Grygoruk Piotr 260299

**Sztuczna Inteligencja i Inżynieria Wiedzy**

Prowadzący: Mgr inż. Michał Karol

1. **Eksploracja danych – podstawowe dane statystyczne i uwagi dotyczące cech i etykiet zbioru danych.**

Eksploracja zbioru danych "Glass" z UCI Machine Learning Repository:  
Ten zbiór danych zawiera informacje na temat różnych typów szkła, a naszym zadaniem będzie analiza statystyczna cech i etykiet.  
Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznieObraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Zbiór danych składa się z 214 rekordów i 11 kolumn. Kolumna "Type" zawiera etykiety, a pozostałe kolumny to cechy. Kolumna "Type" jest typu int64, a pozostałe kolumny są typu float64. Wszystkie rekordy zawierają pełne wartości (214 non-null entries).

1. **Przygotowanie danych – podział danych na zestaw uczący i walidacyjny. Zbadanie wpływu różnego typu przetworzenia danych na wyniki klasyfikacji.**
2. PCA jest techniką używaną do redukcji wymiarowości danych. Polega na przekształceniu zbioru zmiennych oryginalnych w nowy zestaw zmiennych, nazywanych składowymi głównymi. Te składowe główne są kombinacją liniową oryginalnych zmiennych, a każda kolejna składowa ma coraz mniejszą wariancję. Dzięki temu można zredukować ilość wymiarów danych, zachowując jednocześnie jak najwięcej informacji.
3. Normalizacja jest techniką używaną do skalowania danych do określonego zakresu. Najpopularniejszą metodą normalizacji jest skalowanie min-max. Polega ona na przekształceniu wartości każdej zmiennej w zakresie od 0 do 1, gdzie 0 odpowiada najmniejszej wartości danej zmiennej, a 1 największej wartości. Ta technika jest przydatna, gdy zależy nam na zachowaniu względnych proporcji między wartościami zmiennych.
4. Standaryzacja jest techniką używaną do przekształcenia danych w taki sposób, aby miały średnią równą 0 i odchylenie standardowe równe 1. Polega na odjęciu średniej wartości danej zmiennej od każdej obserwacji, a następnie podzieleniu przez odchylenie standardowe. Standaryzacja pozwala zachować względne proporcje między wartościami zmiennych, ale eliminuje ich różnice w skali. Jest szczególnie przydatna, gdy zmienne mają różne jednostki miary lub różnią się zakresem wartości.
5. Dyskretyzacja polega na podziale wartości ciągłej zmiennej na dyskretne przedziały lub kategorie. Może być używana, gdy chcemy przekształcić zmienne ciągłe na zmienne kategoryczne. Istnieje wiele technik dyskretyzacji, takich jak równa szerokość przedziałów, równa liczba obserwacji w przedziałach, czy wykorzystanie algorytmów grupowania, np. algorytmu k-średnich. Dyskretyzacja może pomóc w analizie danych, budowaniu modeli predykcyjnych lub wizualizacji.

**Podział danych na zestaw uczący i walidacyjny**  
Obraz zawierający tekst, Czcionka, zrzut ekranu, linia

Opis wygenerowany automatycznie

**Bez przetwarzania danych:**

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie

**Standaryzacja:**

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

**Dyskretyzacja:**

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie  
  
**PCA:**

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie  
**Selekcja cech:**

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

1. **Test klasyfikatorów.**Naiwny klasyfikator bayesowski to probabilistyczny model statystyczny, który opiera się na twierdzeniu Bayesa i przyjmuje założenie o niezależności cech (stąd "naiwny"). Model ten jest stosowany do klasyfikacji, a szczególnie dobrze radzi sobie z problemem klasyfikacji tekstu. Naiwny klasyfikator bayesowski opiera się na obliczeniu prawdopodobieństwa przynależności do każdej z klas na podstawie występowania poszczególnych cech w próbce.  
     
   Drzewo decyzyjne to graficzna struktura w formie drzewa, która reprezentuje zestaw reguł decyzyjnych do podejmowania decyzji na podstawie danych wejściowych. Składa się z węzłów, które reprezentują testy na atrybutach danych, oraz krawędzi, które łączą węzły i wskazują na możliwe wartości tych atrybutów. Na podstawie wartości atrybutów wchodzących do drzewa, od korzenia (początkowego węzła) przechodzimy przez kolejne węzły, aż dotrzemy do liścia, który zawiera przewidywaną etykietę lub wartość. Drzewo decyzyjne może być stosowane zarówno do klasyfikacji, jak i do regresji.

Hiperparametry drzewa decyzyjnego:

* Kryterium podziału: Określa miarę, na podstawie której dokonywane jest podziały w drzewie. Najpopularniejszymi kryteriami są entropia (informacyjna), nieczystość Gini i błąd klasyfikacji.
* Maksymalna głębokość: Określa maksymalną liczbę poziomów drzewa. Kontroluje złożoność modelu i może pomóc w uniknięciu przeuczenia.
* Min. liczba próbek w liściu: Określa minimalną liczbę próbek wymaganą w liściu drzewa. Jeśli liczba próbek w liściu jest mniejsza niż ta wartość, to podziały nie będą kontynuowane, co pomaga uniknąć przeuczenia.
* Min. przyrost ważoności: Określa minimalny wzrost ważoności (miary jakości podziału) wymagany do kontynuowania podziału w węźle.
* Maksymalna liczba cech: Określa maksymalną liczbę cech branych pod uwagę podczas poszukiwania najlepszego podziału w węźle.

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

1. **Ocena klasyfikacji oraz interpretacja wyników**

**Precision** to miara proporcji poprawnie przewidzianych pozytywnych instancji (true positives, TP) w stosunku do sumy poprawnie przewidzianych pozytywnych instancji i fałszywie przewidzianych pozytywnych instancji (false positives, FP). Wartość precision koncentruje się na dokładności predykcji pozytywnych instancji. Wyższa wartość precision oznacza mniejszą liczbę fałszywie pozytywnych przewidywań.

**Recall**, znany również jako true positive rate (TPR) lub sensitivity, to miara proporcji poprawnie przewidzianych pozytywnych instancji (TP) w stosunku do sumy poprawnie przewidzianych pozytywnych instancji i błędnie nieprzewidzianych pozytywnych instancji (false negatives, FN). Recall koncentruje się na zdolności modelu do wykrywania wszystkich pozytywnych instancji. Wyższa wartość recall oznacza mniejszą liczbę fałszywie negatywnych przewidywań.

**F1-score** to miara średniej harmonicznej między precision a recall. Przy wyborze modelu często chcemy znaleźć równowagę między precision a recall, dlatego F1-score jest przydatnym wskaźnikiem, łączącym obie miary. Wartość F1-score jest wysoka, gdy zarówno precision, jak i recall są wysokie. F1-score jest szczególnie przydatne, gdy dane są niezbalansowane, czyli jedna z klas dominuje nad drugą.

1. **Podsumowanie**

**Redukcja wymiarowości (PCA):**

* Precision: Najlepsze wyniki osiągnięto dla klasy 5 (100%), 6 (75%) i 7 (88%), natomiast klasy 1, 2 i 3 uzyskały wyniki nieco niższe.
* Recall: Klasy 6 (100%) i 7 (88%) osiągnęły najlepsze wyniki, natomiast klasy 1 i 2 uzyskały niższe wyniki.
* F1-score: Klasy 6 (86%) i 7 (88%) uzyskały najlepsze wyniki, natomiast klasy 1, 2 i 3 miały wyniki nieco niższe.

**Dyskretyzacja:**

* Precision: Najlepsze wyniki osiągnięto dla klasy 5 (100%), 6 (67%) i 7 (88%). Klasy 1 i 2 uzyskały wyniki średnie, a klasa 3 miała wynik równy zero.
* Recall: Klasy 5 (75%), 6 (67%) i 7 (88%) osiągnęły dobre wyniki, natomiast klasy 1 i 2 miały wyniki średnie.
* F1-score: Klasy 5 (86%), 6 (67%) i 7 (88%) osiągnęły dobre wyniki, natomiast klasy 1, 2 i 3 miały wyniki niskie lub równe zeru.

**Standaryzacja:**

* Precision: Najlepsze wyniki osiągnięto dla klasy 7 (89%), a klasy 1, 2, 3, 5 i 6 miały wyniki średnie.
* Recall: Klasy 3 (100%), 7 (100%) i 1 (91%) osiągnęły najlepsze wyniki, natomiast klasy 2, 5 i 6 uzyskały wyniki nieco niższe.
* F1-score: Klasy 7 (94%), 1 (80%) i 6 (80%) osiągnęły najlepsze wyniki, a klasy 2, 3 i 5 miały wyniki średnie.

**Bez przetwarzania danych:**

* Precision, recall i f1-score mają takie same wartości jak dla standaryzacji. Oznacza to, że bez przetwarzania danych (czyli korzystając z oryginalnych danych) osiągnięto podobne wyniki do tych po standaryzacji.

Podsumowując, przetwarzanie danych za pomocą redukcji wymiarowości (PCA) nieznacznie obniżyło wyniki klasyfikacji w porównaniu do innych metod. Dyskretyzacja wykazała mieszane rezultaty, ze znacznym obniżeniem wyników dla klasy 3. Standaryzacja i brak przetwarzania danych dawały podobne wyniki, z wyższymi wynikami dla klasy 3 w standaryzacji. Ostateczny wybór metody przetwarzania danych powinien być oparty na analizie konkretnego problemu i dostępnych danych.

**Drzewo decyzyjne:**

* Precision: Klasy 6 (100%) i 7 (100%) osiągnęły najlepsze wyniki, a klasy 1, 2, 3 i 5 uzyskały wyniki średnie.
* Recall: Klasy 1 (91%), 2 (71%), 6 (67%) i 7 (100%) osiągnęły dobre wyniki, natomiast klasy 3 i 5 uzyskały niższe wyniki.
* F1-score: Klasy 7 (100%) i 6 (80%) uzyskały najlepsze wyniki, a klasy 1, 2, 3 i 5 miały wyniki średnie.

**Naiwny klasyfikator Bayesa:**

* Precision: Klasy 6 (100%) i 7 (89%) osiągnęły najlepsze wyniki, a klasy 1, 2, 3 i 5 uzyskały wyniki średnie.
* Recall: Klasy 6 (100%) i 7 (100%) osiągnęły najlepsze wyniki, a klasy 1, 2, 3 i 5 uzyskały wyniki średnie.
* F1-score: Klasy 6 (100%) i 7 (94%) uzyskały najlepsze wyniki, a klasy 1, 2, 3 i 5 miały wyniki średnie.

Podsumowując, drzewo decyzyjne osiągnęło lepsze wyniki w przypadku precision i recall dla większości klas. Naiwny klasyfikator Bayesa osiągnął najwyższy wynik F1-score dla klasy 6, ale drzewo decyzyjne osiągnęło wyższe wyniki F1-score dla innych klas. Ostateczny wybór między tymi dwoma klasyfikatorami powinien być oparty na specyfice problemu, danych wejściowych i oczekiwaniach co do miar jakości.

1. **Źródła**

<https://pl.wikipedia.org/wiki/Naiwny_klasyfikator_bayesowski>

<https://chat.openai.com/>

Materiały z wykładu