BÀI KIỂM TRA – B1
Môn học: HỌC MÁY (7080510)

A. MÔ TẢ TẬP DỮ LIỆU:

Tập dữ liệu Data_Diabetes.csv bao gồm kết quả xét nghiệm bệnh tiểu đường của 723 bệnh nhân toàn bộ là nữ giới. Mỗi bản ghi tương ứng với một bệnh nhân, bao gồm 9 thuộc tính:

1. **Pregnancies:** Số lần mang thai
2. **Glucose:** Chỉ số Gluco
3. **BloodPressur:** Huyết áp (đơn vị: mm Hg)
4. **SkinThickness:** Mức độ dày da (đơn vị: mm)
5. **Insulin:** Chỉ số Insulin (đơn vị: mu U/ml)
6. **BMI:** Chỉ số BMI của cơ thể (được tính bằng = Cân nặng / chiều cao^2)
7. **DiabetesPedigreeFunction**: Chức năng phả hệ của bệnh tiểu đường
8. **Age**: Tuổi
9. **Outcome**: Thuộc tính cho biết bệnh nhân có mắc bệnh tiểu đường hay không?
   Không bị tiểu đường **(0)** – Bị tiểu đường **(1)**

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35.0	126.0	33.6	0.627	50	1
1	85	66	29.0	126.0	26.6	0.351	31	0
8	183	64	29.0	126.0	23.3	0.672	32	1
1	89	66	23.0	94.0	28.1	0.167	21	0
0	137	40	35.0	168.0	43.1	2.288	33	1
5	116	74	29.0	126.0	25.6	0.201	30	0
3	78	50	32.0	88.0	31.0	0.248	26	1
2	197	70	45.0	543.0	30.5	0.158	53	1
4	110	92	29.0	126.0	37.6	0.191	30	0

Tập dữ liệu đã được làm sạch để sử dụng cho mô hình học máy

B.YÊU CẦU:

1. Đọc file dữ liệu, quan sát dữ liệu và các đặc trưng thống kê của tập dữ liệu

2. Đánh giá mức độ cân bằng và thực hiện cân bằng dữ liệu *(Tham khảo phương pháp cân bằng dữ liệu theo link)*

3. Phân tách các biến Độc lập (X) - Phụ thuộc (Y) tương ứng

4. Chia tập dữ liệu thành 2 phần Train - Test

5. Sinh viên sử dụng một trong số các thuật toán đã được học: **kNN, Decision Tree, Naïve bayes** để huấn luyện và đánh giá độ chính xác của mô hình; Tùy chỉnh các tham số của mô hình để đạt được độ chính xác cao nhất.

Hiển thị các kết quả sau của model:

a) *Độ chính xác của model (accuracy) trên tập Train – Test*
b) *Tổng số mẫu dự đoán đúng - sai trên tập Test*
c) *Ma trận confusion matrix trên tập Test*
d) *Tìm hiểu và tính toán thông số **F1-Score, Recall** của model*

6. Sử dụng model xây dựng được, dự đoán Bệnh nhân với các thông số như sau có bị mắc bệnh tiểu đường hay không?

- Pregnancies: 0
- Glucose: 128
- BloodPressure: 78
- SkinThickness: 30
- Insulin: 138
- BMI: 30
- DiabetesPedigreeFunction: 1.18
- Age: 23

C.HƯỚNG DẪN LÀM BÀI VÀ NỘP BÀI:

1. Sinh viên làm trên jupyter notebook, đặt tên file theo quy định sau: Nhóm_Mã SV_Họ tên sinh viên_BaiB1, ví dụ: ***03_1921050330_DoHieuAnh_BaiB1.ipynb***
2. Sử dụng Cell markdown, Cell code để thực hiện theo các yêu cầu ở trên sao cho logic, dễ hiểu, dễ theo dõi.
3. Sinh viên nộp bài theo link sau, chọn mục Bài Tập B1: https://foBrms.gle/qjAkgkomNHwhWj9z6

D.LƯU Ý:

Sinh viên không sử dụng, sao chép bài của nhau, các bài giống nhau 0 điểm