**MỤC LỤC**

[PHẦN 1: GIỚI THIỆU VÀ PHẦN TÍCH BÀI TOÁN 2](#_Toc88403460)

[*1. Giới Thiệu: 2*](#_Toc88403461)

[*2. Bài Toán 2*](#_Toc88403462)

[PHẦN 2: THỰC HIỆN BÀI TOÁN 2](#_Toc88403463)

[*1. DATASET 2*](#_Toc88403464)

[*2. Mô Tả: 2*](#_Toc88403465)

[*3. Thực hiện 2*](#_Toc88403466)

[PHẦN 3: TỔNG KẾT 6](#_Toc88403467)

[*1. Thuận Lợi 6*](#_Toc88403468)

[*2. Thách Thức 6*](#_Toc88403469)

[*3. Thành viên 6*](#_Toc88403470)

**BÁO CÁO ĐỒ ÁN CUỐI KÌ**

# PHẦN 1: GIỚI THIỆU VÀ PHẦN TÍCH BÀI TOÁN

## Giới Thiệu:

Bài toán xác định giới tính qua tên có thể ứng dụng ở nhiều lĩnh vực khác nhau trong cuộc sống như: giúp nhận biết được giới tính thông việc biết được chữ ký, họ tên; giúp xác định ngôi xưng hô trong văn bản khi dịch thuật tự động, ...

## Bài Toán

- Bài Toán: Phân loại nhị phân

- Yêu cầu: Xác định giới tính dựa trên họ tên đầy đủ.

- Input: Họ tên đầy đủ của một người.

- Output: Xác định được giới tính (Nam/Nữ) của người đó: 1(Nam), 0(Nữ).

PHẦN 2: THỰC HIỆN BÀI TOÁN

## DATASET

- Có tộng cộng 26854 cái tên được chia làm ba tập:

+Tập train : 18796 tên (~70%)

+Tập dev : 2687 tên (~10%)

+Tập Test : 5371 tên (~20%)

## Mô Tả:

- Bằng cách chi nhỏ tên ra thành các từ nhỏ hơn, rồi xác định qua tỷ lệ xác xuất các từ xuất hiện trong tên và cho ra kết quả.

## Thực hiện

- Bọn em thử nghiệm qua ba mô hình SVM, Decision Tree Classifier Logistic Regression:

**SVM:** Mất 2.643404245376587 để huấn luyện mô hình

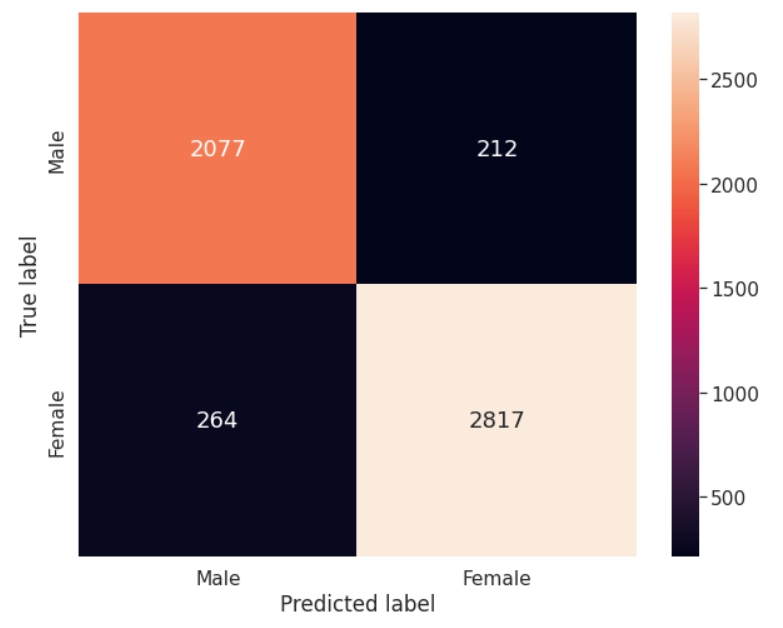
SVM Accuracy Score -> 91.13594040968343

F1-score: 0.909643575486138

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.89 | 0.91 | 0.90 | 2289 |
| 1 | 0.93 | 0.91 | 0.92 | 3081 |
| accuracy |  |  | 0.91 | 5370 |
| macro avg | 0.91 | 0.91 | 0.91 | 5370 |
| weighted avg | 0.91 | 0.91 | 0.91 | 5370 |

[[2077 212]

[ 264 2817]]



→Chậm nhất nhưng chính xác nhất

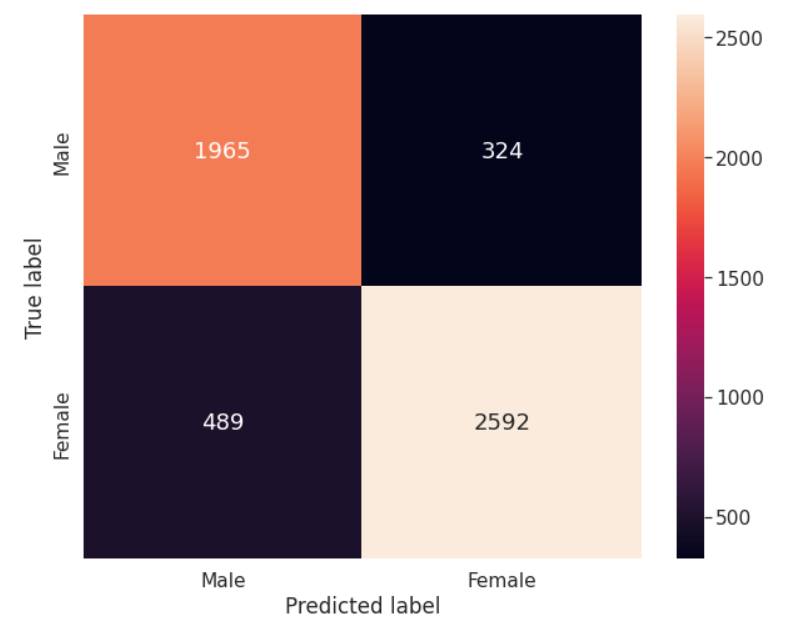
**Decision Tree Classifier:** Mất 0.06429314613342285 để huấn luyện mô hình

DT Accuracy Score -> 84.86033519553072

F1-score: 0.8465108582121548

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.80 | 0.86 | 0.83 | 2289 |
| 1 | 0.89 | 0.84 | 0.86 | 3081 |
| accuracy |  |  | 0.85 | 5370 |
| macro avg | 0.84 | 0.85 | 0.85 | 5370 |
| weighted avg | 0.85 | 0.85 | 0.85 | 5370 |

[[1965 324]

[ 489 2592]]

→ Nhanh nhất nhưng tỉ lệ chính xác thấp nhất

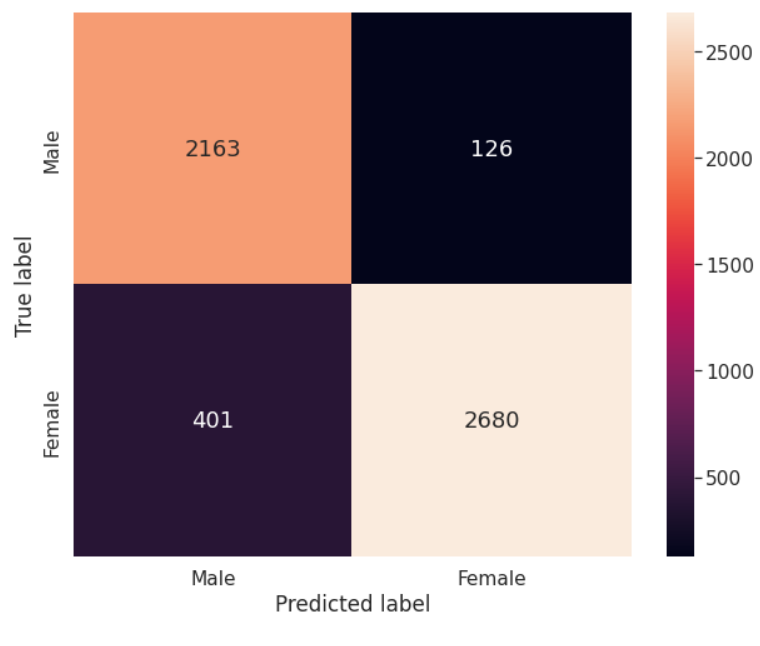
**Logistic Regression:** Mất 0.22625350952148438 để huấn luyện mô hình

LR Accuracy Score -> 90.18621973929235

F1-score: 0.9009440485556488

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.84 | 0.94 | 0.89 | 2289 |
| 1 | 0.96 | 0.87 | 0.91 | 3081 |
| accuracy |  |  | 0.90 | 5370 |
| macro avg | 0.90 | 0.91 | 0.90 | 5370 |
| weighted avg | 0.91 | 0.90 | 0.90 | 5370 |

[[2163 126]

[ 401 2680]]  
→Vừa phải nhất

# PHẦN 3: TỔNG KẾT

## Thuận Lợi

## Thách Thức

## Thành viên

Lê Việt Thắng

Phạm Ngô Hoàng Sinh