High School Graduation Prediction

Maria Belen Acosta Vera

2022

Contents

Chapter 1

Summary

In this machine learning project we used two models: SVM and Decision Tree to predict high-school graduation in Paraguay. We used the 2019 and 2020 Continuous Permanent House Poll (EPHC) from the National Institute of Statistics of Paraguay. As a result, we obtained an accuracy higher than 80% and an F1 score higher than 85%. With this, we can say that socioeconomic factors such as level of poverty, language spoken at home most of the time, area and age are relatively significant variables to explain whether or not someone will graduate high-school.

Chapter 2

Introduction

Education is a human right that shouldn't be denied to anyone. Even still, there are still many who cannot exercise their rights. Such a consequence must have identifiable causes. Many researches[1] suggest that it is hard to generalize which factors influence student performance.

In this project, we'll focus on the socioeconomic factors to understand the influence the environment has in high-school graduation.

Paraguay is a South-American country with about 7 million people, with large social inequality that is easily seen in its educational system: the quality of education is low and this can be proved with the results of standardized tests: Paraguayan students are not learning what they should learn, and that's if they stay within the educational system. There's high dropout.

With this in mind, the need to prepare a Capstone project to achieve the Data Science Professional Certificate from HarvardX became a perfect opportunity to create a model that predicts whether a Paraguayan student will graduate high-school or not. In this project, we try to determine this given the language they speak most of the time, the area where they live, their age and whether or not they're considered poor.

For this, I've used the dataset of the 2019 and 2020 Continuous Permanent House Poll (EPH) from the National Institute of Statistics (INE) of Paraguay, since it has some variables that influence academic success or graduation.

[1] https://www.mec.gov.py/cms_v2/adjuntos/9163

Chapter 3

Data Analysis

Initially, we'll load all the packages we're going to use

3.1 Load packages

```
if(!require(caret)) install.packages("caret", repos="http://cran.us.r-project.org")
if(!require(e1071)) install.packages("e1071", repos="http://cran.us.r-project.org")
if(!require(haven)) install.packages("haven", repos="http://cran.us.r-project.org")
if(!require(ggpmisc)) install.packages("ggpmisc", repos="http://cran.us.r-project.org")
if(!require(ggplot2)) install.packages("ggplot2", repos="http://cran.us.r-project.org")
if(!require(ggrepel)) install.packages("ggrepel", repos="http://cran.us.r-project.org")
if(!require(ggthemes)) install.packages("ggthemes", repos="http://cran.us.r-project.org")
if(!require(grid)) install.packages("grid", repos="http://cran.us.r-project.org")
if(!require(gridExtra)) install.packages("gridExtra", repos="http://cran.us.r-project.org")
if(!require(kableExtra)) install.packages("kableExtra", repos="http://cran.us.r-project.org'
if(!require(MLmetrics)) install.packages("MLmetrics", repos="http://cran.us.r-project.org")
if(!require(rpart)) install.packages("rpart", repos="http://cran.us.r-project.org")
if(!require(stringr)) install.packages("stringr", repos="http://cran.us.r-project.org")
if(!require(tidyverse)) install.packages("tidyverse", repos="http://cran.us.r-project.org")
```

3.2 Load database

Here, we'll load the 2019 and 2020 Continuous Permanent House Poll (EPH) from the National Institute of Statistics (INE) from Paraguay.

```r
set.seed(1, sample.kind = "Rounding")
#2019
ephc2019 <- read_sav("reg02_ephc2019.sav")

#2020
ephc2020 <- read_sav("reg02_ephc2020.sav")
```

3.3 Variables in the data set

```r
ncol(ephc2019)
```

```
## [1] 260
```

```r
nrow(ephc2019)
```

```
## [1] 18233
```

```r
ncol(ephc2020)
```

```
## [1] 263
```

```r
nrow(ephc2020)
```

```
## [1] 17582
```

As we can see, the EPH dataset has many variables. Each year, it gathers information related to the well-being of Paraguayan households such as education, health, employment and income. For 2019, we have **260** variables and **18233** observations.

For 2020, we have three more columns (giving a total of **263**) because there were some questions related to the COVID-19 pandemic. There are **17582** observations.

Now, we will create a subset of both databases to simplify our work. The variables I'm keeping are crucial for our analysis and subsequent prediction model. Here is a brief description of each variable.

- AREA <dbl>: geographic region in which the person lives. "Urbana" means urban areas and "Rural" means rural areas. Urban is represented as 1 on the database and Rural as 6.

- P02 <dbl>: age.
- P03 <dbl>: 1 - Household Head, 2 - Wife, partner, 3 - Son, 4 - Step-son, 5 - Grandson, 6 - daughter/son-in-law, 7 - Father/Mother, 8 - Father/Mother-in-law, 9 - Other relative, 10 - Non-relative, 11 - housekeeper, 12 - housekeeper's family.
- P06 <dbl>: gender, 1 - male, 2 - female.
- E01A <dbl>: main income.
- ED01 <dbl>: language spoken at home most of the time. 1 - Guaraní, 2 - Guaraní and Spanish, 3 - Spanish, 4 - Another language, 5 - doesn't speak, 9 - doesn't answer.
- ED02 <dbl>: literacy. 1 - can read and write, 6 - CANNOT read and write, 9 - doesn't answer.
- ED0504 <dbl>: approved level and grade.
- añoest <dbl>: years of study
- ED09 <dbl>: sector of the institution that they attend. 1 - Public, 2 - Private, 3 - Private subsidized, 9 - doesn't answer.
- pobnopoi <dbl>: poverty condition. 0 - non-poor, 1 - poor.

Let's create our train and validation sets.

```
# The data from 2019 will be our train set
eph_2019 <- ephc2019 %>%
  select(AREA, P02, P03, P06, E01A,
         ED01, ED02, ED0504, añoest, ED09, pobnopoi)

# The data from 2020 will be our validation set
eph_2020 <- ephc2020 %>%
  select(AREA, P02, P03, P06, E01A,
         ED01, ED02, ED0504, añoest, ED09, pobnopoi)


# Let's remove the NA values, since they will hinder our prediction.
eph_2019 <- eph_2019 %>%
  filter(!is.na(AREA)) %>%
  filter(!is.na(añoest)) %>%
  filter(!is.na(P02)) %>%
  filter(!is.na(pobnopoi)) %>%
  filter(!is.na(ED01))
```

```r
eph_2020 <- eph_2020 %>%
  filter(!is.na(AREA)) %>%
  filter(!is.na(añoest)) %>%
  filter(!is.na(P02)) %>%
  filter(!is.na(pobnopoi)) %>%
  filter(!is.na(ED01))

#For now, we'll remove the full datasets so that our environment is less cluttered.
rm(ephc2019, ephc2020)
```

3.4 Data preparation

After deleting all the NA, both databases consist of 11 columns. `eph_2019` has 16,611 observations and `eph_2020` has 16,096. Now, we're going to create a variable that groups people into two categories: those who finished high-school (">=12", because to finish highschool one has to at least complete 12 years of education), and those who didn't ("<12").

We're naming this variable `glyst_hs`, which stands for 'graduation: last year of study - high school'. It's important to mention that we delete the years equal to "99" in `añoest`, because 99 means that there was no answer.

We're also creating a variable called `graduate` so that we can classify our observations according to the structure of education in Paraguay. It's as follows:

Basic School Education (Educación Escolar Básica - EEB):

It is organized in three cycles of three years each: First cycle (1st, 2nd and 3rd grade), Second cycle (4th, 5th and 6th grade) and Third cycle (7th, 8th and 9th grade). It is compulsory and free in official management institutions. We'll divide it into two groups `EEB_1_2` and `EEB3`. The reason for this division is that, historically, there's an inflection point of sorts there and it's relevant for comparison analysis. Similar inflection points also occur at the end of the third cycle (9th grade) and our main point of interest, which is 12th grade or third year of secondary education - the end of high-school.

Secondary Education (Educación Media - EM):

Secondary Education lasts three years, consisting of three courses. It is offered in the modalities of Scientific and Technical Baccalaureate and is intended for the attention of the population between 15 and 17 years of age. It is compulsory and free in official management institutions as of 2010. In our database, it's named `EM`.

Higher Education (Educación Superior - ES):

Higher Education is developed through universities, higher institutes and other third-level professional training institutions, Teacher Training Institutes (IFD)

and technical institutes. Includes university and non-university degrees. In our database it's named `HED`.

Source: Public Financing of Education in Paraguay, notes for the debate and construction of public policies[1]

```
# Create 'glyst_hs' variable
eph_2019<- eph_2019 %>%
  filter(!is.na(añoest)) %>%
  filter(!añoest %in% c("99")) %>%
  mutate(glyst_hs = ifelse(añoest %in%
                             c("1","2","3","4","5","6", "7", "8", "9", "10", "11"), "<12",
                           ifelse(añoest %in%
                                    c("12", "13", "14", "15", "16", "17", "18"), ">=12", "<1
eph_2019 <- eph_2019 %>% mutate(glyst_hs = as.factor(glyst_hs))

eph_2020<- eph_2020 %>%
  filter(!is.na(añoest)) %>%
  filter(!añoest %in% c("99")) %>%
  mutate(glyst_hs = ifelse(añoest %in%
                             c("1","2","3","4","5","6", "7", "8", "9", "10", "11"), "<12",
                           ifelse(añoest %in%
                                    c("12", "13", "14", "15", "16", "17", "18"), ">=12", "<1
eph_2020 <- eph_2020 %>% mutate(glyst_hs = as.factor(glyst_hs))

#Creating the 'graduate' variable
eph_2019 <- eph_2019 %>%
  filter(!is.na(añoest)) %>%
  mutate(graduate = ifelse(añoest %in% c("12"), "HSgrad",
                           ifelse(añoest %in% c("1","2","3","4","5","6"), "EEB_1_2",
                                  ifelse(añoest %in% c("7", "8", "9"), "EEB3",
                                         ifelse(añoest %in% c("10", "11"), "EM",
                                                ifelse(añoest %in%
                                                         c("13", "14", "15","16", "17", "18"


eph_2020 <- eph_2020 %>%
  filter(!is.na(añoest)) %>%
  mutate(graduate = ifelse(añoest %in%
                             c("12"), "HSgrad",ifelse(añoest %in%
                                                        c("1","2","3","4","5","6"), "EEB_1_2
                                                      ifelse(añoest %in% c("7", "8", "9"), "
                                                             ifelse(añoest %in% c("10", "11"
                                                                    ifelse(añoest %in% c("13
```

[1] https://observatorio.org.py/uploads/report_file/url/43/1578060133-Financiamiento_p%C3%BAblico_de_la_educaci%C3%B3n_en_el_Paraguay.pdf

Chapter 4

Data exploration

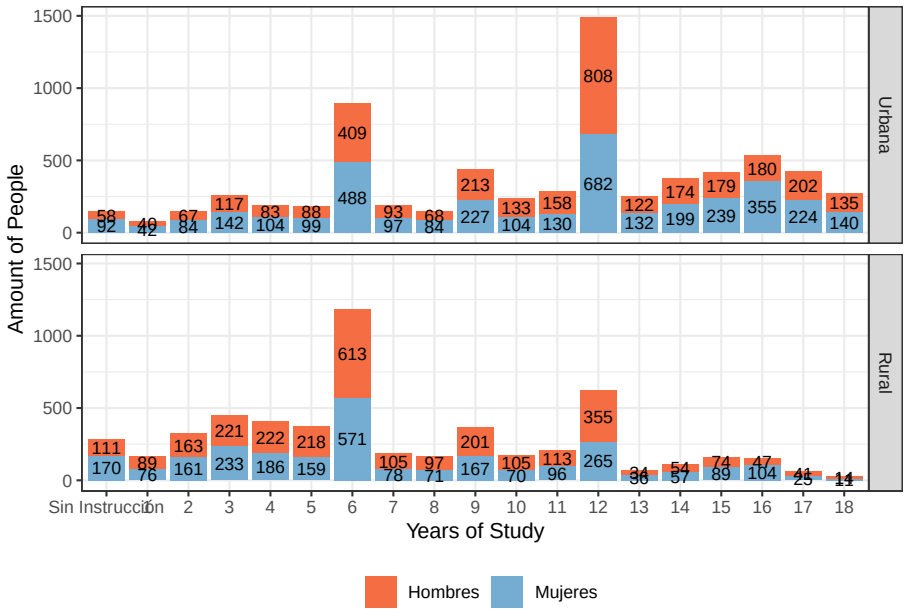
It's always important to see how our data is distributed. The following graphs will help us get a better picture of the Paraguayan education system.

4.1 Characteristics of high-school graduates.

To begin, let's see how many years of study does a Paraguayan has in average. We'll filter by those who are 17 years old or higher, since that's the part of our population that we're interested in.

```
#Average years of study
avg_yr_stdy <- eph_2019 %>%
  filter(P02 >= 17)%>%
  group_by(as_factor(añoest), as_factor(P06), as_factor(AREA)) %>%
  summarise(amount = n())

avg_yr_stdy %>%
  ggplot(aes(x=`as_factor(añoest)`, y = amount, fill = `as_factor(P06)`, label = amount)) +
  geom_bar(stat = "identity",position = "stack") +
  xlab("Years of Study") +
  ylab("Amount of People") +
  labs(fill = "") +
  geom_text(size=3, position = position_stack(vjust = 0.5)) +
  scale_fill_manual(values=c("#F46D43", "#74ADD1")) +
  theme_bw() + facet_grid(`as_factor(AREA)`~.) +
  theme(legend.position = "bottom")
```



Most people over the age of 17 study between six and twelve years. This means that they at least finish sixth grade. We can see that on rural areas, there's more people whose last degree obtained is sixth grade, followed by twelfth grade - finishing high school.

On urban areas, it's the opposite, most seventeen-or-older people finish high-school, followed by sixth grade. From this, we could infer that there's significantly more access to higher education on urban areas.

As mentioned in Public Financing of Education in Paraguay, notes for the debate and construction of public policies[1], **in general, it's harder for women of rural areas to access education.** The mentioned study refers to 15-year-old students and older, but with this graph we can see that it also holds for seventeen-and-older students.

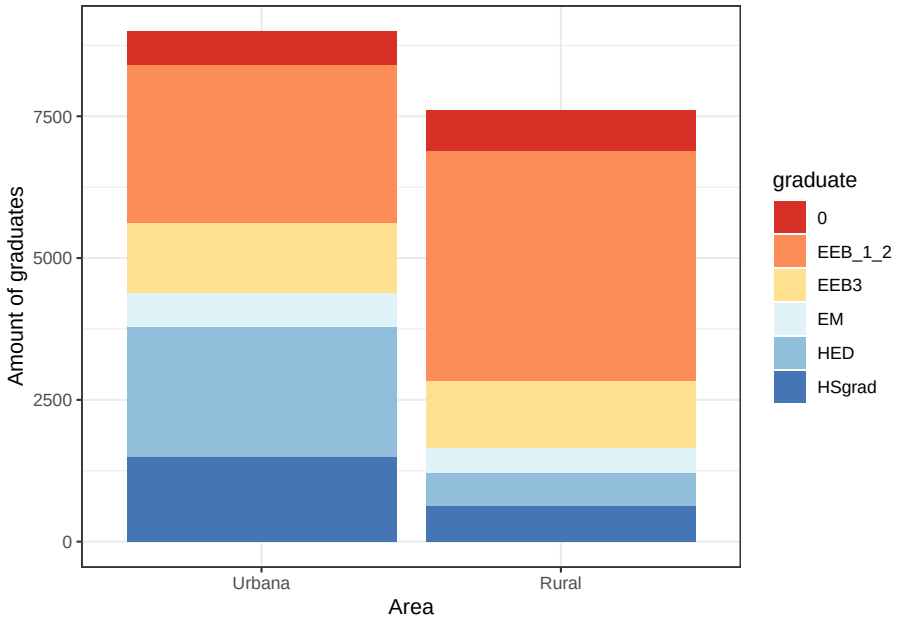
It's important to mention that this study from 2014[2] gives reasons as to why there's this huge differene between urban and rural areas. In the last ones, most activities correspond to agricultural activities, where children between the ages of 10 to 17 work an average of 35.3 hours per week. Undoubtedly, this leads to delayed schooling resulting in an average of 5.1 to 6.4 years of study, which is verifiable in our graph.

We know how many years Paraguayans study in average, let's move on to what's their highest level of education achieved by area.

[1] https://observatorio.org.py/uploads/report_file/url/43/1578060133-Financiamiento_p%C3%BAblico_de_la_educaci%C3%B3n_en_el_Paraguay.pdf

[2] https://mec.gov.py/talento/archivo/materiales-concurso-sup-2014/modulos-manuales/mod_politica.pdf

```r
#Graduation by area
eph_2019 %>%
  group_by(graduate) %>%
  ggplot(aes(as_factor(AREA)))+
  geom_bar(aes(fill=graduate)) +
  labs(x = "Area", y ="Amount of graduates") +
  scale_fill_brewer(palette = "RdYlBu")+
  theme_bw()
```

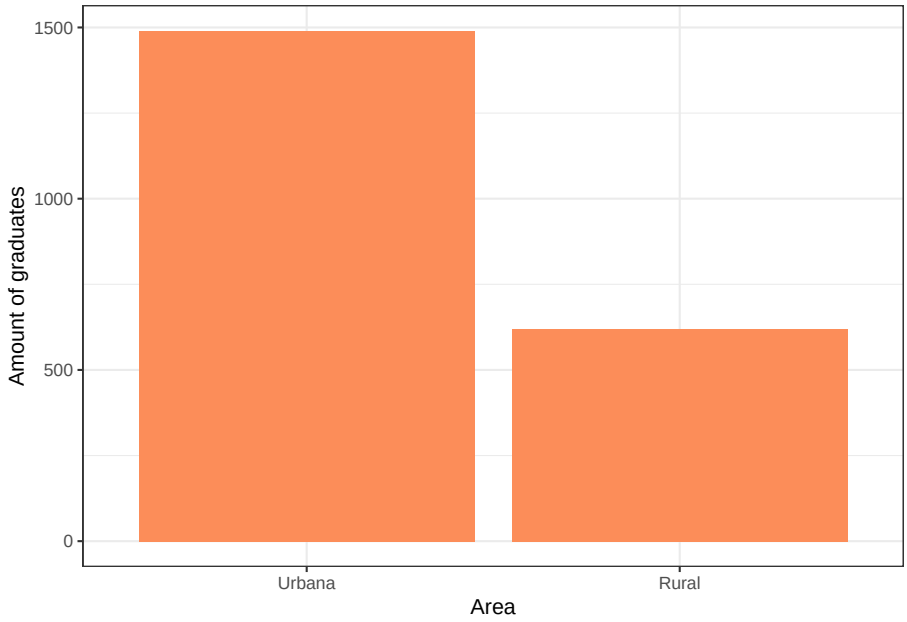


This agrees with what we've seen in the previous graph: **most people from rural areas study, at most, six years, whereas those who live in urban areas often reach secondary education and many of them finish high-school.** As it was mentioned before, Paraguay is a country with a lot of inequality. In comparison, more people from urban areas reach Higher Education than those from rural areas.

For the next, graph, let's focus only on high-school graduation.

```r
#High School graduation per area
eph_2019 %>%
  group_by(graduate) %>%
  filter(añoest %in% c("12")) %>%
  ggplot(aes(as_factor(AREA)))+
  geom_bar(aes(fill=graduate), show.legend = FALSE) +
  labs(x = "Area", y ="Amount of graduates") +
```

```
  scale_fill_brewer(palette = "RdYlBu")+
  theme_bw()
```



This graph is very enlightening in that it shows that the high-school graduates from urban areas more than double those from rural areas.

Let's add the poverty variable into the mix and see what it can tell us.

```
#Poverty Level
eph_2019 %>%
  group_by(as_factor(pobnopoi)) %>%
  filter(añoest ==  12) %>%
  count(pobnopoi)
```

```
## # A tibble: 2 x 3
## # Groups:   as_factor(pobnopoi) [2]
##   `as_factor(pobnopoi)`     pobnopoi     n
##   <fct>                    <dbl+lbl> <int>
## 1 NO POBRE              0 [NO POBRE]  1849
## 2 POBRE                 1 [POBRE]      261
```

As expected, **poverty plays a huge role in high-school graduation.** Of the 2,110 high-school graduates in 2019, 87.6% were considered not poor. In Paraguay, poverty is calculated on whether or not the person can at least afford the Basic Food and Non-Food Basket (values for urban and rural areas mentioned below).

In 2019, someone from urban areas was considered poor if they earned 699,634 guaraníes (Paraguay's currency - PYG) or less. This equals to 112 USD, roughly.

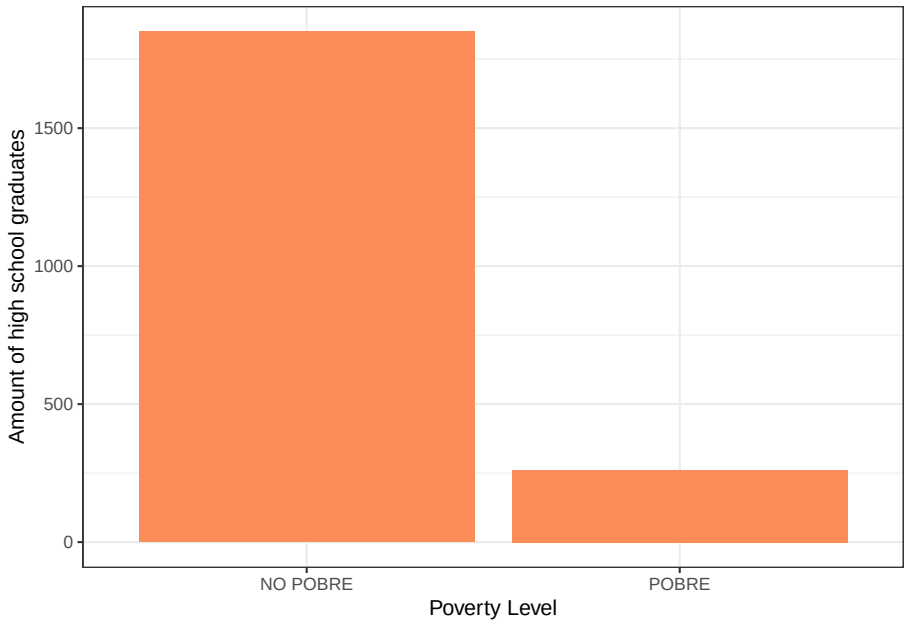
Someone from rural areas was considered poor if they earned 497,049 guaraníes or less, which is approximately 79 dollars.

The amount in dollars was calculated with an exchange rate of 6,237 guaraníes per USD, an average of the exchange rate of PYG/USD in 2019.

This data was obtained from the National Institute of Statistics of Paraguay, specifically from the document "Valores mensuales (guaraníes) de la línea de pobreza extrema y pobreza total por área de residencia, según año. Periodo: 1997/98-2020"[3].

This difference in poor and non-poor graduates is best seen in the next graph. "NO POBRE" means "not poor" and "POBRE" means poor.

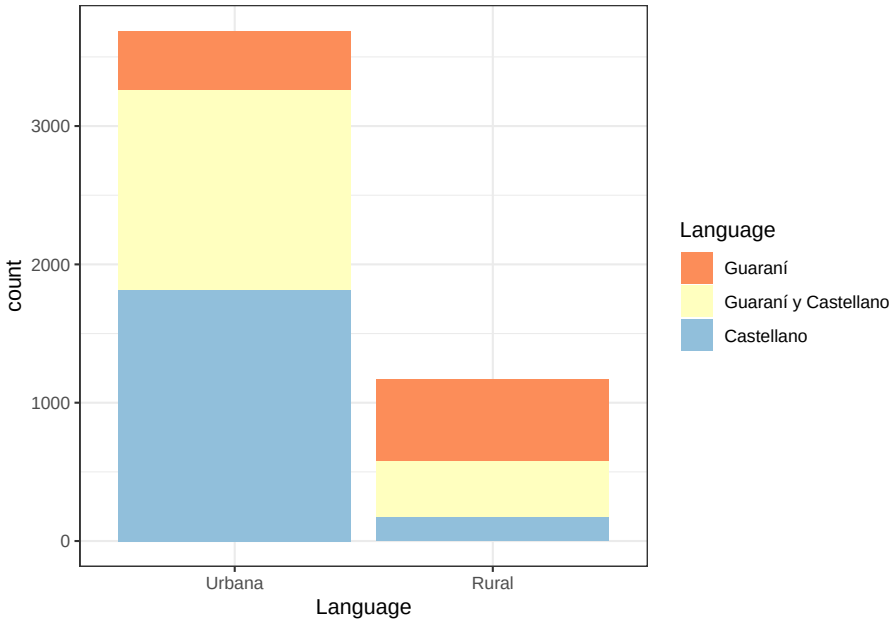
```
#High-school graduates by poverty level.
eph_2019 %>%
  group_by(as_factor(pobnopoi)) %>%
  filter(añoest %in% c("12")) %>%
  ggplot(aes(as_factor(pobnopoi))) +
  geom_bar(aes(fill= "#FC8C58"),show.legend = FALSE) +
  labs(x = "Poverty Level", y ="Amount of high school graduates") +
  scale_fill_brewer(palette = "RdYlBu") +
  theme_bw()
```



[3] https://www.ine.gov.py/default.php?publicacion=4

It's obvious that when someone's poor, they have little to no chances of finishing high-school, even when it's supposed to be free. Another element that could be significant for high-school graduation is the language spoken at home most of the time. Let's see how it behaves according to the area.

```
#Area and language spoken
eph_2019 %>%
  filter(glyst_hs == ">=12") %>%
  filter(ED01 %in% c("1","2","3")) %>%
  ggplot(aes(x=as_factor(AREA)))+
  geom_bar(aes(fill=as_factor(ED01))) +
  xlab("Language") +
  labs(fill="Language") +
  scale_fill_brewer(palette = "RdYlBu") +
  theme_bw()
```



Disregarding the amount of graduates we can see that, according ,to the language spoken at home most of the time, urban and rural areas are complete opposites. In urban areas, more than half of the graduates speak Spanish (Castellano) in some level. Very little of them speak Guaraní the most, whereas in rural areas, the language spoken in some level by more than half of the graduates is Guaraní. A small portion speak Spanish the most.

A researcher[4] once defined Guaraní a majority language with a minority language profile. Minority languages, according to the researcher Eva Núñez (2013,

[4] https://doi.org/10.15665/.v16i01.1401

cited in Von Streber, 2017) *"are intended for the private space with preferably oral transmission and are not valued in the public sector."*

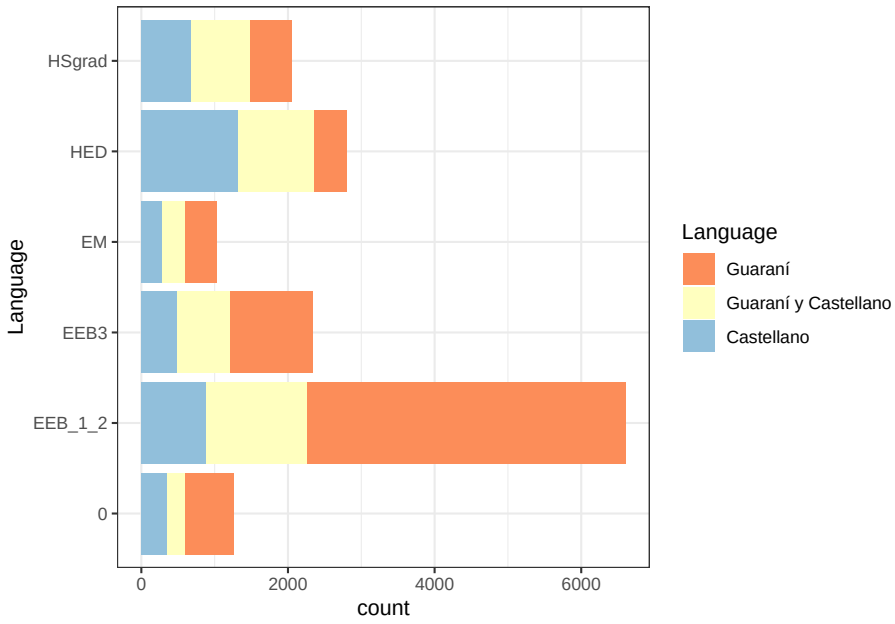
Moreover, there are few and far in between institutions that teach in Guaraní. Paraguayans, no matter their mother tongue, are forced to learn in Spanish even though Guaraní is also an official language, recognized by the Constitution.

According to The Guaraní Language in the Educational System in the last two decades (2014)[5], the existence of educational institutions whose main language is Guaraní is known since the year 1998. For the year 2009, according to the Department of Statistics of the Ministry of Education and Sciences, there were 404 schools that have sections where the main language is Guaraní, which corresponded to a total enrollment of 18,449 students.

Let's see how language influences the last degree obtained.

```
#Last degree obtained and language spoken
eph_2019 %>%
  filter(ED01%in%c("1","2","3")) %>%
  ggplot(aes(graduate)) +
  geom_bar(aes(fill=as_factor(ED01)))+
  labs(fill="Language") + xlab("Language") +
  theme_bw() +
  coord_flip()+
  scale_fill_brewer(palette = "RdYlBu") +
  theme_bw()
```

[5] https://www.mec.gov.py/cms_v2/recursos/9797-la-lengua-guarani-en-el-sistema-educativo-en%20-las-dos-ultimas-decadas



Most of those who study between 1-9 years speak Guaraní the most, specially those who study between 1-6 years. As the years of study increase, so does the predominance of Spanish. This can lead us to hypothesize that to achieve higher education in Paraguay, one must speak Spanish relatively well.

In summary, no matter the area, most people study between 6-12 years. In rural areas, it's more likely that people study up to six years and in urban areas, 12 years. Poverty levels influence greatly in graduation, since mosr people who graduate high-school are NOT considered poor. Language is also important, since as people speak more Spanish, they're more likely to study for more years.

Chapter 5

Models

5.1 Variables used.

We're using some of the variables that we visualized in the previous section. These are: `AREA`, `P02`, `ED01`, `pobnopoi` and `glyst_hs`.

5.2 Load database

Let's load again the full 2019 and 2020 editions of the Continuous Permanent House Poll (EPHC).

```
set.seed(1, sample.kind = "Rounding")

#2019
ephc2019 <- read_sav("reg02_ephc2019.sav")

#2020
ephc2020 <- read_sav("reg02_ephc2020.sav")
```

5.3 Data Preparation

In this section, we'll get our database ready for the modeling. To delete the NA, we'll use a loop that turns all the "Sin instrucción" into zeroes, that way they'll be easier to handle. Also, we're turning all the observations that don't have any information into NA (remember that NA are represented as 9, 99, 999 in our database)

As mentioned, the 2019 data will be our train set, which will again be divided intro train and test sets to tests our models. The 2020 data will be our validation.

```
#Set seed

set.seed(1, sample.kind = "Rounding")

#Select columns and change 'sin instrucción' into 0.

eph_2019 <- eph_2019 %>% select(AREA, añoest, P02, pobnopoi, ED01, glyst_hs)
eph_2019 <- eph_2019 %>% mutate_at(
  vars("AREA", "añoest", "pobnopoi", "ED01", "glyst_hs"),
  funs(as_factor(.))
)
for (i in 1:length(eph_2019$añoest))
  {
  eph_2019$añoest[i]<-ifelse(eph_2019$añoest[i]=="Sin instrucción", 0, eph_2019$añoest[i])
}
eph_2019$añoest<-as.numeric(eph_2019$añoest)
eph_2019$P02<-as.numeric(eph_2019$P02)
eph_2019<- eph_2019 %>%  # Create glyst_hs variable
  filter(!is.na(añoest))

# Here we split the 2019 data into train and test sets
eph_2019 <- eph_2019 %>% select(AREA, añoest, P02, pobnopoi, ED01, glyst_hs)
test_index <- createDataPartition(eph_2019$glyst_hs, times=1, p=0.2, list = F)
train_set <- eph_2019[-test_index,]
test_set <- eph_2019[test_index,]

#Given that our data is imbalanced, we have to 'weight' the models.
model_weights <- ifelse(train_set$glyst_hs == "<12",
                        (1/table(train_set$glyst_hs)[1]) * 0.5,
                        (1/table(train_set$glyst_hs)[2]) * 0.5)
sum(model_weights)#The sum MUST equal 1
```

```
## [1] 1
```

```
rm(test_index)
```

We're going to use two models for our predictions: **SVM Model and Decision Tree**, which will be evaluated using these two metrics: **Accuracy and F1 Score**.

Accuracy is a metric for classification models that measures the number of predictions that are correct as a percentage of the total number of predictions

that are made. Its formula is:

$$\frac{\text{true positives} + \text{true negatives}}{\text{true positives} + \text{false negatives} + \text{true negatives} + \text{true negatives}}$$

F1 Score is a proposed improvement of two simpler performance metrics: Precision and Recall.

Precision counts the percentage that is correct from a prediction. A model is considered *precise* when even if it doesn't find all the positives, the ones that the model does class as positive are very likely to be correct.

Recall can be understood as: within everything that actually is positive, how many did the model succeed to find. A model has *high recall* when succeeds well in finding all the positive cases in the data, even though they may also wrongly identify some negative cases as positive cases. Its formula is the following:

$$2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Source: The F1 Score[1]

5.4 SVM Model

The goal of the Support Vector Machine[2] algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

```
#Applying SVM to our data
svm.adult = svm(glyst_hs ~AREA+P02+pobnopoi+ED01, data = train_set)
test_set$pred.value = predict(svm.adult, newdata = test_set,type="response")
confusionMatrix(test_set$glyst_hs, test_set$pred.value)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  <12 >=12
##       <12  2109  217
```

[1] https://towardsdatascience.com/the-f1-score-bec2bbc38aa6
[2] https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

```
##       >=12  346  590
##
##                Accuracy : 0.8274
##                  95% CI : (0.814, 0.8402)
##     No Information Rate : 0.7526
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5601
##
##  Mcnemar's Test P-Value : 6.869e-08
##
##             Sensitivity : 0.8591
##             Specificity : 0.7311
##          Pos Pred Value : 0.9067
##          Neg Pred Value : 0.6303
##              Prevalence : 0.7526
##          Detection Rate : 0.6465
##    Detection Prevalence : 0.7131
##       Balanced Accuracy : 0.7951
##
##        'Positive' Class : <12
##
```

Let's add the result from the model to a data frame called `results`.

```
results <- data.frame(
  Model="SVM (Support Vector Machine)",
  Accuracy=
    Accuracy(test_set$glyst_hs, test_set$pred.value),
  F1Score=
    F1_Score(test_set$glyst_hs, test_set$pred.value))
results
```

```
##                          Model  Accuracy   F1Score
## 1 SVM (Support Vector Machine) 0.8274065 0.8822422
```

With SVM, we have really good accuracy and F1 score.

5.5 Decision Tree

```
# Applying Decision Tree Model
detree <- rpart(glyst_hs ~
```

```
                  AREA + P02 + pobnopoi + ED01,
                data = train_set)
# Prediction of data and Confusion Matrix
test_set$pred.value2 = predict(detree, newdata = test_set, type="class")
confusionMatrix(test_set$glyst_hs, test_set$pred.value2)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  <12 >=12
##       <12  2082  244
##       >=12  336  600
##
##                Accuracy : 0.8222
##                  95% CI : (0.8086, 0.8352)
##     No Information Rate : 0.7413
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5523
##
##  Mcnemar's Test P-Value : 0.0001577
##
##             Sensitivity : 0.8610
##             Specificity : 0.7109
##          Pos Pred Value : 0.8951
##          Neg Pred Value : 0.6410
##              Prevalence : 0.7413
##          Detection Rate : 0.6383
##    Detection Prevalence : 0.7131
##       Balanced Accuracy : 0.7860
##
##        'Positive' Class : <12
##
```

Let's add it to `results` and compare with the SVM model.

```
results<- bind_rows(
  results,
  data.frame(Model="Decision Tree",
             Accuracy=Accuracy(test_set$glyst_hs, test_set$pred.value2),
             F1Score=F1_Score(test_set$glyst_hs, test_set$pred.value2)))
results
```

```
##                          Model  Accuracy   F1Score
```

```
## 1 SVM (Support Vector Machine) 0.8274065 0.8822422
## 2                Decision Tree 0.8221950 0.8777403
```

Even if the results are slightly lower, they're still very good. Let's see our models on the validation set.

Chapter 6

Validation

We'll proceed in the same way as we did with the test set (2019 data) for our validation (2020 data)

6.1 Data Preparation

```
#Set seed

set.seed(1, sample.kind = "Rounding")

#Select columns

eph_2020 <- eph_2020 %>% select(AREA, añoest, P02, pobnopoi, ED01, glyst_hs)
eph_2020 <- eph_2020 %>% mutate_at(
  vars("AREA", "añoest", "pobnopoi", "ED01", "glyst_hs"),
  funs(as_factor(.))
)
for (i in 1:length(eph_2020$añoest))
  {
  eph_2020$añoest[i]<-ifelse(eph_2020$añoest[i]=="Sin instrucción", 0, eph_2020$añoest[i])
}
eph_2020$añoest<-as.numeric(eph_2020$añoest)
eph_2020$P02<-as.numeric(eph_2020$P02)
eph_2020<- eph_2020 %>%  # Create glyst_hs variable
  filter(!is.na(añoest))


#Add weights to our imbalanced data
```

```
eph_2020 <- eph_2020 %>% select(AREA, añoest, P02, pobnopoi, ED01, glyst_hs)
model_weights <- ifelse(eph_2020$glyst_hs == "<12",
                        (1/table(eph_2020$glyst_hs)[1]) * 0.5,
                        (1/table(eph_2020$glyst_hs)[2]) * 0.5)
sum(model_weights)#The sum MUST equal 1
```

```
## [1] 1
```

6.2 SVM Model

```
svm.adult = svm(glyst_hs ~
                  AREA +  P02 + pobnopoi + ED01,
                data = eph_2019)
eph_2020$pred.value = predict(svm.adult, newdata = eph_2020,type="response")
ConfusionMatrix(eph_2020$glyst_hs, eph_2020$pred.value)
```

```
##       y_pred
## y_true   <12  >=12
##   <12  10224  1834
##   >=12  1095  2702
```

```
results<- bind_rows(
  results,
  data.frame(Model="Validation - SVM",
             Accuracy=Accuracy(eph_2020$glyst_hs, eph_2020$pred.value),
             F1Score=F1_Score(eph_2020$glyst_hs, eph_2020$pred.value)))
results
```

```
##                          Model  Accuracy   F1Score
## 1 SVM (Support Vector Machine) 0.8274065 0.8822422
## 2                Decision Tree 0.8221950 0.8777403
## 3             Validation - SVM 0.8152633 0.8747059
```

6.3 Decision Tree

```
# Applying Decision Tree Model
detree <- rpart(glyst_hs ~
                  AREA  + P02 + pobnopoi + ED01,
```

```
                data = eph_2019)
# Prediction of data and Confusion Matrix
eph_2020$pred.value2 = predict(detree, newdata = eph_2020, type="class")
ConfusionMatrix(eph_2020$glyst_hs, eph_2020$pred.value2)
```

```
##        y_pred
## y_true   <12  >=12
##   <12  10112  1719
##   >=12  1207  2817
```

```
results<- bind_rows(
  results,
  data.frame(Model="Validation - Decision Tree",
             Accuracy=Accuracy(eph_2020$glyst_hs, eph_2020$pred.value2),
             F1Score=F1_Score(eph_2020$glyst_hs, eph_2020$pred.value2)))
results
```

```
##                           Model  Accuracy   F1Score
## 1 SVM (Support Vector Machine) 0.8274065 0.8822422
## 2                Decision Tree 0.8221950 0.8777403
## 3             Validation - SVM 0.8152633 0.8747059
## 4   Validation - Decision Tree 0.8154525 0.8736069
```

Once again, Accuracy and F1 Score decrease in marginal quantities. This could be due to a number of reasons, one of which could be the recent pandemic. Still, this doesn't mean that our scores are bad, since for both models we have higher than 80% accuracy and an F1 Score higher than 85%.

Chapter 7

Results

In fact, we could see that the accuracy and F1 score for both models decreased when modeling on the validation set. This could be due to the pandemic, which influenced societies and education in many ways and wasn't taken into account for our models.

```
results
```

```
##                           Model  Accuracy   F1Score
## 1 SVM (Support Vector Machine) 0.8274065 0.8822422
## 2                Decision Tree 0.8221950 0.8777403
## 3             Validation - SVM 0.8152633 0.8747059
## 4   Validation - Decision Tree 0.8154525 0.8736069
```

According to the F1 score and accuracy, the SVM model is marginally better than the decision tree. Still, we can say that the variables we chose are relatively good indicators to predict high-school graduation.

Chapter 8

Conclusion

There are many factors which influence high-school graduation. Our modeling with SVM and Decision Tree had led us to conclude that the language spoken at home most of the time, the area in which they live, poverty level and age are good indicators to predict whether or not someone will complete 12 or more years of education.

Also, that even though there were many differences in the education system between 2019 and 2020, this difference wasn't large enough to influence greatly on our graduation prediction. We could still make a good prediction for 2020 training our model with 2019 data.

8.1 Limitations

We only used two machine learning models and two metrics. We could get better predictions using other models.

8.2 Future work

Our database has a lot of potential and contains a lot of information. For future works, we could try to predict how many Paraguayans graduate high-school in time, which is at around 17-19 years of age.

Also, we could try and figure out if the fact that our models F1 score and accuracy decreasing is due to the influence of the COVID-19 pandemic or if there's another reason that wasn't taken into account.

Finally, we could also add other factors such as tests that measure whether students are learning or not, the impact of the school supply kits and free school lunch.