Actividades Bootcamp DS- 202403

December 3, 2024

Las actividades serán evaluadas de la siguiente forma:

- Cada alumno informará acerca de su repositorio Github para la revisión. En casos de fuerza mayor (sólo con una justificación razonable) de no tener cuenta de Github podrán enviar la actividad al correo: `dafbustosus@unal.edu.co`
- La revisión y retroalimentación de las actividades será máximo una semana luego de las fechas de Deadline
- Para la ponderación de la nota final las actividades 1 a 7 tendrán un peso del 60%, mientras que la actividad final (Proyecto final) un peso de 40%.
- La escala de notas será de 1.0 a 5.0. Para aprobar se deberá tener al menos 3.5 (70% del puntaje total)

1 Actividad I- Elección potenciales datasets

Consigna

1. Identificar 3 datasets que cumplan con las siguientes condiciones: a) al menos 2000 filas y b) al menos 15 columnas. Pueden buscar en las siguientes fuentes: GitLab, Github, Kaggle, Google Dataset Search (Si desean trabajar con un archivo propio se puede también). Algunas API recomendadas para obtener información: Marvel,PokeApi,CovidTracking,Nomics (Criptomonedas),Wheater API
2. Cargar los archivos correspondientes por medio de la librería pandas
3. Describir las variables potencialmente interesantes en cada archivo teniendo en cuenta el contexto comercial y analítico involucrado

Deadline: **10 Dic 2024**

2 Actividad II- Visualizaciones y EDA

Consigna

1. Deberán entregar el segundo avance de su proyecto final. Elegirán uno de los datasets del desafío I.
   Posteriormente crearán un notebook donde cargaran el archivo utilizando funciones de pandas para luego proceder a realizar 3 gráficos diferentes con Matplotlib/Seaborn/Plotly se debe tener al menos un gráfico con cada librería.
   Finalmente cada gráfico será interpretado con el fin de obtener insights relevantes que permitan dar respuesta a la pregunta problema.

Deadline: **7 Ene 2024**

3 Actividad III- Estructurando proyecto DS I

Consigna

1. Generar preguntas de interés o hipótesis de interés sobre el dataset elegido para el proyecto final.
2. Crear visualizaciones (univariados, bivariados o trivariados) junto con resúmenes numéricos básicos acordes con los tipos de variables disponibles.
3. Interpretar los resultados obtenidos

Deadline: **14 Ene 2025**

4 Actividad IV- Estructurando proyecto DS II

Consigna

1. Abstracto con motivación y audiencia
2. Preguntas/Hipótesis que queremos resolver mediante el análisis de datos
3. Análisis Exploratorio de Datos (EDA)
4. Con base en las visualizaciones y resúmenes numéricos generados del desafío anterior dar recomendaciones basados en los insights observados.
5. Para esta oportunidad se deberán tener avances en los apartados: Definición de objetivo, Contexto comercial, Problema Comercial, Contexto analítico, Exploración de datos (EDA)

Deadline: **28 Ene 2025**

5 Actividad V- Conexión a BD

Consigna

1. Deberán generar una instancia de conexión a Base de datos preferiblemente PostgreSQL donde deberá existir el dataset elegido para el proyecto
2. Se deben realizar consultas SQL desde el lenguaje Python utilizando las librerías (psycopg2 o sqlalchemy) y quedar evidenciadas en el archivo .ipynb que se entregue

Deadline: **11 Feb 2025**

6 Actividad VI- Obtención datos desde API

Consigna

1. Buscar información en APIs públicas (i.e Twitter, NewsAPI, Spotify, Google Apis, etc).
2. Extraer datos e importarlos a un dataframe realizando una exploración simple (i.e filas, columnas, tipos de datos). Se sugiere que estos datos complementen el dataset elegido en el Elección de potenciales Datasets e importe con la librería Pandas

Deadline: **18 Feb 2025**

7 Actividad VII - Data Wrangling (Limpieza datos)

Consigna

1. Se deberá realizar una limpieza de problemas (outliers, duplicados, valores nulos)
2. En caso de existir valores nulos decidir si se puede utilizar alguna técnica de imputación ya sea sensible o múltiple
3. Se deberán resolver el problema de outliers, para esto se deberá generar una estrategia de identificación ya sea con métodos tradicionales (IQ, Z score, etc) o utilizando aprendizaje de máquina (Isolation Forest, LoF, SVM one Class, etc).

Deadline: **4 Mar 2025**

8 Actividad VIII - Proyecto final

Objetivo general

Utilizar modelos de Machine Learning para resolver un problema de una industria o negocio

Objetivos específicos

1. Retomar el trabajo realizado en la segunda pre entrega, sumando el trabajo con Machine Learning
2. Modelar la situación como un problema de Machine Learning
3. Entrenar modelos de Machine Learning
4. Realizar ingeniería de atributos y normalización/estandarización de variables
5. Seleccionar el modelo con mejor performance

Requisitos base

1. Abstracto con motivación y audiencia: Descripción de alto nivel de lo que motiva a analizar los datos elegidos y audiencia que se podría beneficiar de este análisis.
2. Preguntas/Problema que buscamos resolver: Si bien puede haber más de una problemática a resolver, la problemática principal debe encuadrarse como un problema de clasificación o regresión.
3. Breve Análisis Exploratorio de Datos (EDA): Análisis descriptivo de los datos mediante visualizaciones y herramientas estadísticas, análisis de valores faltantes.
4. Ingeniería de atributos: Creación de nuevas variables, transformación de variables existentes (i.e normalización de variables, encoding, etc.)
5. Entrenamiento y Testeo: Entrenamiento y testeo de al menos 2 modelos distintos de Machine Learning utilizando algún método de validación cruzada.
6. Optimización: Utilizar alguna técnica de optimización de hiperparámetros (e.g gridsearch, randomizedsearch, etc.)
7. Selección de modelos: utilizar las métricas apropiadas para la selección del mejor modelo (e.g AUC, MSE, etc.)

Deadline: **18 Mar 2025**