BÁO CÁO CÔNG VIÊC THỰC HIÊN

PHÂN TÍCH CHUỖI THỜI GIAN PODCAST BẰNG KALMAN

# Họ và tên: Nguyễn Anh Đức

# Lớp: 64TTNT1

# MSV: 2251262589

# Link GitHub:

<https://github.com/nguyenanhducIT2/THUC-HANH-01---KALMAN.git>

**1. Giới thiệu đề tài**

Trong bài thực hành này, em triển khai ba mô hình Kalman Filter để phân tích chuỗi thời gian dữ liệu nghe Podcast, sử dụng dữ liệu theo bài thầy gửi. Mục tiêu chính là làm mượt tín hiệu theo ngày phát hành và dự báo xu hướng lượng người nghe theo thời gian. Theo mã số sinh viên của em, đề bài tương ứng là đề số 4 (Thursday).

**2. Dữ liệu sử dụng**

Dữ liệu sử dụng là file train.csv từ bài toán phân tích Podcast, trong đó có các cột như Listening\_Time\_minutes, Likes, Publication\_Day. Nhóm chọn lọc dữ liệu theo ngày trong tuần, ví dụ: thứ Năm (Thursday).

**3. Các mô hình Kalman Filter đã triển khai**

* **Kalman Filter cơ bản**: Làm mượt chuỗi thời gian bằng tham số mặc định.
* **Kalman Filter dùng EM (Expectation-Maximization)**: Tối ưu tham số với thuật toán EM để cải thiện độ chính xác của mô hình.
* **Kalman Filter đa biến**: Phân tích kết hợp thêm đặc trưng Likes để tăng độ chính xác và có cái nhìn tổng quan hơn về sự biến động của dữ liệu.

**4. Tiền xử lý dữ liệu**

Trước khi triển khai các mô hình Kalman Filter, em đã thực hiện một số bước tiền xử lý như:

* **Nội suy các giá trị thiếu**: Sử dụng phương pháp nội suy tuyến tính để điền giá trị thiếu.
* **Mã hóa các biến phân loại**: Các cột phân loại như Podcast\_Name, Episode\_Title, Genre, Publication\_Time, Episode\_Sentiment được mã hóa bằng LabelEncoder.
* **Chuẩn hóa dữ liệu**: Dữ liệu đầu vào được chuẩn hóa bằng StandardScaler để đảm bảo các giá trị đầu vào có cùng quy mô.

**5. Triển khai và đánh giá các mô hình Kalman Filter**

* **Kalman Filter 1D**: Mô hình này giúp làm mượt chuỗi thời gian của Listening\_Time\_minutes bằng cách sử dụng Kalman Filter một chiều. Đánh giá mô hình dựa trên các chỉ số như MSE, RMSE, MAE, và R²:
  + MSE: 410.1238
  + RMSE: 20.2515
  + MAE: 16.6921
  + R²: 0.4702
* **Kalman Regression**: Mô hình Kalman Filter với hồi quy giúp kết hợp các đặc trưng đầu vào và dự đoán chính xác hơn. Đánh giá mô hình:
  + MSE: 410.1238
  + RMSE: 20.2515
  + MAE: 16.6921
  + R²: 0.4702
* **Kalman Multivariate**: Mô hình đa biến sử dụng Likes kết hợp với các đặc trưng khác để cải thiện độ chính xác. Đánh giá mô hình:
  + MSE: 2787.4065
  + RMSE: 52.7959
  + MAE: 44.8697
  + R²: -2.6008

**6. Trực quan hóa kết quả**

* Biểu đồ trực quan hóa kết quả của các mô hình Kalman Filter cho thấy:
  + **Kalman Filter 1D** làm mượt tốt nhưng chưa phản ánh đầy đủ xu hướng thay đổi.
  + **Kalman Regression** và **Kalman Multivariate** mang lại kết quả ổn định và phù hợp hơn với biến động thực tế.

**7. Liên kết GitHub**

Mã nguồn GitHub: <https://github.com/nguyenanhducIT2/THUC-HANH-01---KALMAN.git>

**Kết luận**

Mô hình Kalman Filter đã giúp làm mượt chuỗi thời gian và dự báo xu hướng lượng người nghe podcast. Sau này em định triển khai trên Streamlit và upload models lên Hugiface