|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| TOMASZ CHMIEL | |  | 2024 | PSI | | | | 1 | |
|  | **AUTOR** | **ROK AKAD.** |  | | | **SPEC.** |  | **GRUPA** |
| Wprowadzenie do uczenia nadzorowanego. Problem przewidywania klas. Klasyfikacja binarna i wieloklasowa. | | | | |  | | | 3 | |
| **TEMAT** | | | **NR SPRAWOZDANIA** | |
| 03.11.2024 | Laboratorium maszynowej analizy danych | | | | | Anna Czaderna-Lekka | | | |
| **DATA WYKONANIA** | **PRZEDMIOT** | | | | | **PROWADZĄCY** | | | |

# Cel

***Zadanie 1****: Celem zadania jest zbudowanie modelu regresji logistycznej do przewidywania, czy pasażerowie statku Titanic przeżyli katastrofę. Proces ten obejmuje analizę danych, w tym sprawdzenie kompletności i podstawowych statystyk, a także wizualizację rozkładu przeżycia pasażerów w różnych kategoriach, takich jak płeć, klasa biletu i wiek. Następnie następuje przetwarzanie danych, które obejmuje uzupełnianie brakujących wartości, usuwanie niepotrzebnych kolumn oraz przekształcanie danych kategorialnych na format numeryczny. Celem jest stworzenie modelu, który skutecznie przewiduje przeżycie pasażerów na podstawie dostępnych cech oraz ocena skuteczności modelu za pomocą macierzy błędów, raportu klasyfikacji oraz krzywych ROC i czułość/precyzja.*

***Zadanie 2****: Celem zadania jest przeprowadzenie klasyfikacji wieloklasowej za pomocą regresji logistycznej na zbiorze danych o irysach, który zawiera różne gatunki roślin. Proces ten obejmuje wczytanie danych, sprawdzenie ich kompletności oraz wykonanie wizualizacji, takich jak wykresy punktowe i macierze par, aby zrozumieć relacje między cechami. Następnie dane są standaryzowane, co ma na celu poprawienie wydajności modeli predykcyjnych. Model regresji logistycznej jest trenowany na danych z szerokości płatka jako cechy oraz gatunku jako etykiety. Celem jest zbadanie, jak skutecznie model może klasyfikować gatunki irysów oraz ocena wyników modelu przy użyciu raportu klasyfikacji, macierzy błędów oraz krzywych ROC i czułość/precyzja.*

# Materiały i metody

***Zadanie 1****: Do realizacji zadania wykorzystywany jest zbiór danych "train.csv" oraz "test.csv" z zestawu Titanic, który zawiera informacje o pasażerach, ich cechach demograficznych oraz statusie przeżycia. Zbiór danych składa się z 891 próbek, z których każda opisuje jednego pasażera.*

*Techniki analizy:*

* *Regresja logistyczna do przewidywania przeżycia pasażerów na podstawie cech takich jak płeć, klasa biletu, wiek oraz liczba członków rodziny na pokładzie.*
* *Ewaluacja modelu na podstawie macierzy błędów, raportu klasyfikacji oraz krzywych ROC i czułość/precyzja.*
* *Procedury przygotowania danych: Wczytanie danych z plików CSV, uzupełnienie brakujących wartości (np. wiek), usunięcie niepotrzebnych kolumn (np. 'Cabin'), oraz konwersja danych kategorialnych (np. płeć) na format numeryczny.*

***Zadanie 2****: Do realizacji zadania wykorzystywany jest zbiór danych Iris, który zawiera informacje o różnych gatunkach irysów oraz ich cechach morfologicznych. Zbiór danych składa się z 150 próbek, z których każda opisuje jeden egzemplarz rośliny.*

*Techniki analizy:*

* *Standaryzacja danych z użyciem StandardScaler, przekształcając wartości cech tak, aby miały zerową średnią i jednostkowe odchylenie standardowe.*
* *Model regresji logistycznej do przewidywania gatunków irysów na podstawie cech takich jak długość i szerokość płatka oraz kielicha.*
* *Ewaluacja modelu na podstawie raportu klasyfikacji, macierzy błędów oraz krzywych ROC i czułość/precyzja.*
* *Procedury przygotowania danych: Wczytanie danych z wbudowanego zbioru Iris w bibliotece scikit-learn, podział danych na zestawy treningowe i testowe, oraz standaryzacja danych cech przed treningiem modelu.*

# Wyniki i dyskusja

**Zadanie 1:**

A graph on a screen

Description automatically generated

A graph on a screen

Description automatically generated

A graph on a screen

Description automatically generated A graph with a bar

Description automatically generatedA screenshot of a computer

Description automatically generated

A graph with a line

Description automatically generated

***Zadanie 2:***

A screen shot of a graph

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A graph on a screen

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

# Podsumowanie

***Zadanie 1***

*Obserwacje: Analiza wpływu cech społeczno-gospodarczych na przeżywalność pasażerów Titanica za pomocą regresji logistycznej wykazała istotne zależności między różnymi cechami a szansą na przeżycie. Modele regularyzacji, takie jak regresja grzbietowa i Lasso, pozwoliły na lepsze dopasowanie do danych, a ich zastosowanie zredukowało ryzyko overfittingu. Wizualizacja współczynników modelu ujawniła, które cechy, takie jak płeć i klasa biletu, miały największy wpływ na przewidywane przeżycie.*

*Wnioski: Zastosowanie regularyzacji w modelach regresji logistycznej zwiększa stabilność wyników oraz ich interpretowalność. Analiza danych z wykorzystaniem regularyzacji podkreśla znaczenie wyboru odpowiednich technik przed przystąpieniem do predykcji, co może poprawić dokładność modeli w rzeczywistych zastosowaniach.*

***Zadanie 2***

*Obserwacje: Standaryzacja danych cech przyczyniła się do uzyskania lepszych wyników w modelach regresyjnych dotyczących klasyfikacji gatunków irysów, zwłaszcza w przypadku regresji wieloklasowej. Analiza porównawcza między przewidywanymi a rzeczywistymi gatunkami ujawniła, że zastosowanie standaryzacji wpłynęło na jakość predykcji w porównaniu do modeli na oryginalnych danych. Wykresy pokazują, że modele te były bardziej skuteczne w przewidywaniu gatunków dla zestawów testowych.*

*Wnioski: Standaryzacja jest istotnym krokiem w analizie danych, umożliwiającym lepsze dopasowanie modeli do złożonych relacji w danych. Zastosowanie standaryzacji w regresji logistycznej i jej wariantach przyczyniło się do poprawy dokładności predykcji oraz lepszej interpretacji wyników.*

# Bibliografia

Youtube: Precision, Recall, F1 score, True Positive|Deep Learning Tutorial 19 (Tensorflow2.0, Keras & Python)