PROJEKT 1. WYKORZYSTANIE KLASYCZNYCH MODELI UCZENIA MASZYNOWEGO DO ROZWIĄZYWANIA PROBLEMÓW ML.

Cel: należy określić cel/problem projektu. Materiały: złożony zestaw danych (wielowymiarowa macierz cech), który należy odpowiednio przygotować.

**Metody:** należy stworzyć co najmniej 4 modele uczenia maszynowego.

**SVM**

Dla modelu SVM (Support Vector Machine) skonstruowano wersję domyślną oraz zoptymalizowaną. SVM jest algorytmem klasyfikacyjnym, który poprzez odpowiednie ustawienie hiperparametrów stara się wyznaczyć optymalną hiperpowierzchnię rozdzielającą próbki danych należące do różnych klas. W modelu domyślnym użyto liniowego jądra, co pozwala na proste rozdzielenie danych, natomiast w modelu zoptymalizowanym zastosowano jądro RBF (z parametrami C=10 i gamma='auto'), które pozwala na lepsze dostosowanie się do bardziej złożonej struktury danych.

Optymalizacja hiperparametrów w wersji zoptymalizowanej odbyła się przy użyciu wyszukiwania siatki (Grid Search) i walidacji krzyżowej, co umożliwiło znalezienie najlepszych ustawień dla parametrów C i gamma. Parametr C kontroluje margines decyzyjny, pozwalając modelowi lepiej zróżnicować kategorie w przypadku klasy "Good" i "Bad", a gamma wpływa na zakres działania jądra RBF, pomagając modelowi lepiej uchwycić wzorce w danych.

**Drzewo decyzyjne**

W projekcie zastosowano dwa modele drzewa decyzyjnego – jeden z domyślnymi ustawieniami, a drugi zoptymalizowany przy użyciu wyszukiwania siatki parametrów (Grid Search). Drzewo decyzyjne jest klasyfikacyjnym algorytmem uczenia maszynowego, który poprzez budowanie hierarchii decyzji na podstawie cech wejściowych grupuje dane do określonych klas. Modele te zostały stworzone w celu przewidywania zmiennej docelowej GVB, która informuje, czy klub piłkarski jest "dobry" (1), czy "zły" (0).

**Las losowy**

Dla modelu lasu losowego (random forest) zbudowano wersję domyślną oraz zoptymalizowaną z dostosowanymi ustawieniami hiperparametrów. Random Forest to algorytm klasyfikacyjny oparty na zespole drzew decyzyjnych, który klasyfikuje dane na podstawie uśrednionych wyników z wielu drzew.

Ustawienia zoptymalizowanego modelu obejmują zwiększenie liczby drzew do 10,000 (n\_estimators=10000), co stabilizuje uśredniony wynik oraz zmniejsza wariancję predykcji. Głębokość drzew została ustawiona na 50 (max\_depth=50), co umożliwia modelowi lepsze uchwycenie złożoności danych. Parametr min\_samples\_split=2 pozwala na bardziej szczegółowe podziały w drzewach, a min\_samples\_leaf=1 zwiększa elastyczność modelu, umożliwiając każdemu liściowi większą precyzję. Dodatkowo, ustawienie max\_features='sqrt' ogranicza liczbę używanych cech w każdym podziale do pierwiastka z całkowitej liczby cech, co sprzyja różnorodności drzew i minimalizuje ryzyko nadmiernego dopasowania modelu do danych treningowych. Natomiast parametr bootstrap=True pozwala na losowe próbkowanie z powtórzeniami, co zwiększa stabilność modelu, uśredniając wyniki i ograniczając wpływ pojedynczych, nietypowych obserwacji na wynik końcowy.

**Klasyfikator głosujący**

Klasyfikator z głosowaniem większościowym (Voting Classifier) został przygotowany w dwóch wariantach: wersji bazowej z domyślnymi ustawieniami modeli oraz wersji zoptymalizowanej, w której dostosowano hiperparametry dla poszczególnych modeli składowych. Voting Classifier łączy prognozy z różnych algorytmów, co umożliwia uzyskanie stabilniejszych i bardziej precyzyjnych wyników.

W zoptymalizowanej wersji klasyfikatora zastosowano GridSearchCV i RandomizedSearchCV do doboru najlepszych hiperparametrów, co pozwoliło na znaczącą poprawę dokładności predykcji. W skład klasyfikatora weszły różne modele, w tym Random Forest, Gradient Boosting, SVM oraz K-Nearest Neighbors (kNN). Dzięki wykorzystaniu mechanizmu głosowania miękkiego (voting='soft') i nadaniu większych wag modelom Random Forest i Gradient Boosting, Voting Classifier lepiej uśrednia prognozy, co przekłada się na większą dokładność i stabilność modelu w radzeniu sobie ze złożonymi danymi.

**Wyniki i dyskusja:** należy stworzyć modele przy ustawieniach domyślnych i następnie je porównać z modelami zoptymalizowanymi przy użyciu odpowiednich wskaźników ewaluacji.

**SVM**

Analiza wyników modeli SVM z ustawieniami domyślnymi i zoptymalizowanymi opierała się na wskaźnikach oceny klasyfikacji oraz na wizualizacji wyników przy użyciu wykresów. Oba modele osiągnęły wysoką dokładność na testowym zbiorze danych. Poniżej przedstawiono szczegółowe wyniki:

**Wyniki modelu SVM z ustawieniami domyślnymi:**

* **Dokładność**: 91.89%
* **Raport klasyfikacji**:
  + Klasa "Bad" (0):
    - Precyzja: 0.93
    - Czułość: 0.91
    - F1-score: 0.92
  + Klasa "Good" (1):
    - Precyzja: 0.91
    - Czułość: 0.93
    - F1-score: 0.92

**Wyniki modelu SVM zoptymalizowanego (z parametrami C=10, gamma='auto', kernel='rbf'):**

* **Dokładność**: 92.34%
* **Raport klasyfikacji**:
  + Klasa "Bad" (0):
    - Precyzja: 0.93
    - Czułość: 0.92
    - F1-score: 0.92
  + Klasa "Good" (1):
    - Precyzja: 0.92
    - Czułość: 0.93
    - F1-score: 0.93

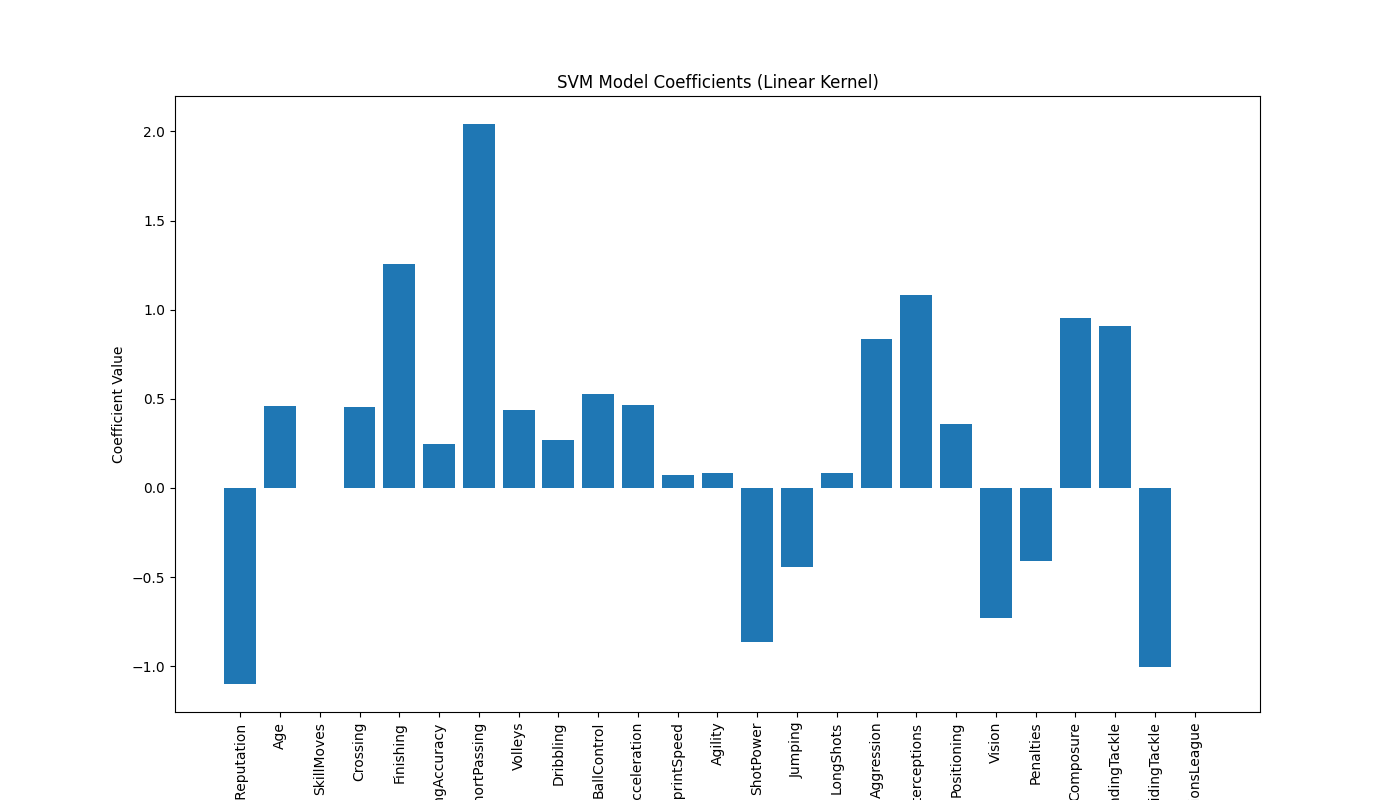
W wersji zoptymalizowanej model uzyskał minimalnie wyższe wartości w precyzji oraz czułości w przypadku klasy "Good", co pokazuje, że zmiana jądra z liniowego na RBF przyczyniła się do nieznacznej poprawy skuteczności. Wskaźnik F1-score również wzrósł o 0.01 dla klasy "Good", co jest szczególnie istotne w kontekście zbalansowania klasyfikacji między obiema klasami.

A collage of graphs

Description automatically generated

Rysunek 1 Macierz par atrybutów

Na [Rysunek 1] przedstawiono wykres zależności między wybranymi cechami względem zmiennej docelowej GVB. Kolory oznaczają klasy zmiennej GVB: kluby oznaczone jako „0” (kolor niebieski) oraz „1” (kolor pomarańczowy). Na przekątnych wykresu znajdują się histogramy rozkładu wartości danej cechy dla obu klas, co pozwala zobaczyć różnice między nimi. Ten wykres pomaga ocenić, które cechy mogą mieć znaczący wpływ na klasyfikację i zrozumieć, jak różne cechy są ze sobą powiązane.



Rysunek 2 Współczynniki modelu SVM

Wykres [Rysunek2] przedstawia współczynniki przypisane różnym cechom przez model SVM z jądrem liniowym. Wartości dodatnie oznaczają, że wzrost danej cechy zwiększa prawdopodobieństwo przypisania klasy „1” (dobry klub), a wartości ujemne oznaczają, że wzrost cechy zwiększa prawdopodobieństwo przypisania klasy „0” (słaby klub). Wykres ten pozwala zrozumieć, które cechy mają największy wpływ na klasyfikację oraz w jaki sposób wpływają one na wynik końcowy modelu. Na przykład cechy takie jak „Passing” i „Finishing” mają najwyższe dodatnie współczynniki, co sugeruje, że mają znaczący wpływ na przypisanie klasy „1”.

A diagram of blue and orange dots

Description automatically generated

Rysunek 3 Granica decyzyjna zoptymalizowanego modelu SVM

Na wykresie [Rysunek 3] przedstawiono granicę decyzyjną zoptymalizowanego modelu SVM w przestrzeni dwóch głównych składowych uzyskanych z analizy PCA (Principal Component Analysis). Obserwacje są oznaczone kolorami zgodnie z ich przynależnością do klasy GVB: „0” (niebieskie kółka) oraz „1” (pomarańczowe krzyżyki). Granica decyzyjna oddziela obszary przypisane do klas „0” i „1” i pokazuje, jak model SVM rozdziela dane w tej uproszczonej przestrzeni. Wykres ten umożliwia wizualną ocenę skuteczności klasyfikacji i pokazuje, jak dobrze model radzi sobie z rozdzielaniem klas w danych.

**Drzewo decyzyjne**

A diagram of a structure

Description automatically generated

Rysunek 4 Drzewo decyzyjne

Powyższy [Rysunek 4] przedstawia drzewo decyzyjne dla założonego problemu. Każdy węzeł drzewa reprezentuje decyzję opartą na jednej z cech, a każda gałąź prowadzi do kolejnych decyzji, aż do liści, które reprezentują ostateczną klasyfikację.

Wyniki klasyfikacji dla drzewa decyzyjnego, zarówno z ustawieniami domyślnymi, jak i po optymalizacji, zostały przedstawione za pomocą macierzy pomyłek. Każda macierz pokazuje liczbę prawidłowych i błędnych klasyfikacji dla klas "Bad" i "Good".

1. **Macierz pomyłek dla drzewa decyzyjnego z ustawieniami domyślnymi**: w tej macierzy widzimy, że model z ustawieniami domyślnymi przewiduje klasy z dużą dokładnością. Klasa "Bad" została poprawnie zaklasyfikowana 101 razy, a klasa "Good" – 103 razy, przy niskiej liczbie błędnych klasyfikacji (8 dla "Bad" i 10 dla "Good").

A blue squares with white text

Description automatically generated

1. **Macierz pomyłek dla drzewa decyzyjnego po optymalizacji**: zoptymalizowany model wykazuje niemal identyczne wyniki jak model z ustawieniami domyślnymi, co wskazuje, że dodatkowe dostrajanie parametrów nie przyniosło znaczącej poprawy. Liczba błędnych klasyfikacji pozostała na podobnym poziomie (9 błędów dla "Bad" i 9 dla "Good").

A blue squares with white text

Description automatically generated

Wyniki uzyskane zarówno dla modelu drzewa decyzyjnego z ustawieniami domyślnymi, jak i zoptymalizowanego, są niemalże identyczne, z dokładnością wynoszącą 91.89% w obu przypadkach. Co więcej, raporty klasyfikacji dla obu modeli wskazują na bardzo podobne wartości precyzji, recall oraz F1-score, zarówno dla klasy "Bad" (0), jak i "Good" (1). Przyczyny identycznych wyników można upatrywać w tym, że struktura drzewa decyzyjnego, nawet przy domyślnych ustawieniach, była wystarczająca do skutecznego rozdzielenia obu klas. Zoptymalizowanie parametrów, takich jak głębokość drzewa czy minimalna liczba próbek w liściu, nie miało wpływu na wyniki, ponieważ drzewo decyzyjne bez tych zmian również świetnie radziło sobie z zadaniem. Dane były na tyle proste, że dodatkowa optymalizacja nie przyniosła znaczących korzyści.

**Las losowy**

Analiza wyników dla modelu lasu losowego wykazała wysoką dokładność zarówno w wersji domyślnej (92.79%), jak i zoptymalizowanej (93.69%). Oba modele efektywnie klasyfikowały dane, osiągając precyzję i czułość na bardzo wysokim poziomie. Poniżej przedstawiono macierze pomyłek obu modeli.

A blue squares with white text

Description automatically generatedA blue squares with white text

Description automatically generated

Wyniki zaoptymalizowanego modelu wzbogacono o wykres ważności cech [Rysunek 5], który przedstawia cechy najbardziej wpływające na decyzję modelu. Jasno można odczytać, że najistotniejszymi cechami przy klasyfikacji jest kontrola and piłką oraz krótkie podania. To niewątpliwie niezwykle ważne aspekty piłki nożnej.

A graph of blue lines

Description automatically generated with medium confidence

Rysunek 5

**Klasyfikator głosujący**

W celu oceny efektywności klasyfikatora głosującego (Voting Classifier) przeprowadzono eksperymenty z modelami o ustawieniach domyślnych oraz zoptymalizowanych. Wyniki analizowano, korzystając z metryk ewaluacyjnych, takich jak dokładność, precyzja, czułość (recall) oraz wynik F1, co pozwoliło na obiektywne porównanie obu wersji klasyfikatora.

Poniżej przedstawiono szczegółowe zmiany wprowadzone w drugim, tak zwanym zoptymalizowanym modelu:

1. **Random Forest (Las losowy)**:
   * **n\_estimators**: Zwiększono liczbę drzew z domyślnej (100) do maksymalnie 500, co pozwala modelowi lepiej uśredniać wyniki, zwiększając dokładność i stabilność klasyfikacji.
   * **max\_depth**: Przetestowano i ustawiono maksymalną głębokość drzew na 15, dzięki czemu model zyskał możliwość dokładniejszego uchwycenia złożoności danych, ale z kontrolą głębokości, co redukuje ryzyko nadmiernego dopasowania.
   * **min\_samples\_split** oraz **min\_samples\_leaf**: Zoptymalizowano minimalną liczbę próbek wymaganych do podziału (2, 5, 10) oraz liczbę próbek na liść (1, 2, 4), co zapewniło bardziej zrównoważone drzewo o większej ogólności.
2. **Gradient Boosting**:
   * **n\_estimators**: Zwiększono liczbę estymatorów do 300, aby poprawić stabilność modelu.
   * **learning\_rate**: Przetestowano mniejsze wartości współczynnika uczenia (0.01, 0.05, 0.1), co zwiększyło precyzję, umożliwiając modelowi lepsze dopasowanie do danych bez nadmiernego skoku na każdym kroku.
   * **max\_depth**: Ograniczenie maksymalnej głębokości do 7 pozwoliło modelowi uchwycić kluczowe wzorce bez nadmiernego dopasowania.
3. **Support Vector Machine (SVM)**:
   * **C** oraz **kernel**: Przetestowano wartości C (0.1, 1, 10) i różne funkcje jądra (linear, rbf) w celu lepszego dopasowania marginesów do danych. Umożliwiło to modelowi bardziej elastyczne wyznaczanie granic decyzyjnych.
   * **probability=True**: Włączono obliczanie prawdopodobieństw, co jest kluczowe w soft voting.
4. **K-Nearest Neighbors (kNN)**:
   * **n\_neighbors** oraz **weights**: Zoptymalizowano liczbę sąsiadów (3, 5, 7, 9) oraz sposób ich ważenia (uniform, distance). Dopasowanie tej liczby do danych pozwoliło na lepsze odwzorowanie lokalnych wzorców, co poprawiło precyzję klasyfikacji.
5. **Voting Classifier z głosowaniem miękkim wagami**:
   * Wersja zoptymalizowana korzysta z soft voting, gdzie każdy model dostarcza prawdopodobieństwa dla klasy zamiast samej klasyfikacji. Dodatkowo przypisano wagi dla najlepszych modeli (2 dla Random Forest i Gradient Boosting, 1 dla SVM i kNN, oraz 0.5 dla Logistic Regression), co zwiększyło ich wpływ na końcowy wynik.

W przypadku klasyfikatora domyślnego uzyskano dokładność na poziomie 93.7%. Zarówno dla klasy 0, jak i klasy 1, wskaźniki precyzji, czułości i F1-score wynosiły 0.94, co wskazuje na równą jakość klasyfikacji dla obu klas. Średnie wartości dla obu klas (macro avg oraz weighted avg) również wynosiły 0.94, co świadczy o stabilności modelu.

Po przeprowadzeniu optymalizacji klasyfikatora wynik dokładności wzrósł do 94.6%, a wynik F1 dla obu klas wyniósł 0.95. Dzięki temu zoptymalizowany model charakteryzuje się większą spójnością i dokładnością, co przekłada się na lepsze wyniki we wszystkich wskaźnikach. Precyzja dla klasy 1 wzrosła do 0.95, a czułość dla klasy 0 również osiągnęła 0.95, co podkreśla bardziej zrównoważoną i dokładną klasyfikację obu klas.

Optymalizacja poprawiła wyniki klasyfikacji, ale jednocześnie zwiększyła czas obliczeń do 33 sekund. Z tego względu należy rozważyć kompromis między wyższą dokładnością, a czasem przetwarzania, szczególnie w przypadku dużych zbiorów danych lub systemów wymagających szybkiej predykcji.

A graph of different colored squares

Description automatically generated with medium confidence

Wykres pokazuje, że Voting Classifier osiągnął najwyższą dokładność w porównaniu do innych modeli, co czyni go najbardziej efektywnym wyborem spośród testowanych metod. Pozostałe modele, takie jak Random Forest, Gradient Boosting, SVM, kNN oraz Logistic Regression, uzyskały podobne wyniki dokładności, mieszczące się w przedziale od 0.90 do 0.93.

Podsumowanie: należy wskazać model działający najlepiej i najgorzej. Co wpływa na polepszenie./pogorszenie predykcji modelu?