Proyecto I

Universidad de Medellín

Nicolás Gómez

MSc en Física

Especialista en Analítica y Ciencia de Datos

DS Engineer @ Mercado Libre

Hobbies: Videojuegos, pasear con mis perritas, leer

Objetivos del curso

- Desarrollar un proyecto de Ciencia de Datos con los conocimientos adquiridos durante la especialización hasta ahora
- El proyecto debe estar enfocado en solucionar una temática real
- Más que clases, la idea es realizar un acompañamiento a sus proyectos

Pasos para construir el proyecto

1. Seleccionar un proyecto realista (e.g. trabajo, grupo de investigación)
   Consejo: Busquen un tema que consideren interesante
2. Preguntarse: ¿Hay datos?, si los hay... ¿Es fácil acceder a ellos?, ¿Tienen buena calidad?
3. Definir el alcance y objetivo, es decir, **¿qué se quiere hacer?** y **¿hasta dónde se desea llegar?** *(construcción de un tablero con un buen storytelling que haga seguimiento de métricas importantes de negocio con una periodicidad dada, construcción de modelo con una métrica de desempeño definida, etc)*
4. Evaluar si el proyecto es realizable con los conocimientos y tiempos del curso
5. **Iterar...**

Definición de métricas: KPIs

- ¿Cómo se desea evaluar los resultados obtenidos?
- Paso crucial para definir el avance e impacto del proyecto
- Un proyecto de DS no es más que un proceso en el que se intentará **optimizar** “algo”, si ese algo no se tiene definido muy probablemente estaremos gastando energía en cosas que no nos ayudarán con nuestro objetivo

Preguntarse:

- ¿Es mi métrica muy general?, ¿Realmente mide lo que deseo medir?
- ¿Cómo se afecta la métrica por otros factores? ¿Tengo control sobre esos factores?
- ¿Qué tan difícil es medir mi KPI?, ¿Hay fuentes para construirla?
- ¿Que periodicidad tendrá dicha métrica?
- ¿Los KPIs estarán relacionados con la métricas para optimizar el modelo?
- ¿Los *stakeholders* están de acuerdo con esta métrica?

Recordar: ¡Esto es un proceso iterativo!

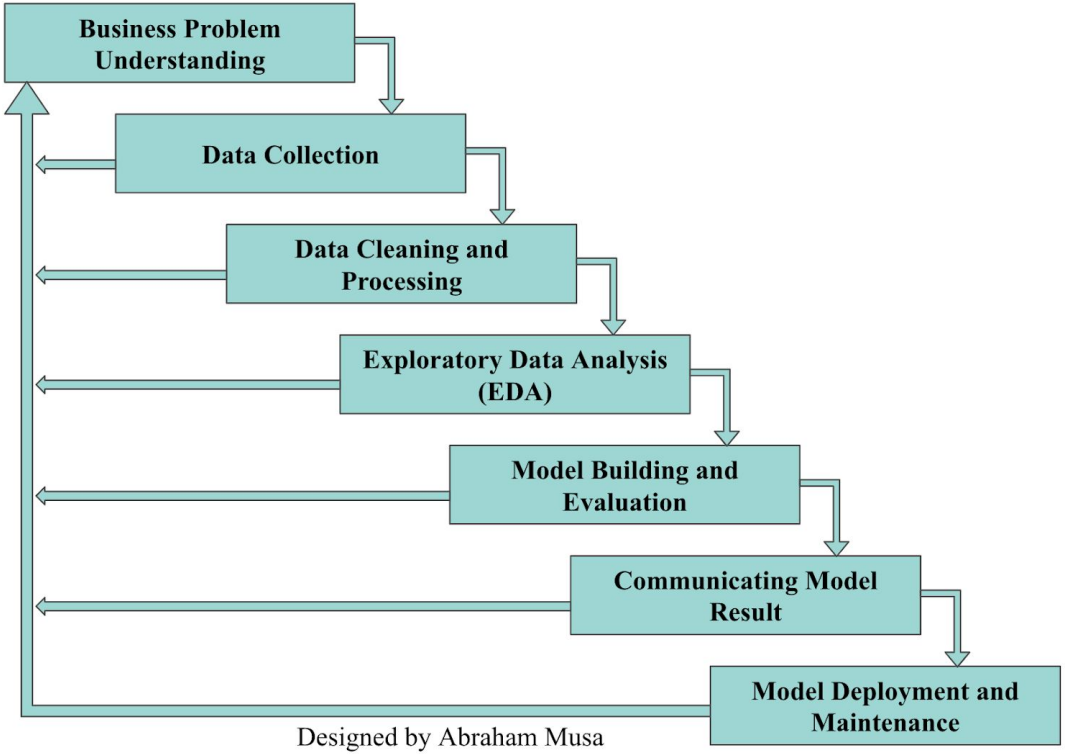
Fechas

- **30 de abril:** Definición de equipo (3 personas, máximo 4 con mayores expectativas) con temática (ya tener claro con qué dataset se va a trabajar)
- **5 y 6 de abril**: EDA, presentación de 20 minutos con los resultados (saber sintetizar la información en el tiempo data, se hará por equipos)
- **12 y 13 de abril:** Primeros avances de la solución (prototipo dashboard, baseline modelo, etc)
- **4 de Junio:** Entrega final con despliegue de la solución y presentación de resultados (tiempo a definir según cantidad de equipos)

Proyecto I

Universidad de Medellín

Ciclo de vida de un proyecto de Ciencia de Datos



Entendimiento del problema de negocio

- Definir y entender el problema que se quiere resolver
- Traducir los requerimientos del negocio en preguntas de ciencia de datos y accionables
- ¿Es posible resolver este problema haciendo uso de ciencia de datos?
- Hay antecedentes a este problema (¿qué resultados obtuvieron?, ¿se puede reutilizar algo?, ¿hubo aprendizajes para tener en cuenta?)
- Definir expectativas e impacto de la solución

Importante:

- Tener a la gente de negocio del lado es crucial para esto

Recolección de datos

Utilizar datos de una fuente confiable, ya que afectarán directamente el resultado de su modelo. Los datos de calidad son relevantes, contienen muy pocos valores faltantes y repetidos, y representan adecuadamente las diversas subcategorías/clases presentes.

GOV.CO https://www.datos.gov.co/

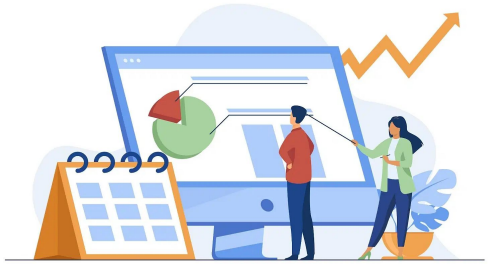
UCI Machine Learning Repository https://archive.ics.uci.edu/datasets

Google Trends https://trends.google.com/trends/

Limpieza y procesamiento de los datos

"Data cruda es de poco uso".

- Datos aleatorios para que no afecte en el aprendizaje de los modelos.
- Limpieza de los datos para eliminar valores no deseados, valores faltantes, filas, columnas, datos duplicados, conversión de datos.
- Visualización de los datos, correlación de datos, análisis multi-variable, feature engineering, etc.



pandas

plotly

NLTK
Express Your Creativity

seaborn

scikit learn

NumPy

matplotlib

statsmodels

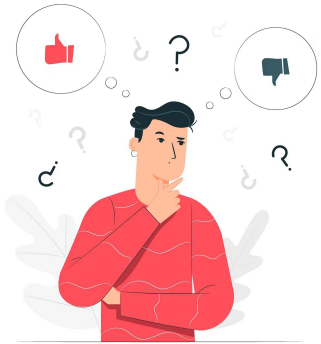
Análisis exploratorio (EDA)

1. Tener clara la pregunta ¿Qué queremos responder?
2. Tener una idea general del dataset.
3. Definir los tipos de datos que tenemos.
4. Elegir el tipo de estadística descriptiva.
5. Visualizar los datos.
6. Analizar posibles interacciones entre las variables del dataset.
7. Saber qué procesamiento se le hará a los datos en base a un análisis previo.
8. Extraer algunas conclusiones de todo el análisis.

https://codificandobits.com/blog/analisis-exploratorio-de-datos/

Construcción del modelo y evaluación

Es importante elegir un modelo relevante para la **tarea en cuestión** y saber elegir las **métricas adecuadas** para el modelo (existen métricas para clasificación, regresión, segmentación de imágenes, clasificación de imágenes, reconocimiento de voz, etc). NO se debe olvidar **optimizar el modelo** e identificar en base a que vamos a optimizarlo (accuracy, f1-score, etc).



Keras

scikit learn

mlflow

Hugging Face

XGBoost

OPTUNA

TensorFlow

spaCy

HYPEROPT

Comunicar los resultados del modelo

- Después de construir la solución es crucial compartir tus resultados a los stakeholders
- Procura saber a quién va dirigida la presentación, por ejemplo, a alguien de negocio, en general, no le importa los algoritmos que usaste, más bien están interesados en cómo los resultados pueden ayudar a propulsar los objetivos del negocio, en cambio, una persona más técnica verá mejor que expliques si usaste cierto modelo y cómo tuneaste hiperparámetros
- Un buen storytelling te ayudará a mostrar cómo los resultados del modelo ayudan a resolver el problema de negocio. Procura comunicarte de manera sencilla y directa, también usa ayudas visuales.

Despliegue del modelo

¿Cómo vamos a disponibilizar el modelo para que sea accesible a un usuario final?

En tiempo real

- Las predicciones son generadas y devueltas al usuario en el menor tiempo posible después de recibir la solicitud.
- Menor latencia.
- Consume recursos constantemente.

Predicción por lotes

- Se procesa gran cantidad de datos de entrada y el modelo genera las predicciones en forma asincrónica (l respuesta no es inmediata)
- Mayor latencia.
- Procesamiento diario, semanal, mensual, etc.

Mantenimiento de los modelos

¿Cómo sabemos que el modelo es acorde al negocio y sigue siendo funcional?

1. Evaluar el rendimiento del modelo para detectar posibles desviaciones ej: data drift.
2. Establecer estrategias de labelling en los datos y evaluar el performance.
3. Establecer un protocolo para identificar los errores de predicción y corregir si es posible.
4. Planificar ciclos de re-entrenamiento para adaptar el modelo a nuevos datos y tendencias emergentes. (La automatización de los pipelines puede ser de gran utilidad.)

EVIDENTLY AI

PANDAS PROFILING

Definición de métricas: KPIs

- ¿Cómo se desea evaluar los resultados obtenidos?
- Paso crucial para definir el avance e impacto del proyecto
- Un proyecto de DS no es más que un proceso en el que se intentará **optimizar** “algo”, si ese algo no se tiene definido muy probablemente estaremos gastando energía en cosas que no nos ayudarán con nuestro objetivo

Preguntarse:

- ¿Es mi métrica muy general?, ¿Realmente mide lo que deseo medir?
- ¿Cómo se afecta la métrica por otros factores? ¿Tengo control sobre esos factores?
- ¿Qué tan difícil es medir mi KPI?, ¿Hay fuentes para construirla?
- ¿Que periodicidad tendrá dicha métrica?
- ¿Los KPIs estarán relacionados con la métricas para optimizar el modelo?
- ¿Los *stakeholders* están de acuerdo con esta métrica?

Recordar: ¡Esto es un proceso iterativo!

Definición de métricas: KPIs

- ¿Cómo se desea evaluar los resultados obtenidos?
- Paso crucial para definir el avance e impacto del proyecto
- Un proyecto de DS no es más que un proceso en el que se intentará **optimizar** “algo”, si ese algo no se tiene definido muy probablemente estaremos gastando energía en cosas que no nos ayudarán con nuestro objetivo

Preguntarse:

- ¿Es mi métrica muy general?, ¿Realmente mide lo que deseo medir?
- ¿Cómo se afecta la métrica por otros factores? ¿Tengo control sobre esos factores?
- ¿Qué tan difícil es medir mi KPI?, ¿Hay fuentes para construirla?
- ¿Que periodicidad tendrá dicha métrica?
- ¿Los KPIs estarán relacionados con la métricas para optimizar el modelo?
- ¿Los *stakeholders* están de acuerdo con esta métrica?

Recordar: Esto es un proceso iterativo!