



Inteligencia artificial avanzada para la ciencia de datos

Reto

Etapas Big Data

Diego Arturo Padilla Domínguez - A01552594

Keyuan Zhao - A01366831

Carolina Herrera Martínez - A01411547

Cutberto Arizabalo Nava - A01411431

Jose Pablo Cobos Austria - A01274631

Campus Querétaro

14 de octubre de 2022

Herramientas y tecnologías aplicadas

Para poder trabajar con los datos de nuestro reto se hizo uso de las siguientes herramientas y tecnologías: Excel, Python, VsCode, Github, Oracle Cloud, entre otros. Y a continuación se explicará de forma más detallada el uso de cada herramienta.

Excel:

Esta herramienta fue utilizada como herramienta de visualización. Los datos que se manejaron venían en un archivo csv, y excel nos permite visualizar los datos de forma inicial sin la necesidad de tener que cargar los datos en código.

Python:

Seleccionamos este lenguaje de programación por que entre muchos otros, este tiene una gran afinidad con la ciencia de datos, por su simplicidad de sintáxis y todas las librerías (como Pandas y Numpy) que nos facilitan mucho la exploración, análisis y creación de datos.

Jupyter Notebook y Google Colab:

Para poder programar y elaborar nuestros scripts, se hizo uso de editores de textos e idle donde cada una de ellas poseía ventajas y limitaciones dependiendo de las tareas que queríamos realizar. En el caso de Jupyter Notebook, se montó una instancia en el servidor montado en Oracle para ejecutar el procesamiento de datos con el dataset completo. Hablando de Google Colab, este fue un entorno utilizado por nosotros para hacer pruebas y explorar el dataset tomando únicamente una porción pequeña de los datos.

Oracle Cloud

Para contar con un equipo potente de procesamiento, montamos un servidor en la nube de Oracle utilizando la capa de prueba que nos brinda un crédito de 6,000 MXN con vigencia de un mes. Elegimos esta nube sobre AWS debido a que, al finalizar la prueba de un mes, la capa gratuita de Oracle ofrece prestaciones muy superiores a las disponibles en la capa gratuita de AWS. Un ejemplo de esto lo vemos al crear instancias de servidores, ya que AWS nos ofrece máquinas gratuitas de 1 núcleo y 1Gb de RAM, mientras que Oracle en esta misma capa nos ofrece hasta 4 núcleos con 16 GB de RAM.

Github y Google Drive:

Para poder guardar nuestros datos, scripts y documentación del proyecto se utilizarán servicios en la nube para que se pueda trabajar de forma colaborativa y simultánea.

Todas estas herramientas que se mencionaron se usaron en combinación en diferentes maneras tareas, por ejemplo, se usó excel para poder analizar los datos de forma inicial antes de empezar a programar porque era lo más sencillo, no obstante por el tamaño del archivo no podíamos obtener mucha información de este, entonces cargamos el archivo en una jupyter notebook, donde mediante python con sus librerías pudimos obtener mayor información sobre nuestros datos, limpiarlos y transformarlos. Además, todos nuestros scripts y documentos del código y los datos se cargaban al final de cada etapa. De esa y otras formas se utilizaron las herramientas en conjunto.

Modelo de almacenamiento

Para el almacenamiento de los datos se optó por utilizar una unidad SSD en la nube de Oracle. En nuestra infraestructura actual solo contamos con una máquina virtual, por lo que se está utilizando el volumen de almacenamiento principal de esta máquina para alojar los datos del reto, incluyendo los datos iniciales y los datos transformados para su mejor procesamiento.

El acceso al servidor y su almacenamiento se encuentra restringido mediante el uso de claves privadas SSH para la conexión con el servidor. Estas claves no se encuentran publicadas en ningún medio, por lo que solo los integrantes del equipo cuentan con acceso a ellas.

El otro lugar en donde los datos son almacenados son en unidades compartidas de almacenamiento de Google drive, en donde su acceso está restringido a solo aquellos usuarios a los que les brindemos permisos.

Scripts para análisis, procesamiento y carga de datos

Los scripts utilizados para el análisis y procesamiento para determinar las comunas de acuerdo con las coordenadas se encuentran en el siguiente repositorio:

<https://github.com/A01411547/RetoMovilidadUrbana>.

Debido a razones de seguridad y de confidencialidad, los scripts utilizados para ejecutar nuestro análisis y procesamiento de todo el dataset se ejecutan en la nube de Oracle mencionado anteriormente, ya que gracias al rendimiento del servidor optimizamos el tiempo para el procesamiento de los datos.

Separación de los datos

En razón del estado actual del desarrollo de nuestro reto, por el momento todavía no realizaremos la separación de los datos.

Para la creación de los modelos optamos por crear dos datasets, uno para entrenamiento contando con un 85% del dataset completo y otro para la prueba del modelo con un 15%, en la parte de validación la haremos con las herramientas que nos brindan las librerías para hacer la segmentación previo al inicio del entrenamiento.

Determinación de si es o no Big Data

Después trabajar con nuestros datos, analizándolos, limpiándolos, transformándolos, entre otras tareas, hemos llegado a la conclusión de tomar un enfoque de Big Data por dos puntos principales:

- **Capacidad de procesamiento:** Al momento de programar ciertas tareas, debido a que los datos eran demasiados, cuando utilizábamos librerías como pandas para poder realizar tareas de manejo de dataframes, en ciertas ocasiones no teníamos la suficiente capacidad de procesamiento, por lo que los códigos se tardaban en ejecutar tiempos muy largos o incluso se llegaban a crashear.