PROYECTO FINAL

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 tenga un contacto directo con algún integrante de la 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. El proyecto está dividido en dos entregas (dos sprints). En cada sprint se realizará una iteración de la metodología ASUM-DM, con mayor énfasis en algunas de sus fases.

El alcance general del proyecto debe incluir los siguientes aspectos:

- La solución planteada debe responder a las problemáticas u oportunidades de una organización real.
- El producto de datos debe estar constituido por al menos 3 de los siguientes componentes dependiendo de lo que sea más apropiado para cada caso particular: modelo de machine learning, API REST (o equivalente), aplicación web/mobile y/o dashboard. **Si se opta por utilizar técnicas o herramientas no cubiertas durante la clase, debe ser informado a los docentes de manera oportuna**.
- Se debe documentar apropiadamente el proceso y los resultados obtenidos en cada fase de la metodología, incluyendo la fase de retroalimentación proporcionada por los *stakeholders* de la organización.
- **El equipo debe estar conformado por 3 o 4 integrantes**. Se recomienda que sea lo más interdisciplinar posible. Al inicio del proyecto debe estar claramente definido y documentado el rol de cada integrante del equipo así como sus responsabilidades principales. Estas responsabilidades principales no indican bajo ninguna circunstancia que un integrante particular no deba aportar durante todas las fases del proyecto.

PRIMERA ENTREGA

Octubre 13, 11:59 PM

OBJETIVOS

- Proponer una solución a una problemática de una organización la cuál pueda ser abordada mediante la ciencia de datos y la elaboración de un producto de datos.
- Realizar un entendimiento del negocio y de la problemática a solucionar.
- Definir una primera propuesta de enfoque analítico a seguir, así como los elementos básicos del producto de datos a construir.
- Recolectar los datos necesarios y hacer un análisis exploratorio de los mismos buscando validar su calidad y suficiencia para la solución planteada.

ACTIVIDADES DEL SPRINT Y ENTREGABLES

En este sprint se realizará una iteración de la metodología ASUM-DM con énfasis en las fases de entendimiento del negocio, definición del enfoque analítico, recolección y entendimiento de los datos. Como entregable del sprint se debe incluir como mínimo el siguiente contenido:

1. **[10%] Definición de la problemática y entendimiento del negocio**: Seleccionar la organización con la cual se trabajará así como la problemática a resolver o la oportunidad a aprovechar a través de la ciencia de datos y la construcción de un producto de datos. Documentar la información clave del negocio (estrategia, datos del sector, etc.) que sustenta la relevancia del problema o la oportunidad. Definir los objetivos del proyecto y métricas de negocio (KPIs) que se usarán para su evaluación.
2. **[10%] Ideación**: Diseñar el producto de datos. Identificar sus potenciales usuarios, los procesos que desempeñan actualmente y sus dolores relacionados con la problemática o la oportunidad. Establecer los requerimientos del producto de datos a construir, los componentes que tendrá desde el punto de vista tecnológico así como un *mockup* del mismo.
3. **[10%] *Responsible*:** Identificar las posibles implicaciones éticas, de privacidad, confidencialidad, transparencia, aspectos regulatorios, entre otros, a considerar con el uso de datos y técnicas de IA en el contexto particular de la problemática abordada. No olvide citar claramente las fuentes consultadas.
4. **[15%] Enfoque analítico**: Definir las hipótesis o preguntas de negocio que guiarán el proceso de experimentación. Proponer las técnicas estadísticas, de visualización de datos y/o de *machine learning* que se aplicarán para dar respuesta a dichas

preguntas. Plantear las métricas que se utilizarán para evaluar la calidad del modelo.

5. **[10%] Recolección de datos**: Describir las fuentes de datos a utilizar en función de su estructura y utilidad. Se recomienda utilizar técnicas como los diccionarios de datos, al menos para las entidades o atributos más relevantes si es que se dispone de muchos.
6. **[35%] Entendimiento de los datos**: Generar un reporte de análisis exploratorio y calidad de los datos. Debe ser evidente el uso de diferentes técnicas de análisis de datos (univariadas/multivariadas/gráficas/no gráficas), así como el análisis de calidad desde sus diferentes dimensiones.
7. **[10%] Conclusiones iniciales:** Identificar un primer conjunto de conclusiones, *insights* y acciones próximas a ser ejecutadas.

ENTREGA FINAL

DICIEMBRE 1, 11:59 PM

OBJETIVOS

- Realizar la preparación y limpieza de datos requerida para la construcción del modelo de machine learning y/o dashboard planteado.
- Construir una primera versión del producto de datos y realizar una primera evaluación de resultados.
- Finalizar las etapas de modelado y evaluación, teniendo en cuenta la retroalimentación brindada durante la sustentación de la segunda entrega.
- Construir el producto de datos con los diferentes componentes establecidos durante la actividad de ideación.
- Presentar los resultados del análisis y el producto de datos a los *stakeholders* de la organización y obtener retroalimentación respecto a los aspectos positivos logrados y elementos a mejorar.

ACTIVIDADES DEL SPRINT Y ENTREGABLES

En este sprint se realizará una segunda iteración de la metodología ASUM-DM, con énfasis en las fases de modelado, operacionalización y entrega de resultados. Dentro del entregable se debe incluir los siguiente:

8. **[20%] Preparación de datos:** Describir el proceso y mostrar evidencia de los datos preparados previos al entrenamiento de los modelos y/o a la construcción del

dashboard. Si realiza procesos de transformación como creación de nuevas características, codificación de variables categóricas, normalización, entre otros, además de procesos de limpieza como imputaciones de datos faltantes, estandarización, entre otros, reportarlos y justificarlos adecuadamente. Para la descripción de todo el flujo de preparación de los datos, se recomienda realizar un diagrama de bloques funcional con los diferentes procesos implementados.

9. **[10%] Estrategia de validación y selección de modelo:** Definir la estrategia de experimentación que seguirá para entrenar y seleccionar el mejor modelo que hará parte del producto de datos planteado. A partir de esta estrategia, separar los datos en conjuntos de entrenamiento, validación y prueba. Realizar un breve reporte verificando que la distribución de los subconjuntos de datos se conservan respecto al conjunto original**.**
10. **[20%] Construcción y evaluación del modelo:** Entrenar múltiples modelos utilizando al menos tres algoritmos y diferentes conjuntos de hiper-parámetros. Reportar apropiadamente los resultados obtenidos realizando la evaluación cuantitativa de los diferentes modelos en los diferentes conjuntos de datos, teniendo en cuenta las métricas seleccionadas durante el enfoque analítico. En la medida de lo posible, realizar una evaluación cualitativa y establecer oportunidades de mejora de los modelos.
11. **[20%] Construcción del producto de datos:** A partir de lo establecido durante la actividad de ideación, construir un prototipo funcional del producto de datos el cuál debe estar compuesto por el mejor modelo de *machine learning*, API REST (o equivalente), aplicación web/mobile y/o dashboard. No olvide incluir el mecanismo de despliegue, así como un diagrama con la arquitectura de la solución.
12. **[5%] Retroalimentación por parte de la organización:** Presentar a modo de bitácora un resumen de las diferentes interacciones con los *stakeholders* de la organización, en donde se detallen los diferentes acuerdos llevados a cabo a nivel de definición de la problemática u oportunidad a abordar, producto de datos, enfoque analítico y resultados. Deben reportarse al menos tres (3) interacciones con los *stakeholders* durante el transcurso del semestre.
13. **[15%] Conclusiones:** Realizar un resumen ejecutivo con los resultados más relevantes del proyecto. Algunas respuestas a preguntas que se pueden incluir son:
    a. ¿Se cumplieron los objetivos del proyecto?
    a. ¿Cuáles fueron las mayores dificultades que se obtuvieron durante su desarrollo?
    b. ¿Qué estimación se puede dar respecto a cómo se impactarían las métricas de negocio (KPIs) definidas?
    c. ¿Qué condiciones considera que deberían tener los datos para obtener mejores resultados? Más datos, nuevas características, menor sesgo, etc.
    d. ¿El mejor modelo obtenido es suficiente para dar solución al problema u

oportunidad de negocio abordado?

14. **[10%] Autoevaluación y evaluación grupal:** Cada integrante debe completar la autoevaluación y evaluación de sus compañeros. La nota otorgada en este punto corresponderá al promedio de calificaciones otorgadas a título personal y por parte de sus compañeros.

FORMATO DE ENTREGA

- Todos los recursos generados deben entregarse mediante un repositorio de GitHub con su estructura apropiadamente documentada mediante el archivo Readme en el que se debe incluir como mínimo:
    - Resumen, conclusiones (*insights*)
    - Integrantes
    - Librerías / dependencias
    - Instrucciones de ejecución
- Debe incluirse un documento en formato PDF a una columna y con letra Arial 12 en donde se describa claramente los resultados de cada una de las actividades y entregables del sprint correspondiente. La longitud máxima permitida es.
    - Primera entrega: 10 páginas.
    - Entrega final: 15 páginas.
- **El documento debe tener un enfoque netamente ejecutivo** y debe ser elaborado de forma incremental. Es decir, para la entrega final se deben incluir los capítulos de la primera entrega ajustados de acuerdo a las recomendaciones realizadas.
- Durante las dos jornadas de sustentación, cada equipo dispondrá de máximo 10 minutos para realizar su presentación, más 5 minutos adicionales para preparación y preguntas. **Las diapositivas deben ser enviadas al correo de los docentes antes de iniciar la clase**. Todos los integrantes del equipo deben participar.
- Se penalizará a los grupos que no cumplan con las pautas aquí establecidas.