Proyecto de Bases de datos

Diana Isabel Arévalo Durán
Universidad Central
Maestría en Analítica de Datos
Curso de Bases de Datos
Bogotá, Colombia
darevalod@ucentral.edu.co

October 8, 2022

Contents

1 Introducción

El presente trabajo se realiza con base en los datos recolectados en la gran encuesta de los hogares Colombianos 2020, con el objetivo de predecir la probabilidad de que los niños de Colombia, menores de 12 años al 2020, en el futuro accedan a educación superior o universitaria.

Para esto, inicialmente se realiza normalización de la base de datos obtenida del Archivo Nacional de Datos (ANDA) que permite la consulta por temática, operación estadística y año, de los diferentes microdatos anonimizados dispuestos al público en la página web del DANE, haciendo uso del sistema manejador de base de datos MySQL.

Posteriormente se aplican métodos de estadística exploratoria multidimensional, resumiendo el conjunto de variables en unas pocas nuevas variables, buscando agrupaciones en los datos y clasificando para identificar relaciones entre variables. Finalmente se contrastan los resultados contra la hipótesis planteada.

2 Características del proyecto

2.1 Titulo del proyecto

Predicción de la probabilidad que tienen los niños de los hogares colombianos menores de 12 años al 2020, de acceso a la educación superior, de acuerdo con condiciones de empleo, fuentes de ingreso y caracteristicas generales de los hogares.

2.2 Objetivo general

Predecir la probabilidad de que los niños de Colombia, menores de 12 años al 2020, en el futuro accedan a educación superior o universitaria.

2.2.1 Objetivos especificos

- A través del uso de métodos de reducción de dimensionalidad generar visualizaciones de datos que permitan la fácil comprensión de los datos de la encuesta general de hogares colombianos.

- Por medio de la clasificación y regresión, analizar comportamientos para predecir la probablidad de acceso a educación superior de los niños menores de 12 años.

2.3 Alcance

El alcance de este proyecto será la predicción probabilidad de que los niños de Colombia, menores de 12 años al 2020, en el futuro accedan a educación superior o universitaria.

2.4 Pregunta de investigación

Entre 0 y 12 años ¿Qué probabilidad tienen los niños de los hogares colombianos de la Gran encuesta integrada de hogares 2020, de acceder a la educación superior o universitaria de acuerdo con condiciones de empleo, fuentes de ingreso y caracteristicas generales de los hogares?.

2.5 Hipotesis

La probabilidad de acceso a la educación superior o universitaria de los niños en Colombia está determinada por el nivel educativo de los padres.

3 Reflexiones sobre el origen de datos e información

La Gran encuesta integrada de hogares es una encuesta mediante la cual se solicita información sobre las condiciones de empleo de las personas, además de las características generales de la población como sexo, edad, estado civil y nivel educativo, se pregunta sobre sus fuentes de ingresos. Esta encuesta proporciona indicadores a nivel nacional del mercado laboral en Colombia que permiten conocer entre otros aspectos: la tasa de ocupación, la tasa de desocupación, la rama de actividad en que se desempeñan los colombianos y su remuneración, así como, el comportamiento del mercado laboral para jóvenes, mujeres y otros grupos poblacionales específicos.

3.1 ¿Cual es el origen de los datos e información?

Los datos corresponden a la Gran Encuesta Integrada de Hogares Colombianos 2020 realizada por el Departamento Administrativo Nacional de Estadística - DANE, a la población civil no institucional residente en hogares particulares. Esta población se estima con base en los censos de población, las estadísticas vitales y de migración.

Estos datos fueron obtenidos del Archivo Nacional de Datos (ANDA) que permite la consulta por temática, operación estadística y año, de los diferentes microdatos anonimizados dispuestos al público en la página web del DANE.

3.2 ¿Cuales son las consideraciones legales o eticas del uso de la información?

Toda la información recolectada para los censos y encuestas de los procesos estadísticos del DANE está protegida por la Ley 79 de 1993 o Ley de Reserva Estadística.

Según esta ley los datos suministrados al DANE a través de censos, encuestas u operaciones estadísticas "no podrán darse a conocer al público ni a las entidades u organismos oficiales, ni a las autoridades públicas, sino únicamente en resúmenes numéricos".

En consonancia con la Ley 1581 de 2012, las bases de datos y archivos regulados por la Ley 79 de 1993 están exceptuadas del régimen de protección de datos personales. Sin embargo, deben aplicarse los principios establecidos para su protección como son: el principio de legalidad en materia de tratamiento de datos, de finalidad, de libertad, de veracidad o calidad, de transparencia, de acceso y circulación restringida, de seguridad y de confidencialidad.

Por lo anterior, las bases anonimizadas y disponibles en el ANDA han surtido un proceso de modificación y transformación de los datos originales, con el

objetivo de garantizar la confidencialidad de las unidades de análisis, por lo cual es necesario tener en cuenta estas modificaciones para el uso de la información.

3.3 ¿Cuales son los retos de la información y los datos que utilizara en la base de datos en terminos de la calidad y la consolidación?

Los datos obtenidos están representados con variables categóricas y el nombre de estas es dado con códigos, por lo que se debe realizar listas de homologación tanto para los nombres como para las categorías, con el objetivo de poder generar información sencilla y entendible.

Debido a que las bases publicadas por el DANE se encuentran anonimizadas y han surtido un proceso de modificación y transformación de los datos originales, la estructura de la base, entregada en el documento técnico de la DIAN, no corresponde con la publicada en la página, es por esto que se debe trabajar en la identificación de variables y homologaciones disponibles.

3.4 ¿Que espera de la utilización de un sistema de Bases de Datos para su proyecto?

Se hace uso del sistema de base de datos con el objetivo de organizar los datos obtenidos para poder acceder a búsquedas eficientes y consultas simplificadas que puedan ser reutilizadas.

Del sistema se espera extraer bases de datos unificadas con las variables más importantes de la operación estadística para realizar análisis de correspondencia.

4 Diseño del Modelo de Datos del SMBD (Sistema Manejador de Bases de Datos)

4.1 Características del SMBD (Sistema Manejador de Bases de Datos) para el proyecto

El Sistema Manejador de Bases de Datos utilizado para el proyecto es MySQL y dentro de sus principales caracteristicas están:

- Sistema de base de datos relacional de uso libre y gratuito.
- Soporta gran cantidad de datos, incluso con más de 50 millones de registros.
- Ejecución de transacciones y uso de claves foráneas.
- Conectividad segura.

4.2 Diagrama modelo de datos

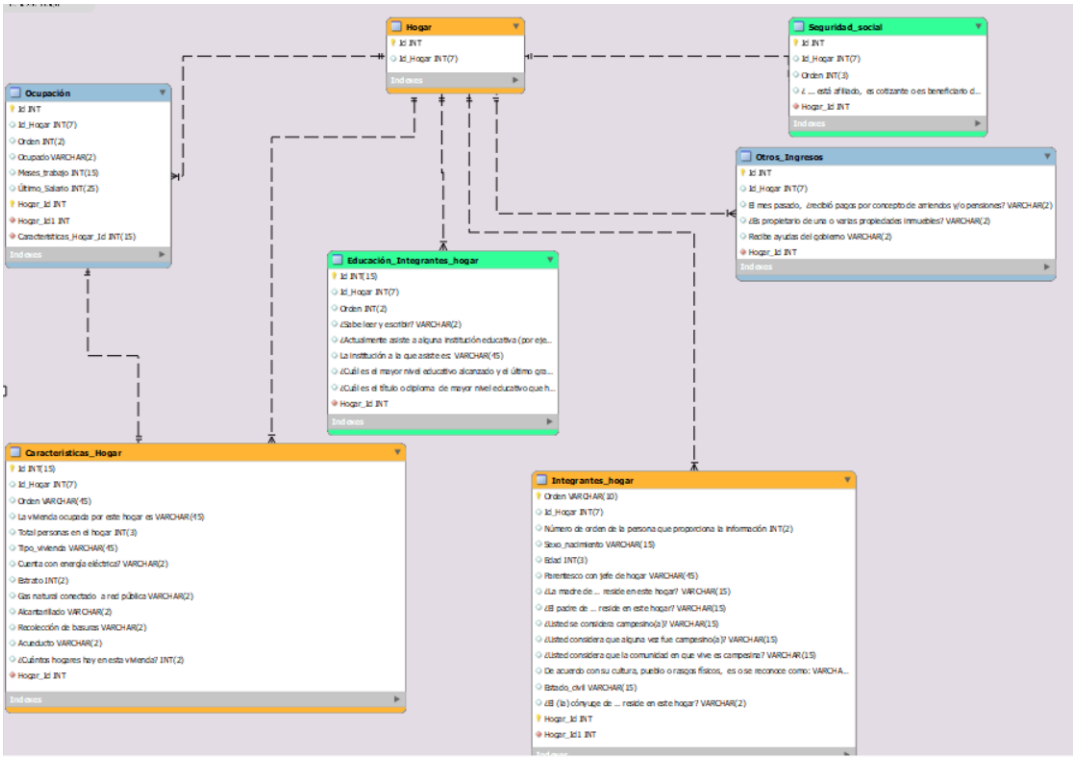


Figure 1: Diagrama Entidad Relación

4.3 Imágenes de la Base de Datos

Integrantes hogar

ï»¿Id	Orden	NÃºmero de orden de la persona que proporciona la informaciÃ³n	Id_Hogar	Sexo_nacimiento	Edad	Parentesco con jefe de hogar	Â¿La madre de â€¦ reside en este hogar?
1	1	1	2120000	Masculino	29	Jefe (a) del hogar	NO
2	2	2	2120000	Femenino	24	Pareja. esposo(a). cÃ³nyuge. compaÃ±ero(a)	NO
3	3	2	2120000	Masculino	7	Hijo(a). hijastro(a)	SI
4	4	2	2120000	Masculino	6	Hijo(a). hijastro(a)	SI
5	5	2	2120000	Masculino	3	Hijo(a). hijastro(a)	SI
6	1	1	2120001	Masculino	61	Jefe (a) del hogar	NO
7	1	1	2120002	Femenino	50	Jefe (a) del hogar	Fallecida
8	2	2	2120002	Femenino	34	Hijo(a). hijastro(a)	SI
9	1	1	2120003	Femenino	52	Jefe (a) del hogar	Fallecida
10	2	2	2120003	Masculino	54	Pareja. esposo(a). cÃ³nyuge. compaÃ±ero(a)	Fallecida
11	1	1	2120004	Femenino	30	Jefe (a) del hogar	NO
12	2	2	2120004	Femenino	12	Hijo(a). hijastro(a)	SI
13	3	1	2120004	Femenino	9	Hijo(a). hijastro(a)	SI
14	4	1	2120004	Femenino	6	Hijo(a). hijastro(a)	SI

integrantes 6 ×

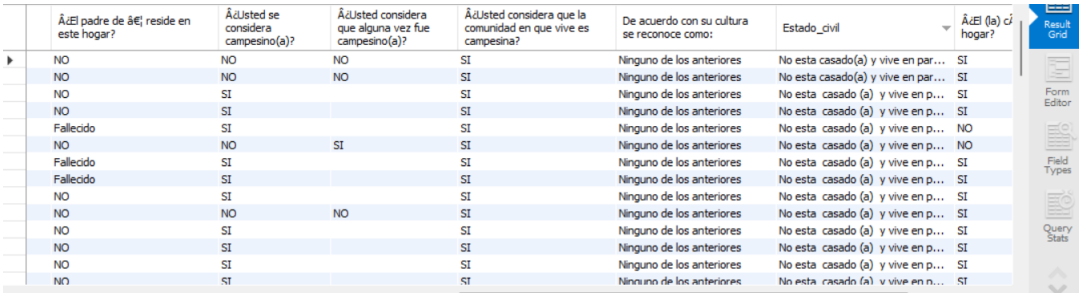
Read Only

Result Grid

Form Editor

Field Types

Query Stats



Â¿El padre de â€¦ reside en este hogar?	Â¿Usted se considera campesino(a)?	Â¿Usted considera que alguna vez fue campesino(a)?	Â¿Usted considera que la comunidad en que vive es campesina?	De acuerdo con su cultura se reconoce como:	Estado_civil	Â¿El (la) cÃ³ hogar?
NO	NO	NO	SI	Ninguno de los anteriores	No esta casado(a) y vive en par…	SI
NO	NO	NO	SI	Ninguno de los anteriores	No esta casado(a) y vive en par…	SI
NO	SI		SI	Ninguno de los anteriores	No esta casado (a) y vive en p…	SI
NO	SI		SI	Ninguno de los anteriores	No esta casado (a) y vive en p…	SI
Fallecido	SI		SI	Ninguno de los anteriores	No esta casado (a) y vive en p…	NO
NO	NO	SI	SI	Ninguno de los anteriores	No esta casado (a) y vive en p…	NO
Fallecido	SI		SI	Ninguno de los anteriores	No esta casado (a) y vive en p…	SI
Fallecido	SI		SI	Ninguno de los anteriores	No esta casado (a) y vive en p…	SI
NO	SI		SI	Ninguno de los anteriores	No esta casado (a) y vive en p…	SI
NO	NO	NO	SI	Ninguno de los anteriores	No esta casado (a) y vive en p…	SI
NO	SI		SI	Ninguno de los anteriores	No esta casado (a) y vive en p…	SI
NO	SI		SI	Ninguno de los anteriores	No esta casado (a) y vive en p…	SI
NO	SI		SI	Ninguno de los anteriores	No esta casado (a) y vive en p…	SI
NO	SI		SI	Ninguno de los anteriores	No esta casado (a) y vive en p…	SI

Result Grid

Form Editor

Field Types

Query Stats

4.4 Código SQL - lenguaje de definición de datos (DDL)

```sql
-- -----------------------------------------------------
-- Table `mydb`.`Hogar`
-- -----------------------------------------------------
CREATE TABLE IF NOT EXISTS `Encuesta_hogares`.`Hogar` (
  `Id` INT NOT NULL AUTO_INCREMENT,
  `Id_Hogar` INT(7) NULL,
  PRIMARY KEY (`Id`))
ENGINE = InnoDB;


-- -----------------------------------------------------
-- Table `mydb`.`Integrantes_hogar`
-- -----------------------------------------------------
CREATE TABLE IF NOT EXISTS `Encuesta_hogares`.`Integrantes_hogar` (
  `Id` INT NOT NULL AUTO_INCREMENT,
  `Orden` VARCHAR(10) NULL,
  `Id_Hogar` INT(7) NULL,
  `Número de orden de la persona que proporciona la información` INT(2) NULL,
  `Sexo_nacimiento` VARCHAR(15) NULL,
  `Edad` INT(3) NULL,
  `Parentesco con jefe de hogar` VARCHAR(45) NULL,
  `¿La madre de … reside en este hogar?` VARCHAR(15) NULL,
  `¿El padre de … reside en este hogar?` VARCHAR(15) NULL,
  `¿Usted se considera campesino(a)?` VARCHAR(15) NULL,
  `¿Usted considera que alguna vez fue campesino(a)?` VARCHAR(15) NULL,
  `¿Usted considera que la comunidad en que vive es campesina?` VARCHAR(15) NULL,
  `De acuerdo con su cultura se reconoce como:` VARCHAR(45) NULL,
  `Estado_civil` VARCHAR(15) NULL,
  `¿El (la) cónyuge de … reside en este hogar?` VARCHAR(2) NULL,
  `Hogar_Id` INT NOT NULL,
  `Hogar_Id1` INT NOT NULL,
  PRIMARY KEY (`Id`, `Hogar_Id`),
  CONSTRAINT `fk_Integrantes_hogar_Hogar1`
    FOREIGN KEY (`Hogar_Id1`)
    REFERENCES `mydb`.`Hogar` (`Id`)
    ON DELETE NO ACTION
    ON UPDATE NO ACTION)
ENGINE = InnoDB;
```

```sql
-- Table `Encuesta_hogares`.`Caracteristicas_Hogar`
-- -----------------------------------------------------
CREATE TABLE IF NOT EXISTS `Encuesta_hogares`.`Caracteristicas_Hogar` (
  `Id` INT(15) NOT NULL AUTO_INCREMENT,
  `Id_Hogar` INT(7) NULL,
  `Orden` VARCHAR(45) NULL,
  `La vivienda ocupada por este hogar es` VARCHAR(45) NULL,
  `Total personas en el hogar` INT(3) NULL,
  `Tipo_vivienda` VARCHAR(45) NULL,
  `Cuenta con energía eléctrica?` VARCHAR(2) NULL,
  `Estrato` INT(2) NULL,
  `Gas natural conectado  a red pública` VARCHAR(2) NULL,
  `Alcantarillado` VARCHAR(2) NULL,
  `Recolección de basuras` VARCHAR(2) NULL,
  `Acueducto` VARCHAR(2) NULL,
  `¿Cuántos hogares hay en esta vivienda?` INT(2) NULL,
  `Hogar_Id` INT NOT NULL,
  PRIMARY KEY (`Id`),
  CONSTRAINT `fk_Caracteristicas_Hogar_Hogar1`
    FOREIGN KEY (`Hogar_Id`)
    REFERENCES `mydb`.`Hogar` (`Id`)
    ON DELETE NO ACTION
    ON UPDATE NO ACTION)
ENGINE = InnoDB;
```

```sql
-- Table `Encuesta_hogares`.`Seguridad_social`
-- -----------------------------------------------------
CREATE TABLE IF NOT EXISTS `Encuesta_hogares`.`Seguridad_social` (
  `Id` INT NOT NULL AUTO_INCREMENT,
  `Id_Hogar` INT(7) NULL,
  `Orden` INT(3) NULL,
  `¿ está afiliado entidad de seguridad social` VARCHAR(45) NULL,
  `A qué regimen` VARCHAR(45) NULL,
  `Hogar_Id` INT NOT NULL,
  PRIMARY KEY (`Id`),
  CONSTRAINT `fk_Seguridad_social_Hogar1`
    FOREIGN KEY (`Hogar_Id`)
    REFERENCES `mydb`.`Hogar` (`Id`)
    ON DELETE NO ACTION
    ON UPDATE NO ACTION)
ENGINE = InnoDB;
```

```
111   -- Table `Encuesta_hogares`.`Educación_Integrantes_hogar`
112   -- -----------------------------------------------------
113   CREATE TABLE IF NOT EXISTS `Encuesta_hogares`.`Educación_Integrantes_hogar` (
114     `Id` INT(15) NOT NULL AUTO_INCREMENT,
115     `Id_Hogar` INT(7) NULL,
116     `Orden` INT(2) NULL,
117     `¿Sabe leer y escribir?` VARCHAR(2) NULL,
118     `¿Actualmente asiste a alguna institución educativa?` VARCHAR(2) NULL,
119     `La institución a la que asiste es:` VARCHAR(45) NULL,
120     `¿mayor nivel educativo alcanzado?` VARCHAR(45) NULL,
121     `¿título o diploma  de mayor nivel educativo?` VARCHAR(45) NULL,
122     `Hogar_Id` INT NOT NULL,
123     PRIMARY KEY (`Id`),
124     CONSTRAINT `fk_Educación_Integrantes_hogar_Hogar1`
125       FOREIGN KEY (`Hogar_Id`)
126       REFERENCES `mydb`.`Hogar` (`Id`)
127       ON DELETE NO ACTION
128       ON UPDATE NO ACTION)
129   ENGINE = InnoDB;
134   -- Table `Encuesta_hogares`.`Ocupación`
135   -- -----------------------------------------------------
136   CREATE TABLE IF NOT EXISTS `mydb`.`Ocupación` (
137     `Id` INT NOT NULL,
138     `Id_Hogar` INT(7) NULL,
139     `Orden` INT(2) NULL,
140     `Ocupado` VARCHAR(2) NULL,
141     `Meses_trabajo` INT(15) NULL,
142     `Último_Salario` INT(25) NULL,
143     `Hogar_Id` INT NOT NULL,
144     `Hogar_Id1` INT NOT NULL,
145     `Caracteristicas_Hogar_Id` INT(15) NOT NULL,
146     PRIMARY KEY (`Id`, `Hogar_Id`),
147     CONSTRAINT `fk_Ocupación_Hogar`
148       FOREIGN KEY (`Hogar_Id1`)
149       REFERENCES `mydb`.`Hogar` (`Id`)
150       ON DELETE NO ACTION
151       ON UPDATE NO ACTION,
152     CONSTRAINT `fk_Ocupación_Caracteristicas_Hogar1`
153       FOREIGN KEY (`Caracteristicas_Hogar_Id`)
154       REFERENCES `mydb`.`Caracteristicas_Hogar` (`Id`)
155       ON DELETE NO ACTION
156       ON UPDATE NO ACTION)
157   ENGINE = InnoDB;
```

```sql
-- Table `Encuesta_hogares`.`Otros_Ingresos`
-- -----------------------------------------------------
CREATE TABLE IF NOT EXISTS `Encuesta_hogares`.`Otros_Ingresos` (
  `Id` INT NOT NULL,
  `Id_Hogar` INT(7) NULL,
  `¿recibió pagos por concepto de arriendos, pensiones?` VARCHAR(2) NULL,
  `¿Es propietario de una o varias propiedades inmuebles?` VARCHAR(2) NULL,
  `Recibe ayudas del gobierno` VARCHAR(2) NULL,
  `Hogar_Id` INT NOT NULL,
  PRIMARY KEY (`Id`),
  CONSTRAINT `fk_Otros_Ingresos_Hogar1`
    FOREIGN KEY (`Hogar_Id`)
    REFERENCES `mydb`.`Hogar` (`Id`)
    ON DELETE NO ACTION
    ON UPDATE NO ACTION)
ENGINE = InnoDB;


SET SQL_MODE=@OLD_SQL_MODE;
SET FOREIGN_KEY_CHECKS=@OLD_FOREIGN_KEY_CHECKS;
SET UNIQUE_CHECKS=@OLD_UNIQUE_CHECKS;
```

4.5 Código SQL - Manipulación de datos (DML)

```sql
-- -----------------------------------------------------
-- Data for table `encuesta_hogares`.`Hogar`
-- -----------------------------------------------------
START TRANSACTION;
USE `encuesta_hogares`;
INSERT INTO `encuesta_hogares`.`Hogar` (`Id`, `Id_Hogar`) VALUES (1, 2120000);
INSERT INTO `encuesta_hogares`.`Hogar` (`Id`, `Id_Hogar`) VALUES (2, 2120001);
INSERT INTO `encuesta_hogares`.`Hogar` (`Id`, `Id_Hogar`) VALUES (3, 2120002);
INSERT INTO `encuesta_hogares`.`Hogar` (`Id`, `Id_Hogar`) VALUES (4, 2120003);
INSERT INTO `encuesta_hogares`.`Hogar` (`Id`, `Id_Hogar`) VALUES (5, 2120004);
INSERT INTO `encuesta_hogares`.`Hogar` (`Id`, `Id_Hogar`) VALUES (6, 2120005);
INSERT INTO `encuesta_hogares`.`Hogar` (`Id`, `Id_Hogar`) VALUES (7, 2120006);
INSERT INTO `encuesta_hogares`.`Hogar` (`Id`, `Id_Hogar`) VALUES (8, 2120007);
INSERT INTO `encuesta_hogares`.`Hogar` (`Id`, `Id_Hogar`) VALUES (9, 2120008);
INSERT INTO `encuesta_hogares`.`Hogar` (`Id`, `Id_Hogar`) VALUES (10, 2120009);
INSERT INTO `encuesta_hogares`.`Hogar` (`Id`, `Id_Hogar`) VALUES (11, 2120010);
INSERT INTO `encuesta_hogares`.`Hogar` (`Id`, `Id_Hogar`) VALUES (12, 2120011);
INSERT INTO `encuesta_hogares`.`Hogar` (`Id`, `Id_Hogar`) VALUES (13, 2120012);
INSERT INTO `encuesta_hogares`.`Hogar` (`Id`, `Id_Hogar`) VALUES (14, 2120013);
INSERT INTO `encuesta_hogares`.`Hogar` (`Id`, `Id_Hogar`) VALUES (15, 2120014);
```

Figure 2: Insertar registros en tabla

5 Bases de Datos No-SQL (*Segunda entrega*)

5.1 Diagrama Bases de Datos No-SQL (*Segunda entrega*)

5.2 SMBD utilizado para la Base de Datos No-SQL (*Segunda entrega*)

6 Aplicación de ETL (Extract, Transform, Load) y Bodega de Datos (*Tercera entrega*)

6.1 Ejemplo de aplicación de ETL y Bodega de Datos (*Tercera entrega*)

7 Lecciones aprendidas (*Tercera entrega*)

8 Bibliografía