**Estadística Computacional**

**Integrantes**

* Christian
* León
* José Luis Pérez Castellanos (CU: 159497)

**Resumen**

Proyecto final, cuya objetivo es el de la utilización de los siguientes softwares:

* Bash
* PostgreSQL
* Docker
* Docker Compose
* Shiny
* R

**Base de datos**

La base de datos que se analizará en este trabajo será la de [AirBnB CDMX](http://insideairbnb.com/mexico-city/).

**Explicación de variables**

* Para la variable de si han estado casados, Yes = 1 y No = 0.
* Para la variable de rural o urbano, Urban = 1, y Rural = 0.
* Para la variable de si ha fumado antes, never smoked = 0, formerly smoked = 1, smokes = 3 y Unknown = 4.
* No utilizamos la variable de working type.
* La variable a predecir es stroke, toma valores de 0 y 1, por lo tanto es un problema de clasificación lo que tenemos.

**Procedimiento de limpieza de datos**

El archivo inicial de la base es app/healthcare-dataset-stroke-data.csv, y en el Dockerfile lo trabajamos con awk, cut y sed para tener la base de datos limpia (líneas 8-20 del Dockerfile). Después de esto, utilizamos un python one-liner en el Dockerfile para pasar estos datos de .csv a .txt, *pero con el formato de json* (esto se hace en la línea 22 del Dockerfile ). Esto es importante para la carga de los datos a la base, ya que lo hacemos con curl especificando que son *json*.

**El Producto de Datos y el problema de negocio**

Se ofrece un producto de datos que se construya a través de los hospedajes vigentes en la plataforma de AirBnB para la Ciudad de México y su respectiva información. Además, se une a la información que genera la Procuraduría de la Ciudad de México, sobre las carpetas de investigación que se abrieron en el 2020 para todos los delitos.

**Problema de negocio**

Nuestro objetivo es el de ofrecer un producto de datos que permita a los inversionistas interesados en el sector inmobiliario en tomar una decisión de inversión. Con la aplicación podrán realizar una predicción sobre el precio promedio con el que podrían rentar a través de la plataforma de AirBnB. Lo anterior, tomando en cuenta las características propias de su propiedad y características asociadas a su geo ubicación en la Ciudad de Mexico. Además, de ofrecerle un panorama de seguridad de su colonia.

**Especificaciones**

Se solicita que la aplicación contenga los siguientes requisitos:

Base de Datos de soporte: Los datos deben de guardarse en un postgreSQL, esto para replicar la necesidad de motores externos de bases de datos en la vida real.

Ingesta inicial. La primer insersión de los datos a su base de datos debe de ser a través de Bash.

Modelo como Servicio Web. La API debe tener un modelo corriendo a manera de servicio.

Reentrenamiento del Modelo . Pasado un tiempo y con suficientes datos nuevos, usualmente se busca un reentrenamiento del modelo, por lo cual su API debe de ser capaz de hacerlo mediante una solicitud.

Ambiente totalmente reproducible. La API debe estar empaquetada por completo para que pueda correr sin ningún problema en cualquier computadora.

Captura de resultados del Modelo. Se busca entender el rendimiento y precisión del modelo, por ello es necasio desplegar sus resultados de desempeño, tal que el equipo de ciencia de datos pueda entender qué tan bien o mal está respondiendo.

**Infraestructura y Ejecución**

Para ejecutar este producto de datos se requiere:

**Para levantar la imagen de docker y la base de datos:**

1. Descargar el archivo…
2. Limpieza de datos:
3. Construir la imagen de docker:

**Para acceder a los servicios del producto de datos:**

1. Abrir el explorador de internet e ir a la siguiente dirección:
   1. localhost
2. Se accede a la página principal que contiene 4 botones con las siguientes funciones:
   1. Mostrar datos: Muestra la tabla disponible en la base de datos con el dataset utilizado para entrenar el modelo. **Nota:** El dataset se creó a partir de todas las propiedades que se encuentran disponibles en la plataforma de AirBnB
   2. Realizar predicción sobre el precio por noche: Permite realizar una predicción, al ingresar los campos más importantes para explicar el precio promedio
   3. Explicación de variables:
   4. Agregar registro a nueva base de datos: Permite agregar observaciones a la base de datos que se crea a partir de la información que proveen los usuarios, sin embargo, esta no se añade a la base de datos de AirBnB.
   5. Mostrar predicciones: Se muestran las predicciones realizadas hasta el momento.
3. Adicionalmente, se puede visualizar y trabajar con la base de datos utilizando el servicio de pgAdmin, para ello, ejecutar lo siguiente:
   1. Abrir el explorador de internet e ir a la siguiente dirección:
      1. localhost:8000
   2. Después de visualizar la pantalla de bienvenida de pgAdmin, ingresar los siguientes datos:
      1. username: [admin@admin.com](mailto:admin@admin.com)
      2. password: admin
   3. Después de entrar al servicio de pgAdmin, dar click derecho sobre Servers en el menú de la izquierda, seleccionar Create y posteriormente Server.
   4. En la ventana que se despliega, capturar la siguiente información:
      1. Pestaña General: Darle nombre al servidor, por ejemplo: Rodent.
      2. Pestaña Connection:
         1. Host name: db
         2. Username: root
         3. Password: root
   5. Estarán disponibles las siguientes tablas:
      1. all\_info: Contiene los registros del dataset de entrenamiento del modelo.
      2. predicted\_results: Contiene las predicciones realizadas.
4. Para salir de este producto de datos, hay que cerrar las pestañas del explorador y ejecutar Ctrl+C en la terminal donde se está corriendo la imagen de Docker.

**Re-Entrenamiento del modelo:**

1. El modelo lo podrán reentrenar solo los administradores del proyecto, esto debido a que el proceso se llevará a cabo a través de la actualización de los hospedajes que se encuentran disponibles en la plataforma de AirBnB, es decir, el reentrenamiento no se lleva a cabo con la información que se genera a través de las predicciones realizadas, sino por un proceso de actualización de la base de datos original.
2. Se crea una nueva base de datos con las proyecciones e intereses de inversión, así se podrá consultar esta información, eliminar registros que no generen información útil para el proyecto y actualizar registros.

**EDA**

Se puede consultar el análisis exploratorio de datos en la carpeta del respositorio.

**Entrenamiento**

Se puede consultar los detalles del entrenamiento del modelo de predicción en el código de R.

**Aspectos a evaluar**

**Limpieza de código**

**Documentación del código**

**Funcionamiento del código**