PROYECTO FINAL
SEGUNDA ENTREGA

El propósito de este proyecto es aplicar alguna metodología de ciencia de datos en conjunto con las diferentes técnicas y herramientas vistas durante el semestre para resolver algún problema o aprovechar alguna oportunidad identificada en una organización de su elección. Se recomienda ampliamente que se intente realizar un acercamiento a esta organización dueña de la problemática y de los datos, particularmente con aquellos *stakeholders* quienes puedan direccionar lo mejor posible los requerimientos de la solución de ciencia de datos a desarrollar.

OBJETIVOS DE LA SEGUNDA ENTREGA

- Finalizar las actividades de entendimiento de los datos y enfoque analítico así como el alcance general del proyecto.
- Realizar la preparación de datos requerida para la construcción de un modelo basado en machine learning.
- Entrenar un primer modelo de machine learning a partir del enfoque analítico definido y realizar una primera evaluación de resultados.

SEGUNDO SPRINT DEL PROYECTO

El segundo sprint del proyecto se debe enfocar en la preparación de datos y construcción de un primer modelo predictivo o explicativo basado en regresión, clasificación o alguna otra técnica analítica previamente discutida con los docentes. Complementariamente, se debe incluir una primera evaluación del mismo y generar nuevas conclusiones que permitan evidenciar la viabilidad y la continuidad del proyecto.

Se recomiendan al menos las siguientes reuniones de grupo:

- *Reunión de lanzamiento y planeación del sprint:* Para definir roles y forma de trabajo del grupo. Se genera una lluvia de ideas sobre las próximas actividades a desarrollar así como los criterios de aceptación de las mismas.
- *Reuniones de seguimiento:* Se recomienda mínimo una reunión de seguimiento

semanal corta. También pueden ser correos de avance según lo defina el grupo. Pueden tener un tablero de control para mayor visibilidad del estado actual de las tareas. Herramientas como Trello, Jira o Project son las más populares.

- *Reunión de finalización:* Para consolidar el entregable y analizar los aspectos a mejorar para el tercer y último sprint.

No olvide consolidar todo el trabajo de preparación y análisis de datos en el repositorio de GitHub. Se recomienda seguir buenas prácticas de versionamiento de documentación.

ACTIVIDADES DEL SPRINT

En este sprint se realizará una segunda iteración de la metodología ASUM-DM, con énfasis en los pasos de preparación de datos, modelado y evaluación. Dentro del entregable se debe incluir los siguiente:

1. **[25%] Preparación de datos**: Describa el proceso y muestre una evidencia de los datos preparados previos al entrenamiento de los modelos. Si realiza procesos de transformación como creación de nuevas características, codificación de variables categóricas, normalización, entre otros, repórtelos y justifíquelos adecuadamente.
2. **[15%] Estrategia de validación y selección de modelo:** Defina la estrategia de experimentación que seguirá para entrenar y seleccionar el mejor modelo que permita dar respuesta al problema planteado y que hará parte de la solución final. A partir de esta estrategia, separe los datos en conjuntos de entrenamiento, validación y prueba. Realice un breve reporte verificando que la distribución de los nuevos conjuntos de datos se conservan respecto al conjunto original.
3. **[20%] Construcción del modelo:** De acuerdo al enfoque analítico definido, entrene al menos tres modelos. Reporte los algoritmos e hiper-parámetros con los que experimentó y parámetros encontrados para cada uno de los modelos.
4. **[25%] Evaluación del modelo:** Realice la evaluación cuantitativa de los modelos teniendo en cuenta las métricas acorde al enfoque analítico y al tipo de modelo Reporte los resultados para los conjuntos de entrenamiento, validación y prueba. Detalle el proceso de análisis que siguió para seleccionar el mejor modelo. En la medida de lo posible, realice un análisis del error (evaluación cualitativa) y

establezca oportunidades de mejora de los modelos.

5. **[15%] Conclusiones**: Realice un resumen ejecutivo con los avances más relevantes del sprint. Algunas respuestas a preguntas que puede incluir en este resumen son: ¿Qué condiciones considera que deberían tener los datos para obtener mejores resultados? Más datos, diferentes características, etc. ¿Cuáles son las mayores dificultades que se han tenido en el proyecto? ¿Qué estrategias se plantean para mitigarlas? ¿El mejor modelo obtenido hasta el momento es suficiente para soportar el problema u oportunidad de negocio identificada? ¿Cómo se usará este modelo dentro del producto o solución que se construirá?

ENTREGA Y EVALUACIÓN

- El proyecto se debe realizar en grupos de 2 o 3 estudiantes.
- Debe entregarse un documento en formato PDF en donde se respondan claramente cada uno de los puntos descritos previamente. Debe tener máximo 8 páginas (incluida tabla de contenido y página de presentación), a una columna y con letra Arial tamaño 12.
- Debe incluirse un repositorio de GitHub con todo el código fuente desarrollado para la preparación de datos, construcción del modelo y evaluación. El repositorio debe estar debidamente documentado. Puede utilizar el archivo Readme para describir el contenido de los diferentes scripts que se hayan creado así como el detalle de las diferentes técnicas y herramientas utilizadas. Si se va a utilizar alguna herramienta no vista en clase, debe discutirse previamente con el profesor. **No se debe crear un repositorio nuevo, utilice el mismo que fue creado para la primera entrega**.
- La fecha máxima de entrega es el jueves 3 de noviembre a las 6:00 p.m.
- Para la sustentación durante la clase se debe contar con una presentación de no más de 5 diapositivas mostrando los avances más relevantes.