|  |  |
| --- | --- |
| ĐẠI HỌC QUỐC GIA TP.HCM  **TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT** | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc lập – Tự do – Hạnh phúc** |

*TP Hồ Chí Minh, ngày 17 tháng 10 năm 2019*

**ĐĂNG KÝ ĐỀ TÀI NGHIÊN CỨU KHOA HỌC SINH VIÊN NĂM 2019**

1. **TÊN ĐỀ TÀI:** Nghiên cứu các công cụ phân tích dữ liệu và ứng dụng khai phá dữ liệu điểm để dự đoán kết quả học tập của sinh viên.
2. **MÃ SỐ: 02**

**Thuộc lĩnh vực nghiên cứu:**

|  |  |  |  |
| --- | --- | --- | --- |
| **TT** | **TÊN LĨNH VỰC** | **CHUYÊN NGÀNH** | **Chọn** |
| 1 | Giáo dục | - Giáo dục học  - Quản lý giáo dục  - Tâm lý giáo dục  - Giáo dục thể chất | **□**  **□**  **□**  **□** |
| 2 | Lĩnh vực Kinh tế | - Tài chính- ngân hàng- chứng khoán- kế toán- kiểm toán, bảo hiểm - tín dụng  - Thương mại- quản trị kinh doanh và du lịch- marketing  - Kinh tế học- kinh tế phát triển – kinh tế chính trị | **□**  **✓**  **□** |
| 3 | Lĩnh vực Pháp lý | - Luật Dân sự  - Luật Hình sự  - Luật Kinh tế  - Luật Hành chính  - Luật Quốc tế | **□**  **□**  **□**  **□**  **□** |
| 4 | Lĩnh vực Tài nguyên và Môi trường | - Tài nguyên  - Môi trường  - Công nghệ môi trường  - Kỹ thuật môi trường | **□**  **□**  **□**  **□** |
| **5** | Lĩnh vực Công nghệ thông tin | - Toán tin  - Công nghệ thông tin | **□**  **□** |

**3. THỜI GIAN THỰC HIỆN: 5** tháng

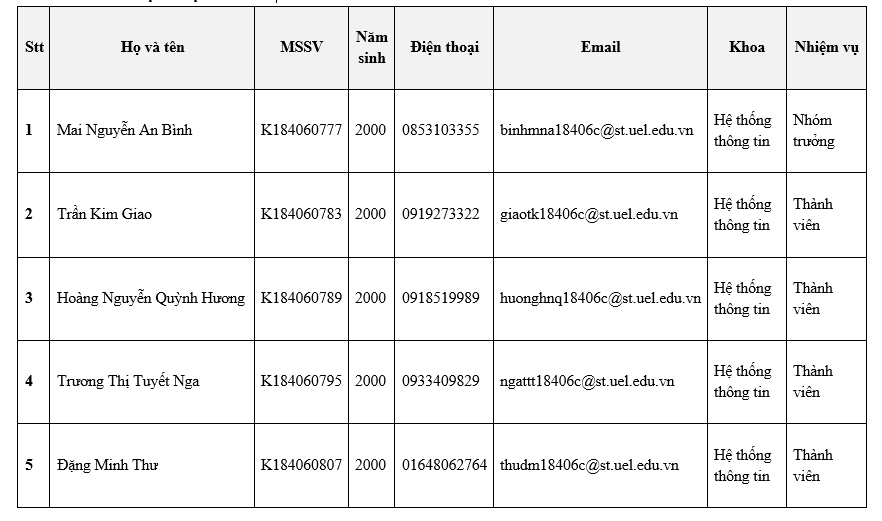
**4. NGƯỜI HƯỚNG DẪN:**

Họ và tên: Th.S Trần Thị Ánh

Đơn vị: Giảng viên khoa Hệ Thống Thông Tin trường Đại học Kinh Tế - Luật

Điện thoại: 0984823596 Email:anhtt@uel.edu.vn

**5. SINH VIÊN THỰC HIỆN ĐỀ TÀI**

****

**6. GIỚI THIỆU Ý TƯỞNG NGHIÊN CỨU**

-         Cách mạng công nghệ đã tạo ra nguồn dữ liệu to lớn với sự bùng nổ của thiết bị kỹ thuật số, các thiết bị cảm biến… tạo ra dữ liệu lớn. Xu thế này sẽ không ngừng phát triển mạnh mẽ. Mục tiêu của các kỹ thuật khai phá dữ liệu lớn là đưa ra được các thông tin yêu cầu hoặc thậm chí phát hiện ra những mối quan hệ ẩn và các mẫu trong hàng loạt các tham số. Ngành phân tích dữ liệu đang trở nên “nóng” hơn bao giờ hết ở cả Việt Nam và trên thế giới. Vì thế mà ngày càng xuất hiện nhiều công cụ phân tích dữ liệu (cả trả phí và miễn phí) với từng ưu điểm và nhược điểm khác nhau. Bởi vì lí do này mà việc chọn được một công cụ phù hợp với nhu cầu chưa bao giờ là dễ dàng.

⇨ Phân tích rõ ràng ưu và nhược điểm của một số công cụ phân tích dữ liệu đang được tin dùng nhất hiện nay. Từ đó giúp định hướng chọn được một công cụ phù hợp nhất với nhu cầu khai phá dữ liệu điểm để dự báo kết quả học tập của sinh viên nhằm định hướng cụ thể kế hoạch học tập trong tương lai.

-        Nhu cầu khai phá dữ liệu điểm để dự báo kết quả học tập của sinh viên. Hiện nay, hầu hết các trường Đại học và Cao đẳng đều đã chuyển sang chế độ đào tạo  theo tín chỉ. Vì vậy việc định hướng một lộ trình học tập hiệu quả đang là vấn đề đáng quan tâm của hầu hết các bạn sinh viên cũng như giảng viên. Bởi lẽ thời gian gần đây số lượng sinh viên buộc phải thôi học đã tăng đáng kể (chẳng hạn như tại Đại học Cần Thơ hàng năm có trên dưới 150 sinh viên bị buộc phải thôi học do kết quả học tập yếu kém). Tuy nhiên, để mỗi sinh viên có thể tự định hướng cho mình một kế hoạch học tập mang lại hiệu quả cao hay với sự tư vấn của cố vấn học tập thì cũng khá khó khăn. Cho nên việc khai phá dữ liệu điểm để dự báo kết quả học tập sẽ giúp đỡ sinh viên cũng như giảng viên rất nhiều trong nhu cầu hoạch định kế hoạch học tập phù hợp, từ đó giúp giảng viên có định hướng giúp đỡ những bạn sinh viên còn yếu kém hay bồi dưỡng các bạn sinh viên có tiềm năng để tham gia vào các cuộc thi.

⇨ Mong muốn hỗ trợ các bạn sinh viên và giảng viên đại học trong việc hình thành kế hoạch học tập cụ thể và hiệu quả.

**7. MỤC TIÊU CỦA ĐỀ TÀI, ĐỐI TƯỢNG, PHẠM VI**

-         Mục tiêu:

* Nghiên cứu, tìm hiểu về các công cụ phân tích dữ liệu hiện nay: Excel, R, Python, SAS, Tableau, KNIME, Looker, QuikView, Splunk, Spark.Qua đó sẽ cung cấp cho mọi người kiến thức, thông tin cụ thể của từng công cụ.
* Sử dụng một công cụ phân tích dữ liệu để khai phá dữ liệu điểm của sinh viên trường Đại học. Từ đó giúp chúng ta thấy được lợi ích to lớn của việc áp dụng công cụ khai phá dữ liệu
* Từ việc khai phá, hỗ trợ chúng ta đưa ra dự đoán về kết quả học tập và lập kế hoạch học tập phù hợp cho từng cá nhân sinh viên, từ đó hỗ trợ lựa chọn môn học phù hợp với năng lực của từng cá nhân.

-         Đối tượng nghiên cứu:

* Các công cụ phân tích dữ liệu “hot” nhất hiện nay: Excel, R, Python, SAS, Tableau, KNIME, Looker, QuikView, Splunk, Spark (Tổng quan, đặc điểm, đặc tính, ưu điểm và nhược điểm cụ thể của từng công cụ để lựa chọn công cụ phù hợp và đạt hiệu quả nhất).
* Công cụ Tableau được đánh giá là phù hợp với đề tài này nhất (ưu điểm của Tableau so với các công cụ khác)
* Dữ liệu điểm của sinh viên Đại học

-         Phạm vi nghiên cứu:sinh viên năm 2,3 tại trường Đại học Kinh Tế - Luật

**8.** **CƠ SỞ LÝ THUYẾT VÀ LỊCH SỬ NGHIÊN CỨU**

* Cơ sở lý thuyết:
* Các công cụ Business Intelligent (BI)
* Lý thuyết về kho dữ liệu, kỹ thuật phân lớp, mô hình dự đoán
* Kỹ thuật khai phá dữ liệu (Data Minning)
* Lịch sử nghiên cứu:
* ***Khai phá dữ liệu điểm để dự đoán kết quả học tập của Sinh viên trường Cao đẳng Sư phạm Hà Nội (2013)***

+ Chương 1: Giới thiệu quá trình quản lý đào tạo của trường , đối tượng, phạm vi và phương pháp nghiên cứu của luận văn.

+ Chương 2: Cơ sở lý thuyết: Kho dữ liệu, Mô hình phân tích trực tuyến OLAP, Khai phá dữ liệu, Công cụ Business Intelligence Development Studio của SQL Server 2008.

+ Chương 3: Ứng dụng: tiến hành phân tích báo cáo đa chiều và xây dựng các mô hình dự đoán dựa trên công cụ Business Intelligence Development Studio của SQL Server 2008, đánh giá các mô hình

+ Chương 4: Kết luận và định hướng phát triển Chương này tóm lược những kết quả đạt được của luận văn. Đồng thời đưa ra định hướng nghiên cứu trong thời gian tới.

* ***Ứng dụng khai phá dữ liệu xây dựng công cụ dự đoán kết quả học tập của sinh viên \_Khoa Thống kê – Tin học\_Trường Đại học Kinh tế\_ĐH Đà Nẵng(2012)***

Tóm tắt: Đề tài tiến hành nghiên cứu các kỹ thuật cho phép phân lớp trong khai phá dữ liệu, ứng dụng các kỹ thuật đó để xây dựng mô hình dự đoán kết quả học tập của sinh viên. Từ đó, xây dựng chương trình dự báo kết quả học tập cho mỗi sinh viên để tư vấn cho sinh viên chọn lộ trình học phù hợp nhất để đạt được kết quả học tập cao nhất.

* ***Phát hiện môn học quan trọng ảnh hưởng đến kết quả học tập của sinh viên ngành công nghệ thông tin tại trường ĐH Cần Thơ (2014)***

Tóm tắt: Tiếp cận khai mỏ dữ liệu để phát hiện môn học quan trọng ảnh hưởng đến kết quả học tập của sinh viên ngành công nghệ thông tin (CNTT). Thu thập,xử lí dữ liệu, đưa dữ liệu về cấu trúc bảng. Đề xuất sử dụng giải thuật rừng ngẫu nhiên học từ dữ liệu để rút trích các môn học quan trọng trong chương trình đào tạo ngành CNTT. Kết quả thu được sau khi rút trích có thể cung cấp thông tin hữu ích cho các nhà quản lý giáo dục trong việc tổ chức giảng dạy để nâng cao hiệu quả đào tạo

* ***A model to Forecast Learning Outcomes for Students in Blended Learning Courses Based On Learning Analytics (2018)***

Tóm tắt:

Bài viết đề xuất mô hình kết quả học tập dự báo dựa trên sự tương tác của người học với các hệ thống học tập trực tuyến bằng cách cung cấp bảng điều khiển phân tích học tập cho người học và giáo viên để theo dõi và định hướng trực tuyến của người học dựa trên một số kỹ thuật khai thác dữ liệu và học máy.

* ***Article Data Science Approach for Simulating Educational Data: Towards the Development of Teaching Outcome Model (TOM) (2018)***

Tóm tắt:

Bài viết đưa ra một phương pháp khoa học dữ liệu kết hợp phân tích và phân phối giải pháp dựa trên dữ liệu để khám phá vai trò của phân tích giảng dạy, bằng cách tạo mã giả mô phỏng dữ liệu để giúp phát triển các trường hợp kiểm tra hoạt động giảng dạy. Kết quả của phương pháp này nhằm thông báo sự phát triển của mô hình kết quả giảng dạy (TOM), có thể được sử dụng để truyền cảm hứng và kiểm tra chất lượng giảng dạy. Phương pháp mô phỏng được báo cáo trong nghiên cứu đã được thực hiện thông qua Splunk. Splunk là một nền tảng Dữ liệu lớn được thiết kế để thu thập và phân tích khối lượng lớn dữ liệu do máy tạo và hiển thị kết quả trên bảng điều khiển trong thời gian thực. Chúng tôi trình bày kết quả dưới dạng một loạt các bảng điều khiển trực quan minh họa các mô hình, xu hướng và kết quả trong hiệu suất giảng dạy. Nghiên cứu của chúng tôi nhằm góp phần phát triển cách tiếp cận khoa học dữ liệu giáo dục để hỗ trợ văn hóa đưa ra quyết định thông tin dữ liệu trong giáo dục đại học.

**9. PHƯƠNG PHÁP NGHIÊN CỨU**

* Phương pháp thống kê: thống kê điểm của các sinh viên trường đại học.
* Phương pháp so sánh: so sánh điểm mạnh yếu giữa các công cụ khai phá dữ liệu, so sánh điểm của các sinh viên trong những học kỳ trước đó.
* Phương pháp phân tích: dùng dữ liệu khai phá được bằng công cụ khai phá dữ liệu để phân tích phân tích lý do dẫn đến sự thay đổi điểm số và đưa ra dự đoán về kết quả học tập của sinh viên trường đại học
* Phương pháp tổng hợp: tổng hợp các công cụ dùng khai phá dữ liệu.

**10.** **TÓM TẮT NỘI DUNG CỦA ĐỀ TÀI**

-         **Phần một: Nghiên cứu các công cụ phân tích dữ liệu như: Excel, R, Python, SAS, Tableau, KNIME, Looker, QuikView, Splunk, Spark.**

***PHẦN MỀM EXCEL***

1. Tổng quan về phần mềm Excel:

Microsoft Excel (hay còn gọi là Excel) là chương trình xử lý bảng tính nằm trong bộ [Microsoft Office](https://vi.wikipedia.org/wiki/Microsoft_Office) của hãng phần mềm [Microsoft](https://vi.wikipedia.org/wiki/Microsoft) được thiết kế để giúp ghi lại, trình bày các thông tin xử lý dưới dạng bảng, thực hiện tính toán và xây dựng các số liệu thống kê trực quan có trong bảng từ Excel. Cũng như các chương trình bảng tính [Lotus 1-2-3](https://vi.wikipedia.org/wiki/Lotus_1-2-3), [Quattro Pro](https://vi.wikipedia.org/wiki/Quattro_Pro)… bảng tính của Excel cũng bao gồm nhiều ô được tạo bởi các dòng và cột, việc nhập dữ liệu và lập công thức tính toán trong Excel cũng có những điểm tương tự, tuy nhiên Excel có nhiều tính năng ưu việt và có giao diện rất thân thiện với người dùng.

Excel là một phần mềm dùng để quản lý và tính toán dữ liệu, điểm mạnh của nó là thống kê và tính toán theo công thức do người dùng lập trình ra, để từ đó nó tự động tiếp nhận dữ liệu khi có thay đổi. Excel thường được dùng trong các doanh nghiệp ở các bộ phận như kế toán, thu ngân, ... Excel hoạt động không giống như những phần mềm chuyên nghiệp bởi lẽ cơ chế bảo mật và backup dữ liệu của nó cũng không tốt bằng những phần mềm đó, tuy nhiên nó vẫn được sử dụng rộng rãi bởi rất ít doanh nghiệp có điều kiện làm ra một phần mềm cho họ.

Excel được xem là công cụ hỗ trợ đắc lực cho quá trình quản trị và hoạt động hiệu quả của doanh nghiệp. Excel cho phép người dùng doanh nghiệp khai thác các dữ liệu tiềm năng bằng cách sử dụng một số công thức trên bảng tính. Dữ liệu được chèn vào từng ô theo hàng hoặc cột, sau đó được sắp xếp, lọc và trình bày trực quan thông qua các biểu đồ theo ý muốn của người dùng. Từ đó giúp gia tăng sự thuyết phục cho các báo cáo kinh doanh của doanh nghiệp. Hơn nữa, Excel còn đề xuất cho bạn các biểu đồ phù hợp nhất cho từng loại dữ liệu. Người dùng Excel có thể định dạng bảng tính của họ bằng các kiểu chữ khác nhau (in đậm, in nghiêng, gạch chân, …), phân biệt dữ liệu giữa cột/hàng, tìm kiếm giá trị lớn nhất của dữ liệu, … Ngoài ra, người dùng còn có thể chọn màu thích hợp bằng nút lệnh Quick Analysis, sau đó sử dụng công cụ Formatting. Khi hiển thị dữ liệu dưới dạng biểu đồ hay đồ thị, Excel giúp người dùng nhận biết các mốc quan trọng trong dữ liệu được phân tích. Điều này giúp các doanh nghiệp nhận biết các xu hướng để đưa ra các dự đoán chính xác nhằm thực hiện chiến lược thích hợp cho doanh nghiệp. Excel còn có thể hợp nhất nhiều thông tin từ các tệp khác nhau để các thông tin này tồn tại ở một vị trí duy nhất. Giống như dữ liệu thô và các thông tin từ các bảng tính, bạn có thể nhập dữ liệu dạng chữ, số hoặc bằng hình ảnh. Bên cạnh đó, Excel là một phần của Office 365 Business và Office 365 Business Premium nên các nhà lãnh đạo doanh nghiệp có thể truy cập chương trình này thông qua những thiết bị khác nhau và ở bất kỳ vị trí nào. Chỉ cần có PC, máy tính xách tay, điện thoại thông minh hoặc máy tính bảng có thể truy cập web, bạn có thể truy cập Excel ở bất cứ đâu. Điều này giúp môi trường làm việc từ xa trở nên phổ biến hơn bao giờ hết.

Sức mạnh của Excel còn được các lập trình viên tích hợp vào các công cụ phân tích và tìm kiếm rất nhiều. Đặc biệt có thể kể đến là các công cụ liên quan đến Facebook. Với sự phát triển mạnh mẽ không ngừng của Facebook cũng như là quảng cáo trên Facebook thì việc lọc ra danh sách dựa theo các tiêu chí và xuất ra file Excel được rất nhiều người quan tâm. Không chỉ dừng lại ở đấy Excel còn được nhiều lập trình viên đánh giá là một công cụ tuyệt vời nhất. Cái mà toàn bộ những tính năng của nó không biết khi nào mới có thể tận dụng hết được. Và với mọi phần mềm khác khi xuất ra bất kỳ file để đánh giá thì Excel chính là lựa chọn hàng đầu bởi hiệu quả mà nó mang lại.

1. Ưu điểm và nhược điểm của phần mềm Excel:

* **Ưu điểm:** Excel là công cụ cơ bản và phổ biến nhất cho Phân tích Dữ liệu và được sử dụng trong phân tích dữ liệu ở tất cả các ngành nghề. Việc sử dụng Excel không mất quá nhiều thời gian để tìm hiểu và không đòi hỏi người dùng là người có kiến thức chuyên sâu về cơ sở dữ liệu. Bên cạnh đó mô hình của excel cũng thuận tiện cho việc tạo danh sách, hỗ trợ tối đa cho người dùng. Excel phân phối rộng rãi, không mất phí sử dụng và được cài đặt sẵn trên máy tính, pc. Điều này giúp người dùng giảm bớt thời gian làm quen, cài đặt và giá rẻ khi sử dụng. Excel còn có nhiều tính năng phân tích, lập báo cáo nhanh chóng mà không cần bất kỳ phần mềm nào khác hỗ trợ.
* **Nhược điểm:** Phải thao tác thành thạo Excel mới làm nhanh và chính xác được. Vì lập dự toán bằng Excel đòi hỏi người dùng phải thành thạo các lệnh trong Excel như SUMIF, Sort and Filter, Vlookup, Pivot Table... chỉ một thao tác không chuẩn thì dữ liệu nhập vào sẽ không tính toán được. Và nếu xuất hiện một công tác bị bỏ sót, người lập sẽ phải chạy lại dự toán từ đầu, tính toán lại từ đầu, cân chỉnh từ đầu... Việc này chiếm rất nhiều thời gian, nhưng lại thường xuyên xảy ra khi sử dụng dự toán bằng Excel. Excel có tính bảo mật thấp và có thể dễ dàng bị xóa file và không thể phục hồi lại dữ liệu. Đa số người dùng các chương trình Microsoft nói chung và Excel nói riêng đều đã từng bị mất dữ liệu vì lỡ tay ấn thoát khi chưa lưu, vì sự cố nguồn điện hay vì máy tính bị lỗi, nhiễm virus,... Một yếu tố quan trọng khác là lượng dữ liệu cần lưu trữ. Khi cơ sở dữ liệu lớn, người dùng không thể kiểm soát hết thì Excel lại trở nên bị động và gây nhiều khó khăn khi làm việc. Bên cạnh đó thì việc lưu trữ các file riêng lẻ khiến cho việc tìm kiếm và tổng hợp dữ liệu trở nên khó khăn và mất nhiều thời gian.

***PHẦN MỀM R***

1. Tổng quan về phần mềm R:

Phần mềm R (hay còn gọi là ngôn ngữ lập trình R) để phân tích dữ liệu được xây dựng bởi Ross Ihaka và Robert Gentleman tại  The University of Auckland, New Zealand, tiếp tục được phát triển bởi nhóm R Development Core Team. Phần lớn các kỹ thuật phân tích trong kinh doanh đều được R hỗ trợ – từ Thống kê đến Học máy hay các kỹ thuật Tối ưu hóa. Phần mềm R là một trong những công cụ phân tích thống kê học cũng như phân tích dữ liệu nói chung. Đây là phần mềm mã nguồn mở (miễn phí) nhưng mang đầy đủ những tính năng của các phần mềm thương mại khác hiện có như SPSS, AMOS, STATA hay EViews.

Phần mềm R có chứa nhiều loại kỹ thuật thống kê (mô hình hóa tuyến tính và phi tuyến, kiểm thử thống kê cổ điển, phân tích chuỗi thời gian, phân loại, phân nhóm, v.v.) và đồ họa. Phần mềm R, giống như S, được thiết kế xoay quanh một ngôn ngữ máy thực thụ, và nó cho phép người dùng thêm các tính năng bổ sung bằng cách định nghĩa các hàm mới. Nhiều hệ thống trong R được viết bằng chính ngôn ngữ của nó, giúp cho người dùng dễ theo dõi các giải thuật. Để thực hiện công việc chuyên về tính toán, R có thể liên kết được với ngôn ngữ C, C++ và Fortran để có thể được gọi trong khi chạy. Người dùng thông thạo có thể viết mã C để xử lý trực tiếp các đối tượng của phần mềm R.

Phần mềm R cũng có tính mở rộng cao bằng cách sử dụng các gói cho người dùng đưa lên cho một số chức năng và lĩnh vực nghiên cứu cụ thể. Do được thừa hưởng từ S, R có nền tảng lập trình hướng đối tượng mạnh hơn đa số các ngôn ngữ tính toán thống kê khác. Việc mở rộng R cũng dễ dàng nhờ các luật đóng khối từ vựng

Một điểm mạnh khác của phần mềm R là nền tảng đồ họa của nó, có thể tạo ra những đồ thị chất lượng cao cùng các biểu tượng toán học. R cũng có định dạng văn bản riêng tương tự như LaTeX, dùng để cung cấp tài liệu hướng dẫn toàn diện, có trực tuyến ở các định dạng khác nhau và cả bản in.

Dù phần mềm R được dùng chủ yếu bởi những nhà thống kê và những người sử dụng khác đòi hỏi một môi trường tính toán thống kê và phát triển phần mềm, nó cũng có thể dùng làm một công cụ tính toán ma trận tổng quát với các kết quả đo đạc cạnh tranh so với GNU Octave và đối thủ thương mại của nó, MATLAB. Giao diện RWeka đã được thêm vào phần mềm khai phá dữ liệu phổ biến Weka, cho phép đọc/ghi định dạng arff vì vậy cho phép sử dụng tính năng khai phá dữ liệu trong Weka và thống kê trong phần mềm R.

Phần mềm R mạnh mẽ cho phân tích dữ liệu vì có thể chạy code mà không cần đến bất cứ compiler nào. R là ngôn ngữ thông dịch ( interpreted language ) do đó code có thể chạy mà không cần compiler. R thông dịch code và làm cho việc viết code đơn giản hơn, dễ phát triển hơn. Bên cạnh đó, R là một vector-language, do đó chúng ta có thể dùng bất kỳ function nào trên một vector mà không cần phải dùng vòng lặp nên bất một phép tính nào cũng có thể thực hiện trên vectors. Và cuối cùng đây là Statistical-Language nên R được dùng trong sinh học, di truyền học và thống kê dữ liệu. R là ngôn ngữ turing-complete có nghĩa nó có thể hoàn thành bất kỳ thuật toán nào.

R cũng được sử dụng nhiều trong kinh doanh, thống kê và lập trình. Lý do quan trọng nhất là vì R là open-source, nên nó cực kỳ "kinh tế". Đồng thời R rất phù hợp cho việc mô phỏng dữ liệu qua bảng biểu. Nhờ một cộng đồng phát triển và hơn 15000+ packages trong mọi lĩnh vực nghiên cứu. Hiện tại, khó có một công cụ nào có thể theo kịp R. Tính toán phân tán (Distributed-Computing) trong lập trình của R là một open source, hiệu năng cao cho ngôn ngữ R. Nó chia công việc thành nhiều task nhỏ cho nhiều node để chạy đồng thời, nhờ đó giảm thời gian để chạy phân tích một tập dữ liệu lớn. Trong việc nghiên cứu dữ liệu, việc khan hiếm nhân tài là một vấn đề rất lớn. Các công ty có thể dùng ngôn ngữ R để làm nền tảng và training nhân viên sử dụng nó.

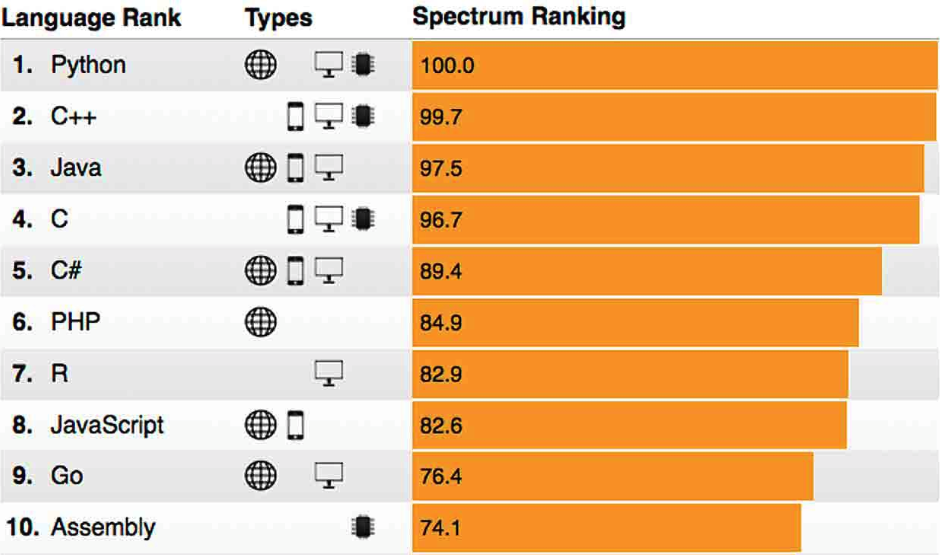
1. Ưu điểm và nhược điểm của phần mềm R:

* **Ưu điểm:** R là một công cụ rất mạnh cho học máy, thống kê và phân tích dữ liệu. Nó là một ngôn ngữ lập trình. Ngôn ngữ R là một platform-independent do đó chúng ta có thể sử dụng nó cho bất kỳ hệ điều hành nào. Việc cài đặt R cũng miễn phí vì thế chúng ta có thể sử dụng mà không cần phải mua bản quyền.  Tuy miễn phí, nhưng chức năng của R không thua kém các phần mềm thương mại. Tất cả những phương pháp, mô hình mà các phần mềm thương mại có thể làm được thì R cũng có thể làm được. R có lợi thế là khả năng phân tích biểu đồ tuyệt vời. Không một phần mềm nào có thể sánh với R về phần biểu đồ! Một lợi thế  khác là R gắn liền với giới học thuật, hầu hết những mô hình thống kê mới nhất đều được hỗ trợ bởi R. Trong các bài báo của tạp chí hàng đầu về các phần mềm thống kê – Journal of Statistical Software – hầu hết là về R. R còn có những package thống kê toàn diện nhất với công nghệ mới nhất, những ý tưởng mới thường xuất hiện đầu tiên trên R và bất kỳ ai cũng được hỗ trợ để đưa ra ý tưởng phát triển, fix bug, phát triển package mới.
* **Nhược điểm:** Nhưng R có cái bất lợi là dùng lệnh (giống như Matlab) chứ không dùng menu như Excel. Điều này có nghĩa là đối với người “lười biếng” thì sẽ thấy R bất tiện. Nhưng với người muốn nắm lấy những cơ chế căn bản của toán học, thì sẽ thích ngôn ngữ này ngay. Một nhược điểm khác là R có rất nhiều “tác giả”, số thư viện/package vào tháng 1/2017 đã hơn 10.000; vấn đề nảy sinh là làm thế nào để chọn thư viện phù hợp nhất (ngoài các thư viện mặc định/có sẵn). Bên cạnh đó thì một vài package của R có thể không hoàn hảo và còn lỗi và R có thể chiếm dụng hết bộ nhớ khả dụng.

***PYTHON***

1. Tổng quan về Python:

* Python là một ngôn ngữ lập trình thông dịch, hướng đối tượng, ngôn ngữ lập trình cấp cao được giải thích với ngữ nghĩa động. Python với triết lý thiết kế của nó rất thuận tiện cho việc đọc hiểu code, đơn giản và rõ ràng được thiết kế 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. Đến nay thì cộng đồng người sử dụng ngôn ngữ này rất đông, nếu so sánh từ bảng xếp hạng các ngôn ngữ năm 2018 thì Python đã leo lên vị trí số 1 trên bảng xếp hạng những ngôn ngữ lập trình phổ biến.
* Python được lấy cảm hứng từ C, Modula-3 và (đặc biệt) ABC.

*Nguồn : IEEE Spectrum The Top Programming languages 2018*

1. Đặc điểm:

* Miễn phí, mã nguồn mở: Bạn có thể tự do sử dụng và phân phối Python, thậm chí là dùng 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 động linh hoạt: Giả sử bạn giả sử bạn đã viết mã Python cho máy Windows của mình. Bây giờ, nếu bạn muốn chạy nó trên máy Mac, bạn không cần phải thay đổi nó như cũ. Nói cách khác, bạn có thể lấy một mã và chạy nó trên bất kỳ máy nào, không cần phải viết mã khác nhau cho các máy khác nhau. Điều này làm cho Python trở thành một ngôn ngữ di động. Tuy nhiên, bạn phải tránh mọi tính năng phụ thuộc hệ thống trong trường hợp này. Nó chạy liền mạch trên hầu hết tất cả các nền tảng như Windows, macOS, Linux.
* Dễ học, dễ đọc: 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 đang trở thành con cưng của nhiều nhà khoa học dữ liệu chỉ vì bộ sưu tập thư viện của nó được thiết kế để phân tích thống kê và phân tích số liệu:
* SciPy – Một bộ sưu tập các gói cho toán học, khoa học và kỹ thuật.
* Pandas – Một thư viện phân tích dữ liệu và mô hình.
* IPython – Một trình bao mạnh mẽ để dễ dàng chỉnh sửa và ghi lại các phiên làm việc. Nó cũng hỗ trợ trực quan hóa và tính toán song song.
* Với Python, bạn có quyền truy cập vào một loạt các thư viện phân tích dữ liệu thông qua mục Python Package như các mô-đun phổ biến NumPy và SciPy. Hai mô-đun này cho phép bạn thực hiện các nhiệm vụ cơ bản trong phân tích dữ liệu là số hóa trên các mảng và ma trận đa chiều cũng như thực hiện tính toán các tín hiệu, hình ảnh. Có rất nhiều thư viện Python khác giúp cho việc phân tích dữ liệu của bạn đơn giản hơn bao giờ hết như Bộ công cụ ngôn ngữ tự nhiên (NLTK), cho phép phân tích và thống kê các ngôn ngữ tự nhiên.

1. Ưu điểm và nhược điểm:

* **Ưu điểm**:
* Python đã có được một phản ứng tích cực từ các nhà khoa học dữ liệu liên quan đến học máy. Do đường cong học tập thấp đối với người dùng, sức mạnh thực sự của Python nằm ở sự đơn giản, dễ đọc và linh hoạt - tất cả được cung cấp bởi một cú pháp chính xác và hiệu quả. Vì là ngôn ngữ lập trình chính thức, Python rất tốt để triển khai các thuật toán để sử dụng sản xuất cũng như tích hợp các ứng dụng web trong các tác vụ phân tích dữ liệu.
* Miễn phí, mã nguồn mở.
* Python có các thư viện rộng lớn giúp giảm đáng kể khoảng thời gian giữa khi bắt đầu dự án và kết quả có ý nghĩa. Kho lưu trữ phần mềm cho ngôn ngữ lập trình Python rất phong phú đến mức Chỉ số gói Python ( [PyPI](https://pypi.python.org/pypi) ) hiện bao gồm 130.641 gói. Thư viện có nhiều môi trường để kiểm tra và so sánh các thuật toán học máy.
* Các gói cung cấp các giải pháp không chỉ trực quan mà còn linh hoạt. Một ví dụ điển hình là [PyBrain](http://pybrain.org/) , đây là một thư viện máy học mô-đun cung cấp các thuật toán mạnh mẽ cho các nhiệm vụ học máy. Được coi là một thư viện máy học phổ biến, [scikit-learn](https://pypi.python.org/pypi/scikit-learn/0.13.1) cung cấp các công cụ khai thác dữ liệu để củng cố khả năng sử dụng máy học vượt trội hiện có của Python.
* **Nhược điểm:**
* Tốc độ chậm hơn so với C hay Java, do rằng mã Python được thực thi theo từng dòng.
* Đối với những tác vụ nào cần chiếm nhiều bộ nhớ thì Python không phải là một lựa chọn tốt.

***SAS***

1. Tổng quan về SAS:

* SAS (viết tắt của **Statistical Analysis System)**  là một bộ phần mềm thống kê được phát triển bởi Viện SAS vào năm 1966 để phân tích nâng cao, phân tích đa biến, kinh doanh thông minh, điều tra tội phạm và quản lý dữ liệu.
* SAS được phát triển thêm vào những năm 1980 và 1990. SAS có thể dễ dàng truy cập, quản lý và có thể phân tích dữ liệu từ bất kỳ nguồn nào. SAS đã giới thiệu một bộ sản phẩm lớn vào năm 2011 cho trí thông minh của khách hàng và nhiều module cho web, phương tiện truyền thông xã hội và phân tích tiếp thị được sử dụng rộng rãi để thu hút khách hàng và khách hàng tiềm năng.

1. Đặc điểm:

* Toàn bộ nghiệp vụ của SAS được xây dựng trên các phân tích. Nó cung cấp một giải pháp quản lý dữ liệu gốc gọi là Quản trị dữ liệu SAS để giúp các tổ chức chuẩn bị và quản lý cả nguồn dữ liệu lớn và dữ liệu truyền thống. Nó cho phép người dùng duy trì và quản lý các thuộc tính dữ liệu thông qua một mô hình dữ liệu chung, gắn cờ thay đổi siêu dữ liệu, tạo ảnh chụp nhanh, lưu trữ và quản lý danh sách, phân cấp và tạo báo cáo về tình trạng dữ liệu và mọi khắc phục cần thiết.
* SAS được viết bằng C, chạy trên  các hệ điều hành: Windows, IBM mainframe, Unix/Linux, OpenVMS Alpha.
* SAS cho phép thực hiện:
* Nhập, truy cập, quản lý và khai thác dữ liệu
* Xuất báo cáo bằng văn bản và đồ họa
* Phân tích thống kê
* Lập kế hoạch kinh doanh, dự báo, và ra quyết định hỗ trợ
* Hoạt động nghiên cứu và quản lý dự án
* Cải tiến chất lượng
* Các ứng dụng phát triển
* Kho dữ liệu (trích xuất, chuyển đổi, tải)
* Nền tảng điện toán độc lập và từ xa
* Ngoài ra, SAS còn cung cấp nhiều giải pháp phần mềm cho các mảng khác như: quản trị  IT, quản trị nguồn nhân lực, quản lý tài chính, kinh doanh thông minh, quản lý quan hệ khách hàng…
* Hầu hết các tập đoàn tài chính – ngân hàng và những người làm PhD (tại Mỹ) sử dụng SAS.

1. Ưu điểm và nhược điểm:

* **Ưu điểm:**
* SAS rất mạnh trong lĩnh vực quản lý dữ liệu, cho phép người sử dụng thao tác dữ liệu hầu như với mọi cách có thể
* SAS có thể làm việc với nhiều file dữ liệu cùng một lúc; điều này giảm đi tính phức tạp trong chuẩn bị dữ liệu đối với những nhiệm vụ phân tích đòi hỏi phải làm việc với nhiều file dữ liệu cùng một lúc.
* SAS có thể quản lý những file dữ liệu khổng lồ lên đến 32.768 biến và số lượng bản ghi là rất lớn chỉ phụ thuộc vào kích cỡ của đĩa cứng.
* Đối với phân tích thống kê, sức mạnh lớn nhất của SAS có thể tìm thấy trong phân tích ANOVA, phân tích mô hình hỗn hợp và phân tích nhiều chiều.
* SAS có các công cụ vẽ đồ thị mạnh nhất (SAS/Graph) so với hai bộ chương trình còn lại.
* SAS có dịch vụ chăm sóc khách hàng khá tận tâm cho nên nếu bạn gặp vấn đề trong việc cài đặt hay những vấn đề kỹ thuật nào khác đều có thể liên hệ với họ.
* **Nhược điểm:**
* Mất nhiều thời gian để học và hiểu được cách quản lý dữ liệu của SAS và nhiều nhiệm vụ quản lý phức tạp của nó lại được thực hiện bằng những lệnh đơn giản trong SPSS và STATA.
* SAS là một phần mềm thương mại cho nên nó sẽ khá đắt và vượt xa tầm với của hầu hết các chuyên gia (trong phạm vi cá nhân).
* Yếu đối với hồi quy logistic kiểu thứ tự và kiểu phạm trù (vì các lệnh này là đặc biệt khó) và các phương pháp ước lượng mạnh.
* Nó cũng có hỗ trợ một ít cho phân tích dữ liệu theo lược đồ mẫu, nhưng lại hạn chế hơn so với STATA.

***TABLEAU***

1. Tổng quan:

* Lịch sử: Tableau được thành lập vào năm 2003 bởi Pat Hanrahan, Christian Chabot và Chris Stolte, và chuyển trụ sở chính đến khu phố [Fremont](https://en.wikipedia.org/wiki/Fremont,_Seattle) của [Seattle](https://en.wikipedia.org/wiki/Seattle), Washington. Vào tháng 6 năm 2018, Tableau đã mua lại Empirical Systems, một công ty khởi nghiệp [trí tuệ nhân tạo](https://en.wikipedia.org/wiki/Artificial_intelligence) có trụ sở tại [Cambridge, Massachusetts](https://en.wikipedia.org/wiki/Cambridge,_Massachusetts) , với kế hoạch tích hợp công nghệ của công ty vào nền tảng Tableau. Vào ngày 10 tháng 6 năm 2019, Tableau đã được [Salesforce](https://en.wikipedia.org/wiki/Salesforce.com) mua lại. Với ý chí "Making data easy to see and understand is one of the great opportunities of our time".
* Đặc điểm:
* Tableau là công cụ thực hiện các nghiệp vụ phân tích dữ liệu một cách nhanh chóng, đơn giản và trực quan dành cho tất cả mọi người. Đối với các phòng ban nghiệp vụ, để sử dụng dữ liệu để định hướng các hoạt động kinh doanh phải cần một môi trường có thể  phân tích và xử lý dữ liệu dưới dạng đồ họa, tuy nhiên nhiều trường hợp, để làm được điều đó thì cần phải có kỹ năng và kiến thức lập trình, nên có rất nhiều yêu cầu phân tích, report gửi đến phòng IT hoặc phòng ban chuyên môn .
* Với Tableau, ngay cả những người không có chuyên môn hoặc kỹ năng lập trình, đều có thể lấy dữ liệu cần thiết từ nguồn dữ liệu, chỉ cần thao tác trực quan bằng cách kéo và thả, thể hiện và xử lý dữ liệu dưới dạng dễ nhận biết, có thể hình dung dữ liệu một cách nhanh chóng là các đặc điểm của Tableau.
* Tốc độ xử lý nhanh, việc phân tích dữ liệu liên tục không bị gián đoạn.
* Hỗ trợ cả những người dùng cá nhân, bằng cách cài đặt trên PC, bạn có thể nhanh chóng xây dựng một môi trường phân tích dựa trên dữ liệu đang có.
* Các dữ liệu được thể hiện một cách đa dạng và trở thành thông tin có ích
* Mục tiêu của Tableau rất đơn giản: "help people see and understand their data" - và Tableau đã phát triển giá trị cốt lõi của mình dựa trên:

1. Ưu điểm, nhược điểm

* **Ưu điểm:**
* Có phiên bản miễn phí.
* Có thể tương tác với bất kì loại dữ liệu nào từ Excel, Data Warehouse cho tới Dữ liệu Website.
* Khả năng cập nhật Dữ liệu theo thời gian thực.
* Làm Dữ liệu trở nên trực quan bằng nhiều cách như Biểu đồ hay thậm chí là cả một Dashboard – Tốt hơn bất kỳ phần mềm mềm nào khác trên thị trường.
* Hệ thống xử lý Big Data của Tableau rất mạnh mẽ.
* **Nhược điểm:** Những nhà cung cấp khác cũng đang bắt kịp công nghệ phân tích trực quan, và do đó nó không còn là ưu thế độc quyền của Tableau nữa. Ngoài ra, các đối thủ cũng đang gây sức ép với chính sách giá mềm hơn.

***KNIME***

1. Tổng quan

* Lịch sử: KINIME tên đầy đủ là Konstanz Information Miner, là một nền tảng phân tích, tích hợp và báo cáo dữ liệu mã nguồn mở. Nó tích hợp các thành phần khác nhau của data mining và machine learning thông qua khái niệm module data pipeline. Giao diện người dùng đồ họa cho phép lắp ráp các node để xử lý trước dữ liệu (bao gồm trích xuất, chuyển đổi và tải), mô hình hóa dữ liệu, trực quan hóa và phân tích dữ liệu. Từ năm 2006, KNIME đã được sử dụng rộng rãi trong nghiên cứu dược phẩm và hiện tại mở rộng sang các lĩnh vực như phân tích dữ liệu khách hàng trong CRM, phân tích dữ liệu tài chính và kinh doanh thông minh.
* KNIME là một nền tảng có thể giúp chúng ta giải quyết bất kỳ vấn đề nào mà chúng ta có thể nghĩ đến trong ranh giới của khoa học dữ liệu ngày nay. Từ các trực quan cơ bản nhất hoặc hồi quy tuyến tính đến học sâu nâng cao, KNIME có thể làm tất cả.
* KNIME cho phép bạn thao tác, phân tích và mô hình hóa dữ liệu một cách vô cùng trực quan thông qua visual programming. Về cơ bản, thay vì viết các khối mã lệnh, bạn chỉ cần kéo thả các nodes lên canvas và kết nối các node này lại với nhau để tạo ra tiến trình phân tích dữ liệu. Quan trọng hơn, KNIME có thể nhúng mã lệnh R, python, khai thác văn bản, dữ liệu hóa học. Ngoài ra KNIME còn cung cấp cho bạn những tùy chọn để vọc những mã lệnh theo hướng tiếp cận tiên tiến hơn.

1. Ưu điểm, nhược điểm

* **Ưu điểm:** Đối với những người chưa quen với nền tảng của công cụ khai thác dữ liệu này, KNIME đã tập hợp một loạt các khóa học ngắn để hiểu rõ hơn về khoa học dữ liệu và cách sử dụng nền tảng một cách hiệu quả.
* **Nhược điểm**:KNIME, là một công cụ nguồn mở rất mạnh mẽ, có những hạn chế riêng. Những cái chính là: Các hình ảnh trực quan không gọn gàng và bóng bẩy như một số phần mềm nguồn mở khác là (ví dụ RStudio); Cập nhật phiên bản không được hỗ trợ tốt; bạn sẽ phải cài đặt lại phần mềm (tức là để cập nhật KNIME từ phiên bản 2 lên phiên bản 3, bạn sẽ cần cài đặt mới và cập nhật sẽ không hoạt động); Cộng đồng đóng góp không lớn như cộng đồng Python hoặc CRAN, vì vậy phải mất một thời gian dài để bổ sung mới cho KNIME

***SPLUNK***

1. Tổng quan

* Splunk là một tập đoàn đa quốc gia của Mỹ có trụ sở tại San Francisco, California, làm phần mềm để thu thập Log, tìm kiếm, theo dõi và phân tích dữ liệu lớn (big data) do máy tạo ra, thông qua một giao diện web nhằm giải quyết nhiều bài toán khác nhau của các tổ chức, doanh nghiệp như trong việc giám sát, vận hành hệ thống, điều tra sự cố.v.v...
* Phần mềm Splunk thu thập, đánh chỉ mục dữ liệu, tìm kiếm trong thời gian thực trong một kho lưu trữ dữ liệu có thể tìm kiếm, từ đó nó tạo ra các đồ thị, báo cáo, cảnh báo, biểu đồ.
* Splunk có sứ mệnh làm cho dữ liệu máy có thể truy cập được trong một tổ chức bằng cách xác định mô hình dữ liệu, cung cấp số liệu, chẩn đoán vấn đề và cung cấp thông tin cho các hoạt động kinh doanh. Splunk được sử dụng để quản lý ứng dụng, bảo mật và tuân thủ theo chính sách tổ chức, cũng như kinh doanh và phân tích web. Tính đến đầu năm 2016, Splunk có hơn 10.000 khách hàng trên toàn thế giới.

1. Chi tiết

Chính sách bản quyền: Splunk cung cấp 2 bộ sản phẩm miễn phí và trả phí

* Sản phẩm trả phí: Có tất cả các chức năng của Splunk, không hạn chế kích thước dữ liệu.
* Sản phẩm miễn phí: Hạn chế một số chức năng, hạn chế khối lượng dữ liệu mỗi ngày là 500MB. Bao gồm các chức năng: Đánh chỉ mục dữ liệu, tìm kiếm trong thời gian thực, thống kế và kết xuất báo cáo.

Tính năng định dạng Log: Splunk hỗ trợ hầu như tất cả các loại log của hệ thống, thiết bị hạ tầng mạng, phần mềm, Firewall, IDS/IPS, Log Event, Register của các máy trạm… Các hình thức thu thập dữ liệu: Splunk có thể thực hiện việc thu thập log từ rất nhiều nguồn khác nhau.

\* Từ một file hoặc thư mục (kể cả file nén) trên server

\* Qua các kết nối UDP, TCP từ các Splunk Server khác trong mô hình Splunk phân tán

\* Từ các Event Logs, Registry của Windows…

Splunk cũng kết hợp rất tốt với các công cụ thu thập log khác để tăng hiệu năng của hệ thống.

Cập nhật dữ liệu: Splunk cập nhật dữ liệu liên tục khi có thay đổi trong thời gian thực. Giúp cho việc phát hiện và cảnh báo chính xác trong thời gian thực.

Đánh chỉ mục dữ liệu: Splunk được xây dựng trên Lucence, có thể đánh chỉ mục dữ liệu với một khối lượng dữ liệu rất lớn trong một khoảng thời gian ngắn. Giúp việc tìm kiếm diễn ra nhanh chóng và thuận tiện.

Tìm kiếm thông tin: Splunk làm việc rất tốt với dữ liệu lớn và cập nhật liên tục. Nó cung cấp cơ chế tìm kiếm với một “Splunk Language” cực kỳ thông minh bao gồm các từ khóa, các hàm và cấu trúc tìm kiếm giúp người sử dụng có thể truy xuất mọi thứ, theo rất nhiều tiêu chí từ tập dữ liệu rất lớn. Những nhà quản trị mạng cao cấp và chuyên nghiệp thường gọi Splunk với cái tên “Splunk toàn năng” hay “Splunk as Google for Log files” để nói lên sức mạnh của Splunk.

Giám sát an ninh mạng và cảnh báo: Splunk cung cấp cho người dùng một cơ chế cảnh báo dựa trên việc tìm kiếm các thông tin do chính người sử dụng đặt ra. Khi có vấn đề liên quan tới hệ thống phù hợp với các tiêu chí mà người dùng đã đặt ra thì hệ thống sẽ cảnh báo ngay tới người dùng (cảnh bảo trực tiếp qua giao diện, gửi Email).

Khắc phục sự cố: Splunk còn cung cấp một cơ chế tự động khắc phục với các vấn đề xảy ra bằng việc cấu hình để tự động chạy các file Script mà người dùng tự tạo (Ví dụ như: Chặn IP, đòng Port …) khi có các cảnh báo xảy ra.

Hiển thị thông tin: Splunk cung cấp một cơ chế hiển thị trực quan giúp người sử dụng có thể dễ dàng hình dung về tình trạng của hệ thống, đưa ra các đánh giá về hệ thống. Splunk còn từ động kết xuất ra các báo cáo với nhiều loại định dạng một cách chuyên nghiệp.

Phát triển: Cũng cung cấp các API hỗ trợ việc tạo các ứng dụng trên Splunk của người dùng. Một số bộ API điển hình như Splunk SDK (Cung cấp các SDK trên nền tảng Python, Java, JS, PHP), Shep (Splunk Hadoop Intergration – Đây là sự kết hợp giữa Splunk và Hadoop), Shuttl (Là một sản phẩm hỗ trợ việc sao lưu dữ liệu trong Splunk), Splunkgit (Giúp bạn hình dung dữ liệu tốt hơn), Splunk power shell resource Kit (Bộ công cụ hỗ trợ việc mở rộng và quản lý hệ thống).

*Lưu ý*: Splunk mạnh về khả năng phân tích và cảnh báo tuy nhiên nó lại không mạnh và không đảm bảo về việc thu thập và truyền tải log. Cụ thể là nó chưa có cơ chế bảo mật trên đường truyền, không phù hợp với những hệ thống đòi hỏi bảo mật cao. Để phát huy hết được sức mạnh của Splunk cần có thời gian tìm hiểu và sử dụng. Splunk chưa có cơ chế giúp tự động phát hiện ra các tấn công hay các vấn đề từ bên ngoài. Những điều này phụ thuộc vào kinh nghiệm sử dụng và vốn hiểu biết về bảo mật của người quản trị. Đề triển khai được một hệ thống sử dụng Splunk hiệu quả chúng ta cũng cần có một hệ thống riêng, đây cũng là một trở ngại không nhỏ với các hệ thống có quy mô trung bình và nhỏ.

1. Ưu và nhược điểm

* **Ưu điểm:**
* Hỗ trợ đa dạng trên các máy trạm, Firewall, IDS/IPS, Log Event..
* Không ngừng nghỉ cập nhật dữ liệu trong thời gian thực
* Cơ chế tìm kiếm thông minh bao gồm các từ khóa, các hàm và cấu trúc tìm kiếm, từ đó bạn có thể truy xuất mọi thứ theo mong muốn
* Tự động khắc phục sự cố
* **Nhược điểm:**
* Không thích hợp với các hệ thống có bảo mật cao
* Thời gian tìm hiểu, sử dụng và vận hành khá lâu
* Đặc biệt phải có một hệ thống riêng đủ lớn. Và tất nhiên, Splunk không phù hợp với các hệ thống có quy mô trung bình và nhỏ.

***SPARK***

1. Tổng quan về Apache Spark

* Apache Spark là một open source cluster computing framework được phát triển sơ khởi vào năm 2009 bởi AMPLab tại đại học California. Sau này, Spark đã được trao cho Apache Software Foundation vào năm 2013 và được phát triển cho đến nay. Nó cho phép xây dựng các mô hình dự đoán nhanh chóng với việc tính toán được thực hiện trên một nhóm các máy tính, có có thể tính toán cùng lúc trên toàn bộ tập dữ liệu mà không cần phải trích xuất mẫu tính toán thử nghiệm. Tốc độ xử lý của Spark có được do việc tính toán được thực hiện cùng lúc trên nhiều máy khác nhau. Đồng thời việc tính toán được thực hiện ở bộ nhớ trong (in-memories) hay thực hiện hoàn toàn trên RAM.
* Thành phần của Apache Spark
* Matei Zaharia, cha đẻ của Spark, sử dụng Hadoop từ những ngày đầu. Đến năm 2009 ông viết Apache Spark để giải quyết những bài toán học máy ở đại học UC Berkely vì Hadoop MapReduce hoạt động không hiệu quả cho những bài toán này. Rất sớm sau đó ông nhận ra rằng Spark không chỉ hữu ích cho học máy mà còn cho cả việc xử lý luồng dữ liệu hoàn chỉnh.
* Thành phần trung của Spark là Spark Core: cung cấp những chức năng cơ bản nhất của Spark như lập lịch cho các tác vụ, quản lý bộ nhớ, fault recovery, tương tác với các hệ thống lưu trữ…Đặc biệt, Spark Core cung cấp API để định nghĩa RDD (Resilient Distributed DataSet) là tập hợp của các item được phân tán trên các node của cluster và có thể được xử lý song song.
* Spark có thể chạy trên nhiều loại Cluster Managers như Hadoop YARN, Apache Mesos hoặc trên chính cluster manager được cung cấp bởi Spark được gọi là Standalone Scheduler.
* Spark SQL cho phép truy vấn dữ liệu cấu trúc qua các câu lệnh SQL. Spark SQL có thể thao tác với nhiều nguồn dữ liệu như Hive tables, Parquet, và JSON
* Spark Streaming cung cấp API để dễ dàng xử lý dữ liệu stream,
* MLlib Cung cấp rất nhiều thuật toán của học máy như: classification, regression, clustering, collaborative filtering…
* GraphX là thư viện để xử lý đồ thị.
* Trong các thư viện mà Spark cung cấp thì có 69% người dùng Spark SQL, 62% sử dụng DataFrames, Spark Streaming và MLlib + GraphX là 58%

1. Tính năng nổi bật

* “Spark as a Service”: Giao diện REST để quản lí (submit, start, stop, xem trạng thái) spark công việc, spark nội dung; Tăng tốc, giảm độ trễ thực thi công việc xuống mức chỉ tính bằng giây bằng cách tạo sẵn spark context cho các job dùng chung; Stop job đang chạy bằng cách stop spark context; Bỏ bước upload gói jar lúc start job làm cho job được start nhanh hơn; Cung cấp hai cơ chế chạy job đồng bộ và bất đồng bộ; Cho phép cache RDD theo tên , tăng tính chia sẻ và sử dụng lại RDD giữa các job; Hỗ trợ viết spark job bằng cú pháp SQL; Dễ dàng tích hợp với các công cụ báo cáo như: Business Intelligence, Analytics, Data Integration Tools; Quản lý bộ nhớ của Apache Spark
  + Về bộ nhớ, Spark giải quyết các vấn đề xung quanh định nghĩa Resilient Distributed Datasets (RDDs). RDDs hỗ trợ hai kiểu thao tác: transformations và action. Thao tác chuyển đổi(tranformation) tạo ra dataset từ dữ liệu có sẵn. Thao tác actions trả về giá trị cho chương trình điều khiển (driver program) sau khi thực hiện tính toán trên dataset.
  + Spark thực hiện đưa các thao tác RDD chuyển đổi vào DAG (Directed Acyclic Graph) và bắt đầu thực hiện. Khi một action được gọi trên RDD, Spark sẽ tạo DAG và chuyển cho DAG scheduler. DAG scheduler chia các thao tác thành các nhóm (stage) khác nhau của các task. Mỗi Stage bao gồm các task dựa trên phân vùng của dữ liệu đầu vào có thể pipline với nhau và có thể thực hiện một cách độc lập trên một máy worker. DAG scheduler sắp xếp các thao tác phù hợp với quá trình thực hiện theo thời gian sao cho tối ưu nhất. Ví dụ: các thao tác map sẽ được đưa vào cùng một stage do không xảy ra shuffle dữ liệu giữa các stage. Kết quả cuối cùng của DAG scheduler là một tập các stage. Các Stages được chuyển cho Task Scheduler. Task Scheduler sẽ chạy các task thông qua cluster manager (Spark Standalone/Yarn/Mesos). Task scheduler không biết về sự phụ thuộc của các stages. Nó chỉ chịu trách nhiệm thực hiện sắp xếp các task một cách tối ưu nhất.
  + Mỗi Worker bao gồm một hoặc nhiều Excuter. Các excuter chịu trách nhiệm thực hiện các task trên các luồng riêng biệt. Việc chia nhỏ các task giúp đem lại hiệu năng cao hơn, giảm thiểu ảnh hưởng của dữ liệu không đối xứng (kích thước các file không đồng đều).
  + Quá trình Spark xây dựng DAG: Có hai kiểu chuyển đổi có thể áp dụng trên các RDDs đó là chuyển đổi hẹp và chuyển đổi rộng. Chuyển đổi hẹp: không yêu cầu xáo trộn dữ liệu vượt qua các phân vùng (partition).Ví dụ như các thao tác map, filter,..

***LOOKER***

1. Tổng quan

* Looker là một ứng dụng khám phá dữ liệu cung cấp các chức năng khám phá dữ liệu sáng tạo cho các doanh nghiệp cả lớn và nhỏ. Với nó, họ có thể truy cập vào giao diện dựa trên web nơi họ có thể dễ dàng nhận được thông tin chi tiết theo thời gian thực về hoạt động của mình thông qua phân tích dữ liệu. Họ có thể tạo báo cáo khi đang di chuyển và *làm cho tất cả các bên quan tâm có thể truy cập được*, vì vậy các thành viên khác trong nhóm có thể đóng góp cho các cuộc thảo luận về một số nhiệm vụ nhất định và liên tục phát triển trong dự án của họ. Do đó, Looker có thể giúp tất cả các công ty sử dụng dữ liệu để điều khiển các quyết định và hoạt động kinh doanh của họ đi đúng hướng.
* Với một chút kiến ​​thức về SQL, bạn sẽ không gặp phải bất kỳ vấn đề nào khiến Looker hoạt động cho bạn hoặc xây dựng các mô-đun phân tích của riêng bạn. Bạn có thể thiết kế trực quan hóa với một mã duy nhất, vì độ phức tạp chỉ phụ thuộc vào mức độ bạn dự định đi để tinh chỉnh trải nghiệm của bạn. Mã ML của Looker có mặt để giúp bạn sản xuất và tối ưu hóa các truy vấn của mình, điều này đặc biệt tiện dụng vì hầu hết các hệ thống BI mong muốn bạn trở thành một chuyên gia trong khu vực.
* Looker vẫn hoạt động khi bạn không có kiến ​​thức về SQL, vì chúng có cơ sở dữ liệu phong phú về video và tài liệu học tập, giống như các bản ghi trực tiếp và các bài giảng diễn viên trên màn hình. Tài liệu cũng bao gồm các câu đố tương tác sẽ làm hài lòng các nhóm sáng tạo đang tìm cách chuyển đổi phân tích thành một hoạt động thú vị.
* Như bạn sẽ đọc trong phần Lợi ích, có rất nhiều Looker có thể cung cấp ngoài các khối tích hợp API mở và giá cả linh hoạt. Nó phân tích cả thông tin SQL được lưu trữ trên web và cung cấp hơn 25 biến thể dữ liệu, trong đó có Hive, Vertica và BigQuery của Google. Nền tảng được thiết kế để chấm dứt sự hỗn loạn dữ liệu và tắc nghẽn, và làm như vậy trong một giao diện trực quan, nhân viên không mất thời gian để hiểu. Hiện tại, Looker là ứng dụng kinh doanh thông minh ưa thích của hơn 700 công ty trên toàn thế giới, giúp họ khiến khách hàng hài lòng và khám phá cách biến lưu lượng truy cập thành thông tin Thương mại điện tử có giá trị. Trước khi bạn liên hệ với họ để báo giá và mua hàng, chúng tôi khuyên bạn nên yêu cầu bản demo miễn phí và khám phá các tính năng trước.

1. Tổng quan về lợi ích của Looker

* Looker là một nền tảng khám phá dữ liệu giúp các công ty đưa ra quyết định kinh doanh tốt hơn thông qua truy cập dữ liệu theo thời gian thực. Dữ liệu của tất cả các kích thước có thể được phân tích trong cơ sở dữ liệu 100% trong cơ sở dữ liệu và 100% dựa trên trình duyệt. Hiện tại, Looker Analytics tích hợp với bất kỳ cơ sở dữ liệu hoặc kho dữ liệu SQL nào như Greenplum và Amazon Redshift. Được phát triển với khả năng sử dụng trong tâm trí, Looker đáp ứng nhu cầu của các công ty khởi nghiệp, doanh nghiệp và doanh nghiệp cỡ trung bình.
* Nhưng chính xác thì Looker có thể giúp đội của bạn như thế nào? Chúng ta hãy xem xét một số lợi ích quan trọng nhất thường được liên kết với hệ thống này:*Dễ sử dụng:* Để bắt đầu, Looker cực kỳ dễ vận hành. Bạn không phải là một nhà phân tích dữ liệu có kinh nghiệm hoặc có kiến ​​thức về SQL để quản lý các quy trình công việc từ đó, vì hầu hết các hoạt động đều đi xuống một mã tự trực quan duy nhất. Không có mã hóa phức tạp và bạn có thể đi sâu vào dữ liệu trên các biểu đồ và bảng tương tác và chính xác. Nền tảng này cũng bao gồm các tính năng tự phục vụ như lọc, xoay vòng và tạo trực quan hóa và bảng điều khiển; *Hình dung tiện dụng:* Tất cả các hoạt động chung liên quan đến khám phá dữ liệu được đối chiếu trong menu bên trái, nơi bạn cũng có thể quản lý cài đặt của mình và phát triển các không gian làm việc được cá nhân hóa. Việc tạo trực quan hóa, có một sơ đồ dữ liệu đặc biệt nơi bạn có thể kéo và thả các yếu tố mong muốn và gán các vai trò như Quản trị viên, Người dùng và Nhà phát triển. Một tính năng lập bản đồ cũng có sẵn cho người dùng thiếu kinh nghiệm để chuẩn bị trực quan hóa kỹ thuật cao, nhập hoặc xuất những thứ bên ngoài bảng điều khiển của Looker. Looker cũng cho phép sử dụng Webhooks để kết nối các đối tượng dữ liệu với quy trình công việc của bên thứ ba, đó là cách nó loại bỏ nhu cầu bổ sung chức năng của nó với các ứng dụng bên ngoài; *Các tính năng cộng tác mạnh mẽ:* Để làm cho vấn đề trở nên tốt hơn, Looker cho phép và hợp lý hóa sự hợp tác giữa những người dùng. Ứng dụng có thể truy cập trên mọi trình duyệt cũng như trên thiết bị di động, điều đó có nghĩa là nhân viên sẽ được thông báo và đóng góp khi ra khỏi văn phòng. Ví dụ: dữ liệu và báo cáo có thể được chia sẻ qua email hoặc URL hoặc dễ dàng tích hợp vào các ứng dụng khác nhau như Excel, Google Docs và thậm chí các ứng dụng của bên thứ ba và tùy chỉnh. Cuối cùng, dữ liệu cũng có thể được xuất thủ công trong tệp văn bản hoặc tệp .csv nếu người dùng thích như vậy.

***QLIK VIEW***

1. Tổng quan

* QlikView là một nền tảng khám phá kinh doanh cung cấp BI tự phục vụ cho tất cả người dùng doanh nghiệp trong các tổ chức. Với QlikView, bạn có thể phân tích dữ liệu và sử dụng các khám phá dữ liệu của mình để hỗ trợ ra quyết định. QlikView cho phép bạn hỏi và trả lời các câu hỏi của riêng bạn và đi theo con đường của riêng bạn để hiểu rõ hơn và nó cho phép bạn cùng đồng nghiệp của mình đưa ra quyết định hợp tác.
* Cốt lõi của QlikView là công cụ phần mềm được cấp bằng sáng chế của chúng tôi, tạo ra các chế độ xem dữ liệu mới một cách nhanh chóng. QlikView nén dữ liệu và giữ nó trong bộ nhớ, nơi nó có sẵn để khám phá ngay lập tức bởi nhiều người dùng. Đối với các tập dữ liệu quá lớn để phù hợp với bộ nhớ, QlikView kết nối trực tiếp với nguồn dữ liệu. QlikView cung cấp trải nghiệm kết hợp trên tất cả các dữ liệu được sử dụng để phân tích, bất kể nó được lưu trữ ở đâu. Bạn có thể bắt đầu ở bất cứ đâu và đi bất cứ đâu; và không giới hạn ở đường dẫn khoan được xác định trước hoặc bảng điều khiển được cấu hình sẵn
* QlikView là một trong những công cụ BI hàng đầu theo báo cáo Gartner Magic Quadrant cho năm 2019. QlikView cung cấp tính năng lưu trữ trong bộ nhớ giúp thu thập, tích hợp và xử lý dữ liệu rất nhanh. Các báo cáo được tạo bằng các công cụ trực quan và mối quan hệ giữa dữ liệu được lấy tự động bởi phần mềm QlikView. Nói cách khác*, QlikView là một công cụ khám phá dữ liệu tạo điều kiện cho việc tạo các ứng dụng động để phân tích dữ liệu.*
* QlikView chủ yếu là một công cụ khám phá dữ liệu và do đó nó có một số tính năng trực quan hóa dữ liệu riêng biệt. Data Discovery là một tìm kiếm hướng đến người dùng cho các mẫu và xu hướng trong các tập dữ liệu. Nó giúp người dùng hiểu và nhìn thấy các mẫu này bằng cách cung cấp các công cụ hỗ trợ trực quan như biểu đồ, bảng biểu, bản đồ, v.v. QlikView cũng là duy nhất vì tính linh hoạt, tính năng trong bộ nhớ và hỗ trợ cộng tác.

2. Các tính năng của QlikView: Khám phá dữ liệu độc đáo và tìm kiếm toàn cầu; Hình dung tương tác; Hợp tác; Kiểm soát tuyệt đối dữ liệu; Môi trường làm việc an toàn; Tính linh hoạt và tích hợp; Báo cáo nhất quán.

**QlikView hỗ trợ các định dạng sau:** QVW, QVR, QVA, HTML, TXT, XLS,..

**Ưu điểm:**

* Khả năng hiển thị dữ liệu
* Miễn phí cho sử dụng cá nhân
* Giao diện người dùng ấn tượng
* khám phá các xu hướng / mẫu dữ liệu và cung cấp phạm vi cho cộng tác trong thời gian thực. (Nó cũng cho phép bạn tìm kiếm các loại kết quả dữ liệu bạn muốn, ngay lập tức thể hiện mối quan hệ dữ liệu).
* Sử dụng QlikView cho công việc ngoại tuyến nếu bạn có thể truy cập dữ liệu cần thiết.

**Nhược điểm:**

* Một số kiến thức lập trình cần thiết
* Biến thể thương mại hơi đắt tiền

***QLIK SENSE***

Qlik Sense cũng là một phần mềm phân tích và hiển thị dữ liệu phổ biến. Tại cốt lõi của nó, nó hoạt động với một công cụ QIX kết hợp. Công cụ này cho phép người dùng liên kết và liên kết dữ liệu từ các nguồn khác nhau để thực hiện phân tích. Qlik Sense phục vụ như một nền tảng phân tích dữ liệu cho nhiều người dùng, từ người dùng không có kỹ thuật đến người dùng kỹ thuật.

Qlik Sense tập trung nhiều hơn vào trực quan hóa dữ liệu vì nó có đồ họa tăng cường. Tuy nhiên, trong QlikView, bạn có thể thao tác dữ liệu theo nhiều cách kỹ thuật thông qua kịch bản. Nếu động cơ của bạn khi sử dụng Qlik Sense là trực quan hóa và phân tích dữ liệu trong đồ họa tốt nhất có thể, thì bạn đã lựa chọn đúng.

Qlik Sense cung cấp rất nhiều tính linh hoạt cho người dùng khi họ có thể thực hiện các hoạt động hoàn toàn độc lập với các phân tích và trực quan hóa tự phục vụ. Ngoài ra, chúng có thể được hướng dẫn bởi phân tích hướng dẫn máy tự động bằng động cơ nhận thức của Qlik Sense. Qlik Sense sử dụng Mô hình kết hợp trong đó người dùng có thể tự do khám phá dữ liệu rộng lớn và phức tạp và rút ra những hiểu biết trực quan từ nó. Có thể tích hợp các tệp dữ liệu lớn từ nhiều nguồn trong Qlik Sense. Các khách hàng có thể chia sẻ các ứng dụng dữ liệu và báo cáo về một trung tâm tập trung. Cùng với điều này, họ có thể chia sẻ các mô hình dữ liệu an toàn, xuất các câu chuyện dữ liệu, vv để tăng cường kinh doanh của họ.

***Các tính năng của Qlik Sense***: Mô hình liên kết; Trực quan và phân tích thông minh; Tự tạo; Chia sẻ và hợp tác tập trung; Kể chuyện và báo cáo dữ liệu; Ứng dụng di động; Chuẩn bị và tích hợp dữ liệu; Động cơ QIX; Quản trị doanh nghiệp và khả năng mở rộng.

-         **Phần hai: Từ những nghiên cứu về công cụ khai phá dữ liệu và lựa chọn công cụ khai phá dữ liệu Tableau để để khai thác dữ liệu điểm của sinh viên trường đại học.**

* So sánh điểm của các sinh viên trong những học kỳ trước đó.
* Điểm số của những học kỳ trước ảnh hưởng như thế nào đến kết quả của những học kỳ sau?
* Có điểm khác biệt giữa điểm số của những môn chuyên ngành và điểm số của những môn đại cương không? Lý do dẫn đến sự khác biệt này (nếu có)?
* Phân tích lý do dẫn đến sự thay đổi theo hướng tích cực (hay tiêu cực) thấy được qua điểm số.
* Phân tích những yếu tố tác động đến thành tích học tập của sinh viên trường đại học.
* Cách sinh viên trường đại học nâng cao điểm số?

-         **Phần ba: Đưa ra dự đoán kết quả học tập của sinh viên qua dữ liệu đã khai thác được.**

Qua phân tích, sinh viên thấy được dự đoán kết quả học tập nhằm kế hoạch học tập hiệu quả. Bên cạnh đó, cố vấn học tập sẽ có những định hướng phát triển cũng như các chiến lược để củng cố nguồn lực tiềm năng của đất nước.

**11.****TÀI LIỆU THAM KHẢO**

**Tiếng việt:**

[1] Nguyễn Thái Nghe, Khoa Công nghệ thông tin và Truyền thông, Trường Đại học Cần Thơ - Ứng dụng các kỹ thuật trong khai phá dữ liệu hỗ trợ sinh viên lập kế hoạch học tập.

[2]  Nguyễn Thị Thanh Thủy - Ứng dụng khai phá dữ liệu xây dựng công cụ dự đoán kết quả học tập của Sinh viên, Tuyển tập Báo cáo Hội nghị Sinh viên Nghiên cứu Khoa học lần thứ 8 Đại học Đà Nẵng năm 2012.

[3] Trịnh Thị Nhị (2011) - Nghiên cứu, khai thác kho dữ liệu điểm tại trường Đại học Sư phạm kỹ thuật Hưng Yên dựa trên bộ công cụ BI của hệ quản trị CSDL SQL Server 2008, Luận văn Thạc sĩ, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội.

[4] Nguyễn Bảo Nam (2012) - Nghiên cứu ứng dụng khai phá dữ liệu chuỗi thời

gian áp dụng vào bài toán dự đoán kết quả học sinh - luận văn thạc sĩ.

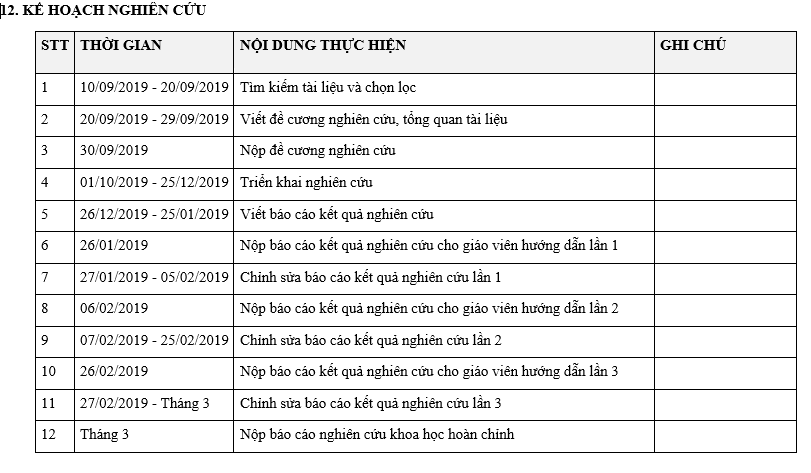
[5]  Hà Quang Thụy (2010) - Bài giảng “kho dữ liệu và khai phá dữ liệu”.

[6] Huỳnh Lý Thanh Nhàn, Nguyễn Thái Nghe. 2013a. Hệ thống dự đoán kết quả học tập và gợi ý lựa chọn môn học. Kỷ yếu hội thảo quốc gia lần thứ XVI: Một số vấn đề chọc lọc của CNTT&TT (@2013), trang 110-118. Nhà xuất bản Khoa học kỹ thuật. ISBN: 987-604-67-0251-1.

**Tiếng anh:**

[7] Brijesh Kumar Baradwaj, Saurabh Pal: Mining Educational Data to Analyze Students Performance, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

[8] Sajadin Sembiring, Prediction of student academic performance by an application of datamining techniques, 2011 International Conference on Management and Artificial Intelligence IPEDR vol.6 (2011) © (2011) IACSIT Press, Bali, Indonesia

**13. ĐÓNG GÓP CỦA NGHIÊN CỨU**

**13. ĐÓNG GÓP NGHIÊN CỨU**

Hoạt động nghiên cứu khoa học là một trong những nhiệm vụ quan trọng hàng đầu trong các trường Đại học Kinh Tế - Luật. Nghiên cứu khoa học là hoạt động đi tìm cái mới. Cái mới ở đây không chỉ thể hiện trên kết quả nghiên cứu mà còn ở phương pháp khoa học và khả năng ứng dụng phương pháp mới để giải quyết một vấn đề. Nếu được triển khai dự án: *“Nghiên cứu các công cụ phân tích dữ liệu và ứng dụng khai phá dữ liệu điểm để dự đoán kết quả học tập của sinh viên trường Đại học”* thì kết quả không chỉ cung cấp những kiến thức, thông tin về các loại công cụ khai phá dữ liệu mà còn phân tích được ưu điểm và nhược điểm của từng loại công cụ từ đó đưa ra quyết định sẽ lựa chọn một công cụ nào phù hợp với việc phân tích dữ liệu điểm để đưa ra dự đoán về kết quả học tập của sinh viên trường Đại học. Bên cạnh việc sử dụng các công cụ khai phá dữ liệu để đưa ra các dự đoán kết quả học tập của sinh viên một các chính xác nhằm phát hiện sớm các sinh viên yếu kém để giúp họ lập kế hoạch học tập phù hợp là một nhu cầu cần thiết ở trường đại học. Nhằm dự đoán kết quả học tập cho từng cá nhân sinh viên, từ đó hỗ trợ lựa chọn môn học phù hợp với năng lực của từng cá nhân. Kết quả từ các nghiên cứu này đã cho thấy sử dụng phương pháp máy học trong dự đoán kết quả học tập của sinh viên là khả thi và có thể ứng dụng trong thực tế tại trường đại học.

|  |  |
| --- | --- |
| *Ngày \_\_ tháng \_\_ năm\_*  **Người đăng ký**  (Ký và ghi rõ họ tên) | *Ngày \_\_ tháng \_\_ năm\_*  **Người hướng dẫn**  (Ký và ghi rõ họ tên) |
| *Ngày \_\_ tháng \_\_ năm\_*  **Lãnh đạo đơn vị**  (Ký và ghi rõ họ tên) | |