Competitie TIA 2023

Echipa: OANCEA BEATRICE SI FOTA ROBERT

1. Înțelegerea Seturilor de Date

Setul de Date de Antrenament (train\_df): Conține coloane pentru identificatorul unic (id), mai multe caracteristici numerice (feature\_0, feature\_1, etc.), și o coloană is\_anomaly care indică dacă un punct de date este o anomalie (1) sau nu (0).

Setul de Date de Test (test\_df): Similar cu setul de antrenament, dar fără coloana is\_anomaly.

1. Metodologia Aplicată

Am ales algoritmul Local Outlier Factor (LOF) pentru a identifica anomaliile în setul de date dupa urmatoarele criterii:

Natura Setului de Date: LOF este eficient în cazul seturilor de date cu multe dimensiuni, așa cum este cazul aici. Algoritmul lucrează bine cu caracteristici numerice continue.

Identificarea Anomaliilor: LOF este special conceput pentru a detecta outlierii într-un set de date, comparând densitatea locală a unui punct cu densitățile vecinilor săi. Acest lucru îl face ideal pentru a identifica punctele care sunt semnificativ diferite de celelalte.

Flexibilitate: LOF permite ajustarea sensibilității la anomaliile prin schimbarea numărului de vecini (k) și a pragului pentru scorurile LOF.

Implementarea algoritmului LOF pentru detectarea anomaliilor în acest context include următorii pași:

Pre-procesarea: Eliminarea coloanelor neesențiale (cum ar fi id și is\_anomaly din setul de antrenament) pentru a obține caracteristicile relevante.

Calculul Matricei de Distanțe: Determinarea distanței Euclidiene între fiecare pereche de puncte în seturile de date.

Găsirea Vecinilor Apropiați: Identificarea celor mai apropiați 'k' vecini pentru fiecare punct (unde 'k' este un hiperparametru).

Calculul Densității Locale: Calculul unei medii a distanțelor la vecinii apropiați pentru a estima densitatea locală a fiecărui punct.

Calculul Scorurilor LOF: Evaluarea gradului în care un punct este o anomalie prin compararea densității sale cu densitățile vecinilor săi.

Detectarea Anomaliilor: Folosind un prag pentru a clasifica punctele cu scoruri LOF ridicate ca anomalii.

Discuție și Concluzii

Eficacitatea Metodei: Algoritmul LOF pare să fi identificat o proporție semnificativă de date ca anomalii. Fără a avea un set de etichete de adevăr pentru setul de test, nu putem determina cu exactitate acuratețea acestor predicții. În contexte reale, este crucial să se stabilească praguri și parametri (cum ar fi 'k' pentru vecinii apropiați) în funcție de natura specifică a datelor și de cerințele problemelor.

A screen shot of a graph

Description automatically generated

A screenshot of a computer program

Description automatically generated

A computer screen shot of a program

Description automatically generated A screen shot of a computer code

Description automatically generated