**ĐỒ ÁN MÔN HỌC BASIC MACHINE LEARNING**

**Deadline: 20/11/2021**

**ĐỒ ÁN 1**

BÀI TOÁN DỰ ĐOÁN CHI PHÍ Y TẾ CÁ NHÂN

HỌC VIÊN THỰC HIỆN:

Bài toán dự đoán chi phí y tế cá nhân là bài toán xác định các yếu tố ảnh hưởng đến chi phí cá nhân và sau đó dự đoán chi phí y tế mà một người sẽ chi trả.

Dữ liệu thu thập bao gồm các thuộc tính sau:

Tuổi (Tuoi), Giới tính(Gioi\_tinh), Chỉ số khối cơ thể (BMI), Số con (So\_con) , Tình trạng hút thuốc(Hut\_thuoc), Nơi sinh sống (Noi\_song), Chi phí y tế cá nhân (Chi\_phi)

Dữ liệu thu thập được chia thành 2 tập tin Medical\_Train.csv và Medical\_Test.csv dùng để huấn luyện và kiểm tra mô hình dự đoán.

Yêu cầu:

1. Đọc và phân tích các dữ liệu và lý giải các yếu tố sẽ ảnh hưởng đến chi phí y tế cá nhân.
2. Tiền xử lý dữ liệu
3. Xây dựng mô hình dự đoán bằng 2 thuật toán Linear Regression và Polynomial Regression (học viên tự tìm hiểu thêm), lập trình bằng Python.
4. Kiểm tra và đánh giá kết quả dự đoán các mô hình thông qua tính giá trị các độ đo RMSE(Root Mean Square Error), MAE(Mean Absolute Error), SSE (Sum of Squared Estimate of Errors), R2 (R-Squared)
5. So sánh kết quả dự đoán của 2 mô hình.
6. Làm slide trình bày bao gồm:
   1. Giới thiệu bài toán và dữ liệu, tiền xử lý dữ liệu.
   2. Cơ sở toán học của mô hình Linear Regession, Polynomial Regression, công thức toán học tính các độ đo
   3. Giải thích Source code chương trình.
   4. Kết quả dự đoán và so sánh giữa 2 mô hình
   5. Kết luận

**ĐỒ ÁN 2**

BÀI TOÁN DỰ ĐOÁN GIÁ CĂN HỘ

HỌC VIÊN THỰC HIỆN: Minh Hoàng

Bài toán dự đoán giá căn hộ là bài toán xác định các yếu tố ảnh hưởng đến giá của một căn hộ và sau đó xây dựng mô hình dự đoán giá căn hộ.

Dữ liệu thu thập từ một website bất động sản được lưu trong file Can\_ho.csv bao gồm các thuộc tính sau:

* DiaChi: địa chỉ của căn hộ
* TinhTrangBDS: đã bàn giao/chưa bàn giao
* DienTich: diện tích sử dụng, đơn vị: m2.
* PhongNgu: số phòng ngủ.
* TenPhanKhu: tên block của căn hộ.
* SoTang: căn hộ ở tầng thứ mấy.
* PhongTam: số lượng nhà tắm/ nhà vệ sinh
* Loại: chung cư / nhà xã hội.
* GiayTo: tình trạng giấy tờ pháp lý của căn hộ
* MaCanHo: mã số căn hộ
* TinhTrangNoiThat: đã có hoặc chưa có nội thất
* HuongCuaChinh: hướng cửa chính
* HuongBanCong: hướng ban công
* DacDiem: Đặc điểm căn hộ
* Gia: giá bán

Yêu cầu:

1. Đọc và phân tích các dữ liệu và lý giải các yếu tố sẽ ảnh hưởng đến giá nhà
2. Tiền xử lý dữ liệu
3. Xây dựng mô hình dự đoán bằng thuật toán Linear Regression và lập trình bằng Python.
4. Kiểm tra và đánh giá các mô hình thông qua các độ đo RMSE, MAE, SSE, R2
5. Đánh giá kết quả dự đoán của mô hình.
6. Làm slide trình bày bao gồm:
   1. Giới thiệu bài toán và dữ liệu
   2. Giải thích quá trình tiền xử lý dữ liệu
   3. Cơ sở toán học của mô hình Linear Regession và công thức toán học tính các độ đo
   4. Giải thích Source code chương trình.
   5. Kết quả dự đoán và đánh giá thông qua các độ đo
   6. Kết luận

**ĐỒ ÁN 3**

BÀI TOÁN DỰ BÁO THỜI TIẾT

HỌC VIÊN THỰC HIỆN:

Bài toán dự báo thời tiết yêu cầu dự báo thời tiết sau 3 giờ dựa trên thông tin thời tiết đã có. Dữ liệu thu thập được lưu trong file Thoi\_tiet.csv

Dữ liệu thu thập bao gồm các thuộc tính sau:

* time: thời gian (hh:mm)
* month: tháng
* temperature: nhiệt độ môi trường (độ C)
* feelslike: nhiệt độ cảm nhận được (độ C)
* wind: tốc độ gió (km/h)
* direction: hướng gió
* gust: tốc độ gió tối đa (km/h)
* cloud: độ che phủ của mây (%)
* humidity: độ ẩm không khí (%)
* precipitation: lượng mưa (mm)
* pressure: áp suất không khí bm)
* weather: tình hình thời tiết hiện tại

Yêu cầu:

1. Đọc và phân tích, lý giải các yếu tố sẽ ảnh hưởng đến tình hình thời tiết.
2. Tiền xử lý dữ liệu
3. Xây dựng mô hình dự đoán bằng thuật toán Support Vector Machine và Decision Tree
4. So sánh độ chính xác dự đoán của 2 mô hình.
5. Làm slide trình bày bao gồm:
   1. Giới thiệu bài toán, dữ liệu, tiền xử lý dữ liệu
   2. Cơ sở lý toán học của mô hình Support Vector Machine và Decision Tree
   3. Giải thích Source code chương trình.
   4. So sánh kết quả dự đoán của 2 mô hình
   5. Kết luận

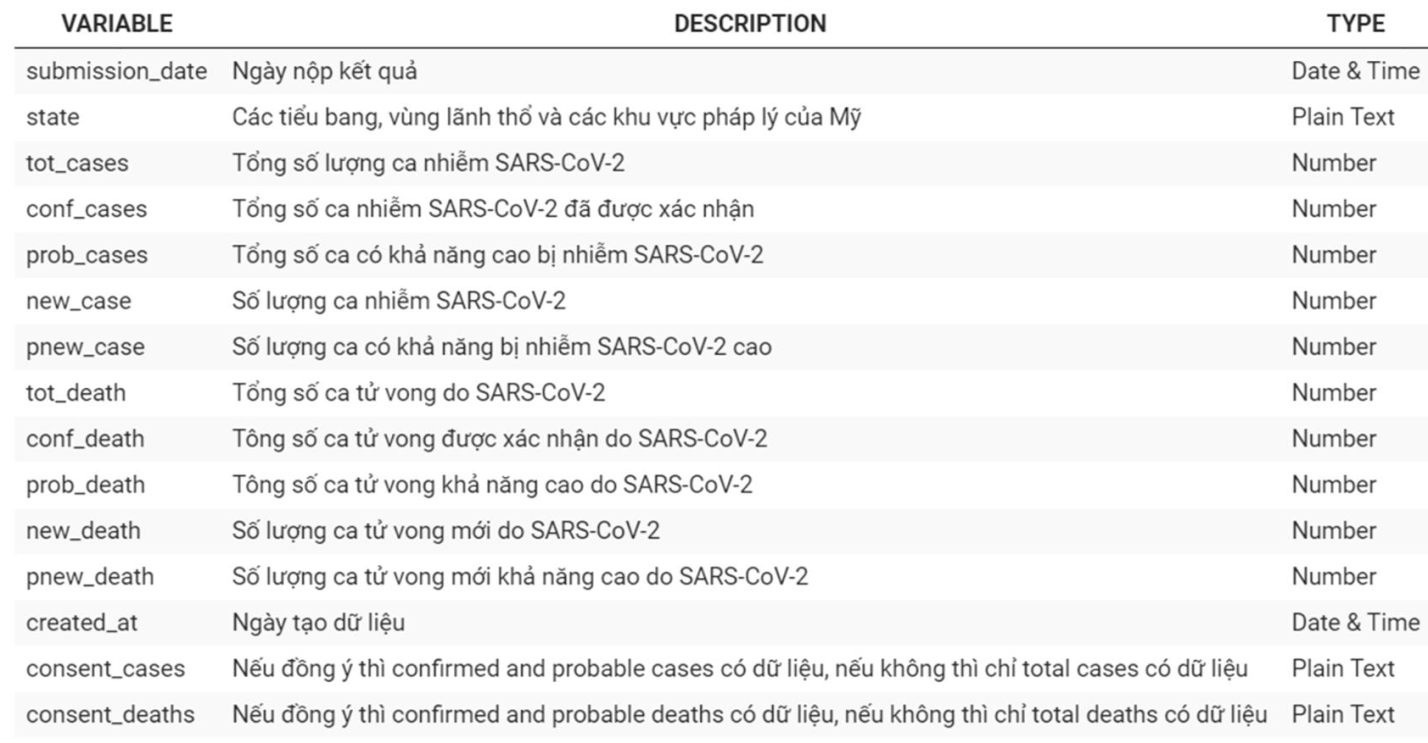
**ĐỒ ÁN 4**

BÀI TOÁN DỰ ĐOÁN SỐ CA NHIỄM COVID

HỌC VIÊN THỰC HIỆN: Quốc Thắng

Trong đồ án này, chúng ta dự đoán số ca nhiễm trong tương lai gần nhằm giúp cơ quan có thẩm quyền tham khảo để đưa ra những giải pháp phù hợp, kịp thời.

Dataset là file covid\_data.csv chứa dữ liệu mô tả liên quan đến đại dịch COVID-19.



Yêu cầu:

1. Đọc và phân tích các yếu tố ảnh hưởng đến số ca nhiễm mới.
2. Tiền xử lý dữ liệu
3. Xây dựng mô hình dự đoán bằng thuật toán Linear Regression và lập trình bằng Python.
4. Kiểm tra và đánh giá mô hình thông qua các độ đo RMSE, MAE, SSE, R2
5. Đánh giá kết quả dự đoán của mô hình.
6. Làm slide trình bày bao gồm:
   1. Giới thiệu bài toán và dữ liệu
   2. Giải thích quá trình tiền xử lý dữ liệu
   3. Cơ sở toán học của mô hình Linear Regession và công thức toán học tính các độ đo
   4. Giải thích Source code chương trình.
   5. Kết quả dự đoán và đánh giá thông qua các độ đo
   6. Kết luận