WSI Laboratorium 4
Algorytm SVM

Filip Misztal 310276

13 grudnia 2023

1 Opis Implementowanych Algorytmów

Algorytm trenowania SVM (pseudokod)

1. Inicjalizuj wagi i obciążenia na początku treningu
2. Dla każdej epoki:
   (a) Oblicz funkcję decyzyjną dla wszystkich obserwacji ($decision\_function$)
   (b) Oblicz funkcję straty z uwzględnieniem marginesu i regularyzacji ($hinge\_loss$)
   (c) Oblicz całkowitą funkcję straty ($loss$)
   (d) Oblicz gradient funkcji straty względem wag ($gradient\_weights$)
   (e) Oblicz gradient funkcji straty względem obciążenia ($gradient\_bias$)
   (f) Zaktualizuj wagi i obciążenia
3. Zwróć wytrenowany model SVM

2 Opis planowanych eksperymentów numerycznych

W celu przetestowania możliwości zaimplementowanego algorytmu skorzystałem ze zbioru danych "Wine Quality Data Set"z repozytorium *ucimlrepo*. Ze względu na to, iż kolumna *quality* będąca targetem nie była binarna tylko przymowała wartości od 0 do 10, zdecydowałem się przeprowadzić badania w trzech wariantach:

1. Klasyfikacja danych według kolumny *quality* podzielonej na dwie klasy:
   (a) podział w połowie zakresu: 1-5 oraz 6-10
   (b) podział względem średniej arytmetycznej z wartości w kolumnie *quality*: 1-6 oraz 7-10
2. Klasyfikacja danych według kolumny *color* - klasy *red* i *white*

We wszystkich musiałem poczynić odpowiednie przekształcenia zbioru danych, aby nadawał się do wytrenowania modelu. W pierwszym i drugim wariancie zmieniłem etykiety na wartości -1 i 1 dla odpowiednio klasy przedziału 1-5 i 6-10 oraz 1-6 i 7-10. W trzecim wariancie przypisałem wartość -1 dla koloru czerwonego i 1 dla białego, a także nie brałem pod uwagę kolumny *quality*.

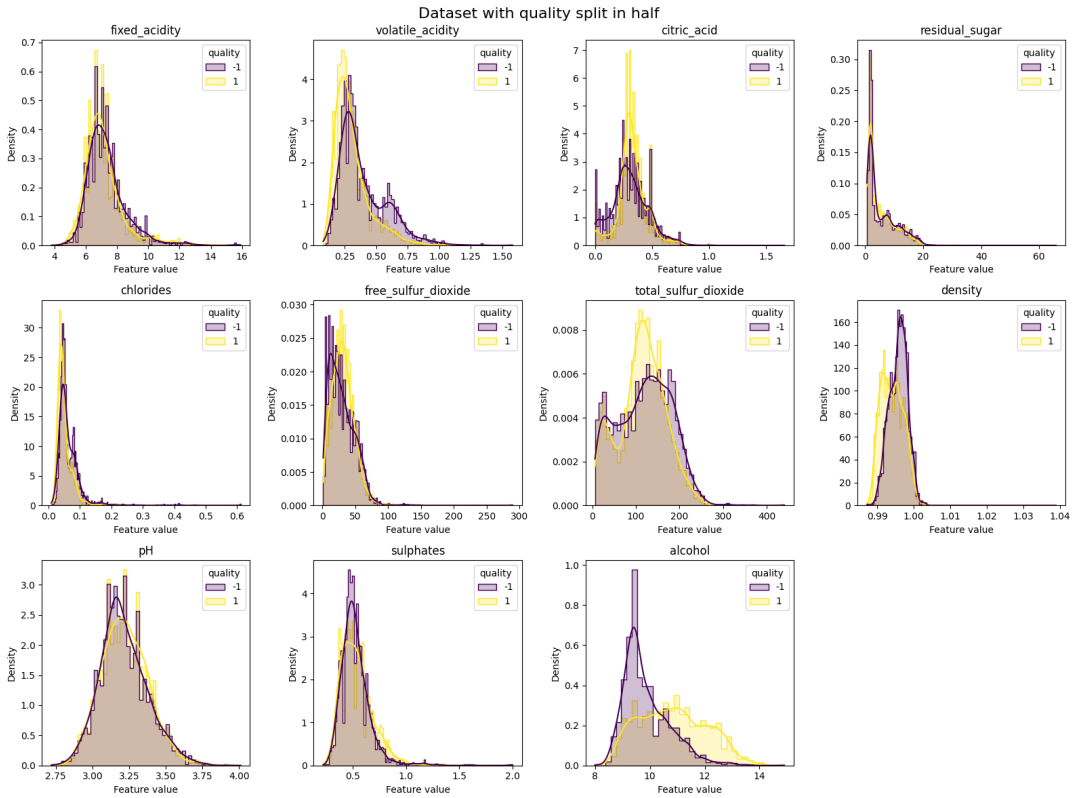
Następnie zbiory cech poddałem skalowaniu aby wyrównać zakres wartości cech i poprawić w ten sposób stabilność procesu uczenia. Konieczne było także podzielenie zbioru na zbiór treningowy i uczący - u mnie w stosunku 80% i 20%.
Po takim przygotowaniu danych zacząłem badać efektywność działania algorytmu oraz dokładność, czułość i specyficzność uzyskanego modelu w zależności od:

1. Wartości learning_rate
2. Ilości iteracji procesu uczenia
3. Wartości parametru regulacyjnego
4. Użytego przekształcenia jądrowego (liniowe i gausowskie)

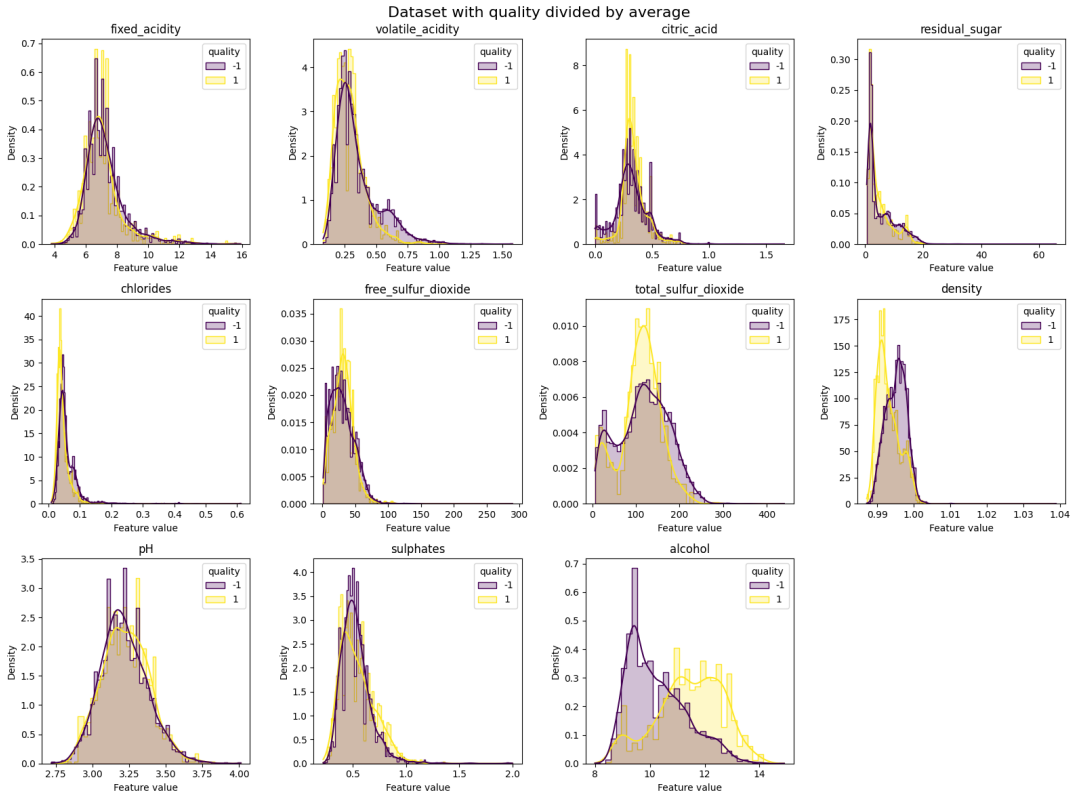
3 Opis uzyskanych Wyników

3.1 Analiza wariantów danych treningowych

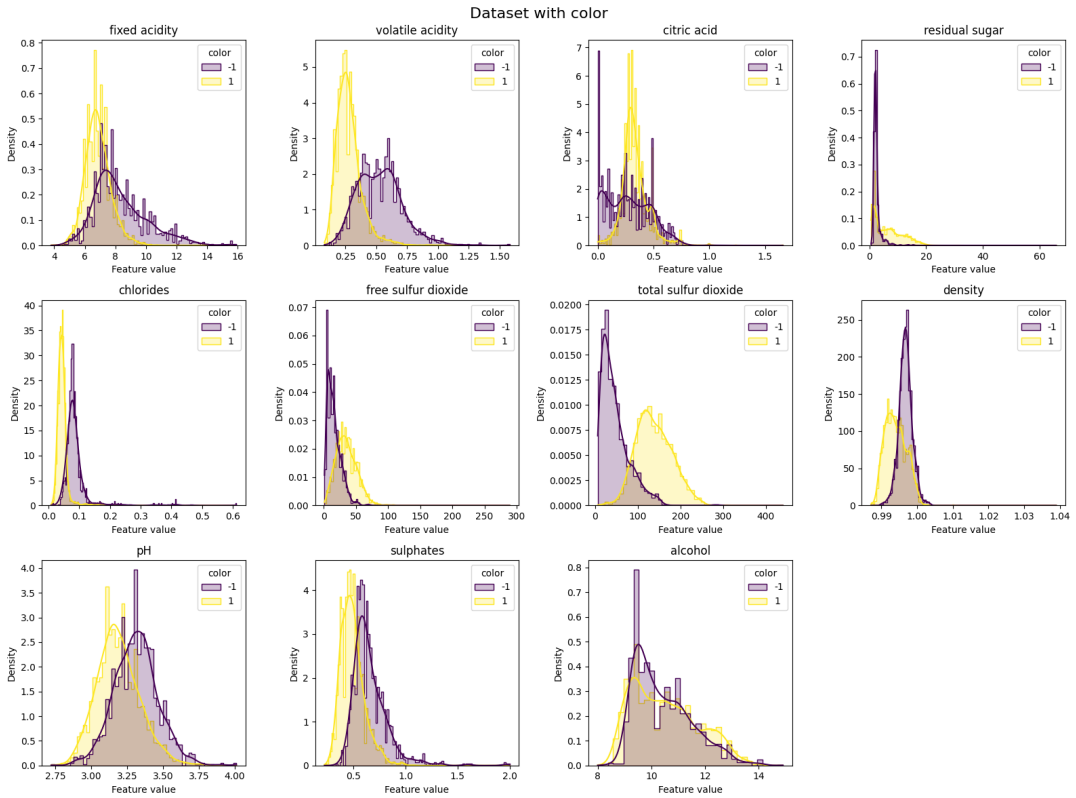
Sposoby, w jakie dokonuje się podziału zbiorów danych na różne klasy, mają istotny wpływ na jakość rozgraniczenia cech w analizowanym materiale. Różnorodność tych metod może prowadzić do subtelnych lub wyraźnych różnic w interpretacji danych, co z kolei wpływa na precyzję analizy i ddziałania klasyfikatora na tych danych. Na poniższych wykresach widać, jak W zależności od przyjętych klas lepiej albo gorzej widoczne były różnice między nimi w konkretnych cechach.



Rysunek 1: Wykresy gęstości dla poszczególnych cech w zbiorze z *quality* podzielonym na pół



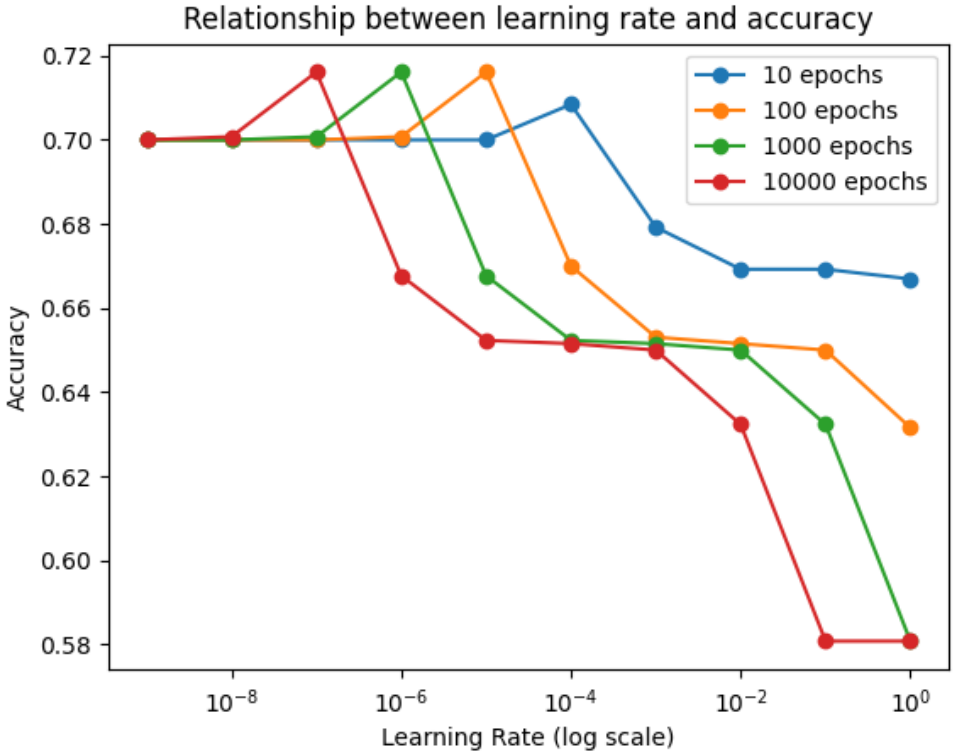
Rysunek 2: Wykresy gęstości dla poszczególnych cech w zbiorze z *quality* podzielonym według średniej



Rysunek 3: Wykresy gęstości dla poszczególnych cech w zbiorze z *color*

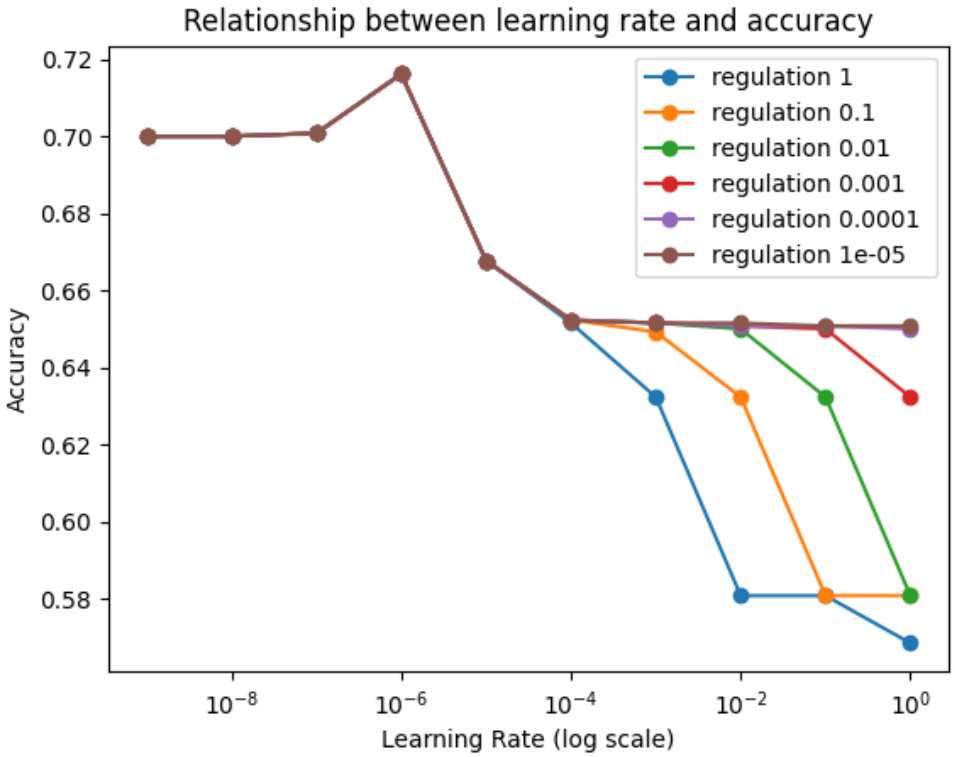
3.2 Wpływ parametrów SVM na dokładność modelu

Badania przeprowadzone na datasecie z klasyfikacją według quality podzielonego na przedziały 1-5 i 6-10, dla parametru regulacyjnego 0.01 i liniowej funkcji jądrowej:



Rysunek 4: Wykres zależności dokładności od wartości learning rate'u i ilości użytych epok

Jak widać na powyższym wykresie, większa ilość epok użytych w procesie jest odwrotnie skorelowana do dokładności modelu. Może to być spowodowane zjawiskiem przeuczenia. Powinien temu zapobiegać parametr regulacyjny, którego wpływ przeanalizowałem poniżej:

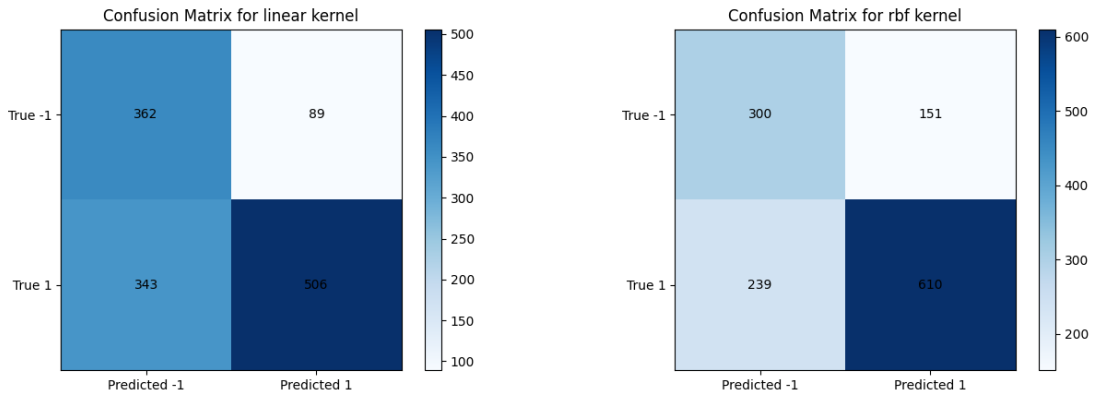


Rysunek 5: Wykres zależności dokładności od wartości learning rate'u i wartości parametru regulacyjnego

Można tutaj zauważyć, że dla każdej wartości learning rate'u istniała pewna górna granica dopasowania modelu. Im wyższa była wartość parametru regulacji, tym niższa musiała być szybkość uczenia, aby model osiągnął tą granicę.

3.3 Wpływ użytej funkcji jądrowej

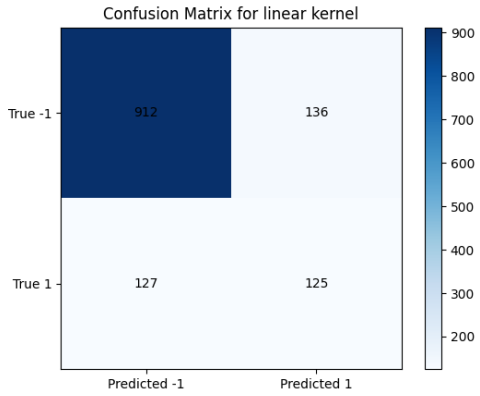
Te badania zdecydowałem się przeprowadzić na wszystkich 3 wcześniej opisanych datasetach, wszędzie przy stałych ustalonych parametrach SVM: learning_rate 10^{-6}, epochs 10^{3}, regulation_param 10^{-5}. W przeprowadzonej ocenie zwracałem uwagę głównie na osiąganą przez model dokładność, czułość i specyficzność.



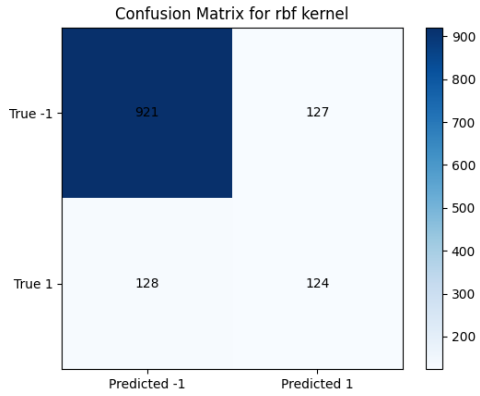
(a) Linear
Accuracy: 66.77% Sensitivity: 59.60% Specificity: 80.27%

(b) Rbf
Accuracy: 70.00% Sensitivity: 71.85% Specificity: 66.52%

Rysunek 6: Macierze pomyłek uzyskane dla zbioru z quality podzielonym na pół na dwie klasy

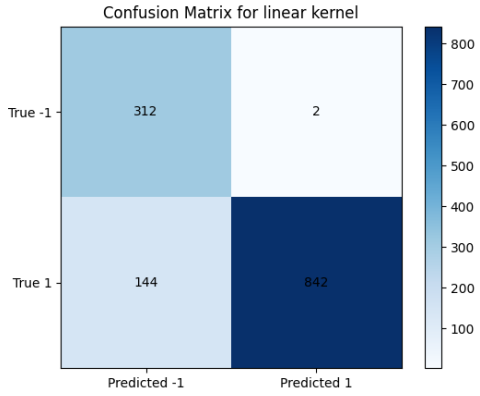


(a) Linear
Accuracy: 79.77% Sensitivity: 49.60% Specificity: 87.02%

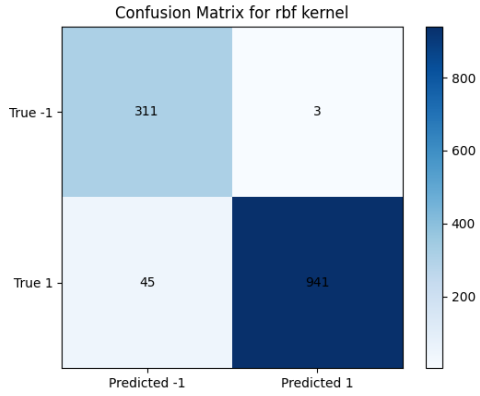


(b) Rbf
Accuracy: 80.38% Sensitivity: 49.21% Specificity: 87.88%

Rysunek 7: Macierze pomyłek uzyskane dla zbioru z quality podzielonym na dwie klasy według średniej



(a) Linear
Accuracy: 88.77% Sensitivity: 85.40% Specificity: 99.36%



(b) Rbf
Accuracy: 96.31% Sensitivity: 95.44% Specificity: 99.04%

Rysunek 8: Macierze pomyłek uzyskane dla zbioru z kolumną color warunkującą klasy

Widać na powyższych wykresach, że bardzo duży wpływ na skuteczność klasyfikacji przez SVM miał sposób, w jaki zdecydowaliśmy się podzielić dane. Najlepsze efekty przyniosła klasyfikacja względem koloru. Ponadto użycie przekształcenia jądrowego gausa niewątpliwie poprawiło osiąganą przez model dokładność.

4 Wnioski z przeprowadzonych badań

Podsumowując przeprowadzone badania, można wyróżnić kilka istotnych wniosków:

- **Wpływ wariantów danych treningowych:** Wybór sposobu grupowania danych na klasy istotnie wpływa na skuteczność modelu. Klasyfikacja względem koloru (czerwony/biały) przyniosła najlepsze rezultaty, co sugeruje, że dobór właściwego kryterium podziału ma kluczowe znaczenie dla skuteczności klasyfikacji.
- **Wpływ parametrów SVM na dokładność modelu:** Z przeprowadzonych badań wynika, że istnieje delikatna równowaga pomiędzy wartością learning rate a ilością epok. Wartość parametru regulacyjnego także ma istotny wpływ na skuteczność modelu, zbyt duża wartość może prowadzić do utraty dokładności.

- **Wpływ użytej funkcji jądrowej:** Użycie funkcji jądrowej typu RBF (radial basis function) poprawiło znacząco skuteczność klasyfikacji w porównaniu do funkcji liniowej. Ponadto, klasyfikacja względem koloru (czerwony/biały) osiągnęła najwyższą dokładność, co potwierdza, że odpowiednie dostosowanie funkcji jądrowej do charakterystyki danych może znacznie poprawić wyniki modelu SVM.

Ostatecznie, wybór odpowiednich parametrów i prezentacja danych mają kluczowe znaczenie dla skuteczności klasyfikacji przy użyciu algorytmu SVM.