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. Para desarrollar este producto usted debe hacer uso de los datos enlazados en este enunciado. Identifique un usuario final que puede estar interesado en un producto de analítica basado en estos datos 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 y mantenimiento.

Cada miembro del grupo debe seleccionar 2 roles (3 si son un equipo de dos personas), y realizar las tareas asociadas a estos roles. La calificación de cada miembro del equipo estará asociada a las tareas específicas de los roles tomados y al resultado global del proyecto. **Los roles de cada miembro del equipo deben ser diferentes a los seleccionados en el proyecto 1**.

Si un estudiante realiza un rol esperado en este proyecto (porque ya lo realizó en el proyecto anterior), su nota será cero en ese rol.

3. Comprensión de negocio y datos: 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 basado en **redes neuronales** (puede incluir adicionalmente otros modelos si lo considera pertinente).

Roles involucrados: Análisis de negocio, Análisis de datos, Ciencia de datos.

Lidera (responsable): Análisis de negocio.

Tarea 2 - Selección, limpieza y alistamiento de 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 abril de 2024).

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. Deberá extraer el conjunto de datos que se le **asigne** a su grupo. Cada grupo tendrá un conjunto de datos diferente. 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. Genere nuevas características de ser necesario. 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. Genere nuevas características de ser necesario. Documente el análisis realizado.

Roles involucrados: Análisis de datos.

4. 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 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.

5. Producto y evaluación

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 de AWS, usando contenedores Docker. 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.

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 - Evaluación

Evalúe los resultados, considerando las preguntas de negocio, los *insights* obtenidos del análisis descriptivo, los modelos predictivos y el tablero desarrollado.

Roles involucrados: Análisis de negocio, Análisis de datos, Ciencia de datos, Tablero de datos.

Lidera: Análisis de negocio.

Tarea 7 - Despliegue y mantenimiento

El tablero debe quedar desplegado en la nube empleando 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 y mantenimiento.

6. Entregables

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

1. Resultado 1: reporte de **máximo 6 páginas** con la documentación de las tareas, los resultados principales del análisis exploratorio de datos y la modelización. Cada sección del proyecto debe indicar qué rol y miembro del equipo la realizó. El reporte debe describir cada tarea listada en el reporte.

2. Resultado 2: presentación de **máximo 10 minutos** con los resultados principales del proyecto, donde cada integrante debe presentar los resultados asociados a su rol. Esta presentación debe incluir también un espacio para demostrar el tablero desarrollado.

3. Resultado 3: tablero desarrollado en Dash y desplegado en la nube.

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

5. Soporte 1: análisis de negocio (soporte incluido en el reporte).

6. Soporte 2: 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.

7. Soporte 3: fuentes de análisis (cuadernos de jupyter o archivos .py con el análisis exploratorio).

8. Soporte 4: 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.

9. Soporte 5: fuentes del tablero (archivos .py del tablero desarrollado).

10. Soporte 6: snapshots de los recursos lanzados para el despliegue (AWS y terminal), y URL del tablero en ejecución.

11. Soporte 7: **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 cada una de 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, incluyendo un archivo 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**.

La calificación individual del proyecto se realizará de la siguiente manera:

1. **25 puntos**: contribución individual al reporte.

2. **25 puntos**: contribución individual a la presentación.

3. **50 puntos**: contribución individual de acuerdo con su rol, reflejado en los entregables y soportes asociados.

7. 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: domingo 25 de mayo de 2025