Southeast Europe Journal of Soft Computing ISSN 2233 – 1859
Available online: http://scjournal.ius.edu.ba

Vol 11, No. 1, March 2022, pp. 19-27

THE EFFECTS OF VIDEO GAMES ON SCHOOL ACHIEVEMENT IN PRIMARY EDUCATION BASED ON SARAJEVO CITY: A DATA SCIENCE CASE STUDY

[1*] Adnan Bahtić, [1,2]Ali Abd Almisreb, [2]Mohammed A. Saleh, [3]Musab A. M. Ali

[1]Faculty of Engineering and Natural Sciences,
International University of Sarajevo, Hrasnicka Cesta 15, Ilidža 71210 Sarajevo,
Bosnia and Herzegovina
[2]Department of Cybersecurity, IITU, Almaty 050000, Kazakhstan
[3]Faculty of Engineering, Halic University, Istanbul, Turkey

**Corresponding Author*: adnanbahtic@hotmail.com

Article Info

Article history:

Article received on 25 February 2022

Received in revised form 5 March 2022

Keywords:

Video games; covid-19; education; data science;

ABSTRACT: Video games are in recent years a big part of our daily life, especially for young people, and for most of them they have a very important role in their life. The video game industry and video games development are growing every day, and especially in the last period with Covid-19 pandemic and lockdown lot of people find comfort and spend their free time playing games, most of the consumers of the video games are young people or to be precise kids. The Idea is to show and investigate the linking effects of video games to school achievement in primary education. Or to be precise by data analysis and data mining the aim is to investigate and show the results of the analysis. that are collected. Datapoints are collected from several elementary schools that are participating in this investigation. Data exploration and analysis consisted of exploring most important features, relations between different features in order to better understand the data that we are dealing with. All analyses in this thesis are done in R programing language and RStudio as IDE.

1. INTRODUCTION

Today's is largely composed of technology. In a relatively short span of time world has been immersed with high-definition television, Facebook, YouTube, internet radio," green" cars, outrageous thrill rides, 3-D technology, etc. But no area of technology has become as prominent as that of video gaming [1]. Considering the popularity of video games among young people this research about effects that children have in their primary education with playing video games sounds interesting. Most people (assuming younger generation) played video game in same shape or form, either by Mobile phone, PC or Console. Then can we ask our self what it is enough time for children (if they have from 8-14 years) time spend on video games, are they having positive effects, are they having negative effects, does this affect their grades, education, and overall learning process? in discovering the complete bacterial diversity that exists in a complex environment. In this project, R language is used, which is a programing language and environment for statistical computing and graphics [2] that are mostly used for data analysis, machine and deep learning and data mining. All these techniques and tools are used for this data analysis project. Data analytics is the science of analyzing raw data to make conclusions about that information. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption [3].

2. METHODOLOGY

2.1. Data collection

One of most important roles for every data analysis and data mining project is data gathering. When gathering information data is not necessarily in the correct format for the model execution. Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes [4]. Data for this study were collected from elementary schools in Sarajevo through a survey created in Google Forms. The questions were asked to the participants are aimed to collect a group of details as listed in Table 1. 238 pupils participated in study and after cleaning the data, left 224 participants in which 115 (52%) of them are males and 109 (48%) females. Table 2 shows more details about the participants. The answer is rewritten in the R, from Bosnian language to the English for better understanding data and the results. Therefore, some of the results are written in short form.

Table 1 Information about participants

Name	Summary
Gender	Gender of children (Male or Female)
Age	Age of the participant
Grade	Which grade are they currently in
GLS	Final grade in last semester
GDS	Final grade in this semester
PVG	Are they playing video games
DIW	How many days in week do they play video games
HPD	How many hours do they play video games per day
Scale	On scale from 1 to 10 hom much do they love video games
PT	Do they want to play more games thes usual
OPN_TPG	In there opinion - are they playing games to much
OPN_AYG	In there opinion – are games affecting their grades
PP	Are they ever grounded by ther parents

Grade	Male	Female	Total
5	28 (48%)	30 (52%)	58
6	**35 (71%)**	14 (29%)	49
7	29 (47%)	32 (53%)	**61**
8	23 (42%)	**33 (58%)**	56

Table 2: Ratio of boys and girls in grades

Table 2 illustrates that in almost every group we have similar number of male and female participants, except in group that goes to the 6th grades, we see smaller number of female pupils are participating this research.

As presented in Figure 1, the histogram shows in what columns do we have missing values and data and what is the percentage of them.

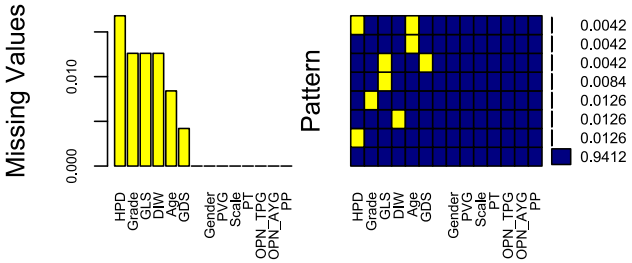


Figure 1: Histogram of missing data and patterns

With inform which data is missing, for this analysis it would be removed entirely.

2.2 Data exploration

One of the most important parts of every data analysis project is data exploration. With visual data exploration, we can find out that Data visualization is a critical tool in the data analysis process. Visualization tasks can range from generating fundamental distribution plots to understanding the interplay of complex influential variables in machine learning algorithms [5].

2.2.1 Density plot

"The peaks of the density plot are at the locations where there is the highest concentration of points "[6]. The simplest explanation of what is the density plot is show in the line above, what is interesting here is which values are we using for density plot. This observation is done with the two main rows in this data set, first one is the grade in last semester, and the other one is the grades in this semester. Figures 2 and 3 show the where are the highest concentration of the student grades and the mean which is represented in blue dashed line. Formula for mean can be see below.

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \quad (1)$$

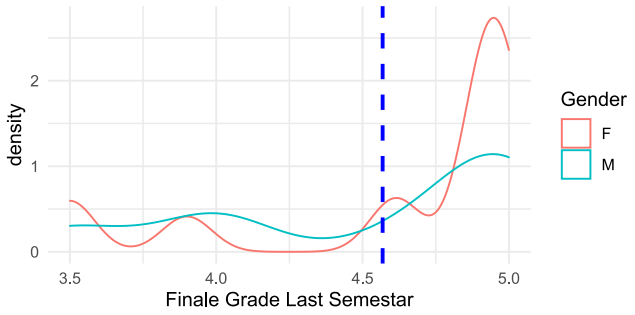


Figure 2: Density plot for Final grades last semester

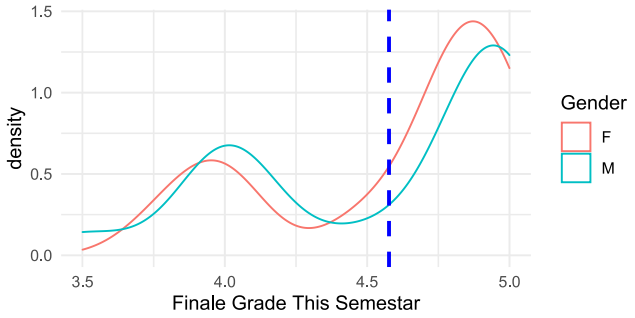


Figure 3: Density plot for Final grades this semester

2.2.2 Two value sample test

before starting with two value sample test, short explanation of what is it is necessary. Two-Sample t-Test is a method used to test whether the unknown population means of two groups are equal or not [7] and it is done by the formula:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \quad (2)$$

For Two-Sample t-Test formula which is shown above, there are $\bar{y}_1, \bar{y}_2$ which represent Means of sample size, n_1, n_2 that are sample size for this observation and S pooled standard deviation.

P values are statistical hypothesis testing based on a critical level of significance is a dichotomous test [8]. To be precise the null hypothesis is rejected when p < .05 and not rejected when p > .05, or in other words the p is significant when p < .05 and not significant when p > .05.

Table 3 Two-Sample t-test

Type	Male Mean	Female Mean	T	df	P-value	95 percent confidence interval	
Age	12.4	12.3	-0.082	196.24	0.9345	-0.3545	0.3261
GLS	4.48	4.55	0.941	221.99	0.3475	-0.0788	0.2230
GDS	4.57	4.55	-0.299	211.44	0.7649	-0.1425	0.1049
DIW	3.81	2.65	-3.094	221.95	0.0001	-1.7545	-0.577
HPD	2.31	0.99	-6.618	192.69	3.514e-	-1.7162	-0.928
Scale	7.20	4.64	-7.010	220.67	2.851e-	-3.2768	-1.838

Table 3 provides an overview for the two-sample t-test that is done for all continuous variables in this dataset. Before getting to this part, here is just short explanation of this parameters and how they are represented in Table 3:

- t represents the statistic value in t-test.
- df represent the degrees of freedom.
- P-value we already mention that show is their significance in these values.
- Confidence interval 95 show the interval mean at 95%

Table 4 represent p-values for continuous values in this data set.

Parameters: Age, Grade last semester (GDS), Grade this semester, Days in week (DIW) and Hours per day (HPD).

Table 4 P-Values for Age, GDS, GLS, DIW, HPD

Female				
Type	**P**	**p.adj**	**p.format**	**p.signif**
Age	0.965	1.00	0.96	ns
GDS	0.856	1.00	0.86	ns
GLS	0.575	1.00	0.57	ns
DIW	0.021	0.04	0.02	*
HPD	0.001	0.0003	1.5e-05	***
Male				
Age	0.958	1.00	0.96	ns
GDS	0.879	1.00	0.87	ns
GLS	0.600	1.00	0.60	ns
DIW	0.033	0.04	0.03	*
HPD	0.00158	0.0016	0.0016	**

Table 4 has some new features that needed to be explained and explanation can be seen bellow.

- P represent the p- values.
- P.adj represent the adjusted p-value.
- P.format represent the formatted p-value.
- P.signif represent the significance level in this observation.*

After results for two sample t-test and p values are presented, the next step is to present the results for GLS of two genders in Figure 4. GDS in Figure 5, DIW in Figure 6. HPD in Figure 7.

In Figure 4,5,6, and 7 T-test are done for 4 different categorical variables. Every Figure contains two plots, and in each of them box plot is represented. On the left side of plots, t-test is done for all 4 categories, which show the dots (which are representing the Female and male students). Each of box plot contains the line (in the middle of the box) that is representing the mean for this observation and line (above and below) that is showing us the 95% interval for both male and female.

On the other hand, on right side, p-values that is represented for all 4 categorical variables, or to be precise this plot is just showing the box plot of exactly the same as for the t-test with just one slightly different detail, on top of every box is represented what is the p value for this category, for example in Figure 3.9 both box on top have “ns” which indicates that GLS is not significant in this data. In example HDP, there is a one star of significance.

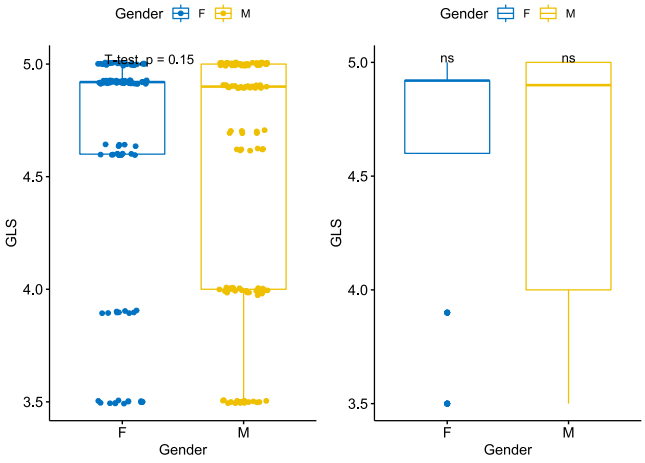


Figure 4: T-test for GLS

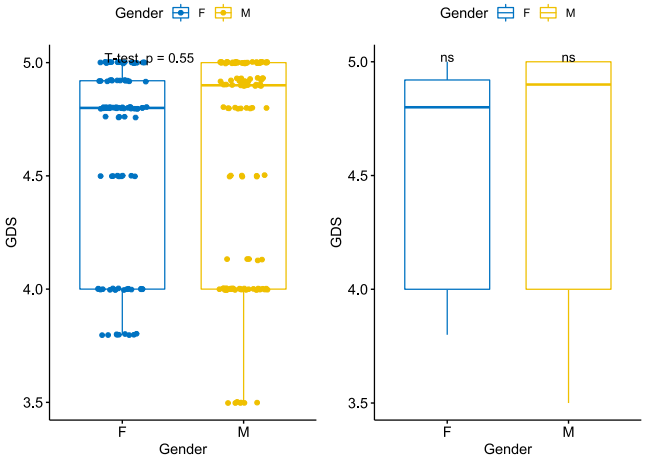


Figure 5: T-test for GDS

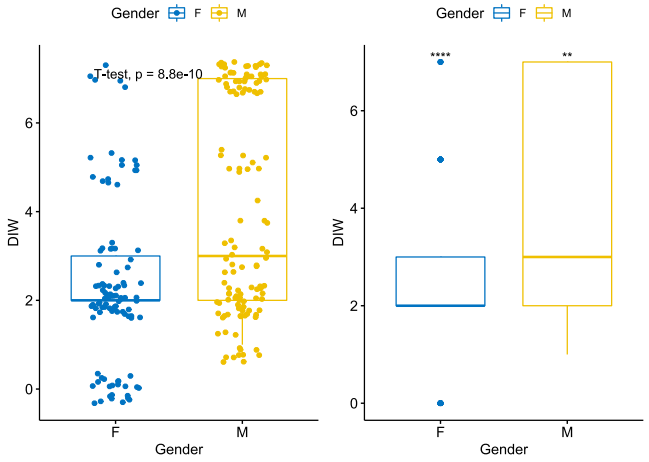


Figure 6: T-test for DIW

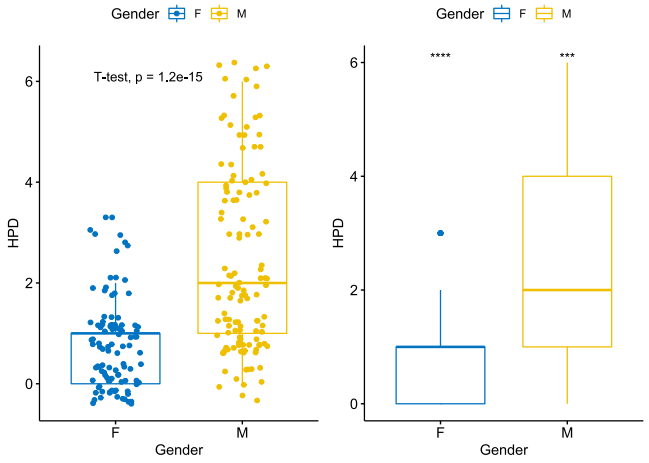


Figure 7: T-test for HPD

2.3 Correlation and collinearity analysis

Continuous variables are used mostly for creating observation and plot. Same process will be also done here where talk will be about correlation and collinearity analysis here, first indication is what is correlation between different variables. Simply, correlation is a statistical expression to two variables and are they linearly related.

After knowing that, this is how the correlation looks like:

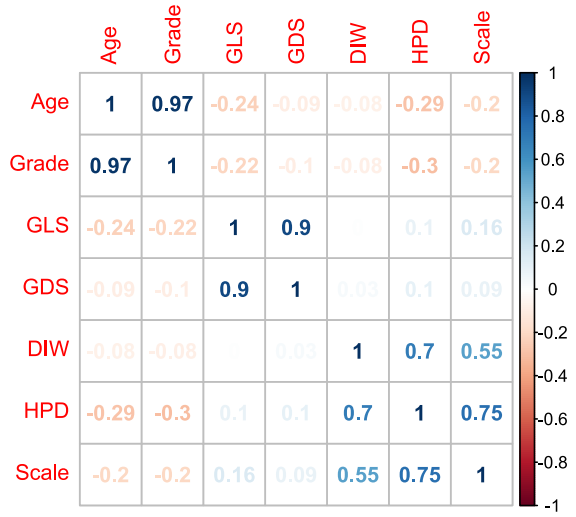


Figure 8 Correlation

From Figure 8 The results of the correlational analysis are shown. Like mention above the continues variables in this dataset, what can be addressed is that the results that can be seen. Correlation can from -1.00 to 1.00 or to be precise negative and positive correlation. In Figure 8 the positive correlation is shown by the blue color while the negative correlation is shown by the red color. What stands out in this figure is positive correlation in this plot is between the Age and grade. On the other hand, the figure shows that the lowest negative correlation is between grade and HPD.

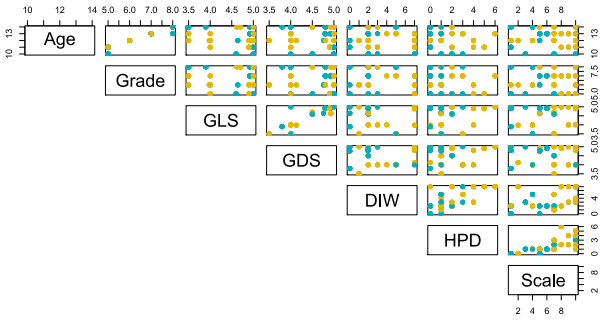


Figure 9 pair plot

Figure 9 represents correlation between all the values that are used in this observation. This is just representing all the variables in this dataset and how there are presented between al columns in this dataset.

This test will use all continuous variables in data set. Using R-Studio library called “psych” for visualization and representation of correlation between variables, this plot will show if there are correlation (which can be between -1 and 1) and represent the scatter plot of all values with the critical points, and regression line.

In the Figure 10 correlation is presented.

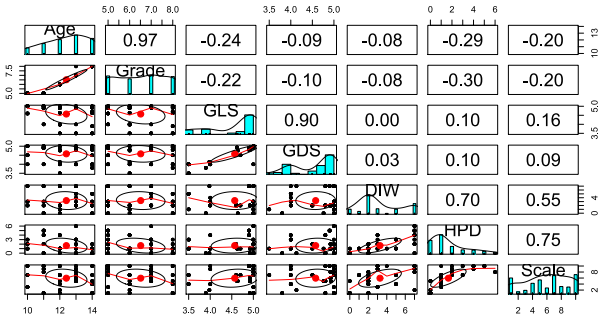


Figure 10: “Psych” Correlation

2.4 Decision tree

One of most popular algorithms that can be used in data mining and data analysis is Decision tree. Decision tree is mostly known for easy implementation and user friendly. Decision tree is a type of supervised learning algorithm that can be used in both regression and classification problems. It works for both categorical and continuous input and output variables. For this example, decision tree while have gender as the main category with Age, GLS, GDS, DIW, HPD are used as inputs for creating this decision tree.

Results:

```
        Conditional inference tree with 5 terminal nodes

Response:  Gender
Inputs:  Age, GLS, GDS, DIW, HPD
Number of observations:  224

1) HPD <= 3; criterion = 1, statistic = 36.028
  2) HPD <= 0; criterion = 0.981, statistic = 8.401
    3) DIW <= 4; criterion = 0.994, statistic = 10.542
      4)*  weights = 50
    3) DIW > 4
      5)*  weights = 7
  2) HPD > 0
    6) Age <= 10; criterion = 0.998, statistic = 12.734
      7)*  weights = 22
    6) Age > 10
      8)*  weights = 108
1) HPD > 3
  9)*  weights = 37
```

Figure 11: Decision tree results

Results can be better represented with the decision tree created bellow; the visual illustration of the tree is given in Figure 12.

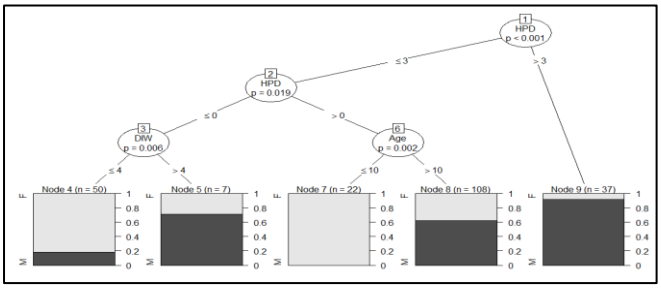


Figure 12: Decision tree

Closer inspection of the Figure 12 show interesting results. First thing that can be see is Hours per day in node 1, that show if greater than 3 hours per day, sample size is n=37. Next what is that p in this example is significant, or to be precise p<0.001.

In node 3 is p=0.006 and that with node 4 (n=50) which suggest that DIW is less or equal to four and on opposite side node 5 (n= 7) is bigger to 4 DIW.

And node 8 represent Age. Giving that node 6 show the p value is equal to 0.002 and that in node 7 n is 22, while in node 8 n is much larger, representing 108 children from data.

2.5 Linear regression

Regression analysis is a statistical technique for estimating the relationship among variables which have reason and result relation [9]. Different multiple models of linear regression can be crated. First thing needed to be check if data is ready for observation, so the first part is to create a Training and Testing set. As the names say we will use one part of date for training our Model and one for testing

After creating data train and data test samples, first step is creating 2 different models of multiple linear regression and they are:

1) For GLS for columns that are significant

2) For GDS for columns that are significant

Now after listing all models that will be used, next process is start working with the models. Before starting models, short list of explanation for tables that will be used for all models. Table 5 implies how regression model is done with and results show:
- Residuals which represent the error.
- Coefficients that show interceptor of the value and the slop.
- Std error that is the just standard deviation of distribution.
- t measure how many standard deviations in our estimation coefficients are far away from 0.
- Pr (> | t |) showing the p values and significant.

After that for every model you will see how prediction test is done, followed by the table that will consist of:
SSE (Sum of Errors) – Formula:

$$SSE = \sum (\hat{y}_{i-} y_i)^2 \quad (3)$$

In SSE $\hat{y}_i$ represent predicted value of y for observation i and y_i value for observation i
SST (Sum of Squares Total) - Formula:

$$SST = \sum (y_{i-} y)^2 \quad (4)$$

R2 (R squared) – Formula:

$$R2 = SSR/SST \quad (5)$$

In R2 formula we have two values, first one is presenting the Sum of square roots and other is showing us the sum of squares in total
RMSE (Root-Mean-Square Error)

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_{i-} y_i)^2}{n}} \quad (6)$$

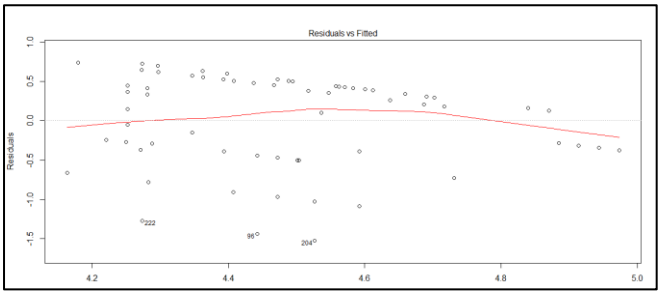
Just like in the first two equations the $\hat{y}_i$ represent predicted value of y for observation *i* and y_i value for observation *i*, with *n* showing the number of observations.

2.5.1 Model 1

First test for the model is done for GLS. This model will contain all the values that are obtained in this data set, first what is represented is models are coefficients in this observation.

Table 5 GLS Linear regression model

	Estimate	**St.error**	**T Value**	**Pr(>|t|)**	
Intercept	6.39151	0.86648	7.376	1.05e-1	***
Gender M	-0.1198	0.12374	-0.968	0.3344	
Age	-0.211	0.16877	-1.256	0.2111	
Grade	0.116	0.19257	0.605	0.5459	
PVG No	-0.0850	0.18545	-0.459	0.6472	
DIW	-0.0295	0.02680	-1.10	0.2720	
HPD	0.07429	0.03781	1.965	0.0513	.




```
Residuals:
    Min      1Q  Median      3Q     Max
-1.5266 -0.3710  0.1981  0.4391  0.7413
```

Figure 13 Model 1 Residuals

Table 6 GLS Linear regression model table

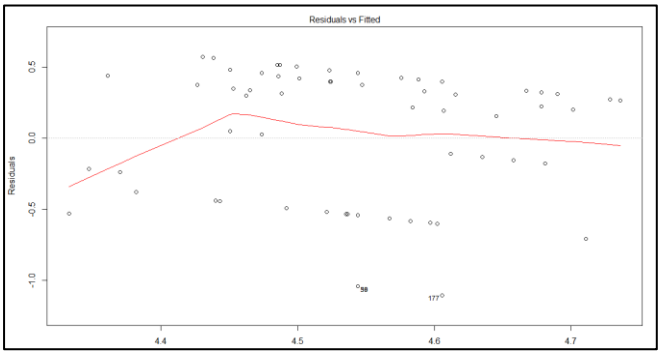
Type	Results
SSE	45.88492
SST	20.50675
RMSE	0.542
Residual Standard error	0.5549 on 149 degrees of freedom
F-Statistic	3.728 on 6 and 149 DF
Multiple R-squared	0.1305
Adjusted R-Squared	0.0955

2.5.2 Model 2

This model will represent the values that would be significant in observation.

Table 7 GDS Linear regression model

	Estimate	St.error	T Value	Pr(>|t|)	
Intercept	4.66623	0.74044	6.302	3.14e-	***
Gender M	-0.0613	0.10574	-0.580	0.5625	
Age	0.01904	0.14422	0.1442	0.8952	
Grade	-0.0542	0.16455	-0.330	0.7421	
PVG No	-0.1659	0.15847	-1.047	0.2966	
DIW	-0.0229	0.02290	-1.003	0.3176	
HPD	0.05921	0.03231	1.832	0.0689	.




```
Residuals:
    Min      1Q  Median      3Q     Max
-1.1054 -0.4553  0.2217  0.3786  0.5697
```

Figure 14 Model 2 Residuals

Table 8 GDS Linear regression model table

Type	Results
SSE	33.5065
SST	13.600071
RMSE	0.4634
Residual Standard error	0.4742 on 149 degrees of freedom
F-Statistic	1.261 on 6 and 149 DF
Multiple R-squared	0.04831
Adjusted R-Squared	0.09984

2.6 Random Forest algorithm

The random forest algorithm works by aggregating the predictions made by multiple decision trees of varying depth. Every decision tree in the forest is trained on a subset of the dataset called the bootstrapped dataset. This will be done with initial data, data that is not cleaned and have missing values. Creating a model that would predict genders by using the Random Forest algorithm.

First process is to create input function called rfInput. It is illustrated in Figure 15.

```
ntree      OOB      1      2
  300:   0.00%  0.00%  0.00%
ntree      OOB      1      2
  300:   0.00%  0.00%  0.00%
ntree      OOB      1      2
  300:   0.00%  0.00%  0.00%
ntree      OOB      1      2
  300:   0.00%  0.00%  0.00%
ntree      OOB      1      2
  300:   0.00%  0.00%  0.00%
ntree      OOB      1      2
  300:   0.00%  0.00%  0.00%
```

Figure 15 Rfinput

6 number of iterations are created on this test whereby the with 300 number of trees in every iteration. OOB represent Out-Of-Bag error ration that in this example is approximate goes here 0% as we can see. Which says that estimation is promising. Next, RF algorithm can be created as in Figure 16.

```
Call:
 randomForest(formula = Gender ~ ., data = NewData, proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 0%
Confusion matrix:
    F   M class.error
F 106   0           0
M   0 132           0
```

Figure 16 Rf algorithm

From Figure 16 can be seen that from this output of random forest algorithm with number of trees 500 that estimation (OOB) is at 0. Usually, the Estimation error rate algorithm, as in Figure 17, is used to show how the error is changed at every tree but in this example is shown that there is not any significant error from results above, but after creating plot, what can be seen is a small error ration in approximately first 50 trees iteration.

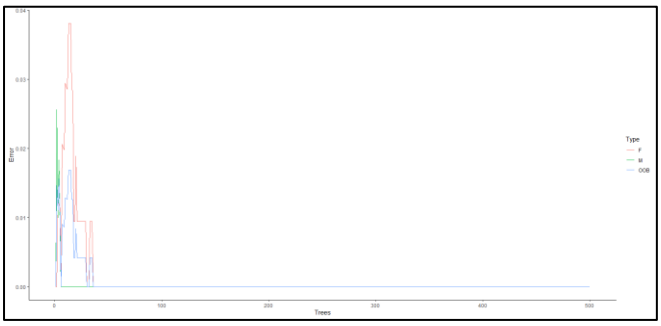


Figure 17 OOB estimation plot

3. CONCLUSION

Data analysis that is provided in this article give some interesting results. Briefly the summery and results will be represented in conclusion part. Fist starting with information about children, whereby observing the data, more male students participate in this project then female. Table 1 present the summary of all values that are used in analysis, while table 2 represent which grade children's go which participate in study. Data collection was the trickiest part, with all restriction, collecting data was challenge and in figure 1, the missing values and histogram of missing patterns are created. Regarding density plot between grades last semester and the grades this semester the critical point is bigger for the male students this semester. In part where the Two-sample t-test is done in table 3 and 4 with showing interesting thigs, first is that p-value are significant in part of HPD and students scale for they interest in video games, followed by the 95 precent confidences interval that give us the slightly accurate information for mean of all continuous variables. Also, the plots that are given in Figure 4 to 7 are plot of t-test which indicates with visual representation of boxplot that help us first to see how variables are shown in this data and is there p-values. The correlation part is just showing is their correlation between values in this data, there are plots that show and indicate where the correlation is positive and where is negative. With grate library from R called "psych". After that three machine learning algorithms are created. Decision tree, linear regression, and Random Forest algorithm. Decision tree part show us the how Gender is affected by the 5 categories that are chosen and firs thing that is show is conditional inference tree and then decision tree with nods representing Male and Female students. Two model of linear regression are done in this analysis, one is done for GLS, and one is done for GDS, in which not all values are used And last part that is done is Random Forest algorithm, that with use of algorithm that is made by multiple decision trees shot the OOB errors ration that in this estimation are pretty good. Random Forest algorithm is done with the initial data, without removing values. After all processes are done some interesting results can be interpreted from this data analysis. First is that male students are playing video games more then female which is expected. Also, in what can be seen in tables is that grade for male students is bigger this semester then the last one, which can be do the online classes and online education. By looking the three values (Hours per day, days in week and scale) we have a big difference between male and female students but the grade for this semester (in average) are bigger for the male students, which can indicate either the online education are more suitable for male children, or that increment in small amount of video games can increase the productivity of male children. One

more interesting thing is that Days in week for male and female participant have the same significance but, in Hours per day they are the great significance for male students, and slightly less for female participants. In part of two sample t-test 4 plot show the results and the significance that are mentioned above and boxplot with confidential interval of the results. Before starting to talk about linear regressions models, not all values are used in two models only ones that can be significant and get real results for prediction.

REFERENCES

[1] [J. Wright, "The effects of video game play on academic performance," Mod. Psychol. Stud., vol. 17, no. 1, p. 6, 2011.

[2] M. M. Skoric, L. L. C. Teo, and R. L. Neo, "Children and video games: Addiction, engagement, and scholastic achievement," Cyberpsychology Behav., vol. 12, no. 5, pp. 567–572, 2009, doi: 10.1089/cpb.2009.0079.

[3] https://www.r-project.org/about.html

[4] https://www.investopedia.com/terms/d/data-analytics.asp

[5] M. E. Megel and J. A. Heermann, "Methods of data collection," Plast. Surg. Nurs., vol. 14, no. 2, pp. 109–110, 1994, doi: 10.1097/00006527-199406000-00014

[6] D. Plot, "Density Plots," Density Plots, pp. 1–7, 2021, doi: 10.4135/9781529772364.

[7] D. Rasch, K. D. Kubinger, and K. Moder, "The two-sample t test: Pre-testing its assumptions does not pay off," Stat. Pap., vol. 52, no. 1, pp. 219–231, 2011, doi: 10.1007/s00362-009-0224-x.

[8] P. Sedgwick, "Understanding P values," BMJ, vol. 349, no. January, pp. 10–12, 2014, doi: 10.1136/bmj.g4550.

[9] G. K. Uyanık and N. Güler, "A Study on Multiple Linear Regression Analysis," Procedia - Soc. Behav. Sci., vol. 106, pp. 234–240, 2013, doi: 10.1016/j.sbspro.2013.12.027.

[10] A. Malek, M. Ninčević, and D. Jurić Vukelić, "The Role of Playing Video Games on School Achievement," C

[11] E. L. S. Sara Prot, Craig A.Anderson, Douglas A.Gentile, Stephanie C. Brown, "The Positive and Negative Effects of Video Games," Smart Kids, pp. 2–4, 2014, [Online]. Available: http://www.raisesmartkid.com/3-to-6-years-old/4-articles/34-the-good-and-bad-effects-of-video-games

[12] A. Ayeni, "Empirics of Standard Deviation," no. May, pp. 1–8, 2014, doi:

[13] http://uc-r.github.io/gda