Reporte de Laboratorio Nro. 1

Tipantiza Cumbal Nayeli Michelle[L00073321]

Universidad de las Fuerzas Armadas
nmtipantiza@espe.edu.ec

Tema: Creación de datos sintéticos

Resumen

El presente laboratorio trata de cómo generar datos sintéticos que satisfacen las necesidades y ciertas condiciones enfocadas en la atención médica del instituto de seguro IESS en el Ecuador, ya que su problema radica en que las personas buscan una buena atención médica y el solicitar una cita es muy complicado. Para la creación y el desarrollo del mismo se tomó en cuenta a las siguientes entidades en donde se almacenará la información que se va a controlar: ***administrador, call center, permisos y servicio de salud***, cada una de ellas cuenta con características o rasgos de entidad, denominados atributos, mediante la generación de los datos sintéticos se podrá compartir información a los demás y siempre con protección de privacidad.

1. Introducción

La creación de datos sintéticos o conocido en inglés como ***Synthetic Data*** son conocidos como un procedimiento o método de generar información artificialmente para reemplazar datos reales y emplear modelos de Inteligencia Artificial [1]. Estos datos se generan por medio de un algoritmo informático, es decir, un agrupación de instrucciones ordenadas para resolver problemas, procesar datos y llevar a cabo ciertas actividades. En este caso se necesita generar datos artificiales para crear un entorno realista con ayuda de las entidades del proyecto. Esto va a ayudar a modo de capacitación adquiriendo habilidades avanzadas y se podrá aumentar la confianza sin tener que practicar con datos e información real.

Crear datos sintéticos puede ser sencillo, gran parte de científicos de datos usan paquetes preconstruidos para producir conjuntos de los mismos [2]. La finalidad del laboratorio es reconocer los modelos ya existentes en los datos y así poder reproducirlos. Además, el estudiante va a experimentar con los conocimientos adquiridos en clase y conocer nuevas maneras de cómo generar una agrupación de datos para resolver problemas en la vida real. Esto beneficia tanto al estudiante en su vida universitaria como su vida laboral en un futuro.

La creación de entidades y atributos también es parte de este laboratorio, cabe mencionar que un modelo o tipo de entidad va a corresponder a una o muchas tablas en la BD. Por otro lado, los atributos son los que van a definir o identificar las características de cada entidad. Los datos sintéticos serán de ayuda para simular el futuro o futuros alternativos de dichas entidades, lo cual facilitará el trabajo. Existe varios casos en donde esta estrategia es de gran beneficio, ya sea porque hay datos difíciles o costosos de adquirir o simplemente porque son de gran privacidad y no pueden ser revelados.

2. Método

En base al problema del proyecto, se ha pensado en cuatro entidades para poder crear a los conjuntos de datos, por medio de la Tabla 1, se ha podido clasificar a las entidades seleccionadas con sus respectivos atributos.

Cuadro 1: Entidades y atributos.

Administrador	**Call Center**	**Permisos**	**Servicios de Salud**
IdAdmi	IdCallCenter	IdPermiso	IdServicio
ClaveAdmi	NombreCallCenter	NombrePermiso	NombreServicio
	TelefonoCallCenter	DescripcionPermiso	DisponibleServicio
		StatusPermiso	DetalleServicio

Colab es un producto de Google Search, el cual ha permitido ejecutar, escribir y modificar el código de Python en el navegador. Para la generación de datos sintéticos se ha usado:

```
!pip install Faker
```

Faker es un paquete de Python que genera datos falsos para el usuario. Se ha definido un tamaño para el conjuntoo de datos, se ha asignado la cantidad de 5000 usuarios para este laboratorio, para definir esto, hay que modificar la línea de la siguiente manera:

```
num_users = 5000
```

También está las librerías que serán de gran ayuda para todo el laboratorio y nos ayudará con la generación de los datos:

```
import pandas as pd
import uuid
import random
from faker import Faker
import datetime
```

2.1. Administrador

Se procede a crear la lista de atributos de la primera entidad que es ***Administrador,*** el mismo tiene su respectivo ID y su Contraseña, para ello se genera sus respectivos códigos, para los dos se ha hecho uso de de la biblioteca **uuidd** que permite generar una cadena aleatoria de caracteres.

(Uuidd será usado para todos los atributos que tienen y solicitan la ID).

```
df['idadmi'] = [uuid.uuid4().hex for i in range(num_iess)]
print(df['idadmi'].nunique()==num_iess)
```

Para visualizar los datos, sin necesidad de descargar el excel, usamos:

```
df.head(5)
```

Dentro de los paréntesis de puede colocar cualquier valor, ya que solo es un ejemplo para visualizar si se pudo generar a los datos, como se puede observar en la Figura 1, se ha colocado el número 5, por lo tanto, tendremos 5 columnas.

```
df.head(5)
```

	idadmi	claveadmi
0	e52ca469f07c4a9ca09812e7dc41c983	50ff9dfb38e74353bb2a462d2f126957
1	d15bd4028e354e1f9ae0a86a36523051	fa30bcaf62404935ab3093c6aac7cdd1
2	012bc3bc871342a082a79d640c75a2da	e3e0cada8b534e9f82cceb97005a6f2a
3	af74eefda18a4e78bc2df505a94104ff	2beca8d68ccc4d62b26db2ca6e20010b
4	7fc4601da1114a45ba19e19622430f7f	7719399f81cd4c18ae28c282f78a9a4a

Figura 1: Datos administrador.

2.2. Call Center

La siguiente entidad es ***Call Center,*** para ella se han creado los atributos de id, nombre y teléfono, cabe recalcar que para añadir al teléfono se necesita la siguiente librería aparte de la principal:

```
import random as r
from random import seed
from datetime import datetime
```

Para crear nombres aleatorios primero se debe definir un genero, y seguido de eso, se podrá crear le código para que nos den nombres tanto de mujer, como de hombre:

```
faker = Faker()
def name_gen(genero):
    if genero=='male':
        return faker.name_male()
    elif genero=='female':
        return faker.name_female()
    return faker.name()
df['nombrecallcenter'] = [name_gen(i) for i in df['genero']]
```

Para generar un número telefónico que contiene diez dígitos, se colocan los npumeros opcionales del 0 a 9 y así se creará una cadena aleatoria:

```
numero=[]
size = 10
for i in range(0, num_iess):
  random.seed(datetime.now())
  valores = [0,1,2,3,4,5,6,7,8,9]
  numero=(''.join([str(random.choice(valores)) for i in range(size)]))
  df.telefonocallcenter[i]=numero
```

Se puede observar en la Figura 2, las cinco columnas de los datos generados con su ID, nombres de hombres y mujeres y un teléfono celular aleatorio.

```
df.head(5)
```

	idcallcenter	nombrecallcenter	telefonocallcenter
0	39d44ab6f4ad4d7fa6795bb018e67c9e	Ann Fry	4330085873
1	f2dd7c6e91de4944bcac897e3867a196	Deanna Smith	4538288101
2	6b150578cc7a4f93bd1216b7b6295d30	Savannah Potts	1968500932
3	725c2668a10f4f4f9942cc64a8edc061	Cody Zamora	1758864540
4	9996c92c08314a48b8df528156db03bc	Lisa Ramirez	7560286713

Figura 2: Datos Call Center.

2.3. Permisos

Otra entidad que ha sido seleccionada es la de ***Permisos,*** las cuales se van a asignar distintamente a los usuarios, para esto es útil las librerías de las dos anteriores entidades, y se repiten algunos pasos, los atributos que fueron asignados son: id, nombre, descripción y el status, con ayuda de los códigos anteriores se pudo modificar y usarlos para algunos de los aributos, pero se ha creado uno más que es para el status:

```
statuspermiso = ["Activo", "Desactivo"]
df['statuspermiso']= random.choices(
    statuspermiso,
    weights=(50,20),
    k=num_iess)
```

En la Figura 3 se observa cinco columnas de los datos generados con su ID, nombre de las opciones de permisos la descripción y el status del permiso.

```
df.head(5)
```

	idpermiso	nombrepermiso	descripcionpermiso	statuspermiso
0	0898dc501a5a4968981a193f78cb5281	Delete	2woN3vGPwSlW8shYMAKcvU2z8	Desactivo
1	deead44aafc042759e675e79e0c50eb4	Read	G3oGO2W1iqycxniYe7i1RES88	Activo
2	693c222180e44fb294a9dcc062cfd51f	Delete	HNbVET4NnVh67q674iqkboyVo	Activo
3	abb2db28ae344d81a9a5cebca9f361b9	Create	uZGjSb6VognEtQRAnKg6GoLjT	Desactivo
4	d7c49c878bab4e9fb398e266efc6edee	Read	CD7FoeWgI7vUaKUeASektpjBp	Activo

Figura 3: Datos Call Permisos.

2.4. Servicios de Salud

Finalmente, tenemos la entidad de ***Servicios de salud,*** las cuales va a disponer el sistema, se usarán las mismas librerías de las dos anteriores entidades, y algunos pasos se repiten nuevamente, los atributos que fueron asignados son: id, nombre, descripción y disponibilidad del servicio, por

medio de los códigos anteriores se pudo modificar y usarl para algunos de los aributos, se creó uno más para ver la disponibilidad del servicio que es el siguiente código:

```
disponibleservicio = ['Si','No']
df['disponibleservicio'] = random.choices(
    disponibleservicio,
    k=num_iess)
```

En la Figura 4 se observa cinco columnas de los datos generados con su ID, nombre de los servicios que dispone y su disponibilidad.

```
[81] df.head(5)
```

	idservicio	nombreservicio	descripcionservicio	disponibleservicio
0	1eea59ccc4f94294a36dcb74ab3d97f7	Enfermedad	GckL4exwzgQl0rW6RJIRRfgvg	Si
1	d35aceeebe4d465d87d5c9c468fe7271	Maternidad	wg8cS1i27J1wbzx9FMl6nTYWU	Si
2	fe2412bca6a54bbc8eeeb6259cf4aa7e	Enfermedad	mYCh9mG6ichKp5N8M38sUBZ4D	No
3	557e0431be7a4729b5223f1a00ee063a	Enfermedad	b8UobjvXCFwEsWSZSbwsUfE88	No
4	122ed52763c84b48a8d678787c140f79	Maternidad	vGcsVm3QOcJDt60qiKYympVxm	Si

Figura 4: Datos Servicios de Salud.

No olvidar que para generar el documento .xlsx y poder descargarlo ya con todas los datos datos sintéticos que han sido creados, antes se debe instalar paquete de Python que genera a los datos falsos y finalmente ejecutar todas las líneas.

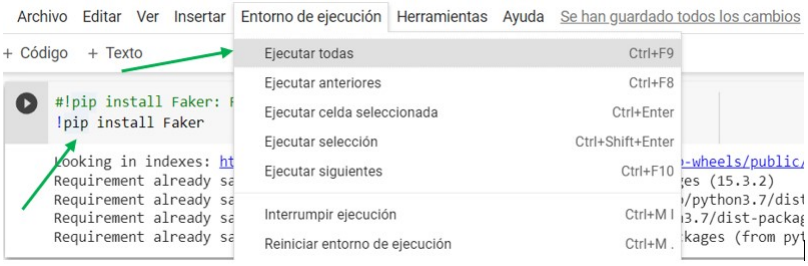

Figura 5: Paso Necesario.

3. Results and Analysis

Se ha podido obtener datos sintéticos para cada una de las características que se han seleccionado, se ha generado una tabla con una gran cantidad de datos que serán útiles para este laboratorio, ya que como son falsos, no afectarán a ninguna entidad real. Como se puede observar en la Figura 6, tenemos el archivo excel que se ha generado con el código principal antes de modificarlo en base a nuestro proyecto con los 100000 datos.

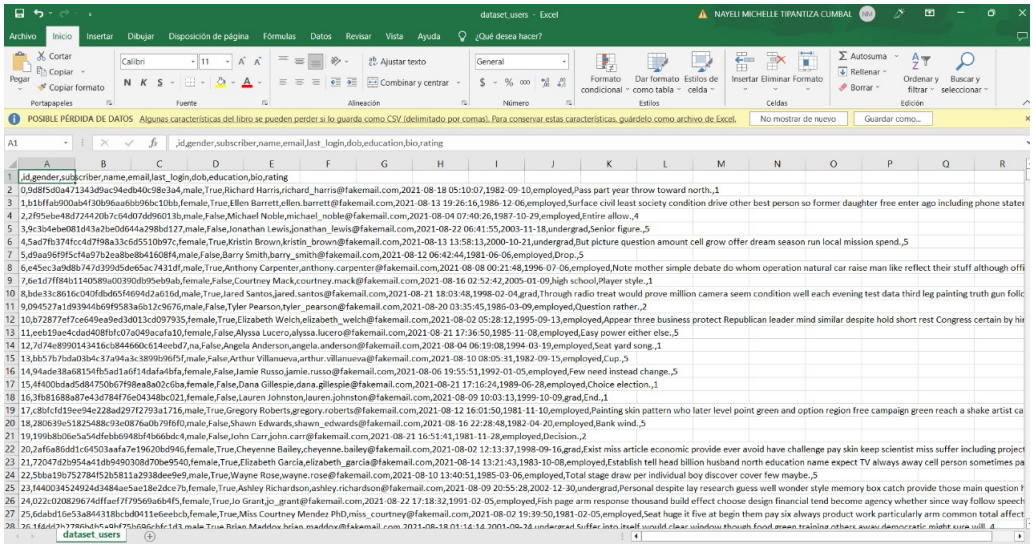

Figura 6: Excel principal.

4. Discusión

La creación de datos sintéticos ayudaron a modo de capacitación en crear datos para los atributos sin poner en peligro datos reales, son un ejemplo para crear una BD en el futuro generando datos aleatorios y probando diferentes maneras de manejar funciones.

5. Conclusión

Gran parte de los datos sintéticos ayudan a las empresas a poner en práctica ciertas características antes de implementarlos en una base de datos.

Es muy necesario generar datos artificiales para poder crear un entorno realista con ayuda de las entidades que se han seleccionado para el proyecto.

5.1. Link del repositorio GitHub

https://github.com/NayeliMTC/Sistemas-de-Bases-De-Datos-8393

Referencias

[1] Delgado Tenorio Manuel. ¿qué son los datos sintéticos? [Accessed: Nov 18, 2022].

[2] Korolov Maria. Datos sintéticos definidos. [Accessed: Nov 18, 2022].