TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KỲ**

**KHAI THÁC DỮ LIỆU & KHAI PHÁ TRI THỨC**

**FINAL PROJECT**

*Người hướng dẫn*: **Thầy LÊ CUNG TƯỞNG**

*Người thực hiện*: **NGUYỄN MINH NHÂN – 51703147**

**Lim Hiệp Tiến – 51703195**

Lớp **: 17050302**

Khoá  **: 21**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KỲ**

**KHAI THÁC DỮ LIỆU & KHAI PHÁ TRI THỨC**

**FINAL PROJECT**

*Người hướng dẫn*: **Thầy LÊ CUNG TƯỞNG**

*Người thực hiện*: **NGUYỄN MINH NHÂN – 51703147**

**Lim Hiệp Tiến – 51703195**

Lớp **: 17050302**

Khoá  **: 21**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

LỜI CẢM ƠN

Em chân thành cảm ơn cảm ơn bố,mẹ đã tạo điều kiện tốt nhất có thể để nhóm em tiếp tục theo học và cũng rất biết ơn quý thầy cô ở khoa Công Nghệ Thông Tin đặc biệt là thầy Lê Cung Tưởng đã tạo điều kiện và hỗ trợ em trong suốt quá trình giảng dạy cũng như trong quá trình làm bài báo cáo cuối kỳ môn Khai Thác Dữ Liệu & Khai Phá Tri Thức

Trong quá trình làm bài em đã cố gắng thảo luận, nghiên cứu và hoàn thiện hết mức có thể project này, tuy nhiên vẫn không thể tránh những sai sót về vấn đề trình bày và nội dung đề tài. Vì vậy, em rất mong sự góp ý của các thầy cô để em hiểu rõ hơn về đề tài này.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng tôi / chúng tôi và được sự hướng dẫn của thầy Lê Cung Tưởng. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Nguyễn Minh Nhân*

*Lim Hiệp Tiến*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

Tìm hiểu về nội dung các giải thuật và giải quyết final project

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc72062735)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc72062736)

[TÓM TẮT iv](#_Toc72062737)

[MỤC LỤC 1](#_Toc72062738)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 2](#_Toc72062739)

[CHƯƠNG 1 – TỔNG QUAN 3](#_Toc72062740)

[1. Công nghệ sử dụng 3](#_Toc72062741)

[2. Tính năng 3](#_Toc72062742)

[3. Ứng dụng 4](#_Toc72062743)

[4. Hướng dẫn cài đặt 5](#_Toc72062744)

[CHƯƠNG 2 – LÝ THUYẾT GIẢI THUẬT 8](#_Toc72062745)

[1. Data mining 8](#_Toc72062746)

[1.1. Data mining là gì? 8](#_Toc72062747)

[1.2. Các bước thực hiện 8](#_Toc72062748)

[1.3. Ứng dụng 9](#_Toc72062749)

[1.4. Công cụ khai phá 9](#_Toc72062750)

[2. Giải thuật 12](#_Toc72062751)

[2.1. NAIVE BAYES 12](#_Toc72062752)

[2.1.1. Thuật toán Naive Bayes là gì? 12](#_Toc72062753)

[2.1.2. Ưu & Nhược điểm 12](#_Toc72062754)

[2.1.3. Thuật toán Naive Bayes hoạt động như thế nào? 13](#_Toc72062755)

[2.2. SUPPORT VECTOR MACHINE (SVM) 14](#_Toc72062756)

[2.2.1. SVM là gì? 14](#_Toc72062757)

[2.2.2. Ưu nhược điểm 14](#_Toc72062758)

[2.3. LOGISTIC REGRESSION 15](#_Toc72062759)

[2.3.1. Logistic regression là gì 15](#_Toc72062760)

[2.3.2. Sigmoid 16](#_Toc72062761)

[2.4. DECISION TREE 16](#_Toc72062762)

[2.4.1. Decision tree là gì? 17](#_Toc72062763)

[2.4.2. Các loại Decision tree 17](#_Toc72062764)

[2.4.3. Information Gain 17](#_Toc72062765)

[2.4.4. Ưu nhược điểm 17](#_Toc72062766)

[2.5. RANDOM FOREST 18](#_Toc72062767)

[2.5.1. Random Forest là gì 18](#_Toc72062768)

[2.5.2. Hoạt động của thuật toán Random Forest 19](#_Toc72062769)

[CHƯƠNG 3 – DEMO 21](#_Toc72062770)

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

**DANH MỤC HÌNH**

**No table of figures entries found.**

CHƯƠNG 1 – TỔNG QUAN

1. Công nghệ sử dụng

Python là ngôn ngữ lập trình hướng đối tượng, cấp cao, mạnh mẽ, được tạo ra bởi Guido van Rossum. Nó dễ dàng để tìm hiểu và đang nổi lên như một trong những ngôn ngữ lập trình nhập môn tốt nhất cho người lần đầu tiếp xúc với ngôn ngữ lập trình. Python hoàn toàn tạo kiểu động và sử dụng cơ chế cấp phát bộ nhớ tự động. Python có cấu trúc dữ liệu cấp cao mạnh mẽ và cách tiếp cận đơn giản nhưng hiệu quả đối với lập trình hướng đối tượng. Cú pháp lệnh của Python là điểm cộng vô cùng lớn vì sự rõ ràng, dễ hiểu và cách gõ linh động làm cho nó nhanh chóng trở thành một ngôn ngữ lý tưởng để viết script và phát triển ứng dụng trong nhiều lĩnh vực, ở hầu hết các nền tảng.

Python là một ngôn ngữ khá cũ được tạo ra bởi Guido Van Rossum. Thiết kế bắt đầu vào cuối những năm 1980 và được phát hành lần đầu tiên vào tháng 2 năm 1991.

1. Tính năng

**Ngôn ngữ lập trình đơn giản, dễ học:** Python có cú pháp rất đơn giản, rõ ràng. Nó dễ đọc và viết hơn rất nhiều khi so sánh với những ngôn ngữ lập trình khác như C++, Java, C#. Python làm cho việc lập trình trở nên thú vị, cho phép bạn tập trung vào những giải pháp chứ không phải cú pháp.

**Miễn phí, mã nguồn mở:** Bạn có thể tự do sử dụng và phân phối Python, thậm chí là dùng nó cho mục đích thương mại. Vì là mã nguồn mở, bạn không những có thể sử dụng các phần mềm, chương trình được viết trong Python mà còn có thể thay đổi mã nguồn của nó. Python có một cộng đồng rộng lớn, không ngừng cải thiện nó mỗi lần cập nhật.

**Khả năng di chuyển:** Các chương trình Python có thể di chuyển từ nền tảng này sang nền tảng khác và chạy nó mà không có bất kỳ thay đổi nào. Nó chạy liền mạch trên hầu hết tất cả các nền tảng như Windows, macOS, Linux.

**Khả năng mở rộng và có thể nhúng:** Giả sử một ứng dụng đòi hỏi sự phức tạp rất lớn, bạn có thể dễ dàng kết hợp các phần code bằng C, C++ và những ngôn ngữ khác (có thể gọi được từ C) vào code Python. Điều này sẽ cung cấp cho ứng dụng của bạn những tính năng tốt hơn cũng như khả năng scripting mà những ngôn ngữ lập trình khác khó có thể làm được.

**Ngôn ngữ thông dịch cấp cao:** Không giống như C/C++, với Python, bạn không phải lo lắng những nhiệm vụ khó khăn như quản lý bộ nhớ, dọn dẹp những dữ liệu vô nghĩa,... Khi chạy code Python, nó sẽ tự động chuyển đổi code sang ngôn ngữ máy tính có thể hiểu. Bạn không cần lo lắng về bất kỳ hoạt động ở cấp thấp nào.

**Thư viện tiêu chuẩn lớn để giải quyết những tác vụ phổ biến:** Python có một số lượng lớn thư viện tiêu chuẩn giúp cho công việc lập trình của bạn trở nên dễ thở hơn rất nhiều, đơn giản vì không phải tự viết tất cả code. Ví dụ: Bạn cần kết nối cơ sở dữ liệu MySQL trên Web server? Bạn có thể nhập thư viện MySQLdb và sử dụng nó. Những thư viện này được kiểm tra kỹ lưỡng và được sử dụng bởi hàng trăm người. Vì vậy, bạn có thể chắc chắn rằng nó sẽ không làm hỏng code hay ứng dụng của mình.

**Hướng đối tượng:** Mọi thứ trong Python đều là hướng đối tượng. Lập trình hướng đối tượng (OOP) giúp giải quyết những vấn đề phức tạp một cách trực quan. Với OOP, bạn có thể phân chia những vấn đề phức tạp thành những tập nhỏ hơn bằng cách tạo ra các đối tượng.

1. Ứng dụng

**Lập trình ứng dụng web:** Bạn có thể tạo web app có khả năng mở rộng (scalable) được bằng cách sử dụng framework và CMS (Hệ thống quản trị nội dung) được tích hợp trong Python. Vài nền tảng phổ biến để tạo web app là: Django, Flask, Pyramid, Plone, Django CMS. Các trang như Mozilla, Reddit, Instagram và PBS đều được viết bằng Python.

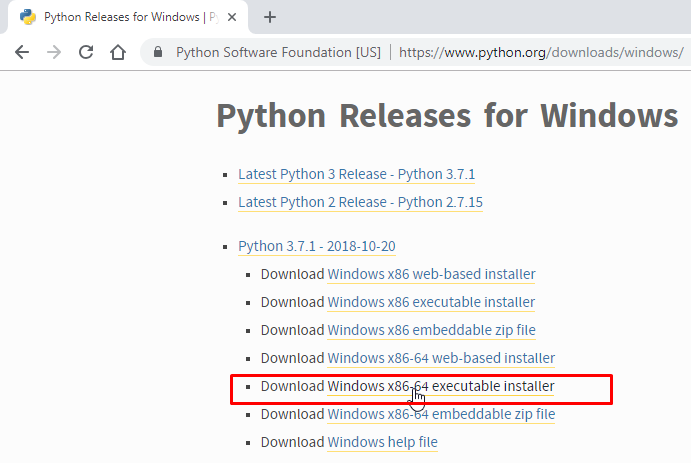
**Khoa học và tính toán:** Có nhiều thư viện trong Python cho khoa học và tính toán số liệu, như SciPy và NumPy, được sử dụng cho những mục đích chung chung trong tính toán. Và, có những thư viện cụ thể như: EarthPy cho khoa học trái đất, AstroPy cho Thiên văn học,... Ngoài ra, Python còn được sử dụng nhiều trong machine learning, khai thác dữ liệu và deep learning.

**Tạo nguyên mẫu phần mềm:** Python chậm hơn khi so sánh với các ngôn ngữ được biên dịch như C++ và Java. Nó có thể không phải là lựa chọn tốt nếu nguồn lực bị giới hạn và yêu cầu về hiệu quả là bắt buộc. Tuy nhiên, Python là ngôn ngữ tuyệt vời để tạo những nguyên mẫu (bản chạy thử - prototype). Ví dụ, bạn có thể sử dụng Pygame (thư viện viết game) để tạo nguyên mẫu game trước. Nếu thích nguyên mẫu đó có thể dùng C++ để viết game thực sự.

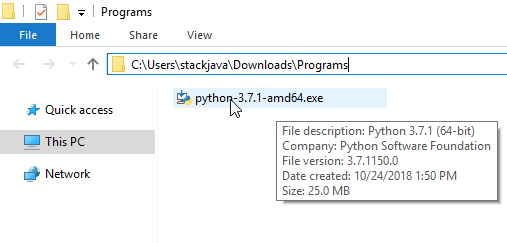
**Ngôn ngữ tốt để dạy lập trình:** Python được nhiều công ty, trường học sử dụng để dạy lập trình cho trẻ em và những người mới lần đầu học lập trình. Bên cạnh những tính năng và khả năng tuyệt vời thì cú pháp đơn giản và dễ sử dụng của nó là lý do chính cho việc này.

1. Hướng dẫn cài đặt
2. Download file cài đặt python tại

<https://www.python.org/downloads/windows/>

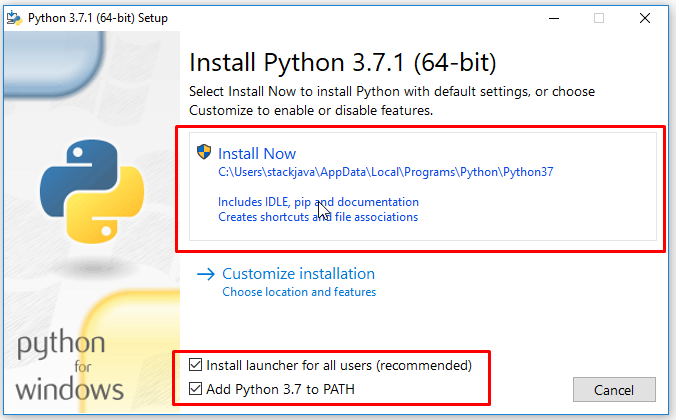


Click đúp vào file .exe vừa tải về để cài đặt

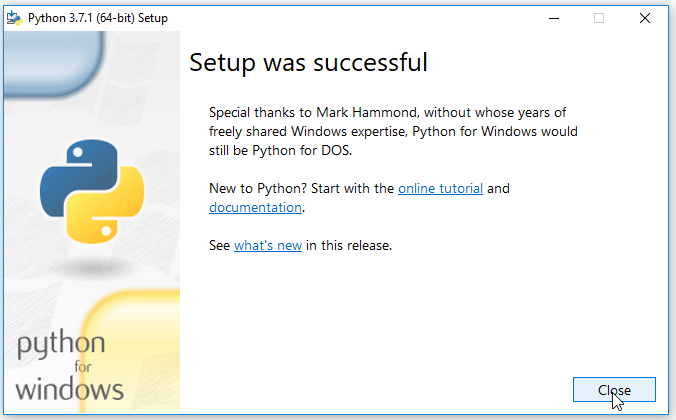


Chọn tick vào ô Add Python 3.7 to PATH (lựa chọn này giúp bạn chạy lệnh python trên cmd, powershell trên windows)

Và click vào Install Now (Bạn có thể đổi đường dẫn folder cài đặt bằng cách chọn phần Customize installation ở dưới)



Thành công



CHƯƠNG 2 – LÝ THUYẾT GIẢI THUẬT

1. Data mining
   1. Data mining là gì?

Data mining – khai phá dữ liệu là quá trình phân loại, sắp xếp các tập hợp dữ liệu lớn để xác định các mẫu và thiết lập các mối liên hệ nhằm giải quyết các vấn đề nhờ phân tích dữ liệu. Các MCU khai phá dữ liệu cho phép các doanh nghiệp có thể dự đoán được xu hướng tương lai.

Quá trình khai phá dữ liệu là một quá trình phức tạp bao gồm kho dữ liệu chuyên sâu cũng như các công nghệ tính toán. Hơn nữa, Data Mining không chỉ giới hạn trong việc trích xuất dữ liệu mà còn được sử dụng để chuyển đổi, làm sạch, tích hợp dữ liệu và phân tích mẫu.

Có nhiều tham số quan trọng khác nhau trong Data Mining, chẳng hạn như quy tắc kết hợp, phân loại, phân cụm và dự báo. Một số tính năng chính của Data Mining:

* Dự đoán các mẫu dựa trên xu hướng trong dữ liệu.
* Tính toán dự đoán kết quả
* Tạo thông tin phản hồi để phân tích
* Tập trung vào cơ sở dữ liệu lớn hơn.
* Phân cụm dữ liệu trực quan
  1. Các bước thực hiện

Bước 1: Làm sạch dữ liệu – Trong bước này, dữ liệu được làm sạch sao cho không có tạp âm hay bất thường trong dữ liệu.

Bước 2: Tích hợp dữ liệu – Trong quá trình tích hợp dữ liệu, nhiều nguồn dữ liệu sẽ kết hợp lại thành một.

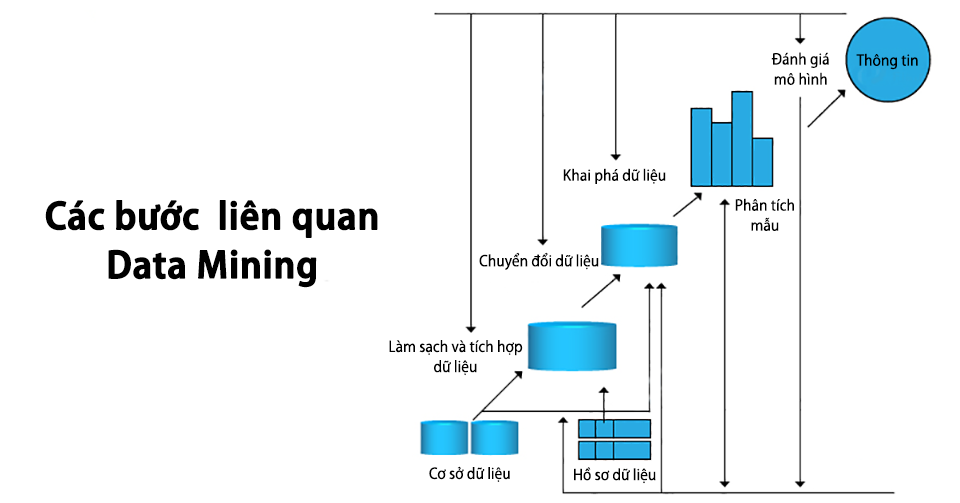
Bước 3: Lựa chọn dữ liệu – Trong bước này, dữ liệu được trích xuất từ cơ sở dữ liệu.

Bước 4: Chuyển đổi dữ liệu – Trong bước này, dữ liệu sẽ được chuyển đổi để thực hiện phân tích tóm tắt cũng như các hoạt động tổng hợp.

Bước 5: Khai phá dữ liệu – Trong bước này, chúng tôi trích xuất dữ liệu hữu ích từ nhóm dữ liệu hiện có.

Bước 6: Đánh giá mẫu – Chúng tôi phân tích một số mẫu có trong dữ liệu.

Bước 7: Trình bày thông tin – Trong bước cuối cùng, thông tin sẽ được thể hiện dưới dạng cây, bảng, biểu đồ và ma trận.



* 1. Ứng dụng
* Phân tích thị trường và chứng khoán
* Phát hiện gian lận
* Quản lý rủi ro và phân tích doanh nghiệp
* Phân tích giá trị trọn đời của khách hàng
* Khám phá thêm 10 ứng dụng khai phá dữ liệu
  1. Công cụ khai phá

**RapidMiner**

Là một trong những công cụ phổ biến nhất để khai phá dữ liệu, RapidMiner được viết trên nền tảng Java nhưng không yêu cầu mã hóa để vận hành. Hơn nữa, nó cung cấp các chức năng khai thác dữ liệu khác nhau như tiền xử lý dữ liệu, biểu diễn dữ liệu, lọc, phân cụm, v.v.

**Weka**

Weka là một phần mềm khai thác dữ liệu mã nguồn mở được phát triển tại Đại học Wichita. Giống như RapidMiner, Weka không có mã hóa và sử dụng GUI đơn giản.

Sử dụng Weka, bạn có thể gọi trực tiếp các thuật toán học máy hoặc nhập chúng bằng mã Java. Nó cung cấp một loạt các công cụ như trực quan hóa, tiền xử lý, phân loại, phân cụm, v.v.

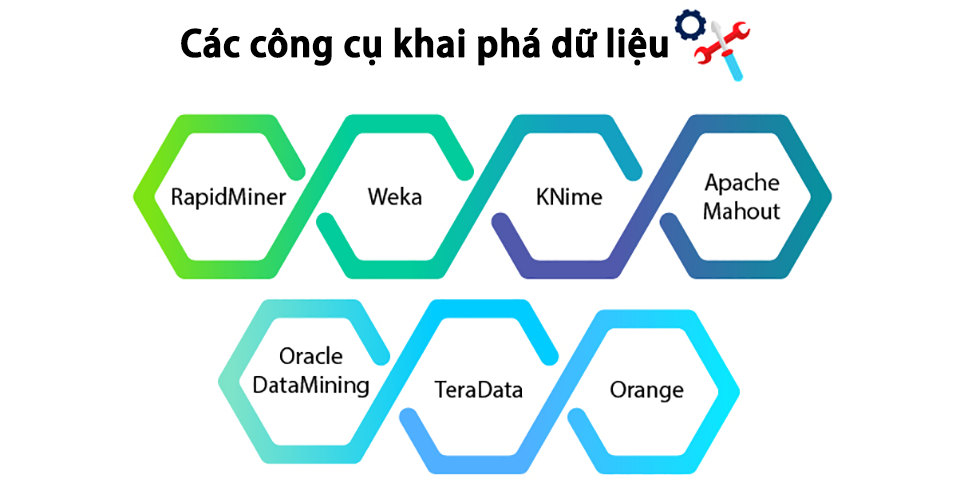
**KNime**

KNime là một bộ khai phá dữ liệu mạnh mẽ, chủ yếu được sử dụng cho tiền xử lý dữ liệu, đó là, ETL: Trích xuất, Chuyển đổi & Tải. Hơn nữa, nó tích hợp nhiều thành phần khác nhau của khoa học máy và khai phá dữ liệu để cung cấp một nền tảng bao gồm cho tất cả các hoạt động phù hợp.

**Apache Mahout**

Apache Mahout là một phần mở rộng của Nền tảng Big Data Hadoop. Các nhà phát triển tại Apache đã phát triển Mahout để giải quyết nhu cầu ngày càng tăng về khai phá dữ liệu và hoạt động phân tích trong Hadoop.

Kết quả là, nó chứa các chức năng học máy khác nhau như phân loại, hồi quy, phân cụm, v.v.



**Oracle DataMining**

Oracle DataMining là một công cụ tuyệt vời để phân loại, phân tích và dự đoán dữ liệu. Nó cho phép người dùng thực hiện khai phá dữ liệu trên cơ sở dữ liệu SQL để trích xuất các khung hình và biểu đồ.

**TeraData**

Đối với dữ liệu, nhập kho là một yêu cầu cần thiết. TeraData, còn được gọi là Cơ sở dữ liệu TeraData cung cấp dịch vụ kho chứa các công cụ khai phá dữ liệu.

Nó có thể lưu trữ dữ liệu dựa trên mức độ sử dụng của chúng, nghĩa là, nó lưu trữ dữ liệu ít được sử dụng trong phần ‘slow’ và cho phép truy cập nhanh vào dữ liệu được sử dụng thường xuyên.

**Orange**

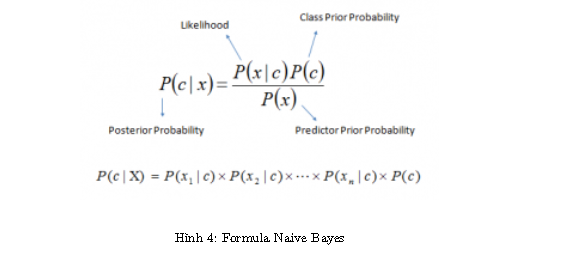
Phần mềm Orange được biết đến bởi việc tích hợp các công cụ khai phá dữ liệu và học máy. Nó được viết bằng Python và cung cấp trực quan tương tác và thẩm mỹ cho người dùng.

1. Giải thuật
   1. NAIVE BAYES
      1. Thuật toán Naive Bayes là gì?

Thuật toán Naive Bayes là kỹ thuật phân loại dựa trên giả định của Định lý Bayes với sự độc lập giữa các yếu tố dự đoán. Nói một cách dễ hiểu, một bộ phân loại Naive Bayes giả định rằng sự hiện diện của một đối tượng cụ thể trong một lớp không liên quan đến sự hiện diện của bất kỳ đối tượng địa lý nào khác.

Mô hình Naive Bayes dễ xây dựng và đặc biệt hữu ích cho các tập dữ liệu rất lớn. Cùng với sự đơn giản, Naive Bayes được biết đến là làm tốt hơn các phương pháp phân loại thậm chí rất phức tạp.

Định lý Bayes cung cấp một cách tính xác suất hậu nghiệm P (c | x) từP (c), P (x) và P (x | c). Nhìn vào phương trình bên dưới



P (c|x) là xác suất sau của lớp (c, mục tiêu) đã cho trước dự đoán (x, tại các cống phẩm).

P (c) là xác suất trước của hạng.

P (x|c) là khả năng xảy ra là xác suất của bộ dự đoán đã cho.

P (x) là xác suất dự đoán trước

* + 1. Ưu & Nhược điểm

**Ưu điểm:**

Dễ dàng và nhanh chóng để dự đoán lớp của tập dữ liệu thử nghiệm. Nó cũng hoạt động tốt trong dự đoán đa kính

Khi giả định về tính độc lập, bộ phân loại Naive Bayes hoạt động tốt hơn so với các mô hình khác như hồi quy logistic và bạn cần ít dữ liệu huấn luyện hơn.

Nó hoạt động tốt trong trường hợp các biến đầu vào phân loại so với (các) biến thể số. Đối với biến số, phân phối chuẩn được giả định (đường cong hình chuông, giả định mạnh mẽ).

**Nhược điểm:**

Nếu biến phân loại có một danh mục (trong tập dữ liệu thử nghiệm), không được quan sát trong tập dữ liệu huấn luyện, thì mô hình sẽ chỉ định xác suất 0 (không) và sẽ không thể đưa ra dự đoán. Điều này thường được gọi là "Tần số không". Để giải quyết vấn đề này, chúng ta có thể sử dụng kỹ thuật làm mịn. Một trong những kỹ thuật làm mịn đơn giản nhất được gọi là ước lượng Laplace.

Mặt khác, Bayes ngây thơ còn được biết đến như một công cụ ước lượng tồi, vì vậy các kết quả đầu ra xác suất từ dự đoán-proba không được coi trọng quá nhiều.

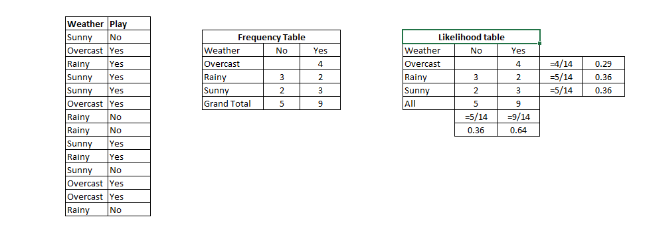
Một hạn chế khác của Naive Bayes là giả định về các mã dự đoán độc lập. Trong cuộc sống thực, hầu như không thể có được một tập hợp các yếu tố dự đoán hoàn toàn độc lập.

* + 1. Thuật toán Naive Bayes hoạt động như thế nào?

**Bước 1**: Chuyển đổi tập dữ liệu thành bảng tần số

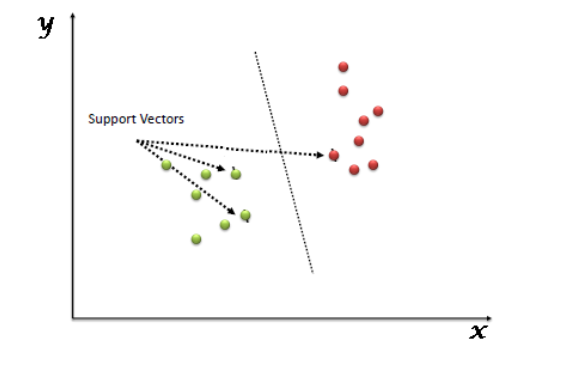
**Bước 2**: Tạo bảng Khả năng bằng cách tìm các xác suất

**Bước 3:** Bây giờ, sử dụng phương trình Naive Bayesian để tính xác suất sau cho mỗi lớp. Lớp có xác suất sau cao nhất là kết quả của sự suy đoán.



* 1. SUPPORT VECTOR MACHINE (SVM)
     1. SVM là gì?

(SVM) là một thuật toán học máy được giám sát có thể được sử dụng cho cả các thử thách phân loại hoặc hồi quy. Tuy nhiên, nó được sử dụng nhiều nhất trong các bài toán phân loại.



* + 1. Ưu nhược điểm

**Ưu điểm:**

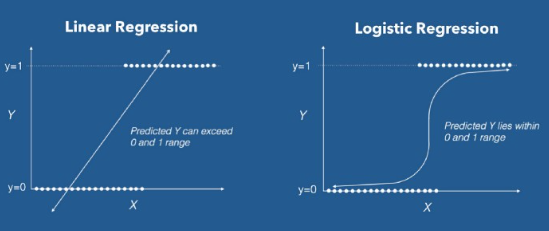
* Nó hoạt động thực sự tốt với biên độ tách biệt rõ ràng
* Nó có hiệu quả trong không gian chiều cao.
* Nó có hiệu quả trong trường hợp số lượng kích thước lớn hơn số lượng mẫu.
* Nó sử dụng một tập hợp con các điểm đào tạo trong hàm quyết định (được gọi là trình hỗ trợ), vì vậy nó cũng hiệu quả về bộ nhớ

**Nhược điểm:**

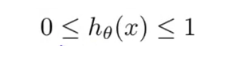
* Nó không hoạt động tốt khi chúng tôi có tập dữ liệu lớn vì thời gian đào tạo bắt buộc cao hơn
* Nó cũng hoạt động không tốt lắm, khi tập dữ liệu có nhiều nhiễu hơn, tức là các lớp đích chồng chéo lên nhau
* SVM không trực tiếp cung cấp các ước tính xác suất, những ước tính này được tính bằng cách sử dụng xác thực chéo năm lần đắt đỏ. Nó được bao gồm trong thư viện liên quan đến phương pháp SVC củaPython scikit-learning.
  1. LOGISTIC REGRESSION
     1. Logistic regression là gì

Hồi quy logistic được đặt tên cho hàm được sử dụng ở cốt lõi của phương pháp, hàm thần học

Logistic Regression là một thuật toán Máy học được sử dụng cho các vấn đề phân loại nhóm, nó là một thuật toán phân tích dự đoán và dựa trên xác suất conceptof.

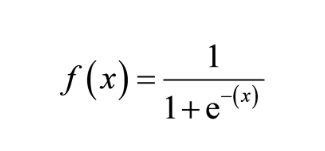


Hàm logistic, còn được gọi là hàm sigmoid được phát triển bởi các nhà thống kê để mô tả các đặc tính của sự gia tăng dân số trong hệ sinh thái, tăng nhanh và đạt cực đại theo khả năng mang của môi trường. Đó là một đường cong hình chữ S có thể lấy bất kỳ số nào có giá trị thực và ánh xạ nó thành một giá trị từ 0 đến 1, nhưng bất cứ khi nào chính xác ở các giới hạn đó



* + 1. Sigmoid

Hàm sigmoid / hàm logistic là một hàm tương tự như một đường cong hình chữ “S” khi được vẽ trên đồ thị. Nó nhận các giá trị từ 0 đến 1 và "bình phương" chúng hướng vào lề ở trên cùng và dưới cùng, gắn nhãn chúng là 0 hoặc 1



* 1. DECISION TREE
     1. Decision tree là gì?

Thuật toán Cây quyết định thuộc họ thuật toán học có giám sát, không giống như các thuật toán học có giám sát khác, thuật toán cây quyết định cũng có thể được sử dụng để giải các bài toán hồi quy và phân loại.

Mục tiêu của việc sử dụng Cây quyết định là tạo ra một mô hình đào tạo có thể sử dụng dự đoán hàng đầu về lớp hoặc giá trị của biến mục tiêu bằng cách học các quy tắc quyết định đơn giản được đưa ra từ dữ liệu trước đó (dữ liệu đào tạo).

Trong Cây quyết định, để dự đoán nhãn lớp cho một bản ghi, chúng ta bắt đầu từ gốc của cây. Chúng tôi so sánh các giá trị của thuộc tính gốc với thuộc tính của bản ghi. Trên cơ sở so sánh, chúng tôi theo dõi nhánh tương ứng với giá trị đó và chuyển đến nút tiếp theo.

* + 1. Các loại Decision tree

Cây quyết định biến phân loại: có một biến mục tiêu phân loại, sau đó nó được gọi là cây quyết định biến phân loại.

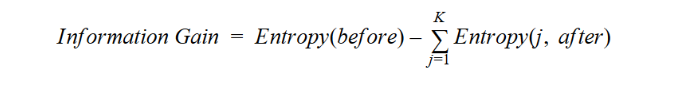
Cây quyết định biến liên tục: có một biến mục tiêu liên tục thì được gọi là Cây quyết định biến liên tục

* + 1. Information Gain

**Ưu điểm**

Khi chúng ta sử dụng một nút trong cây quyết định để phân vùng các cá thể huấn luyện trong tập hợp con intosmaller thì entropy sẽ thay đổi. Mức tăng thông tin là thước đo của entropy thay đổi này

Giả sử S là tập các cá thể, A là một thuộc tính, Sv là tập con của S với A = v và Giá trị (A) là tập tất cả các giá trị có thể có của A, thì



* + 1. Ưu nhược điểm

Dễ hiểu và dễ hiểu. Tại mỗi nút, chúng ta có thể thấy chính xác quyết định mà mô hình của chúng ta đang thực hiện. Trong thực tế, chúng tôi sẽ có thể hiểu đầy đủ về nguyên nhân và lỗi đến từ đâu, loại dữ liệu nào mà mô hình sẽ hoạt động tốt và kết quả đầu ra bị ảnh hưởng như thế nào bởi các giá trị của các tính năng. Công cụ hình ảnh của Scikit learning là một lựa chọn tuyệt vời để hình dung và hiểu các ý kiến quyết định.

Yêu cầu chuẩn bị dữ liệu rất ít. Nhiều mô hình ML có thể yêu cầu xử lý trước dữ liệu nặng như chuẩn hóa và có thể yêu cầu các quy trình quản lý phức tạp. Mặt khác, cây quyết định hoạt động khá tốt sau khi chỉnh sửa một vài thông số.

Chi phí sử dụng cây để suy luận là logarit trong số điểm dữ liệu được sử dụng để huấn luyện cây. Đó là một điểm cộng rất lớn vì nó có nghĩa là có nhiều dữ liệu hơn sẽ không nhất thiết làm giảm tốc độ suy luận của chúng ta

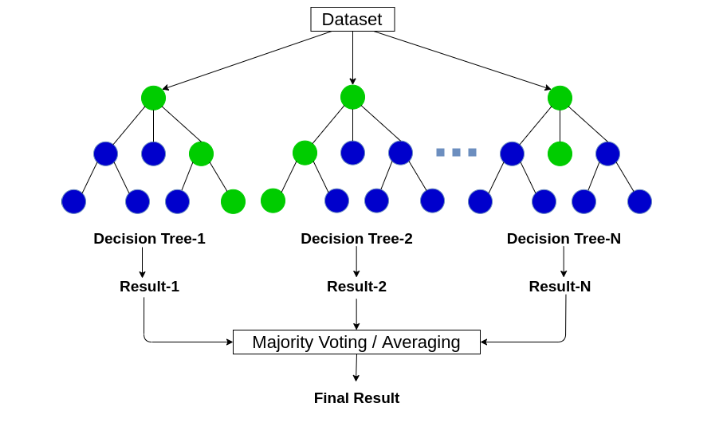
**Nhược điểm**

Việc trang bị quá mức là khá phổ biến với các cây quyết định đơn giản là do bản chất của việc đào tạo chúng. Bạn thường nên thực hiện một số loại giảm kích thước như PCA để cây không phải học các phép phân chia trên nhiều tính năng

Vì những lý do tương tự như trường hợp overfitting, cây quyết định cũng dễ bị sai lệch đối với các lớp chiếm đa số trong tập dữ liệu. Luôn luôn là một ý tưởng hay khi thực hiện một số loại cân bằng lớp như trọng lượng lớp, lấy mẫu, hoặc chức năng giảm chuyên dụng

* 1. RANDOM FOREST
     1. Random Forest là gì

Rừng ngẫu nhiên (RF) về cơ bản là một túi chứa n Cây quyết định (DT) có một tập hợp các siêu tham số khác nhau và được huấn luyện trên các tập con dữ liệu khác nhau. Giả sử tôi có 100 cây quyết định trong túi Rừng ngẫu nhiên của mình !! Như tôi vừa nói, những cây quyết định này có một tập hợp các siêu tham số khác nhau và một tập hợp con dữ liệu huấn luyện khác nhau, vì vậy quyết định hoặc dự đoán được đưa ra bởi những cây này có thể khác nhau rất nhiều



Random Forest là một thuật toán học tập có giám sát được sử dụng cho cả phân loại cũng như hồi quy. Tuy nhiên, nó chủ yếu được sử dụng cho các bài toán phân loại, như chúng ta biết rằng một khu rừng được tạo thành từ cây cối và càng nhiều cây cối đồng nghĩa với việc rừng càng vững chắc. Tương tự, thuật toán rừng ngẫu nhiên tạo ra các cây quyết định trên các mẫu dữ liệu và sau đó nhận dự đoán từ mỗi mẫu và cuối cùng chọn ra giải pháp tốt nhất bằng cách bỏ phiếu. Đây là một phương pháp tổng hợp tốt hơn một phương pháp quyết định đơn lẻ bởi vì nó làm giảm sự phù hợp quá mức bằng cách lấy trung bình kết quả

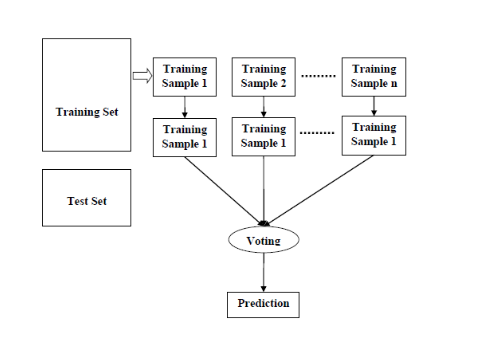
* + 1. Hoạt động của thuật toán Random Forest

Bước 1: Đầu tiên, hãy bắt đầu với việc chọn các mẫu ngẫu nhiên từ một tập dữ liệu nhất định.

Bước 2: Tiếp theo, thuật toán này sẽ xây dựng một cây quyết định cho mọi mẫu. Thenit sẽ lấy kết quả dự đoán từ mọi cây quyết định.

Bước 3: Ở bước này, sẽ thực hiện bình chọn cho mọi kết quả dự đoán.

Bước 4: Cuối cùng, chọn kết quả dự đoán được bình chọn nhiều nhất làm kết quả dự đoán cuối cùng.



CHƯƠNG 3 – DEMO

python codes : Final\_project\_Datamining.ipynb

collected data : Comment1, Comment2, Comment3.csv

Presentation 51703147\_51703195.pptx

project report: report.docx

Folder-data: chứa file comment đã chèn label emotion

**TÀI LIỆU THAM KHẢO**