Proyecto de Bases de datos para la estimacion de acceso y uso de programas distritales en habitantes de calle de Bogotá

Julián David Parra[1]

Facultad de ingeniería y ciencias básicas
Universidad Central
Maestría en Analítica de Datos
Curso de Bases de Datos
Bogotá, Colombia
[1]jparram6@ucentral.edu.co

19 de noviembre de 2022

Contents

1 Introducción

Las entidades locales, han dispuesto campañas y programas destinados a los habitantes de calle con el fin de mejoras sus condiciones de vida e incorporarlos a la vida social; con el acompañamiento de profesionales sociales, médicos, psicólogos, antropólogos y otros, tratan sus diferentes problemas. En muchos de los casos, personas que acceden a alguno de los programas al cabo de un tiempo, poco o mucho, éstos desertan y recaen en sus antiguos malos hábitos: drogas, robos, etc. Gran parte de esta población con el paso del tiempo generan problemas psicosociales que dificulta la relación y comunicación, limitando la capacidad de los expertos quienes se encargan de realizar el seguimiento de detectar las razones que hacen que un habitante de calle recaiga; si identificamos las personas mas propensas a recaer se podrá actuar con precisión y diferenciación estos casos más críticos, así aumentar la efectividad de los programas y el número de habitantes de calle rehabilitados en su totalidad. En la base de datos se identifican los habitantes de calle que han accedido en algún momento a los programas de atención y si continúan en ellos, con esas dos preguntas podemos inferir la deserción, la cual es la variable dependiente que utilizaremos en nuestro modelo y que nos permitirá determinar la probabilidad de riesgo de recaída de cada uno de los habitantes nuevos que sean identificados y registrados. Según [1] "En el Censos Sectorial se encuentran variables demográficas como el sexo y la edad de las personas, y variables que definen directamente las condiciones de vida de los habitantes de la calle, concordantes con las necesidades de información para la atención que prestan las Instituciones como el **IDIPRON** (Instituto Distrital para la Protección de la Niñez y la Juventud), la **SDIS** (Secretaría Distrital de Integración Social) y las oficinas de las Alcaldías municipales encargadas del tema".

2 Características del proyecto de investigación

Se cuenta con la información del censo de habitantes de calle (llamado por sus siglas CHC) realizado el año 2007, donde se recoge información de 9538 habitantes de calle de la ciudad de Bogotá recolectados entre el 27 de octubre y el 8 de noviembre. Mediante técnicas de análisis estadístico se espera estimar la probabilidad de deserción que tienen los habitantes de calle que acceden a alguno de los programas distritales de rehabilitación e inclusión social dispuestos por las entidades distritales, este insumo se utilizará para poder enfocar los recursos económicos y humanos en la retención de los habitantes en cada programa, priorizando y realizando un seguimiento más cercano a quienes su probabilidad de deserción sea más alta. Se generarán segmentos de priorización en donde cada uno y según sus características se tomarán medidas específicas y se definirán planes de acción diferencial.

2.1 Titulo del proyecto de investigación

Estimación de uso y permanencia de programas distritales en habitantes de calle de Bogotá

2.2 Objetivo general

Identificar las caracteristicas que determinan la permanencia en los programas distritales para habitantes de calle, por medio de un modelo de arboles de decisión, con el fin de focalizar los recursos economicos y humanos en su rehabilitación.

2.2.1 Objetivos especificos

- Identificar las causas por las que se presenta deserción.
- Identificar la distribución poblacional de habitantes de calle por variables sociodemográficas.
- Describir el perfil demográfico de habitantes de calle desertores.
- Describir el perfil demográfico de habitantes de calle quienes nunca han participado en alguno de los programas dispuestos por la alcaldia.

2.3 Alcance

Conociendo y entendiendo la problemática entorno a los habitantes de calle, este proyecto estará enfocado en dar información importante basada en el análisis de los datos del censo CHC 2017, el cual cuenta con 123 variables el cual le servirá a las entidades distritales para definir sus estrategias. El periodo de análisis, procesamiento y entrega de información está proyectado para realizarse en 4 meses, durante la duración del semestre actual de la maestría en analítica de datos de la Universidad Central.

2.4 Pregunta de investigación

Teniendo en cuenta la sensibilidad del tema en cuestión ¿Es posible estimar la probabilidad de riesgo de recaída en los habitantes de calle con información demográfica, de consumo y actividad económica?

2.5 Hipotesis

La segmentación demográfica permite ver cuál es el perfil donde se debe focalizar los esfuerzos tanto econónicos como profesionales para que sea más efectivas las actividades escogidas para el desarrollo de la rehabilitación de los habitantes de calle.

3 Reflexiones sobre el origen de datos e información

El Departamento Administrativo Nacional de Estadística en conjunto con la Secretaria Distrital de Integración Social (SDIS) se unen para hacer el censo de habitantes de calle (CHC), la base de datos cuenta con 9538 registros, en términos de variables, existen 123 variables, 10 de ellas son variables continuas y 110 son variables categóricas. La base de datos está en formato .SAV, el cual es un archivo de spss, también contamos con un archivo donde se ubican las etiquetas en la que está el dominio de cada variable. Por el tipo de grupo objetivo y por el contenido de las preguntas que son altamente sensible, la base de datos contiene registros e información faltante, aquellas encuestas que contienen información faltante está marcada con un flag en una variable que está dentro de la base de datos.

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

La base de datos es descargada del archivo nacional de datos (ANDA), la cual es una página oficial del DANE, se obtiene desde el siguiente link. Para efectos de este proyecto y con el fin de delimitar la información arbitrariamente se van a analizar las siguientes variables: edad, género, actividad económica y el consumo de sustancias.

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

Por el tratamiento de datos sensibles y por políticas de publicación de información La base de datos está debidamente anonimizada por lo que su tratamiento es completamente reservado.

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

Organizar y guardar de forma ordenada y eficiente la información que está contenida en la base de datos, con el fin de poderla usar y conservar durante y después de proceso de análisis. La base está guardada en dos tablones diferentes, uno que contiene los códigos y otra que contiene las etiquetas, por lo que el reto está en poder conectar estas dos bases de forma adecuada.

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

Poder guardar, ordenar, normalizar y disponibilizar la base de datos para el trabajo durante la fase de análisis de la información y para que pueda ser usada por otros usuarios que puedan generar valor agregado al tratamiento de habitantes de calle.

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

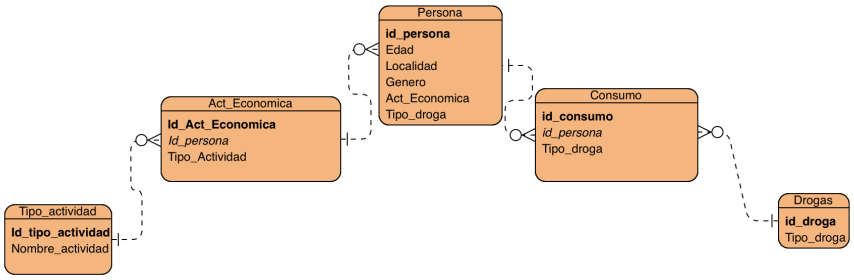
El diseño de la base de datos estará basado en el modelo relacional.

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

El sistema manejador de bases de datos a utilizar es SQL Server, las siguientes son las principales características.

- Capacidad de trabajar con grandes volumenes de datos de forma eficiente.
- Control de acceso a las bases de datos.
- Posibilidad de crear vistas seleccionando otras vista o tablas de la base de datos.
- Gestiona bases de datos relacionales.

4.2 Diagrama modelo de datos



4.3 Imágen de la Base de Datos

	DIRECTORIO	TIP_FOR	P1	P1S1	P2	P2S1	P5	CTL_1	P8	P9R	P10R	P11R	P12	P13	P15R	P16S1	P16S2	P16S3	P16S4	P16S5	P16S6	P16S7	P16S8	P16S9	P17	P
0	101000	1	11	11001	1	16	2	1	38.0	1.0	3	2.0	1.0	3	2	2.0	3.0	4.0	3.0	2.0	3.0	3.0	3.0	3.0	2.0	
1	101001	1	11	11001	1	16	2	1	38.0	2.0	3	2.0	1.0	3	2	4.0	4.0	3.0	4.0	4.0	2.0	4.0	4.0	4.0	2.0	
2	101002	1	11	11001	1	16	2	1	25.0	2.0	3	1.0	1.0	3	2	4.0	4.0	4.0	3.0	4.0	1.0	3.0	4.0	4.0	2.0	
3	101003	1	11	11001	1	16	2	1	52.0	1.0	3	2.0	1.0	3	2	1.0	1.0	1.0	3.0	3.0	1.0	3.0	2.0	2.0	2.0	
4	101004	1	11	11001	1	16	2	1	27.0	2.0	3	1.0	1.0	3	2	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	2.0	

5 rows × 123 columns

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

```
/*!40101 SET NAMES utf8 */;

/*!40101 SET SQL_MODE=''*/;

/*!40014 SET @OLD_UNIQUE_CHECKS=@@UNIQUE_CHECKS, UNIQUE_CHECKS=0 */;
/*!40014 SET @OLD_FOREIGN_KEY_CHECKS=@@FOREIGN_KEY_CHECKS, FOREIGN_KEY_CHECKS=0
/*!40101 SET @OLD_SQL_MODE=@@SQL_MODE, SQL_MODE='NO_AUTO_VALUE_ON_ZERO' */;
/*!40111 SET @OLD_SQL_NOTES=@@SQL_NOTES, SQL_NOTES=0 */;
CREATE DATABASE /*!32312 IF NOT EXISTS*/`habitantescalle` /*!40100 DEFAULT CHARA

USE `habitantescalle`;

/*Table structure for table `persona` */

DROP TABLE IF EXISTS `persona`;

CREATE TABLE `persona` (
  `id_persona` int(11) NOT NULL,
  `edad` int(11) DEFAULT NULL,
  `localidad` varchar(50) NOT NULL,
  `genero` varchar(50) NOT NULL,
  `act_economica` varchar(50) NOT NULL,
  `tipo_droga` decimal(10,2) DEFAULT NULL,
  PRIMARY KEY (`id_persona`),
  KEY `act_economica` (`act_economica`),
  CONSTRAINT `persona_ibfk_1` FOREIGN KEY (`act_economica`) REFERENCES `act_econo
  CONSTRAINT `persona_ibfk_2` FOREIGN KEY (`tipo_droga`) REFERENCES `consumo` (`
  ) ENGINE=InnoDB DEFAULT CHARSET=latin1;


/*Table structure for table `act_economica` */

DROP TABLE IF EXISTS `act_economica`;

CREATE TABLE `act_economica` (
  `id_act_economica` int(11) NOT NULL,
  `id_persona` int(11) NOT NULL,
  `tipo_actividad` int(11) NOT NULL,
  PRIMARY KEY (`id_act_economica`),
  KEY `id_persona` (`id_persona`),
  KEY `tipo_actividad` (`tipo_actividad`),
  CONSTRAINT `act_economica_ibfk_1` FOREIGN KEY (`id_persona`) REFERENCES `perso
  CONSTRAINT `act_economica_ibfk_2` FOREIGN KEY (`tipo_actividad`) REFERENCES `a
```

```
) ENGINE=InnoDB DEFAULT CHARSET=latin1;


/*Table structure for table `offices` */

DROP TABLE IF EXISTS `tipo_actividad`;

CREATE TABLE `tipo_actividad` (
  `id_tipo_actividad` int(10) NOT NULL,
  `nombre_actividad` int(10) NOT NULL,
  PRIMARY KEY (`id_tipo_actividad`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1;



/*Table structure for table `act_economica` */

DROP TABLE IF EXISTS `consumo`;

CREATE TABLE `consumo` (
  `id_consumo` int(11) NOT NULL,
  `id_persona` int(11) NOT NULL,
  `tipo_droga` int(11) NOT NULL,
  PRIMARY KEY (`id_consumo`),
  KEY `id_persona` (`id_persona`),
  KEY `tipo_droga` (`tipo_droga`),
  CONSTRAINT `consumo_ibfk_1` FOREIGN KEY (`id_persona`) REFERENCES `persona` (`
  CONSTRAINT `consumo_ibfk_2` FOREIGN KEY (`tipo_droga`) REFERENCES `drogas` (`i
) ENGINE=InnoDB DEFAULT CHARSET=latin1;


/*Table structure for table `offices` */

DROP TABLE IF EXISTS `drogas`;

CREATE TABLE `drogas` (
  `id_droga` int(10) NOT NULL,
  `tipo_droga` int(10) NOT NULL,
  PRIMARY KEY (`id_droga`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1;


/*!40101 SET SQL_MODE=@OLD_SQL_MODE */;
/*!40014 SET FOREIGN_KEY_CHECKS=@OLD_FOREIGN_KEY_CHECKS */;
/*!40014 SET UNIQUE_CHECKS=@OLD_UNIQUE_CHECKS */;
/*!40111 SET SQL_NOTES=@OLD_SQL_NOTES */;
```

4.5 Código SQL - Manipulación de datos (DML)

```
/*Data for the table `persona` */

INSERT INTO `persona`(`id_persona`,`edad`,`localidad`,`genero`,`act_economica`,`
(101000,38,16, hombre , 9 ,NULL),
(101001,38,16, mujer , 8 , 3 ),
(101002,25,16, mujer , 8 ,NULL),
(101003,52,16, hombre , 9 ,NULL),
(101004,27,16, mujer , 8 , 4 ),
(101005,NULL,16,NULL,NULL,NULL),
(101006,62,16, hombre , 9 ,NULL),
(101007,61,16, hombre , 9 ,NULL),
(101008,50,16, hombre , 9 ,NULL),
(101009,57,16, hombre , 5 ,NULL),
(101010,57,16, hombre , 1 ,NULL),
(101011,34,16, mujer , 8 ,NULL),
(101012,29,16, hombre , 6 , 3 ),
(101013,54,16, mujer , 9 ,NULL),
(101014,56,16, mujer , 2 , 2 ),
(101015,59,16, mujer , 4 ,NULL),
(101016,60,16, mujer , 9 ,NULL),
(101017,58,16, hombre , 2 , 6 ),
(101018,63,16, mujer , 9 ,NULL),
(101019,40,16, hombre , 1 , 3 ),
(101020,57,16, mujer , 8 ,NULL),
(101021,58,16, hombre , 3 , 1 ),
(101022,49,16, hombre , 5 ,NULL),
(101023,42,16, hombre , 9 ,NULL),
(101024,33,16, hombre , 4 ,NULL),
(101025,34,16, mujer , 5 , 6 ),
(101026,41,16, hombre , 4 , 1 ),
(101027,28,16, hombre , 9 , 3 ),
(101028,58,16, hombre , 9 ,NULL),
(101029,30,16, hombre , 8 , 6 ),
(101030,34,16, hombre , 4 , 3 ),
(101031,54,16, mujer , 9 ,NULL),
(101032,45,16, mujer , 1 , 1 ),
(101033,33,16, mujer , 8 , 6 ),
(101034,44,16, mujer , 9 ,NULL),
(101035,61,16, hombre , 1 , 6 ),
(101036,59,16, mujer , 1 , 3 ),
(101037,70,16, mujer , 4 ,NULL),
(101038,56,16, hombre , 9 ,NULL),
(101039,53,16, mujer , 5 , 1 ),
```

```
(101040,59,16, hombre , 9 , 6 ),
(101041,48,16, mujer , 1 ,NULL),
(101042,57,16, hombre , 9 ,NULL),
(101043,50,16, mujer , 4 , 6 ),
(101044,56,16, hombre , 4 , 6 ),
(101045,57,16, hombre , 9 , 1 ),
(101046,57,16, hombre , 1 ,NULL),;

/*Data for the table `tipo_actividad_economica` */

INSERT INTO `act_economica`(`id_act_economica`,`id_persona`,`tipo_actividad`) va

(1,101000, 9 ),
(2,101001, 8 ),
(3,101002, 8 ),
(4,101003, 9 ),
(5,101004, 8 ),
(6,101005,NULL),
(7,101006, 9 ),
(8,101007, 9 ),
(9,101001, 9 ),
(10,101003, 5 ),
(11,101010, 1 ),
(12,101011, 8 ),
(13,101012, 6 ),
(14,101003, 9 ),
(15,101014, 2 ),
(16,101002, 4 ),
(17,101016, 9 ),
(18,101017, 2 );

/*Data for the table `tipo_actividad` */

INSERT INTO `tipo_actividad`(`id_tipo_actividad`,`nombre_actividad`) values

(1,'Limpiando vidrios, cuidando carros, tocando llantas, vendiendo en la calle u
(2,'Cantando, haciendo malabares, cuenter a artesan a u otras similares '),
(3,'Carpinter a, electricidad, construcci n u otras similares '),
(4,'Pidiendo, retacando, mendigando '),
(5,'Recogiendo material reciclable '),
(6,'Como campanero, taquillero, vendiendo o transportando sustancias psicoactiva
(7,'Robando o atracando'),
(8,'Ejerciendo la prostituci n ');

/*Data for the table `consumo` */
```

```
INSERT INTO `consumo`(`id_consumo`,`id_persona`,`tipo_droga`) values

(1,101000,NULL),
(2,101001,3),
(3,101002,NULL),
(4,101003,NULL),
(5,101004,4),
(6,101005,NULL),
(7,101006,NULL),
(8,101007,NULL),
(9,101001,NULL),
(10,101003,NULL),
(11,101010,NULL),
(12,101011,NULL),
(13,101012,3),
(14,101003,NULL),
(15,101014,2),
(16,101002,NULL),
(17,101016,NULL),
(18,101017,6);



/*Data for the table `drogas` */

INSERT INTO `drogas`(`id_tipo_actividad`,`nombre_actividad`) values
(1,'Cigarrillo'),
(2,'Alcohol'),
(3,'Marihuana'),
(4,'Inhalantes'),
(5,'Coca na'),
(6,'Basuco'),
(7,'Hero na'),
(8,'Pepas');
```

4.6 Código SQL + Resultados: Vistas

```
create view 'no_consumidores' as

Select * from persona
where tipo_droga is null;


create view 'atracadores' as
```

```
Select * from persona
where act_economica = 7;
```

4.7 Código SQL + Resultados: procedimientos almacenados

```
create procedure distribucion_tercera_edad()
 begin
   select genero,
   count(*) conteo,
   sum(case when genero = 'mujer' then 1 else 0 end)/count(*) porc_mujeres,
   sum(case when genero = 'hombre' then 1 else 0 end)/count(*) porc_hombres,
   from persona
   where stock<=10;
 end
```

5 Bases de Datos No-SQL

5.1 Diagrama Bases de Datos No-SQL

Se genera una base de datos No-SQL, con el tabulado del censo de habitantes de calle donde se encuentran los 9537 registros y las 14 variables que se van a utilizar en el análisis, allí están incluidas las variables identificadoras, las variables predictoras y la variable objetivo. El siguiente gráfico presenta el diagrama de la base de datos:

5.2 SMBD utilizado para la Base de Datos No-SQL

el sistema manejador de bases de datos utilizado fue MongoDB, ya que es uno de los SMBD más populares en el manejo de bases de datos NoSQL, con una interfaz gratuita y de código abierto. Según el portal de MongoDB, ubicado en el siguiente link MongoDB, estas son las caracteristicas más relevantes del SMBD.

- MongoDB almacena datos en documentos flexibles similares a JSON, por lo que los campos pueden variar entre documentos y la estructura de datos puede cambiarse con el tiempo
- El modelo de documento se asigna a los objetos en el código de su aplicación para facilitar el trabajo con los datos
- Las consultas ad hoc, la indexación y la agregación en tiempo real ofrecen maneras potentes de acceder a los datos y analizarlos
- MongoDB es una base de datos distribuida en su núcleo, por lo que la alta disponibilidad, la escalabilidad horizontal y la distribución geográfica están integradas y son fáciles de usar

- MongoDB es de uso gratuito. Las versiones lanzadas antes del 16 de octubre de 2018 se publican bajo licencia AGPL. Todas las versiones posteriores al 16 de octubre de 2018, incluidos los parches lanzados para versiones anteriores, se publican bajo Licencia pública del lado del servidor (SSPL) v1.

6 Lecciones aprendidas

- Validación de objetivos generales y particulares, que guarden la relación con el análisis que se prevé hacer, siempre viendo que estos objetivos sean alcanzables y que sean abordados correctamente desde las metodologías propuestas.
- Acotar los objetivos y definir el alcance del proyecto.
- En la analítica de datos tiene mucha importancia mantener una base de datos bien estructurada, incluso, más importante que las mismas técnicas de análisis; por esta razón fue muy interesante conocer las diferentes formas administrar una base de datos, desde bases SQL hasta NoSQL.
- Conocer como se mueve el mercado en términos de herramientas para el manejo de base de datos y de tecnología en general, para poder tener un mejor panorama y elegir la mejor opción que se requiera en proyectos personales o laborales.
- Mantener siempre a la vanguardia con las novedades que la tecnología ofrece.
- La importancia del idioma inglés en la investigación, el poder entender las ideas desde la fuente aportan mucha más imformación de lo que se está leyendo.
- El trabajo y aportes en equipo, poder conocer las ideas de otras personas, entendiendo temáticas desde otras perspectivas, ayuda a tener una visión más general del conocimiento.
- Compartir y socializar información es de mucha ayuda para cerrar brechas.

7 Bibliografía

References

[1] Departamento administrativo nacional de estadísticas DANE. Censo de habitantes de la calle 2017. 2018.