AKADEMIA INNOWACYJNYCH ZASTOSOWAŃ TECHNOLOGII CYFROWYCH (AI TECH)

„Uczenie maszynowe" – laboratorium

Laboratorium 1 - Przetwarzanie Danych

Data aktualizacji: 03.03.2024

Cel ćwiczenia

Przetwarzanie danych dla celów budowy modeli uczenia maszynowego.

Główne zagadnienia:

- Transformacje danych
- Czyszczenie danych
- Problem wartości odstających (outlierów)
- Redukcja wymiarów

Wprowadzenie do zadania

Kontekst

Zadanie skupia się na analizie i przetwarzaniu zbioru danych „Polish companies bankruptcy" [2], który ze względu na swoją złożoność i specyfikę, wymaga szczególnego podejścia przed przystąpieniem do budowy modeli klasyfikacji lub klasteryzacji. Zbiór ten, charakteryzujący się wysokim poziomem trudności, dostarcza danych dotyczących fundamentalnej analizy finansowej polskich spółek, co stanowi podstawę do przewidywania ich upadłości.

Zawartość Zbioru

Zbiór składa się z 5 plików, z których każdy odpowiada danych z innego roku analizy. Dane opisane są za pomocą 64 atrybutów numerycznych, w tym między innymi dochód, zysk, kapitał własny, sprzedaż, co pozwala na głęboką analizę kondycji finansowej firm. Wyzwanie stanowią brakujące dane, wartości ujemne oraz różnorodność dziedzin atrybutów.

Cel

Celem zadania jest zbudowanie modelu klasyfikacj lub grupowania. Aby to osiągnąć, konieczne jest dokładne zapoznanie się z zawartością zbioru, jego przetworzenie i analiza pod kątem wykorzystania w modelowaniu.

Wytyczne

Analiza Zbioru: Przed przystąpieniem do modelowania konieczna jest dokładna analiza zbioru, co wymaga zapoznania się z wstępnym opisem dostępnym w źródle [2] oraz dokładniejszym opisem i przykładami analiz i klasyfikacji podanymi w pracy [3].

Obróbka Danych: Wyzwaniem jest poradzenie sobie z „niedogodnościami" danych, takimi jak braki w danych, wartości ujemne i różnorodność dziedzin atrybutów, co wymaga zastosowania zaawansowanych technik przetwarzania danych.

Przebieg ćwiczenia

1. Zapoznanie się z opisem zbioru danych, wybór odpowiedniego zakresu danych, eksploracyjna analiza danych.
2. Wyczyszczenie zbioru danych o brakujące wartości, wartości ujemne, różna skala wartości atrybutów (normalizacja, standaryzacja). Patrz literatura [4, 5].
3. Analiza wartości odstających z wykorzystaniem np. Z-Score lub jednego z algorytmów Outlier Detection. Patrz literatura [6].
4. Analiza zbiorów danych przy wykorzystaniu dwóch algorytmów redukcji wymiarów, np. PCA, t-SNE, UMAP. Patrz literatura [7-11].

5. Uruchomienie wybranego modelu klasyfikacji lub grupowania. **Analiza porównawcza wyników oraz decyzji podjętych w trakcie przygotowania danych do modelowania.**

Punktacja

Przy realizacji zadania student może otrzymać **max 10 punktów** wedle poniższej tabeli.

2	Zapoznanie się z opisem zbioru danych, wybór odpowiedniego zakresu danych, eksploracyjna analiza danych.
3	Wyczyszczenie zbioru danych o brakujące wartości, wartości ujemne, różna skala wartości atrybutów (normalizacja, standaryzacja).
1	Analiza wartości odstających z wykorzystaniem np. Z-Score lub jednego z algorytmów Outlier Detection.
2	Analiza zbiorów danych przy wykorzystaniu dwóch algorytmów redukcji wymiarów, np. PCA, t-SNE, UMAP (do wyboru).
2	Uruchomienie wybranego modelu klasyfikacji lub grupowania. Analiza porównawcza wyników oraz decyzji podjętych w trakcie przygotowania danych do modelowania.

Pytania dodatkowe

1. Na czym polega standaryzacja danych oraz normalizacja danych? Jakie są różnice pomiędzy tymi metodami? Jaki wpływ mają poszczególne transformacje danych na ostateczne wyniki modeli?
2. Na czym polega wybrana metoda detekcji obserwacji odstających? Jaki wpływ na wyniki ma wybrana metoda? Jak wybór metody obserwacji odstających wpływa na podział zbioru danych na zbiór treningowy oraz zbiór testowy?
3. Na czym polega wybrana metoda redukcji wymiarowości? Jakie są różnice pomiędzy wybranymi metodami?
4. Na czym polegają wybrane metody klasyfikacji lub grupowania danych?

Literatura

1. Materiały do wykładu
2. Zbiór danych
   https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data
3. Zieba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble Boosted Trees with Synthetic Features Generation in Application to Bankruptcy Prediction. Expert Systems with Applications
   https://www.ii.pwr.edu.pl/~tomczak/PDF/%5BMZSTJT%5D.pdf
4. https://scikit-learn.org/1.4/modules/preprocessing.html
5. https://scikit-learn.org/1.4/modules/impute.html
6. https://scikit-learn.org/1.4/modules/outlier_detection.html
7. https://scikit-learn.org/1.4/modules/generated/sklearn.decomposition.PCA.html
8. https://scikit-learn.org/1.4/modules/decomposition.html#pca
9. https://scikit-learn.org/1.4/modules/generated/sklearn.manifold.TSNE.html#sklearn.manifold.TSNE
10. https://scikit-learn.org/1.4/modules/manifold.html#t-sne
11. https://umap-learn.readthedocs.io/en/latest/