Họ và tên thí sinh :..... ; MSSV: .....

Sử dụng tệp dữ liệu *Dethi_data.xlsx* được cung cấp. Sinh viên thực hiện trả lời các yêu cầu bên dưới trong hệ thống Learning Management System (LMS) của đại học Công Nghiệp TP.HCM (IUH).

- Với câu trả lời điền đáp án thì **chỉ điền số** cho kết quả và **làm tròn 2 chữ số**. Ví dụ: **120.00** hoặc **57.89**
- Với câu chọn đáp án thì chọn câu trả lời đúng nhất, **chọn câu trả lời sai sẽ bị trừ điểm.**

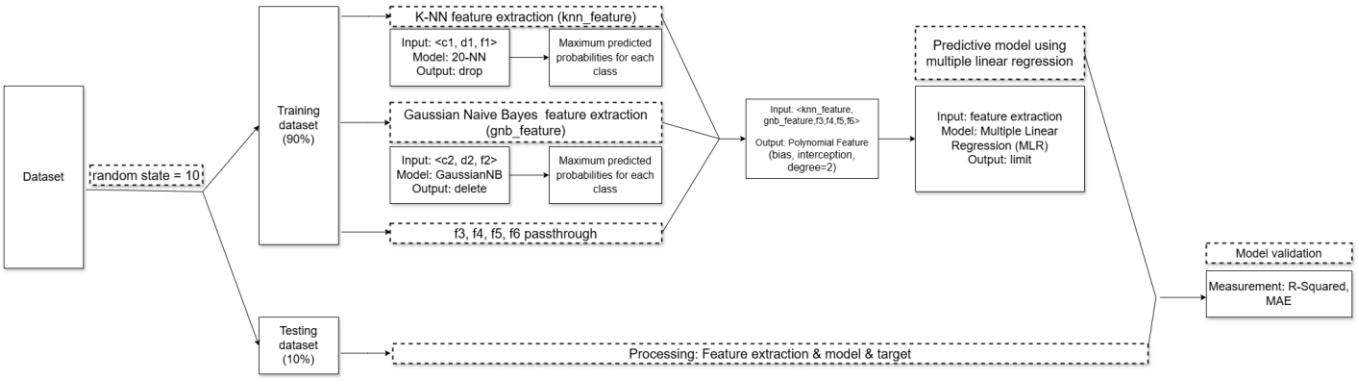
Giai đoạn 1: Tại công ty X, nhóm A tiến hành xây dựng mô hình dự báo giá trị **limit** thông qua các đặc trưng <c1, d1, f1, c2, d2, f2, f3, f4, f5, f6> bằng mô hình Multiple Linear Regression (MLR). Tập dữ liệu ban đầu sẽ được chia theo hệ số ngẫu nhiên là 10, thành 2 phần tương ứng: tập dữ liệu huấn luyện (df_train) chiếm 90% và tập dữ liệu kiểm thử (df_test) chiếm 10%. Nhóm A sử dụng các đại lượng R-Squared và Mean Absolute Error (MAE) để đánh giá mô hình MLR.

(*) Sử dụng mô tả 'giai đoạn 1' để giải quyết các câu hỏi từ **1, 2, 3, 4, 5**.

Giai đoạn 2: Sau một thời gian, để cải tiến mô hình MLR, nhóm A đề xuất phát triển mô hình Hybrid Model (HM) để dự báo giá trị **limit** thông qua các đặc trưng biến đổi Polynomial. Biết rằng việc phân chia tập tệp dữ liệu cũng giống như 'giai đoạn 1' và để đánh giá mô hình cũng sử dụng các đại lượng R-Squared, MAE. Thiết kế của Hybrid Model được mô tả như sơ đồ bên dưới:

- knn_feature: được trích xuất dựa trên giá trị xác suất lớn nhất trên nhãn phân lớp (maximum predicted probabilities for each class) theo mô hình K-Nearest Neighbors (KNN).
- gnb_feature: được trích xuất dựa trên giá trị xác suất lớn nhất trên nhãn phân lớp (maximum predicted probabilities for each class) theo mô hình Gaussian Naive Bayes (GNB).
- Các đặc trưng <f3,f4,f5,f6> vẫn giữ nguyên giá trị (không biến đổi).
- Các đặc trưng bao gồm <knn_feature,gnb_feature,f3,f4,f5,f6> được đưa sang biến đổi đặc trưng Polynomial với degree=2 , có bias và interception.

(*) Sử dụng mô tả 'giai đoạn 2' để giải quyết các câu hỏi từ **6, 7, 8, 9, 10.**



(*) Lưu ý khi xây dựng các mô hình trong cả hai giai đoạn: Với các hyper-parameter không có yêu cầu giá trị thì xem như là sử dụng giá trị mặc định.

Câu 1: (*1.0 điểm, CLO 01*): Hãy mô tả tổng quan về dữ liệu.

Câu 2: (*1.0 điểm, CLO 01*): Sau khi chia tách dữ liệu làm 2 phần là tập dữ liệu huấn luyện (df_train) và tập dữ liệu kiểm thử (df_test). Hãy cho biết số lượng mẫu của df_train, df_test.

Câu 3: (*1.5 điểm, CLO 02*): Sau khi chia tách dữ liệu làm 2 phần là tập dữ liệu huấn luyện (df_train) và tập dữ liệu kiểm thử (df_test). Hãy cho biết danh sách những giá trị order nào tương ứng thuộc về df_train, df_test.

Câu 4: (*1.5 điểm, CLO 03*): Hãy cho biết giá trị R-Squared và MAE đạt được trên tập dữ liệu kiểm thử theo mô hình MLR.

Câu 5: (*1.5 điểm, CLO 04*): Hãy cho biết với các giá trị đầu vào lần lượt như sau: c1 = -1.0, d1 = 9640.0, f1 = 15134.0, c2 = -2.0, d2 = 7404.0, f2 = 0.0, f3 = 7002.0, f4 = 8167.0, f5 = 3996.0, f6 = 2000.0. Thì giá trị dự báo của **limit** bởi mô hình MLR sẽ là bao nhiêu?

Câu 6: (*1.0 điểm, CLO 04*): Hãy cho biết kết quả của đại lượng độ chính xác (accuracy) trong mô hình trích xuất đặc trưng KNN, với k = 20 trên tập dữ liệu huấn luyện.

Câu 7: (*0.5 điểm, CLO 03*): Hãy cho biết giá trị trung bình của đặc trưng knn_feature được trích xuất từ tập dữ liệu huấn luyện.

Câu 8: (*0.5 điểm, CLO 03*): Hãy cho biết giá trị nhỏ nhất, trung bình, phương sai, lớn nhất trong ma trận biểu diễn các đặc trưng của tập dữ liệu huấn luyện trước khi đi qua biến đổi đặc trưng Polynomial. Lưu ý: không tính các giá trị thiếu.

Câu 9: (*0.5 điểm, CLO 05*): Hãy cho biết dựa vào giá trị R-Squared và MAE đạt được trên tập dữ liệu kiểm thử thì mô hình Hybrid và MLR cái nào tốt hơn?

Câu 10: (*1.0 điểm, CLO 04*): Hãy cho biết với các giá trị đầu vào lần lượt như sau: c1 = -1.0, d1 = 9640.0, f1 = 15134.0, c2 = -2.0, d2 = 7404.0, f2 = 0.0, f3 = 7002.0, f4 = 8167.0, f5 = 3996.0, f6 = 2000.0. Thì giá trị dự báo của **limit** bởi mô hình HM sẽ là bao nhiêu?

---------- Hết ----------

Lưu ý: - Giáo viên ra đề và trưởng bộ môn ký duyệt vào mặt sau của đề.

- *Được sử dụng sử dụng tài liệu off-line để tham khảo,*

  -Đề thi KHÔNG ĐƯỢC sử dụng các công cụ AI hỗ trợ học tập.

  -Không được sử dụng điện thoại.

  -Không được sử dụng bất kì phương thức liên lạc trao đổi.

  -Sinh viên nộp lại đề thi sau khi thi xong.

  -Cán bộ coi thi không giải thích gì thêm.