**Ảnh có chứa quảng trường

Mô tả được tạo tự độngĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA HỆ THỐNG THÔNG TIN**

**BÁO CÁO ĐỒ ÁN**

**MÔN ĐIỆN TOÁN ĐÁM MÂY**

**Đề tài: SỬ DỤNG GOOGLE COLLAB ĐỂ PHÂN TÍCH VÀ DỰ ĐOÁN KHÁCH HÀNG RỜI BỎ DỊCH VỤ NGÂN HÀNG**



GVHD: ThS. Hà Lê Hoài Trung

Nhóm sinh viên thực hiện:

1. Nguyễn Cao Khoa MSSV: 19521694

2. Hồ Trọng Khang MSSV: 19521661

3. Trần Nhật Tân MSSV: 19522177

4. Huỳnh Quốc Khánh MSSV: 19521677

Tp. Hồ Chí Minh, 12/2022

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

*……., ngày……...tháng……năm 2022…*

**Người nhận xét**

*(Ký tên và ghi rõ họ tên****)***

**LỜI CẢM ƠN**

Nhóm xin chân thành cảm ơn thầy Hà Lê Hoài Trung - giảng viên môn Điện Toán Đám Mây đã nhiệt tình chỉ bảo và hướng dẫn tận tình trong suốt thời gian thực hiện đồ án môn học. Nhờ đó, chúng em đã có được nhiều kiến ​​thức bổ ích trong việc ứng dụng cũng như kỹ năng làm đồ án.

Nhóm em tìm hiểu về một thư viện mã nguồn mở cho machine learning là Tensorflow từ đó phân tích và dự đoán khả năng khách hàng rời bỏ dịch vụ dựa trên tệp dữ liệu Churn Modeling. Dựa trên những kiến ​​thức được giáo viên cung cấp trên lớp, kết hợp với việc tự nghiên cứu các công cụ và kiến ​​thức mới, nhóm cố gắng thực hiện dự án một cách tốt nhất có thể. Trong thời gian thực hiện dự án, nhóm chúng em đã sử dụng kiến ​​thức nền tảng đã tích lũy được, kết hợp học tập và nghiên cứu kiến ​​thức mới. Từ đó, nhóm của chúng em sử dụng đầy đủ các thông tin thu thập được để đưa ra báo cáo dự án tốt nhất có thể. Tuy nhiên, trong quá trình thực hiện, nhóm chúng em không tránh khỏi những thiếu sót. Vì vậy, nhóm rất mong nhận được sự góp ý của các thầy để nhóm em hoàn thiện hơn kiến ​​thức và chuẩn bị cho các đề tài khác trong tương lai.

Sau cùng, nhóm xin kính chúc thầy thật dồi dào sức khỏe, niềm tin để tiếp tục thực hiện sứ mệnh cao đẹp là truyền đạt kiến thức cho các bạn sinh viên.

Xin chân thành cảm ơn Thầy!

**DANH MỤC CÁC BẢNG, HÌNH ẢNH:**

[**Hình 1: Tổng quan về dataset** 7](#_Toc122953174)

[**Hình 2: Thành phân chính của Tensorflow** 11](https://uithcm.sharepoint.com/sites/inTon/Shared%20Documents/General/Cloude_Final.docx#_Toc122953175)

[**Hình 3: Cấu trúc cơ bản của mạng nơ ron nhân tạo ANN** 14](#_Toc122953176)

[**Hình 4: Mô hình ANN đơn giản** 15](#_Toc122953177)

[**Hình 5: Kết nối google drive** 16](#_Toc122953178)

[**Hình 6: Import các thư viện** 16](#_Toc122953179)

[**Hình 7: Đọc dữ liệu** 16](#_Toc122953180)

[**Hình 8: Xóa các cột không liên quan** 17](#_Toc122953181)

[**Hình 9: Kiểm tra dữ liệu Null** 17](#_Toc122953182)

[**Hình 10: Kiểm tra kiểu dữ liệu các cột** 17](#_Toc122953183)

[**Hình 11: Kiểm tra outliers các thuộc tính** 17](#_Toc122953184)

[**Hình 12: Biểu đồ quan hệ giữa Age và Exit** 18](#_Toc122953185)

[**Hình 13: Biểu đồ quan hệ giữa Balance và Exit** 19](#_Toc122953186)

[**Hình 14: Biểu đồ quan hệ giữa Gender và Exit** 19](#_Toc122953187)

[**Hình 15: Biểu đồ quan hệ giữa Geography và Exit** 20](#_Toc122953188)

[**Hình 16: Dùng hot coding chuyển dữ liệu cột Gender và Age thành dạng 1;0** 20](#_Toc122953189)

[**Hình 17: Sử dụng MinMaxScaler để chuyển đổi dữ liệu** 21](#_Toc122953190)

[**Hình 18: Tách dữ liệu và dùng train\_test\_split** 21](#_Toc122953191)

[**Hình 19: Dùng TensorFlow và Keras để format** 21](#_Toc122953192)

**MỤC LỤC:**

[**LỜI CẢM ƠN** 3](#_Toc124884217)

[**PHẦN 1: TỔNG QUAN ĐỀ TÀI** 7](#_Toc124884218)

[**1.** **Lý do chọn đề tài** 7](#_Toc124884219)

[**2.** **Mục tiêu** 7](#_Toc124884220)

[**3.** **Hướng tiếp cận** 7](#_Toc124884221)

[**PHẦN 2: TỔNG QUAN VỀ TẬP DỮ LIỆU** 8](#_Toc124884222)

[**1.** **Giới thiệu tập dữ liệu** 8](#_Toc124884223)

[**2.** **Mô tả tập dữ liệu** 9](#_Toc124884224)

[**PHẦN 3: GIỚI THIỆU VỀ TENSORFLOW** 10](#_Toc124884225)

[**1.** **Khái niệm** 10](#_Toc124884226)

[**2.** **TensorFlow là gì?** 10](#_Toc124884227)

[**3.** **Cách hoạt động** 11](#_Toc124884228)

[**4.** **Các thành phần chính** 12](#_Toc124884229)

[**5.** **Lợi ích từ TensorFlow** 13](#_Toc124884230)

[**6.** **Các thuật toán nổi bật được hỗ trợ bởi TensorFlow** 13](#_Toc124884231)

[**PHẦN 4: GIỚI THIỆU VỀ ARTIFICIAL NEURAL NETWORK** 14](#_Toc124884232)

[**1.** **Artificial Neural Network là gì ?** 14](#_Toc124884233)

[**2.** **Cấu trúc và thành phần của ANN** 14](#_Toc124884234)

[**3.** **Ứng dụng của ANN vào đời sống** 16](#_Toc124884235)

[**PHẦN 5: XỬ LÝ VÀ MÔ HÌNH HÒA TRỰC QUAN DỮ LIỆU** 17](#_Toc124884236)

[**1.** **Kết nối Drive** 17](#_Toc124884237)

[**2.** **Nhập các thư viện** 17](#_Toc124884238)

[**3.** **Đọc dữ liệu** 17](#_Toc124884239)

[**4.** **Xóa các cột không liên quan** 17](#_Toc124884240)

[**5.** **Kiểm tra dữ liệu Null** 18](#_Toc124884241)

[**6.** **Kiểm tra kiểu dữ liệu các thuộc tính** 18](#_Toc124884242)

[**7.** **Kiểm tra outliers các thuộc tính** 18](#_Toc124884243)

[**8.** **Trực quan hóa dữ liệu** 20](#_Toc124884244)

[**8.1.** **Age và Exit** 20](#_Toc124884245)

[**8.2.** **Balance và Exit** 21](#_Toc124884246)

[**8.3.** **Genders và Exit** 21](#_Toc124884247)

[**8.4.** **Geography và Exit** 22](#_Toc124884248)

[**9.** **Sử dụng one hot coding chuyển đổi dữ liệu cho 2 cột Gender và Age** 22](#_Toc124884249)

[**10.** **Sử dụng MinMaxScaler để chuyển đổi dữ liệu cho 4 cột CredictScore, EstimateSalary, Balance, Age.** 23](#_Toc124884250)

[**11.** **Tách dữ liệu vào X,Y và sử dụng train\_test\_plit** 23](#_Toc124884251)

[**12.** **Sử dụng thư viện TensorFlow và Keras để format cho mô hình** 23](#_Toc124884252)

[**PHẦN 6: WEB DEPLOY PREDICTION** 26](#_Toc124884253)

[**1.** **Tổng quan về công nghệ sử dụng** 26](#_Toc124884254)

[**1.1** **Thuật toán sử dụng** 26](#_Toc124884255)

[**1.2** **Các thuộc tính sử dụng** 27](#_Toc124884256)

[**1.3** **Tổng quan về Google Cloud Platform.** 27](#_Toc124884257)

[**2.** **Giao diện và testing** 27](#_Toc124884258)

[**2.1** **Giao diện** 27](#_Toc124884259)

[**2.2** **Testing** 28](#_Toc124884260)

[**3.** **Mã nguồn phần mềm** 29](#_Toc124884261)

[**a.** **Code phần giao diện:** 29](#_Toc124884262)

[**b.** **Code phần xử lý:** 29](#_Toc124884263)

[**PHẦN 7: KẾT LUẬN** 29](#_Toc124884264)

[**1.** **Kết quả đạt được** 29](#_Toc124884265)

[**2.** **So sánh, đánh giá** 29](#_Toc124884266)

[**3.** **Ưu điểm** 30](#_Toc124884267)

[**4.** **Khuyết điểm** 30](#_Toc124884268)

[**5.** **Hướng phát triển** 31](#_Toc124884269)

[**Tài liệu tham khảo** 32](#_Toc124884270)

[**Hết.** 32](#_Toc124884271)

# **PHẦN 1: TỔNG QUAN ĐỀ TÀI**

1. **Lý do chọn đề tài**

Hiện nay các công ty, doanh nghiệp bỏ ra một số tiền lớn để chi trả cho quảng cáo, marketing, khuyến mãi để có thể thu hút được nhiều khách hàng mới tham gia và cũng như là sử dụng dịch vụ hay sản phẩm của công ty. Ví dụ như các ví điện tử (momo, zalopay,..) thì họ luôn có các khuyến mãi kiểu như là : những ai chưa dùng dịch vụ thì mở ví và liên kết ngân hàng sẽ được tặng 200.000đ - 500.000đ. Và số tiền bỏ ra sẽ luôn lớn hơn gấp 4 đến 5 lần so với việc chúng ta bỏ ra để giữ chân khách hàng cũ. Vì vậy việc giữ chân khách hàng hiện tại là cách tối ưu chi phí và hợp lí để có thể duy trì một lượng khách hàng lớn.

Xác định được vấn đề chính, nhóm nhận thấy việc phân tích khách hàng nào sẽ rời bỏ dịch vụ là cần thiết. Từ các phân tích này các các công ty sẽ có thể dự đoán và biết được khách hàng nào sắp từ bỏ dịch vụ từ đó có những biện pháp chăm sóc, khuyến mãi, ưu đãi nhằm giữ chân khách hàng lại với công ty.

1. **Mục tiêu**

* Hiểu và nắm bắt được dữ liệu. Từ đó có thể xử lý được dữ liệu nhằm phục vụ cho
* yêu cầu đồ án.
* Biết cách sử dụng thư viện TensorFlow và xây dựng được mô hình dự đoán với độ
* chính xác lên tới 85%.
* So sánh và đánh giá sự hiệu quả giữa các mô hình.
* Khám phá, phân tích và trực quan hóa tập dữ liệu.

1. **Hướng tiếp cận**

Để hoàn thiện dự án này em sẽ thực hiện theo các bước như sau:

* Đầu tiên, xác định rõ mục tiêu trước và sau của dự án.
* Tiếp theo, tiến hành tìm nguồn và thu thập dữ liệu.
* Sau khi đã thu thập được dữ liệu, thực hiện chọn và xử lý dữ liệu chuyển dữ liệu thô thành dữ liệu có giá trị đối với bài toán. Cụ thể em sẽ tiến hành tìm kiếm dữ liệu bị lỗi, dữ liệu bị thiếu hoặc ngoại lệ sau đó xử lí chúng.
* Dữ liệu sau khi được xử lí và làm sạch, tiến hành phân tích và khám phá dữ liệu bằng cách trực quan hóa, biểu diễn mối liên hệ giữa mức lương với các thuộc tính khác.
* Sau đó, tiến hành xây dựng thuật toán và chỉ số đánh giá mô hình dự đoán mức lương trong lĩnh vực Công nghệ thông tin.
* Cuối cùng, so sánh kết quả, tính chỉnh xác giữa các mô hình và tổng kết dự án.

# **PHẦN 2: TỔNG QUAN VỀ TẬP DỮ LIỆU**

1. **Giới thiệu tập dữ liệu**

Để có thể dự đoán được khách hàng nào có nguy cơ yêu cầu xoá tài khoản của họ tại ngân hàng, nhóm sẽ lấy dữ liệu từ trên trang chủ Kaggle. Đây là tập dữ liệu về Dự đoán sự rời đi của khách hàng.

Tên tập dữ liệu: Customer Churn Prediction ANN, kích thước: 684.86 kB.

Tên file csv sử dụng: Churn Modeling.csv, kích thước: 684.86 kB.

Nguồn: [Churn Modeling Dataset](https://www.kaggle.com/code/rithinashikh/customer-churn-prediction-ann/data?select=Churn+Modeling.csv)

Chi tiết tập dữ liệuCột “Class” là dữ liệu dự đoán gian đoạn. Với 2 kết quả 0 và 1.

* 0 thể hiện không có sự gian lận qua các giao dịch.
* 1 thể hiện có sự gian lận qua các giao dịch.

Table

Description automatically generated

**Hình 1: Tổng quan về dataset**

1. **Mô tả tập dữ liệu**

Tập dữ liệu Churn Modeling.csv bao gồm 14 cột dữ liệu với 2931 dòng dữ liệu với các đặc trưng như sau:

**Nhóm thuộc tính thông tin của khách hàng:**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Ý nghĩa | Kiểu dữ liệu | Miền giá trị |
| CustomerId | Id của khách hàng trong tập dữ liệu. | Numeric | 350-850 |
| Surname | Họ của khách hàng. | String |  |
| Geography | Tỉnh thành hoặc khu vực khách hàng sinh sống. | String | 3 |
| Gender | Giới tính. | String | Female/Male |
| Age | Độ tuổi. | Numeric | 18-92 |
| EstimatedSalary | Mức lương ước tính. | Numeric | 11.6-200.000 |

**Nhóm thuộc tính dữ liệu ngân hàng của khách hàng:**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Ý nghĩa | Kiểu dữ liệu | Miền giá trị |
| CreditScore | Điểm tín dụng của khách hàng. | Numeric | 350-850 |
| Tenure | Số năm khách hàng sử dụng dịch vụ. | Numeric | 0-10 |
| Balance | Tỉnh thành hoặc khu vực khách hàng sinh sống. | Numeric | 0-251.000 |
| NumOfProducts | Số lượng sản phẩm tiện ích của ngân hàng mà khách hàng đang sử dụng. | Numeric | 1-4 |
| HasCrCard | Khách hàng có thẻ tín dụng không ? | Binary | 1: có  0: không |
| IsActiveMember | Khách hàng có còn hoạt động không ? | Binary | 1: có  0: không |
| Exited | Khách hàng rời bỏ ? | Binary | 1: có  0: không |

**Các thuộc tính khác:**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Ý nghĩa | Kiểu dữ liệu | Miền giá trị |
| RowNumber | Số lượng dữ liệu trong tập dữ liệu | Numeric | 1-10.000 |

# **PHẦN 3: GIỚI THIỆU VỀ TENSORFLOW**

1. **Khái niệm**

Sự phát triển của trí tuệ nhân tạo dẫn đến việc tìm hiểu về Machine Learning và Deep Learning đã trở thanh xu thế hiện nay.

Việc sử dụng các thư viện có sẵn để tính toán đã giúp việc tiếp cận các bài toán trở nên đơn giản hơn. Vì thế mà Google đã sử dụng Deep Neural Network để cả thiện các dịch vụ của mình, trong đó có gmail, hình ảnh, google search engine.

TensorFlow được phát triển bởi đội Google Brain. Đây ra một thư viện phần mềm mã nguồn mở và ra mắt lần đầu vào cuối năm 2015 và mãi đến năm 2017 nó mới có phiên bản ổn định.

1. **TensorFlow là gì?**

TensorFlow được biết đến như là thư viện mã nguồn mở, được ứng dụng để tính toán Machine Learning trên quy mô lớn. Thư viện này kết hợp nhiều mô hình, thuật toán Machine Learning và Deep Learning.

Từ các hàm được dựng sẵn trong thư viện cho từng bài toán, TensorFlow có thể xây dựng nhiều Neural Network khác nhau. Đồng thời, các hàm này còn giúp người dùng tính toán song song trên các máy tính hay thậm chí là trên nhiều CPU, GPU trên cùng một máy. Ngoài ra, bạn còn có thể tạo ra các đồ thị luồng dữ liệu để dựng lên các model (data flow graph) từ các hàm của Tensorflow.

Phần hiệu năng của Tensorflow cực kỳ tốt do được viết từ ngôn ngữ lập trình C++ và thao tác Interface bằng Python nhanh nhạy. Đối tượng sử dụng của Tensorflow cũng cực kỳ đa dạng, từ các nhà nghiên cứu cho tới các nhà khoa học dữ liệu và lập trình viên hiện nay.

1. **Cách hoạt động**

Tensorflow cho phép bạn xây dựng cấu trúc và biểu đồ luồng dữ liệu để miêu tả cách mà data di chuyển qua một seri các node đang xử lý hoặc qua biểu đồ. Mỗi node trong đồ thị lại đại diện cho một operation toán học, hay còn được gọi là tensor hay dữ liệu đa chiều.

Kiến trúc hoạt động của Tensorflow được chia thành 3 phần chính là:

* Tiền xử lý dữ liệu.
* Dựng model.
* Train và ước tính model cụ thể.

Bằng ngôn ngữ lập trình Python, Tensorflow sẽ cung cấp toàn bộ thông tin cho các lập trình viên. Python sẽ điều phối các luồng công việc cũng như kết nối chúng lại, node và tensor trong Tensorflow là các đối tượng của Python. Chính nhờ sự hỗ trợ đắc lực từ Python, quá trình phát triển phần mềm của lập trình viên được đơn giản hoá đi rất nhiều.

1. **Diagram, schematic

   Description automatically generatedCác thành phần chính**

**Hình 2: Thành phân chính của Tensorflow**

Gồm 2 thành phần chính là Tensor và Graph

1. Tensor

Tensor: là framework cốt lõi của TensorFlow. Toàn bộ các phép tính toán đều liên quan tới Tensor. Mỗi Tensor là một vector hoặc ma trận của “n” chiều không gian đại diện cho toàn bộ dữ liệu.

Các giá trị trong cùng một tensor đều chứa một loại dữ liệu giống hệt nhau với một shape (chiều của ma trận hay mảng) dữ liệu đã biết.

Mỗi tensor có thể được bắt nguồn từ dữ liệu Input hay một kết quả tính toán. Phần edge của các node chính là tensor, một cách để nhập operation với dữ liệu này.

1. Graph

Bên trong Tensorflow toàn bộ hoạt động được tiến hành trong một biểu đồ (graph). Mỗi biểu đồ là một tập hợp tính toán liên tiếp, mỗi operation được gọi là một op node và được kết nối lại với nhau.

Graph là một thành phần quan trọng của Tensorflow, tập hợp và mô tả toàn bộ chuỗi tính toán được thực hiện trong quá trình training. Nó mang đến nhiều lợi thế như:

* Chạy trên nhiều GPU hoặc CPU ngay cả với hệ điều hành trên điện thoại.
* Bảo toàn các tính toán quan trọng ngay sau đó nhờ tính năng di động của biểu đồ. Đồng thời, biểu đồ này cũng có thể được lưu trữ và sử dụng lại trong tương lai.
* Tất cả tính toán trong biểu đồ đều được thực hiện bằng cách kết nối các tensor lại với nhau, bao gồm 1 tensor, 1 node và 1 edge. Node mang operation toán học và sản xuất các output ở đầu cuối, còn các edge giải thích mối quan hệ input/output giữa các node.

1. **Lợi ích từ TensorFlow**

Lợi ích dễ thấy nhưng quan trọng nhất mà TensorFlow cung cấp cho việc lập trình machine learning chính là abstraction. Thay vì phải đối phó với những tình huống rườm rà từ việc thực hiện triển khai các thuật toán, hay tìm ra cách hợp lý để chuyển output của 1 chức năng sang input của 1 chức năng khác, giờ đây bạn có thể tập trung vào phần logic tổng thể của 1 ứng dụng hơn. TensorFlow sẽ chăm sóc phần còn lại thay cho bạn.

Ngoài ra TensorFlow còn ung cấp các tiện ích bổ sung cho các lập trình viên cần debug cũng như giúp bạn tự suy xét các ứng dụng TensorFlow. Chế độ Eagerexecution cho phép bạn đánh giá và sửa đổi từng operation của biểu đồ 1 cách riêng biệt và minh bạch, thay vì phải dựng toàn bộ biểu đồ dưới dạng 1 đối tượng độc lập vốn khá mơ hồ hay phải đánh giá chung tổng thể. Cuối cùng, 1 tính năng khá độc đáo của TensorFlow là TensorBoard. TensorBoard cho phép bạn quan sát 1 cách trực quan những gì TensorFlow đang làm.

TensorFlow còn có nhiều cải tiến từ sự hậu thuẫn từ các ekíp thương mại hạng A tại Google. Google không những tiếp lửa cho tiến độ nhanh chóng cho sự phát triển đằng sau dự án, mà còn tạo ra nhiều phục vụ độc đáo xung quanh TensorFlow để nó dễ dàng deploy và sử dụng: như silicon TPU mình đã nói ở trên để tăng tốc hiệu suất đám mây Google, 1 online hub cho việc chia sẻ các model được tạo với framework, sự hiện diện của in-browser và gần gũi với mobile của framework, và nhiều hơn thế nữa…

1. **Các thuật toán nổi bật được hỗ trợ bởi TensorFlow**

* **Linear regression:** tf.estimator.LinearRegression.
* **Classification:** tf.estimator.LinearClassifier.
* **Deep learning classification:** tf.estimator.DNNClassifier.
* **Deep learning wipe and deep:** tf.estimator.DNNLinearCombinedClassifier.
* **Boosted tree regression:** tf.estimator.BoostedTreesRegressor.
* **Bossted tree classification:** tf.estimator.BoostedTreesClassifier

# **PHẦN 4: GIỚI THIỆU VỀ ARTIFICIAL NEURAL NETWORK**

1. **Artificial Neural Network là gì ?**

ANN là mạng neural nhân tạo và là mô hình toán học hoặc mô hình toán được xây dựng thông qua các neural sinh học. Nó bao gồm các nhóm việc, neural nhân tạo có thể nối hút với nhau và xử lý các thông tin bằng biện pháp truyền theo các kết nối rồi tính toán giá trị mới tại các nút.

Ở nhiều trường hợp, mạng Artificial Neural Network là hệ thống thích ứng có thể tự thay đổi cấu trúc của mình dựa vào thông tin bên ngoài hoặc bên trong chảy qua mạng ở quá trình học. Nhiều mạng Artificial Neural Network còn là công cụ giúp mô hình hóa dữ liệu thống kê phi tuyến. Chúng còn được sử dụng để mô hình hóa cho các mối quan hệ có tính phức tạp giữa các dữ liệu vào hoặc giữa kết quả để kiếm tìm mẫu trong dữ liệu.

1. **Cấu trúc và thành phần của ANN**

ANN lấy ý tưởng từ cách hoạt động của bộ não con người - tạo ra các kết nối phù hợp. Do đó, ANN đã sử dụng các silicon và dây điện để làm nơ ron và đuôi gai sống cho mình.Trong cơ thể con người, 1 phần não đã bao gồm 86 tỉ tế bào thần kinh và chúng được kết nối với hàng nghìn tế bào khác thông qua Axons. Bởi vì con người có rất nhiều đầu vào thông tin khác nhau từ các giác quan, nên cơ thể cũng có nhiều đuôi gai để giúp truyền thông tin này.Chúng sẽ tạo ra xung điện để di chuyển, truyền thông tin trong mạng lưới nơ ron thần kinh này. Và điều này cũng tương tự cho mạng nơ ron nhân tạo ANN - Khi cần xử lý các vấn đề khác nhau, nơ-ron sẽ gửi một thông điệp đến một nơ-ron khác.

Diagram

Description automatically generated

**Hình 3: Cấu trúc cơ bản của mạng nơ ron nhân tạo ANN**

Do đó, chúng ta có thể nói rằng ANN sẽ bao gồm nhiều nút bên trong, chúng bắt chước các tế bào thần kinh sinh học bên trong não người. Các mạng ANN sẽ kết nối các nơ-ron này bằng các liên kết và chúng có tương tác với nhau.

Các nút trong ANN được sử dụng để lấy dữ liệu đầu vào. Hơn nữa, việc thực hiện các thao tác trên dữ liệu cũng rất đơn giản. Sau khi thực hiện những thao tác với dữ liệu, các hoạt động này được chuyển cho các tế bào thần kinh khác. Đầu ra tại mỗi nút được gọi là giá trị kích hoạt hoặc giá trị nút của nó.

Mỗi liên kết trong mạng ANN đều có liên quan với trọng lượng. Ngoài ra, chúng có khả năng học hỏi. Điều đó sẽ diễn ra bằng cách thay đổi các giá trị trọng lượng. Dưới đây là một hình minh họa về một ANN đơn giản:

**Chart, diagram, bubble chart

Description automatically generated**

**Hình 4: Mô hình ANN đơn giản**

Artificial Neural Network (ANN) gồm 3 thành phần chính: Input layer và Output layer chỉ gồm 1 layer , Hidden layer có thể có 1 hay nhiều layer tùy vào bài toán cụ thể. ANN hoạt động theo hướng mô tả lại cách hoạt động của hệ thần kinh với các neuron được kết nối với nhau.

Trong ANN, trừ Input layer thì tất cả các node thuộc các layer khác đều full-connected với các node thuộc layer trước nó. Mỗi node thuộc Hidden layer nhận vào ma trận đầu vào từ layer trước và kết hợp với trọng số để ra được kết quả.

1. **Ứng dụng của ANN vào đời sống**

* Trong không gian vũ trụ: sử dụng ANN cho máy bay lái tự động, dùng để phát hiện các lỗi còn sót trên máy bay.
* Trong quân sự: định hướng và sử dụng vũ khí, theo dõi mục tiêu.
* Trong thiết bị điện tử: dự đoán chuỗi mã, cách bố trí chip IC và phân tích lỗi chip.
* Trong y khoa: phân tích tế bào ung thư, phân tích điện tâm đồ.
* Trong vận tải: xây dựng hệ thống chuẩn đoán phan xe tải hoặc hệ thống lập trình cho xe di chuyển.
* Giọng nói: nhận dạng giọng nói và phân loại giọng nói

# **PHẦN 5: XỬ LÝ VÀ MÔ HÌNH HÒA TRỰC QUAN DỮ LIỆU**

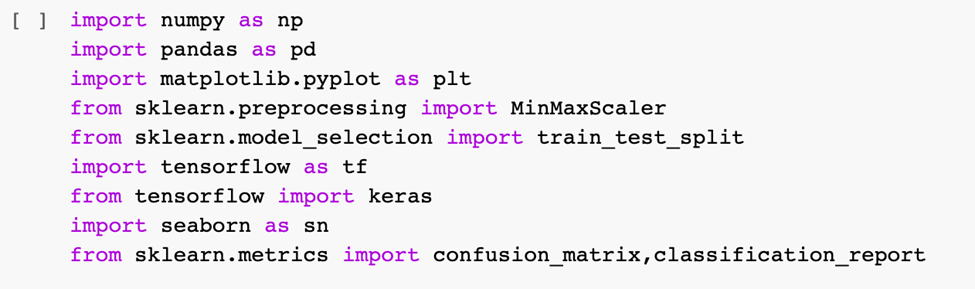
1. **Kết nối Drive**

Graphical user interface, text, application

Description automatically generated

**Hình 5: Kết nối google drive**

1. **Nhập các thư viện**



**Hình 6: Import các thư viện**

1. **Đọc dữ liệu**

A picture containing table

Description automatically generated

**Hình 7: Đọc dữ liệu**

1. **Xóa các cột không liên quan**

A picture containing graphical user interface

Description automatically generated

**Hình 8: Xóa các cột không liên quan**

1. **Kiểm tra dữ liệu Null**

Graphical user interface, application

Description automatically generated

**Hình 9: Kiểm tra dữ liệu Null**

1. **Kiểm tra kiểu dữ liệu các thuộc tính**

8.Table

Description automatically generated

**Hình 10: Kiểm tra kiểu dữ liệu các cột**

1. **Kiểm tra outliers các thuộc tính**

Graphical user interface, text, application, email

Description automatically generated

Diagram

Description automatically generated

**Hình 11: Kiểm tra outliers các thuộc tính**

1. **Trực quan hóa dữ liệu**
   1. **Age và Exit**

Chart

Description automatically generated

**Hình 12: Biểu đồ quan hệ giữa Age và Exit**

**Nhận xét:** Có thể thấy số lượng khách hàng trong tầm 30 đén 45 tuổi là những khách hàng sẽ tiếp tục sử dụng dịch vụ cao nhất. Nhưng bên cạnh đó đây cũng là tệp khách có số lượng rời bỏ cao nhất => có thể do độ tuổi này họ muốn cố định về việc sử dụng tài khoản ngân hàng do đã có công việc ổn định hoặc cũng có thể do công ty yêu cầu họ sử dụng.

* 1. **Balance và Exit**

Chart, bar chart

Description automatically generated

**Hình 13: Biểu đồ quan hệ giữa Balance và Exit**

**Nhận xét:** Có thể thấy số lượng khách hàng có số dư tài khoản thấp từ khoản 0 đến 250000 là những khách hàng sẽ tiếp tục sử dụng dịch vụ cao nhất => đây có thẻ là những người vừa mới tham gia dịch vụ.

* 1. **Genders và Exit**

**Chart

Description automatically generated**

**Hình 14: Biểu đồ quan hệ giữa Gender và Exit**

**Nhận xét:** Có thể thấy rằng tỉ lệ khách hàng nữ vẫn còn sử dụng các dịch vụ cao hơn khách hàng nam. Và ở chiều ngược lại thì tỉ lệ khách hàng nam rút khỏi các dịch vụ cao hơn khách hàng nữ.

* 1. **Geography và Exit**

**Chart

Description automatically generated**

**Hình 15: Biểu đồ quan hệ giữa Geography và Exit**

**Nhận xét:** Có thể thấy rằng ở Pháp có số lượng khách hàng sẽ sử dụng dịch vụ lâu hơn so với 2 nước còn lại. Có thể do người dân ở đây không quá khắt khe về việc chất lượng dịch vụ.

1. **Sử dụng one hot coding chuyển đổi dữ liệu cho 2 cột Gender và Age**

**Table

Description automatically generated**

**Hình 16: Dùng hot coding chuyển dữ liệu cột Gender và Age thành dạng 1;0**

1. **Sử dụng MinMaxScaler để chuyển đổi dữ liệu cho 4 cột CredictScore, EstimateSalary, Balance, Age.**

**Text

Description automatically generated**

**Hình 17: Sử dụng MinMaxScaler để chuyển đổi dữ liệu**

1. **Tách dữ liệu vào X,Y và sử dụng train\_test\_plit**

**Graphical user interface, text, application

Description automatically generated**

**Hình 18: Tách dữ liệu và dùng train\_test\_split**

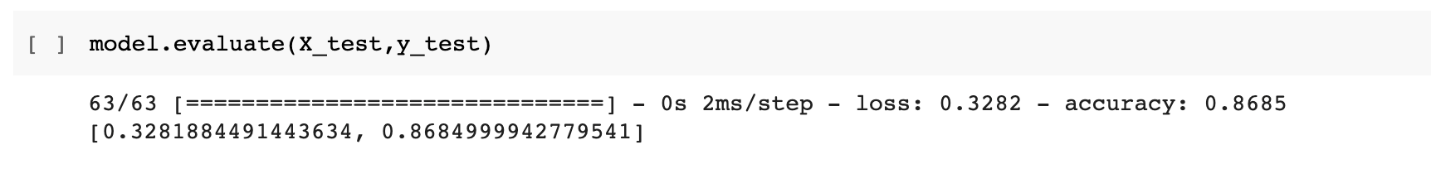
1. **Sử dụng thư viện TensorFlow và Keras để format cho mô hình**

Graphical user interface, text

Description automatically generated

**Hình 19: Dùng TensorFlow và Keras để format**

1. **Training Model**



**Hình 20: Training model**

* **Độ chính xác khoảng 86%**

1. **Dự đoán trên Test Set và sử dụng Ma trận nhầm lẫn và Báo cáo phân loại để đánh giá hiệu suất của mô hình**

Text

Description automatically generated

1. **Các tiêu chí đánh giá**

Table

Description automatically generated

Recall: Cho biết có bao nhiêu điểm dữ liệu thực sự ở lớp Positive được mô hình phân lớp đúng trong mọi điểm dữ liệu thực sự ở lớp Positive.

Text

Description automatically generated

Precision: Cho biết trong số các điểm dữ liệu được mô hình phân loại hay dự đoán vào lớp Positive có bao nhiêu điểm thực sự thuộc về lớp Positive

Text

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with low confidence

* Precision và Recall có giá trị trong khoảng [0,1], hai giá trị này càng gần với 1 thì mô hình càng chính xác.



Là một kỹ thuật đánh giá hiệu năng của mô hình dự đoán cho các bài toán phân lớp thể hiện số lượng điểm dữ liệu thuộc vào 1 lớp và được dự đoán thuộc về 1 lớp. Confusion Matrix cung cấp thêm thông tin về tỉ lệ phân lớp đúng giữa các lớp, hay giúp phát hiện các lớp có tỉ lệ phân lớp nhầm cao nhờ vào các khái niệm True (False), Positive (Negative).

* **True Positive (TP)**: đối tượng ở lớp Positive, mô hình phân đối tượng vào lớp Positive (dự đoán đúng).
* **True Negative (TN)**: đối tượng ở lớp Negative, mô hình phân đối tượng vào lớp Negative (dự đoán đúng).
* **False Positive (FP)**: đối tượng ở lớp Negative, mô hình phân đối tượng vào lớp Positive (dự đoán sai) – Type I Error.
* **False Negative (FN)**: đối tượng ở lớp Positive, mô hình phân đối tượng vào lớp Negative (dự đoán sai) – Type II Error.

Chart, treemap chart

Description automatically generated

# **PHẦN 6: WEB DEPLOY PREDICTION**

* 1. **Tổng quan về** **công nghệ sử dụng**
  2. **. Thuật toán sử dụng**

**Dựa trên kết quả thu được ở phần trước, nhóm sử dụng thuật ANN cho phần mềm này. Theo kết quả, thuật toán này có độ chính xác ổn.**

* 1. **. Các thuộc tính sử dụng**

Text

Description automatically generated

* Sau các bước trước, nhóm thu được 5 thuộc tính: CreditScore, Tenure, HasCrCard, IsActiveMember, EstimatedSalary là những thuộc tính đầu vào và thuộc tính Exited là thuộc tính dự đoán để đưa vào deploy.
  1. **.** **Tổng quan về Google Cloud Platform.**

****

**Google Cloud hay còn gọi là Google Cloud Platform (GCP)** là một nền tảng của kỹ thuật điện toán đám mây cho phép các cá nhân, tổ chức, các doanh nghiệp, các cơ quan có thể xây dựng, phát triển, và hoạt động các ứng dụng của mình trên hệ thống phần mềm do google tạo ra. Các ứng dụng rất phổ biến hiện nay được mọi người sử dụng rất nhiều như: Trình duyệt Chrome, ứng dụng bản đồ Google Map, Google Apps, kênh Youtube…

Google Cloud cung cấp tất cả các giải pháp quản lý cho doanh nghiệp, giúp doanh nghiệp có thể phát triển hệ thống công nghệ của mình một cách chính xác, hiện đại. Bên cạnh đó, GC còn giúp người dùng và doanh nghiệp giải quyết các vấn đề như: Developer (phát triển), Management (Quản lý), Computer Engine, Mobile, Storage, Big Data…

Một điểm khác biệt nữa mà GC mang lại so với các dịch vụ đám mây khác đó chính là hệ thống DataCenter luôn ổn định và có độ bảo mật dữ liệu cực cao, giúp bảo vệ dữ liệu người dùng và khách hàng trước sự dòm ngó và xâm nhập trái phép của các hacker công nghệ.

*Google Cloud cung cấp những sản phẩm chính sau đây:*

* **Big Data:** BigQuery, Cloud Dataproc, Cloud Dataflow…
* **Services:** Translate API, Prediction API…
* **Storage**: Cloud Storage, Cloud Datastore, Cloud SQL…
* **Compute**: App Engine, Compute Engine, …

Những dịch vụ Google Cloud cấp cao:

* **Dịch vụ IoT (Internet of things hay còn gọi là internet vạn vật):** đây chính là dịch vụ cho phép người dùng có thể dễ dàng quản lý, sử dụng dữ liệu từ những thiết bị IoT.
* **Dịch vụ Cloud Machine Learning Engine (máy tìm kiếm đám mây):** được sử dụng để phát triển ứng dụng AI – trí tuệ nhân tạo. AI đã cho thấy sức mạnh vượt trội của mình trong những năm gần đây, AI đã giúp con người xử lý những dữ liệu và thông tin khổng lồ dễ dàng và đơn giản hơn bao giờ hết.
* **Dịch vụ Hadoop và Apache Spark:** bao gồm cả Google Cloud Dataproc: giúp người dùng quản lý dữ liệu có hệ thống, bảo mật, nhanh hơn và an toàn hơn.
* **Dịch vụ Google Big Query:**đây là dịch vụ giúp xử lý và phân tích các tệp dữ liệu cực kỳ lớn. Nó hoạt động với chức năng truy vấn tương tự cơ sở dữ liệu SQL truyền thống. Những tệp dữ liệu mà nó xử lý có thể lên đến hàng trăm triệu terabyte.

**VPS được viết tắt của cụm từ Virtual Private Server**. Hiểu một cách đơn giản nhất, đây chính là máy chủ ảo, được tạo ra từ việc phân chia một máy chủ vật lý thành nhiều máy chủ khác nhau có tính năng và hoạt động giống như một máy chủ riêng biệt.

Cách thức hoạt động của nó dựa trên việc chia sẻ tài nguyên từ máy chủ vật lý ban đầu. Mỗi VPS là một hệ thống riêng, có CPU xử lý riêng, có dung lượng RAM và ổ cứng lưu trữ tách biệt, người dùng có toàn quyền quản lý root và cập nhật, restart hệ thống bất cứ lúc nào họ muốn.

Graphical user interface, application

Description automatically generated

* 1. **Các bước deploy machine learning model trên GCP**

Bước 1: đăng nhập và vào GCP console

Graphical user interface

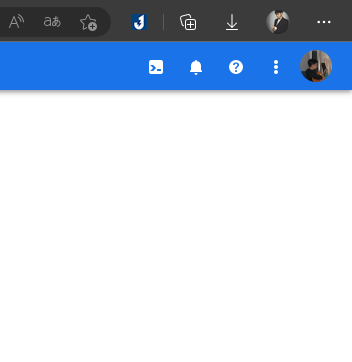
Description automatically generated

Bước 2: Tạo một project để tiến hành deploy

Graphical user interface, application

Description automatically generated

Bước 3: vào Activate cloud shell để tiến hành deploy.



Bước 4: chọn Open Editor góc bên phải.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, Word

Description automatically generated

Các thư mục GCDeploy bao gồm giao diện trang web và các file deploy.

Graphical user interface, text, application

Description automatically generated

Chọn lại Open terminal để tiến hành build web app sử dụng công nghệ App Engine.

Graphical user interface, text, application

Description automatically generated

Câu lệnh gcloud init để khởi tạo.

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Câu lệnh gcloud deploy app.yaml –project dùng để deploy project hiện tại

Text

Description automatically generated

Project build thành công.

Text

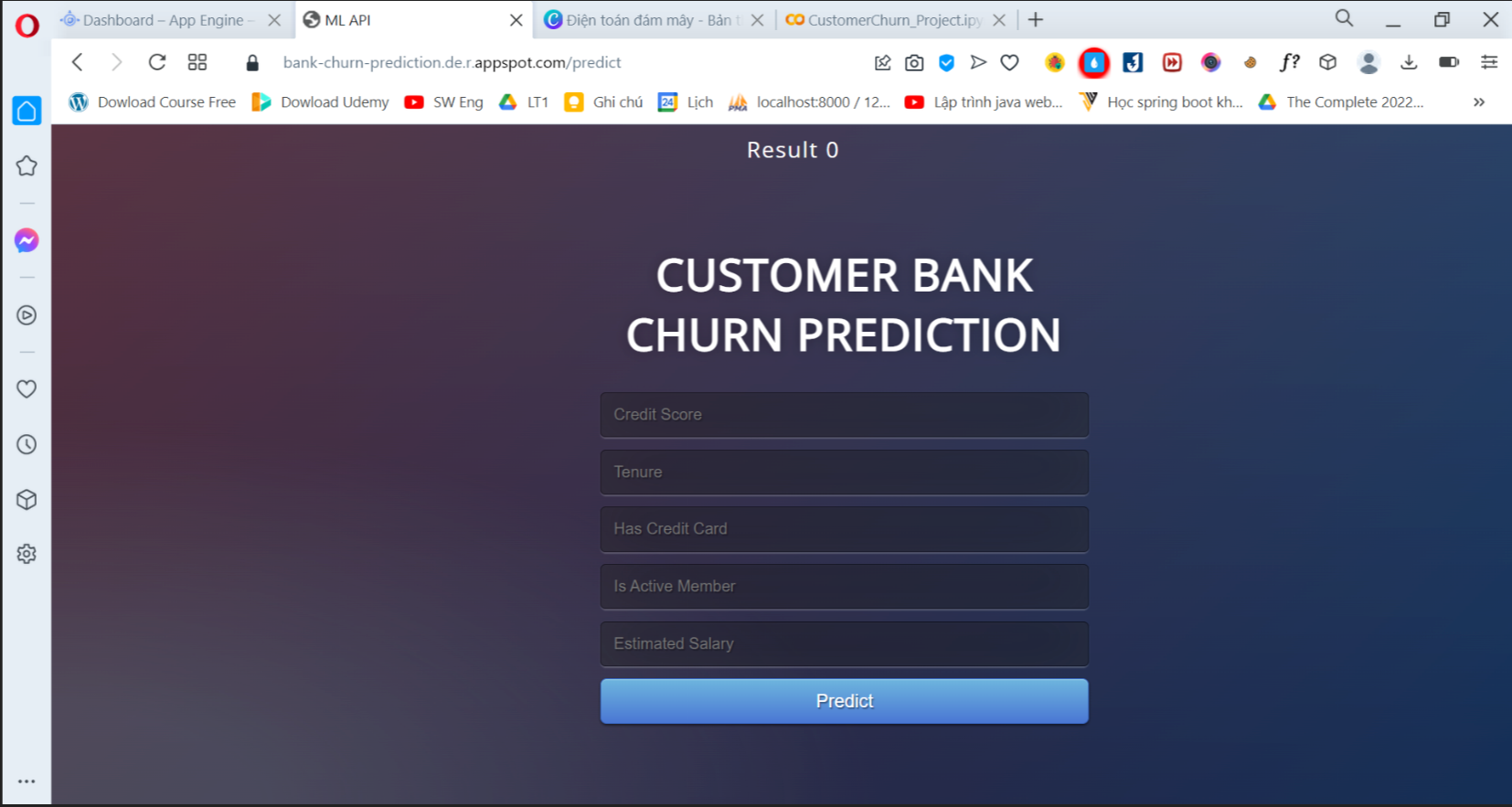
Description automatically generated

Câu lệnh gcloud app browse để mở trang web prediction.

Text

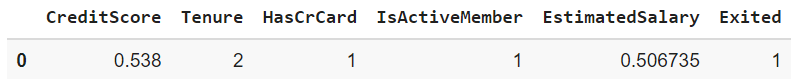
Description automatically generated

* 1. **Giao diện và testing**
  2. **. Giao diện**



* 1. **.** **Testing**

Test dataset:



Kết quả:

Graphical user interface, text, application

Description automatically generated

* 1. **Mã nguồn phần mềm**

**Graphical user interface, text, application

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

Phần mềm bao gồm:

* File app.yaml dùng để cấu hình ngôn ngữ và cấu hình deploy của ứng dụng.
* File ANN.pkl để khởi tạo và chạy thuật toán
* File main.py chứa code python để thực hiện các bước tiền xử lý và áp dụng thuật toán
* Static và template là code giao diện của trang dự đoán (html, javascript và css)
  1. **Code phần giao diện:**

**Home:**

<!DOCTYPE html>

<html >

<!--From https://codepen.io/frytyler/pen/EGdtg-->

<head>

  <meta charset="UTF-8">

  <title>ML API</title>

  <link href='https://fonts.googleapis.com/css?family=Pacifico' rel='stylesheet' type='text/css'>

<link href='https://fonts.googleapis.com/css?family=Arimo' rel='stylesheet' type='text/css'>

<link href='https://fonts.googleapis.com/css?family=Hind:300' rel='stylesheet' type='text/css'>

<link href='https://fonts.googleapis.com/css?family=Open+Sans+Condensed:300' rel='stylesheet' type='text/css'>

<link rel="stylesheet" href="{{ url\_for('static', filename='css/style.css') }}">

</head>

<body>

 <div class="login">

  <h1>CUSTOMER BANK CHURN PREDICTION</h1>

     <!-- Main Input For Receiving Query to our ML -->

    <form action="{{ url\_for('predict')}}"method="post">

      <input type="text" name="interview\_score3" placeholder="Credit Score" required="required" /><br>

      <input type="text" name="experience" placeholder="Tenure" required="required" /><br>

        <input type="text" name="test\_score" placeholder="Has Credit Card" required="required" /><br>

        <input type="text" name="interview\_score5" placeholder="Is Active Member" required="required" /><br>

        <input type="text" name="interview\_score6" placeholder="Estimated Salary" required="required" /><br>

        <button type="submit" class="btn btn-primary btn-block btn-large">Predict</button>

    </form>

   <br>

   <br>

 </div>

 {{prediction\_text}}

</body>

</html>

**Result:**

<!DOCTYPE html>

<html>

<head>

<title> CUSTOMER BANK CHURN PREDICTION </title>

</head>

<body>

<h1> Results </h1>

</body>

</html>

* 1. **Code phần xử lý:**

**Main.py:**

import numpy as np

from flask import Flask, request, jsonify, render\_template, url\_for

import pickle

app = Flask(\_\_name\_\_)

model = pickle.load(open('12\_19521694.pkl','rb'))

@app.route('/')

def home():

    #return 'Hello World'

    return render\_template('home.html')

    #return render\_template('index.html')

@app.route('/predict',methods = ['POST'])

def predict():

    int\_features = [float(x) for x in request.form.values()]

    final\_features = [np.array(int\_features)]

    prediction = model.predict(final\_features)

    print(prediction[0])

    #output = round(prediction[0], 2)

    return render\_template('home.html', prediction\_text="Result {}".format(prediction[0]))

@app.route('/predict\_api',methods=['POST'])

def predict\_api():

    '''

    For direct API calls trought request

    '''

    data = request.get\_json(force=True)

    prediction = model.predict([np.array(list(data.values()))])

    output = prediction[0]

    return jsonify(output)

if \_\_name\_\_ == '\_main\_':

    app.run(host="0.0.0.0", port=8080, debug=True)

# **PHẦN 7: KẾT LUẬN**

1. **Kết quả đạt được**

Kết quả mà nhóm đạt được sau khi hoàn thành đồ án:

* Hiểu được vấn đề và cách giải quyết bài toán phân loại.
* Áp dụng được các phương pháp để tiền xử xử lý dữ liệu.
* So sánh và đánh giá được kết quả các mô hình.
* Tối ưu hóa được các thuật toán sử dụng
* Đưa ví dụ thực tế vào trong bài toán.

1. **So sánh, đánh giá**

|  |
| --- |
| **Thuật toán Artificial Neutral Network** |
| Giá trị Precision = 0.46 và Recall = 0.75 của mô hình còn thấp do đó độ chính xác của mô hình dự đoán Artificial Neutral Network tạm được. |
| Có 1529 khách hàng được dự đoán không có khả năng rời bỏ sử dụng dịch vụ đúng với thực tế. |
| Có 184 khách hàng được dự đoán có khả năng rời bỏ sử dụng dịch vụ đúng với thực tế. |
| Có 216 khách hàng thực tế có khả năng rời bỏ sử dụng dịch vụ nhưng mô hình dự báo đưa ra không có khả năng. (FN) (Type II Error).  Có 62 khách hàng thực tế không rời bỏ sử dụng dịch vụ nhưng mô hình dự báo kết luận có khả năng. (FP) (Type I Error). |

**Nhận xét:**

Thuật toán ANN cho kết quả dự đoán chưa tối ưu.

1. **Ưu điểm**

Những ưu điểm của đồ án:

* Áp dụng được nhiều phương pháp so sánh và đánh giá mô hình.
* Ứng dụng được ví dụ thực tế vào trong đồ án là dự đoán người dùng rời bỏ sử dụng dịch vụ .

1. **Khuyết điểm**

Vì thời gian để nghiên cứu và hiện thực đề tài còn giới hạn do đó còn tồn tại một số hạn chế, như sau:

* Độ chính xác thuật toán không cao.
* Thời gian thực thi thuận toán chậm.
* Chưa nghiên cứu tìm ra các thuật toán tối ưu hơn

1. **Hướng phát triển**

Nhóm sẽ tiếp tục phát triển đồ án để:

* Ứng dụng hệ thống vào các phần mềm chuẩn đoán trong kinh tế.
* Sử dụng nhiều tập dataset để đánh giá hiệu năng của mô hình.
* Tiếp tục nguyên cứu tìm ra các thuật toán mới tối ưu tốt hơn.
* Nghiên cứu sử dụng thêmnhiều thuật toán khác để đánh giá và so sánh từ đó tối ưu kết quả đạt được.
* Tiếp tục tối ưu các thuật toán sử dụng để đưa ra kết quả chính xác và thời gian nhanh hơn.
* Kết hợp các thuật toán lại thành một mô hình hoàn chỉnh.

# **Tài liệu tham khảo**

[1]. Link Dataset “Customer Churn Prediction ANN” - [Customer Churn Prediction ANN | Kaggle](https://www.kaggle.com/code/rithinashikh/customer-churn-prediction-ann/data?select=Churn+Modeling.csv)

[2]. “Artificial Neutral Network” - [Tổng quan về Artificial Neural Network (viblo.asia)](https://viblo.asia/p/tong-quan-ve-artificial-neural-network-1VgZvwYrlAw)

[3]. “ANN - Wikipedia” - [Mạng thần kinh nhân tạo – Wikipedia tiếng Việt](https://vi.wikipedia.org/wiki/M%E1%BA%A1ng_th%E1%BA%A7n_kinh_nh%C3%A2n_t%E1%BA%A1o)

[4]. “Các phương pháp đánh giá mô hình dự báo”- [Các phương pháp đánh giá mô hình học máy, học sâu (Machine learning & Deep learning) (rabiloo.com)](https://rabiloo.com/vi/blog/cac-phuong-phap-danh-gia-mo-hinh-machine-learning-va-deep-learning)

[5]. “Confusion matrix” - [Confusion Matrix / Ma trận nhầm lẫn / Ma trận lỗi - Viblo](https://viblo.asia/p/confusion-matrix-ma-tran-nham-lan-ma-tran-loi-p1-V3m5WQB7ZO7)

[6] Tổng quan về GCP - [Google Cloud Là Gì? Những Dịch Vụ Mà Google Cloud Cung Cấp (itnavi.com.vn)](https://itnavi.com.vn/blog/google-cloud-la-gi-nhung-dich-vu-ma-google-cloud-cung-cap)

# **Hết.**