PROYECTO FINAL

El propósito de este proyecto es aplicar el conjunto de técnicas y herramientas vistas durante el semestre para desarrollar un proyecto de ciencia de datos en una organización de su elección. Se recomienda ampliamente que se mantenga un contacto directo y constante con algún integrante (*stakeholder*) dentro de la organización escogida, particularmente alguien quien pueda direccionar el alcance y expectativas del proyecto así como gestionar el acceso oportuno a los datos. El proyecto está dividido en dos entregas haciendo énfasis en diferentes etapas de la metodología ASUM-DM.

Dentro del alcance general del proyecto se deben incluir los siguientes aspectos:

- La solución planteada (producto de datos) debe responder a una problemática u oportunidad real dentro de una organización conocida.
- 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: (i) modelo de machine learning, (ii) API REST (o equivalente), (iii) aplicación web/mobile, o (iv) 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 4 integrantes**. Se recomienda que este sea lo más interdisciplinar posible.

PRIMERA ENTREGA

Octubre 19, 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 Y ENTREGABLES

En esta entrega se hará énfasis en las fases de entendimiento del negocio, definición del enfoque analítico, recolección y entendimiento de los datos. Se debe incluir lo siguiente:

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 mediante la ciencia 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 a abordar. Establecer los requerimientos del producto de datos a construir, los componentes que tendrá desde el punto de vista analítico y 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 analíticas en el contexto particular de la problemática u oportunidad 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 e importancia. Se recomienda utilizar técnicas como los diccionarios de datos, al menos para las entidades o atributos más relevantes.
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).
7. **[10%] Conclusiones iniciales:** Identificar un primer conjunto de conclusiones, *insights* y acciones próximas a ser ejecutadas.

ENTREGA FINAL

NOVIEMBRE 30, 11:59 PM

OBJETIVOS

- Realizar la preparación y limpieza de datos requerida para la construcción del modelo de *machine learning* y/o dashboard planteado.
- Desarrollar las etapas de modelado y evaluación de modelos.
- Construir el producto de datos funcional 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 logros alcanzados y próximos pasos.

ACTIVIDADES Y ENTREGABLES

En esta entrega se hará énfasis en las fases de modelado, evaluación y despliegue del producto de datos. Se debe incluir lo 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. No olvide detallar los diferentes procesos de transformación implementados como creación de nuevas características, codificación de variables categóricas, normalización, imputación de datos faltantes, estandarización, entre

otros. Se recomienda realizar un diagrama de bloques funcional con los diferentes procesos implementados.

9. **[5%] 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 teniendo en cuenta las métricas seleccionadas durante el enfoque analítico. En la medida de lo posible, realizar una evaluación cualitativa complementaria 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 obtenido, API REST (o equivalente), aplicación web/mobile y/o dashboard. No olvide incluir el mecanismo de despliegue así como el diagrama de arquitectura del producto de datos.
12. **[15%] 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) una vez el producto de datos sea utilizado por usuarios reales?
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. Estas evaluaciones se realizarán mediante un formulario el cual será enviado el día posterior a la fecha de entrega. La nota otorgada a cada integrante del equipo corresponderá al promedio de calificaciones otorgadas a título personal y por parte de sus compañeros.

INSTRUCCIONES DE ENTREGA

- El proyecto debe desarrollarse en grupos de 4 integrantes.
- Todos los recursos generados deben entregarse mediante un repositorio de GitHub. **El repositorio debe ser público.**
- Documente en el archivo Readme información relevante como integrantes del equipo, objetivo, alcance, conclusiones (*insights*), instrucciones de ejecución y dependencias.
- El repositorio debe ser auto-contenido, es decir, debe incluir de una forma clara y concreta el desarrollo de cada una de las actividades solicitadas (implementación + interpretación).
- El o los notebooks generados deben poderse ejecutar secuencialmente sin ningún error. En el caso de haber creado varios notebooks, se debe clarificar el orden en que deben ser ejecutados.
- Debe incluirse un documento en formato PDF a una columna y con letra Arial 12 en donde se describan claramente los resultados más importantes de cada una de las actividades. **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. La longitud máxima permitida es:
    - Primera entrega: 5 páginas.
    - Entrega final: 10 páginas.
- Adicional al documento, cada entrega debe incluir una sustentación en video en el que participen todos los integrantes del equipo. **Esta sustentación también debe ser de carácter ejecutivo. No olvide incluir las diapositivas en el repositorio**. El tiempo máximo permitido es:
    - Primera entrega: 5 minutos.
    - Entrega final: 10 minutos.
- Se penalizará a los estudiantes que no cumplan con estos criterios de entrega.