Proyecto Final: Modelado y Evaluación de Datos

OBJETIVO GENERAL:

Desarrollar un proyecto de minería de datos basado en el ciclo CRISP-DM que permita a los alumnos aplicar técnicas de modelado y evaluación de datos para resolver un problema de negocio , utilizando herramientas de programación y visualización en Python, como pandas, numpy, matplotlib, seaborn, scikit-learn.

OBJETIVOS ESPECÍFICOS:

1. Implementar y evaluar diferentes modelo de regresión lineal para resolver el problema planteado en la primera etapa del proyecto.
2. Visualizar el resultados del modelado y la evaluación utilizando herramientas gráficas.
3. Documentar cada etapa del proceso, justificando las decisiones tomadas.
4. Presentar el análisis y los resultados de manera estructurada

DESCRIPCIÓN DEL PROYECTO:

Para la segunda entrega del proyecto final, los estudiantes deben abordar las siguientes etapas del modelo CRISP-DM:

1. **Descripcion del Problema y Pregunta de Investigacion (3 ptos).**
2. **Limpieza de los Datos (10 ptos)**
3. **Análisis de Datos (8 ptos):**
    - **Análisis de Datos para Justificar la Elección de los Modelos:**
      Tipo de Datos: Describir los tipos de datos presentes en el dataset (numéricos, categóricos, etc.). Esto incluye una clasificación detallada de los datos y su relevancia para el problema de negocio.
        - **Distribución de Datos:** Visualizar la distribución de los datos utilizando gráficos de seaborn y matplotlib. Incluir histogramas, gráficos de barras y gráficos de dispersión para mostrar cómo se distribuyen los datos y detectar posibles anomalías.
4. **Modelado de Datos y evaluación (40 ptos):**
    - **Selección y definición del modelo:** Definir al menos dos modelos para resolver el problema de minería de datos planteado en la primera etapa del proyecto. Por ejemplo:

- **Modelo de Regresión Lineal:** Utilizado para predecir valores continuos basados en una relación lineal entre las variables.
- **Modelo de Regresión Múltiple:** Extensión del modelo de regresión lineal que utiliza múltiples variables predictoras.
- **Modelo de Regresión Logística:** Utilizado para problemas de clasificación binaria.
- **Modelo k-Vecinos más Cercanos (k-NN):** Algoritmo de clasificación basado en la proximidad de los datos.
- **Clasificador Bayesiano Ingenuo (Naive Bayes):** Basado en el teorema de Bayes, adecuado para problemas de clasificación y regresión.
- **Árboles de Decisión:** Modelo de clasificación y regresión que utiliza un árbol de decisiones.
- **Redes Neuronales:** Modelos complejos que imitan el funcionamiento del cerebro humano para el reconocimiento de patrones.
- **Modelo K-Means:** Algoritmo de clustering que agrupa datos en k clusters basados en su similitud.

Definir las variables seleccionadas para cada modelo, y justificar su elección con base en su relevancia para el problema planteado.

- **Entrenamiento y Predicción:** Describir el proceso de entrenamiento y predicción para cada modelo implementado. Incluir detalles sobre la división del dataset en conjuntos de entrenamiento y prueba, y los parámetros utilizados para cada modelo.
- **Evaluación del Desempeño:** Presentar los resultados del entrenamiento y predicción, y la evaluación del desempeño del modelo utilizando al menos dos medidas de evaluación (accuracy, matriz de confusión, error cuadrático medio, curva AUC). Incluir gráficos y tablas que muestren el rendimiento de cada modelo.

5. **Conclusiones (10 ptos):**
   - **Conclusión General del Proyecto:**
     - Resultados obtenidos con respecto al problema planteado. Describir cómo los modelos implementados ayudaron a resolver el problema de negocio y qué hallazgos importantes se obtuvieron.
     - Descripción del modelo con mayor desempeño y las razones de su efectividad. Comparar los modelos implementados y explicar por qué uno de ellos tuvo un mejor rendimiento.

- Propuestas de mejoras o siguientes pasos a realizar como trabajo futuro. Sugerir posibles mejoras en el análisis y modelado de datos, y describir los pasos futuros que podrían tomarse para continuar con el proyecto.

- **Conclusiones por Equipo:**
  - Expresar su opinión sobre el desarrollo del proyecto, incluyendo:
  - Qué les pareció el desarrollo del proyecto. Reflexionar sobre la experiencia de trabajar en el proyecto y los aprendizajes obtenidos.
  - Qué se les facilitó y qué fue lo más complicado. Identificar las partes del proyecto que fueron más fáciles y las que presentaron mayores desafíos.
  - Explicar si les gustaría seguir trabajando en el proyecto y las razones detrás de su decisión.

ENTREGABLES:

1. **Reporte (6 ptos):**
   - Considerando el documento de la primera entrega, este deberá ser complementado con los puntos descritos en la descripción del proyecto: *Análisis de datos, modelado de datos y evaluación, y conclusiones*. Además, se debe incluir una *introducción* donde se resalte la importancia de resolver el problema o el impacto que tiene el tipo de problema que se resolverá.
   - Asegurarse de que el documento esté bien estructurado y presente la información de manera clara y concisa.
2. **Archivos (20 ptos):**

-00 Data Clean

Data_Set.csv

```
— 00_Data_Clean
│   ├── Data_Set.csv
│   ├── Data_Clean.ipnyb
— 01_preprocessing_results
│   ├── preprocessing
│   │   ├── 01_A_preprocessing.ipynb
│   │   ├── T_test_PCA.csv
```

```
        ├── T_test_final.csv
        ├── T_test_final_objetivo.csv
        ├── T_train_PCA.csv
        ├── T_train_final.csv
        ├── T_train_final_objetivo.csv
        ├── boxplots_numericas.png
        ├── histogramas_numericas.png
        ├── mi_pca
        │   └── pca_artifacts_bundle.zip
        ├── pca_metadata.json
        ├── pca_pipe_num.joblib
        ├── preprocessor_cat.joblib
        ├── scaled_boxplots_pca.png
        └── scaled_histogramas_pca.png
    └── preprocessing_production
        ├── 01_A_preprocessing_production.ipynb
        └── T_new_final.csv
├── 02_lineal_regression_results
    ├── regression_lineal
        ├── 02_A_regresion_lineal.ipynb
        ├── expected_columns.json
        ├── mi_regresion_lineal
        │   └── mi_reg_lin_artifacts_bundle.zip
        └── modelo_reg_lineal.pkl
    └── regression_lineal_production
        ├── 02_B_regresion_lineal_production.ipynb
        └── Regresion_lineal_nuevos_predicciones.csv
├── 03_model_evaluation
    ├── 03_A_model_evaluation.ipynb
    ├── expected_columns.json
    ├── mi_regresion_lineal
        ├── mi_reg_lin_artifacts_bundle
            ├── expected_columns.json
            └── modelo_reg_lineal.pkl
        └── mi_reg_lin_artifacts_bundle.zip

    └── modelo_reg_lineal.pkl
├── 04_regularization
    ├── 04_A_regularization.ipynb
    ├── 04_A_regularization_production.ipynb
    ├── Regresion_ganadora_nuevos_predicciones.csv
    ├── betas_post_pre_numericas.csv
    ├── coeficientes_modelos.csv
    ├── comparacion_regularizacion.csv
    └── reg_lin_ganador
```

```
│   ├── reg_lin_ganador_bundle
│   │   ├── metadata_modelo.json
│   │   └── modelo_elasticnet.pkl
│   └── reg_lin_ganador_bundle.zip
```

3. **Presentación (3 ptos):**
    - Realizar una presentación que contenga:
        - Introducción, el problema de negocio y de minería de datos. Explicar claramente el problema que se pretende resolver y su relevancia.
        - El análisis de datos y la selección de modelos. Describir el proceso de análisis de datos y cómo se seleccionaron los modelos implementados.
        - Los resultados del entrenamiento y predicción. Presentar los resultados obtenidos de los modelos y su evaluación.
        - La evaluación del desempeño de los modelos. Comparar el rendimiento de los modelos utilizando las métricas de evaluación seleccionadas.
        - Visualizaciones y gráficos de apoyo. Incluir gráficos y visualizaciones que ayuden a entender los resultados y la evaluación de los modelos.
        - Conclusiones generales y particulares.

NOTA:

1. Definir un título para el proyecto que sea coherente con el tópico seleccionado.
2. Asegurarse de que los documentos entregados estén organizados y presenten un formato limpio.
3. Incluir una portada en todos los documentos entregables.
4. Solo un miembro del equipo será responsable de subir la documentación.
5. Todos los integrantes del equipo deben estar familiarizados con el desarrollo del proyecto, ya que durante la presentación se les harán preguntas.
6. La retroalimentación hacia los otros equipos al momento de su presentación es vital, formará parte de rubrica de evaluación.
7. Algunos recursos para descargar datasets:

   Outscraper - get any public data from the internet

Kaggle: Your Machine Learning and Data Science Community