**Sprawozdanie**

**Weronika Domczewska**

**Projekt na zaliczenie laboratorium z przedmiotu Metody i Narzędzia Big Data**

**6.06.2023**

1. **Wprowadzenie**
2. Cel ćwiczenia

Celem ćwiczenia było rozwiązanie zadania klasyfikacji binarnej dla zbioru danych pacjentek diagnozowanych pod kątem PCOS oraz porównanie skuteczności modeli otrzymanych w rozwiązaniu. Modele różnią się między sobą zastosowanymi algorytmami klasyfikacji oraz rodzajem danych diagnostycznych, które były użyte do trenowania ich. Dane zostały podzielone na możliwe do zebrania w trakcie wywiadu lekarskiego (związane z wyglądem i trybem życia pacjentki) oraz zebrane przy badaniach specjalistycznych (morfologia, pomiary ciśnienia, saturacja, stężenie hormonów, opis USG brzucha). Modele były trenowane także na obu rodzajach danych jednocześnie.

1. Spodziewane wyniki  
   W wyniku powinnam otrzymać klasyfikatory określające, czy pacjentka o danych parametrach choruje na PCOS. Otrzymane modele powinny mieć skuteczność na poziomie 70 procent (na wykładzie ta wartość była podana jako akceptowalna dla klasyfikatorów związanych z danymi medycznymi). Spodziewam się, że najwyższą skuteczność osiągną klasyfikatory, które były trenowane na połączeniu danych pozyskanych w wywiadzie z pacjentem oraz wyników badań laboratoryjnych.
2. Cel dydaktyczny

Podczas wykonywania projektu powinnam nauczyć się, jak wykorzystać w praktyce wiadomości związane z przygotowaniem zbioru danych etykietowanych do modelowania z wykorzystaniem elementów podejścia EDA (ang. Exploratory Data Analysis), rozwiązywania zadania klasyfikacji oraz oceny jej jakości.

1. **Opis przeprowadzonych badań**
2. Narzędzia

W projekcie wykorzystałam zbiór danych, pochodzący z platformy kaggle.com, który był plikiem w formacie programu Excel (.xlsx). Do analizy danych oraz budowania modeli wykorzystałam język Python oraz biblioteki: pandas, matplotlib, numpy, seaborn oraz scikit-learn.

1. Sposób wykorzystania narzędzi
2. Wyniki
3. Analiza wyników
4. Wnioski