**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

Logo

Description automatically generated**NĂM HỌC:** 2024 *–* 2025

BÁO CÁO ĐỒ ÁN CUỐI KỲ

**Phân tích dữ liệu bảo toàn tính riêng tư**

**Giảng viên hướng dẫn:** Lê Phúc Lữ

                            Nguyễn Ngọc Toàn

**Sinh viên thực hiện:** Trần Anh Tú 22120400

Hoàng Ngọc Tuệ 22120407

**Mục Lục**

[I. Thông tin nhóm 3](#_Toc201445160)

[II. Phân công công việc và đánh giá. 3](#_Toc201445161)

[1. Phân công 3](#_Toc201445162)

[2. Đánh giá 3](#_Toc201445163)

[III. Tổng quan và đánh giá về các yêu cầu đã hoàn thành 3](#_Toc201445164)

[1. Tổng quan 3](#_Toc201445165)

[2. Đánh giá bài làm 3](#_Toc201445166)

[IV. Yêu cầu về tập dữ liệu 4](#_Toc201445167)

[V. Yêu cầu về tiền xử lý dữ liệu 4](#_Toc201445168)

[1. Lựa chọn các thuộc tính để phục vụ hồi quy 4](#_Toc201445169)

[2. Xử lý các cột dữ liệu 4](#_Toc201445170)

[VI. Mồ hình tuyến tính và ứng dụng Differential Privacy 5](#_Toc201445171)

[1. Mồ hình tuyến tính 5](#_Toc201445172)

[2. Ứng dụng Differential Privacy (Áp dụng ở Server-level) 5](#_Toc201445173)

[VII. Privacy Preserving cho mô hình Federated Learning. 6](#_Toc201445174)

[VIII. Tài liệu tham khảo 7](#_Toc201445175)

1. Thông tin nhóm

|  |  |
| --- | --- |
| MSSV | Họ và tên |
| 22120400 | Trần Anh Tú |
| 22120407 | Hoàng Ngọc Tuệ |

1. Phân công công việc và đánh giá.
   1. Phân công

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Công việc | Thành viên phụ trách | |
| Tú | Tuệ |
| 1 | Đọc tập dữ liệu | X |  |
| 2 | Tiền xử lý dữ liệu |  | X |
| 3 | Triển khai mô hình hồi quy |  | X |
| 4 | Triển khai Differential Privacy cho mô hình | X |  |

* 1. Đánh giá

|  |  |  |
| --- | --- | --- |
| **Thành viên** | **Đánh giá mức độ hoành thành công việc (%)** | **Phần trăm đóng góp cho đồ án** |
| Trần Anh Tú | 100% | 50% |
| Hoàng Ngọc Tuệ | 100% | 50% |

1. Tổng quan và đánh giá về các yêu cầu đã hoàn thành
   1. Tổng quan

|  |  |  |
| --- | --- | --- |
| **STT** | **Công việc** | **Mức độ hoàn thành** |
| 1 | Đọc dữ liệu | 100% |
| 2 | Tiền xử lý dữ liệu | 100% |
| 3 | Triển khai mô hình hồi quy | 100% |
| 4 | Áp dụng Differential Privacy cho mô hình | 100% |

* 1. Đánh giá bài làm
* Đối với yêu cầu xử lý dữ liệu: dữ liệu đã được đọc vào và xử lý đúng theo yêu cầu được để ra.
* Đối với mô hình hồi quy: Mô hình hồi quy được triển khai đơn giản đáp ứng đủ 4 công thức hồi quy cũng như dễ dàng áp dụng Differential Privacy.
* Đối với việc áp dụng Differential Privacy: Áp dụng thành công và thấy rõ được sự khác biết về kết quả trước và sau khi áp dụng Differential Privacy vào Model.
* Tổng kết: Mô hình hồi quy được triển khai trong bài làm đáp ứng đầy đủ các yêu cầu của đề bài. Mô hình tuy đơn giản nhưng cũng đủ để thấy rõ được cách triển khai Differential Privacy cũng như thấy được mối quan hệ giữa sự chính xác và tính riêng tư.

1. Yêu cầu về tập dữ liệu

* Yêu cầu đọc dữ liệu: sử dụng pandas để đọc dữ liệu trên file csv.
* Yêu cầu in ra thông tin của file csv: sử dụng hàm len() và thuộc tính columns để in ra số dòng và chi tiết các thuộc tính.
* Yêu cầu đọc dữ liệu và in ra 5 dòng đâu: sử dụng head để in ra 5 dòng đầu tiên.

1. Yêu cầu về tiền xử lý dữ liệu
   1. Lựa chọn các thuộc tính để phục vụ hồi quy

* Cần lựa chọn các thuộc tính có một liên hệ rõ ràng đến với Price trong Data để có thể đạt được mô hình hồi quy có tính chính xác cao.
* Các cột dữ liệu được chọn bao gồm: 'Year', 'Kilometer', 'Fuel Tank Capacity', 'Engine', 'Max Power', 'Owner'.
* Các cột trên được lựa chọn do có mối liên hệ rất rõ ràng với thuộc tính Price được thể hiện qua đồ thị:

A group of graphs on a white background

AI-generated content may be incorrect.

* Đối với các giá trị không được lựa chọn: 'Fuel Type', 'Transmission', 'Location', 'Color', 'Seller Type', 'Drivetrain', 'Length', 'Width', 'Height', 'Seating Capacity'. Đa phần đều không thể hiện rõ ràng mối liên hệ với Price. Đối với 'Max Torque' do Max Torque có mối liên hệ đối với Max Power nên chỉ cần chọn một là đủ. Còn với các thuộc tính loại thì rất khó để thấy được sự liên hệ với Price chỉ có Owner là thấy rõ ràng nhất nên được chọn.
  1. Xử lý các cột dữ liệu
* Các cột dữ liệu được lấy ra từ Data và tiến hành xử lý theo đúng yêu cầu:
  + Đối dữ liệu thiếu: Thêm vào dữ liệu mean() của thuộc tính đó.
  + Đối với dữ liệu vừa chữ vừa số: Cắt chuỗi đề lấy được số.
  + Đối với dữ liệu loại: Đánh số cho loại đó.

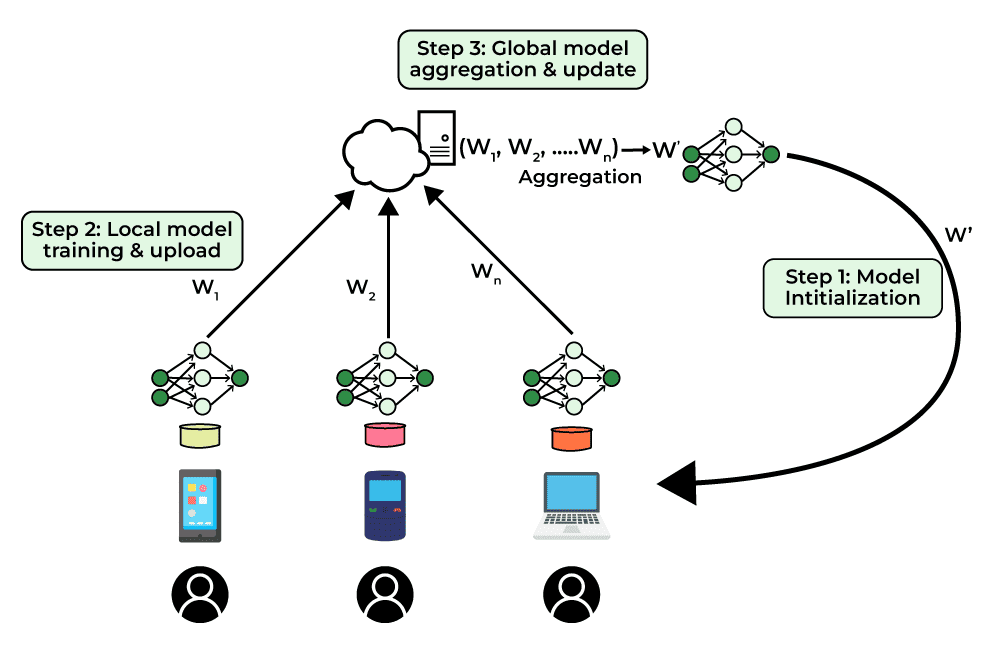
1. Mồ hình tuyến tính và ứng dụng Differential Privacy
   1. Mô hình tuyến tính

* Mồ hình hồi quy được triển khai trên các thuộc tính đã được xử lý ở trên.
* Để tính toán đường hồi quy nhóm em sử dụng **Normal Equation** do dữ liệu được cho không quá lơn và việc sử dụng Normal Equation sẽ đảm bảo tính chính xác cao hơn.
* Các công thức được sử dụng trong mô hình hồi quy:
  + Công thức 1**:** . Công thức này được sử dụng để có một mô hình hồi quy cơ bản nhằm phục vụ so sánh và cải tiến các đường hồi quy sau này.
  + Công thức 2: . Dựa vào đồ thị thấy rằng Fuel Tank Capacity, Engine, Max Power có mối liên hệ phi tuyến tính với Price nên công thức trên được sử dụng bình phương các thuộc tính trên để làm giảm thiểu tính phi tuyến tính.
  + Công thức 3: . Công thức này nhằm để đơn giản hóa thuật toán hồi quy chỉ giữ lại các tham số có ảnh hưởng lớn đến phương trình hồi quy.
  + Công thức 4: . Công thức trên thay các cột Engine, Max Power, Fuel Tank Capacity bằng các thuộc tính thể hiện sự tương tác giữa chúng qua đó thấy được giá trị của một chiếc xe không phụ thuộc vào một thuộc tính đơn lẻ.
* Đối với hàm để đọc một file csv từ một đường dẫn và in ra độ chính xác của mô hình evaluate\_on\_csv hàm sẽ nhận vào 2 tham số file path đường dẫn tới file csv, và formula type số nguyên biểu thị công thức nào sẽ được sử dụng.
  1. Ứng dụng Differential Privacy (Áp dụng ở Server-level)
* Về ý tưởng: Nhóm sử dụng Laplace mechanism để sinh noise và sau đó thêm noise đó vào câu trả lời.
* Sau khi tính sensitivity và chọn epsilon thì hai tham số trên sẽ được truyền vào phương thức apply\_laplace\_mechanism của model trong phương thức đó biến noise sẽ được sinh ra dựa trên laplace distribution sau đó dược cộng vào hai hệ số coefficients và intercept. Do dữ liệu của hồi quy được dự đoán dựa trên hai hệ số coefficients và intercept nên model trên sẽ thỏa DP.
* Đánh giá về hiệu xuất:
  + Đối với Epsilon = 0.1: Tính riêng tư được đảm bảo nhưng độ chính xác khá thấp.
  + Đối với Epsilon = 5.0: Tính riêng tư không qua được đảm bảo nhưng độ chính xác cao.
  + Đối với Epsilon = 50: Gần như không có sự khác biệt với không áp dụng DP.

A screenshot of a computer program

AI-generated content may be incorrect.

1. Privacy Preserving cho mô hình Federated Learning.

* Mô hình Federated Learning: là một phương pháp học máy trong đó quá trình huấn luyện mô hình được thực hiện trên nhiều thiết bị hoặc máy khách (clients) thay vì tập trung toàn bộ dữ liệu tại một máy chủ trung tâm. Dữ liệu không cần được chia sẻ ra ngoài, giúp tăng cường quyền riêng tư và bảo mật. 
* Trong source code, mô hình Federated Learning đã được triển khai một cách đơn giản nhưng đầy đủ các ý tưởng chính của mô hình.
* Mô hình Federated Learning đã được triển khai theo các bước:
  + Khởi tạo một Server và sử dụng dữ liệu mẩu (giả sử dữ liệu này là public) để train model của Server.
  + Sau đó các Client tham giá vào mô hình sẽ tiến hành lấy các thông tin khởi tạo cần thiết (trong mô hình bài làm này là trung bình và độ lêch chuẩn) sau đó tiến hành sử dụng dữ liệu local để huấn luyện mô hình Client.
  + Sau khi hoàn tất huấn luyện Client tiến hành gửi đến Server các số liệu gồm có hai hệ số là coefficient và intercept.
  + Server sau đó sẽ tiến hành tổng hợp và nâng cấp mô hình dựa trên các hệ số nhân được ở Client (trong bài làm này Server sẽ lấy trung bình các tham số)

1. Tài liệu tham khảo

[1]

“Python Machine Learning Multiple Regression,” [*www.w3schools.com*. https://www.w3schools.com/python/python\_ml\_multiple\_regression.asp](file:///C:\Users\trana\Downloads\www.w3schools.com.%20https:\www.w3schools.com\python\python_ml_multiple_regression.asp)

[2]

GeeksforGeeks, “Normal Equation in Linear Regression,” *GeeksforGeeks*, Sep. 27, 2018. [https://www.geeksforgeeks.org/machine-learning/ml-normal-equation-in-linear-regression/](https://www.geeksforgeeks.org/machine-learning/ml-normal-equation-in-linear-regression/%20) (accessed Jun. 14, 2025).

[3]

GeeksforGeeks, “What is Federated Learning?,” *GeeksforGeeks*, Jan. 20, 2021. [https://www.geeksforgeeks.org/machine-learning/collaborative-learning-federated-learning/](https://www.geeksforgeeks.org/machine-learning/collaborative-learning-federated-learning/%20) (accessed Jun. 21, 2025).