**ANALYZING STUDENT ACHIEVEMENT IN SECONDARY EDUCATION**

**FUNDAMENTALS OF DATA ANALYTICS**

**CSIS 3360-001**

Submitted To: Project Team 4:

Samuel Otim Nicy Taneja-300297904

Instructor, Ankit Dahiya- 300297997

Faculty of CSIS Harjot Bawa- 300302572

**INTRODUCTION:**

The modern academic system is totally transformed in last 2 decades. The emerging technology is helping the students to learn the different fields. Apart from the technology there are a lot of social and economical factors that are affecting the study result. Analyzing student’s performance has become more challenging due to the large volume of data in educational databases. So, in our project we will be analyzing various factors to predict student’s performance.

**DATA SOURCE:**

In case of trying to get the data, we used many data sets that available over the internet by the different academic and paid institution. But finally, we find The dataset from UCI Machine Learning Repository has been used. The dataset is attached, and the link is https://archive.ics.uci.edu/ml/datasets/student+performance.

In this project we are using a sample of 649 students with their results based on the following parameters:

1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)   
2 sex - student's sex (binary: 'F' - female or 'M' - male)   
3 age - student's age (numeric: from 15 to 22)   
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)   
5 family size - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)   
6 Parents status - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)   
7 Mother education - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)   
8 Father education - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)   
9 Mother job - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other')   
10 Father job - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other')   
11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')   
12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')   
13 travel time - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)   
14 study time - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)   
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)   
16 schools up - extra educational support (binary: yes or no)   
17 family support - family educational support (binary: yes or no)   
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)   
19 activities - extra-curricular activities (binary: yes or no)   
20 nursery - attended nursery school (binary: yes or no)   
21 higher - wants to take higher education (binary: yes or no)   
22 internet - Internet access at home (binary: yes or no)   
23 romantic - with a romantic relationship (binary: yes or no)   
24 family relationship - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)   
25 free time - free time after school (numeric: from 1 - very low to 5 - very high)   
26 go out - going out with friends (numeric: from 1 - very low to 5 - very high)   
27 alcohol consumption - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)   
28 Weekend alcohol consumption - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)   
29 health - current health status (numeric: from 1 - very bad to 5 - very good)   
30 absences - number of school absences (numeric: from 0 to 93)   
  
# these grades are related with the course subject, Math or Portuguese:   
31 G1 - first period grade (numeric: from 0 to 20)   
31 G2 - second period grade (numeric: from 0 to 20)   
32 G3 - final grade (numeric: from 0 to 20, output target)

**Hardware and Software used**

Operating System - Microsoft windows 8.1 and above.

Ram used – minimum of 4 gigabyte DDR1

Rom Used- minimum of 160 gigabytes.

Software requires -R Studio, Microsoft office

Technologies - R- programming,

**ANALYSIS TECHNIQUES TO BE USED:**

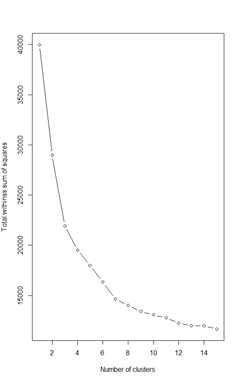
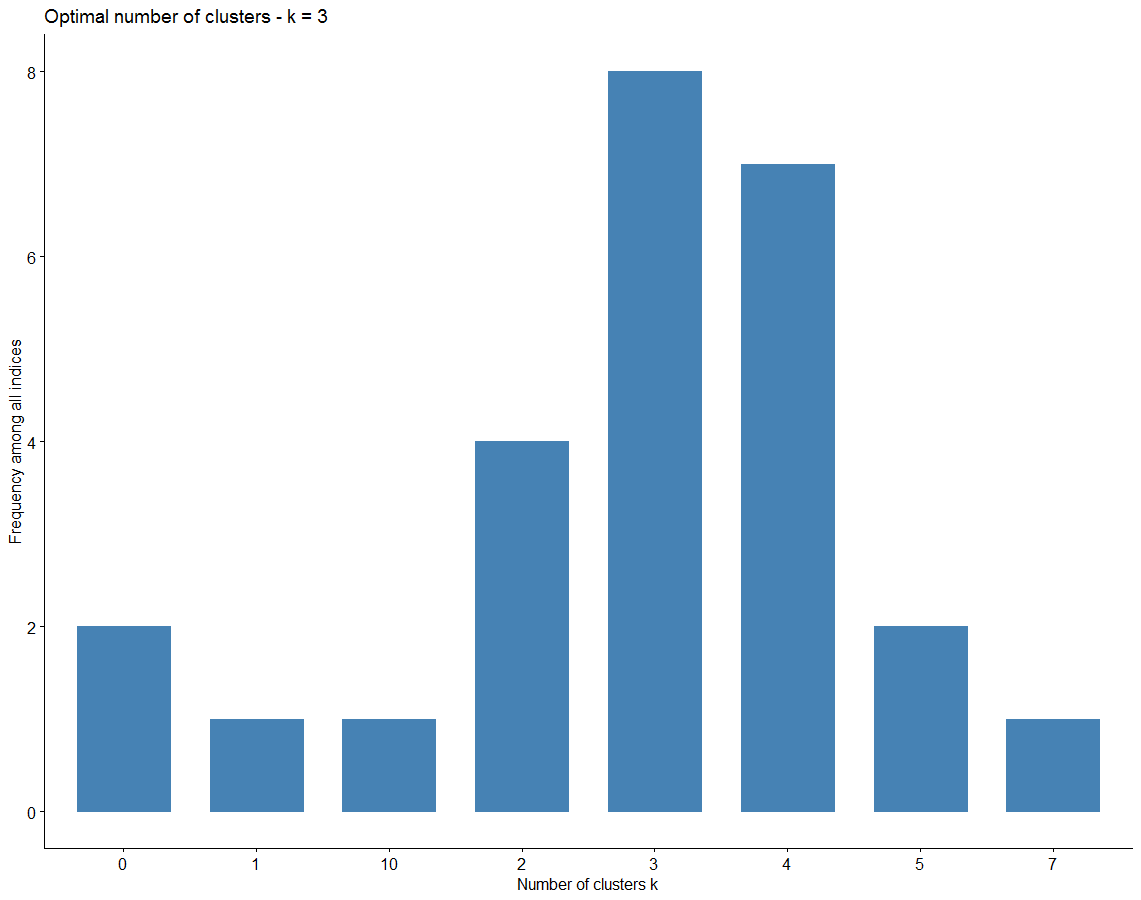
In this project we will divide the data in some groups using the “kmeans” and try to perform the regression analysis based on these factors. Using regression analysis, the relationship between dependent and independent variables will be estimated. The first period and the second period grade (G1 and G2) will be predicted based on the parameters such as father’s and mother’s profile, activities, age and health conditions. The third period grade G3 will be predicted based on G1 and G2 as G3 has a strong correlation with them. This method will help us to predict the student’s performance based on the variables chosen. This can help the instructors and the family to work more on their child and take proactive measures to improve the grades and understanding of the subject for better performance.

**ANALYSIS:**

1. **KMeans Clustering:**

**Results:**

The elbow method was used. The curve was bending at 3. The optimal number of clusters came out to be 3.

****

Among all indices:

===================

\* 2 proposed 0 as the best number of clusters

\* 1 proposed 1 as the best number of clusters

\* 4 proposed 2 as the best number of clusters

\* 8 proposed 3 as the best number of clusters

\* 7 proposed 4 as the best number of clusters

\* 2 proposed 5 as the best number of clusters

\* 1 proposed 7 as the best number of clusters

\* 1 proposed 10 as the best number of clusters

Conclusion

=========================

\* According to the majority rule, the best number of clusters are 3.

1. **Linear Regression and Prediction:**

**Step 1.**

Data Preparation

In the data preparation firstly, we clean the data by removing the existing outliers and null values in the data set.

We divide the data set into the 6.5 vs 4.5. The first set we used for the training of the analysis to perform the linear regression while the second data set we used to verify the analysis that we performed.

**Step 2.**

Data Processing.

In the Data preparation we convert the categorical data into the binary data in the form of zero and one. If there is more than one category, we convert the data with respect to the one category.

Linear Regression is a linear approach to model the relationship between the independent input variable and dependent output variable. In our model, the following categorical variables were made as dummy variables: sex, address, guardian, school support, family support, paid, activities, nursery, higher and internet.

Linear Regression on G3:

**Step 3.**

Liner regreesion

lm(formula = G3 ~ ., data = finalnumericdata)

Residuals:

Min 1Q Median 3Q Max

-8.8372 -0.5133 -0.0169 0.6064 5.6715

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.0006693 0.9259720 -0.001 0.999423

age 0.0222177 0.0474924 0.468 0.640082

Pstatus -0.0866044 0.1601915 -0.541 0.588957

Medu -0.0373899 0.0611778 -0.611 0.541313

Fedu 0.0304015 0.0598311 0.508 0.611549

traveltime 0.1241927 0.0737015 1.685 0.092479 .

studytime 0.0509324 0.0653211 0.780 0.435852

failures -0.2447822 0.0982942 -2.490 0.013025 \*

famrel -0.0236495 0.0541330 -0.437 0.662353

freetime -0.0418228 0.0521594 -0.802 0.422961

goout -0.0241505 0.0495568 -0.487 0.626197

Dalc -0.0699276 0.0707945 -0.988 0.323659

Walc -0.0064709 0.0546809 -0.118 0.905836

health -0.0454001 0.0355448 -1.277 0.201988

absences 0.0171964 0.0113831 1.511 0.131377

G1 0.1326353 0.0371462 3.571 0.000384 \*\*\*

G2 0.8747815 0.0347310 25.187 < 2e-16 \*\*\*

internetyes 0.1174885 0.1256736 0.935 0.350220

higheryes 0.2003229 0.1814631 1.104 0.270052

nurseryyes -0.0732449 0.1267589 -0.578 0.563590

activitiesyes 0.0282219 0.1035868 0.272 0.785369

paidyes -0.1831286 0.2147892 -0.853 0.394213

familysupport 0.1239321 0.1062758 1.166 0.244009

schoolsupport -0.1501707 0.1686312 -0.891 0.373529

guardianother 0.2403286 0.2250423 1.068 0.285971

reasonrepo -0.1620956 0.1367574 -1.185 0.236363

reasonother -0.4119268 0.1694597 -2.431 0.015348 \*

resonhome -0.0902566 0.1327717 -0.680 0.496893

famlsizedum 0.0348521 0.1140587 0.306 0.760040

address 0.1826722 0.1181534 1.546 0.122602

sexdum -0.1004888 0.1154785 -0.870 0.384532

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.25 on 618 degrees of freedom

Multiple R-squared: 0.8573, Adjusted R-squared: 0.8504

F-statistic: 123.8 on 30 and 618 DF, p-value: < 2.2e-16

The given model having a p value is less than 0.05. Which is efficient to accept the given model. When we are performing the regression we find out that the final grade is mostly dependent on the grade G1, and The grade G2.

Linear Regression on G1:

lm(formula = G1 ~ ., data = dataforg1)

Residuals:

Min 1Q Median 3Q Max

-11.0106 -1.4372 -0.1118 1.5210 6.5992

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 10.23687 1.66000 6.167 1.26e-09 \*\*\*

age -0.08647 0.08708 -0.993 0.321153

Pstatus 0.29655 0.29625 1.001 0.317216

Medu 0.22609 0.11286 2.003 0.045580 \*

Fedu 0.16735 0.11048 1.515 0.130340

traveltime -0.09063 0.13638 -0.665 0.506600

studytime 0.43227 0.11966 3.612 0.000328 \*\*\*

failures -0.99602 0.17602 -5.658 2.34e-08 \*\*\*

famrel 0.04297 0.09973 0.431 0.666676

freetime -0.08150 0.09648 -0.845 0.398564

goout -0.01733 0.09170 -0.189 0.850182

Dalc -0.20593 0.13079 -1.574 0.115885

Walc -0.05526 0.10118 -0.546 0.585148

health -0.02522 0.06558 -0.385 0.700634

absences -0.04879 0.02098 -2.325 0.020370 \*

internetyes 0.27715 0.23236 1.193 0.233407

higheryes 1.57908 0.32949 4.792 2.06e-06 \*\*\*

nurseryyes -0.22036 0.23449 -0.940 0.347709

activitiesyes 0.23855 0.19151 1.246 0.213387

paidyes -0.61172 0.39649 -1.543 0.123374

familysupport -0.13841 0.19666 -0.704 0.481817

schoolsupport -1.06299 0.30917 -3.438 0.000625 \*\*\*

guardianother 0.16987 0.41654 0.408 0.683542

reasonrepo 0.56512 0.25206 2.242 0.025311 \*

reasonother -0.14877 0.31347 -0.475 0.635253

resonhome 0.34646 0.24532 1.412 0.158358

famlsizedum 0.25399 0.21090 1.204 0.228923

address 0.45127 0.21788 2.071 0.038754 \*

sexdum -0.35336 0.21330 -1.657 0.098106 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.313 on 620 degrees of freedom

Multiple R-squared: 0.3207, Adjusted R-squared: 0.29

F-statistic: 10.45 on 28 and 620 DF, p-value: < 2.2e-16

Linear Regression on G2:

Call:

lm(formula = G2 ~ ., data = dataforg2)

Residuals:

Min 1Q Median 3Q Max

-13.4327 -1.3762 -0.0357 1.5310 6.8427

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.06365 1.77543 3.979 7.75e-05 \*\*\*

age 0.10486 0.09314 1.126 0.26069

Pstatus 0.34900 0.31685 1.101 0.27113

Medu 0.23577 0.12071 1.953 0.05124 .

Fedu 0.20864 0.11816 1.766 0.07794 .

traveltime -0.09549 0.14587 -0.655 0.51296

studytime 0.35551 0.12798 2.778 0.00564 \*\*

failures -1.22158 0.18826 -6.489 1.77e-10 \*\*\*

famrel 0.18821 0.10666 1.765 0.07814 .

freetime -0.10517 0.10319 -1.019 0.30851

goout -0.05816 0.09807 -0.593 0.55339

Dalc -0.19137 0.13989 -1.368 0.17180

Walc -0.08373 0.10821 -0.774 0.43936

health -0.10469 0.07014 -1.493 0.13603

absences -0.04025 0.02244 -1.794 0.07335 .

internetyes 0.29509 0.24851 1.187 0.23552

higheryes 1.62223 0.35240 4.603 5.05e-06 \*\*\*

nurseryyes -0.17777 0.25079 -0.709 0.47870

activitiesyes 0.18123 0.20483 0.885 0.37661

paidyes -0.26265 0.42406 -0.619 0.53590

familysupport -0.11873 0.21034 -0.564 0.57263

schoolsupport -0.82228 0.33067 -2.487 0.01315 \*

guardianother 0.18676 0.44550 0.419 0.67521

reasonrepo 0.59612 0.26958 2.211 0.02738 \*

reasonother -0.29224 0.33527 -0.872 0.38373

resonhome 0.20446 0.26237 0.779 0.43613

famlsizedum 0.22010 0.22557 0.976 0.32955

address 0.49298 0.23303 2.116 0.03478 \*

sexdum -0.32175 0.22813 -1.410 0.15894

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.474 on 620 degrees of freedom

Multiple R-squared: 0.3101, Adjusted R-squared: 0.279

F-statistic: 9.954 on 28 and 620 DF, p-value: < 2.2e-16

We can utilize the model because in both cases the value of the p is less than 0.05.

In the both the case there are four major reasons that are affecting the results of the G1 and G2 grades. Whether a student wants to study in higher classes, he or she failed in a class or not, how much time a student spent on studies in a week and last one whether a student has a school support or not.

Predict grade 1 when a male student wants to do higher studies, studies for 8 hours, has school support, has never failed, has internet at home, lives in urban area and consumes very less alcohol daily, mother has secondary education.

**Accuracy of the liner regression using step model for grade one**

accuracy(pridictg1,testforg1$G1)

ME RMSE MAE MPE MAPE

Test set 0.0599252 3.171156 2.50191 -Inf Inf

**Accuracy of the liner regression using step model for grade Two**

accuracy(pridictg2,testforg2$G2)

ME RMSE MAE MPE MAPE

Test set -0.03640495 2.758307 2.017264 -Inf Inf

**Prediction Actual and residual for both G2 and G1 using step model**



predict\_df <-data.frame(higheryes=1,studytime=3,schoolsupport=1,failures=0,Medu=3, Dalc=1, address=1, internetyes=1, sexdum=1)

> #predict

> predit\_out <-predict(stepg1, predict\_df, interval="prediction", level=0.95)

> #print output

> print(predit\_out)

fit lwr upr

12.71672 8.248602 17.18484

#Predict grade 1 when a female student wants to do higher studies, studies for 12 hours, has school support, has never failed, has internet at home, lives in urban area and consumes very less alcohol daily, mother has higher education.

predict\_df <-data. frame(higheryes=1,studytime=4,schoolsupport=1, failures=0,Medu=4, Dalc=1, address=1, internetyes=1, sexdum=0)

#predict

predit\_out <-predict (stepg1, predict\_df, interval="prediction", level=0.95)

#print output

print(predit\_out)

fit lwr upr

13.88117 9.399138 18.36321

pridictg12 = data.frame(higheryes=0,studytime=3,schoolsupport=1,failures=0,Medu=3, Dalc=1, address=1, internetyes=1, sexdum=1)

outg12 =predict(stepg1, pridictg12, interval="prediction", level=0.95)

print(outg12)

while we compare a student for grade g1 that don’t want to go for the higher studies in compare to a student who wants to go for the higher studies

fit lwr upr

1 11.2759 6.742857 15.80895



**Conclusions:**

With the help of the regression and using step model analysis we find some critical factors that are impacting the grades of the student.

1. Whether the student want to go for the higher studies or not.
2. The number of the hours a student study per week.
3. A student fails in last classes or not.
4. A student has a school support or not.

**References:**

[1] https://archive.ics.uci.edu/ml/datasets/student+performance

[2] https://www.ece.uvic.ca/~rexlei86/SPP/otherswork/1-s2.0-S1877050915036182-main.pdf

[3] https://www.ece.uvic.ca/~rexlei86/SPP/otherswork/1-s2.0-S0360131512002102-main.pdf