1. Descripción general

El objetivo de este proyecto es desarrollar un producto de analítica de datos sobre los resultados de las pruebas Saber 11 en el país. Se han identificado dos usuarios finales interesados en este producto:

1. Rectores y administradores de colegios (públicos y privados) interesados en conocer factores que afectan el desempeño de los estudiantes en las pruebas Saber 11.

2. Secretarías de educación de alcaldías y gobernaciones interesadas en conocer cómo sus estudiantes e instituciones se comparan entre sí y cómo se comparan con las de otros municipios y departamentos.

Seleccione uno de los dos usuarios finales y diseñe su producto pensando especialmente en ese usuario. Esto quiere decir que su desarrollo debe ser completo, pero debe estar especialmente enfocado en los intereses de este usuario. El nivel esperado de desarrollo de este producto es de **prototipo funcional**.

2. Roles

Para la realización de este proyecto se han contemplado los siguientes roles:

1. Ingeniería de datos.

2. Análisis de datos.

3. Ciencia de datos.

4. Análisis de negocio.

5. Tablero de datos.

6. Despliegue.

Cada miembro del grupo debe seleccionar 2 roles (3 si son un equipo de dos personas), y realizar las tareas asociadas a estos 2 roles. La calificación de cada miembro del equipo estará asociada no solo al resultado global del proyecto, sino a las tareas específicas de los roles tomados. **Los roles de cada miembro del equipo deben ser diferentes a los seleccionados en los proyectos 1 y 2 (excepto para equipos de 2 personas, en los que deben asegurar que los dos miembros del equipo hayan realizado todos los roles en los 3 proyectos)**.

3. Preguntas de negocio y plan de acción

Tarea 1

Determine dos preguntas de negocio que quiere resolver para su cliente seleccionado (solo 2). Identifique cómo puede resolver estas preguntas a través de visualizaciones de

los datos (descriptivo) y un modelo predictivo de clasificación o regresión basado en redes neuronales (puede incluir adicionalmente otros modelos si lo considera pertinente).

Roles involucrados: todos, con liderazgo de Análisis de negocio.

4. Datos

Para desarrollar este producto usted debe hacer uso de los datos disponibles en el portal de Datos Abiertos: `https://www.datos.gov.co/Educaci-n/Resultados-nicos-Saber-11/kgxf-xxbe` (note que los datos están actualizados a agosto de 2023).

Tarea 2 - Selección, limpieza y alistamiento de datos

Como el conjunto de datos original es relativamente grande emplee AWS Glue y AWS Athena para extraer un subconjunto de datos **relevante** para responder a la pregunta de negocio definida en la Tarea 1. Defina el conjunto en términos de valores específicos para fecha, ubicación y otras variables que haya seleccionado. Seleccione un conjunto de por lo menos 100mil datos. **Anuncie su selección por Slack**. No puede usar la misma selección de otro grupo. Cargue el subconjunto de datos en python, explore los datos disponibles y realice una limpieza cuidadosa. Identifique datos faltantes y decida un estrategia para su gestión. Asegúrese de que los datos queden en un formato que permita su posterior análisis. Documente los procedimientos de extracción, limpieza y alistamiento realizados.

Roles involucrados: Ingeniería de datos.

Tarea 3 - Exploración de datos

Realice un análisis exploratorio que permita describir estadística y visualmente el comportamiento de las variables a considerar. Calcule estadísticas descriptivas, realice histogramas, diagramas de caja, diagramas de dispersión, diagramas de violín y otros que permitan comprender cómo se comportan las variables. Documente el análisis realizado.

Roles involucrados: Análisis de datos.

5. Modelos

Tras explorar en detenimiento los datos y tener claras la preguntas de negocio, es hora de pasar a construir los modelos. Se espera que sean modelos de clasificación o regresión basados en redes neuronales. Tenga presente lo aprendido en la exploración de datos, así como el usuario final seleccionado y las preguntas a resolver.

Tarea 4 - Modelamiento

Aquí deberá explorar diferentes configuraciones de modelo, realizar ingeniería de características, emplear diferentes métodos de estimación, comparar y seleccionar las mejores

alternativas. Consulte bibliografía que le permita contar con elementos para proponer los modelos. No es necesario emplear todas las variables disponibles, pero todas las variables incluidas y sus relaciones deben estar correctamente justificadas. Como hay un buen número de variables, clientes y preguntas de negocio diferentes, se espera que el modelo desarrollado por cada equipo sea **único**. Evalúe su modelo usando métricas apropiadas. **Documente el modelamiento realizado y sus experimentos empleando MLflow**.

Roles involucrados: Ciencia de datos.

6. Producto

Tras explorar los datos y construir los modelos, es hora de diseñar y desarrollar el producto. El producto debe ser un tablero en Dash desplegado en la nube, usando una máquina virtual. El tablero debe ser de fácil uso y le debe permitir al usuario acceder a **3 visualizaciones** relevantes y emplear el modelo predictivo ingresando los datos apropiados. El tablero debe quedar desplegado empleando máquinas de EC2 o contenedores Docker en AWS.

Roles involucrados: Análisis de datos, Ciencia de datos.

Tarea 5 - Diseño y desarrollo del tablero

Empiece por diseñar el tablero: ¿qué valores debe permitir ingresar? ¿qué resultados genera? ¿qué visualizaciones incluye? ¿cómo mostrará las instrucciones? ¿cómo dispondrá estos elementos en el tablero? Para esta tarea es buena idea hacer un wireframe (un diseño sencillo que puede hacer en papel o digitalmente), que le permite tener una visión clara de su tablero y todos sus elementos. Recuerde no perder de vista al usuario y su necesidad. Piense siempre en la experiencia del usuario. Una vez haya terminado el diseño, desarrolle su tablero en Dash.

Roles involucrados: Tablero de datos.

Tarea 6 - Despliegue

El tablero debe quedar desplegado en la nube empleando máquinas de EC2 o contenedores Docker en AWS. Además, los **modelos** que se desplieguen deben estar **serializados** en archivos. Es decir, al ejecutar su tablero los modelos se deben cargar pre-entrenados de archivos locales. Asegúrese de que su tablero sea accesible y quede en ejecución.

Roles involucrados: Despliegue e Ingeniería de Datos.

7. Entregables

Como resultado de las tareas anteriores deberá entregar los siguientes resultados y soportes:

1. (**30 puntos**) Resultado 1: reporte de **máximo 7 páginas** con la documentación de las tareas, los resultados principales del análisis exploratorio de datos y la modelización.

2. (**30 puntos**) Resultado 2: presentación de **máximo 10 minutos** con los resultados principales del análisis exploratorio de datos y la modelización. Esta presentación debe incluir también un espacio para demostrar el tablero desarrollado.

3. (**40 puntos**) Resultado 3: tablero desarrollado en Dash y desplegado en la nube empleando contenedores.

4. (**5 puntos**) Reporte de trabajo en equipo: incluya un pequeño reporte de cómo se dividieron los roles entre los miembros del equipo.

5. Soporte 1: fuentes de extracción y limpieza (cuadernos de jupyter o archivos .py con la limpieza de datos). Pantallazos de AWS Glue y Athena correspondientes a la extracción de datos.

6. Soporte 2: fuentes de análisis (cuadernos de jupyter o archivos .py con el análisis exploratorio).

7. Soporte 3: fuentes de modelización (cuadernos de jupyter o archivos .py con la modelización desarrollada). Aquí debe incluir los pasos de entrenamiento, prueba y evaluación del modelo, así como los pantallazos de MLflow con los resultados de los experimentos realizados.

8. Soporte 4: fuentes del tablero (archivos .py del tablero desarrollado).

9. Soporte 5: snapshots de los recursos lanzados para el despliegue (AWS y terminal), y URL del tablero en ejecución.

10. Soporte 6: **repositorio Git** en Github, con un historial de commits que claramente refleje el aporte de cada miembro del grupo (de acuerdo con su rol). El repositorio debe estar estructurado con carpetas que reflejen las tareas definidas en el proyecto. Debe incluir una carpeta "despliegue" que contenga la última versión del tablero y permita lanzarlo, replicando el despliegue en AWS. Si emplea un despliegue con contenedores debe incluir un Dockerfile.

Nota: los soportes son parte fundamental de la entrega. Su no entrega lleva a una alta penalización.

Nota 2: si bien el trabajo es en equipo (de 2 o 3 personas), la nota es individual, luego es necesario que cada miembro del equipo demuestre su contribución al proyecto, tanto a través de los **commits en el repositorio**, como a través del **reporte** de trabajo en equipo y la **sustentación**.

8. Recomendaciones

1. El objetivo del proyecto es lograr un buen producto, bien soportado y claramente desarrollado. Justifique adecuadamente sus decisiones, observaciones y conclusiones.

2. Sea conciso y eficiente con el espacio. Ni el reporte ni la presentación deben ser largos. Al contrario, en un buen reporte cada gráfica y afirmación importa, y en un buena presentación cada diapositiva cuenta.

3. Es un trabajo en equipo. Defina los ítems de trabajo, asígnelos entre los miembros del equipo, defina fechas de entrega y revisión interna. Discuta los resultados, observaciones y conclusiones. Priorice tareas y resultados a incluir.

4. Empiece a trabajar prontamente y discuta con el instructor su avance y resultados.

Fecha de entrega: miércoles 29 de mayo