

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TPHCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

**MÔN: KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI:**

**DỰ ĐOÁN PHÂN LOẠI KHÁCH HÀNG**

|  |  |  |
| --- | --- | --- |
| **GVHD: Ths. QUÁCH ĐÌNH HOÀNG** | |  |
| **SVTH: TRẦN VĂN DUY** | **19133016** |
| **NGUYEN DUY PHƯỚC** | **19133003** |
| **TRẦN CÔNG TRƯỜNG** | **19133062** |
| **CAO ANH VĂN** | **19133067** |

**TP. Hồ Chí Minh, ngày 13 tháng 06 năm 2022**

|  |  |
| --- | --- |
| **ĐH SƯ PHẠM KỸ THUẬT TP.HCM**  **KHOA CNTT**  **\*\*\*\*\*\*\*** | **XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc lập – Tự do – Hạnh phúc**  **\*\*\*\*\*\*\*** |

**PHIẾU NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

Họ và tên Sinh viên 1: Trần Văn Duy MSSV: 19133016

Họ và tên Sinh viên 2: Nguyễn Duy Phước MSSV: 19133003

Họ và tên Sinh viên 3: Trần Công Trường MSSV: 19133062

Họ và tên Sinh viên 4: Cao Anh Văn MSSV: 19133067

Ngành: Kỹ thuật dữ liệu

Tên đề tài: **DỰ ĐOÁN PHÂN LOẠI KHÁCH HÀNG**.

Họ và tên Giáo viên hướng dẫn: Quách Đình Hoàng

**NHẬN XÉT:**

1. Về nội dung đề tài và khối lượng thực hiên:

...………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

1. Ưu điểm:

…………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

1. Khuyết điểm:

…………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

1. Đề nghị cho bảo vệ hay không?
2. Đánh giá loại:
3. Điểm:

*Tp. Hồ Chí Minh, ngày tháng năm 202*

Giáo viên hướng dẫn

*(Ký & ghi rõ họ tên)*

|  |  |
| --- | --- |
| **ĐH SƯ PHẠM KỸ THUẬT TP.HCM**  **KHOA CNTT**  **\*\*\*\*\*\*\*** | **XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc lập – Tự do – Hạnh phúc**  **\*\*\*\*\*\*\*** |

**PHIẾU NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN**

Họ và tên Sinh viên 1: Trần Văn Duy MSSV: 19133016

Họ và tên Sinh viên 2: Nguyễn Duy Phước MSSV: 19133003

Họ và tên Sinh viên 3: Trần Công Trường MSSV: 19133062

Họ và tên Sinh viên 4: Cao Anh Văn MSSV: 19133067

Ngành: Kỹ thuật dữ liệu

Tên đề tài: **DỰ ĐOÁN PHÂN LOẠI KHÁCH HÀNG**.

Họ và tên Giáo viên phản biện: ………………………………………………………….

**NHẬN XÉT:**

1. Về nội dung đề tài và khối lượng thực hiên:

...………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

1. Ưu điểm:

…………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

1. Khuyết điểm:

…………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

1. Đề nghị cho bảo vệ hay không?
2. Đánh giá loại:
3. Điểm:

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

Giáo viên phản biện

*(Ký & ghi rõ họ tên)*

**Mục Lục**

[LỜI NÓI ĐẦU 1](#_Toc106045190)

[1. Tóm tắt (abstract) 2](#_Toc106045191)

[2. Giới thiệu (introduction) 2](#_Toc106045192)

[3. Dữ liệu (data) 3](#_Toc106045193)

[3.1. Nguồn gốc của tập dữ liệu : 3](#_Toc106045194)

[3.2. Các thuộc tính của tập dữ liệu : 3](#_Toc106045195)

[3.3. Tập dữ liệu gồm 1 file dữ liệu dạng csv. 4](#_Toc106045196)

[3.4. Một vài dòng của tập dữ liệu 4](#_Toc106045197)

[3.5. Tiền xử lý dữ liệu 4](#_Toc106045198)

[4. Phương pháp (method) 4](#_Toc106045199)

[4.1. Decision tree - cây quyết định 4](#_Toc106045200)

[4.2. Random Forest - Rừng cây ngẫu nhiên 6](#_Toc106045201)

[4.3 Mô hình K-NN 7](#_Toc106045202)

[4.4 Mô hình Logistic Regression 8](#_Toc106045203)

[5. Thực nghiệm, kết quả, và thảo luận (experiments, results, and discussions) 8](#_Toc106045204)

[5.1. Các độ đo mà nhóm sử dụng 8](#_Toc106045205)

[5.2. Mô hình decision tree 9](#_Toc106045206)

[5.3. Mô hình Random Forest 14](#_Toc106045207)

[5.4. Mô hình K-NN 17](#_Toc106045208)

[5.5. Mô Hình Logistic Regression 20](#_Toc106045209)

[6. Kết luận (conclusion) 23](#_Toc106045210)

[7. Đóng góp 24](#_Toc106045211)

[8. Tham khảo (references) 26](#_Toc106045212)

# LỜI NÓI ĐẦU

Lời đầu tiên, nhóm em xin gửi lời cảm ơn chân thành và sâu sắc đến thầy Quách Đình Hoàng đã tận tình chỉ bảo và giúp đỡ nhóm chúng em để có thể hoàn thành được đồ án cuối kì.

Chúng em cũng xin gửi lời cảm ơn tới quý thầy cô Khoa Công nghệ Thông tin, trường đại học Sư Phạm Kỹ Thuật TP Hồ Chí Minh đã cung cấp các kiến thức nền tảng để từ đó chúng em thực hiện tiểu luận này.

Dù nhóm đã cố gắng nhưng trong quá trình thực hiện bài làm vẫn còn những điểm sai sót chưa khắc phục được, rất mong thầy có thể thông cảm và cho chúng em nhứng ý kiến quý báu để khắc phục và hoàn thiện tốt hơn.

TP. Hồ Chí Minh, ngày 13 tháng 06 năm 2022

**Nhóm sinh viên**

**Trần Văn Duy Nguyễn Duy Phước Trần Công Trường Cao Anh Văn**

# Tóm tắt (abstract)

Theo nghiên cứu đã chỉ ra rằng việc thu hút 1 khách hàng mới tốn nhiều chi phí hơn việc giữ khách hàng hiện tại. Trên thực tế, việc tăng tỷ lệ giữ chân khách hàng chỉ 5% cũng có thể tạo ra lợi nhuận tăng ít nhất 25%. Bài toán nhóm muốn giải quyết là dự đoán khách hàng có thuộc vào nhóm khách hàng sẽ rời đi hay không để đưa ra phương án giữ khách hàng. Việc dự đoán được khách hàng nào có khả năng rời đi thì sẽ giúp cho người quản lý đưa ra giải pháp kịp thời để không mất đi một khách hàng tiềm năng.

Phương pháp nhóm sử dụng để thực hiện bao gồm: xử lý dữ liệu Na, phân tích mô tả để để trực quan hóa dữ liệu thông qua các biểu đồ nhằm loại bỏ các biến có mối tương quan cao và phương pháp phân tích chuẩn đoán. Trong phương pháp phân tích chuẩn đoán nhóm sử dụng các mô hình machine learning: Random Forest, Logistic Regression, KNN và Decision Tree để dự đoán việc khách sẽ rời đi hay ở lại từ đó đưa ra biện pháp giải quyết.

# Giới thiệu (introduction)

Trong thời đại 4.0, việc ứng dụng công nghệ vào mọi lĩnh vực đời sống đóng vai trò vô cùng quan trọng, đặc biệt là ứng dụng các công nghệ trong phân tích dữ liệu giúp chúng ta dự đoán được những kết quả sẽ xảy ra trong tương lai và tìm một giải pháp để đưa ra kết quả tốt nhất. Vì vậy, trong bài phân tích này, chúng em sẽ ứng dụng ngôn ngữ R để phân tích các yếu tố ảnh hưởng tới quyết định rời đi của khách hàng ở một công ty viễn thông. Từ đó có thể đưa ra các giải pháp tối ưu để giữ chân khách hàng, nâng cao các dịch vụ từ đó đem lại hiệu quả kinh tế cao cho công ty

Câu hỏi nghiên cứu của nhóm là dự đoán các khách hàng có khả năng rời đi cao nhất dựa trên các dịch vụ, thông tin các nhân của khách hàng. Việc tìm ra các khách hàng có khả năng rời đi cao là một vấn đề thiết yếu và quan trọng, ảnh hưởng tới các quyết định phát triển của công ti. Bởi vì khi tìm ra được các khách hàng có khả năng rời đi cao, công ti có thể tìm được các yếu tố ảnh hưởng lớn đến quyết định rời đi của khách hàng. Từ đó có thể có các phương hướng phù hợp để giữ chân khách hàng

Input của bài toán là các yếu tố cá nhân, và các dịch vụ mà khách hàng sử dụng tại công ty. Output của bài toán là biến “Churn” mang giá trị “yes” hoặc “no” cho thấy quyết định có rời đi hay không của khách hàng. Các bài toán được sử dụng trong mô hình là: K-NN, random forest, Decision tree, Logistic Regression

# Dữ liệu (data)

## 3.1. Nguồn gốc của tập dữ liệu :

* Tập dữ liệu được lấy từ trang trang kaggle. Link đến tập dữ liệu : <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>
* Dữ liệu được cung cấp bởi một công ty viễn thông.

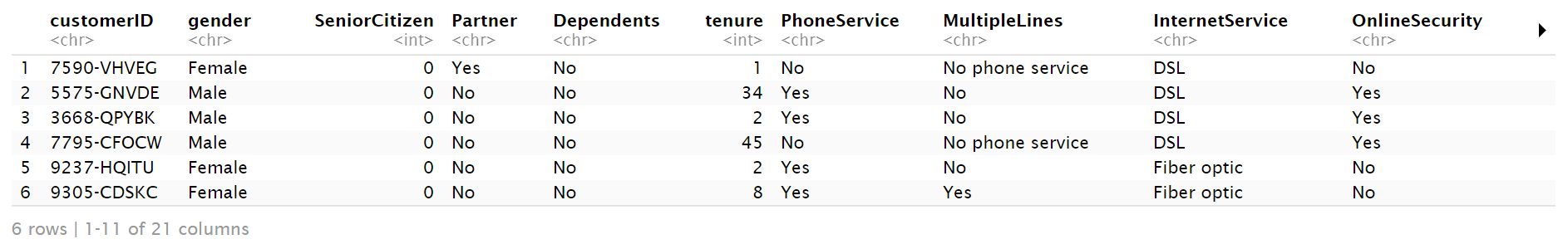
## 3.2. Các thuộc tính của tập dữ liệu :

1. CustomerID: id của mỗi khách hàng.
2. Gender: giới tính ( male, female).
3. SeniorCitizen: người cao tuổi (0,1).
4. Partner: Có cộng sự hay không ( yes, no).
5. Dependents: Có người phụ thuộc hay không ( yes, no).
6. tenure: số tháng ở lại với công ty.
7. PhoneService: có sử dụng dịch vụ điện thoại hay không ( yes, no).
8. MultipleLines: có nhiều đường dây hay không ( yes, no, no phone).
9. InternetService: nhà cung cấp dịch vụ điện thoại (DSL, có, không).
10. Online Security : Cho biết liệu khách hàng có đăng ký dịch vụ bảo mật trực tuyến bổ sung do công ty cung cấp hay không (yes, no).
11. OnlineBackup: dịch vụ sao lưu online ( yes, no, no internet ).
12. DeviceProtection: dịch vụ bảo vệ thiết bị ( yes, no , no internet).
13. TechSupport: hỗ trợ kĩ thuật ( yes, no , no internet).
14. StreamingTV: truyền hình trực tuyến ( yes, no, no internet).
15. Contract: Cho biết loại hợp đồng hiện tại của khách hàng : Month-to-Month, One Year, Two Year.
16. StreamingMovies: xem phim trực tuyến ( yes, no , no internet).
17. PaperlessBilling: thanh toán trực tuyến ( yes, no) .
18. PaymentMethod: phương thức thanh toán.
19. Monthly Charges: thanh toán hằng tháng.
20. TotalCharges: tổng thanh toán.
21. Churn: quyết định rời đi hay tiếp tục sử dụng dịch vụ ( yes, no).

## Tập dữ liệu gồm 1 file dữ liệu dạng csv.

Trong quá trình làm nhóm sẽ chia tập dữ liệu thành 2 tập là train và test.

## 3.4. Một vài dòng của tập dữ liệu

****

Hình 3.4.1. 5 dòng dữ liệu đầu tiên của tập dữ liệu.

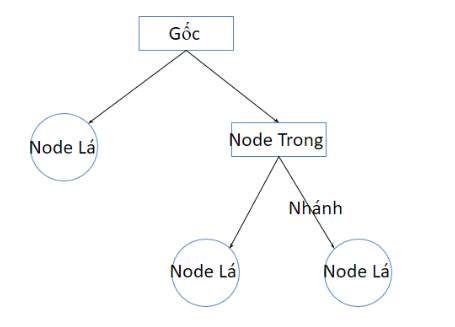
## 3.5. Tiền xử lý dữ liệu

* Tập dữ liệu có 11 dòng dữ liệu có biến TotalCharges nhận giá trị null, nên nhóm sẽ loại bỏ khỏi tập dữ liệu .
* Đối với việc chuẩn hóa dữ liệu thì mỗi cá nhân trong nhóm sẽ có cách chuẩn hóa riêng để phù hợp với model của mình áp dụng.

# 4. Phương pháp (method)

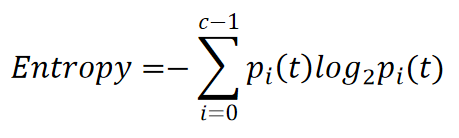
## 4.1. Decision tree - cây quyết định

* Cây quyết định là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật. Khi cho dữ liệu về các đối tượng gồm các thuộc tính cùng với lớp của nó, cây quyết định sẽ sinh ra các luật để dự đoán lớp của các đối tượng chưa biết.
* Cấu trúc của một cây quyết định :
* Root (Gốc): Là nút trên cùng của cây.
* Node trong: nút trung gian trên một thuộc tính đơn (hình Oval).
* Nhánh: Biểu diễn các kết quả của kiểm tra trên nút.
* Node lá: Biểu diễn lớp hay sự phân phối lớp (hình vuông hoặc chữ nhật)



Hình 4.1.1. Hình ảnh minh họa cây quyết định.

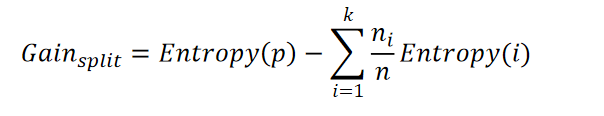
* Có nhiều hệ số được sử dụng trong thuật toán để đưa ra quyết định phân chia. Thường được sử dụng là Information Gain và Gini index.
* **Sử dụng Information Gain**
* **Entropy** trong học máy là thước đo tính ngẫu nhiên của thông tin đang được xử lý. Entropy càng cao, càng khó rút ra bất kỳ kết luận nào từ thông tin đó.



Hình 4.1.2. Công thức tính Entropy

Trong đó p\_i (t) là tần suất của lớp i tại nút t và c là tổng số lớp.

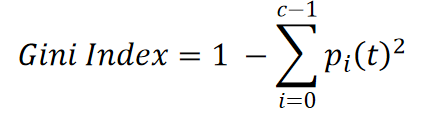
* **Information Gain** dựa trên sự giảm của hàm Entropy khi tập dữ liệu được phân chia trên một thuộc tính. Để xây dựng một cây quyết định, ta phải tìm thuộc tính trả về Infomation gain cao nhất.



Hình 4.1.3. Công thức tính information gain.

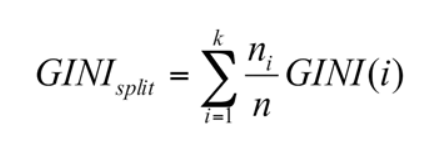
Sẽ chọn thuộc tính làm node khi Gain lớn nhất.

* **Sử dụng Gini index**
* Gini index là chỉ số thể hiện mức độ phân loại sai khi ta chọn ngẫu nhiên một phần tử từ tập data.



Hình 4.1.4. Công thức tính Gini Index.

Trong đó p\_i (t) là tần suất của lớp i tại nút t và c là tổng số lớp.



Hình 4.1.5. Công thức tính GINIsplit

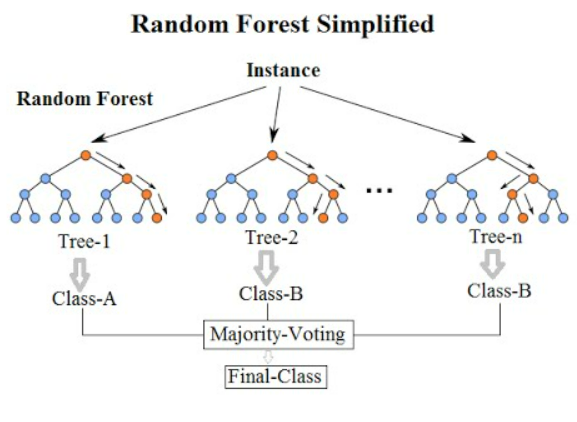
trong đó, n\_i = số bản ghi ở con i,

n = số bản ghi tại nút cha p.

* Ta sẽ chọn biến có GINIsplit nhỏ nhất để phân nhánh.

## 4.2. Random Forest - Rừng cây ngẫu nhiên

* Random là ngẫu nhiên, Forest là rừng, nên ở thuật toán Random Forest mình sẽ xây dựng nhiều cây quyết định bằng thuật toán Decision Tree, tuy nhiên mỗi cây quyết định sẽ khác nhau (có yếu tố random). Sau đó kết quả dự đoán được tổng hợp từ các cây quyết định.
* Mỗi cây thì phải thực hiện việc dự báo phân loại, sau đó kết quả dự báo chung cho cả nhóm cây (group overall prediction) thì sẽ được quyết định bằng majority vote.

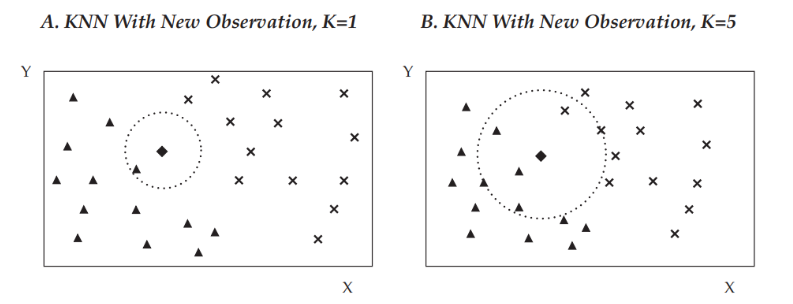


Hình 4.2.1 Hình ảnh minh họa về Random Forest

* Giả sử bộ dữ liệu của mình có n dữ liệu (sample) và mỗi dữ liệu có d thuộc tính (feature). Để xây dựng mỗi cây quyết định mình sẽ làm như sau:
* Lấy ngẫu nhiên n dữ liệu từ bộ dữ liệu với kỹ thuật Bootstrapping, hay còn gọi là random sampling with replacement.
* Sau khi sample được n dữ liệu từ bước 1 thì mình chọn ngẫu nhiên ở k thuộc tính (k < d). Giờ mình được bộ dữ liệu mới gồm n dữ liệu và mỗi dữ liệu có k thuộc tính.
* Dùng thuật toán Decision Tree để xây dựng cây quyết định với bộ dữ liệu ở bước 2.
* Ở trong quá trình huấn luyện lần này chúng ta sẽ sử dụng 2 siêu tham số là số lượng cây quyết định sẽ xây dựng (ntree), số lượng thuộc tính dùng để xây dựng cây (mtry).

## 4.3 Mô hình K-NN

* Khái niệm: Thuật toán knn hoạt động bằng cách tìm ra label của k điểm gần nhất với điểm dữ liệu cần được dự đoán label. Những điểm gần nhất được tìm ra thông qua việc tính khoảng cách giữa điểm đó với điểm cần dự đoán. Label của điểm dự đoán sẽ là label của những điểm có số lượng lớn hơn
* VD



Hình 4.3.1 Sự khác nhau khi lựa chọn siêu tham số k.

- Nếu k = 1, hình thoi sẽ được phân loại vào cùng loại với điểm dữ liệu gần nhất của nó (tức là hình tam giác trong bảng bên trái - bảng A).

- Bảng bên phải (bảng B) thể hiện trường hợp k = 5, thuật toán sẽ xem xét 5 điểm dữ liệu gần hình thoi nhất, đó là 3 hình tam giác và 2 hình chữ thập. Quy tắc quyết định là chọn phân loại có số lượng lớn nhất trong 5 điểm dữ liệu được xem xét. Vì vậy, trong trường hợp này, hình thoi cũng được xếp vào phân loại tam giác.

## 4.4 Mô hình Logistic Regression

- Khái niệm : Logistic Regression là 1 thuật toán phân loại được dùng để gán các đối tượng cho 1 tập hợp giá trị rời rạc (như 0, 1, 2, …). Một ví dụ điển hình là phân loại Email, gồm có email công việc, email gia đình, email spam, … Giao dịch trực tuyến có là an toàn hay không an toàn, khối u lành tính hay ác tình. Thuật toán trên dùng hàm sigmoid logistic để đưa ra đánh giá theo xác suất. Ví dụ: Khối u này 80% là lành tính, giao dịch này 90% là gian lận, …

- Mặc dù có tên là Regression, tức một mô hình cho fitting, Logistic Regression lại được sử dụng nhiều trong các bài toán Classification. Sau khi tìm được mô hình, việc xác định class y cho một điểm dữ liệu x được xác định bằng việc so sánh hai biểu thức xác suất:

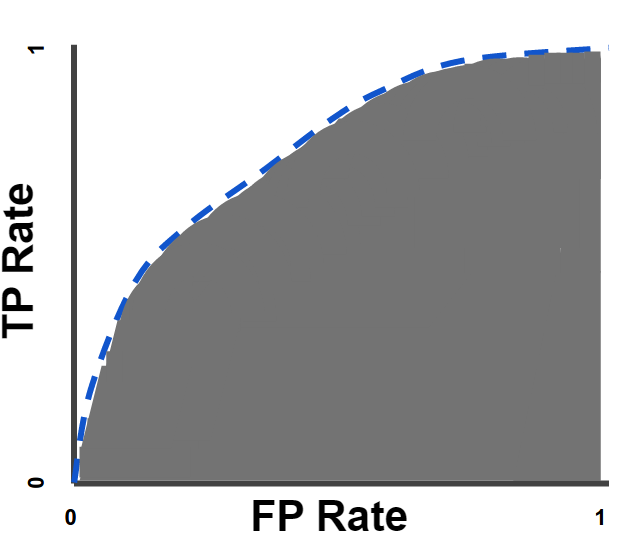


Nếu biểu thức thứ nhất lớn hơn thì ta kết luận điểm dữ liệu thuộc class 1, ngược lại thì nó thuộc class 0. Vì tổng hai biểu thức này luôn bằng 1 nên một cách gọn hơn, ta chỉ cần xác định xem P(y=1|x;w) lớn hơn 0.5 hay không. Nếu có, class 1. Nếu không, class 0.

# 5. Thực nghiệm, kết quả, và thảo luận (experiments, results, and discussions)

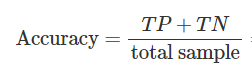
## 5.1. Các độ đo mà nhóm sử dụng

* **Độ đo AUC** : The Area Under the Curve (AUC) - Diện tích dưới đường cong (AUC) là thước đo khả năng phân biệt giữa các lớp của một bộ phân loại và được sử dụng như một bản tóm tắt của đường cong ROC. AUC càng cao, hiệu suất của mô hình càng tốt trong việc phân biệt giữa các lớp tích cực và tiêu cực.



Hình 5.1.1 Mô tả hình ảnh của AUC.

* **Accuracy :** Cách đơn giản và hay được sử dụng nhất là *accuracy* (độ chính xác). Cách đánh giá này đơn giản tính tỉ lệ giữa số điểm được dự đoán đúng và tổng số điểm trong tập dữ liệu kiểm thử.



Hình 5.1.2 Công thức tính Accuracy.

Trong đó:

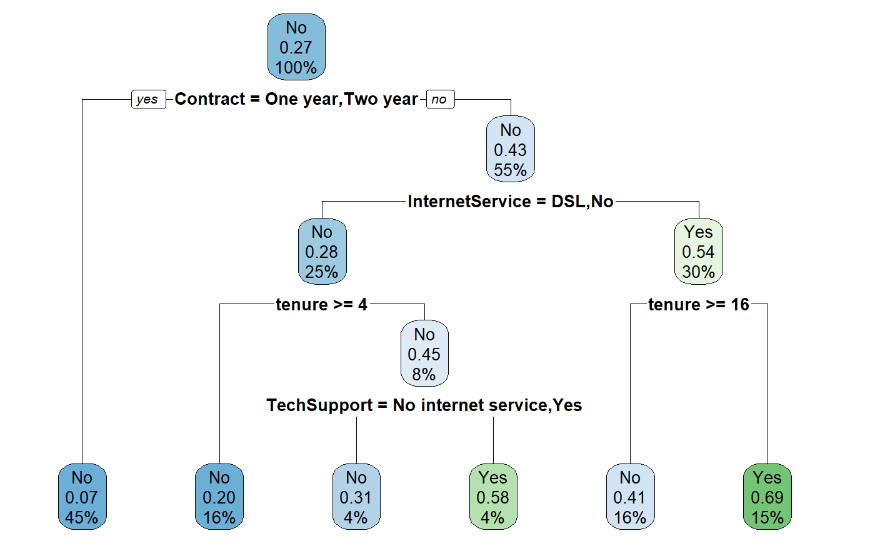
TP: Dương tính thật.

TN: Âm tính thật.

Total sample: số dòng của tập dữ liệu.

## 5.2. Mô hình decision tree

* Kết quả dự đoán của mô hình :
* Trước khi pruning - cắt tỉa
* Kết quả cây quyết định.

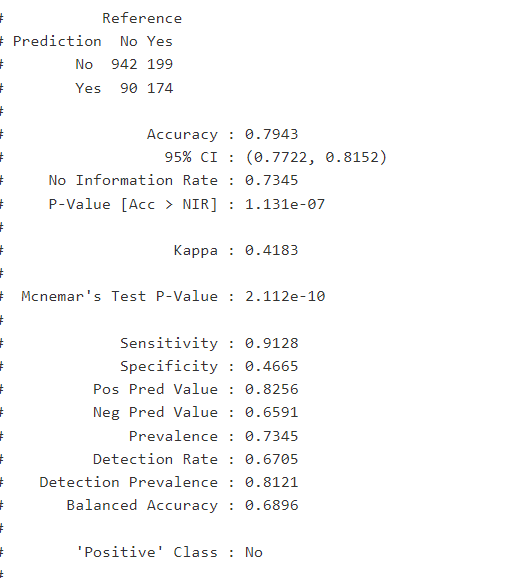


Hình 5.2.1 Hình ảnh cây quyết định.

Giải thích về cây : ví dụ xét tại nút gốc 0.27 thể hiện có 0.27% trong tổng thể nhận churn : yes.

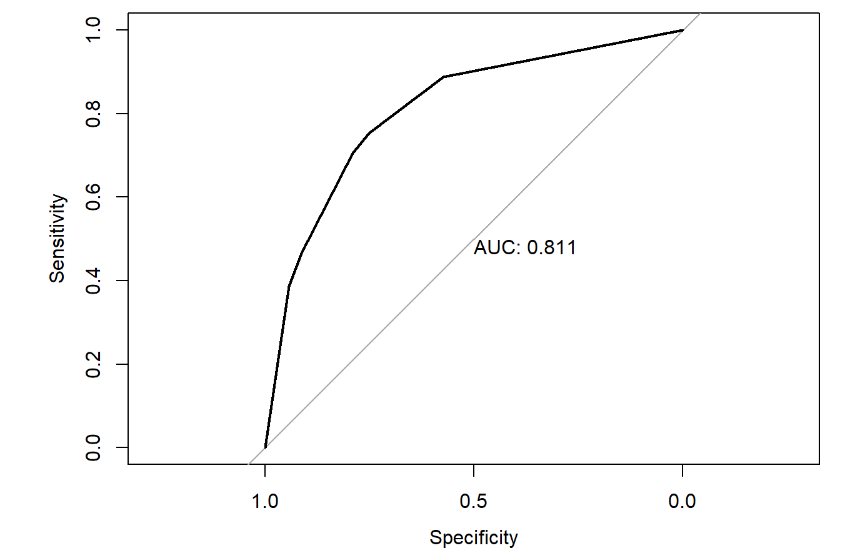
Khi xét contract ta thấy nếu yes thì sẽ có 445% contract=one year, two year có xác xuất nhận churn : yes là 0.07.

* Confusion Matrix và kết quả độ đo Accuracy



Hình 5.2.2. Confusion Matrix và kết quả độ đo Accuracy

* Kết quả của độ đo AUC

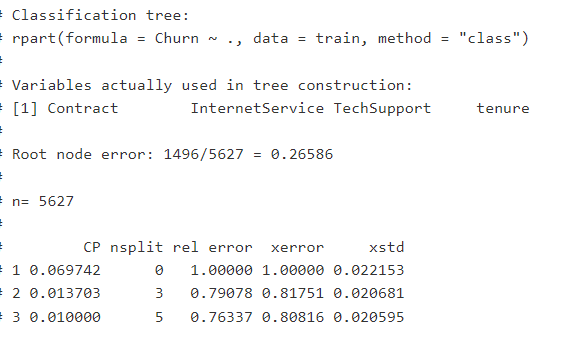


Hình 5.2.3. Kết quả độ đo AUC.

**Nhận xét :** Confusion Matrix sẽ nhận nhãn No là Positive.

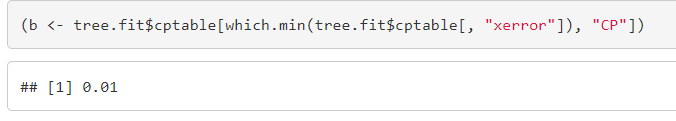
Dựa vào AUC=0.811 và Accuracy=0.7943, độ tin cậy 95% là (0.7722, 0.8152). Ta có thể thấy với mô hình trước khi pruning - cắt tỉa chỉ ở mức tương đối tốt.

* Sau khi pruning - cắt tỉa
* **Complexity parameter (cp)** : tham số độ phức tạp trong r để điều chỉnh độ sâu của cây.
* CP sẽ quyết định có tiếp tục tách nút hay không .
* CP sẽ so sánh tỷ lệ lỗi. Nếu sau khi tách mà lớn hơn trước khi tách thì sẽ dừng việc tách cây.
* Giá trị mặc định của cp là 0.0.
* Để cắt tỉa cây, chúng ta sẽ xem xét bảng *CP* được trả về trong hàm rpart (hàm được sử dụng trong model này)để tìm giá trị của tham số độ phức tạp với sai số dự đoán ước tính tối ưu. Sai số dự đoán ước tính của mỗi cây con (tương ứng với mỗi giá trị của CP) được chứa trong cột xerror và độ lệch chuẩn liên quan nằm trong cột xstd. Chúng ta sẽ muốn tìm giá trị của CP mang lại một cây được cắt tỉa tương ứng với sai số dự đoán ước tính nhỏ nhất.Ta sẽ chọn ra cây nhỏ nhất và có xerror nhỏ nhất.
* Bảng cp



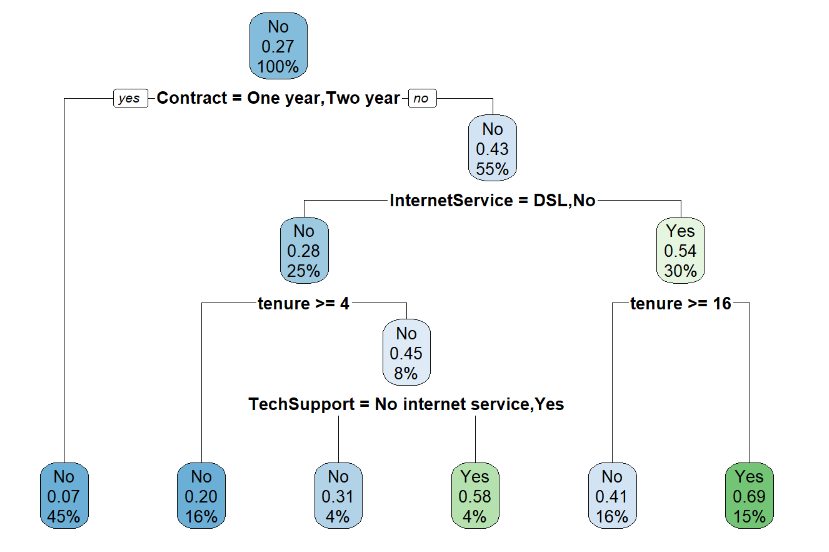
Hình 5.2.4. Bảng CP.

* Ta sẽ tìm ra cp tối ưu nhất - có xerror là nhỏ nhất.



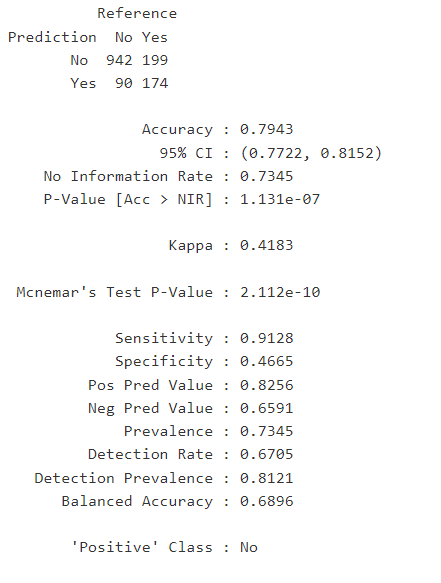
Hình 5.2.4. Cp tối ưu nhất.

* Với giá trị cp=0.01 ta sẽ vẽ lại cây và tính toán các độ đo
* Cây quyết định.



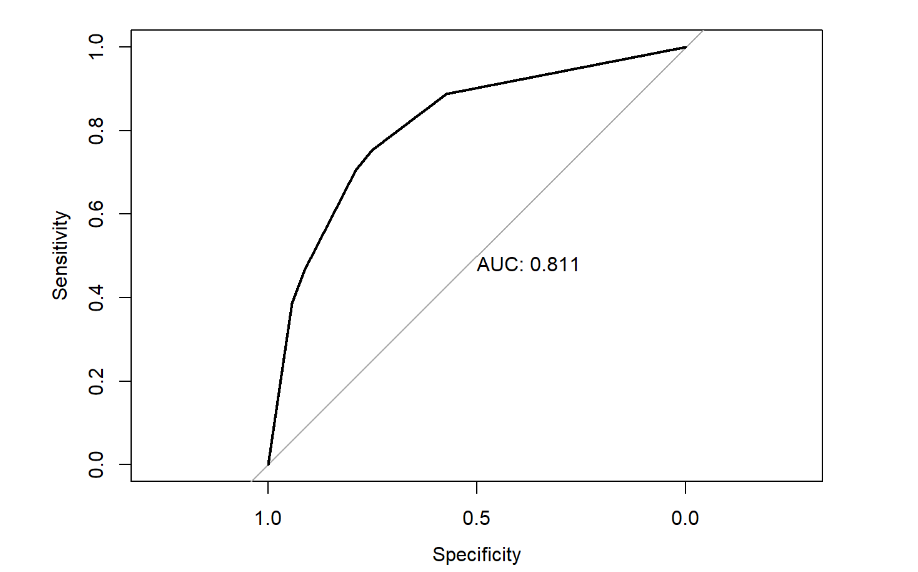
Hình 5.2.5. Cây quyết định sau khi cắt tỉa.

* Confusion Matrix và kết quả độ đo Accuracy



Hình 5.2.6. Confusion Matrix và Accuracy sau khi cắt tỉa.

* Đọ đo AUC



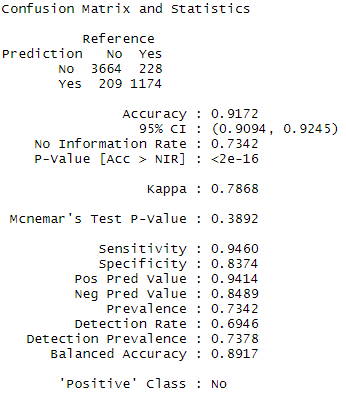
Hình 5.2.7. Độ đo AUC.

**Nhận xét :** ta thấy giá trị dùng để cắt tỉa cây cp nhận giá trị 0.01, trùng với giá trị mặc định của hàm được sử dụng trong model. Vì thế kết quả sau khi cắt tỉa cây sẽ không khác với trước khi cắt tỉa cây. Tóm lại với mô hình này thì ta nhận lại được kết quả chỉ ở mức tương đối tốt.

## 5.3. Mô hình Random Forest

* Kết quả dự đoán của mô hình:
* Trước khi cải thiện mô hình.

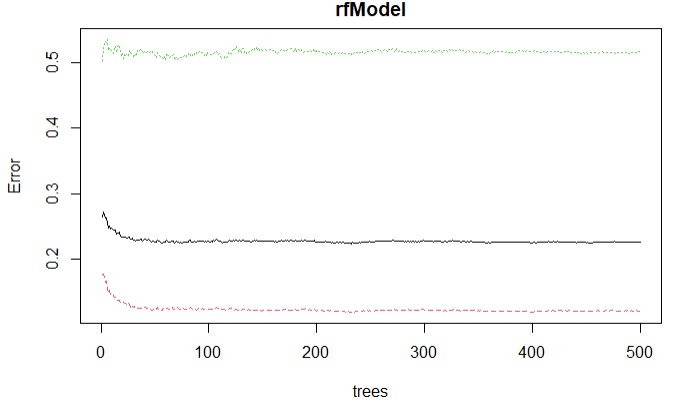
Kết quả của ma trận Confusion và Accuracy trước cải tiến .



Hình 5.3.1 Kết quả dự đoán trước khi cải thiện mô hình trên tập train.

* Kết quả sau khi cải thiện mô hình.
* Cải thiện mô hình.

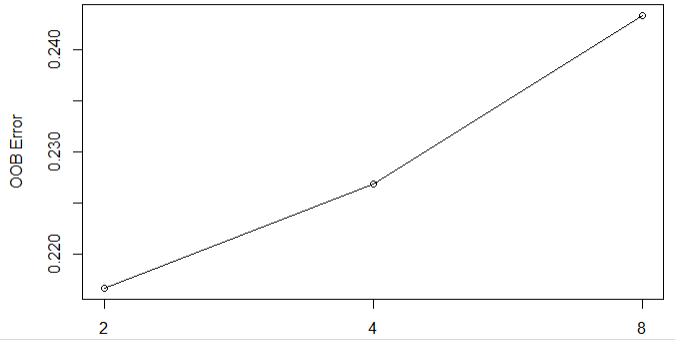
Với thuật toán Random Forest khi dự đoán mô hình với 2 siêu tham số ntree và mtry, ntree sẽ được mặc định là 500 còn mtry sẽ là căn bậc 2 của số lượng thuộc tính đầu vào. Ta cần tìm ntree và mtry nào để cho error của mô hình là bé nhất.



Hình 5.3.2 Tỷ lệ error tương ứng với từng ntree.

Với ntree bằng 200 thì tỷ lệ error trên mô hình hường như là không thay đổi nên chọn ntree= 200.

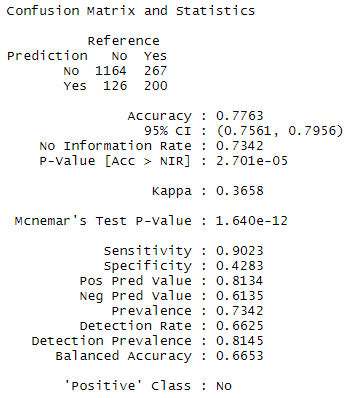
Sau khi xác định được ntree cần tìm mtry nào sẽ là tốt nhất cho mô hình.



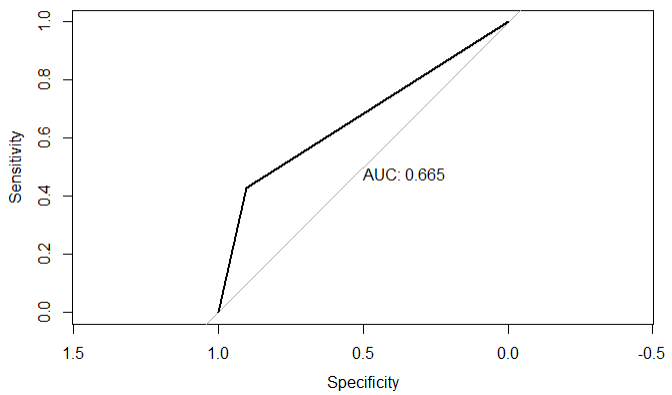
Hình 5.3.3 Tỷ lệ OOB Error ứng với từng mtry.

Với mtry = 2 sẽ cho kết quả tốt nhất.

* Kết quả sau khi cải thiện mô hình.



Hình 5.3.4 Kết quả dự đoán sau khi cải thiện mô hình trên tập test.



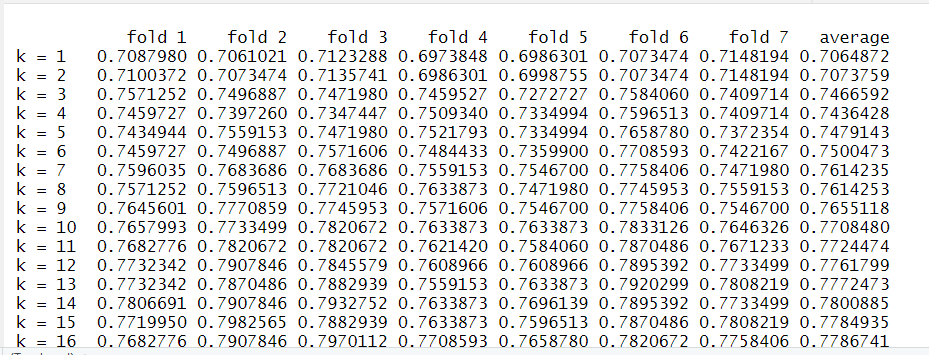
Hình 5.3.5 Độ đo AUC.

**Nhận xét:** Sau khi cải thiện mô hình với ntree = 200 và mtry = 2 thì cho kết quả Accurancy = 0.7763 và AUC = 0.665. Với kết quả này mô hình thật sự chưa tốt.

## 5.4. Mô hình K-NN

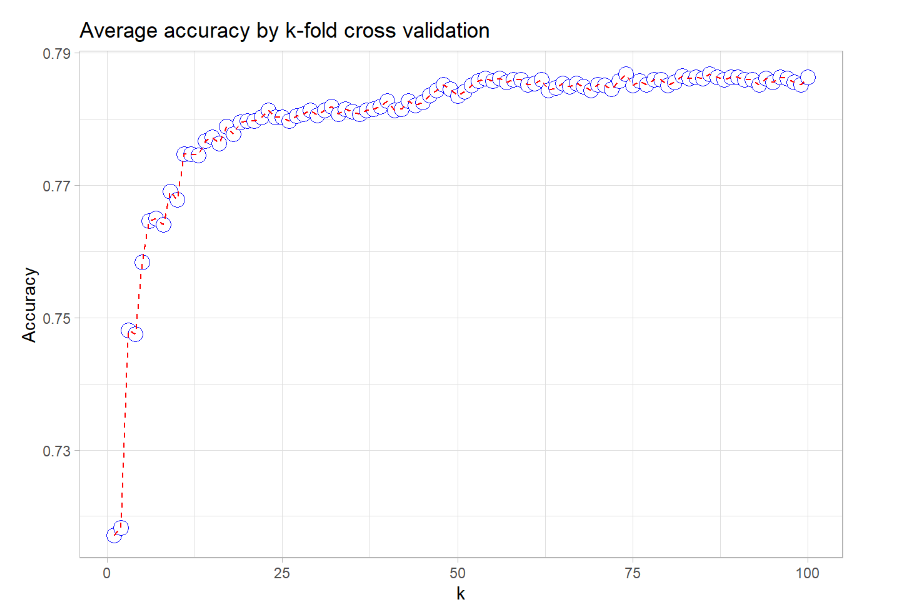
- Tạo mô hình

* Từ tập dữ liệu ta sẽ chia thành 2 phần là tập train và tập test. Sau đó ta tiến hành huấn luyện mô hình trên tập train bằng phương pháp cross validation.
* Khi huấn luyện bằng phương pháp cross validation trên tập train, ta sẽ chia tập train thành 7 fold. Khi đó 1 fold sẽ dùng để dự đoán và 6 fold còn lại sẽ dùng đề huấy luyện
* Với mỗi fold ta sẽ tiến hành huấn luyện với các giá trị k, tức là số điểm dữ liệu xung quanh điểm cần xét chạy từ 1 tới 100 điểm
* Sau khi thực hiện trên mỗi fold ta sẽ lưu lại accuracy của fold đó, và tính trung bình của cả 7 fold với mỗi giá trị k
* Sau đó sẽ tìm giá trị k cho kết quả accuracy lớn nhất và dùng nó để làm siêu tham số cho mô hình và tiến hành dự đoán trên tập test
* Huấn luyện trên tập train
* Kết quả sau khi huấn luyện bằng phương pháp cross validation: giá trị accuracy tính được với mỗi k trên mỗi fold và giá trị trung bình của mỗi k



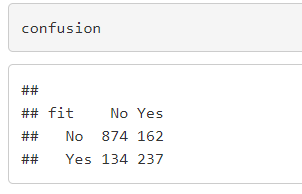
Hình 5.4.1 Biểu đồ thể hiện accuracy của mỗi fold trên từng k.

* Tìm giá trị k cho accuracy lớn nhất

****

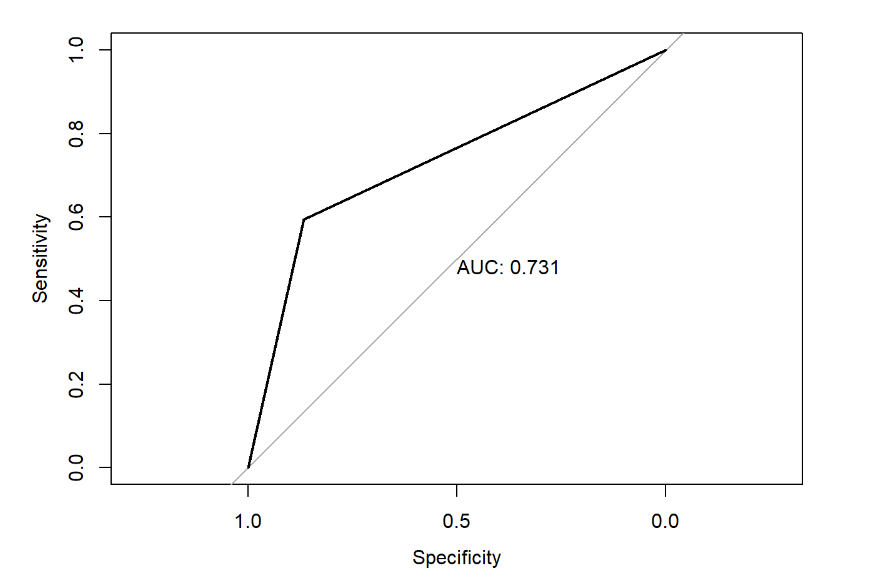
Hình 5.4.2 Biểu đồ thể hiện accuray của tửng k.

* Giá trị k tìm được là k = 74
* Kết quả dự đoán trên tập test
* Kết quả confusion matrix dự đoán được



Hình 5.4.3 Biểu đồ thể hiện confusion matrix.

* Kết quả accuracy: 0.7896
* Kết quả AUC



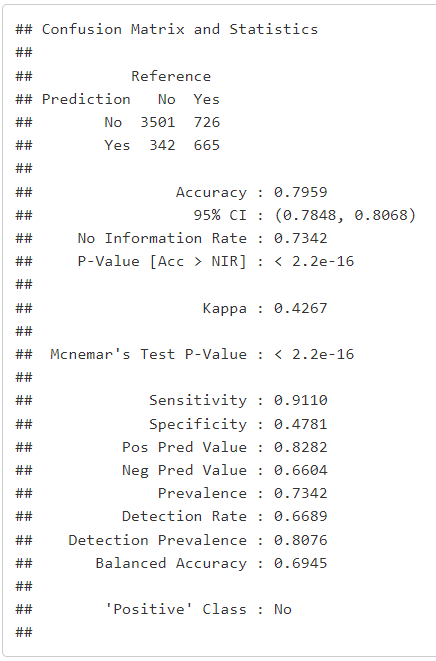
Hình 5.4.4 Biểu đồ thể hiện AUC

* Nhận xét
* Với mô hình K-NN, độ chính xác Accuracy = 0.7896 và chỉ số AUC= 0.731 vẫn còn tương đối thấp, chưa thể áp dụng vào thực tế
* Mô hình chưa đem lại hiệu quả cao

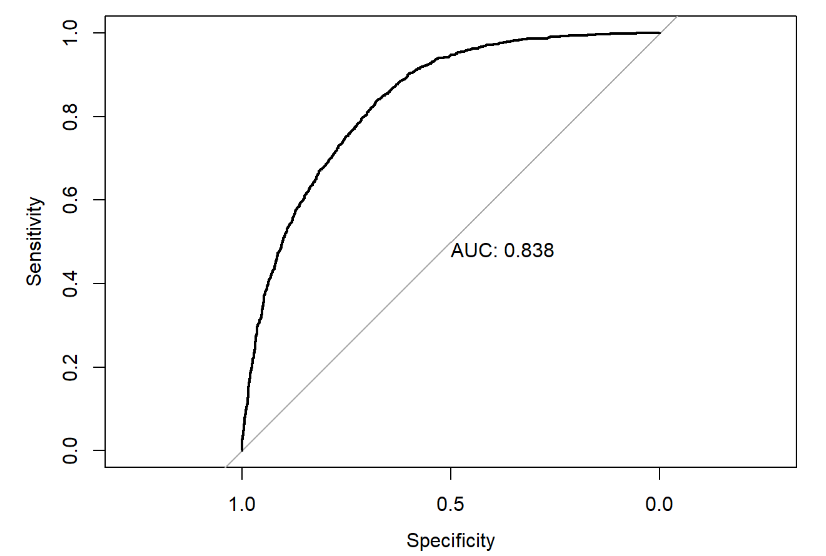
# 5.5. Mô Hình Logistic Regression

* Kết quả dự đoán của mô hình:

**Trên tập train :**

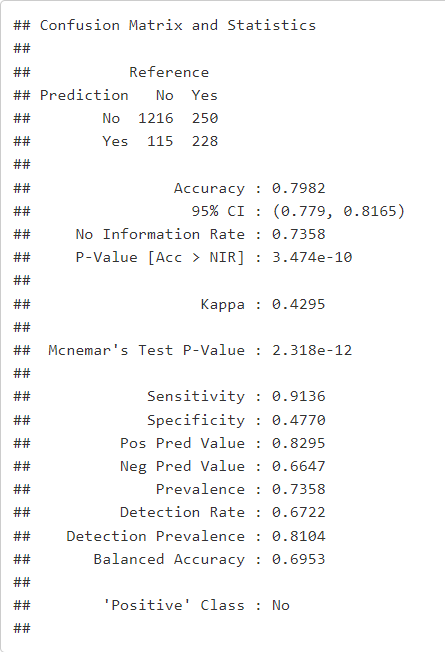


Hình 5.5.1 Kết quả trên tập Train

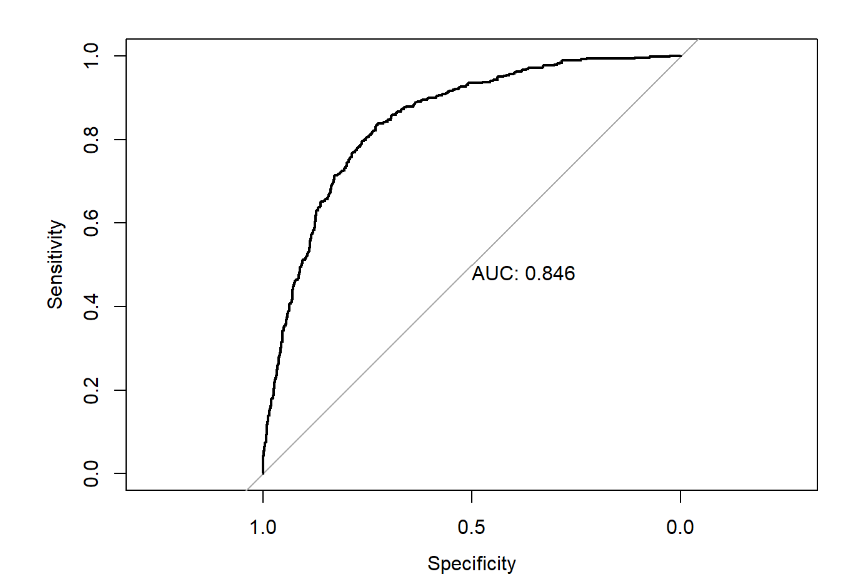


Hình 5.5.2 Chỉ số AUC trên Tập Train

**Trên tập Test :**



Hình 5.5.3 Kết quả trên Tập Test



Hình 5.5.4 Chỉ số AUC trên tập test

* Nhận xét :

- Trên tập train có acc khoảng 0.8 và AUC khoảng 0.85 và Tập test có acc khoàng 0.79 và AUC khoảng 0.82

- Model này rất tốt vì acc và AUC không có sự khác biệt lớn trên 2 tạp train test

- Nhưng chỉ số Specificities ở 2 tập còn rất thấp khoảng 0.46

- Mô hình hồi quy logistic cuối cùng (với ngưỡng = 0,5) có Độ chính xác là 0,79 và AUC là 0,82. Dựa trên các giá trị P cho các biến, PhoneService, InternetServiceDSL, OnlineBackup, Contract, PaperleslsBilling, PaymentMethodElectronic.check, Các khoản phí hàng tháng, thời hạn sử dụng trong 0-1 năm và 1-2 năm có ảnh hưởng đáng kể hơn đến việc dự đoán thời gian ngừng hoạt động.

# 6. Kết luận (conclusion)

* Kết quả chính :
* Nhóm đã thực hiện các bước để kiểm tra dữ liệu, xóa bỏ các giá trị null (đối với một số mô hình).
* Tìm ra các biến tương quan trong tập dữ liệu để loại bỏ khỏi tập dữ liệu trước khi dự đoán.
* Thực hiện dự đoán với 4 mô hình là decision tree, Random Forest, K-NN và Logistic Regression
* Từ các kết quả trên đưa ra được các nhận xét, kết luận và các định hướng để doanh nghiệp viễn thông có thể áp dụng để tăng tỷ lệ giữ chân khách hàng của mình.
* Mô hình hoạt động tốt nhất :
* Hồi quy logistic có độ chính xác là 0,79 và AUC là 0,82.
* Những dự định nhóm sẽ làm :
* Tìm hiểu thêm các mô hình khác và áp dụng thử xem có tăng hiệu suất so với các mô hình đã áp dụng hay không.
* Rút ra thêm các nhận xét, kết luận để đưa ra thêm các chiến lược phát triển tăng khả năng giữ chân khách hàng.

# 7. Đóng góp

|  |  |  |
| --- | --- | --- |
| **Thành viên** | **Công việc** | **Mức độ hoàn thành** |
| Trần Văn Duy | Model Random Forest | 100% |
| Nguyễn Duy Phước | Mô hình Logistic Regression | 100% |
| Trần Công Trường | Mô hình KNN | 100% |
| Cao Anh Văn | Model decision tree | 100% |

Bảng 7.1 Phân công nhiệm vụ và mức độ hoàn thành.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tên\Người đánh giá** | **Trần Văn Duy** | **Nguyễn Duy Phước** | **Trần Công Trường** | **Cao Anh Văn** |
| **Trần Văn Duy** |  | **10** | **10** | **10** |
| **Nguyễn Duy Phước** | **10** |  | **10** | **10** |
| **Trần Công Trường** | **10** | **10** |  | **10** |
| **Cao Anh Văn** | **10** | **10** | **10** |  |

Bảng 7.2 Bảng đánh giá tinh thần nhóm các thành viên của mỗi người trên thang 10.

# 8. Tham khảo (references)

* Quách Đình Hoàng, slide, video bài giảng môn Data mining , đại học Sư phạm Kỹ thuật thành phố Hồ Chí Minh, năm 2021.
* “10 Pruning regression trees with rpart, Daniel J. McDonald”, Matías Salibán-Barrera,2021/09/17, <https://ubc-stat.github.io/stat-406-worksheets/pruning-regression-trees-with-rpart.html>
* “Churn analysis - decision tree & random forest”, DANILO DA SILVA , 2018-09-14, <https://www.kaggle.com/code/danilodiogo/churn-analysis-decision-tree-random-forest>
* “Telco Customer Churn”, BLASTCHAR , <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>