Proyecto de Big Data y Machine Learning

¿Están listos para poner a prueba sus habilidades de Big Data y Machine Learning? Preparé cuatro proyectos super interesantes. Cada proyecto ofrece una oportunidad única para aplicar sus conocimientos y obtener experiencia práctica en áreas relevantes de la industria.

1. Análisis de Sentimiento en Twitter: ¡Descifren el Pulso de la Satisfacción del Cliente!

Objetivo: Descifrar el sentimiento detrás de los tweets de los clientes y predecir su nivel de satisfacción.

Datos:

- **Kaggle:** https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis

Tareas:

- **Exploración de Datos (EDA):** Visualicen la distribución de tweets por sentimiento, identifiquen palabras clave y hashtags relevantes.
- **Preprocesamiento de Datos:** Limpien y preparen los tweets para el análisis, eliminando ruido y normalizando el texto.
- **Extracción de Características:** Extraigan características relevantes del texto, como palabras clave, emoticones y hashtags.
- **Entrenamiento del Modelo:** Entrenen un modelo de aprendizaje automático (Naive Bayes, SVM) para clasificar tweets en categorías de sentimiento.
- **Evaluación del Modelo:** Prueben el rendimiento del modelo con métricas como precisión, exactitud y F1-score.
- **Visualización de Resultados:** Presenten los resultados de forma atractiva, mostrando la distribución del sentimiento por tweet y comparando productos o marcas.

2. Recomendación de Películas: ¡Conviértanse en Gurús del Cine!

Objetivo: Desarrollen un sistema de recomendación de películas personalizado para cada usuario.

Datos:

- **MovieLens:** https://grouplens.org/
- **IMDb:** https://www.imdb.com/

Tareas:

- **Exploración de Datos (EDA):** Analicen las calificaciones de películas por parte de los usuarios, identifiquen géneros populares y patrones de preferencias.
- **Preprocesamiento de Datos:** Limpien y preparen los datos de calificaciones para el análisis.
- **Cálculo de Similitud:** Comparen las preferencias de los usuarios para determinar similitudes entre ellos.
- **Recomendaciones:** Recomienden películas a cada usuario en función de las preferencias de usuarios similares.
- **Evaluación del Sistema:** Evalúen el rendimiento del sistema de recomendación con métricas como precisión, exactitud y recall.

3. Detección de Fraude en Tarjetas de Crédito: ¡Protejan las Finanzas del Mundo!

Objetivo: Desarrollen un modelo de aprendizaje automático para detectar transacciones fraudulentas con tarjetas de crédito.

Datos:

- **UCI Machine Learning Repository:** https://archive.ics.uci.edu/
- **Kaggle:** https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Tareas:

- **Exploración de Datos (EDA):** Analicen las características de las transacciones, identifiquen patrones y anomalías relacionadas con el fraude.
- **Preprocesamiento de Datos:** Limpien y preparen los datos de transacciones para el análisis.
- **Extracción de Características:** Extraigan características relevantes de las transacciones, como monto, ubicación, hora, tipo de transacción, etc.

- **Entrenamiento del Modelo:** Entrenen un modelo de aprendizaje automático (Random Forest, XGBoost) para clasificar transacciones como legítimas o fraudulentas.
- **Evaluación del Modelo:** Prueben el rendimiento del modelo con métricas como precisión, exactitud y F1-score.
- **Interpretación del Modelo:** Identifiquen los factores que más contribuyen a la detección de transacciones fraudulentas.

4. Análisis de Mantenimiento Predictivo: ¡Predigan Fallos y Eviten Costosos Reparaciones!

Objetivo: Desarrollen un modelo de aprendizaje automático para predecir el fallo de máquinas industriales.

Datos:

- **UCI Machine Learning Repository:** https://archive.ics.uci.edu/
- **Kaggle:** https://www.kaggle.com/datasets/shivamb/machine-predictive-maintenance-classification

Tareas:

- **Exploración de Datos (EDA):** Analicen los datos de sensores de las máquinas, identifiquen patrones y tendencias que indiquen un posible fallo.
- **Preprocesamiento de Datos:** Limpien y preparen los datos de sensores para el análisis.
- **Análisis de Series Temporales:** Analicen las series temporales de los datos de sensores

¡Bonus Track! Análisis de Datos de Spotify: ¡Descifren las Preferencias Musicales del Mundo!

Objetivo: Desentrañar las preferencias musicales de los usuarios de Spotify mediante el análisis de datos de canciones y escuchas.

Datos:

- **Spotify API:** https://developer.spotify.com/
- **Conjuntos de datos públicos de Spotify:** https://www.kaggle.com/datasets

Tareas:

- **Exploración de Datos (EDA):** Analicen las características de las canciones, como género, artista, popularidad, duración, etc.

- **Preprocesamiento de Datos:** Limpien y preparen los datos de canciones y escuchas para el análisis.
- **Análisis de Preferencias Musicales:** Identifiquen patrones en las preferencias musicales de los usuarios, como géneros favoritos, artistas populares, tendencias temporales.
- **Visualización de Resultados:** Presenten los resultados de forma atractiva, utilizando gráficos, mapas y otras herramientas de visualización.
- **Recomendaciones Musicales:** Desarrollen un sistema de recomendación de música personalizado para cada usuario en función de sus preferencias.

Observaciones.

- El trabajo es individual.
- Pueden obtener sus propios datasets, buscando en google dataset search o con algún método de scraping qué les parezca conveniente.

Entrega.

- El proyecto se entrega en una notebook jupyter (anaconda o colab). Es decir cada punto del proyecto debe estar escrito y comentado en la notebook.

- Deben subir la notebook correctamente comentada a su repositorio de github. Lo qué deben subir al classroom es el link a su repositorio.

Evaluación:

1. Comprensión del Concepto (20 puntos):

- Demuestra una comprensión clara de los principios y técnicas de Big Data y Machine Learning utilizados en el proyecto (10 puntos).

- Justifica adecuadamente la elección de las herramientas, métodos y algoritmos utilizados (10 puntos).

2. Aplicación Efectiva (30 puntos):

- Implementa correctamente las técnicas de Big Data y Machine Learning para abordar el problema planteado (20 puntos).

- Pre-procesa y manipula los datos de manera efectiva para prepararlos para el análisis (5 puntos).

- Selecciona y configura adecuadamente los parámetros del modelo (5 puntos).

3. Exploración de Datos (EDA) y Visualización (15 puntos):

- Realiza una EDA completa para comprender la distribución, patrones y tendencias de los datos (10 puntos).
- Crea visualizaciones informativas que comunican claramente los hallazgos de la EDA (5 puntos).

4. Entrenamiento y Evaluación del Modelo (20 puntos):

- Entrena el modelo de Machine Learning de manera adecuada utilizando los datos preparados (10 puntos).
- Evalúa el rendimiento del modelo utilizando métricas relevantes e interpreta los resultados (10 puntos).

5. Comunicación y Documentación (15 puntos):

- Redacta un informe claro y conciso que describe el proyecto, la metodología, los resultados y las conclusiones (10 puntos).
- Presenta el trabajo de manera efectiva utilizando presentaciones visuales y demostraciones (5 puntos).

6. Originalidad e Impacto (10 puntos):

- Demuestra creatividad en el enfoque del problema y la aplicación de las técnicas (5 puntos).
- Discute el impacto potencial del proyecto en el contexto del problema abordado (5 puntos).