Proyecto - Big Data

Objetivo

Aplicar técnicas para extracción, transformación, carga de datos a datos realistas de la vida cotidiana y generar predicciones a partir de esos datos depurados.

Descripción general

La realización de este proyecto busca que los estudiantes se expongan a las complejidades que implican obtener datos reales que provienen de múltiples fuentes. Se espera que realicen una investigación preliminar donde buscarán conjuntos de datos, abiertos o de su ámbito laboral, que provengan de múltiples fuentes.

Posteriormente, deberán preprocesar e integrarlos de manera que puedan ser utilizados para propósitos analíticos y predictivos.

Los resultados serán presentados en la clase final del módulo.

Entregable #1: Investigación preliminar (5%)

A diferencia de las asignaciones anteriores, en esta ocasión los estudiantes tendrán mayor libertad para definir los detalles del proyecto. Se espera que los estudiantes analicen múltiples fuentes de datos, ya sea relacionadas con su trabajo o bien fuentes abiertas. Dado que se utilizarán los datos para realizar cruces interesantes y, además, una predicción con un modelo automático, los estudiantes deberán analizarlos a la luz de los requerimientos.

De forma detallada, deberá abordarse lo siguiente:

- Fuentes de datos analizadas. Los estudiantes deben documentar qué fuentes de datos analizaron. Deberán escoger al menos dos fuentes de datos que se puedan cruzar exitosamente, para obtener un conjunto de datos de mayor riqueza de información. Los estudiantes deberán argumentar por qué realizaron la selección final. A manera de sugerencia, los estudiantes pueden tratar de utilizar datos del INEC, Programa Estado de la Nación, Ministerios de Gobierno (e.g. Economía, Educación), por nombrar algunos. El requerimiento estricto, eso si, es que se puedan cruzar unos con los otros.

- Descripción detallada de los datos. Solamente para los datos escogidos, deberán describir cada uno de los atributos contenidos. También deberá explicarse cómo se une un conjunto de datos a otro (e.g. por número de cédula). Los estudiantes podrán utilizar las técnicas ya aprendidas para mejorar el entendimiento de los lectores, por ejemplo, estadística descriptiva, distribuciones, etc.
- Objetivo predictivo. Deberá explicarse en detalle qué atributo de los datos se utilizarán como variable objetivo del modelo de aprendizaje automático. Esto servirá como el planteamiento del objetivo de investigación que se plantea, antes de iniciar la realización del proyecto.

Entrega: Archivo PDF en TEC Digital a más tardar el 15 de enero de 2020 a las 11:59 PM

Entregable #2: Proyecto programado (20%)

Una vez que se haya escogido los dos (o más) conjuntos de datos y se haya definido el objetivo de predicción, se procederá a la implementación de todo el código para respaldarlo.

El desglose funcional y por puntaje es:

- Cargado y preprocesamiento de datos (antes de cruzarlos). Se espera que los estudiantes desarrollen módulos en Python para cargar los datos escogidos y que se realice un proceso similar a la tarea 3. Este apartado cubre el desarrollo del código para cargarlos, crear las transformaciones de Spark necesarias a nivel de dataframe. Al no incluir escritura a la base de datos, en este punto, los estudiantes deberán realizar todas las pruebas unitarias necesarias para demostrar que los datos han sido cargados y preprocesados correctamente. La evaluación de esta sección se enfocará mayoritariamente en las pruebas.
  (7%)
- Materialización en Postgresql. Similar a la tarea 3, una vez que los datos estén preparados, deberán escribirse a una base de datos Postgresql. Para ello se desarrollará un programa principal en Python, cuyo uso deberá documentarse detalladamente en el PDF de instrucciones. Este programa utilizará todos los módulos anteriores y agrega la funcionalidad necesaria para escribir los conjuntos de datos antes y después de cruzados. Se espera que la estructura de la base de datos sea expuesta con claridad en las instrucciones, ya que la validación para calificación incluirá correr consultas SQL. Para ellos se requiere entender la estructura.
  (3%)
- Modelo de predicción. En un Jupyter Notebook, deberá desarrollarse al menos dos modelos de predicción sobre el objetivo planteado en el primer entregable. Los elementos típicos para el desarrollo correcto de modelos (vistos en el módulo de

aprendizaje) aplicarán aquí. Se espera que se agregue suficiente información de rastreo en el notebook para determinar que la implementación fue realizada correctamente.
(7%)

- Análisis de resultados. Derivado del entrenamiento de modelos, deberá analizarse los resultados de cada uno por separado, primero, y posteriormente una comparación entre ambos. Debe explicarse con detalle por qué alguno funciona mejor o no. Se deja a criterio de los estudiantes agregar cualquier elemento que consideren enriquezca este análisis de resultados.
  (3%)

Entrega: Código en Python con pruebas, Jupyter Notebook con análisis de resultados y Archivo PDF con instrucciones de ejecución en TEC Digital a más tardar el 21 de enero de 2020 a las 11:59 PM

Entregable #3: Presentación (5%)

Como fase final del módulo, los estudiantes presentarán los resultados a sus compañeros de clase. Los estudiantes tendrán 10 minutos para desarrollarlas. Deberán cubrir los dos temas primarios abordados, a saber, la integración de datos y la predicción.

Entrega: Presentación en PDF y/o Jupyter Notebook en TEC Digital a más tardar el 22 de enero de 2019 a las 11:59 PM