**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN**

**VÕ HOÀNG THÔNG -   18521462**

**NGUYỄN THỊ THANH KIM -   18520963**

**ĐỖ HÙNG DŨNG -   18520629**

**NGUYỄN TẤN PHONG -   18521239**

**NGUYỄN XUÂN VINH -   18521655**

**BÁO CÁO MÔN HỌC**

**MÔN : PHÂN TÍCH THIẾT KẾ THỰC NGHIỆM**

**LỚP : DS304.K21**

**ẢNH HƯỞNG CỦA CÁC YẾU TỐ LÊN ĐIỂM CUỐI KỲ CỦA SINH VIÊN**

**SINH VIÊN KHOA HỌC DỮ LIỆU 2018**

**TP. HỒ CHÍ MINH, 2020**

**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN**

**VÕ HOÀNG THÔNG -   18521462**

**NGUYỄN THỊ THANH KIM -   18520963**

**ĐỖ HÙNG DŨNG -   18520629**

**NGUYỄN TẤN PHONG -   18521239**

**NGUYỄN XUÂN VINH -   18521655**

**BÁO CÁO MÔN HỌC**

**MÔN : PHÂN TÍCH THIẾT KẾ THỰC NGHIỆM**

**LỚP : DS304.K21**

**ẢNH HƯỞNG CỦA CÁC YẾU TỐ LÊN ĐIỂM CUỐI KỲ CỦA SINH VIÊN**

**GIẢNG VIÊN HƯỚNG DẪN**

**TS Đỗ Trọng Hợp**

**TP. HỒ CHÍ MINH, 2020**

**LỜI CẢM ƠN**

Để hoàn thành báo cáo môn học này, chúng tôi xin tỏ lòng biết ơn sâu sắc đến thầy Đỗ Trọng Hợp đã tận tình hướng dẫn trong suốt quá trình viết báo cáo môn học. Với sự nhiệt huyết của mình, thầy đã mang lại những bài học lý thuyết quý báu trong môn Phân tích và thiết kế thực nghiệm nói riêng và cả những bài học cuộc sống nói chung.

Cuối cùng, chúng tôi kính chúc thầy dồi dào sức khỏe và tiếp tục thành công trong sự nghiệp.

**MỤC LỤC**

[CHƯƠNG 1: GIỚI THIỆU 5](#_Toc46083597)

[1.1 Dữ liệu 5](#_Toc46083598)

[1.2 Nội dung báo cáo 7](#_Toc46083599)

[CHƯƠNG 2: THỐNG KÊ PHÂN TÍCH DỮ LIỆU 8](#_Toc46083600)

[CHƯƠNG 3: PHÂN TÍCH PHƯƠNG SAI CÁC YẾU TỐ ẢNH HƯỞNG LÊN ĐIỂM CUỐI KỲ 14](#_Toc46083601)

[3.1. Cơ sở lý thuyết 14](#_Toc46083602)

[3.1.1. Phân tích phương sai – ANOVA 14](#_Toc46083603)

[3.1.2. Hai loại phân tích ANOVA 14](#_Toc46083604)

[3.2. Trực quan hóa dữ liệu 15](#_Toc46083605)

[3.3. Kiểm định giả thuyết 18](#_Toc46083606)

[3.3.1. Ảnh hưởng các yếu tố lên biến G3 18](#_Toc46083607)

[3.3.2. Ảnh hưởng các Interaction lên biến G3 21](#_Toc46083608)

[3.3.3. Kết luận 23](#_Toc46083609)

[CHƯƠNG 4: PHÂN TÍCH HỒI QUY 24](#_Toc46083610)

[4.1.Mô hình Linear Regression 24](#_Toc46083611)

[4.1.1. Giới thiệu 24](#_Toc46083612)

[4.1.2. Mô hình hồi quy tuyến tính đơn biến 25](#_Toc46083613)

[4.1.3. Mô hình hồi quy đa biến 30](#_Toc46083614)

[4.2. Mô hình Support Vector Regression 37](#_Toc46083615)

[4.2.1 Giới thiệu tổng quan và tính toán độ phù hợp của phương trình hồi quy 37](#_Toc46083616)

[4.2.2 Mô hình SVR với một thuộc tính 38](#_Toc46083617)

[4.2.3.Mô hình SVR với đa thuộc tính 40](#_Toc46083618)

4.2.4. So sánh và nhận xét giữa các mô hình…………………………………….45

[CHƯƠNG 5: KẾT LUẬN 46](#_Toc46083619)

[TÀI LIỆU THAM KHẢO 47](#_Toc46083620)

# CHƯƠNG 1: GIỚI THIỆU

## 1.1 Dữ liệu

Tên bộ dữ liệu: **Student Performance Data Set**

Bộ dữ liệu nghiên cứu về thành tích học tập của học sinh tại hai trường Trung học ở nước Bồ Đào Nha. Dữ liệu chứa nhiều thuộc tính được thu thập từ báo cáo của trường Trung học nơi học sinh đang học tập và qua bản khảo sát các câu hỏi.

Bộ dữ liệu cung cấp được tách thành hai phần cho hai môn học khác nhau là Toán học và tiếng Bồ Đào Nha. Bộ dữ liệu trên có thể được thực hiện cho các loại bài toán khác nhau như Hồi Quy (Regression) và Phân loại (Classification).

Bộ dữ liệu gồm hai file: **student-mat.csv** và **student-por.csv** trong đó student-mat.csv là bộ dữ liệu về điểm môn Toán học của sinh viên và student-por.csv là bộ dữ liệu về điểm tiếng Bồ Đào Nha.

Student-mat.csv chứa 395 điểm dữ liệu, student-por.csv chứa 649 điểm dữ liệu.

**Biến mục tiêu**

G3 - điểm số cuối kỳ(số nguyên trong khoảng từ 0 đến 20)

**Biến yếu tố**

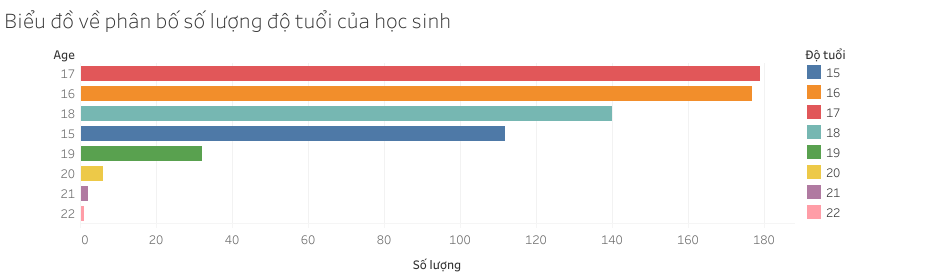
|  |  |
| --- | --- |
| Thuộc tính | Nội dung |
| Sex | Giới tính ('F' – Nữ or 'M' – Nam) |
| Age | Tuổi (số nguyên: từ 15 đến 22) |
| School | Trường học ('GP' - Gabriel Pereira hoặc 'MS' - Mousinho da Silveira) |
| Address | Loại địa chỉ nhà ('U' – thành thị or 'R' – nông thôn) |
| famsize | Số thành viên trong gia đình ('LE3' – <=3 or 'GT3' - >3) |
| Pstatus | Tình trạng phụ huynh ('T' – sống chung or 'A' -xa nhau) |
| Medu | Trình độ giáo dục của mẹ (0 - không, 1 - giáo dục tiểu học (lớp 4), 2 - 5 đến 9, 3 - giáo dục trung học hoặc 4 - giáo dục đại học) |
| Fedu | Trình độ giáo dục của cha (0 - không, 1 - giáo dục tiểu học (lớp 4), 2 - 5 đến 9, 3 - giáo dục trung học hoặc 4 - giáo dục đại học) |
| Mjob | Công việc của mẹ ('teacher'-giáo viên, 'health'-liên quan đến sức khỏe, 'services'-ngành dịch vụ, 'at\_home'-nội trợ, hoặc 'other'-khác) |
| Fjob | Công việc của cha ('teacher'-giáo viên, 'health'-liên quan đến sức khỏe, 'services'-ngành dịch vụ, 'at\_home'-nội trợ, hoặc 'other'-khác) |
| reason | Lý do chọn trường này (‘home’-gần nhà, ‘reputation’-danh tiếng của trường, ‘course’-ưu tiên khóa học or ‘other’-khác) |
| guardian | Người giám hộ (‘mother’-mẹ, ‘father’-cha hoặc 'other'-khác) |
| traveltime | Thời gian di chuyển từ nhà đến trường (1 - <15 phút, 2 - 15 đến 30 phút, 3 - 30 phút đến 1 giờ hoặc 4 -> 1 giờ) |
| studytime | Thời gian học hàng tuần (1 - <2 giờ, 2 - 2 đến 5 giờ, 3 - 5 đến 10 giờ, hoặc 4 -> 10 giờ) |
| failures | số lần thi trượt (số: n nếu 1 <= n <3, khác 4) |
| schoolsup | Hỗ trợ giáo dục từ nhà trường (‘yes’-có hoặc ‘no’-không) |
| famsup | Hỗ trợ giáo dục từ gia đình (‘yes’-có hoặc ‘no’-không) |
| activities | Hoạt động ngoại khóa (‘yes’-có hoặc ‘no’-không) |
| paid | Các lớp học thêm có phí (‘yes’-có hoặc ‘no’-không) |
| Internet | Truy cập Internet tại nhà (‘yes’-có hoặc ‘no’-không) |
| nursery | Học trường mẫu giáo (‘yes’-có hoặc ‘no’-không) |
| higher | Muốn học cao hơn (‘yes’-có hoặc ‘no’-không) |
| romantic | Có mối quan hệ lãng mạn (‘yes’-có hoặc ‘no’-không) |
| famrel | Chất lượng mối quan hệ gia đình (từ 1 - rất tệ đến 5 - rất tốt) |
| freetime | Thời gian rảnh sau giờ học (từ 1 - rất ít đến 5 - rất nhiều) |
| goout | Đi chơi với bạn bè (từ 1 - rất ít đến 5 - rất nhiều) |
| Dalc | Mức tiêu thụ rượu trong ngày (từ 1 - rất thấp đến 5 - rất cao) |
| Walc | Mức tiêu thụ rượu cuối tuần (từ 1 - rất thấp đến 5 - rất cao) |
| health | Tình trạng sức khỏe hiện tại (từ 1 - rất tệ đến 5 - rất tốt) |
| absences | Số lần nghỉ học (số: từ 0 đến 93) |
| G1 | Điểm số đầu tiên (số: từ 0 đến 20) |
| G2 | Điểm số thứ hai (số: từ 0 to 20) |

## 1.2 Nội dung báo cáo

Giáo dục hiện nay đang là một vấn đề được xã hội quan tâm, trong thế giới hiện đại ngày càng phát triển như hiện nay, chất lượng giáo dục có thể bị ảnh hưởng bởi nhiều yếu tố cả về chủ quan học sinh/sinh viên lẫn nhân tố khách quan. Báo cáo này nghiên cứu tác động của các yếu tố nhân khẩu học của học sinh/ sinh viên có ảnh hưởng như thế nào đến điểm số. Từ những kết quả đạt được chúng tôi mong muốn có thể giúp đỡ cải thiện chất lượng giáo dục.

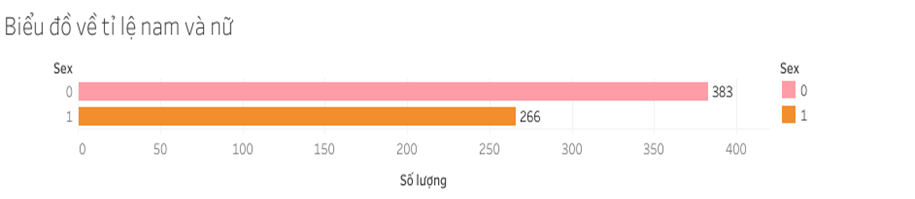
Trong báo cáo này, chúng tôi sẽ nghiên cứu về các yếu tố ảnh hưởng đến điểm cuối kỳ tiếng Bồ Đào Nha(điểm thi lần thứ ba - biến G3) của học sinh tại hai trường Trung học. Lý do chúng tôi chọn bộ dữ liệu về điểm môn tiếng Bồ Đào Nha là bởi vì có nhiều quan sát (Observation) hơn bộ dữ liệu điểm cuối kỳ môn Toán.

# CHƯƠNG 2: THỐNG KÊ PHÂN TÍCH DỮ LIỆU



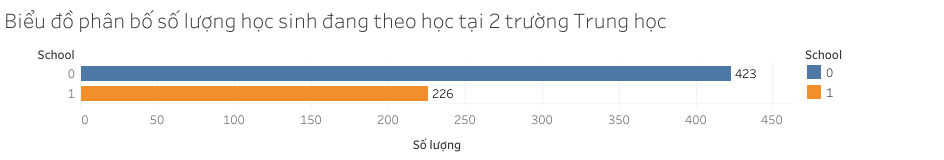
Hình 2.1: Biểu đồ phân bố số lượng độ tuổi của học sinh

Độ tuổi phân bố không đồng đều giữa các nhóm, có sự chênh lệch khá lớn giữa các độ tuổi còn lại. Trong đó, độ tuổi 16 và 17 có số lượng học sinh được nghiên cứu nhất nhiều.



Hình 2.2: Biểu đồ phân bố số lượng nam và nữ trong bộ dữ liệu Student-por

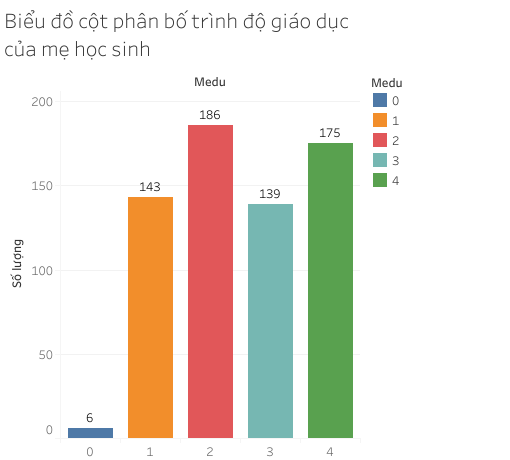
|  |  |
| --- | --- |
| Số lượng nữ chiếm nhiều nhất với 383 học sinh, với tỉ lệ hơn 50%. Còn lại là số học sinh nam với số lượng 266 học sinh.  Hình 2.3: Biểu đồ tròn tỉ lệ nam và nữ |  |



Hình 2.4: Biểu đồ phân bố số lượng học sinh đang theo học tại hai trường

Số 0 đại diện cho trường Gabriel Pereira (GP) và số 1 đại diện cho trường Mousinho da Silveira (MS)

Số lượng học sinh đang theo học tại trường GP nhiều hơn MS và chiếm tỉ lệ hơn 50% bộ dữ liệu.

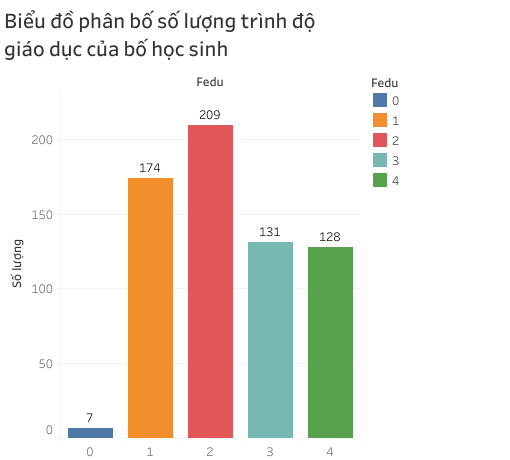


Hình 2.5: Biểu đồ cột phân bố trình độ giáo dục của mẹ học sinh

Tỉ lệ phụ huynh nữ với trình độ THCS và Đại học chiếm cao nhất với 186 và 175 phụ huynh.

Tỉ lệ phụ huynh nữ thất học chiếm ít nhất với 6 phụ huynh.

Thông thường, chúng tôi thấy tỉ lệ nữ giới học cao lên khá là thấp, phần lớn có thể học xong cấp 2 hoặc cấp 3 là đã đi lấy chồng. Ngược lại, nhìn ở biểu đồ này thấy số lượng nữ giới học Đại học và cao học cao hơn các trình độ còn lại.

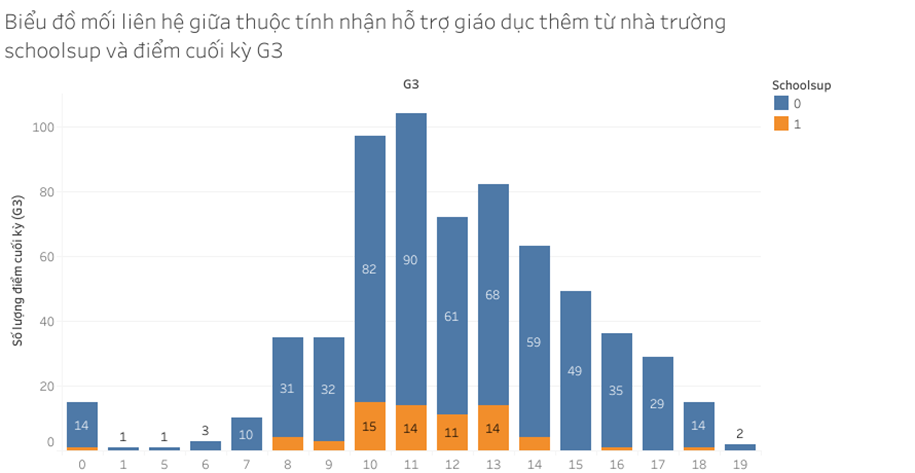


Hình 2.6: Biểu đồ cột phân bố trình độ giáo dục của bố học sinh

Tỉ lệ trình độ THCS chiếm cao nhất với 209 phụ huynh và thấp nhất là 7 phụ huynh với người không học.

Số lượng phụ huynh bố học cao học thấp hơn số lượng phụ huynh mẹ. Có vẻ như ở nước ngoài, nữ giới quan trọng học thức hơn nam.

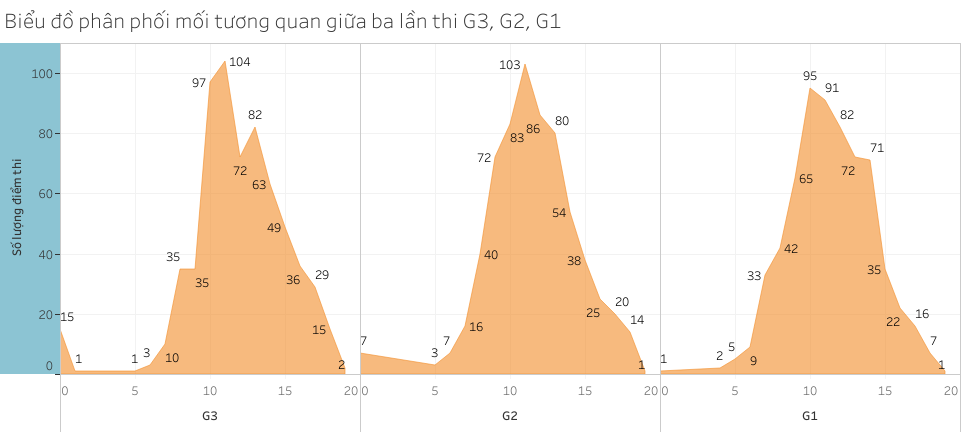
|  |  |
| --- | --- |
| Với 581 học sinh không muốn nhận hỗ trợ từ nhà trường và số còn lại 68 em được hỗ trợ thêm về giáo dục.  Quan sát biểu đồ cho thấy nhiều học sinh có quan điểm dành ít thời gian cho việc học tại trường hơn. Điều này không có nghĩa là những học sinh này lười học hay chán ghét việc học.  Hình 2.7: Biểu đồ tròn thể hiện số lượng học sinh nhận hỗ trợ giáo dục từ nhà trường |  |



Hình 2.8: Biểu đồ mối liên hệ giữa thuộc tính nhận hỗ trợ giáo dục thêm từ nhà trường và điểm cuối kỳ- Trục x: Điểm số G3

Số lượng học sinh nhận hỗ trợ giáo dục thêm từ nhà trường theo lý thuyết sẽ đạt điểm cao hơn các bạn không học thêm. Tuy nhiên, biểu đồ này cho thấy phần lớn những học sinh học thêm có điểm số tương đối trung bình so với các bạn không học thêm từ nhà trường.

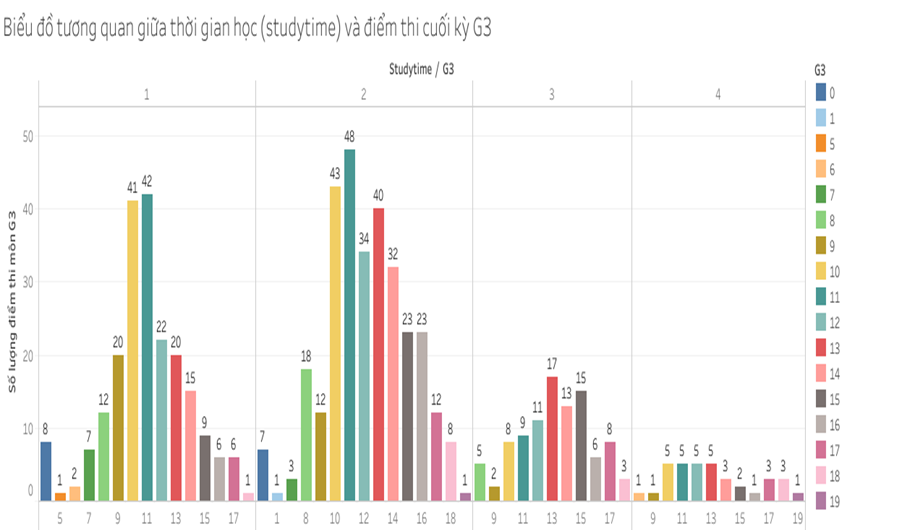
|  |  |
| --- | --- |
| Hình 2.9: Biểu đồ tròn số lượng học sinh theo học cao học  Đa số học sinh mong muốn theo học cao học, cho thấy có rất nhiều học sinh quan trọng học thức. |  |



Hình 2.10: Biểu đồ phân phối mối tương quan giữa số lượng điểm thi của ba lần thi G3, G2 và G1

Trục x là điểm thi

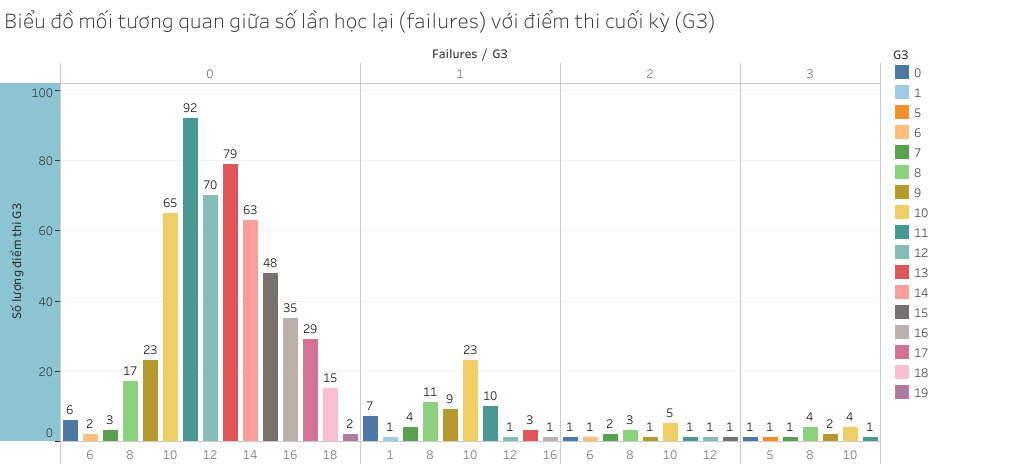
Biểu đồ điểm thi lần thứ hai (G2) có phân phối hình chuông, phân phối chuẩn nhất, với tỉ lệ số lượng điểm thi phân phối đồng đều của tất cả phổ điểm [0;20].



Hình 2.11: Biểu đồ tương quan giữa thời gian học và điểm thi cuối kỳ G3

Trục x là điểm thi, 4 biểu đồ tượng trưng cho 4 thời gian học sinh dành ra để học trong tuần.

Số lượng học sinh dành ít hơn 2h và từ 2-5h mỗi tuần để học chiếm nhiều nhất (biểu đồ thứ nhất và biểu đồ thứ hai), số lượng học sinh dành trên 10h mỗi tuần để học chiếm ít nhất.



Hình 2.12: Biểu đồ mối tương quan giữa số lần học lại với điểm thi cuối kỳ G3

Trục x: Điểm thi, 4 biểu đồ từ trái sang lần lượt là số lần học lại từ 1-4 của học sinh

Số lần học lại là 0 lần, chiếm tỉ lệ số lượng học sinh cao nhất, số lần học lại với điểm thi cuối kỳ tỉ lệ thuận với nhau.

# CHƯƠNG 3: PHÂN TÍCH PHƯƠNG SAI CÁC YẾU TỐ ẢNH HƯỞNG LÊN ĐIỂM CUỐI KỲ

## 3.1. Cơ sở lý thuyết

### 3.1.1. Phân tích phương sai – ANOVA

Phân tích phương sai (Analysis of Variance) hay còn gọi là kiểm định ANOVA là một kỹ thuật thống kê tham số được sử dụng để so sánh các bộ dữ liệu. Nói một cách dễ hiểu, phân tích ANOVA có chức năng đánh giá sự khác biệt tiềm năng trong một biến phụ thuộc mức quy mô bằng một biến mức danh nghĩa có từ 2 loại trở lên. Các nhà phân tích sử dụng thử nghiệm ANOVA để xác định ảnh hưởng của các biến độc lập đối với biến phụ thuộc trong nghiên cứu hồi quy. Kỹ thuật kiểm định ANOVA này được phát triển bởi Ronald Fisher năm 1918.

Ví dụ: Một nhóm bệnh nhân tâm thần đang thử ba liệu pháp khác nhau: tư vấn, dùng thuốc và phản hồi sinh học. Bạn muốn xem liệu một liệu pháp tốt hơn những liệu pháp khác.

### 3.1.2. Hai loại phân tích ANOVA

Có nhiều hơn 2 loại phân tích phương sai, ở đây giới thiệu lý thuyết phân tích phương sai một yếu tố và hai yếu tố.

#### 3.1.2.1. Phân tích phương sai một yếu tố (one-way ANOVA)

One-way ANOVA là một loại thử nghiệm thống kê so sánh phương sai trong nhóm có nghĩa là trong một mẫu trong khi chỉ xem xét một yếu tố hoặc một biến độc lập. Phương sai một yếu tố so sánh ba hoặc nhiều hơn ba nhóm phân loại để xác định xem có sự khác biệt giữa chúng hay không. Trong mỗi nhóm nên có ba hoặc nhiều quan sát và phương tiện của các mẫu được so sánh.

Ví dụ như: Bạn có thể sử dụng phân tích phương sai một yếu tố để tìm hiểu liệu hiệu suất kiểm tra có khác nhau hay không dựa trên mức độ lo lắng giữa các học sinh (chia học sinh thành ba nhóm độc lập: học sinh thấp, trung bình và cao bị căng thẳng)

#### 3.1.2.2 Phân tích ANOVA hai yếu tố (two-way ANOVA)

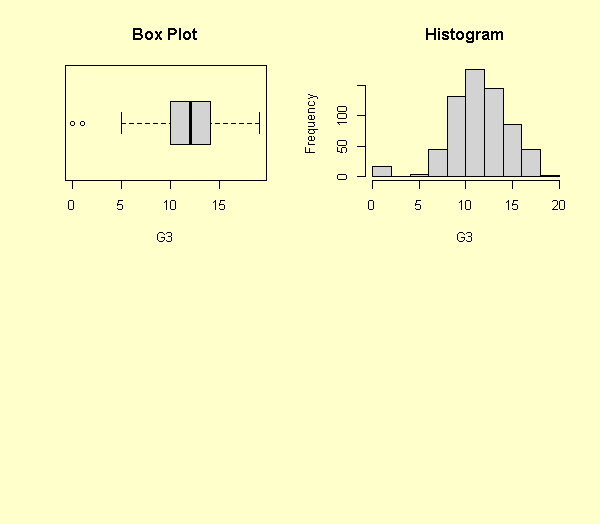
ANOVA hai yếu tố là một phần mở rộng của phân tích phương sai một yếu tố. Với One Way, bạn có một biến độc lập ảnh hưởng đến biến phụ thuộc. Còn với two-way ANOVA, sẽ có 2 biến độc lập.

Ví dụ: bạn có thể sử dụng phân tích ANOVA hai yếu tố để tìm hiểu liệu có sự tương tác giữa giới tính và trình độ học vấn đối với sự lo lắng kiểm tra giữa các sinh viên đại học. Trong đó giới tính (nam / nữ) và trình độ học vấn (đại học / sau đại học) là các biến độc lập của bạn, và kiểm tra lo lắng là biến phụ thuộc của bạn.

Trong phần 3.3, one-way ANOVA được sử dụng để kiểm tra xem liệu những độc lập nào trong 32 biến độc lập có ảnh hưởng đến biến phụ thuộc và chọn ra những biến có ảnh hưởng nhất, sau đó two-way ANOVA được sử dụng để kiểm tra sự liệu có sự tương tác giữa các biến độc lập được chọn đối với biến phụ thuộc.

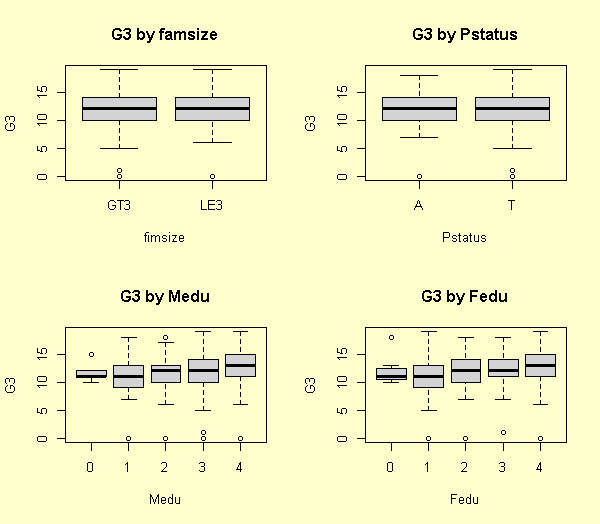
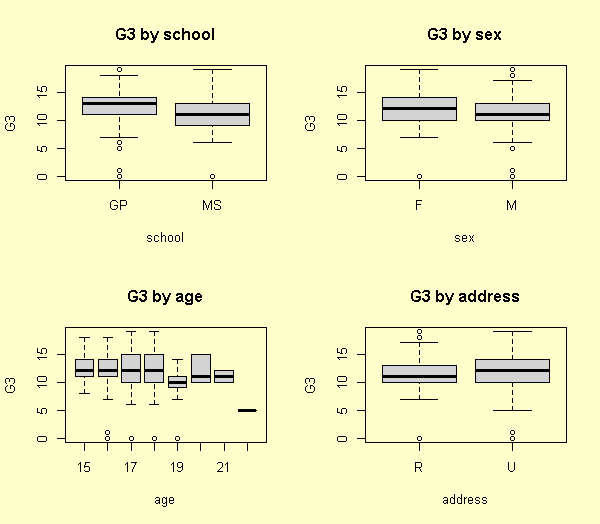
## 3.2. Trực quan hóa dữ liệu

Để tìm hiểu liệu yếu tố nào có ảnh hưởng đến điểm số G3, trước tiên chúng tôi vẽ một số biểu đồ để đánh giá sơ bộ sự phân bố điểm số G3 theo các yếu tố kể trên.



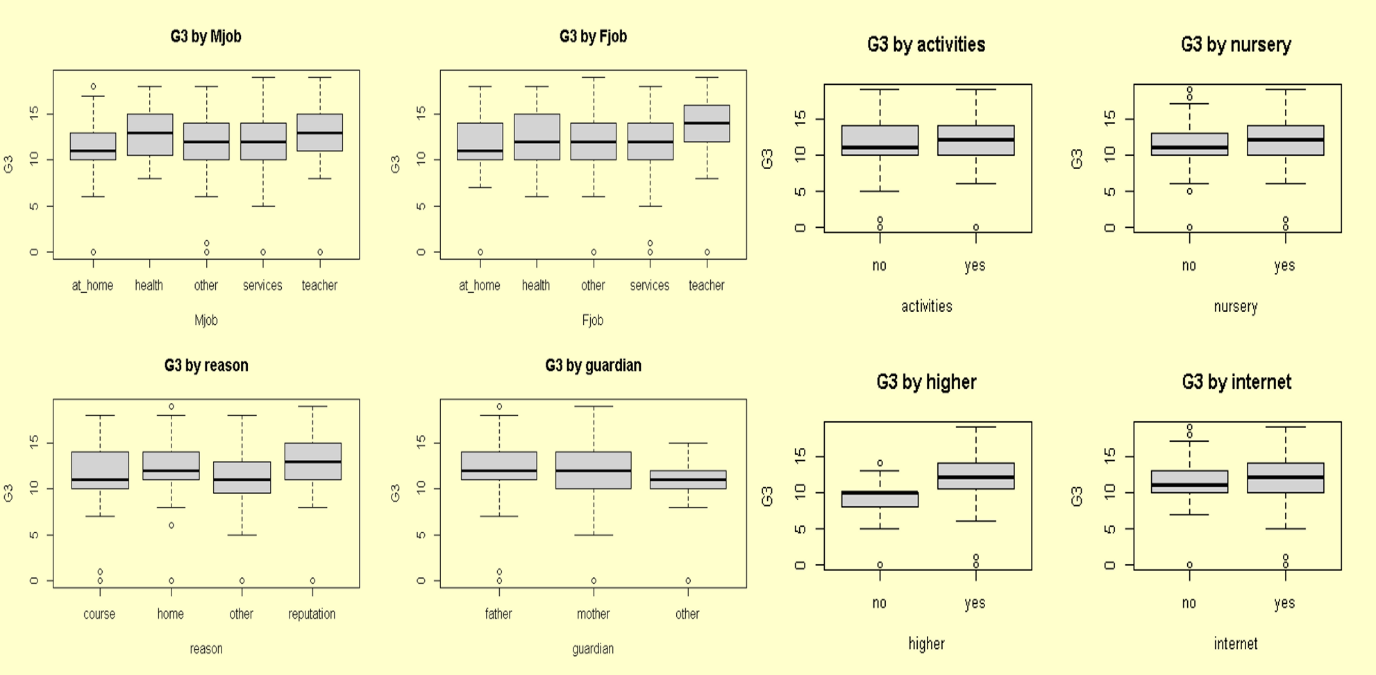
Hình 3.1. Biểu đồ hộp và tần suất điểm số G3.

Dựa vào hai biểu đồ ở hình 1 có thể thấy rằng G3 có phân phối chuẩn với trung bình khoảng 11-12 điểm, khoảng hơn 50% học sinh/sinh viên có điểm số từ 10 trở lên.



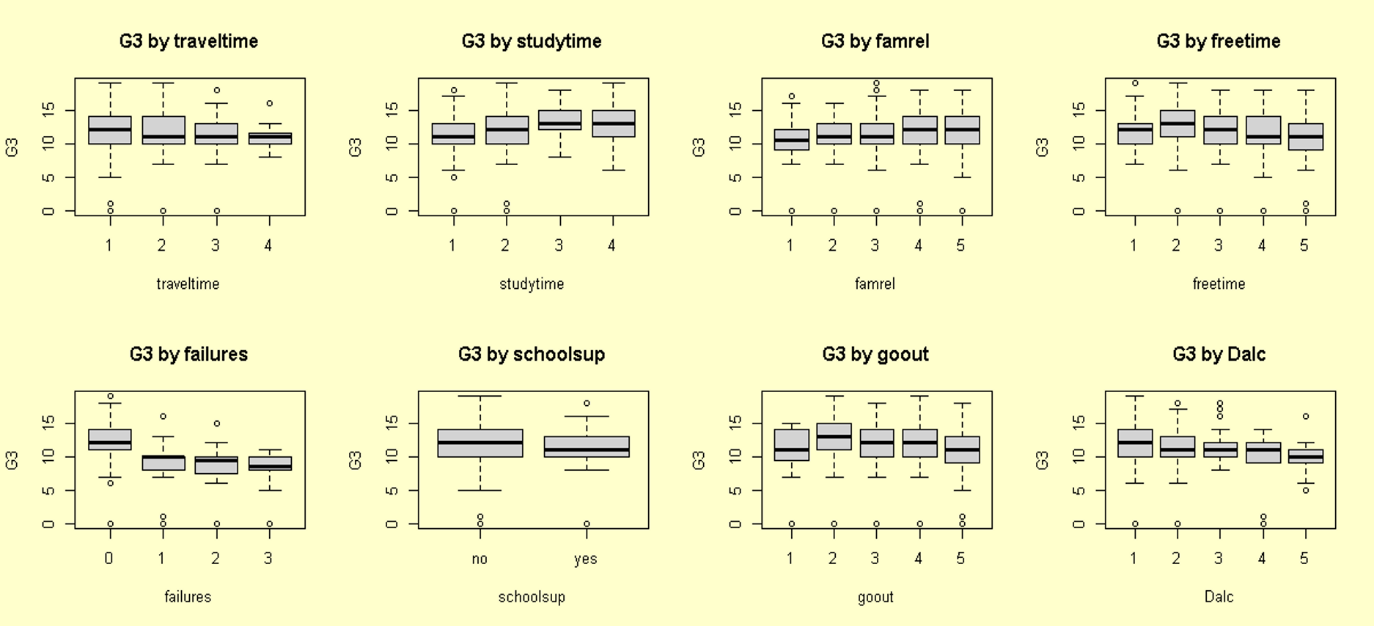
Hình 3.2. Biểu đồ hộp phân bố điểm số G3 theo các yếu tố school, sex, age, address, famsize, Pstatus, Medu, Fedu.

Từ biểu đồ của 8 yếu tố đầu tiên có thể thấy phân bố G3 bởi school ở hai trường GP và MS là khác biệt cả về trung vị và bách phân vị, các giá trị của GP cao hơn đáng kể so với MS vì vậy ta có thể dự đoán G3 bị ảnh hưởng bởi yếu tố school, ta cũng có thể nhận xét tương tự như biểu đồ của G3 by age, Medu và Fedu. Đối với phân bố G3 bởi sex và address, famsize và Pstatus, median và bách phân vị 25% và 75% là gần như nhau nên có thể suy đoán G3 không bị ảnh hưởng nhiều bởi 4 yếu tố này. tương tự với những nhận xét trên, có thể thấy rằng phân bố g3 bởi các mức trong yếu tố Medu và Fedu có khác biệt, có thể suy đoán 2 yếu tố Mdeu và Fedu có ảnh hưởng đến G3. còn phân bố g3 bởi các yếu tố famsize, Pstatus, Mjob, Fjob, guardian, reason không có nhiều khác biệt ở các mức trong yếu tố, có thể suy đoán các yếu tố này không ảnh hưởng lớn đến G3.



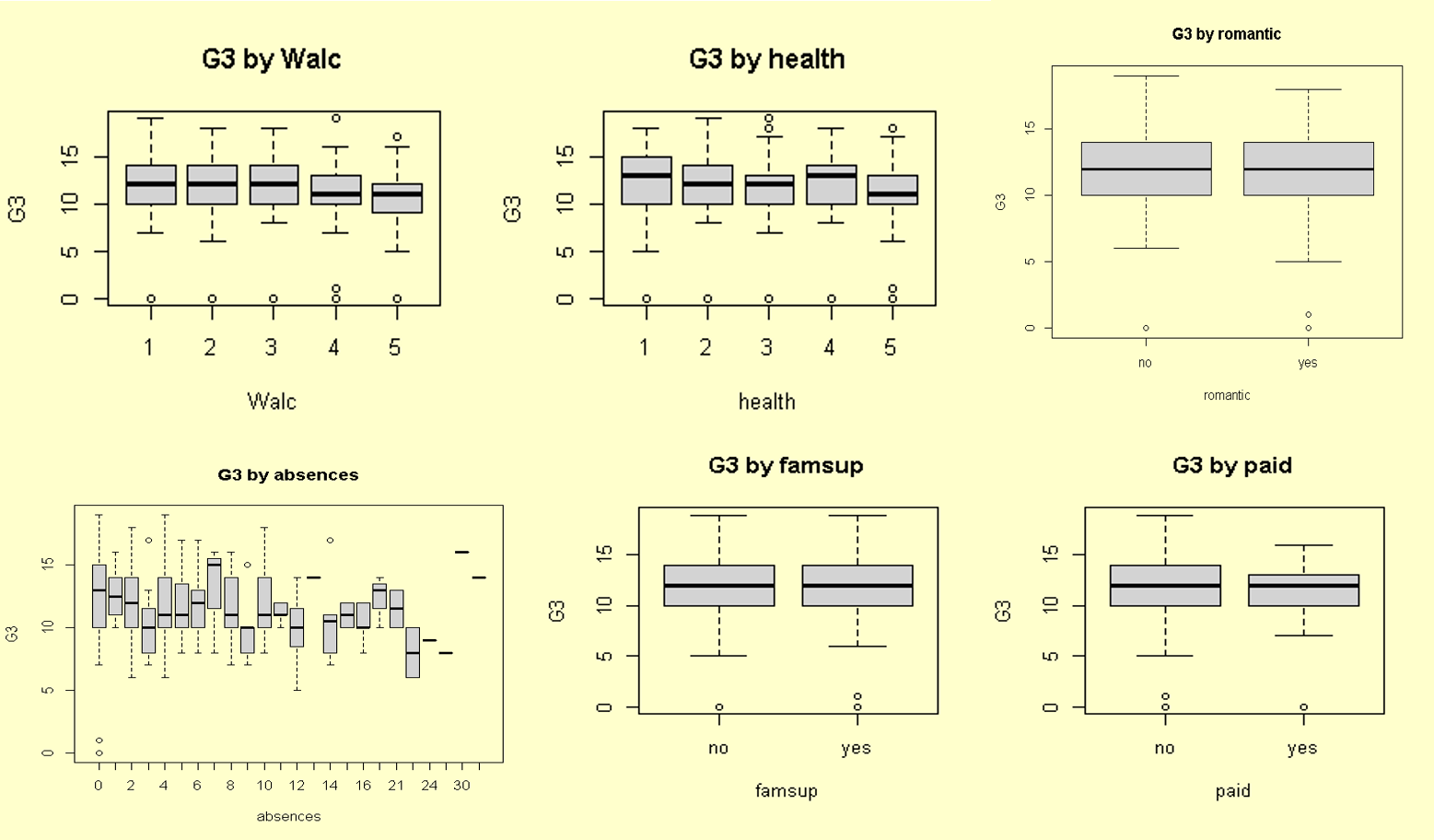
Hình 3.3. Biểu đồ hộp phân bố điểm số G3 bởi các yếu tố Mjob, Fjob, reason, guardian, activities, nursery, higher, internet.

Tương tự như những nhận xét ở Hình 2 từ Hình 3 chúng ta có thể dự đoán rằng yếu tố higher có thể có ảnh hưởng đến điểm số G3, học sinh/sinh viên muốn học lên cao thường có điểm số cao hơn, còn các yếu tố còn lại dường như không gây ra biến thiên đối với G3 quá nhiều.



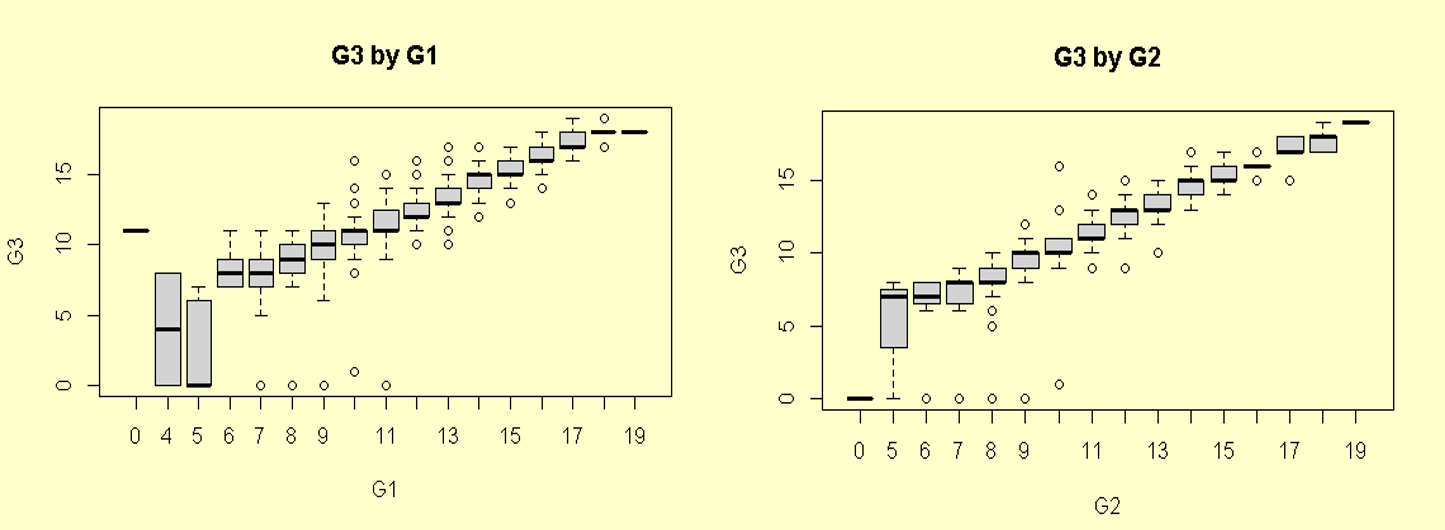
Hình 3.4. Biểu đồ hộp phân bố điểm số G3 bởi các yếu tố traveltime, studytime, failures, schoolsup, famrel, freetime, goout, Dalc, Walc.

Từ Hình 3.4 ta có thể dự đoán rằng traveltime, studytime, failures, Dalc có thể sẽ ảnh hưởng đến G3.



Hình 3.5. Biểu đồ hộp phân bố điểm số G3 bởi các yếu tố Walc, health, romantic, absences, famsup, paid.

Tương tự với những nhận xét trước, ta có thể thấy mối liên hệ giữa G3 và các yếu tố trên dường như không đủ rõ ràng để kết luận G3 có bị ảnh hưởng hay không



Hình 3.6. Biểu đồ hộp thể hiện sự phân bố điểm G3 bởi G1 và G2.

Từ Hình 3.6 có thể nhận xét rằng có mối quan hệ tuyến tính giữa G1 và G2, hai yếu tố này sẽ ảnh hưởng lớn đến G3, điều này là dễ hiểu vì điểm số thường có liên hệ với nhau.

## Kiểm định giả thuyết

### 3.3.1. Ảnh hưởng các yếu tố lên biến G3

Df Sum Sq Mean Sq F value   Pr(>F)

school 1 547 546.6 56.89 1.57e-13 \*\*\*

sex 1 113 112.68 10.96 0.000982 \*\*\*

age 7 262 37.40 3.687 0.000644 \*\*\*

address 1 190 190.06 18.71 1.76e-05 \*\*\*

famsize 1 14 13.71 1.314 0.252

Pstatus 1 0 0.004 0 0.985

Medu 1 390 390.1 39.6 5.75e-10 \*\*\*

Fedu 4 329 82.16 8.223 1.8e-06 \*\*\*

Mjob 4 296 74.01 7.37 8.31e-06 \*\*\*

Fjob 4 135 33.68 3.273 8.31e-06 \*

reason 3 308 102.57 10.25 1.34e-06 \*\*\*

guardian 2 55 27.40 2.638 0.0723 .

traveltime 3 113 37.72 3.659 0.0123 \*

studytime 3 465 155.03 15.88 5.71e-10 \*\*\*

failures 3 1305 434.9 51.39 <2e-16 \*\*\*

famsup 1 24 23.71 2.276 0.132

paid 1 20 20.38 1.956 0.162

activities 1 24 24.18 2.321 0.128

nursery 1 6 5.591 0.535 0.465

higher 1 746 746.2 80.24 <2e-16 \*\*\*

internet 1 152 152.22 14.9 0.000125 \*\*\*

romantic 1 55 55.49 5.353 0.021 \*

famrel 4 151 37.76 3.678 0.00568 \*\*

freetime 4 183 45.86 4.489 0.00139 \*\*

goout 4 281 70.32 6.986 1.65e-05 \*\*\*

Dalc 4 328 81.88 8.193 1.9e-06 \*\*\*

Walc 4 225 56.20 5.535 0.000219 \*\*\*

health 4 101 25.25 2.44 0.0457 \*

absences 23 248 10.76 1.032 0.421

G1 16 4821 301.30 98.03 <2e-16 \*\*\*

G2 15 5762 384.1 242.7 <2e-16 \*\*\*

**Giả thuyết H0** : không có sự ảnh hưởng của yếu tố X lên điểm G3 với mức ý nghĩa 5%

p < 0.05 -> Bác bỏ H0  -> có sự ảnh hưởng

p > 0.05 -> Không thể bác bỏ H0 -> không có sự ảnh hưởng

* School : p = 1.57e-13 < 0.05 -> bác bỏ H0 ->  có sự ảnh hưởng
* Sex : p = 0.000982 <  0.05 -> bác bỏ H0 ->  có sự ảnh hưởng
* Age : p = 0.000644 < 0.05 < 0.05  -> bác bỏ H0 ->   có sự ảnh hưởng
* Address : p = 1.76e-05 < 0.05 -> bác bỏ H0 ->  có sự ảnh hưởng
* Famsize : p = 0.252> 0.05 -> bác bỏ H0 ->  không có ảnh hưởng
* Pstatus : p = 0.985> 0.05 -> bác bỏ H0 ->  không có ảnh hưởng
* Medu : p = 5.75e-10 < 0.05 **-**> bác bỏ H0 ->  có sự ảnh hưởng
* Fedu : p = 1.8e-06 < 0.05 -> bác bỏ H0 ->  có sự ảnh hưởng
* Mjob : p = 8.31e-06 < 0.05 -> bác bỏ H0 ->  có sự ảnh hưởng
* Fjob : p = 8.31e-06 < 0.05  -> bác bỏ H0 ->  có sự ảnh hưởng
* Reason : p = 1.34e-06 < 0.05 -> bác bỏ H0 ->  có sự ảnh hưởng
* Guardian : p = 0.0723 >0.05 -> bác bỏ H0 ->không có ảnh hưởng
* Traveltime : p =0.0123 < 0.05 -> bác bỏ H0 ->  có sự ảnh hưởng
* Studytime : p = 5.71e-10 < 0.05  -> bác bỏ H0 ->  có sự ảnh hưởng
* Failures : p = <2e-16 < 0.05  -> bác bỏ H0 ->  có sự ảnh hưởng
* Schoolsup : p = 0.000125 < 0.05  -> bác bỏ H0 ->  có sự ảnh hưởng
* Famsup : p = 0.132> 0.05  -> không thể bác bỏ H0 ->  không có sự ảnh hưởng
* Paid : p = 0.162> 0.05 -> không thể bác bỏ H0 ->  không có sự ảnh hưởng
* Activities : p = 0.128 > 0.05  -> không thể bác bỏ H0 ->  không có sự ảnh hưởng
* Nursery : p = 0.465> 0.05 -> không thể bác bỏ H0 ->  không có sự ảnh hưởng
* Higher : p = <2e-16 < 0.05 -> bác bỏ H0 ->   có sự ảnh hưởng
* Internet : p = 0.000125 < 0.05  -> bác bỏ H0 ->  có sự ảnh hưởng
* Romantic : p = 0.021 < 0.05 -> bác bỏ H0 ->  có sự ảnh hưởng
* Famrel : p = 0.00568 < 0.05  -> bác bỏ H0 ->  có sự ảnh hưởng
* Freetime : p = 0.00139 < 0.05  -> bác bỏ H0 ->  có sự ảnh hưởng
* Goout : p = 1.65e-05 < 0.05  -> bác bỏ H0 ->  có sự ảnh hưởng
* Dalc : p = 1.9e-06 < 0.05  -> bác bỏ H0 ->  có sự ảnh hưởng
* Walc : p = 0.000219 < 0.05  -> bác bỏ H0 ->  có sự ảnh hưởng
* Health : p = 0.0457 < 0.05  -> bác bỏ H0 ->  có sự ảnh hưởng
* Absences : p = 0.421>0.05  -> không thể bác bỏ H0 ->  không có sự ảnh hưởng
* G1 : p = <2e-16 < 0.05 -> bác bỏ H0 ->  có sự ảnh hưởng
* G2 : p = <2e-16 < 0.05 -> bác bỏ H0 ->  có sự ảnh hưởng

Kết luận :

-         Có 8 yếu tố không có ảnh hưởng đối với G3 lần lượt là: famsize, Pstatus, guardian, famsup, paid, activities, nursery, absences.

-         Có 24 yếu tố còn lại có ảnh hưởng đối với G3 .

-         “G1, G2, higher, studytime, failures, Medu, school, sex” là các yếu tố có F-value cao nhất -> theo One-way thì các thuộc tính này ảnh hưởng đến G3 nhiều nhất .

|  |  |
| --- | --- |
| Có ảnh hưởng | Không ảnh hưởng |
| School, Sex, Age, Address, Medu, Fedu, Mjob, Fjob, reason, traveltime, Studytime, Failures, Schoolsup, Higher, Internet, Romantic, Famrel, freetime, goout, dalc, walc, health, G1, G2 | famsize, Pstatus, guardian, famsup, paid, activities, nursery, absences |

### 3.3.2. Ảnh hưởng các Interaction lên biến G3

Trong phần này chúng ta sẽ phân tích Anova cho hai nhóm biến độc lập với 1 biến phụ thuộc gọi là phân tích two-way Anova. Giả sử chúng ta đã có sự ảnh hưởng giữa các nhóm tuổi đối với điểm G3 thì câu hỏi đặt ra tiếp theo là sự khác biệt đó có đúng cho cả hai nhóm giới tính là nam và nữ hay không. Phân tích one-way Anova không thể trả lời câu hỏi này vì phân tích này được tiến hành cho toàn bộ hai loại giới tính nam và nữ không có sự chia ra.

Sự ưu việt của phân tích two-way Anova như sau: Ví dụ trường hợp chúng ta có thể tìm ra là ảnh hưởng của độ tuổi tới điểm G3 khác nhau giữa Nam và Nữ. Đối với Nam sự ảnh hưởng tăng theo độ tuổi trong khi đối với Nữ lại giảm theo độ tuổi, trường hợp này gọi là hiệu ứng tương tác **Interaction Effect**.

Sau khi chọn ra 7 yếu tố có sự ảnh hưởng lớn nhất tới G3, chúng tôi lập ra bảng Two-way Anova sau:

                     Df Sum Sq Mean Sq F value   Pr(>F)

sex:Medu 1 2 2.1 0.215 0.642838

sex:studytime 3 65 21.61 2.233 0.083230 .

sex:failures 3 11 3.7 0.439 0.725287

sex:higher 1 1 1.2 0.135 0.713359

sex:G1 13 67 5.17 1.713 0.0542 .

sex:G2 13 33 2.5 1.623 0.0743 .

Medu:studytime 3 21 6.9 0.745 0.526

Medu:failures 3 4 1.3 0.154 0.927

Medu:higher 1 34 33.6 3.749 0.0533 .

Medu:G1 13 58 4.4 1.459 0.128

Medu:G2 14 59 4.2 2.743 0.000596 \*\*\*

studytime:failures 6 26 4.4 0.536 0.781

studytime:higher 3 10 3.5 0.388 0.761

studytime:G1 34 102 3.01 0.985 0.495

studytime:G2 35 28 0.8 0.49 0.994

failures:higher 3 28 9.3 1.163 0.323

failures:G1 22 222 10.1 3.715 3.56e-08 \*\*\*

failures:G2 21 97 4.6 3.159 3.18e-06 \*\*\*

higher:G1 8 47 5.9 1.961 0.049 \*

higher:G2 8 59 7.4 4.895 7.03e-06 \*\*\*

G1:G2 55 162 2.94 2.117 1.27e-05 \*\*\*

**Giả thuyết H0** : không có sự ảnh hưởng của interaction của X và Y lên G3 với mức ý nghĩa 5%

p  < 0.05 -> Bác bỏ H0  -> có sự ảnh hưởng

p > 0.05 -> Không thể bác bỏ H0 -> không có sự ảnh hưởng

* Sex:Medu:p = 0.642838 > 0.05 -> không thể bác bỏ H0 -> không có sự ảnh hưởng
* Sex:Studytime:p = 0.083230> 0.05 -> không thể bác bỏ H0 -> không có sự ảnh hưởng
* Sex:Failures:p = 0.725287> 0.05 ->không thể bác bỏ H0 -> không có sự ảnh hưởng
* Sex:Higher:p = 0.713359 > 0.05 -> không thể bác bỏ H0 -> không có sự ảnh hưởng
* Sex:G1:p = 0.0542 > 0.05 -> không thể bác bỏ H0 -> không có sự ảnh hưởng
* Sex:G2:p = 0.0743> 0.05 -> không thể bác bỏ H0 -> không có sự ảnh hưởng
* Medu:Studytime:p = 0.526 > 0.05 -> không thể bác bỏ H0 -> không có sự ảnh hưởng
* Medu:Failures:p = 0.927 > 0.05 ->không thể bác bỏ H0 -> không có sự ảnh hưởng
* Medu:Higher:p = 0.0533> 0.05 -> không thể bác bỏ H0 -> không có sự ảnh hưởng
* Medu:G1:p = 0.128< 0.05 ->không thể bác bỏ H0 ->không có sự ảnh hưởng
* Medu:G2:p = 0.000596< 0.05 -> bác bỏ H0 -> có sự ảnh hưởng
* Studytime:Failures:p = 0.781>0.05 ->  không thể bác bỏ H0 ->không có sự ảnh hưởng
* Studytime:Higher:p = 0.761 > 0.05 -> không thể bác bỏ H0 ->không có sự ảnh hưởng
* Studytime:G1:p = 0.495 < 0.05 -> bác bỏ H0 -> có sự ảnh hưởng
* Studytime:G2:p = 0.994> 0.05 -> không thể bác bỏ H0 -> không có sự ảnh hưởng
* Failures:Higher:p = 0.323 > 0.05 -> không bác bỏ H0 -> không có sự ảnh hưởng
* Failures:G1:p = 3.56e-08< 0.05 -> bác bỏ H0 -> có sự ảnh hưởng
* Failures:G2:p = 3.18e-06 < 0.05 -> bác bỏ H0 -> có sự ảnh hưởng
* Higher:G1:p = 0.049< 0.05 -> bác bỏ H0 -> có sự ảnh hưởng
* Higher:G2:p = 7.03e-06< 0.05 -> bác bỏ H0 -> có sự ảnh hưởng
* G1:G2:p = 1.27e-05 < 0.05 -> bác bỏ H0 -> có sự ảnh hưởng

**Kết luận :**

- Có 14 interaction không có tương tác với G3 lần lượt là:“ Sex:Medu, Sex:Studytime, Sex:Failures, Sex:Higher, Sex:G1, Sex:G2, Medu:Studytime, Medu:Failures, Medu:Higher, Medu:G1, Studytime:Failures, Studytime:Higher, Studytime:G2, Failures:Higher ”

- Có 7 interaction có tương tác với G3 lần lượt là: “ Medu:G2, Studytime:G1, Failures:G1, Failures:G2, Higher:G1, Higher:G2, G1:G2 ”

-  “ Sex:G2, failures:G2, failures:G1, sex:failures, higher:G2 ” là 5 yếu tố ảnh hưởng interaction nhiều nhất lên G3.

|  |  |
| --- | --- |
| Có ảnh hưởng | Không ảnh hưởng |
| Medu:G2, Studytime:G1, Failures:G1, Failures:G2, Higher:G1, Higher:G2, G1:G2 | Sex:Medu, Sex:Studytime, Sex:Failures, Sex:Higher, Sex:G1, Sex:G2, Medu:Studytime, Medu:Failures, Medu:Higher, Medu:G1, Studytime:Failures, Studytime:Higher, Studytime:G2, Failures:Higher |

### 3.3.3. Kết luận

Từ kết quả phân tích phương sai, chọn ra những biến độc lập có ảnh hưởng theo kiểm định giả thuyết ở phần **3.3.1** và có Mean Sq cao để phân tích hồi quy. Những biến độc lập thỏa hai yếu tố trên lần lượt là: G1, G2, higher, studytime, failures, Medu, sex,.. Ngoài ra, một số interaction effect giữa các yếu tố trên cũng có tương tác với biến phụ thuộc G3, những interaction đó lần lượt là Medu:G2, Studytime:G1, Failures:G1, Failures:G2, Higher:G1, Higher:G2, G1:G2. Tiếp theo ta sẽ chạy một mô hình dự đoán xem có chính xác với kết quả này hay không, từ đó nhận xét được hiệu suất mô hình. Cuối cùng đưa ra được mô hình với hiệu suất tối ưu nhất.

# CHƯƠNG 4: PHÂN TÍCH HỒI QUY

## 4.1.Mô hình Linear Regression

### 4.1.1. Giới thiệu

Mô hình hồi qui tuyến tính đơn biến dùng để xem xét mối quan hệ tuyến tính giữa biến phụ thuộc y (biến kết cục) và biến độc lập x (biến dự đoán). Phương trình tuyến tính (đường thẳng) đơn biến có dạng:

*y= a + bxi+ei*

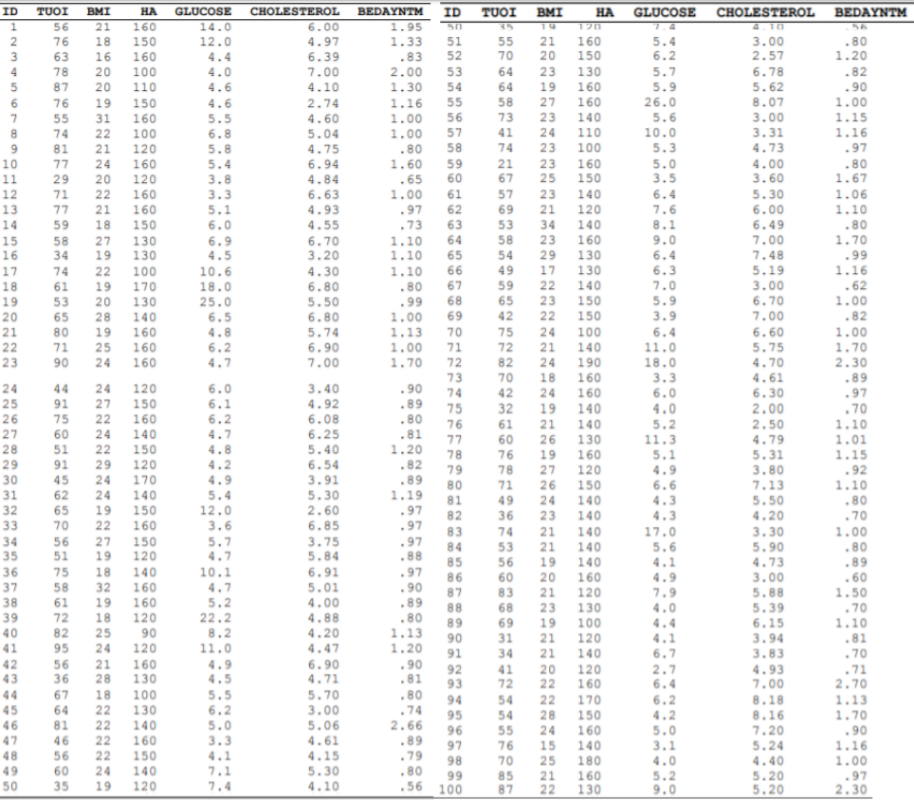
Trong đó a là điểm cắt trên trục tung, β là độ dốc (trong thống kê gọi là hệ số hồi qui) và e là phần dư.

Mô hình hồi qui tuyến tính đa biến có dạng:

*y= a + b1x1i+ b2x2i +b3x3i + ......+........+ bkxki + ei*

Ví dụ: dữ liệu đo bề dày lớp NTM động mạch cảnh trên siêu âm ở 100 bệnh nhân có bệnh tim mạch như sau:

Dữ liệu này có 7 cột: ID bệnh nhân, tuổi (năm), BMI (kg/m2), huyết áp tâm thu (mmHg), glucose máu (mmol/L), cholesterol (mmol/L), bề dày lớp NTM trên siêu âm (mm).



Bảng 4.1. Dữ liệu bề dày NTM của 100 bệnh nhân

Với ví dụ trên ta có thể viết phương trình hồi qui tuyến tính đa biến với 5 yếu tố (x1=tuổi, x2=cholesterol, x3=glucose, x4=huyết áp TT và x5= BMI) như sau:

Bề dày NTM= a + b1(tuổi) + b2(cholesterol) + b3(glucose) + b4(huyết áp TT) + b5 (BMI) + ....+ e

Trong phân tích hồi qui tuyến tính đa biến, ta cần biết mức độ ảnh hưởng của từng yếu tố lên biến kết cục y (bề dày lớp NTM trong ví dụ này). Muốn biết mức độ ảnh hưởng cần lưu ý đến các trị số sau:

1. Hệ số tương quan R (coefficient of correlation): yếu tố nào có R càng lớn thì ảnh hưởng càng nhiều

2. Bình phương của R (R square): yếu tố nào có càng lớn thì mối quan hệ giữa yếu tố đó và biến y càng chặt chẽ.

3. Hệ số hồi qui b (regression coefficient): yếu tố nào có b cao thì ảnh hưởng nhiều hơn, tuy nhiên các yếu tố có đơn vị khác nhau (tuổi, mmol/L, mmHg....) nên không thể so sánh mức ảnh hưởng giữa các yếu tố. Nếu muốn so sánh phải đổi các yếu tố có cùng đơn vị là độ lệch chuẩn, lúc đó ta có hệ số hồi qui chuẩn hóa: = b. ( Với Sx là độ lệch chuẩn của x tương ứng và Sy là độ lệch chuẩn của y). Dựa vào công thức trên, ta tính được hệ số hồi qui chuẩn của tuổi (0,42), cholesterol (0,22), glucose (0,14), huyết áp (0,10) và BMI (0,02)

4. Trị số p (p value): càng nhỏ mức ảnh hưởng càng mạnh

### 4.1.2. Mô hình hồi quy tuyến tính đơn biến

Với mô hình hồi quy tuyến tính đơn biến chúng ta sẽ xem xét sự ảnh hưởng của từng yếu tố tác động đến G3(output target).

· Lưu ý: Lựa chọn tỉ lệ ngẫu nhiên giữa tập train và tập test tỉ lệ 7:3

|  |  |  |
| --- | --- | --- |
| **Tên thuộc tính** | **Accuracy train** | **Accuracy test** |
| school | 0.06914 | 0.10858 |
| sex | 0.01465 | 0.02097 |
| age | 0.01691 | -0.00566 |
| address | 0.02786 | 0.02812 |
| famsize | 0.00188 | 0.00160 |
| Pstatus | 0.00003 | -0.00094 |
| Medu | 0.07708 | 0.00252 |
| Fedu | 0.02901 | 0.07834 |
| Mjob | 0.02726 | 0.00633 |
| Fjob | 0.00387 | -0.00082 |
| reason | 0.01443 | 0.01796 |
| guardian | 0.00744 | 0.00247 |
| traveltime | 0.01512 | 0.01799 |
| studytime | 0.05507 | 0.07904 |
| failures | 0.15233 | 0.15996 |
| schoolsup | 0.00375 | 0.00533 |
| famsup | 0.00755 | -0.01052 |
| paid | 0.00009 | 0.00264 |
| activities | 0.00300 | 0.00420 |
| nursery | 0.00075 | 0.00027 |
| higher | 0.13226 | 0.05234 |
| internet | 0.02508 | 0.01496 |
| romantic | 0.01110 | 0.00126 |
| famrel | 0.01147 | -0.02221 |
| freetime | 0.01399 | 0.01683 |
| goout | 0.00534 | 0.01225 |
| Dalc | 0.03032 | 0.06776 |
| Walc | 0.03106 | 0.03085 |
| health | 0.00749 | 0.01458 |
| absences | 0.01033 | 0.00262 |
| G1 | 0.68262 | 0.68275 |
| G2 | 0.85164 | 0.82148 |

Bảng 4.2. Độ tương quan giữa các biến và Taget

Trực quan sự tương quan dữ liệu:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | |

Trực quan hoá kết quả của một số yếu tố sau khi chạy mô hình hồi quy đơn biến: (Đây là bộ test được chia từ bộ dataset : student-por.csv với tỉ lệ train:test là 7:3)

**Kết luận:** Sau khi chạy xong mô hình hồi quy tuyến tính đơn biến, ta có kết quả độ đo chính xác như Bảng 4.2. Các yếu tố tác động đến G3 (output taget) gồm có: failures 16% (số lần thi trượt), studytime 8% (thời gian học), higher 5% (mong muốn học cao hơn), G1 68% (điểm lần đầu) và G2 82% (điểm lần 2).

### 4.1.3. Mô hình hồi quy đa biến

#### 4.1.3.1 Mô hình hồi quy với tất cả thuộc tính

Trước hết, chúng ta thử chạy mô hình hôi quy đa biến với tất cả các thuộc tính của dataset để biết được tổng quan về bộ dữ liệu này.

Tổng số thuộc tính (biến độc lập) là 32 và taget (biến phụ thuộc) là G3. Ta được kết quả độ đo chính xác accuracy như dưới đây:

Lưu ý: Lựa chọn tỉ lệ ngẫu nhiên giữa tập train và tập test tỉ lệ 7:3

|  |  |  |
| --- | --- | --- |
| **Độ đo** | **Train** | **Test** |
| Accuracy | 0.8664 | 0.8195 |

Nhận xét: Theo kết quả nhận được ta thấy độ đo chính xác là 81.95% gần 82%, cao hơn với tất cả các biến khi hồi quy đơn biến. Chỉ thấp hơn khá là ít so với G2 là 0.0019 tức 0.19% mà thôi.

#### 4.1.3.2 Mô hình hồi quy với 5 thuộc tính quan trọng

Với các kết luận ở mô hình hồi quy tuyến tính đơn biến, ta sử dụng các yếu tố có ảnh hưởng đến G3 để khởi chạy mô hình hồi quy đa biến. Sau khi chạy mô hình ta có kết quả như bảng sau:

Lưu ý: Tương tự mô hình đơn biến ta sử dụng train, test với tỉ lệ 7:3

|  |  |  |  |
| --- | --- | --- | --- |
| **Tên biến** | **Độ đo** | **Train** | **Test** |
| Studytime, higher, failures, G1, G2 | Accuracy | 0.8574 | 0.8276 |

Nhận xét: Kết quả độ chính xác với mô hình hồi quy với 5 biến là 82.76% cao hơn với khi hồi quy đơn biến cho mỗi thuộc tính và khi hồi quy đa biến với tất cả thuộc tính => các biến này có ý nghĩa quan trọng trong việc đánh giá G3

#### 4.1.3.3 Mô hình hồi quy bởi chỉ 2 biến G1, G2 và 1 biến khác

Lưu ý: Lựa chọn tỉ lệ ngẫu nhiên giữa tập train và tập test tỉ lệ 7:3

|  |  |  |
| --- | --- | --- |
| **Tên thuộc tính** | **Accuracy Train** | **Accuracy Test** |
| G1, G2, School | 0.85735 | 0.82417 |
| G1, G2, Sex | 0.85653 | 0.82671 |
| G1, G2, Age | 0.85601 | 0.82454 |
| G1, G2, Address | 0.85661 | 0.82497 |
| G1, G2, Famsize | 0.85609 | 0.82464 |
| G1, G2, Pstatus | 0.85625 | 0.82515 |
| G1, G2, Medu | 0.85598 | 0.82475 |
| G1, G2, Fedu | 0.85598 | 0.82472 |
| G1, G2, Mjob | 0.85606 | 0.82418 |
| G1, G2, Fjob | 0.85678 | 0.82602 |
| G1, G2, Reason | 0.85646 | 0.82532 |
| G1, G2, Guardian | 0.85600 | 0.82428 |
| G1, G2, Traveltime | 0.85653 | 0.82428 |
| G1, G2, Studytime | 0.85629 | 0.82569 |
| G1, G2, Failures | 0.85695 | 0.82668 |
| G1, G2, Schoolsup | 0.85609 | 0.82493 |
| G1, G2, Famsup | 0.85688 | 0.82436 |
| G1, G2, Paid | 0.85613 | 0.82546 |
| G1, G2, Activities | 0.85599 | 0.82468 |
| G1, G2, Nursery | 0.85602 | 0.82482 |
| G1, G2, Higher | 0.85643 | 0.82530 |
| G1, G2, Internet | 0.85612 | 0.82269 |
| G1, G2, Romantic | 0.85599 | 0.82473 |
| G1, G2, Famrel | 0.85607 | 0.82514 |
| G1, G2, Freetime | 0.85606 | 0.82579 |
| G1, G2, Goout | 0.85602 | 0.82388 |
| G1, G2, Dalc | 0.85605 | 0.82270 |
| G1, G2, Walc | 0.85610 | 0.82589 |
| G1, G2, Health | 0.85666 | 0.82534 |
| G1, G2, Absences | 0.85667 | 0.82574 |

**Nhận xét:** Các thuộc tính Sex (82.67%), Fjob(82,60%), Studytime(82.57%), Failures(82.67%), Higher(82.53%) có độ chính xác cao khi kết hợp với G1 và G2, đều cho kết quả trên 82%. Tuy nhiên vẫn không bằng khi chạy hồi quy đa biến với Studytime, higher, failer, G1, G2 là 82.76%.

#### 4.1.3.4 Mô hình hồi quy bởi chỉ 2 biến G1, G2 và Sex, Fjob, Studytime, Failures, Higher

Lưu ý: Lựa chọn tỉ lệ ngẫu nhiên giữa tập train và tập test tỉ lệ 7:3

|  |  |  |
| --- | --- | --- |
| **Tên thuộc tính** | **Accuracy Train** | **Accuracy Test** |
| G1, G2, Sex, Fjob, Studytime, Failures, Higher | 0.8587120152277503 | 0.8298735759324345 |

**Nhận xét:** Ta có thể nhận thấy kết quả độ đo chính xác của 7 thuộc tính này lên G3 lên đến 82.99% gần 83% cao hơn so với cả hồi quy đơn biến từng thuộc tính và cả hôi quy đa biến trước đó.

#### 4.1.3.5 Mô hình hồi quy bởi chỉ G1 và Sex, Fjob, Studytime, Failures, Higher

Lưu ý: Lựa chọn tỉ lệ ngẫu nhiên giữa tập train và tập test tỉ lệ 7:3

|  |  |  |
| --- | --- | --- |
| **Tên thuộc tính** | **Accuracy Train** | **Accuracy Test** |
| G1, Sex, Fjob, Studytime, Failures, Higher | 0.6938375522766445 | 0.6914010521605477 |

**Nhận xét:** Với kết quả độ đo chính chỉ khoảng 69% thấp hơn khá là nhiều so với khi có G2 là 83% từ đó kết luận được biến G2 rất quan trọng trong việc đánh giá G3.

#### 4.1.3.6. Mô hình hồi quy bởi chỉ G1 và từng biến Sex, Fjob, Studytime, Failures, Higher

Lưu ý: Lựa chọn tỉ lệ ngẫu nhiên giữa tập train và tập test tỉ lệ 7:3

|  |  |  |
| --- | --- | --- |
| **Tên thuộc tính** | **Accuracy Train** | **Accuracy Test** |
| G1, Sex | 0.6834508253438332 | 0.686599195277106 |
| G1, Fjob | 0.6830684138386135 | 0.6856800568248224 |
| G1, Studytime | 0.683565444038083 | 0.6846798763664173 |
| G1, Failures | 0.6895480657919607 | 0.688953016768144 |
| G1, Higher | 0.6869288044017896 | 0.6786461845796994 |

**Nhận xét:** Kết quả độ đo chính xác với các biến Sex, Fjob, Studytime, Failures, Higher lần lượt là 68.66%, 68.57%, 68.47%, 68.9%, 67.86% đều thấp hơn so với việc kết hợp G1 với cả 5 thuộc tính là 69% ở trước đó. Vậy 5 biến này có sự tương quan với nhau, khi kêt hợp sẽ cho kết quả tốt hơn.

#### 4.1.3.7 Mô hình hồi quy bởi chỉ G2 và Sex, Fjob, Studytime, Failures, Higher

Lưu ý: Lựa chọn tỉ lệ ngẫu nhiên giữa tập train và tập test tỉ lệ 7:3

|  |  |  |
| --- | --- | --- |
| **Tên thuộc tính** | **Accuracy Train** | **Accuracy Test** |
| G2, Sex, Fjob, Studytime, Failures, Higher | 0.8551812410635053 | 0.827427402654703 |

**Kết luận:** Ở đây thay vì G1 ở 4.1.3.5 thì ta dung G2 và cho kết quả độ đo chính xác cao hơn nhiều lên đến 82.74% trong khi với G1 chỉ 69%. Từ đó ta có thể nhận thấy sự quan trọng của G2 cao hơn G1 khá nhiều trong việc đánh giá G3

#### 4.1.3.8. Mô hình hồi quy bởi chỉ G2 và từng biến Sex, Fjob, Studytime, Failures, Higher

Lưu ý: Lựa chọn tỉ lệ ngẫu nhiên giữa tập train và tập test tỉ lệ 7:3

|  |  |  |
| --- | --- | --- |
| **Tên thuộc tính** | **Accuracy Train** | **Accuracy Test** |
| G2, Sex | 0.8523402911372587 | 0.8235523586069409 |
| G2, Fjob | 0.8523006855613502 | 0.822223952661072 |
| G2, Studytime | 0.8522242813696473 | 0.8229885236985549 |
| G2, Failures | 0.853056096449893 | 0.824141180796137 |
| G2, Higher | 0.8524390217975815 | 0.8226526839059949 |

**Nhận xét: :** Kết quả độ đo chính xác với các biến Sex, Fjob, Studytime, Failures, Higher lần lượt là 82.36%, 82.22%, 82.3%, 82.41%, 82.27% đều thấp so với việc kết hợp 5 thuộc tính G2. Việc này cho thấy sự tương tự như G1: 5 biến này có sự tương quan với nhau, khi kêt hợp sẽ cho kết quả tốt hơn.

## 4.2. Mô hình Support Vector Regression

### 4.2.1 Giới thiệu tổng quan và tính toán độ phù hợp của phương trình hồi quy

#### 4.2.1.1 Giới thiệu

Support Vector Regression (SVR) hoạt động trên các nguyên tắc tương tự như Support Vector Machine (SVM). Người ta có thể nói rằng SVR là một biến thể của của mô hình SVM khi biến phụ thuộc là kiểu dữ liệu số chứ không phải thuộc kiểu dữ liệu phân loại. Một điểm mạnh khi sử dụng mô hình SVR, đó là dựa trên kỹ thuật phi tham số (a non-parametric technique). Không giống như mô hình hồi quy đơn biến (SLR), kết quả phụ thuộc vào các giả thuyết Gauss-Markov, đầu ra của mô hình (output model) SVR không phụ thuộc vào phân phối của các biến độc lập hay các biến phụ thuộc. Thay vào đó, kỹ thuật SVR phụ thuộc vào các hàm kernel. Một ưu điểm khác của SVR là mô hình cho phép xây dựng mô hình phi tuyến tính mà không thay đổi các biến giải thích (explanatory variables, biến giải thích là một loại biến độc lập), giúp giải thích tốt hơn kết quả mô hình. Ý tưởng cơ bản của SVR là không quan tâm đến dự đoán, miễn là lỗi nhỏ hơn giá trị nhất định. Điều này được gọi là nguyên tắc maximal margin. Ý tưởng về maximal margin này cho phép xem SVR như một vấn đề tối ưu hóa lồi. Mô hình hồi quy cũng có thể bị phạt (penalized) bằng cách sử dụng tham số chi phí (cost parameter), để phù hợp với dữ liệu tránh trường hợp quá khớp (over-fitting).

Kỹ thuật SVR dựa vào các hàm kernel để xây dựng mô hình. Một số hàm kernel thông dụng như là Linear, Polynomial, Sigmoid và Radial Basis. Khi sử dụng kỹ thuật SVR, ta nên chọn một hàm kernel tiếp cận phù hợp với dữ liệu. Việc lựa chọn kernel nào là phù hợp là một công việc khó, đòi hỏi các kỹ thuật tối ưu hóa để lựa chọn ra đưa kernel function tốt nhất.

Đối với bài toán với bộ dữ liệu như thế này, chúng tôi sử dụng hàm kernel được cài đặt mặc định do framework sklearn python cung cấp. Radius Basis Function (RBF) kernel được sử dụng cho mô hình trên. Hàm kernel biến đổi dữ liệu từ không gian phi tuyến sang không gian tuyến tính. Kỹ thuật dựa trên hàm kernel này cho phép mô hình SVR tìm ra sự phù hợp và ánh xạ dữ liệu tới không gian ban đầu

Mô hình SVR được xây dựng như sau:



#### 4.2.1.2 Tính toán độ phù hợp của phương trình hồi quy

Sử dụng độ đánh giá R-squared để đo mức độ phù hợp của đường hồi quy với dữ liệu quan sát.

R-squared là một thước đo thống kê biểu thị tỷ lệ phương sai cho một biến phụ thuộc được giải thích bởi một biến độc lập hoặc các biến trong mô hình hồi quy. Nó cũng có thể được gọi là hệ số xác định (coefficient of determination).

Công thức cho R-squared:



#### 4.2.1.3 Đánh giá

Chia dữ liệu theo tỉ lệ 7:3. 70% dòng đầu của tập dữ liệu được lấy để huấn luyện và xây dựng mô hình, 30% còn lại được sử dụng để đánh giá mô hình dựa trên thước đo R-squared.

### 4.2.2 Mô hình SVR với một thuộc tính

|  |  |  |
| --- | --- | --- |
| **Thuộc tính** | **Accuracy Train** | **Accuracy Test** |
| school | 0.02267 | -0.00029 |
| sex | 0.02806 | -0.17981 |
| age | 0.02892 | -0.19799 |
| address | -0.00632 | -0.11577 |
| famsize | -0.02140 | -0.25637 |
| Pstatus | 0.00011 | -0.20588 |
| Medu | 0.06029 | -0.10630 |
| Fedu | 0.02728 | -0.11376 |
| Mjob | 0.03602 | -0.14622 |
| Fjob | 0.02008 | -0.23669 |
| reason | 0.02126 | -0.17710 |
| guardian | 0.00206 | -0.23667 |
| traveltime | 0.01697 | -0.20039 |
| studytime | 0.06377 | -0.13120 |
| failures | 0.15365 | 0.02242 |
| schoolsup | 0.02232 | -0.26644 |
| famsup | 0.00321 | -0.23996 |
| paid | 0.00122 | -0.18255 |
| activities | -0.00398 | -0.17908 |
| nursery | 0.00140 | -0.24838 |
| higher | 0.13061 | -0.10088 |
| internet | 0.00841 | -0.16206 |
| romantic | 0.00342 | -0.23450 |
| famrel | 0.00364 | -0.16627 |
| freetime | 0.01830 | -0.23123 |
| goout | 0.04638 | -0.19057 |
| Dalc | 0.04872 | -0.16216 |
| Walc | 0.06242 | -0.22016 |
| health | 0.00370 | -0.17305 |
| absences | 0.05012 | -0.22391 |
| G1 | 0.68513 | 0.56390 |
| G2 | 0.82158 | 0.64569 |

**Kết luận:** Dựa trên R2 score trên tập test, ta có thể 5 thuộc tính cho kết quả cao nhất theo thứ tự giảm dần là G2, G1, failures, higher và school.

### 4.2.3. Mô hình SVR với đa thuộc tính

#### 4.2.2.1 Mô hình SVR với 5 thuộc tính có độ chính xác cao nhất

|  |  |  |
| --- | --- | --- |
| **Thuộc tính** | **Accuracy train** | **Accuracy test** |
| higher, failures,  school, G1, G2 | 0.8410282238061523 | 0.6187923868963905 |

Kết luận: Sau khi thử nghiệm trên tập dữ liệu huấn luyện với 5 thuộc tính có độ đánh giá cao nhất từ mô hình huấn luyện với một thuộc tính, kết quả cho thấy trên tập dữ liệu kiểm tra, mô hình cho kết quả là 61% trên tổng toàn tập kết quả.

#### 4.2.2.2 Mô hình SVR với tất cả biến độc lập ( trừ biến mục tiêu )

|  |  |
| --- | --- |
| Accuracy Train | Accuracy Test |
| 0.9311991427798605 | 0.5559797156534803 |

Kết luận: Thử nghiệm trên với 32 thuộc tính, ta thấy kết quả của mô hình là 55%. Có thể nhận thấy rằng có khá nhiều thuộc tính nhiễu hay không tương quan với biến đầu ra G3.

#### 4.2.2.3 Mô hình SVR bởi chỉ 2 biến G1, G2 và 1 biến khác

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Accuracy Train** | **Accuracy Test** |
| schoolsup | 0.830513 | 0.660917 |
| absences | 0.830414 | 0.636304 |
| paid | 0.826419 | 0.631134 |
| school | 0.82578 | 0.62307 |
| nursery | 0.829668 | 0.622423 |
| famsize | 0.827231 | 0.614309 |
| failures | 0.833559 | 0.606954 |
| higher | 0.826913 | 0.60345 |
| studytime | 0.827794 | 0.602037 |
| reason | 0.829311 | 0.601552 |
| activities | 0.826133 | 0.597938 |
| Fedu | 0.829861 | 0.597838 |
| internet | 0.825276 | 0.595365 |
| romantic | 0.823873 | 0.59095 |
| guardian | 0.825888 | 0.589286 |
| traveltime | 0.830551 | 0.586406 |
| health | 0.828363 | 0.585457 |
| sex | 0.826013 | 0.576837 |
| Pstatus | 0.824073 | 0.566104 |
| Medu | 0.829176 | 0.566077 |
| Mjob | 0.825566 | 0.563534 |
| famsup | 0.825331 | 0.561965 |
| famrel | 0.828389 | 0.557686 |
| age | 0.833471 | 0.555152 |
| address | 0.827633 | 0.554715 |
| Walc | 0.828115 | 0.552682 |
| freetime | 0.827447 | 0.54661 |
| goout | 0.833163 | 0.538188 |
| Fjob | 0.825739 | 0.537089 |
| Dalc | 0.825725 | 0.524078 |

**Kết luận:** Thử nghiệm nhận thấy, hai biến G1 và G2 cho kết quả khá tương quan và tỉ lệ thuận với biến mục tiêu nên chúng tôi thử khảo sát hai biến này với các biến khác xem kết quả như thế nào. Kết quả cho ra khá bất ngờ, khi độ chính xác của 5 thuộc tính đầu tiên trong bảng có độ chính xác cao hơn 60%.

#### 4.2.2.4 Mô hình SVR bởi chỉ 2 biến G1, G2 và schoolsup, absences, paid, school, nursery, famsize

|  |  |
| --- | --- |
| **Accuracy Train** | **Accuracy Test** |
| 0.8507591633306768 | 0.6839653494843865 |

**Kết luận:** Tiếp tục thử nghiệm, lần này chúng tôi sử dụng 6 thuộc tính cho kết quả cao nhất ban đầu, độ chính xác của kết quả còn cao hơn so với lần trước là xấp xỉ 68%. Cao nhất trong các lần thử nghiệm, có thể thấy mối tương quan thuận giữa 6 thuộc tính schoolsup, absences, paid, school, nursery, famsize và G1,G2.

#### 4.2.2.5 Mô hình SVR bởi chỉ 2 biến G1, G2 và schoolsup, absences, paid, school, nursery

|  |  |
| --- | --- |
| **Accuracy Train** | **Accuracy Test** |
| 0.8431194756716491 | 0.6557203107961298 |

**Kết luận:** Từ mô hình ở mục lục 4.2.2.4, chúng tôi giả sử một trong các thuộc tính trong đó không có mối tương quan thuận nên đã loại bỏ thuộc tính thứ 6 có độ chính xác thấp hơn, thì thấy kết quả cho ra thấp hơn dự đoán. Vậy thuộc tính thứ 6 này ít nhiều cũng có tương quan vả ảnh hưởng đến biến đầu ra G3.

#### 4.2.2.6 Mô hình SVR bởi chỉ 2 biến G1, G2 và schoolsup, absences, paid, shool, nursery, famsize, failures

|  |  |
| --- | --- |
| **Accuracy Train** | **Accuracy Test** |
| 0.8575917904616621 | 0.6437325039068669 |

**Kết luận:** Từ mô hình ở mục lục 4.2.2.4, chúng tôi cũng giả sử như vậy, nếu thêm một thuộc tính vào thì tương quan của thuộc tính đó có khả năng tăng kết quả dự đoán của mô hình lên nữa không. Với độ chính xác 64% giảm so với kết quả của mô hình dự đoán tốt nhất là 68% thì thấy rằng thuộc tính mới được thêm vào là failures không có tương quan thuận và ảnh hưởng nhiều để cho ra kết quả tốt hơn.

#### 4.2.2.7 Mô hình SVR bởi biến G2 và schoolsup

|  |  |
| --- | --- |
| **Accuracy Train** | **Accuracy Test** |
| 0.8276718580358045 | 0.6826368342194155 |

Ngược lại khi làm điều này với G1, nghĩa là kết hợp G1 với một biến bất kỳ trong 5 biến kết hợp với G1 và G2 cho kết quả tương đối và không cao so với G2 và schoolsup (một trong những kết quả cao nhất). Có thể thấy biến G2 có ý nghĩa rất quan trọng trong việc đánh giá điểm cuối kỳ (G3) của sinh viên và nhận hỗ trợ giao dục thêm từ nhà trường (schoollsup) cũng quan trọng không kém.

**Kết luận:**

- 5 thuộc tính có tương quan thuận cao nhất khi được dự đoán với mô hình SVR là schoolsup, absences, paid, school, nursery.

- Thuộc tính schoolsup kết hợp với G2 cho biết ý nghĩa nhiều nhất. Rằng nhận hỗ trợ giao dục thêm từ nhà trường (schoolsup) cũng là một yếu tố quyết định đến điểm cuối kỳ của sinh viên và điểm thi lần thứ 2 (G2) có ảnh hưởng lớn đến điểm cuối kỳ G3.

### 4.2.4. So sánh và nhận xét giữa các mô hình

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên mô hình với biến độc lập | Accuray Train | Accuracy Test |
| **1** | **G2** | **0.82** | **0.64** |
| 2 | G1 | 0.68 | 0.56 |
| 3 | higher, failures, school, G1, G2 | 0.8410 | 0.6188 |
| 4 | 32 biến độc lập | 0.9312 | 0.5560 |
| **5** | **G1, G2, schoolsup** | **0.8305** | **0.6609** |
| 6 | G1, G2, absences | 0.8304 | 0.6363 |
| 7 | G1, G2, paid | 0.8264 | 0.6311 |
| 8 | G1, G2, school | 0.8258 | 0.6231 |
| 9 | G1, G2, nursery | 0.8296 | 0.6224 |
| 10 | G1, G2, famsize | 0.8272 | 0.6143 |
| 11 | G1, G2, failures | 0.8335 | 0.6069 |
| **12** | **G1, G2 và  schoolsup, absences, paid, school, nursery, famsize** | **0.8508** | **0.6840** |
| 13 | **G1, G2 và schoolsup, absences, paid, school, nursery** | **0.8431** | **0.6557** |
| 14 | G1, G2 và schoolsup, absences, paid, shool, nursery, famsize, failures | 0.8575 | 0.6437 |
| 15 | **G2 và schoolsup** | **0.8276** | **0.6826** |

Nhận xét:

Trong 15 mô hình khảo sát thì có khoảng 5 mô hình được đánh giá là cho kết quả tương đối tốt so với các mô hình còn lại.

Hai mô hình cho kết quả cao nhất là mô hình thứ 12 và mô hình thứ 15. Mô hình thứ 15 với chỉ hai thuộc tính nhưng đã cho ra kết quả gần như là tốt nhất trong tất cả những mô hình khảo sát được.

# CHƯƠNG 5: KẾT LUẬN

Ngoài mối liên hệ giữa điểm số của học sinh sinh viên, các yếu tố như giới tính, trình độ học vấn của cha mẹ, số lần thi trượt, ý định học lên cao hơn, thời gian học cũng có ảnh hưởng đến điểm số của học sinh/sinh viên như sau:

Học sinh/sinh viên nữ dường như có điểm số cao hơn

Trình độ học vấn của cha mẹ có ảnh hưởng đến kết quả học tập của con cái

Học sinh/ sinh viên tiêu thụ ít rượu hơn có kết quả học tập cao hơn.

Học sinh/ sinh viên thi trượt càng nhiều điểm số càng thấp.

Học sinh/ sinh viên có định đọc học lên cao hơn thì có điểm số cao hơn.

Số lần thất bại, ý định học lên cao, điểm số lần 1, lần 2 là những yếu tố dự đoán quan trọng cho mô hình hồi quy tuyến tính.

Điểm thi cuối kỳ lần thứ ba (G3) có tương quan thuận mạnh mẽ với yếu tố điểm thi lần thứ nhất (G1) và điểm thi lần thứ hai (G2). Bên cạnh đó, các yếu tố như hỗ trợ giáo dục từ nhà trường (schoolsup), số lần vắng học (absences), yêu cầu học phụ đạo (paid), nơi trường học sinh học (school), trước kia có học mẫu giáo (nursery) là những yếu tố tác động mạnh đến mô hình Support Vector Regression.

Từ đồ án ta có thể thấy khi có một bộ dữ liệu ta có thể thực nghiệm để có cái nhìn khách quan hơn về những yếu tố ảnh hưởng đến biến phụ thuộc, những yếu tố hay interaction nào quan trọng hay không quan trọng để từ đó xây dựng được mô hình dự đoán được hướng phát triển phù hợp với mục đích thực nghiệm.

# TÀI LIỆU THAM KHẢO

[1] P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.