywne, bezsenność) utworzyłam jedną zmienną objaśnianą.
Pomimo tego, że podejmowałam próbę oversamplingu jak i bootstrapingu nie przyniosło to znacznej poprawy.

Zarówno z krzywej ROC jak i z macierzy pomyłek wywnioskowałam, że problemem klasyfikatorów było rozróżnienie średniej kondycji psychicznej od pozostałych.

Bazując na własnej wiedzy, że zdrowie psychiczne jest wynikiem wielu złożonych czynników, mogę przypuszczać, że choć muzyka może mieć wpływ na nasze samopoczucie, to nie jest ona jedynym ani dominującym czynnikiem determinującym nasze zdrowie psychiczne właśnie dlatego trudno wysunąć jednoznaczne wnioski.

Podczas wykonywania ćwiczenia nauczyłam się pracy z nieoczyszczonymi danymi, dużo czasu zajęło mi przygotowanie obu zbiorów danych, a następnie ich połączenie w całość. Poprawiłam moje umiejętności z programowania, ze względu na to, że wykonałam wiele wizualizacji danych, a także oceny jakości klasyfikatorów. Zgłębiłam wiedzę na temat różnych metod klasyfikacji, a także metod pracy z niezbalansowanym zbiorem danych.