Лабораторна робота №2. Метрики

Підгрупа №2

1. Відкрити та зчитати дані з наданого файлу. Файл містить п’ять стовпчиків:
    a. Фактичне значення цільової характеристики.
    b. Результат передбачення моделі №1 у вигляді ймовірності приналежності об’єкту до класу 0.
    c. Результат передбачення моделі №1 у вигляді ймовірності приналежності об’єкту до класу 1.
    d. Результат передбачення моделі №2 у вигляді ймовірності приналежності об’єкту до класу 0.
    e. Результат передбачення моделі №2 у вигляді ймовірності приналежності об’єкту до класу 1.
2. Визначити збалансованість набору даних. Вивести кількість об’єктів кожного класу.
3. Для зчитаного набору даних виконати наступні дії:
    a. Обчислити всі метрики (***Accuracy, Precision, Recall, F-Scores, Matthews Correlation Coefficient, Balanced Accuracy, Youden’s J statistics, Area Under Curve for Precision-Recall Curve, Area Under Curve for Receiver Operation Curve***) для кожної моделі при різних значеннях порогу класифікатора (крок зміни порогу обрати самостійно).
    b. Збудувати на одному графіку в одній координатній системі (*величина порогу; значення метрики*) графіки усіх обчислених метрик, відмітивши певним чином максимальне значення кожної з них.
    c. Збудувати в координатах (*значення оцінки класифікаторів; кількість об’єктів кожного класу*) окремі для кожного класу графіки кількості об’єктів та відмітити вертикальними лініями оптимальні пороги відсічення для кожної метрики.
    d. Збудувати для кожного класифікатору ***PR-криву*** та ***ROC-криву***, показавши графічно на них значення оптимального порогу.
4. Зробити висновки щодо якості моделей, визначити кращу модель.
5. Створити новий набір даних, прибравши з початкового набору $(50 + 5K)\%$ об’єктів класу 1, вибраних випадковим чином. Параметр K представляє собою залишок від ділення дня народження студента на дев’ять та має визначатися в програмі на основі дати народження студента, яка задана в програмі у вигляді текстової змінної формату ***‘DD-MM’***.
6. Вивести відсоток видалених об’єктів класу 1 та кількість елементів кожного класу після видалення.
7. Виконати дії п.3 для нового набору даних.
8. Визначити кращу модель.
9. Пояснити вплив незбалансованості набору даних на прийняте рішення.