Table of Contents

[I . Phát hiện bất thường của dữ liệu 2](#_Toc170941911)

[1. Upload file dữ liệu dạng csv bao gồm Timestamp, tên thiết bị và giá trị dữ liệu của thiết bị theo thời gian 2](#_Toc170941912)

[2. Đọc dữ liệu từ thiết bị thông qua các API để lấy về dữ liệu ( query theo các điều kiện và thời gian) 2](#_Toc170941913)

[ Lựa chọn các dữ liệu đặc trưng ( Feature selection) 2](#_Toc170941914)

[ Tiền xử lý dữ liệu ( Preprocessing data) 2](#_Toc170941915)

[ Phân tích dữ liệu (Analysis) 3](#_Toc170941916)

[ Sau khi các bước xử lý trên thì tới bước áp dụng thuật toán gồm (IsolationForest, LocalOutlierFactor,EllipticEnvelop) + kết hợp với EMA Period, Contamination để phát hiện các bất thường của dữ liệu. 3](#_Toc170941917)

[ Sau chạy phân tích dữ liệu thì sẽ hiện các biểu đồ thể hiện các dữ liệu 3](#_Toc170941918)

[II. Đối với dự đoán dữ liệu 5](#_Toc170941919)

[1. Upload file dữ liệu dạng csv bao gồm Timestamp, tên thiết bị và giá trị dữ liệu của thiết bị theo thời gian. 5](#_Toc170941920)

[2. Đọc dữ liệu từ thiết bị thông qua các API để lấy về dữ liệu ( query theo các điều kiện và thời gian). 5](#_Toc170941921)

[ Lựa chọn các dữ liệu đặc trưng ( Feature selection) 5](#_Toc170941922)

[ Tiền xử lý dữ liệu ( Preprocessing data) 6](#_Toc170941923)

[ Tiền xử lý dữ liệu ( Preprocessing data): Gồm drop missing values, forward fill, backward fill, mean fill). 6](#_Toc170941924)

[ Ở đây có thêm 1 bước chạy để kiểm tra dữ liệu dựa vào Target column để ta có thể chọn ra những đặc trưng có tỷ lệ % theo mong muốn. 6](#_Toc170941925)

[ Phân tích dữ liệu (Analysis) 7](#_Toc170941926)

Các hình ảnh

[Hình 1 Lựa chọn đặc trưng 3](#_Toc170942000)

[Hình 2 Tiền xử lý dữ liệu (Preprocessing) 4](#_Toc170942001)

[Hình 3 So sánh dữ liệu giữa các cột 4](#_Toc170942002)

[Hình 4 Phát hiện dữ liệu bất thường 5](#_Toc170942003)

[Hình 5 Phân bố dữ liệu và bảng giá trị các cột 5](#_Toc170942004)

[Hình 6 Tổng quan và kết quả các giá trị cao nhất và thấp nhất của mô hình 6](#_Toc170942005)

[Hình 7 Lựa chọn đặc trưng 7](#_Toc170942006)

[Hình 8 Tiền xử lý và lựa chọn các đặc trưng 8](#_Toc170942007)

[Hình 9 Các thuật toán để dự báo dữ liệu 8](#_Toc170942008)

[Hình 10 Cột mục tiêu và dự báo 9](#_Toc170942009)

[Hình 11 Phân bố dữ liệu và giá trị cột mục tiêu, dự báo 9](#_Toc170942010)

[Hình 12 Bảng tương quan dữ liệu, kết quả của mô hình 10](#_Toc170942011)

[Hình 13 Tổng quan sau khi chạy dự báo 10](#_Toc170942012)

Nhờ thầy tư vấn để viết khoá luận tốt nghiệp Sau đại học với nội dung về đề tài của em là:

Với bộ dữ liệu IoT từ các thiết bị của công ty theo thời gian, dữ liệu sẽ được training dựa vào các thuật toán để đưa ra các kết quả dự báo và phát hiện bất thường của dữ liệu

# I . Phát hiện bất thường của dữ liệu

gồm 2 phần chính

1. Upload file dữ liệu dạng csv bao gồm Timestamp, tên thiết bị và giá trị dữ liệu của thiết bị theo thời gian
2. Đọc dữ liệu từ thiết bị thông qua các API để lấy về dữ liệu ( query theo các điều kiện và thời gian)

từ 2 bước trên sau khi trích xuất dữ liệu từ thiết bị sử dụng các bước để phát hiện bất thường gồm:

* Lựa chọn các dữ liệu đặc trưng ( Feature selection)

A screenshot of a computer

Description automatically generated

Hình 1 Lựa chọn đặc trưng

* Tiền xử lý dữ liệu ( Preprocessing data)
  + Tiền xử lý dữ liệu ( Preprocessing data): Gồm drop missing values, forward fill, backward fill, mean fill),
  + Select Key column ( so sánh dữ liệu key column với giá trị cao hơn, hoặc thấp hơn)

A screenshot of a computer

Description automatically generated

Hình 2 Tiền xử lý dữ liệu (Preprocessing)

* Phân tích dữ liệu (Analysis)
  + Sau khi các bước xử lý trên thì tới bước áp dụng thuật toán gồm (IsolationForest, LocalOutlierFactor,EllipticEnvelop) + kết hợp với EMA Period, Contamination để phát hiện các bất thường của dữ liệu.
  + Sau chạy phân tích dữ liệu thì sẽ hiện các biểu đồ thể hiện các dữ liệu

A screenshot of a computer

Description automatically generated

Hình 3 So sánh dữ liệu giữa các cột

A screenshot of a graph

Description automatically generated

Hình 4 Phát hiện dữ liệu bất thường

A screenshot of a graph

Description automatically generated

Hình 5 Phân bố dữ liệu và bảng giá trị các cột

A screenshot of a computer

Description automatically generated

Hình 6 Tổng quan và kết quả các giá trị cao nhất và thấp nhất của mô hình

# II. Đối với dự đoán dữ liệu

* 1. Upload file dữ liệu dạng csv bao gồm Timestamp, tên thiết bị và giá trị dữ liệu của thiết bị theo thời gian.
  2. Đọc dữ liệu từ thiết bị thông qua các API để lấy về dữ liệu ( query theo các điều kiện và thời gian).

từ 2 bước trích xuất dữ liệu từ thiết bị sử dụng các bước để phát hiện bất thường gồm

* Lựa chọn các dữ liệu đặc trưng ( Feature selection)

A screenshot of a computer

Description automatically generated

Hình 7 Lựa chọn đặc trưng

* Tiền xử lý dữ liệu ( Preprocessing data)
  + Tiền xử lý dữ liệu ( Preprocessing data): Gồm drop missing values, forward fill, backward fill, mean fill).
  + Ở đây có thêm 1 bước chạy để kiểm tra dữ liệu dựa vào Target column để ta có thể chọn ra những đặc trưng có tỷ lệ % theo mong muốn.

A screenshot of a computer

Description automatically generated

Hình 8 Tiền xử lý và lựa chọn các đặc trưng

* Phân tích dữ liệu (Analysis)

Bước chọn thuật toán để phân tích dữ liệu thì có khác nhiều thuật toán được áp dụng  
A screenshot of a computer

Description automatically generated

Hình 9 Các thuật toán để dự báo dữ liệu

* Sau khi chạy thuật toán ta sẽ được các đồ thi như sau

A screenshot of a graph

Description automatically generated

Hình 10 Cột mục tiêu và dự báo

A screenshot of a computer

Description automatically generated

Hình 11 Phân bố dữ liệu và giá trị cột mục tiêu, dự báo

A screenshot of a computer

Description automatically generated

Hình 12 Bảng tương quan dữ liệu, kết quả của mô hình

A screenshot of a computer

Description automatically generated

Hình 13 Tổng quan sau khi chạy dự báo