Proyecto - Big Data

Objetivo

Aplicar técnicas para extracción, transformación, carga de 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 tareas, 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 **(para dos problemas)**:

- 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 sí, 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 aborda, antes de iniciar el proyecto. **Se recomienda que el objetivo sea de predicción binaria.**

Entrega: Archivo PDF en TEC Digital a más tardar el Miércoles 3 de Julio de 2024 a las 11:00PM

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 ajustarlos de manera que sean fáciles de utilizar en la fase de predicción posterior. Este apartado cubre el desarrollo del código para cargarlos y 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. Una vez que los datos estén preparados, deberán escribirse a una base de datos PostgreSQL, utilizando el esquema de contenedor secundario. 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 ello se requiere entender la estructura.
  (3%)
- Modelo de predicción. Se deberá desarrollar al menos **dos modelos de predicción**, utilizando las bibliotecas de aprendizaje automático de Spark, sobre el objetivo planteado en el primer entregable de este proyecto. Los elementos típicos para el desarrollo correcto de modelos (vistos en el módulo de aprendizaje) aplicarán aquí. Se deberá generar un programa principal que se pueda ejecutar desde el contenedor con suficiente información de rastreo para determinar que la implementación fue realizada

correctamente. En esta sección se permite que los estudiantes utilicen Jupyter notebooks, agregados al repositorio de código entregado. En caso de utilizar esta opción, los estudiantes deberán agregar la configuración necesaria al Dockerfile entregado para que sea accesible desde la máquina real. Para cualquiera de las dos alternativas, deberá ser documentado, en detalle, en el PDF de documentación que se entregará. Todos los pasos para reproducir la ejecución del modelo deberán ser explicados en la misma.
(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: Repositorio de código que incluye configuración de contenedores, código en Python con pruebas, Jupyter Notebook (dentro del repositorio de manera opcional), análisis de resultados y Archivo PDF con instrucciones de ejecución en TEC Digital a más tardar el Viernes 19 de Julio de 2024 a las 11:00PM

Entregable #3: Presentación (5%)

Como fase final del módulo, los estudiantes presentarán los resultados a sus compañeros de clase. Los estudiantes tendrán 12 minutos (más tiempo de preguntas) para desarrollarlas. Deberán cubrir los dos temas primarios abordados, a saber, la integración de datos y la predicción.

La evaluación de la presentación, en base 100, se desglosa así:

- Participación (20)
- Claridad y Manejo Tiempo (10)
- Calidad Material (10)
- Descripción fase integración de datos (20)
- Descripción fase de predicción (20)
- Resultados (20)

Entrega: Presentación en PDF en TEC Digital a más tardar el Lunes 22 de Julio de 2024 a las 11:00PM