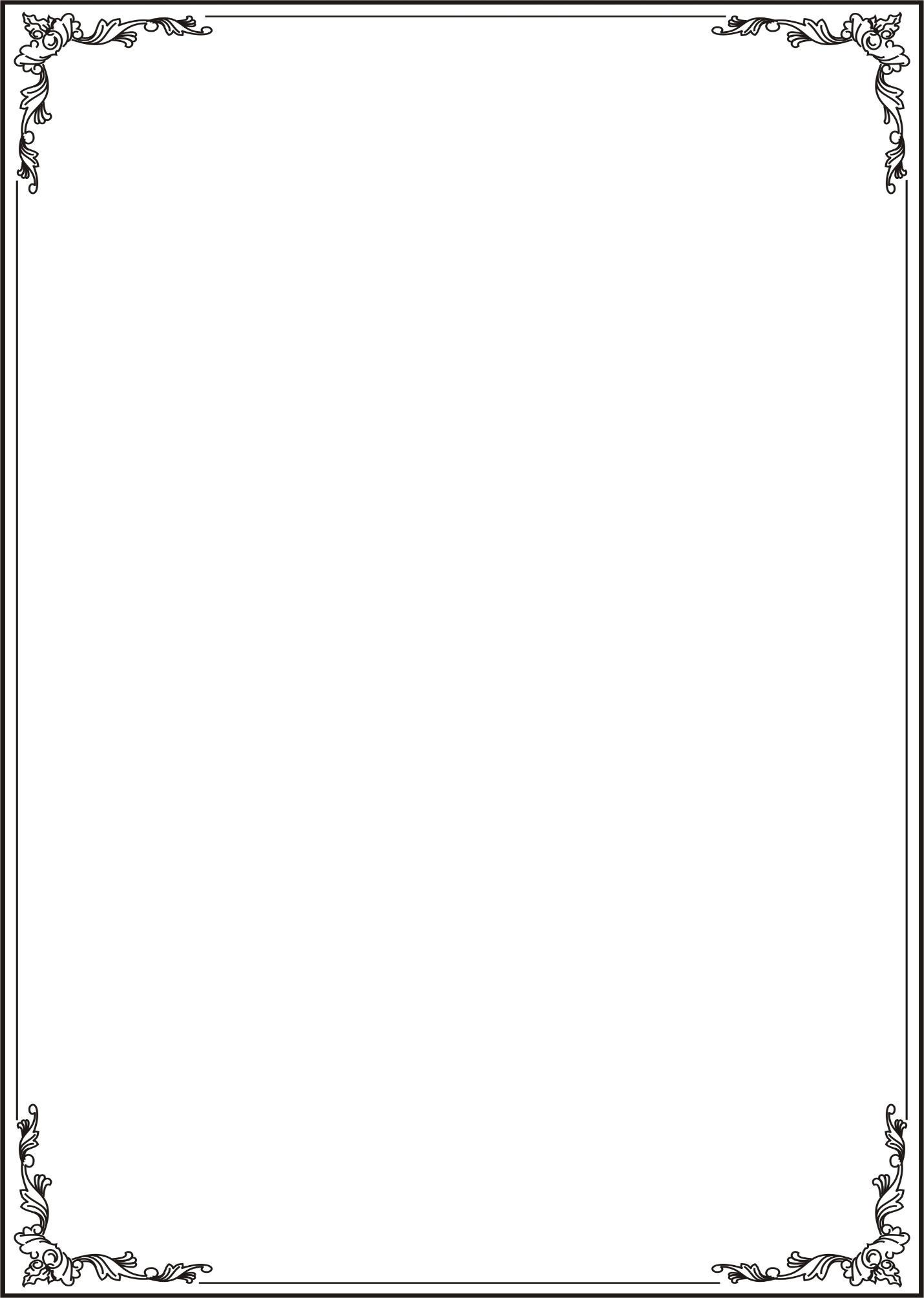
**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

🙜🙞🕮🙜🙞



**ĐỒ ÁN CUỐI KÌ**

**DATA MINING**

**Đề tài: Phân tích và dự đoán giá điện thoại**

**GVHD :** TS. Quách Đình Hoàng

**Nhóm SVTH :**

Lý Quốc Dũng 19133015

Đoàn Trần Đăng Khoa 19133028

Nguyễn Quốc Bảo 19133002

Võ Hoàng Khả Diệu 19133014

**Tp. Hồ Chí Minh, tháng 06 năm 2022**

**Mục lục**

1. **Tóm tắt..……………………………………………………………….3**
2. **Giới thiệu.……………………………………………………………..3**
3. **Dữ liệu..………………………………………………………………..4**
4. **Trực quan hóa dữ liệu………………………………………………..5**
5. **Mô hình hóa dữ liệu…………………………………………………..9**
6. **Thực nghiệm, Kết quả và thảo luận…………………………………10**
7. **Kết luận………………………………………………………………13**
8. **Phụ lục………………………………………………………………..13**
9. **Đóng góp……………………………………………………………..13**
10. **Tham khảo…………………………………………………………..14**

**1. Tóm tắt (abstract)**

Do quá trình công nghiệp hóa, hiện đại hóa ngày nay, nhu cầu liên lạc và giải trí của con người ngày càng tăng, vì thế việc sở hữu một chiếc điện thoại không chỉ để phục vụ cho đời sống con người mà còn là thể hiện cá tính, đam mê của mỗi người. Vì vậy nhóm sẽ phân tích dự đoán giá điện thoại trong tương lai là hữu ích và cần thiết.

Các câu hỏi được đặt ra:

+ So sánh các mô hình khác nhau

+ Lợi ích của việc xác định các tính năng

+ Các tính năng ảnh hưởng đến độ chính xác mô hình

+ Tìm ra các tính năng có mối tương quan với giá

Nhóm sử dụng các phương pháp:

+Phân tích EDA.

+Sử dụng hai mô hình KNN, Linear Regression, Decision Tree, Support Vector Machine để phân tích và dự đoán giá điện thoại trong tương lai.

Kết quả thu được : Sau khi nhóm tiến hành phân tích dữ liệu vẽ các biểu đồ các giá trị để tìm ra điểm tương quan và dùng các mô hình Machine Learning để dự đoán thì rút ra kết luận giá điện thoại tăng dần dựa theo các biến ram, battery power, pixel height càng lớn

**2. Giới thiệu (introduction)**

Do giá điện thoại bị ảnh hưởng bởi các tính năng của điện thoại, trong số đó có yếu tố ram, battery power, pixel height sẽ ảnh hưởng đến giá điện thoại được bán ra, vì điện thoại có ram, battery power, pixel height càng lớn thì giá điện thoại càng cao. Vì vậy nhóm đặt ra các câu hỏi, tìm hiểu mối tương quan giữa giá điện thoại và các yếu tố trên để có thể phân tích một cách chính xác, trực quan hơn về giá điện thoại trong tương lai.

Input: Tệp dữ liệu CSV về xe: tương tác với các cột ram, battery power, pixel height, price\_range

Output:

-Phân tích EDA:

+Phân tích price\_range với ram để tìm hiểu được mối tương quan giá điện thoại với ram

+Phân tích price\_range với battery power để tìm hiểu được mối tương quan giá điện thoại với battery power

+Phân tích price\_range với pixel height để tìm hiểu được mối tương quan giá điện thoại với pixel height

Machine Learning:

Dùng các mô hình KNN, Linear Regression, Decision Tree, Support Vector Machine.

* Gía điện thoại càng cao nếu các tính năng ram, battery power, pixel height càng lớn.

**3.Dữ liệu (data)**

Nguồn dữ liệu từ Kaggle: https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification?datasetId=11167

Tác giả: ABHISHEK SHARMA

Dữ liệu phân tích: train.csv, test.csv, moblie.csv

Tập dữ liệu bao gồm 21 cột và 2000 hàng.

21 cột bao gồm các giá trị:

+ Id

+ battery\_power: Tổng năng lượng pin có thể lưu trữ

+ blue: Có bluetooth hay không

+ clock\_speed:tốc độ mà bộ vi xử lý thực hiện các lệnh

+ dual\_sim: có hỗ trợ hai sim hay không

+ fc: Độ phân giải camera trước

+ four\_g: có 4g hay không

+ int\_memory: dung lượng bộ nhớ trong

+ m\_dep: độ dày điện thoại

+ mobile\_wt: độ nặng điện thoại

+ n\_cores:Số lõi của bộ xử lý

+ pc: Độ phân giải camera sau

+ px\_height: chiều cao của độ phân giải màn hình

+ px\_width: chiều rộng của độ phân giải màn hình

+ ram: Dung lượng của Ram

+ sc\_h: chiều cao màn hình

+ sc\_w: chiều rộng màn hình

+ talk\_time: thời gian dài nhất khi sạc điện thoại

+ three\_g: có 3g hay không

+ touch\_screen: có màn hình cảm ứng hay không

+ wifi: có wifi hay không

+ price\_range: khoảng giá trị chi phí điện thoại (0,1,2,3)

Tập dữ liệu chia làm hai phần: 75% để train và 25% để test, nhóm thực hiện loại bỏ đi các đơn vị đo ở từng cột, và tiến thành lọc bỏ đi các giá trị NA và NULL





**4. Trực quan hóa dữ liệu (data visulization):**

3 dòng đầu

Table

Description automatically generated

3 dòng cuối

Table

Description automatically generated

Phân tích EDA: Các phân tích sẽ làm cơ sở để trả lời các câu hỏi đặt ra và tiến hành phân tích Machine Learning

Chart, scatter chart

Description automatically generated

Nhận xét : Từ hình dung ma trận tương quan này, thấy rằng 2 đặc điểm có tương quan cao nhất là: "price\_range" và "ram"

Liên quan đến price\_range (biến dùng dự đoán), ta cũng có thể thấy rằng "pin\_power" và "px\_height" đều có tương quan nhẹ với "price\_range"

Chart, scatter chart

Description automatically generated

Nhận xét: Các biến Ram tương quan với price\_range lên đến 92%, px\_hight tương quan 14 % và battery\_power tương quan lên đến 24%

Chart, line chart

Description automatically generated

Nhận xét: ram có sự tương quan cao với price\_range dẫn đến khi ram tăng mạnh thì giá cũng tăng theo tương ứng

Chart, line chart

Description automatically generated

- Nhận xét: battery Power có sự tương quan nhẹ với price\_range nên đồ thị EDA thể hiện sự tăng trưởng về giá nhẹ

Chart, line chart

Description automatically generated

- Nhận xét: px\_height có sự tương quan nhẹ với price\_range nên đồ thị EDA thể hiện sự tăng trưởng về giá nhẹ

**5.** **Mô hình hóa dữ liệu (data modeling):**

* **Mô hình Decision Tree:**
* Cây quyết định (Decision Tree) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật. Các thuộc tính của đối tượngncó thể thuộc các kiểu dữ liệu khác nhau như Nhị phân (Binary) , Định danh (Nominal), Thứ tự (Ordinal), Số lượng (Quantitative) trong khi đó thuộc tính phân lớp phải có kiểu dữ liệu là Binary hoặc Ordinal.
* Để xác định các nút trong mô hình cây quyết định, ta thực hiện tính Infomation Gain tại mỗi nút theo trình tự sau:
  + **Bước 1**: Tính toán hệ số Entropy của biến mục tiêu S có N phần tử với Nc phần tử thuộc lớp c cho trước:
* H(S)=  – ∑cc=1 (Nc/N) log(Nc/N)
  + **Bước 2**: Tính hàm số Entropy tại mỗi thuộc tính: với thuộc tính x, các điểm dữ liệu trong S được chia ra K child node S1, S2, …, SK với số điểm trong mỗi child node lần lượt là m1, m2 ,…, mK , ta có:
* H(x, S) = ∑Kk=1 (mk / N) \* H(Sk )
  + **Bước 3**: Chỉ số Gain Information được tính bằng:
* **G(x, S) = H(S) – H(x,S)**
* **Mô Hình Support Vector Machine:**
* VM là một thuật toán giám sát, nó có thể sử dụng cho cả việc phân loại hoặc đệ quy. Tuy nhiên nó được sử dụng chủ yếu cho việc phân loại. Trong thuật toán này, chúng ta vẽ đồi thị dữ liệu là các điểm trong n chiều ( ở đây n là số lượng các tính năng bạn có) với giá trị của mỗi tính năng sẽ là một phần liên kết. Sau đó chúng ta thực hiện tìm "đường bay" phân chia các lớp. Đường bay - nó chỉ hiểu đơn giản là 1 đường thằng có thể phân chia các lớp ra thành hai phần riêng biệt.
* **Mô hình K-Nearest-Neighbor**
* K- Nearest Neighbors (kNN) là một thuật toán học máy có giám sát (supervised learning(2)) có thể được sử dụng trong cả phân loại (classification(3)) và hồi quy (regression(4)).
* Trong bài toán Classification, label của một điểm dữ liệu mới được suy ra trực tiếp từ k điểm dữ liệu gần nhất trong training set. Label của một test data có thể được quyết định bằng cách bầu chọn theo số phiếu giữa các điểm gần nhất, hoặc nó có thể được suy ra bằng cách đánh trọng số khác nhau cho mỗi trong các điểm gần nhất đó rồi suy ra label.
* Trong bài toán Regresssion, đầu ra của một điểm dữ liệu sẽ bằng chính đầu ra của điểm dữ liệu đã biết gần nhất (trong trường hợp K=1), hoặc là trung bình có trọng số của đầu ra của những điểm gần nhất, hoặc bằng một mối quan hệ dựa trên khoảng cách tới các điểm gần nhất đó.
* Ý tưởng: đi tìm đầu ra của một điểm dữ liệu mới bằng cách chỉ dựa trên thông tin của K điểm dữ liệu trong training set gần nó nhất (k-lân cận), không quan tâm đến việc có một vài điểm dữ liệu trong những điểm gần nhất này là nhiễu. Cho K chạy trong 1 khoảng giá trị và tìm ra k sao cho mô hình có kết quả tốt nhất.
* **Mô hình Linear Regression (hồi quy tuyến tính)**
* Công thức chung của thuật toán Linear Regression:
* Phân tích hồi quy tuyến tính là một phương pháp phân tích quan hệ giữa biến phụ thuộc Y với một hay nhiều biến độc lập X. Mô hình hóa sử dụng hàm tuyến tính. Các tham số của mô hình được ước lượng từ dữ liệu. Hồi quy tuyến tính được sử dụng rộng rãi trong thực tế do tính chất đơn giản hóa của hồi quy.
* Ý tưởng: xây dựng mô hình dựa trên dữ liệu huấn luyện (tập train). Sử dụng phương pháp cross validation(1) để điều chỉnh siêu tham số và mô hình. Tiến hành chia tập dữ liệu huấn luyện (train) thành 10 phần bao gồm 9 phần train và 1 phần validate.

**6. Thực Nghiệm, kết quả và thảo luận (experiments, results, and discussion):**

Nhóm chọn mô hình theo cross-validation thì nhóm chia dữ liệu làm 2 phần train 75% và test 25% và chia train làm 10 phần(9 phần để train, 1 phần để validation), và nhóm tìm ra được giữa 4 mô hình:

+ Decision Tree:



+ K-Nearest Neighbors: Thử nghiệm với siêu tham số từ 1 đến 15

Table

Description automatically generated

Trên tập test

Text

Description automatically generated with medium confidence

Trên tập train

Graphical user interface, text, application

Description automatically generated

+ Support Vector Machine (SVM):

* Root Mean Squared Error:
  + Train:



* + Test:



* Mean Squared Error:’
  + Train:



* + Test:



Và Linear Regression:

Table

Description automatically generated

* Root Mean Squared Error:
  + Train:



* + Test:



* Mean Squared Error:’
  + Train:



* + Test:



🡪 Từ kết quả, Nhóm thấy rằng các giá trị của Mô hình Linear Regression cho RMSE và MSE tương ứng nhỏ hơn các giá trị RMSE và MSE cho Mô hình SVM và KNN, Decision tree

**7. Kết Luận (conclusion):**

Đầu tiên, nhìn vào ma trận tương quan, Nhóm thấy rằng ram là biến có tương quan nhất với phạm vi giá, năng lượng pin là biến có tương quan nhiều thứ hai với phạm vi giá và chiều cao px là biến tương quan thứ ba với phạm vi giá.

Bây giờ, nhìn vào phần tóm tắt của Mô hình Linear Regression, thấy rằng cả 3 biến (ram, battery power, pixel height) đều có t-values rất cao, nghĩa là chúng đều rất quan trọng trong việc dự đoán biến y của nhóm. Cụ thể, ram có t-values cao nhất, battery power có t-values cao thứ hai và pixel height có t-values cao thứ ba. “Thứ tự của tầm quan trọng thay đổi” này trực tiếp tương ứng với những phát hiện của nhóm từ ma trận tương quan của chúng tôi ở trên!

Tóm lại, các tính năng quan trọng nhất trong việc dự đoán giá của một chiếc Điện thoại di động là ram, battery power, pixel height.

**8. Phụ Lục:**

(1): ***Cross validation*** là một kỹ thuật lấy mẫu để đánh giá mô hình học máy trong trường hợp dữ liệu không được dồi dào cho lắm.

Tham số quan trọng trong kỹ thuật này là k, đại diện cho số nhóm mà dữ liệu sẽ được chia ra. Vì lý do đó, nó được mang tên k-fold [cross-validation](https://trituenhantao.io/tu-dien-thuat-ngu/cross-validation/). Khi giá trị của k được lựa chọn, người ta sử dụng trực tiếp giá trị đó trong tên của phương pháp đánh giá. Ví dụ với k=10, phương pháp sẽ mang tên **10-fold**[cross-validation](https://trituenhantao.io/tu-dien-thuat-ngu/cross-validation/)**.**

Kỹ thuật này thường bao gồm các bước như sau:

1. Xáo trộn [dataset](https://trituenhantao.io/tu-dien-thuat-ngu/dataset/) một cách ngẫu nhiên
2. Chia [dataset](https://trituenhantao.io/tu-dien-thuat-ngu/dataset/) thành k nhóm
3. Với mỗi nhóm:
   * Sử dụng nhóm hiện tại để đánh giá hiệu quả mô hình
   * Các nhóm còn lại được sử dụng để huấn luyện mô hình
   * Huấn luyện mô hình
   * Đánh giá và sau đó hủy mô hình
4. Tổng hợp hiệu quả của mô hình dựa từ các số liệu đánh giá

(2) ***Supervised learning:*** là thuật toán dự đoán đầu ra (outcome) của một dữ liệu mới (new input) dựa trên các cặp (input, outcome) đã biết từ trước. Cặp dữ liệu này còn được gọi là (data, label), tức (dữ liệu, nhãn). Supervised learning là nhóm phổ biến nhất trong các thuật toán Machine Learning.

(3) **Classification:** Một bài toán được gọi là classification nếu các label của input data được chia thành một số hữu hạn nhóm.

(4) ***Regression:*** Nếu label không được chia thành các nhóm mà là một giá trị thực cụ thể.

**9. Đóng góp (Contributions):**

|  |  |  |  |
| --- | --- | --- | --- |
| **Thành Viên** | **MSSV** | **Công Việc** | **Đóng góp** |
| Lý Quốc Dũng | 19133015 | EDA, Linear Regression | 100% |
| Đoàn Trần Đăng Khoa | 19133028 | EDA, Support Vector Machine | 100% |
| Nguyễn Quốc Bảo | 19133002 | EDA, K- Nearest Neighbor | 100% |
| Võ Hoàng Khả Diệu | 19133014 | EDA, Decision Tree, tương quan | 100% |

**10. Tham Khảo:**

* 1. Bách khoa toàn thư wikipedia, Hồi quy tuyến tính

<https://vi.wikipedia.org/wiki/H%E1%BB%93i_quy_tuy%E1%BA%BFn_t%C3%ADnh>

1. Huy Nguyen, K-Nearest Neighbors (KNN)

<https://datasciencebasic.com/?p=134>

1. Bài 2: Phân nhóm các thuật toán Machine Learning,27/12/2016

<https://machinelearningcoban.com/2016/12/27/categories/>

1. Trí tuệ nhân tạo,Giới thiệu về k-fold cross-validation, 30/01/2020

<https://trituenhantao.io/kien-thuc/gioi-thieu-ve-k-fold-cross-validation/>

1. Bài 6: K-nearest neighbors, 08/01/2017

<https://machinelearningcoban.com/2017/01/08/knn/>