

UNIVERSIDAD INTERNACIONAL DE CIENCIA Y TECNOLOGIA

BASE DE DATOS AVANZADO

PROFESOR: ABDEL G. MARTÍNEZ L.

TAREA CORRESPONDIENTE A: **Trabajo Final**

PRESENTADO POR: INNUALILER ORÁN

ALEX MAYLIN

EMILIO I. CÁRDENAS V.

DICIEMBRE 2, 2015

# Selección del tema

Para esta etapa del proyecto entramos al sitio del Instituto Nacional de Estadística y Censo (INEC).

Al revisar los temas relacionados con educación, nos llamó la atención un estudio realizado en el año 2014 donde presentaban la cantidad de docentes activos que para la fecha tenían o no título universitario.

Descargamos los datos de la investigación y comenzamos analizar la data.

# Fase 1 Integración y Recopilación

Los datos se encontraban en formato Excel (.xls), semi estructurados, tenía el siguiente formato:

* Dependencia,Provincia, Comarca Cadenas
* Total Entero
* Hombres Entero
* Mujer Entero
* Total con titulo en docencia Entero
* Hombres con titulo en docencia Entero
* Mujeres con titulo en docencia Entero
* Total con otro título Entero
* Hombres con otro título Entero
* Mujeres con otro título Entero
* Total sin título Entero
* Hombres sin título Entero
* Mujeres sin título Entero

Nota: la primera columna dentro del Excel se distinguía por una sangría. Ejm:

Bocas del Toro.

Changuinola.

La misma tabla se encontraba 3 veces, la primera vez con todos los totales, la segunda tabla con los datos relacionados a los centros educativos oficiales y la tercera con los datos relacionados a los centros educativos particulares.

# Fase 2 selección limpieza y transformación

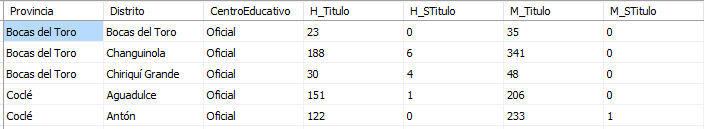
Luego de analizar las tablas descargadas del sitio del INEC, concluimos que las siguientes columnas no eran necesarias tenerlas en nuestra fuente de datos:

* Total Entero
* Hombres Entero
* Mujer Entero
* Total con titulo en docencia Entero
* Total con otro título Entero
* Hombres con otro título Entero
* Mujeres con otro título Entero
* Total sin título Entero

Por lo que se eliminaron de nuestra estructura, dejando una tabla con las siguientes columnas:

* Provincia Cadena
* Distrito Cadena
* Centro Educativo Cadena
* Hombre con titulo Entero
* Hombre sin título Entero
* Mujer con título Entero
* Mujer sin título Entero

Luego se exporto esta última tabla a una base de datos e SQL Server:



Luego de tener la tabla cargada en el SQL Server, se corrió un script que permitiera crear un registro por cada uno de los valores dentro de las columnas Hombre con Titulo (H\_titulo), Hombre sin Titulo (H\_STitulo), Mujer con Titulo (M\_titulo), Mujer sin Titulo (M\_STitulo) y agregara la columna sexo. Dejando nuestra fuente de datos de la siguiente manera:

* Provincia Cadena
* Distrito Cadena
* Tipo Educación Cadena
* Sexo Cadena
* Titulo Entero [0,1]



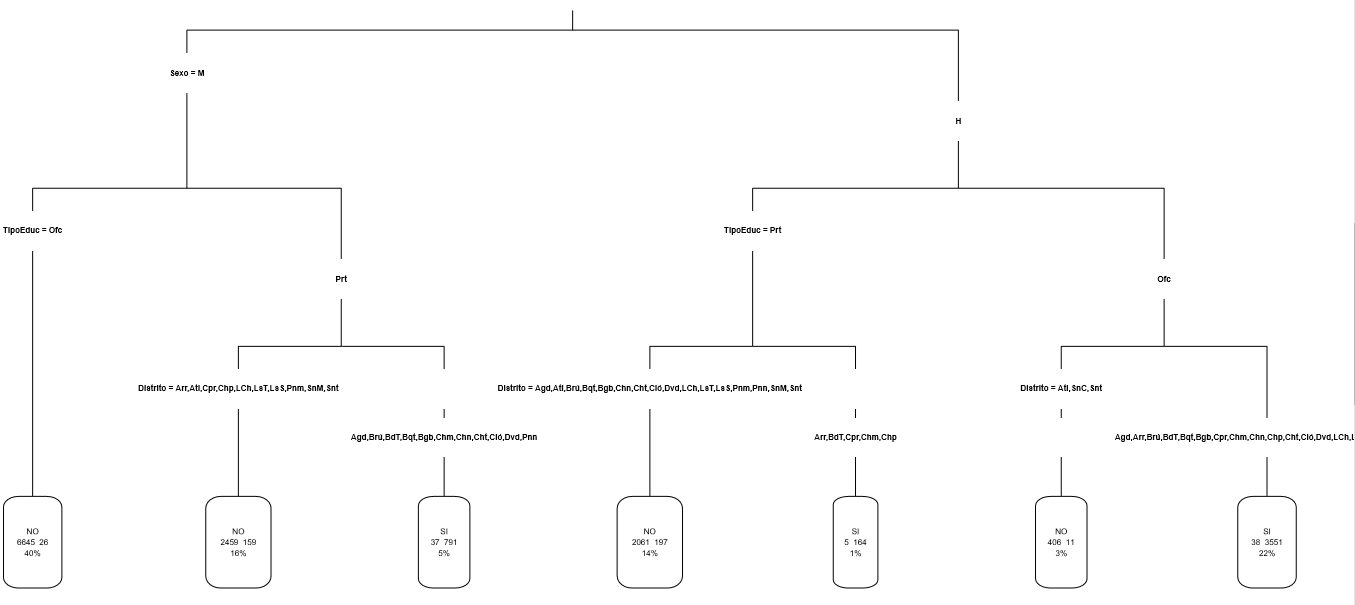
Como último paso se exporto la tabla del SQL Server a un archivo de texto plano (.csv).

# Fase 3 Minería de datos

Aplicamos la técnica de árbol de decisión, específicamente el algoritmo rpart, para la fuente de datos.

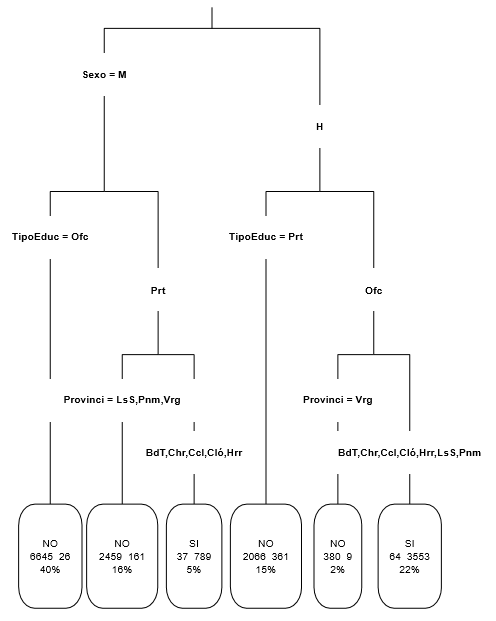
Evaluamos la columna Titulo utilizando como dependencias las columnas Provincia, distrito, Sexo y Centro Educativo.

Obteniendo el siguiente resultado:



Aplicamos el algoritmo para evaluar la columna titulo, pero esta vez omitimos la columna Distrito.

Obteniendo de esta forma el siguiente resultado:



# Fase 4 Evaluación e interpretación

Decidimos omitir la columna de distrito y basarnos para el estudio en las columnas Provincia, Sexo y Centro Educativo como dependencias de si el docente tiene titulo o no.

Como podemos ver en el segundo árbol generado, Si un docente es Mujer y trabaja en colegio oficial hay un alto porcentaje de docentes Mujeres sin título universitario. Sin embargo si es docente en colegio particular vemos que para las provincias Los santos, Panamá y Veraguas hay un gran número de docentes Mujeres que no poseen título universitario, pero para las provincias Bocas del toro, Chiriquí, Coclé, Colón y Herrera la mayoría de los docentes Mujeres tienen título universitario.

Caso contrario que las mujeres, según el análisis realizado, la mayoría de los docentes hombres que dictan clases en escuela particulares No tienen titulo.

Los docentes Hombres que pertenecen a escuelas particulares en la provincia de Veraguas, en su mayoría no poseen título universitario.

Los docentes Hombres que pertenecen a escuelas particulares en las provincias de Bocas del toro, Chiriquí, Coclé, colón, Herrera, Los Santos y Panamá, en su mayoría tienen título universitario.

# Anexos

## Codigo R:

> FuenteDatos <- read.csv("C:/Users/Emilio/Desktop/Final Mineria/FuenteDatos.txt")

> View(FuenteDatos)

> arb\_rpart <- rpart(Titulo ~ ., data=FuenteDatos, method="class")

> plot(arb\_rpart)

> rpart.plot(arb\_rpart, type=3, extra=101, fallen.leaves=T)

Warning messages:

1: In abbreviate(names, minlen) : abbreviate used with non-ASCII chars

2: In abbreviate(names, minlen) : abbreviate used with non-ASCII chars

3: In abbreviate(names, minlen) : abbreviate used with non-ASCII chars

> arb\_rpart2 <- rpart(Titulo ~ Provincia + Sexo + TipoEducacion, data=FuenteDatos, method="class")

> rpart.plot(arb\_rpart2, type=3, extra=101, fallen.leaves=T)

Warning messages:

1: In abbreviate(names, minlen) : abbreviate used with non-ASCII chars

2: In abbreviate(names, minlen) : abbreviate used with non-ASCII chars

## Script de para generar fuente de datos SQL Server:

set rowcount 0

select NULL mykey, \*

into #mytemp

from dbo.Depuracion

create table #TempDep

(

Provincia varchar(50),

Distrito varchar(50),

TipoEduc varchar(50),

Sexo varchar(5),

Titulo int

)

set rowcount 1

update #mytemp set mykey = 1

while @@rowcount > 0

begin

set rowcount 0

DECLARE @Provincia varchar(50),

@Distrito varchar(50),

@TipoEduc varchar(50),

@H\_titulo int,

@H\_Stitulo int,

@M\_titulo int,

@M\_Stitulo int

select @Provincia = Provincia,

@Distrito = Distrito,

@TipoEduc = CentroEducativo,

@H\_titulo =H\_Titulo,

@H\_Stitulo =H\_STitulo,

@M\_titulo =M\_Titulo,

@M\_Stitulo = M\_STitulo

from #mytemp

where mykey = 1

declare @Contador int

Set @Contador = 0

while(@Contador < @H\_titulo)

BEGIN

INSERT INTO #TempDep (Provincia, Distrito, TipoEduc, Sexo, Titulo)

VALUES (@Provincia, @Distrito, @TipoEduc, 'H', 1)

set @Contador = @Contador + 1

END

Set @Contador = 0

while(@Contador < @H\_Stitulo)

BEGIN

INSERT INTO #TempDep (Provincia, Distrito, TipoEduc, Sexo, Titulo)

VALUES (@Provincia, @Distrito, @TipoEduc, 'H', 0)

set @Contador = @Contador + 1

END

Set @Contador = 0

while(@Contador < @M\_titulo)

BEGIN

INSERT INTO #TempDep (Provincia, Distrito, TipoEduc, Sexo, Titulo)

VALUES (@Provincia, @Distrito, @TipoEduc, 'M', 1)

set @Contador = @Contador + 1

END

Set @Contador = 0

while(@Contador < @M\_Stitulo)

BEGIN

INSERT INTO #TempDep (Provincia, Distrito, TipoEduc, Sexo, Titulo)

VALUES (@Provincia, @Distrito, @TipoEduc, 'M', 0)

set @Contador = @Contador + 1

END

delete #mytemp where mykey = 1

set rowcount 1

update #mytemp set mykey = 1

end

set rowcount 0

insert into FuenteDatos

Select \* from #TempDep

drop table #mytemp

drop table #TempDep