BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TPHCM

KHOA CÔNG NGHỆ THÔNG TIN

MÔN: LẬP TRÌNH R CHO PHÂN TÍCH

BÁO CÁO

ĐỀ TÀI:

PHÂN TÍCH CÁC YẾU TỐ NGOẠI CẢNH ẢNH HƯỞNG ĐẾN ĐIỂM SỐ CỦA HỌC SINH

GVHD: QUÁCH ĐÌNH HOÀNG

SVTH: TRẦN VĂN DUY 19133016

LÊ PHƯƠNG NAM 19133036

TRẦN CÔNG TRƯỜNG 19133062

CAO ANH VĂN 19133067

TP. Hồ Chí Minh, ngày 15 tháng 12 năm 2021

Mục lục

1. Tóm tắt (abstract) 3

2. Giới thiệu (introduction) 3

3. Dữ liệu (data) 4

4. Trực quan hóa dữ liệu (data visulization) 5

5. Mô hình hóa dữ liệu (data modeling) 7

6. Thực nghiệm, kết quả, và thảo luận (experiments, results, and discussion) 9

6.1. Dự đoán sử dụng thuật toán kNN 9

6.1.1. Thực nghiệm 9

6.1.2. Kết quả: 9

6.1.3. Thảo luận: 10

6.2. Dự đoán sử dụng thuật toán linear regression 10

6.3. Tìm biến nào ảnh hưởng đến điểm số nhất 13

7. Kết luận (conclusion) 13

8. Phụ lục (appendices) 14

9. Đóng góp (contributions) 14

10. Tham khảo (references) 15

1. Tóm tắt (abstract)

Sau một thời gian dài học tập và rèn luyện, chúng em nhận thấy điểm số là một yếu tố quan trọng đối với quá trình học tập của học sinh. Thông qua điểm số có thể nhìn thấy được khả năng học tập của học sinh đó. Vì vậy chúng em quyết định sẽ tìm ra các yếu tố tác động lên điểm số của học sinh, để hiểu được ảnh hưởng của nền tảng cha mẹ, việc chuẩn bị bài kiểm tra, v.v. đối với kết quả học tập của học sinh. Và nhờ đó chúng em có thể đưa ra các dự đoán về điểm số của học sinh. Và sau khi trả lời được các câu hỏi đó chúng ta sẽ có cách để giúp học sinh sẽ có điểm số tốt hơn trong tương lai.

Phương pháp nhóm sử dụng để thực hiện bao gồm phương pháp phân tích mô tả để để trực quan hóa dữ liệu thông qua các biểu đồ và phương pháp phân tích chuẩn đoán. Trong phương pháp phân tích chuẩn đoán chúng em có sử dụng các mô hình machine learning là linear regression và K-NN để đưa ra dự đoán về điểm trung bình của học sinh.

2. Giới thiệu (introduction)

Trong thời đại 4.0, việc ứng dụng công nghệ vào mọi lĩnh vực đời sống đóng vai trò vô cùng quan trọng, đặc biệt là ứng dụng các công nghệ trong phân tích dữ liệu giúp chúng ta dự đoán được những kết quả sẽ xảy ra trong tương lai và tìm một giải pháp để đưa ra kết quả tốt nhất. Vì vậy, trong bài phân tích này, chúng em sẽ ứng dụng ngôn ngữ R để phân tích thành tích học tập của học sinh qua các kỳ thi, vận dụng kiến thức đã học về phân tích dữ liệu bằng ngôn ngữ R, sau đó tìm những thuộc tính ảnh hưởng tới điểm của học sinh. Từ những phân tích này có thể đưa ra dự đoán về kết quả làm bài thi trong tương lai của học sinh.

Câu hỏi nghiên cứu của nhóm bao gồm yếu tố nào ảnh hưởng nhất đến điểm thi của học sinh và dự đoán điểm thi của học sinh. Việc tìm ra được yếu tố nào ảnh hưởng nhất đến điểm thi của học sinh là rất quan trọng. Bởi vì mỗi học sinh đều có một hoàn cảnh, và các yếu tốt tác động khác nhau lên bản thân, điều này sẽ hình thành nên cách nhận thức và khả năng học tập của học sinh đó. Và kết quả là mỗi người sẽ có một điểm thi khác nhau, có thể thấp, có thể cao. Vì vậy nếu tìm ra được yếu tố nào ảnh hưởng nhất đến khả năng thi cử của học sinh ta có thể tìm ra những phương pháp tối ưu nhất tác động lên các yếu tố đó và giúp nâng cao được chất lượng thi của học sinh sau này.

Câu hỏi nghiên cứu thứ 2 là dự đoán điểm thi của học sinh thông qua các yếu tố tác động lên học sinh. Hãy tưởng tượng những gì một giáo viên có thể làm nếu cô ấy biết vào đầu năm học rằng một học sinh có vẻ học giỏi thực sự được dự đoán chỉ rơi vào tình trạng thiếu thành thạo trong bài kiểm tra cuối năm. Đây có thể là dấu hiệu cảnh báo sớm mà cô ấy cần điều chỉnh việc giảng dạy ngay lập tức, thay vì vài tuần hoặc vài tháng sau đó khi có nhiều dấu hiệu rõ ràng hơn cho thấy học sinh đang tụt lại phía sau? Có thể can thiệp sớm là điều giúp học sinh đó đi đúng hướng để thành công không? Có bao nhiêu sinh viên nữa có thể đạt được trình độ thông thạo nếu giáo viên của họ có quyền dự đoán chính xác về kết quả kiểm tra trước nhiều tháng? Thông qua các dự đoán này, giáo viên hoặc giảng viên sẽ có cái nhìn tổng quan về học sinh của mình, và từ đó các phương pháp dạy, truyền đạt kiến thức cho phù hợp một cách sớm nhất. Và các nhà giáo dục càng biết rõ về học sinh của mình thì sẽ càng có lộ trình đào tạo hiệu quả hơn. Từ đó nâng cao chất lượng của học sinh, sinh viên.

Input của bài toán là các yếu tố cá nhân, xã hội và kinh tế có tác động tương tác lên chúng điểm thi bao gồm giới tính, sắc tộc, trình độ học vấn của cha mẹ, bữa ăn trưa, đã hoàn thành khóa luyện thi hay chưa. Output của bài toán là điểm trung bình, điểm trung bình sẽ bằng trung bình của điểm toán, điểm đọc và điểm viết. Chúng em sử dụng các thuật toán linear regression và K-NN để đưa ra dự đoán về điểm trung bình của học sinh.

3. Dữ liệu (data)

Nguồn dữ liệu của nhóm được được lấy từ một vài trường trung học cơ sở ở Mỹ và được tổng hợp bởi trang kaggle.com:
https://www.kaggle.com/spscientist/studentsperformance-in-exams
Tập dữ liệu này bao gồm 8 cột: bao gồm điểm của ba kỳ thi và nhiều yếu tố cá nhân, xã hội và kinh tế tương tác với chúng:
gender: giới tính
race/ethnicity: dân tộc
parental level of education: trình độ học vấn của bố mẹ
lunch: bữa ăn trưa
test preparation course: khóa học luyện thi
math score: điểm toán
reading score: điểm đọc
writing score: điểm viết
Một vài dòng dữ liệu của tập dữ liệu:

	gender <fctr>	race.ethnicity <fctr>	parental.level.of.education <fctr>	lunch <fctr>
1	female	group B	bachelor's degree	standard
2	female	group C	some college	standard
3	female	group B	master's degree	standard
4	male	group A	associate's degree	free/reduced
5	male	group C	some college	standard
6	female	group B	associate's degree	standard

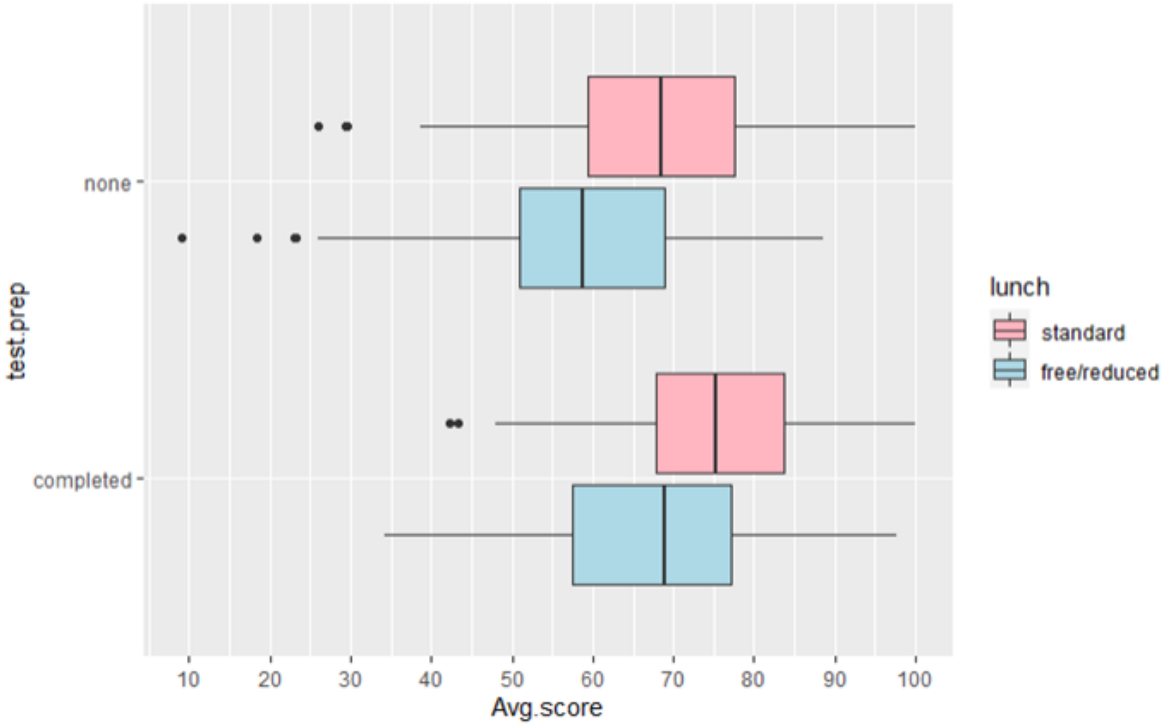
Hình 3.1

Tập dữ liệu của nhóm không có các giá trị null hay bất cứ các giá trị không hợp lệ nào:

```
> sapply(data, function(x)sum(is.na(x)))
                gender         race.ethnicity parental.level.of.education                 lunch
                     0                      0                          0                     0
test.preparation.course             math.score          reading.score         writing.score
                     0                      0                          0                     0
```

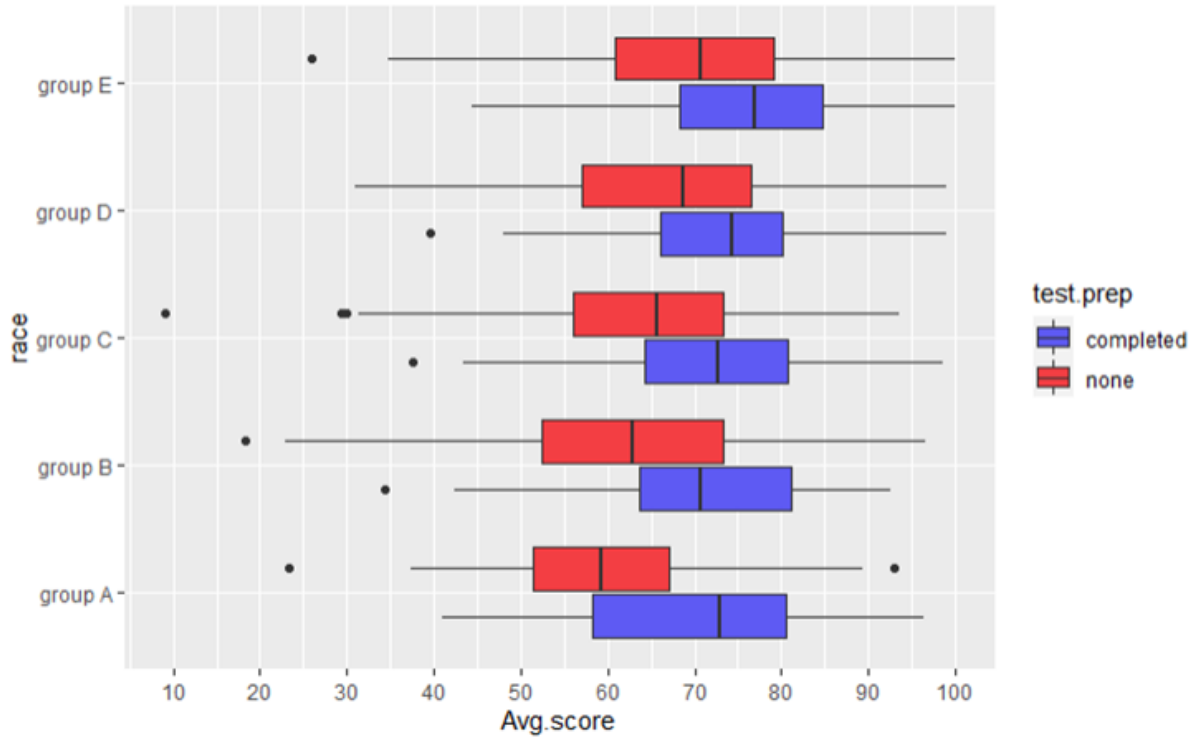
Hình 3.2

4. Trực quan hóa dữ liệu (data visulization)



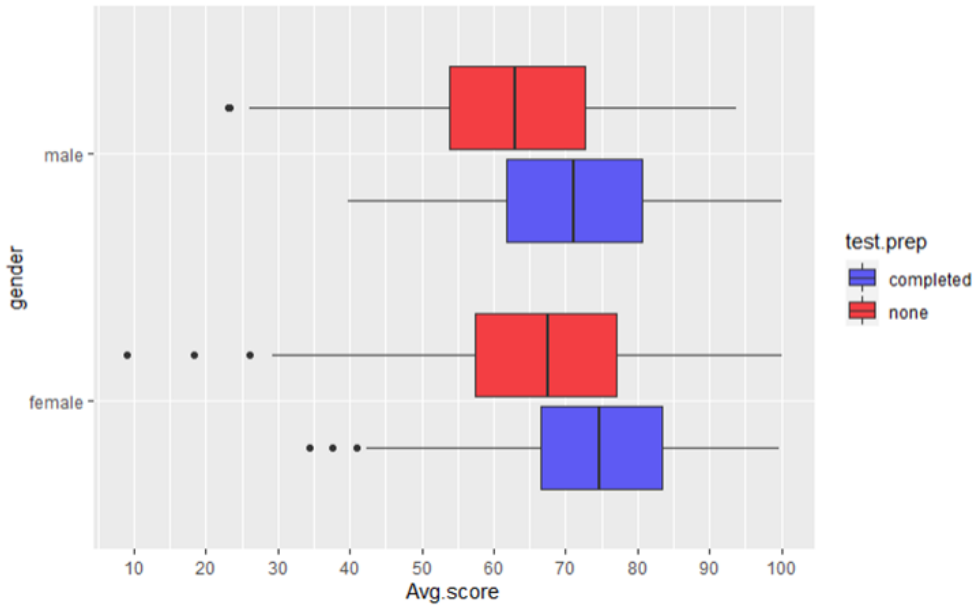
Hình 4.1: Biểu đồ thể thể hiện sự phân bố điểm trung bình theo bữa ăn trưa và khóa học luyện thi.

Nhận xét: Học sinh có bữa theo tiêu chuẩn và đã hoàn thành các khóa học luyện thi sẽ có điểm trung bình cao hơn so với những học sinh chưa hoàn hoặc không hoàn thành các khóa học luyện thi và bữa ăn trưa miễn phí hoặc giảm giá.



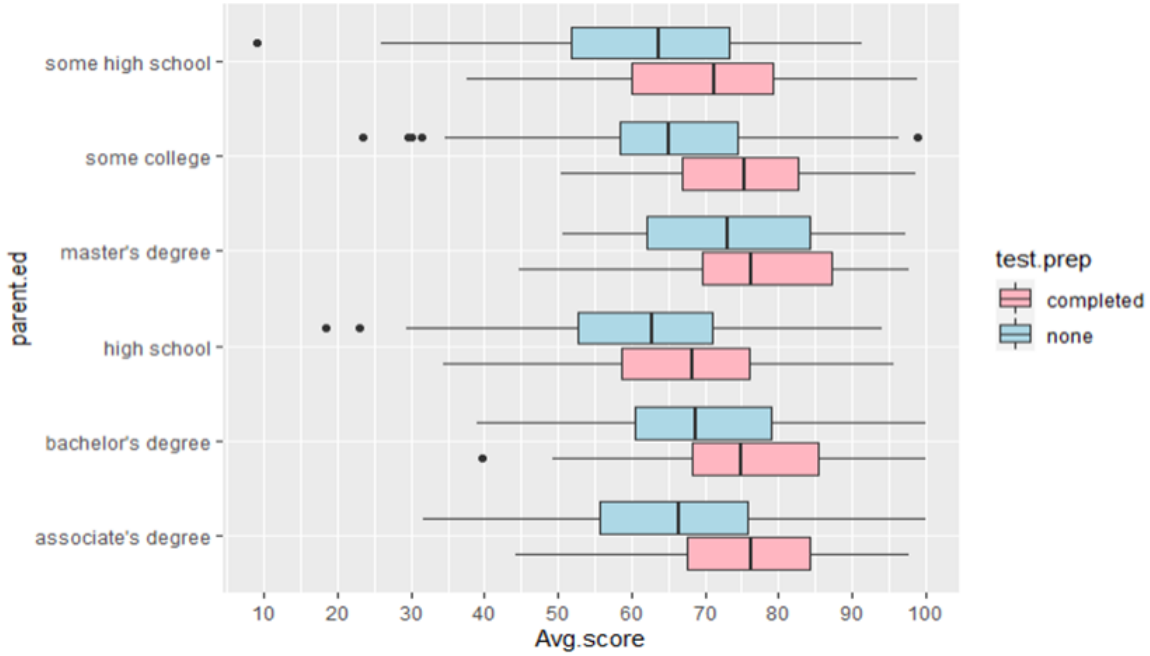
Hình 4.2: Biểu đồ thể hiện sự phân bố điểm trung bình theo các nhóm dân tộc và các khóa luyện thi.

Nhận xét: Điểm trung bình của học sinh nhóm E khi hoàn thành khóa luyện thi cao đến 76,7. Thấp nhất là học sinh nhóm A khi môn luyện thi không có điểm nào bằng 59,2.



Hình 4.3: Biểu đồ thể hiện sự phân bố điểm trung bình theo giới tính và khóa luyện thi.

Nhận xét: Điểm trung bình của nữ với test.prep as Complete cao tới 74,5. Thấp nhất là nam với test.prep không ai bằng 63.



Hình 4.4: Biểu đồ thể hiện sự phân bố điểm trung bình theo trình độ học vấn của bố mẹ và các khóa luyện thi.

Nhận xét: Những sinh viên có trình độ học vấn của cha mẹ là thạc sĩ và đã hoàn thành bài thi test.prep có điểm trung bình cao nhất là 76,3. Học sinh có trình độ học vấn thấp nhất của cha mẹ là trung học phổ thông và test.prep không có điểm số trung bình là 61,6.

*Nhận xét rút ra từ phân tích EDA:

Nữ có điểm trung bình cao hơn nam.

Học sinh ăn trưa theo tiêu chuẩn có điểm số cao hơn.

Khi hoàn thành khóa luyện thi, điểm cao hơn không hoàn thành khóa luyện thi.

Cha mẹ có trình độ học vấn càng cao thì điểm trung bình của con cái càng cao.

Không có nhiều sự khác biệt về điểm số trung bình giữa các sinh viên thuộc các sắc tộc khác nhau.

5. Mô hình hóa dữ liệu (data modeling)

Để trả lời cho câu hỏi biến nào ảnh hưởng nhất đến điểm số em sử dụng trị tuyệt đối của hệ số chuẩn hóa từ mô hình linear regression để đánh giá mức độ ảnh hưởng. Còn về câu hỏi dự đoán điểm trung bình thì tụi em sử dụng hai thuật toán là kNN và linear regression.

Ý tưởng về thuật toán kNN

- kNN là thuật toán supervised-learning đơn giản nhất.
- kNN đi tìm đầu ra của một điểm dữ liệu mới bằng cách chỉ dựa trên thông tin của k điểm dữ liệu trong training set gần nó nhất.
- Hàm hồi quy k-nearest neighbors (kNN) cho một input point x là:

$$f(x) = \frac{1}{k} \sum_{x_i \in N_k(x, D)} y_i$$

Hình 5.1. Công thức tính f(x) với thuật toán kNN

Kết quả đầu ra sẽ là trung bình của tổng các yi tương ứng với k điểm xi gần input x nhất.

Chú thích:

+ k là số điểm gần x nhất trong training dataset D.
+ D là training dataset.

Ý tưởng chính của Linear Regression là cố gắng tìm một đường thẳng để đường thẳng đó gần với tất cả các điểm trên đồ thị của chúng ta nhất có thể (khoảng cách từ đường thẳng đó đến các điểm là nhỏ nhất).

Mô hình hồi quy tuyến tính có dạng:

$$\hat{y} = f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_d x_d = \theta^T x$$

Hình 5.2

Chú thích:

- $\theta^T = (\theta_0, \theta_1, \cdots, \theta_d) \in R^{d+1}$: là tham số của mô hình.
- $x = (1, x_1, \cdots, x_d)^T \in R^{d+1}$: là đầu vào của mô hình
- Ta quy ước x là vector dạng cột, x^T là vector dạng dòng.
- Khi cần nhấn mạnh, ta viết $f(x)$ để mô tả sự phụ thuộc của f vào.
- θ_0 là dự đoán của mô hình khi tất cả các đặc trưng bằng 0.

Sau khi đã có công thức cho hàm dự đoán công việc tiếp theo của ta là tìm ra các trọng số θ_0 và θ_1 để có thể vẽ được đường thẳng mà ta đã đề ra ban đầu. Cost Function là một hàm dùng để đo độ chính xác của hàm dự đoán từ đó giúp tìm các trọng số tối ưu cho hàm dự đoán. Hàm này được định nghĩa như sau:

$$\mathcal{L}(\theta_0, \theta_1) = \frac{1}{2m} * \sum_{i=1}^{m} [y^{(i)} - h_\theta(x^{(i)})]^2$$
$$= \frac{1}{2m} * \sum_{i=1}^{m} [y^{(i)} - (\theta_0 + \theta_1 x^{(i)})]^2$$

Hình 5.3

Với sự giúp đỡ của Cost Function, bây giờ bài toán tìm trọng số tối ưu của ta có thể chuyển thành tìm các trọng số $_0$, $_1$ để giá trị của Cost Function là nhỏ nhất. Mục tiêu ban đầu là tìm một đường thẳng sao cho đường thẳng đó gần các dữ liệu nhất có thể. Khi giá trị Cost Function nhỏ nhất (khoảng cách giữa các giá trị dự đoán và giá trị thực nhỏ nhất) thì nó cũng mang ý nghĩa tương tự.

6. Thực nghiệm, kết quả, và thảo luận (experiments, results, and discussion)

6.1. Dự đoán sử dụng thuật toán kNN

6.1.1. Thực nghiệm

Xây dựng model trên traning dataset.

- Chia tập dữ liệu theo tỷ lệ 8-2, tập train (80%) và tập test(20%).
- Kết quả :
    - Chọn k cho thuật toán: hàm train đã chọn ra k=31 là tốt nhất.

```
k-Nearest Neighbors

801 samples
  5 predictor

Pre-processing: centered (11), scaled (11)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 722, 720, 721, 721, 721, 721, ...
Resampling results across tuning parameters:

  k   RMSE      Rsquared   MAE
   5  13.76883  0.1345269  11.06127
   7  13.72685  0.1268785  11.05080
   9  13.61541  0.1338190  10.98496
  11  13.51235  0.1353655  10.85728
  13  13.48751  0.1358011  10.84222
  15  13.50074  0.1310329  10.85152
  17  13.44781  0.1368179  10.84331
  19  13.35134  0.1474580  10.74545
  21  13.34443  0.1500418  10.71130
  23  13.32992  0.1530089  10.72878
  25  13.30635  0.1552723  10.73844
  27  13.30852  0.1547458  10.73078
  29  13.31767  0.1536350  10.72688
  31  13.29804  0.1568298  10.69552
  33  13.33737  0.1510595  10.74077
  35  13.35958  0.1481704  10.76163
  37  13.35802  0.1476372  10.74933
  39  13.32534  0.1534623  10.70785
  41  13.30526  0.1578998  10.68374
  43  13.30012  0.1594019  10.67299

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 31.
```

Hình 6.1.1.1. Kết quả k cho thuật toán

- k=31 thì RMSE nhỏ nhất là 13.29804.

6.1.2. Kết quả:

- Kết quả dự đoán của mô hình em sẽ lưu vào biến preKnn ở tập test và ta thấy được ở hình ở trên cùng, ví dụ dòng đầu tiên điểm thực tế là 82.33 , dự đoán là 74,75.

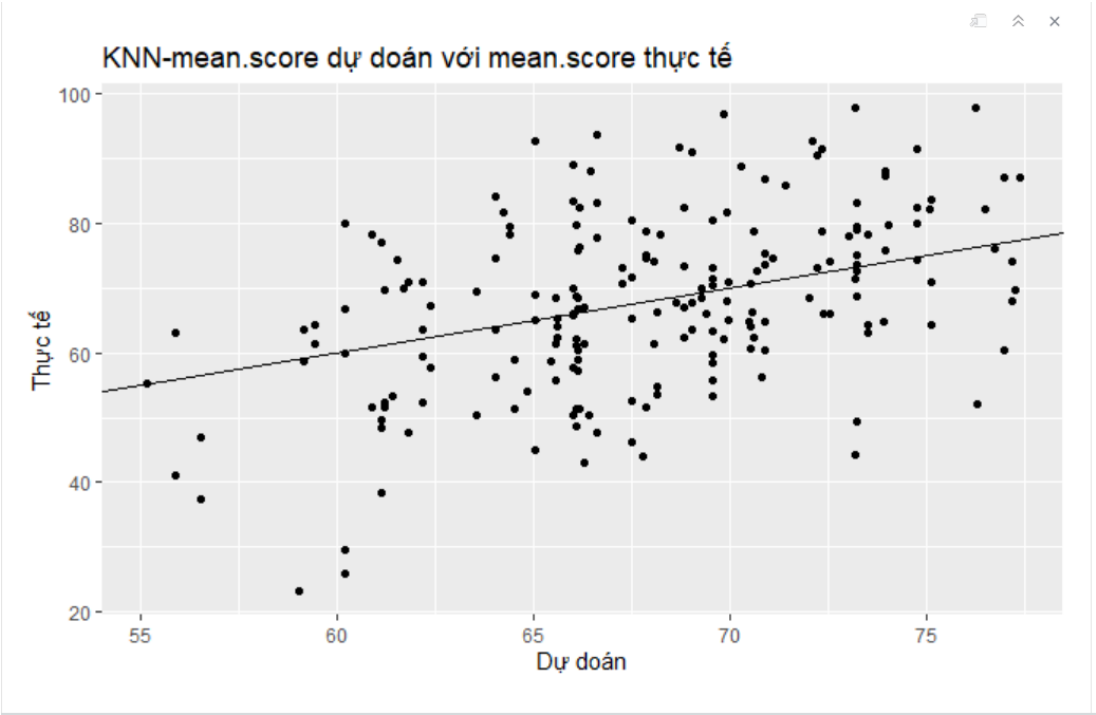
Description: df [199 x 10]

test.preparation.course <fctr>	math.score <int>	reading.score <int>	writing.score <int>	mean.score <dbl>	predKnn <dbl>
completed	69	90	88	82.33333	74.75833
none	90	95	93	92.66667	72.10753
completed	88	95	92	91.66667	68.69792
none	18	32	28	26.00000	60.20202
none	54	58	61	57.66667	62.39024
none	69	54	55	59.33333	62.18280
none	58	73	68	66.33333	68.16000
completed	77	69	68	71.33333	73.18627
none	88	78	75	80.33333	67.50833
none	39	39	34	37.33333	56.53333
none	60	72	74	68.66667	66.07843

1-11 of 199 rows | 6-11 of 10 columns Previous 1 2 3 4 5 6 ... 19 Next

Hình 6.1.2.1 Tập test khi thêm biến preKnn

- RMSE là : 12,20 và R-square là : 0.2002683



Hình 6.1.2.2 Biểu đồ tương quan giữa điểm trung bình thực tế và dự đoán

Nhận xét: Nhìn vào sự phân bố và vị trí của đường thẳng ta có thể thấy mô hình dự đoán cho ra kết quả chưa thực sự tốt.

6.1.3. Thảo luận:

- Với mô hình kNN ta thấy RMSE vẫn còn khá cao và R-square còn khá thấp.
- Mô hình chưa thực sự tốt.

6.2. Dự đoán sử dụng thuật toán linear regression

Đầu tiên em chia tập dữ liệu ra làm 2 phần train và test. Tập train gồm 80% và tập test gồm 20%. Sau đó em lại chia tập train ra làm 2 phần, một phần gồm 80% dữ liệu tập train (train1), phần còn lại là 20% dữ liệu tập train (train2). Sau khi đã chia tập dữ liệu xong em tiến hành train model trên tập train1 và dùng tập train2 để dự đoán. Trong quá trình train tập train1 và dự đoán trên tập train2 thì em có thử điều chỉnh các biến dự đoán khác nhau và nhận được nhiều kết quả khác nhau. Sau quá trình này thì em quyết định chọn model với tất cả các biến của tập dữ liệu trừ các biến điểm toán, điểm đọc, điểm viết

```
Call:
lm(formula = Avg.score ~ . - m.score - r.score - w.score, data = Student_train1)

Residuals:
    Min      1Q  Median      3Q     Max
-48.129  -8.134   0.775   9.242  26.778

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 68.5733     2.3019  29.789  < 2e-16 ***
gendermale                  -3.7386     1.0181  -3.672 0.000261 ***
racegroup B                  0.4109     2.1003   0.196 0.844940
racegroup C                  1.3195     1.9616   0.673 0.501426
racegroup D                  4.3134     1.9928   2.164 0.030813 *
racegroup E                  7.1373     2.2602   3.158 0.001667 **
parent.edbachelor's degree   1.7228     1.7985   0.958 0.338484
parent.edhigh school        -5.0697     1.6229  -3.124 0.001868 **
parent.edmaster's degree     3.2621     2.3051   1.415 0.157519
parent.edsome college       -1.0817     1.5241  -0.710 0.478127
parent.edsome high school   -4.8995     1.5703  -3.120 0.001892 **
lunchstandard                8.5995     1.0639   8.083 3.33e-15 ***
test.prepnone               -7.8647     1.0675  -7.367 5.58e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.72 on 620 degrees of freedom
Multiple R-squared:  0.233,     Adjusted R-squared:  0.2181
F-statistic: 15.69 on 12 and 620 DF,  p-value: < 2.2e-16
```

Hình 6.2.1

với RMSE được tính theo công thức ở Hình 8.1 là bằng 12.13063. Dưới đây là kết quả dự đoán trên tập train2:

test.prep <fctr>	m.score <int>	r.score <int>	w.score <int>	Avg.score <dbl>	pred <dbl>
none	38	60	50	49.33333	56.04979
completed	78	72	70	73.33333	72.35248
none	67	69	75	70.33333	72.35038
none	70	70	65	68.33333	61.81925
none	69	74	74	72.33333	72.53971
completed	77	69	68	71.33333	71.97195
completed	71	84	87	80.66667	64.99323
none	67	64	61	64.00000	61.08094
none	61	57	56	58.00000	53.38994
none	47	49	50	48.66667	69.71902

Hình 6.2.2: Kết quả dự đoán trên tập train2

Với model đã chọn ở trên em tiến hành train lại trên toàn bộ tập train

```
Call:
lm(formula = Avg.score ~ . - m.score - r.score - w.score, data = Student_train)

Residuals:
    Min      1Q  Median      3Q     Max
-48.017  -8.374   0.996   9.090  26.700

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 66.0788     2.0524  32.195  < 2e-16 ***
gendermale                  -2.8804     0.8947  -3.220 0.001337 **
racegroup B                  1.6121     1.8920   0.852 0.394435
racegroup C                  2.4482     1.7762   1.378 0.168508
racegroup D                  5.4653     1.8046   3.029 0.002538 **
racegroup E                  7.7683     2.0323   3.822 0.000143 ***
parent.edbachelor's degree   2.4484     1.6024   1.528 0.126914
parent.edhigh school        -5.4301     1.3962  -3.889 0.000109 ***
parent.edmaster's degree     3.9871     2.0395   1.955 0.050954 .
parent.edsome college       -0.8488     1.3303  -0.638 0.523612
parent.edsome high school   -4.1085     1.3884  -2.959 0.003178 **
lunchstandard                9.2015     0.9300   9.894  < 2e-16 ***
test.prepnone               -7.4020     0.9396  -7.878 1.11e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.57 on 783 degrees of freedom
Multiple R-squared:  0.2422,    Adjusted R-squared:  0.2306
F-statistic: 20.86 on 12 and 783 DF,  p-value: < 2.2e-16
```

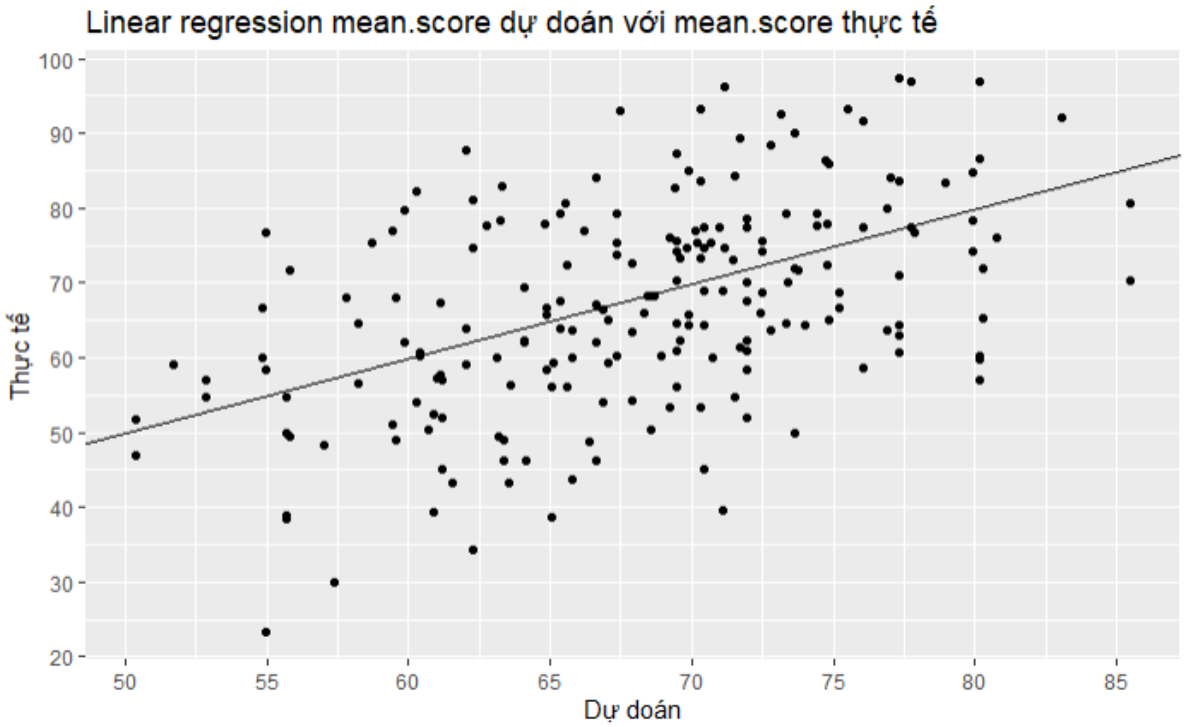
Hình 6.2.3

Sau đó tính lại RMSE nhận được giá trị bằng 12.23255. Và thu được kết quả dự đoán trên tập test như hình bên dưới:

test.prep <fctr>	m.score <int>	r.score <int>	w.score <int>	Avg.score <dbl>	pred <dbl>
none	38	60	50	49.33333	56.04979
completed	78	72	70	73.33333	72.35248
none	67	69	75	70.33333	72.35038
none	70	70	65	68.33333	61.81925
none	69	74	74	72.33333	72.53971
completed	77	69	68	71.33333	71.97195
completed	71	84	87	80.66667	64.99323
none	67	64	61	64.00000	61.08094
none	61	57	56	58.00000	53.38994
none	47	49	50	48.66667	69.71902

Hình 6.2.4: Kết quả dự đoán trên tập train2

Khi đã dự đoán được điểm trên tập test em vẽ được biểu đồ tương quan giữa điểm trung bình thực tế và dự đoán như hình bên dưới:



Hình 6.2.5: Biểu đồ tương quan giữa điểm trung bình thực tế và dự đoán

*Nhận xét: qua biểu đồ trên có thể thấy được sự phân bố của các điểm là khá xa so với đường thẳng nên có thể nói rằng mô hình dự đoán chưa được tốt.

6.3. Tìm biến nào ảnh hưởng đến điểm số nhất

Để tìm được biến nào ảnh hưởng nhất em sử dụng phương pháp đánh giá hệ số chuẩn hóa của model linear regression. Thì việc đánh giá dựa trên hệ số chuẩn hóa là để đưa các hệ số về cùng một thang đo để có thể so sánh với nhau. Việc tìm hệ số chuẩn hóa em sẽ sử dụng thư viện "lm.beta" để chuẩn hóa các hệ số này. Sau khi chuẩn hóa sẽ có được kết quả như hình sau:

```
Call:
lm(formula = Avg.score ~ . - m.score - r.score - w.score, data = Student_train)

Standardized Coefficients::
                (Intercept)                   gendermale                racegroup B                racegroup C
                 0.00000000                -0.10043847                 0.04412383                 0.08023247
                racegroup D                racegroup E parent.edbachelor's degree       parent.edhigh school
                 0.17014461                 0.18202110                 0.05464791                -0.14631696
  parent.edmaster's degree      parent.edsome college  parent.edsome high school              lunchstandard
                 0.06625711                -0.02438909                -0.11216556                 0.30867164
           test.prepnone
                -0.24734747
```

Hình 6.3.1: Kết quả sau khi chuẩn hóa hệ số.

Như ta thấy được kết quả trên hình thì hệ số của lunch standard là lớn nhất. Nên có thể kết luận được rằng biến lunch standard là biến có ảnh hưởng nhất đến điểm số ở trong tập dữ liệu này.

7. Kết luận (conclusion)

Phân tích EDA cho ta biết những yếu tố nào có điểm trung bình cao hơn. Sử dụng mô hình **Linear Regression** và **K-Nearest Neighbors** để tự đoán điểm trung bình. Kết quả cho ra với mô hình **K-Nearest Neighbors** có kết quả tốt

hơn (**K-Nearest Neighbors** có RMSE =12,20 , **Linear Regression** có RMSE = 12.23255).Biến có ảnh hưởng nhất đến điểm số ở trong tập dữ liệu nàylà biến lunch standard. Nếu có thời gian nhóm sẽ tìm hiểm làm cho RMSE ở các mô hình nhỏ hơn và R-square cao hơn để mức độ phù hợp của mô hình cao hơn so với ban đầu.

8. Phụ lục (appendices)

Dự đoán sử dụng thuật toán kNN:

- Chi tiết về các siêu tham số và độ đo :
  - k ở thuật toán kNN : ở đây em sử dụng thư viện caret với hai hàm train sẽ hỗ trợ tìm ra k tốt nhất.
  - Có sử dụng cross validation để chia dữ liệu, với k=10 (fold). Em có tham khảo một số trang trên mạng thì thấy k=10 là thường được sử dụng nhiều nhất và nhận được kết quả thực nghiệm tốt nhất.
  - Độ đo RMSE: RMSE là lỗi trung bình bình phương gốc (RMSE) là thước đo mức độ hiệu quả của mô hình. Thực hiện điều này bằng cách đo sự khác biệt giữa các giá trị dự đoán và giá trị thực tế. R-MSE càng nhỏ tức là sai số càng bé thì mức độ ước lượng cho thấy độ tin cậy của mô hình có thể đạt cao nhất.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Hình 8.1. Công thức tính RMSE

  - Độ đo R-square: Giá trị dao động từ 0 đến 1. R bình phương càng gần 1 thì mô hình đã xây dựng càng phù hợp và ngược lại.

$$R^2 = 1 - (ESS/TSS)$$

Hình 8.2. Công thức tính R-square

Chú thích:

+ Residual Sum of Squares (ESS): tổng các độ lệch bình phương phần dư.

- **Total Sum of Squares (TSS)**: tổng các độ lệch bình phương toàn bộ.

9. Đóng góp (contributions)

Thành viên	**Nhiệm vụ**	**Mức độ hoàn thành**

Trần Văn Duy	Phân tích EDA	100%
Lê Phương Nam	Xây dựng và đánh giá model	100%
Trần Công Trường	Phân tích EDA	100%
Cao Anh Văn	Xây dựng và đánh giá model	100%

Bảng 9.1 Phân công nhiệm vụ và mức độ hoàn thành.

Tên\Người đánh giá	**Trần Văn Duy**	**Lê Phương Nam**	**Trần Công Trường**	**Cao Anh Văn**
Trần Văn Duy		**10**	**10**	**10**
Lê Phương Nam	**10**		**10**	**10**
Trần Công Trường	**10**	**10**		**10**
Cao Anh Văn	**10**	**10**	**10**	

Bảng 9.2 Bảng đánh giá tinh thần nhóm các thành viên của mỗi người trên thang 10.

10. Tham khảo (references)

- Quách Đình Hoàng, slide, video bài giảng môn Lập trình R cho phân tích, đại học Sư phạm Kỹ thuật thành phố Hồ Chí Minh, năm 2021.
- Jim Frost. "Identifying the Most Important Independent Variables in Regression Models." *Statistics by Jim*, https://statisticsbyjim.com/regression/identifying-important-independent-variables/.
- Uyên Đặng. "[Machine Learning] Linear Regression và ứng dụng cho bài toán dự đoán điểm Nhập môn Lập trình – AI CLUB TUTORIALS." *AI CLUB TUTORIALS*, 24 April 2021, http://tutorials.aiclub.cs.uit.edu.vn/index.php/2021/04/24/linear-regression/.