STATISTICS APPLIED WITH R USING ARTIFICIAL INTELLIGENCE

[1]MGS. EDWIN FERNANDO MEJÍA PEÑAFIEL, [2] MSC. RAQUEL VIRGINIA COLCHA ORTIZ, [3]MGS. WILLIAN GEOVANNY YANZA CHAVEZ, [4] MSC. MARCO ANTONIO GAVILANES SAGÑAY, [5]DRA. GLADYS LOLA LUJÁN JOHNSON

[1]ORCID https://orcid.org/ 0000-0001-6888-4621
Facultad de Ciencias, Docente-Investigador de la "Escuela Superior Politécnica de Chimborazo (ESPOCH)"
Riobamba 060103, Ecuador.
efmejia@espoch.edu.ec
[2]ORCID https://orcid.org/ 0000-0002-3252-9158
Facultad de Administración de Empresas
Docente-Investigador de la "Escuela Superior Politécnica de Chimborazo (ESPOCH)" Riobamba 060103, Ecuador.
e-mail: raquel.colcha@espoch.edu.ec.
[3]ORCID ID https://orcid.org/0000-0002-9688-7309
Facultad de Administración de Empresas
Docente-Investigador de la "Escuela Superior Politécnica de Chimborazo (ESPOCH)" Riobamba 060103, Ecuador.
e-mail: willian.yanza@espoch.edu.ec.
[4]ORCID https://orcid.org/0000-0002-7470-3732
Facultad de Administración de Empresas
Docente-Investigador de la "Escuela Superior Politécnica de Chimborazo (ESPOCH)" Riobamba 060103, Ecuador.
e-mail: marco.gavilanes@espoch.edu.ec
[5]Escuela de Posgrado Universidad César Vallejo, Sede Piura Perú
ORCID https://orcid.org/0000-0002-4727-6931
ljohnsongl@ucvvirtual.edu.pe

Abstract: *Machine learning within R using inferential statistics in academic fields today is already a reality, that research is directed towards the world of artificial intelligence as an aid within engineering based on tools such as machine learning, in this case with supervised learningand unsupervised learning. It is important because it can be applied in engineering, such as in statistics and the world of artificial intelligence to automate statistical processesof various samples in this area.*

Objective*. Perform adescriptive statistic using quantitative data in R through structured programming compared with supervised and unsupervised learning algorithms for decision making in engineering.*

Methodology*. This research is descriptive, uses the Methodology of Construction of an Algorithm for the systemic Learning of students of the first semester of the subject of ICTs (Mejía E. et al, 2018) adapted to the use of artificial intelligence algorithms, which provides effective methods that allow to implement adescriptive statistic, to perform programs using functions and procedures within R. Tests were carried out with third-semester students of the ESPOCH Statistics career to determine with structured programming (before) and artificial intelligence algorithms (after).*

Results*. Students using structured programming to obtain these statistics only 36.84% reach n to complete the work, while students using artificial intelligence algorithms reach 84.21%, to conclude the work.*

Conclusion*. It is concluded that under the parameters of use of artificial intelligence algorithms to obtain a descriptive statistic for engineering, makes decisions with percentages that favor these* *techniques provided by the researchersis, this result He tells us that the second option is the best, obtaining in statistical terms very favorableconditions to insert this technique and methodology in engineering environments. Students with these algorithms that use supervised and unsupervised learning will have an extra plus when performing this type of statistics in the professional environment.*

Keywords: *Programming, Program, Programming methodology, Statistics, Artificial intelligence, supervised learning ,unsupervised learning.*

Table of Contents

Introduction

The advances of artificial intelligence, the interest and application of machine learning has experienced such an expansion, that it has become an applied discipline in practically all areas of academic and industrial research (D.H. Kim and T. MacKinnon, 2018) . The growing number of people dedicated to this discipline has resulted in a whole repertoire of tools with which, profiles with medium specialization, manage to access powerful predictive methods (A. Núñez Reiz et all., 2019). The R programming language is an example of this. (Orellana, 2019)

The term machine learning encompasses the set of algorithms that allow identifying patterns present in the data and creating with them structures (models) that represent them (Kai-Qi Li et al., 2022). Once models have been generated, they can be used to predict information about facts or events that have not yet been observed. It is important to remember that machine learning systems are only able to memorize patterns that are present in the data with which they are trained, therefore, they can only recognize what they have seen before(Manisha Koranga et all., 2022). By using systems trained with past data to predict futures, it is assumed that, in the future, the behavior will be the same, which is not always the case.

(Very, 2020)

Although terms such as machine learning, data mining, artificial intelligence, data science are often used synonymously, it is important to note that machine learning methods (Xin Li et al., 2022) are only part of the many strategies that need to be combined to extract information, understand and give value to data. The following document aims to be an example of the type of problem that an analyst usually faces: starting from a more or less processed data set (data preparation is a critical stage that precedes machine learning), you want to create a model that allows you to successfully predict the behavior or value that new observations take. (Amat, 2020)

Unlike other documents, this one is intended to be a practical example with less theoretical development. The reader will realize how easy it is to apply a wide range of predictive methods with R and its libraries. However, it is crucial that any analyst understands the theoretical foundations on which each of them is based for such a project to succeed. Although they are only briefly described here, they will be accompanied by links where you can find detailed information. (Hernandez, 2021)

To understand the origin of the programming language in R, it is important to know some relevant historical points in its evolution. It was created in 1993 by professors and researchers Robert Gentleman and Ross Ihaka (Hernández and Usuga, 2021).

However, its beginnings date back to the previous language called S, by John Chambers and his collaborators at Bell Laboratories, during the seventies. It should be noted that since 1995, the source code of R is available under GNU public license for Windows, MacOS and Linux distributions. This license is maintained with free access by the R Foundation, and can be executed, copied, distributed, studied, modified and improved without any restriction (Vargas y Mesa, 2021).

The R Foundation is a non-profit organization, conceived by the members of the R Development Core Team, whose objective is to provide support to R and the innovations in computational statistics that it requires, ensuring continuous development (R Core Team, 2020). This manual uses R to talk about the programming language, but not the application or statistical package R. This clarification has as exception the download and installation of the program with the same name.

RStudio is the main integrated development environment or IDE (Integrated Development Environment) forR, which is free software available for Windows, MacOS, and Linux operating systems (Equipo RStudio, 2020).

In particular, R.es an interpreted language, which can run on Linux, Mac and Windows because interpreters are available on these operating systems. In any case, being an interpreted language will always be slower than a compiled language. If you are looking for computing speed in complex problems, R is not the right program. However, if you are looking for a compromise between ease of handling and reasonable execution speed in the analysis of datasets, R may be the perfect tool. (Santana and Hernandez, 2016)

In this sense, you should point out that although many R functions are programmed in the R code itself, but many others are programmed in C or Fortran, so their execution is very fast. In many cases (possibly most), the R user does not need to know the complexities of the program (in C or Fortran) that is behind some function, and it is enough simply that this function integrates comfortably with the rest of his data process. (Santana and Hernandez, 2016)

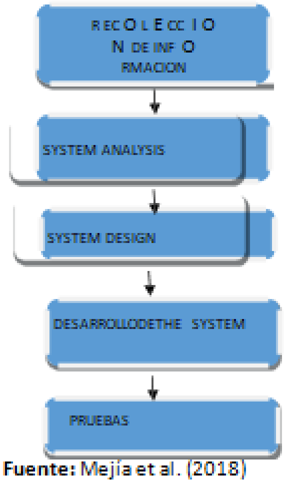
The objective of this research work is to perform a descriptive statistic using quantitative data in R through structured programming compared with supervised and unsupervised learning algorithms for decision making in engineering. The research performs these three exercises with the same number of students within the semester and the resolution of each exercise is verified.

1. Methodology

Itwill use the Methodology for Building an Algorithm for Systemic Learning of first-semester students of the ICT subject (Mejía et al., 2018) that observes en la figura 1 with small changes in the use of structured programming and machine learning:

Figura 1

Itodolog ía for structured programming and machine learning



Fuente: Mejía et al. (2018)

1.1 Proyecto propuesto

S epretendedar solution with a data analysis using descriptive statistics through quantitative data in R with structured programming compared to supervised and unsupervised learning algorithms for decision making in engineering. The research performs these three exercises with the same number of students within the semester and the resolution of each exercise is verified. A program within R for the analysis of quality results thatallow us to makedecisions.

1.2 Program

A program is a set of logical steps written in a programming language that allows us to perform a specific task. (CILSA, 2017).

1.3 R Language

This environment is commonly used for statistical and graphical computing, as it has a wide variety of statistical techniques (linear and nonlinear models, classical statistical tests, time series analysis, classification, clustering, etc.) and graphs. It works on UNIX platforms and similar systems (including FreeBSD and Linux), Windows, and MacOS.

Its current development is the responsibility of the R Development Core Team. It is part of a collaborative and open project where users can publish packages that extend their basic configuration (official package repository). In addition, it can be downloaded for free through the following link: https://www.r-project.org/.

1.4 Machine Learning

Machine learning is a discipline in the field of artificial intelligence. This discipline consists of making a computer learn to perform certain processes automatically, with as little human intervention as possible.

(Sierra, 2022)

1.5 Microsoft Exel 365

Excel is a very powerful tool for obtaining meaningful insights from large amounts of data. It also works great with simple calculations and to keep track of almost any type of information. The key to unlocking all this potential is the grid of cells. Cells can contain numbers, text, or formulas. Data is written to cells and grouped into rows and columns. This allows you to summarize, sort and filter it, put it in tables and create very visual graphs. (Microsoft, 2022)

We will use Excelthat is within office 365 which we have licensed within the Polytechnic School of Chimborazo.

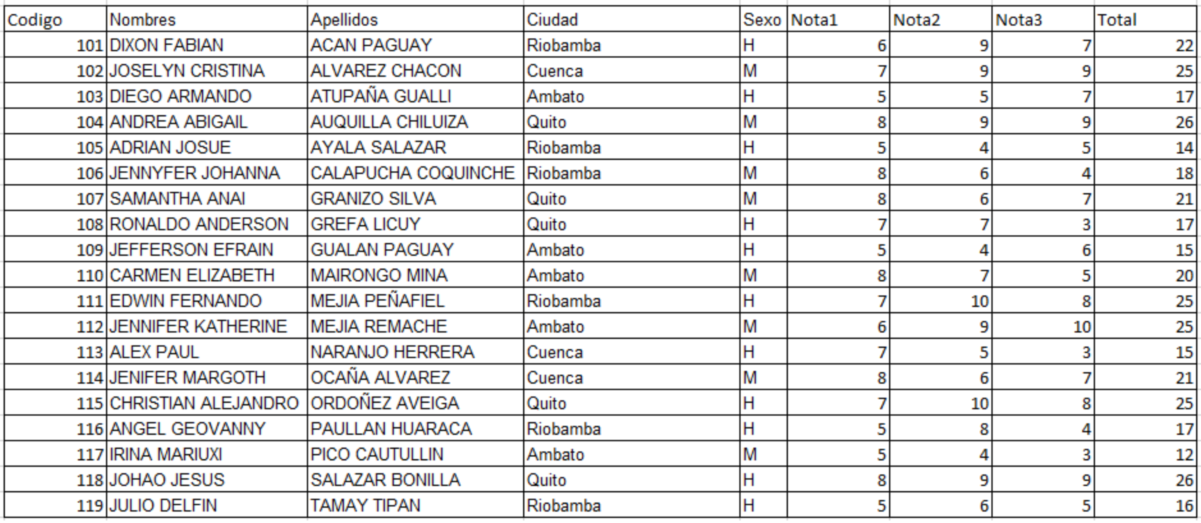
Both to use structured programming and for machine learning we will use a Database (Excel) with all the data required to present the desired results.

1.6 Database in Excel

Figure2 shows the database touse to resolve this issue.

Figura 2

Database in Excel



Codigo	Nombres	Apellidos	Ciudad	Sexo	Nota1	Nota2	Nota3	Total
101	DIXON FABIAN	ACAN PAGUAY	Riobamba	H	6	9	7	22
102	JOSELYN CRISTINA	ALVAREZ CHACON	Cuenca	M	7	9	9	25
103	DIEGO ARMANDO	ATUPAÑA GUALLI	Ambato	H	5	5	7	17
104	ANDREA ABIGAIL	AUQUILLA CHILUIZA	Quito	M	8	9	9	26
105	ADRIAN JOSUE	AYALA SALAZAR	Riobamba	H	5	4	5	14
106	JENNYFER JOHANNA	CALAPUCHA COQUINCHE	Riobamba	M	8	6	4	18
107	SAMANTHA ANAI	GRANIZO SILVA	Quito	M	8	6	7	21
108	RONALDO ANDERSON	GREFA LICUY	Quito	H	7	7	3	17
109	JEFFERSON EFRAIN	GUALAN PAGUAY	Ambato	H	5	4	6	15
110	CARMEN ELIZABETH	MAIRONGO MINA	Ambato	M	8	7	5	20
111	EDWIN FERNANDO	MEJIA PEÑAFIEL	Riobamba	H	7	10	8	25
112	JENNIFER KATHERINE	MEJIA REMACHE	Ambato	M	6	9	10	25
113	ALEX PAUL	NARANJO HERRERA	Cuenca	H	7	5	3	15
114	JENIFER MARGOTH	OCAÑA ALVAREZ	Cuenca	M	8	6	7	21
115	CHRISTIAN ALEJANDRO	ORDOÑEZ AVEIGA	Quito	H	7	10	8	25
116	ANGEL GEOVANNY	PAULLAN HUARACA	Riobamba	H	5	8	4	17
117	IRINA MARIUXI	PICO CAUTULLIN	Ambato	M	5	4	3	12
118	JOHAO JESUS	SALAZAR BONILLA	Quito	H	8	9	9	26
119	JULIO DELFIN	TAMAY TIPAN	Riobamba	H	5	6	5	16

Source: Authors

1.7 Population

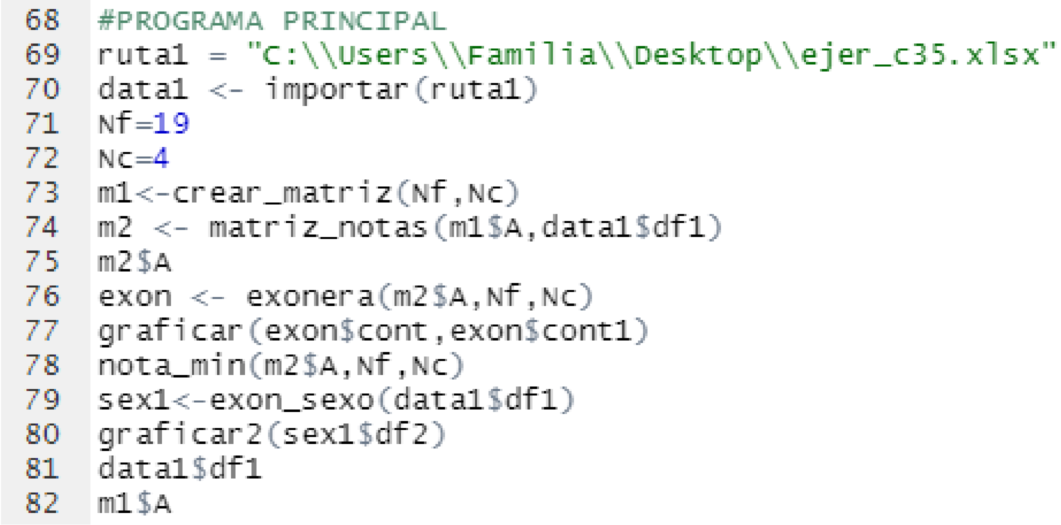
The population that will be used to carry out the three exercises with both structured programming and supervised and unsupervised learning is 19 students, the same as those who are from the subject of statistical programming of the third semester of the Statistics career of ESPOCH, Faculty of Sciences.

2. Results

Scheme to be solved as a prototype with structured programming – This program that is going to be carried out is composed of lines of code in R where the s lines of code are groupedin a main program that calls the functions within it, information is obtained as exonerated students, minimum and maximum of the three partial by Supposed statistical graphics for decision making, in this program and from the point of view of programmer that in this case comes to do the teacher and the 1 9 students, as a whole has several lines of code, which call from the main program to the different functions that allow to create different data stores or dataframes and also matrices for the correct use of the presentation of results as shown in Figure 3. In figure 3 we have in line 69 the call to our database in excel, from where it starts with the creation of our data1 dataframe, an m1 matrix is created using the crear_matriz function and from our data1 the m2 matrix is created with data chosen for our statistics. Then you have the exoneration function where you get how many students are exonerated and how many students are not exonerated. You also have the graph of exonerated, you get the minimum and maximum grades. The exón_sexo function is called to visualize how many men and how many women have been exonerated and its graph. Finally, all the information contained in our database is displayed in data1$df1 as well as the notes of the partials and the sum in m1$A.

Fimouth 3

Core program using structured programming



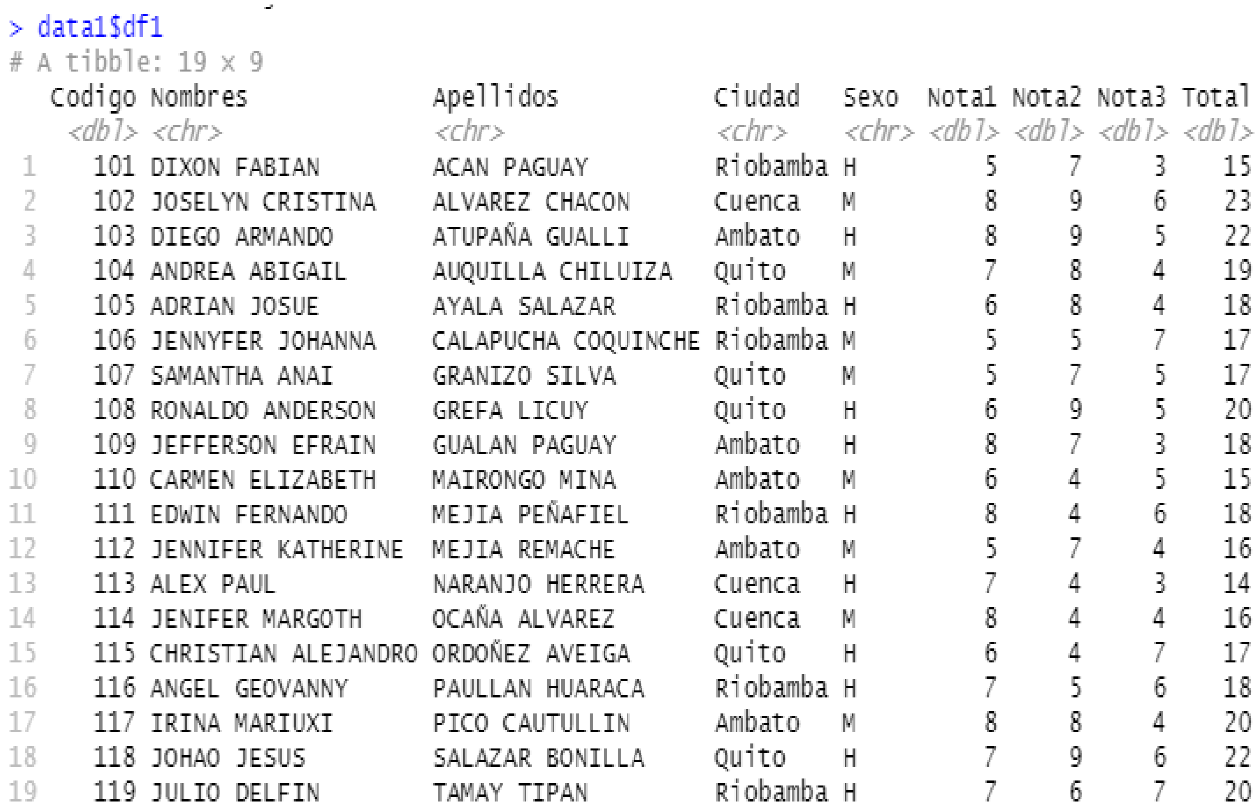
```
68  #PROGRAMA PRINCIPAL
69  ruta1 = "C:\\Users\\Familia\\Desktop\\ejer_c35.xlsx"
70  data1 <- importar(ruta1)
71  Nf=19
72  Nc=4
73  m1<-crear_matriz(Nf,Nc)
74  m2 <- matriz_notas(m1$A,data1$df1)
75  m2$A
76  exon <- exonera(m2$A,Nf,Nc)
77  graficar(exon$cont,exon$cont1)
78  nota_min(m2$A,Nf,Nc)
79  sex1<-exon_sexo(data1$df1)
80  graficar2(sex1$df2)
81  data1$df1
82  m1$A
```

Source: Authors

Figure 4 shows line 81 of Figure 3, showing the data1$df1, which are the data from the database under study.

Fimouth 4

Data frame with all database data



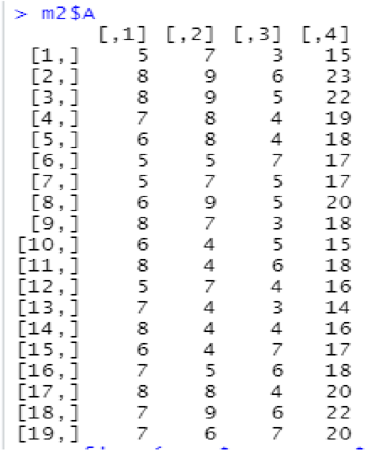
```
> data1$df1
# A tibble: 19 x 9
   Codigo Nombres             Apellidos           Ciudad   Sexo  Nota1 Nota2 Nota3 Total
    <dbl> <chr>               <chr>               <chr>    <chr> <dbl> <dbl> <dbl> <dbl>
 1    101 DIXON FABIAN        ACAN PAGUAY         Riobamba H         5     7     3    15
 2    102 JOSELYN CRISTINA    ALVAREZ CHACON      Cuenca   M         8     9     6    23
 3    103 DIEGO ARMANDO       ATUPAÑA GUALLI      Ambato   H         8     9     5    22
 4    104 ANDREA ABIGAIL      AUQUILLA CHILUIZA   Quito    M         7     8     4    19
 5    105 ADRIAN JOSUE        AYALA SALAZAR       Riobamba H         6     8     4    18
 6    106 JENNYFER JOHANNA    CALAPUCHA COQUINCHE Riobamba M         5     5     7    17
 7    107 SAMANTHA ANAI       GRANIZO SILVA       Quito    M         5     7     5    17
 8    108 RONALDO ANDERSON    GREFA LICUY         Quito    H         6     9     5    20
 9    109 JEFFERSON EFRAIN    GUALAN PAGUAY       Ambato   H         8     7     3    18
10    110 CARMEN ELIZABETH    MAIRONGO MINA       Ambato   M         6     4     5    15
11    111 EDWIN FERNANDO      MEJIA PEÑAFIEL      Riobamba H         8     4     6    18
12    112 JENNIFER KATHERINE  MEJIA REMACHE       Ambato   M         5     7     4    16
13    113 ALEX PAUL           NARANJO HERRERA     Cuenca   H         7     4     3    14
14    114 JENIFER MARGOTH     OCAÑA ALVAREZ       Cuenca   M         8     4     4    16
15    115 CHRISTIAN ALEJANDRO ORDOÑEZ AVEIGA      Quito    H         6     4     7    17
16    116 ANGEL GEOVANNY      PAULLAN HUARACA     Riobamba H         7     5     6    18
17    117 IRINA MARIUXI       PICO CAUTULLIN      Ambato   M         8     8     4    20
18    118 JOHAO JESUS         SALAZAR BONILLA     Quito    H         7     9     6    22
19    119 JULIO DELFIN        TAMAY TIPAN         Riobamba H         7     6     7    20
```

Source: Authors

Now in figure 5 the information of notes of the three partials is shown together with the sum of them in the table m1$A:

Fimouth 5

Table with partials and the sum of them



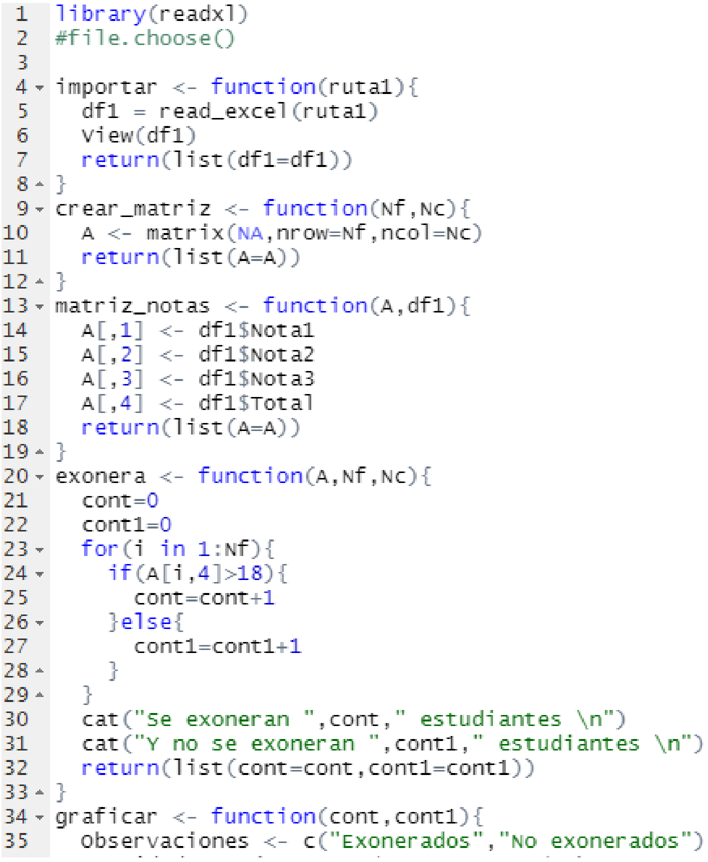
```
> m2$A
      [,1] [,2] [,3] [,4]
 [1,]    5    7    3   15
 [2,]    8    9    6   23
 [3,]    8    9    5   22
 [4,]    7    8    4   19
 [5,]    6    8    4   18
 [6,]    5    5    7   17
 [7,]    5    7    5   17
 [8,]    6    9    5   20
 [9,]    8    7    3   18
[10,]    6    4    5   15
[11,]    8    4    6   18
[12,]    5    7    4   16
[13,]    7    4    3   14
[14,]    8    4    4   16
[15,]    6    4    7   17
[16,]    7    5    6   18
[17,]    8    8    4   20
[18,]    7    9    6   22
[19,]    7    6    7   20
```

Source: Authors

Also in figure 6 we present some functions that are used for this process, where single and double conditional control structures, cyclic control structures such as for loops, arrays, dataframes and packages such as ggplot2 for graphs are used.

Fimouth 6

Functions used for this process



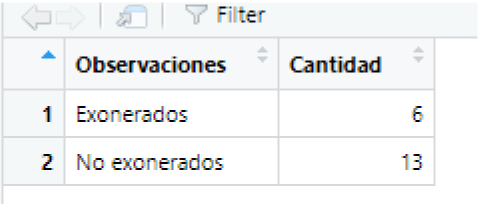
```
library(readxl)
#file.choose()

importar <- function(ruta1){
  df1 = read_excel(ruta1)
  View(df1)
  return(list(df1=df1))
}
crear_matriz <- function(Nf,Nc){
  A <- matrix(NA,nrow=Nf,ncol=Nc)
  return(list(A=A))
}
matriz_notas <- function(A,df1){
  A[,1] <- df1$Nota1
  A[,2] <- df1$Nota2
  A[,3] <- df1$Nota3
  A[,4] <- df1$Total
  return(list(A=A))
}
exonera <- function(A,Nf,Nc){
  cont=0
  cont1=0
  for(i in 1:Nf){
    if(A[i,4]>18){
      cont=cont+1
    }else{
      cont1=cont1+1
    }
  }
  cat("Se exoneran ",cont," estudiantes \n")
  cat("Y no se exoneran ",cont1," estudiantes \n")
  return(list(cont=cont,cont1=cont1))
}
graficar <- function(cont,cont1){
  Observaciones <- c("Exonerados","No exonerados")
```

Source: Authors

Table 1 indicates the number of exempt and non-exonerated students who are in the subject of statistical programming, and then in figure 7 present a frequency diagram graphwith these data.

Table 1

Number of students exonerated and not exempted from PAO3

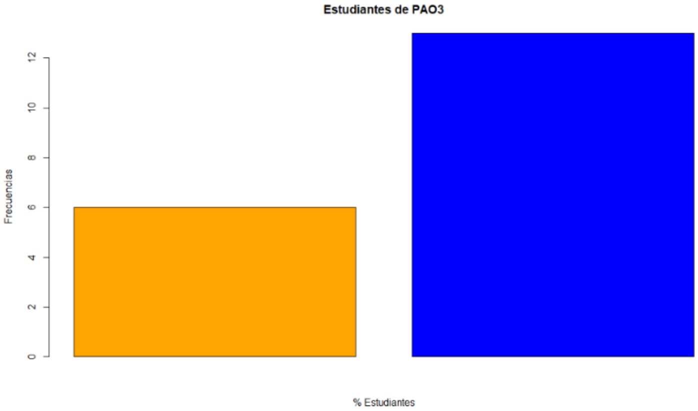


	Observaciones	Cantidad
1	Exonerados	6
2	No exonerados	13

Source: Authors

Fimouth 7

Students exonerated and not exonerated from PAO3

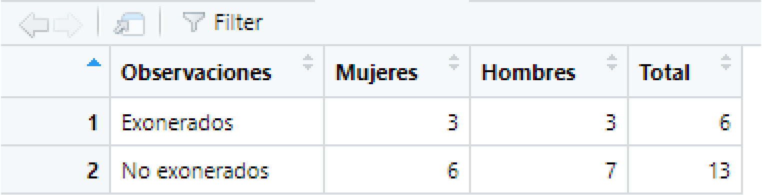


Source: Authors

Table 2 shows the number of students by sex exonerated and Figure 8 shows:

Table 2

Number of students exonerated and not exempted from PAO3

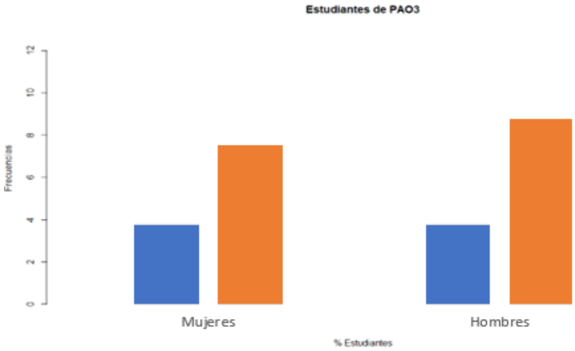


	Observaciones	Mujeres	Hombres	Total
1	Exonerados	3	3	6
2	No exonerados	6	7	13

Source: Authors

Fimouth 8

Students exonerated and not exonerated by sex of PAO3



Source: Authors

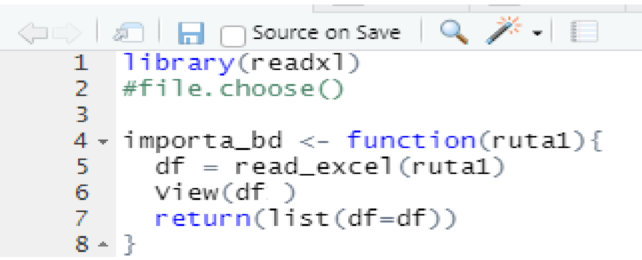
As you can see we have 2 types of graphs one by number of students in exonerated and non-exonerated, and the other graph by number of students belonging to a male or female sex with the same category of exoneration. Which allow to select data through the s functions that generate tables and data frame within RStudio using the programming language R and under structured programming for the decision makingofteachers.

Machine learning - supervised learning with R – It is ta presentation of results using algorithms with supervised learning that is going to be carried out is composed of some functions to perform a study by segmentation of data which groups the information and chains them according to the algorithm that internally handles PowerBI, in this report made and from the point of view of user that in this case comes to do the teacher, as a whole has several objects and that is shown in several presentations within the same sheet, which show in the same s based on the segmentation before said, somes Presentations have been expressed in an ideal way to show the necessary information, which leads us to decision-making within this scope of partial grades within the current semester. In addition, several R-based presentation objects are used to display results using programming. The methodology of building the dashboard is under R scripts, there is structured programming based on packages and libraries of this language.

Data import – First when entering RStudio, we import data from our databasein Excel, as shown in figure 9, a function is performed for this purpose.

Fimouth 9

Datamonitoring from Excel



```
library(readxl)
#file.choose()

importa_bd <- function(ruta1){
  df = read_excel(ruta1)
  View(df )
  return(list(df=df))
}
```

Source: Authors

With this, the Excel database is passed to RStudio within a data frame to be able to carry out the study.

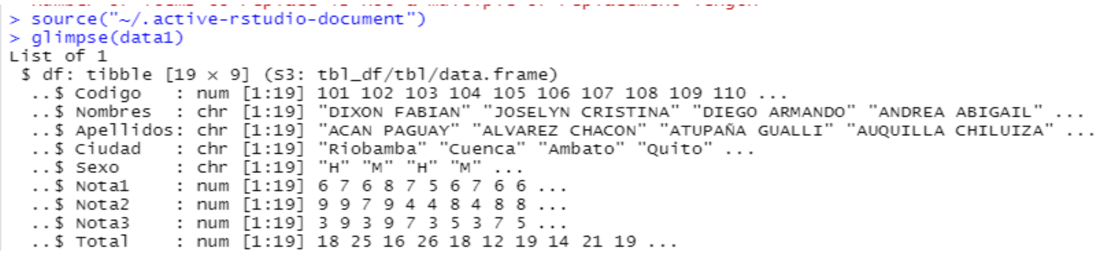
A first check to take into account after inserting the data within RStudio, is to verify that each variable has already been stored with the corresponding value and type, that is:

numerical variables with numbers, character, Boolean and qualitative factor. In R, preferably a qualitative variable, should be stored asa factor type.

When using the following code: glimpse(data1),it displays a table with the following information:

Fimouth 10

Type of Variables



```
> source("~/.active-rstudio-document")
> glimpse(data1)
List of 1
 $ df: tibble [19 × 9] (S3: tbl_df/tbl/data.frame)
  ..$ Codigo   : num [1:19] 101 102 103 104 105 106 107 108 109 110 ...
  ..$ Nombres  : chr [1:19] "DIXON FABIAN" "JOSELYN CRISTINA" "DIEGO ARMANDO" "ANDREA ABIGAIL" ...
  ..$ Apellidos: chr [1:19] "ACAN PAGUAY" "ALVAREZ CHACON" "ATUPAÑA GUALLI" "AUQUILLA CHILUIZA" ...
  ..$ Ciudad   : chr [1:19] "Riobamba" "Cuenca" "Ambato" "Quito" ...
  ..$ Sexo     : chr [1:19] "H" "M" "H" "M" ...
  ..$ Nota1    : num [1:19] 6 7 6 8 7 5 6 7 6 6 ...
  ..$ Nota2    : num [1:19] 9 9 7 9 4 4 8 4 8 8 ...
  ..$ Nota3    : num [1:19] 3 9 3 9 7 3 5 3 7 5 ...
  ..$ Total    : num [1:19] 18 25 16 26 18 12 19 14 21 19 ...
```

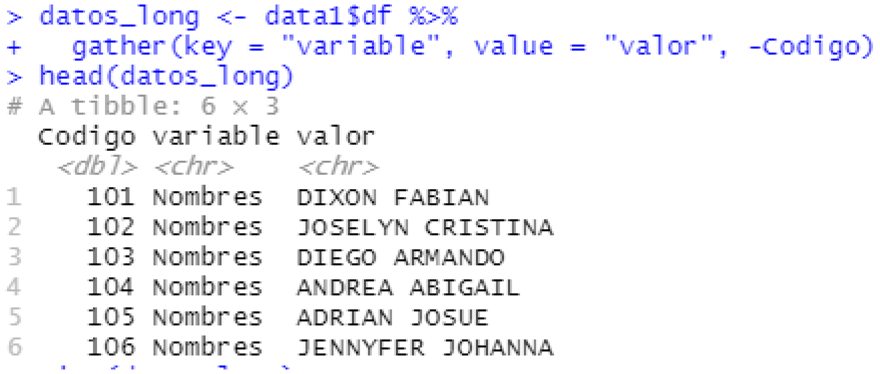
Source: Authors

Figure 10 shows the type of each variable, as well as the values that each one has within the database.

The exploratory analysis of variables is characterized by performing summation calculations, restructuringthe data and graphingthem. The different packages such as ggplot2, tidyr, dplyr and others, which come within tydiverse give us a lot of ease having toprogram few lines. The structure is followed: observation, variable, value to perform the calculations more quickly. A data frame is created with these fields since it is not very large data setwithout having memorystorage problems.

Fimouth 11

Structure observation, variable and value



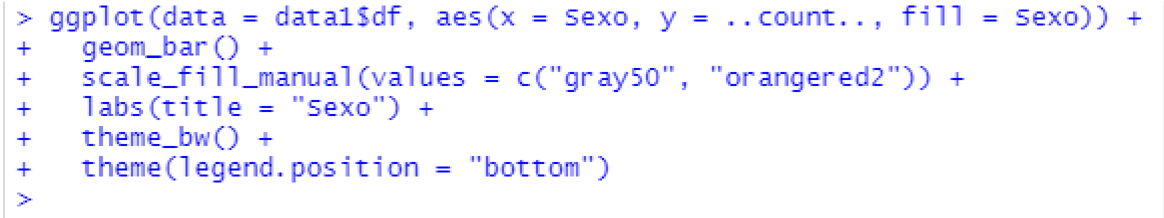
```
> datos_long <- data1$df %>%
+   gather(key = "variable", value = "valor", -Codigo)
> head(datos_long)
# A tibble: 6 x 3
  Codigo variable valor
   <dbl> <chr>    <chr>
1    101 Nombres  DIXON FABIAN
2    102 Nombres  JOSELYN CRISTINA
3    103 Nombres  DIEGO ARMANDO
4    104 Nombres  ANDREA ABIGAIL
5    105 Nombres  ADRIAN JOSUE
6    106 Nombres  JENNYFER JOHANNA
```

Source: Authors

When creating a model, the study must be characterized with a distribution of the response variable, since itis the most important thing in data prediction.

Fimouth 12

Structure observation, variable and value



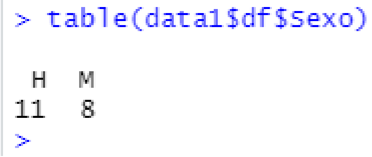
```
> ggplot(data = data1$df, aes(x = Sexo, y = ..count.., fill = Sexo)) +
+   geom_bar() +
+   scale_fill_manual(values = c("gray50", "orangered2")) +
+   labs(title = "Sexo") +
+   theme_bw() +
+   theme(legend.position = "bottom")
>
```

Source: Authors

Figure 12 shows the code to obtain a graph with the number of people of different sex, as we have in table 3 their frequencies:

Table 3

Number of students by sex in PAO3



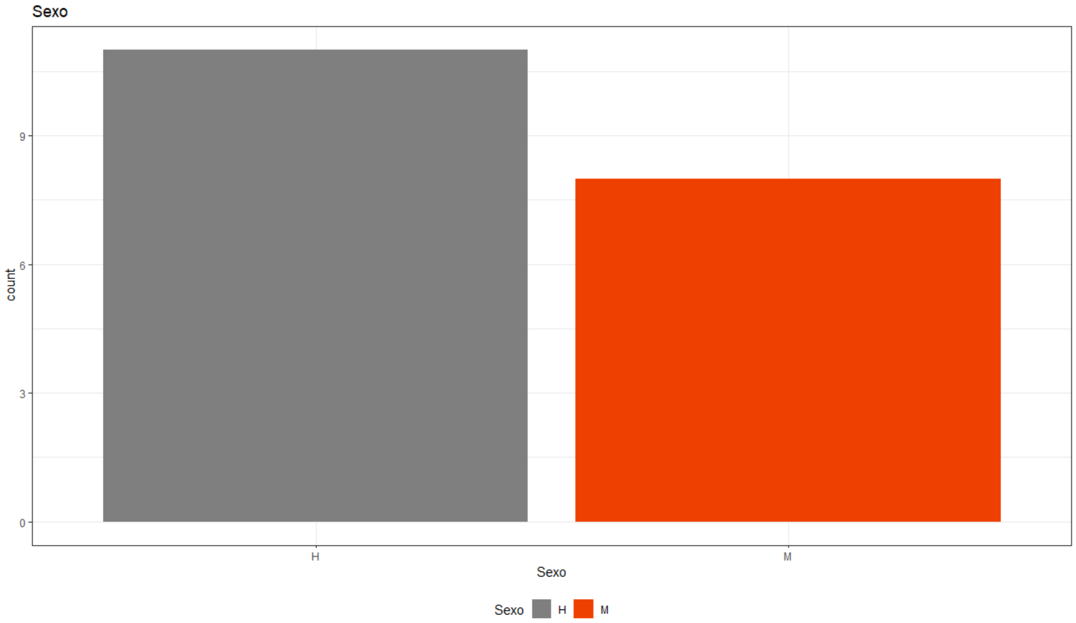
```
> table(data1$df$Sexo)

 H  M
11  8
>
```

Source: Authors

And in figure 13 you can see the graph with these data, from table 3 and figure 12.

Fimouth 13

Students by type of sex

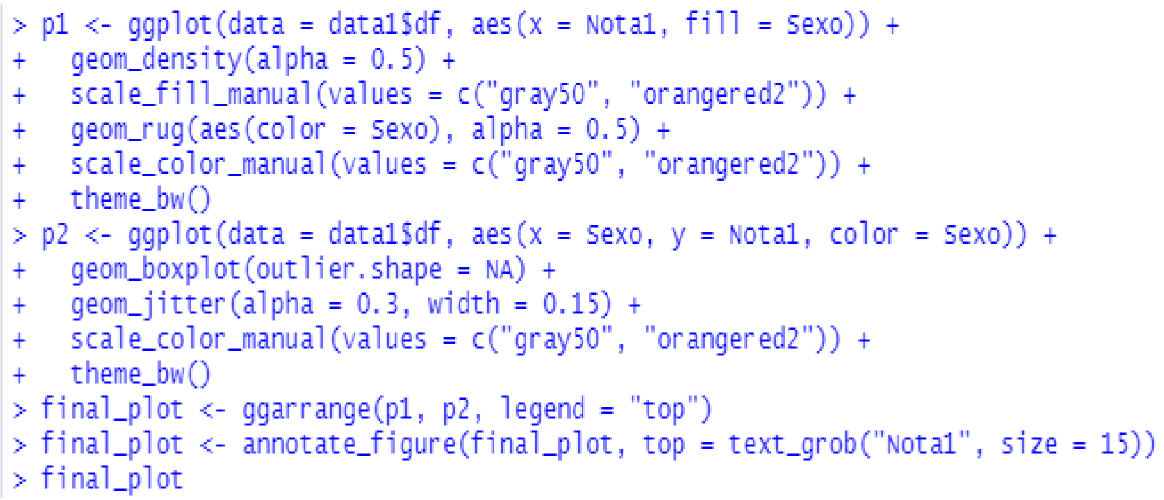


Source: Authors

Now we are going to make apredictionof which students are exonerated and which students are not exonerated, the data analysis is done with respect to the response variable Sex. Performing this analysis, situations are extracted on the variables that are well related with respect to students regarding sex and their grades. Figure 14 shows how to do this using R code.

Fimouth 14

R code for continuous variable distribution



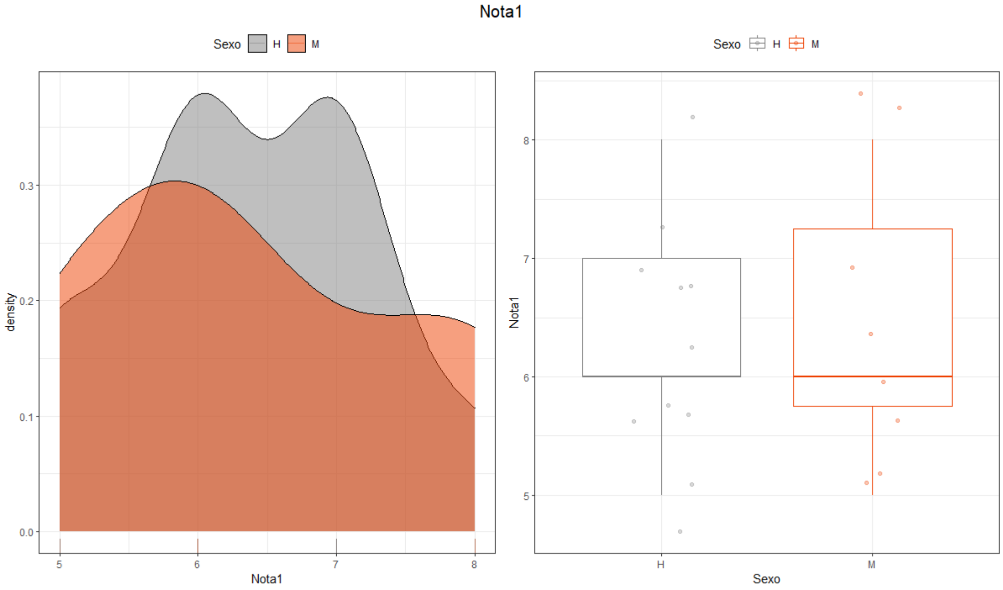
```
> p1 <- ggplot(data = data1$df, aes(x = Nota1, fill = Sexo)) +
+   geom_density(alpha = 0.5) +
+   scale_fill_manual(values = c("gray50", "orangered2")) +
+   geom_rug(aes(color = Sexo), alpha = 0.5) +
+   scale_color_manual(values = c("gray50", "orangered2")) +
+   theme_bw()
> p2 <- ggplot(data = data1$df, aes(x = Sexo, y = Nota1, color = Sexo)) +
+   geom_boxplot(outlier.shape = NA) +
+   geom_jitter(alpha = 0.3, width = 0.15) +
+   scale_color_manual(values = c("gray50", "orangered2")) +
+   theme_bw()
> final_plot <- ggarrange(p1, p2, legend = "top")
> final_plot <- annotate_figure(final_plot, top = text_grob("Nota1", size = 15))
> final_plot
```

Source: Authors

Figure 15 shows exactly how this type of distribution of continuous variables affects the type of sex of each student and its corresponding grade.

Fimouth 15

Distribution of continuous variables



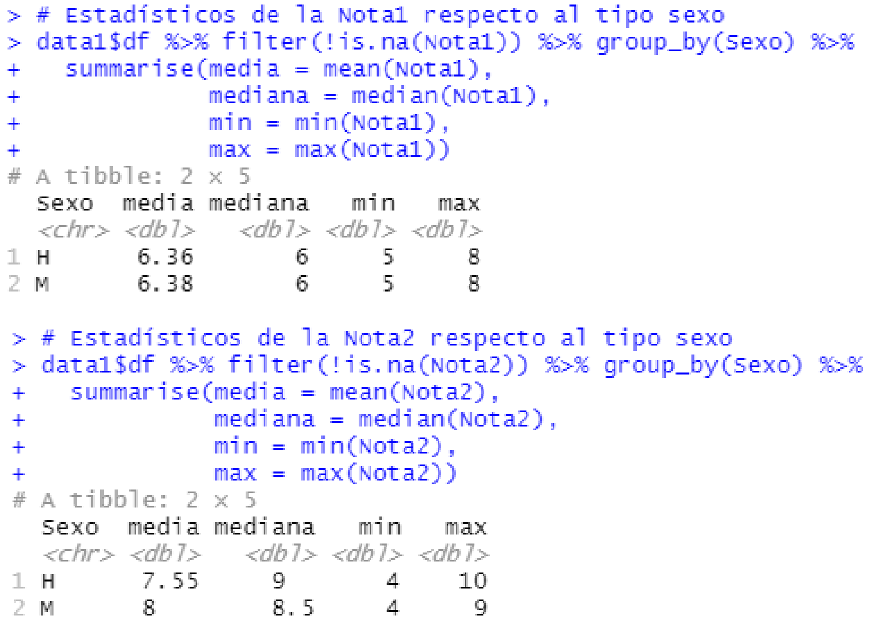
Source: Authors

As can be seen in Figure 15, very few students overcome the barrier of grade point average. So there you have to make some decisions by the teacher.

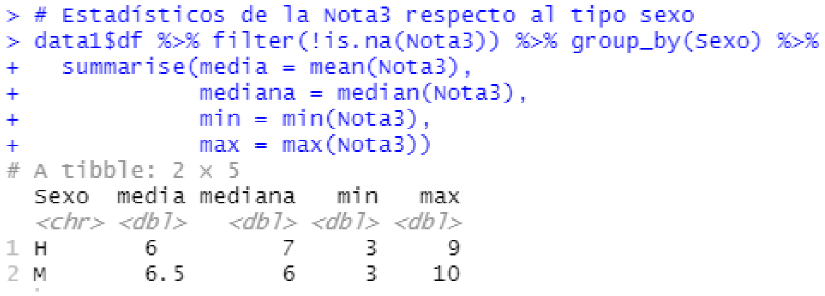
Now we are going to obtain the statistics of the Note1, Note2 and Note3 regarding the type of sex, for them the R code and the response table that visualizes us are indicated in figure 16.

Fimouth 16

Statistics of Note1, Note2 and Note3 with respect to
to the type of sex of each student



```
> # Estadísticos de la Nota1 respecto al tipo sexo
> data1$df %>% filter(!is.na(Nota1)) %>% group_by(Sexo) %>%
+   summarise(media = mean(Nota1),
+             mediana = median(Nota1),
+             min = min(Nota1),
+             max = max(Nota1))
# A tibble: 2 x 5
  Sexo  media mediana   min   max
  <chr> <dbl>   <dbl> <dbl> <dbl>
1 H      6.36       6     5     8
2 M      6.38       6     5     8

> # Estadísticos de la Nota2 respecto al tipo sexo
> data1$df %>% filter(!is.na(Nota2)) %>% group_by(Sexo) %>%
+   summarise(media = mean(Nota2),
+             mediana = median(Nota2),
+             min = min(Nota2),
+             max = max(Nota2))
# A tibble: 2 x 5
  Sexo  media mediana   min   max
  <chr> <dbl>   <dbl> <dbl> <dbl>
1 H      7.55     9       4    10
2 M      8        8.5     4     9
```



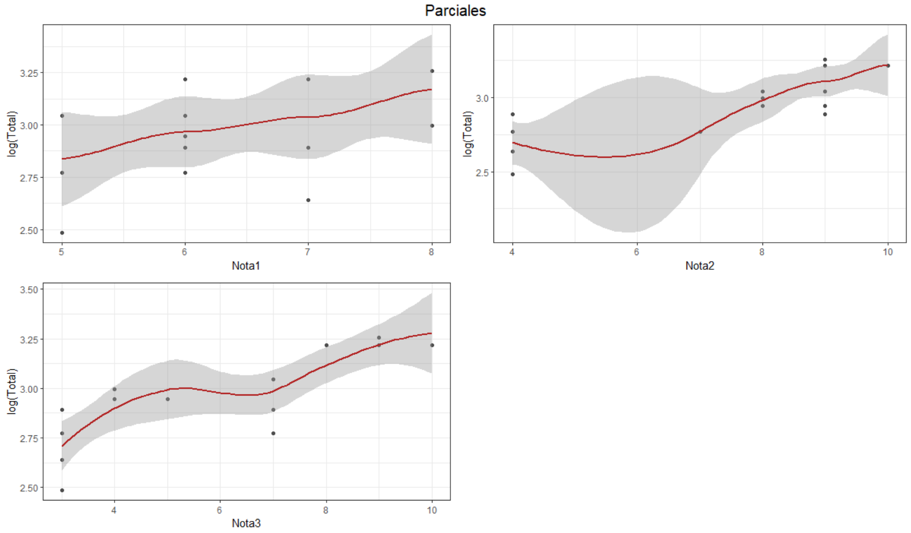
```
> # Estadísticos de la Nota3 respecto al tipo sexo
> data1$df %>% filter(!is.na(Nota3)) %>% group_by(Sexo) %>%
+   summarise(media = mean(Nota3),
+             mediana = median(Nota3),
+             min = min(Nota3),
+             max = max(Nota3))
# A tibble: 2 x 5
  Sexo  media mediana   min   max
  <chr> <dbl>   <dbl> <dbl> <dbl>
1 H       6         7     3     9
2 M       6.5       6     3    10
```

Source: Authors

As can be seen in figure 16, the statistics are as mean, median, minimum and maximum of the grades of the students of all the partials.

Fimouth 17

Correlation of continuous variables with respect to their partials

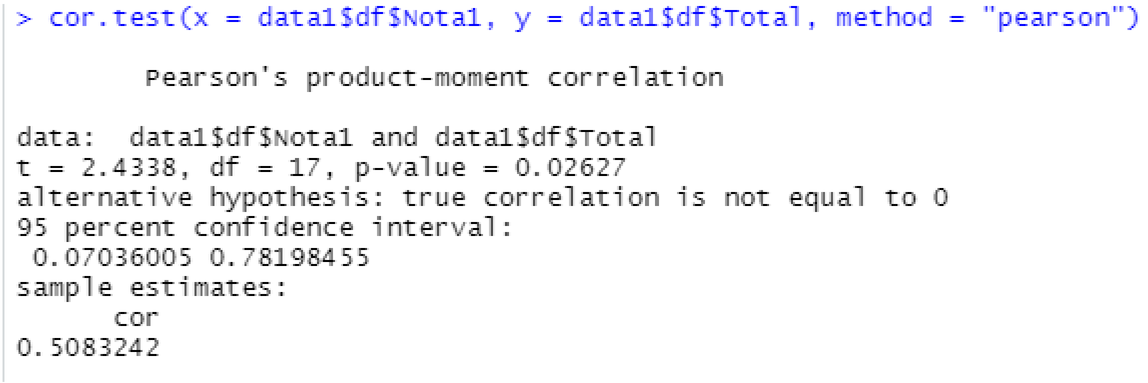


Source: Authors

Figure 17 shows thelinear correlation between thegrades of the midterms and the type of sex of each student, althoughvery significant (p-value = 0.0 26), is minimal (cor = 0.508) as shown in Figure 18. The type of diagram seen in the scatter figure does not point to some kind of obvious nonlinear relationship. It can be concluded that these variables in any case do not contain information that results.

Fimouth 18

Correlation between continuous variables



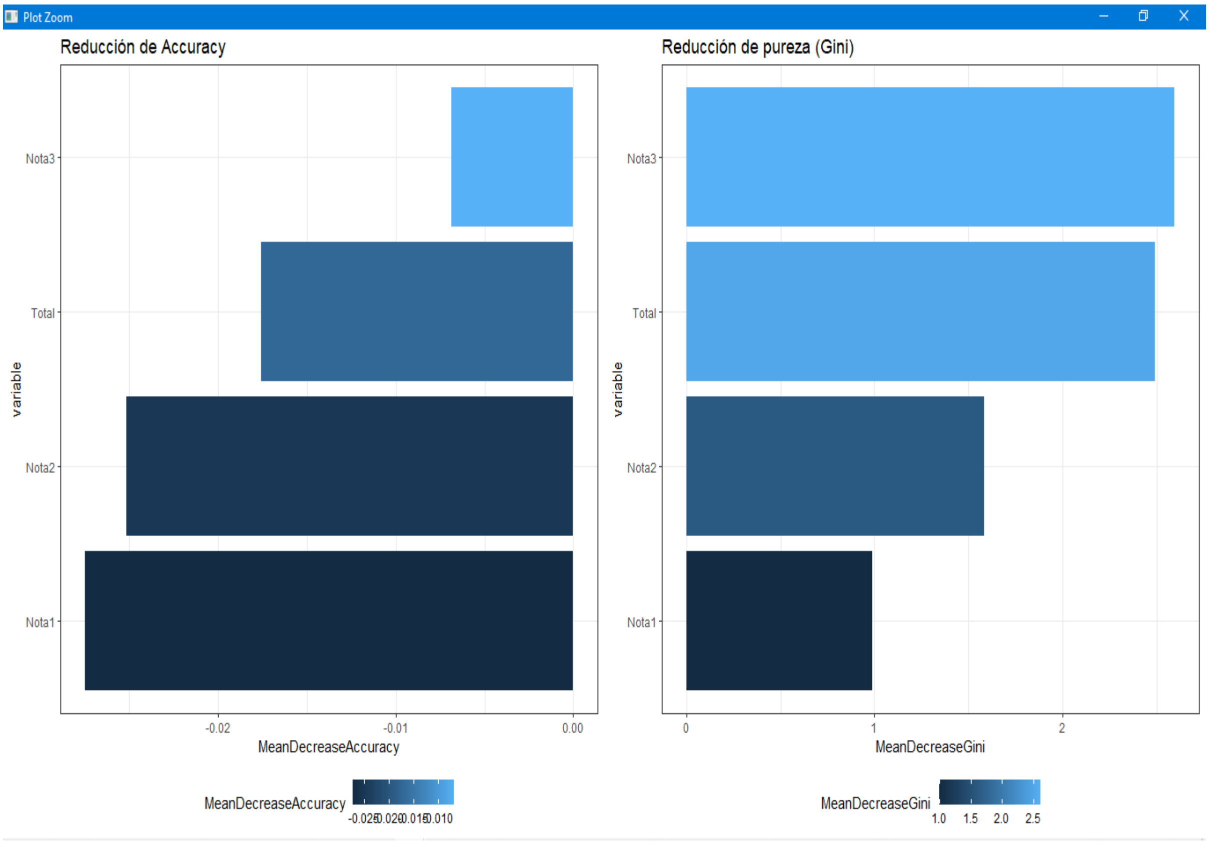
```
> cor.test(x = data1$df$Nota1, y = data1$df$Total, method = "pearson")

	Pearson's product-moment correlation

data:  data1$df$Nota1 and data1$df$Total
t = 2.4338, df = 17, p-value = 0.02627
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.07036005 0.78198455
sample estimates:
      cor
0.5083242
```

Source: Authors

Random Forest - When making the prediction with a Random Forest model, it must be taken into account that it is done to obtain theaverage of all the predictions that are in all the trees it contains, as shown in figure 19.

Fimouth 19

Random Forest for partials



Source: Authors

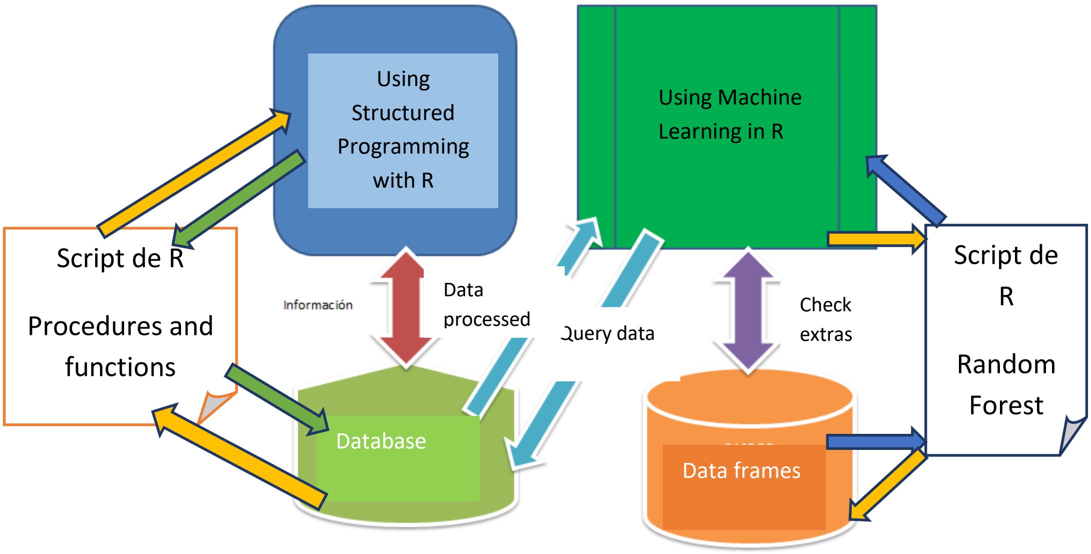
Figure 19 tells us that as we had in reality if students decrease the grades in the first and second partial they will no longer be able to exonerate themselves, while if the grades increase in the first, second and over in the third partial they will automatically be able to exonerate themselves.

The factors that most influence within the aspect of the exploration of the data, together with the study of the distribution, and the relationship that is had with the response variable indicate the different factors that influenced the exoneration or not of the students of the subject of statistical programming of the ESPOCH, Faculty of Sciences, career of Statistics, we have: obtaining low grades in two of the three partials. So, the way you have for the student to be exonerated is to follow up on a personal basis after the first partial so that in some way academic performance can increase. In test environments taken with the students of this subject has given us good results for the next semester followed with a decision making within the environment of first partial grades .

Project infrastructure – For this project we use figure 20 which shows how our project has been carried out. The main idea is that based on this figure the interaction of the different modules is carried out using structured programming and also using machine learning through random forest.

Figure 20

Interconnected systems of the project

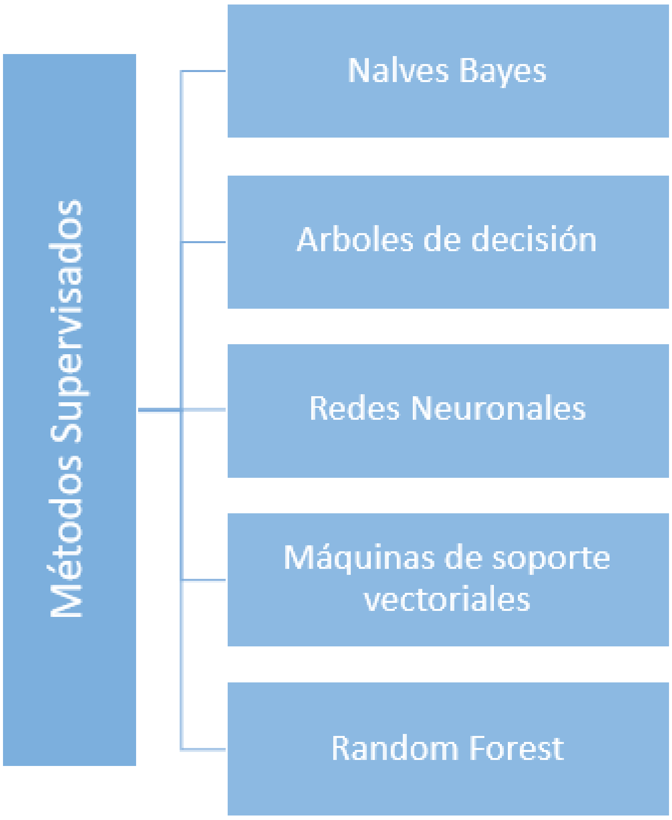


Source: Authors

Supervised learning - Supervised learning is a technique that uses algorithms that work with training sets made up of classes or labels, in this technique algorithms learn based on the history and make output predictions depending on the class (Zhang, 2020). Figure 21 shows several of the algorithmic methods that work under supervised learning according to (Jiang et al. 2020).

Figure 21

Supervised learning methods



Source: (Jiang et al., 2020)

3. Line of code analysis

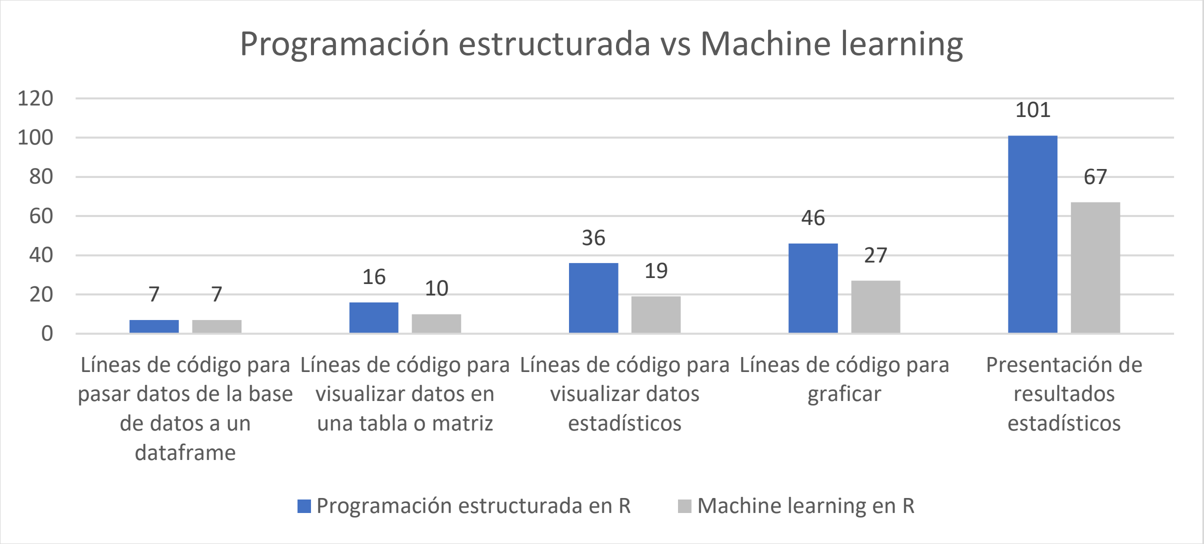
Table 4
Comparing Lines of Code in
Structured Programming vs. Machine Learning

Remarks	**Structured programming in R**	**Machine learning en R**
Lines of code for passing data from the database to a dataframe	7	7
Lines of code for displaying data in a table or array	16	10
Lines of code for displaying statistical data	36	19
Lines of code for graphing	46	27
Presentation of statistical results	101	67

Source: Authors

As can be seen in Table 4 to be able to perform the same situations, more lines of code are used in structured programming than with machine learning, this because R has packages and functions that help reduce these lines. Under this parameter, students adapted better to machine learning than with structured programming to obtain the same results.

Figure 22
Comparison between Structured Programming and Machine Learning



Source: Authors

Results of the elaboration of the work - Table 5 presents the results with the population studied that are the students of the third semester of the subject of statistical programming:

Table 5
Results of the preparation of the work
by third semester students

Observation	**Population**	**Quantityof studies**	**Concludes**	**Quantityof studies**	**Does not conclude**
Structured Programming	19	7	36,84%	12	63,16%
Machine Learning	19	16	84,21%	3	15,79%

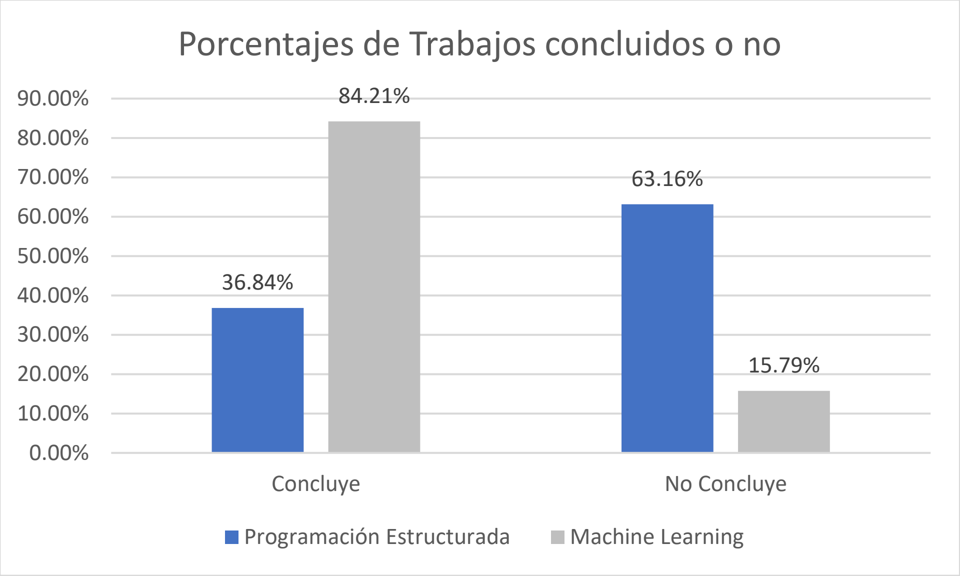
Source: Author

Table 5shows the percentages where students conclude their work using structured programming with 36.84% and using machine learning conclude with 84.21%. It is also reflected that 63.16% do not conclude using structured programming and 15.79 do not conclude using machine learning.

Figure 23showsthe percentages of how they conclude or not the work of presandntation of statistics to know if a student is exonerated or not.

Figure 23

Students conclude their work



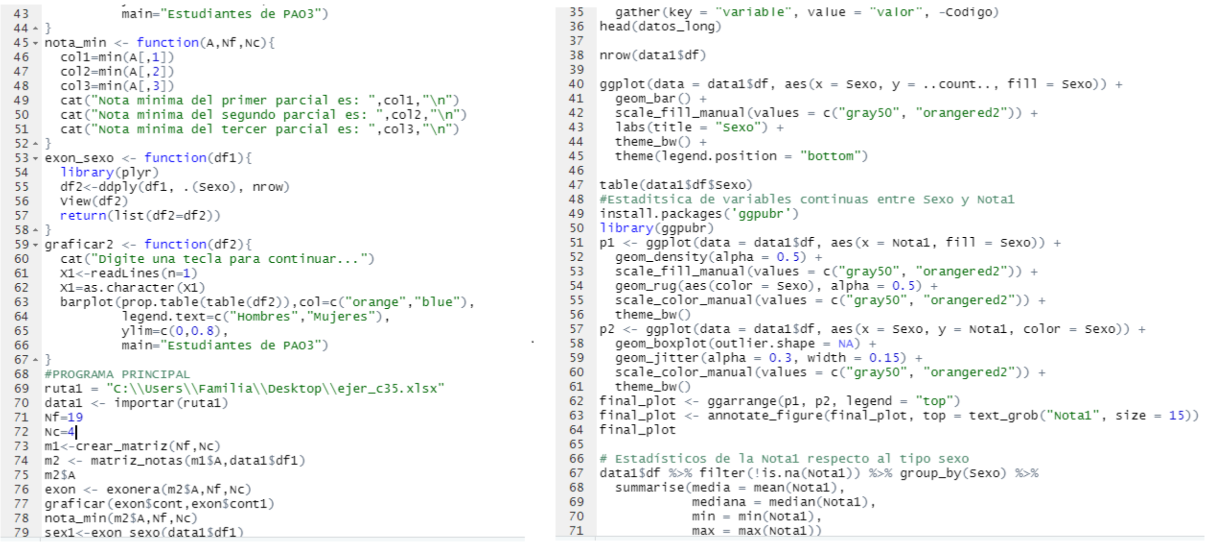
Source: Author

4. Discussion

The presentation of statistics to know if a student can be exonerated or not using structured programming and machine learning was carried out under the methodology As shown in figure 24:

Figure 24

Structured programming (left), machine learning (right)



```
          main="Estudiantes de PAO3")
}
nota_min <- function(A,Nf,Nc){
  col1=min(A[,1])
  col2=min(A[,2])
  col3=min(A[,3])
  cat("Nota minima del primer parcial es: ",col1,"\n")
  cat("Nota minima del segundo parcial es: ",col2,"\n")
  cat("Nota minima del tercer parcial es: ",col3,"\n")
}
exon_sexo <- function(df1){
  library(plyr)
  df2<-ddply(df1, .(Sexo), nrow)
  View(df2)
  return(list(df2=df2))
}
graficar2 <- function(df2){
  cat("Digite una tecla para continuar...")
  X1<-readLines(n=1)
  X1=as.character(X1)
  barplot(prop.table(table(df2)),col=c("orange","blue"),
          legend.text=c("Hombres","Mujeres"),
          ylim=c(0,0.8),
          main="Estudiantes de PAO3")
}
#PROGRAMA PRINCIPAL
ruta1 = "C:\\Users\\Familia\\Desktop\\ejer_c35.xlsx"
data1 <- importar(ruta1)
Nf=19
Nc=4
m1<-crear_matriz(Nf,Nc)
m2 <- matriz_notas(m1$A,data1$df1)
m2$A
exon <- exonera(m2$A,Nf,Nc)
graficar(exon$cont,exon$cont1)
nota_min(m2$A,Nf,Nc)
sex1<-exon_sexo(data1$df1)
```

```
  gather(key = "variable", value = "valor", -Codigo)
head(datos_long)

nrow(data1$df)

ggplot(data = data1$df, aes(x = Sexo, y = ..count.., fill = Sexo)) +
  geom_bar() +
  scale_fill_manual(values = c("gray50", "orangered2")) +
  labs(title = "Sexo") +
  theme_bw() +
  theme(legend.position = "bottom")

table(data1$df$Sexo)
#Estaditsica de variables continuas entre Sexo y Nota1
install.packages('ggpubr')
library(ggpubr)
p1 <- ggplot(data = data1$df, aes(x = Nota1, fill = Sexo)) +
  geom_density(alpha = 0.5) +
  scale_fill_manual(values = c("gray50", "orangered2")) +
  geom_rug(aes(color = Sexo), alpha = 0.5) +
  scale_color_manual(values = c("gray50", "orangered2")) +
  theme_bw()
p2 <- ggplot(data = data1$df, aes(x = Sexo, y = Nota1, color = Sexo)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(alpha = 0.3, width = 0.15) +
  scale_color_manual(values = c("gray50", "orangered2")) +
  theme_bw()
final_plot <- ggarrange(p1, p2, legend = "top")
final_plot <- annotate_figure(final_plot, top = text_grob("Nota1", size = 15))
final_plot

# Estadísticos de la Nota1 respecto al tipo sexo
data1$df %>% filter(!is.na(Nota1)) %>% group_by(Sexo) %>%
  summarise(media = mean(Nota1),
            mediana = median(Nota1),
            min = min(Nota1),
            max = max(Nota1))
```

Source: Author

On the left side you can see lines of code with structured programming and on the right side you can see lines of code with machine learning. As can be seen, there is a similar simile with respect to the form of presentar statistics. But the part of own packages that R has with respect to machine learning minimizes the source code.

A small problem arises when you do not know how to size the variables well, since the use of machine learning appropriates a standard measure within R using its own packages and presenting the proper functions that each one brings. It is necessary to foresee that each dataframe createsor in R fulfills its desired terms; that is why the situation of making uniquedataframes that are our data stores within Structured Programming and machine learning was taken.
As future work it is proposed to carry out deeper studies with random forest and neural networks in the case of supervised learning and in thecase of unsupervised learning clustering by k-means and method of the sum of squares.

5. Conclusion

- In this article, an analysis of the presentation of statistics has been carried out using structured programming and machine learning with supervised learning in the case of random forest. The lines of code are programmed to present the desired results within R which is an optimal language for this type of process and helps us solve through open source scripts the problem of exoneration of student grades of the subject of statistical programming of ESPOCH.
- The use of random forest is of great help for this type of learning within machine learning, since it helps us in making decisions with a multi-dimensional graph and with desired results.
- There were multiple problems when developing the data frames with structured programming to move from our database to a data warehouse within RStudio, but along the way they were solved.
- Students when using structured programming only 36.84% conclude , while students when using machine learning with supervised learning and the random forest method reach 84.21% of them, we also have that students using structured programming do not conclude with 63.16% And with machine learning it does not conclude with 15.79%.
- Random forest shows us that students if they want to be exonerated they must raise the level and their grade from the first partial, but if in two partials they do not rise from the average they could even stay up to semester.
- Regarding lines of code as shown in Table 4, it greatly exceeds the use of functions, procedures, conditional and cyclic control structures within structured programming to the use of functions with Machine learning within R. Even minimizing codein an exaggerated way with structured programming, machine learning greatly surpasses with few lines of code to present the same results.
- The importance of having software like R and that allows us to program in different ways makes it a high-level language for this field of statistics and presentation of results for decision making.

Reference

[1] *A. Núñez Reiz, M.A. Armengol de la Hoz, M. Sánchez García. (2019). Big Data Analysis and Machine Learning in Intensive Medicine. Intensive Medicine. Volume 43, Issue 7, Pages 416-426. ISSN 0210-5691.*

[2] *https://doi.org/10.1016/j.medin.2018.10.007*

[3] *Amat, J. (2020). Machine Learning con R y caret. https://www.cienciadedatos.net/documentos/41_machine_learning_con_r_y_caret*

[4] *Avello, R. & Seisdedo, A. (2017). Statistical processing with R in scientific research. Medisur - Internet Magazine, 15(5), 583-586.*

[5] *RStudio Team. (2020). RStudio: integrated development for R. RStudio, PBC, Boston, MA http://www.rstudio.com/*

*[6] *D.H. Kim, T. MacKinnon. (2018). Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clinical Radiology. Volume 73, Issue 5, Pages 439-445, ISSN 0009-9260,*

[7] https://doi.org/10.1016/j.crad.2017.11.015

[8] Hernández, F. (2021). Modelos Predictivos. Bookdown: Authoring Books and Technical Documents with r Markdown. https://fhernanb.github.io/libro_mod_pred/index.html

[9] Hernandez, F. & Usuga, O. (2021). Manual of R. https://fhernanb.github.io/Manual-de-R/

[10] Jiang, T., Gradus, J. L., yRosellini, A. J. (2020). Supervised Machine Learning: A Brief Primer. Behavior Therapy, 51(5), 675–687.

[11] https://doi.org/10.1016/j.beth.2020.05.002

*[12] *Kai-Qi Li, Yong Liu, Qing Kang. (2022). Estimating the thermal conductivity of soils using six machine learning algorithms. International Communications in Heat and Mass Transfer. Volume 136, 106139, ISSN 0735-1933.*

[13] https://doi.org/10.1016/j.icheatmasstransfer.2022.106139

*[14] *Manisha Koranga, Pushpa Pant, Tarun Kumar, Durgesh Pant, Ashutosh Kumar Bhatt, R.P. Pant. (2022). Efficient water quality prediction models based on machine learning algorithms for Nainital Lake, Uttarakhand, Materials Today: Proceedings. Volume 57, Part 4, Pages 1706-1712. ISSN 2214-7853.*

[15] https://doi.org/10.1016/j.matpr.2021.12.334

[16] Mejia, F., Rosero, R., Luna, W. and Villa, E. (2018). Methodology for Building an Algorithm for Systemic Learning of First Semester Students of the ICT Subject. Engineering KnE, 3(9), 221–234.

[17] https://doi.org/10.18502/keg.v3i9.365

[18] Microsoft. (2022). Basic tasks in Excel. Microsoftsupport.

[19] https://support.microsoft.com/es-es/office/tareas-b%C3%A1sicas-en-excel-dc775dd1-fa52-430f-9c3c-d998d1735fca

[20] Orellana, J. (2019). What is machine learning and why is it so popular? University of Cuenca. https://www.ucuenca.edu.ec/component/content/article/233-espanol/investigacion/blog-de-ciencia/1222-machine-learning

[21] R Core Team. (2020). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

*[22] *SafouraJahedizadeh, AfsanehGhanizadeh. (2021). Continuous flow and better personal outcomes in higher education: combining methods. Journal of Psychodidactics. Volume 26, Issue 2, Pages 96-104. ISSN 1136-1034.*

[23] https://doi.org/10.1016/j.psicod.2020.11.006

[24] Santana, A. & Hernández, C. (2016). Programming in R. Department of Mathematics, ULPGC. https://estadistica-dma.ulpgc.es/cursoR4ULPGC/15-programacionR.html

[25] Sierra, Y. (2020). What is machine learning and what is it for? Lemontech blog - technology. https://blog.lemontech.com/que-es-el-machine-learning-y-para-que-sirve-ejemplos/

[26] Vargas, L. & Mesa, E. (2021). Introduction to data analysis with RStudio. Estudio 45-8 S.A.S

*[27] *Xin Li, Yang Wen, Jiaojiao Jiang, Tugrul Daim, Lucheng Huang. (2022). Identifying potential breakthrough research: A machine learning method using scientific papers and Twitter data. Technological Forecasting and Social Change. Volume 184, 122042. ISSN 0040-1625.*

[28] https://doi.org/10.1016/j.techfore.2022.122042

[29] Zhang, X.-D. (2020). Machine learning. In A Matrix Algebra Approach to Artificial Intelligence (pp. 223-440). Springer.