PRIMERA ENTREGA

PROYECTO ANALITICA DE DATOS

POR

AURA LUZ MORENO DÍAZ

JUAN JOSE MOLINA OCAMPO

UNIVERSIDAD DE ANTIOQUIA

2022

(1) describas el problema predictivo a resolver,

El banco Santander de origen español desea predecir la cantidad de clientes que podrían realizar una transacción conociendo que la cantidad monetaria no posee importancia; dichos resultados se presentarán en datos binarios, donde 1 es el resultado de una transacción exitosa y 0 el cliente no realizó una transacción sin importar el monto de la transacción realizada. Los resultados de este análisis permitirán obtener mucha información para resolver otros problemas como pueden serlo: cantidad de colaboradores para atender la demanda, dinero disponible, espacio de las instalaciones, satisfacción del cliente… entre otros aspectos relativos a la realización de una transacción.

(2) el dataset que vas a utilizar,

Santander Customer Transaction Prediction

https://www.kaggle.com/c/santander-customer-transaction-prediction

Data Description

You are provided with an anonymized dataset containing numeric feature variables, the binary `target` column, and a string `ID_code` column.

The task is to predict the value of `target` column in the test set.

File descriptions

- **train.csv - the training set.**
- **test.csv - the test set. The test set contains some rows which are not included in scoring.**
- **sample_submission.csv - a sample submission file in the correct format.**

(3) las métricas de desempeño requeridas (de machine learning y de negocio);

https://sitiobigdata.com/2019/01/19/machine-learning-metrica-clasificacion-parte-3/#

Presentemos primero las principales métricas de clasificación usadas en Machine Learning:

- **Matriz de confusión o error**
- **Precisión**
- **Recall o sensibilidad o TPR (Tasa positiva real)**
- **Precisión**
- **Especificidad o TNR (Tasa negativa real)**
- **F1-Score**
- **Área bajo la curva de funcionamiento del receptor (ROC) (AUC)**
- **Pérdida logarítmica**
- **Cohen's Kappa**

(4) un primer criterio sobre cual sería el desempeño deseable en producción.

https://www.kaggle.com/datasets/sorelyss/icfes-colombia-20182021

ICFES en Colombia

https://www.kaggle.com/datasets/dorbicycle/world-foodfeed-production

Quien se come la comida que producimos?

https://www.kaggle.com/datasets/mariaren/covid19-healthy-diet-dataset
Dieta saludable y covid 19

https://www.kaggle.com/datasets/ankanhore545/world-suicide-rates-20002019
Indices de suicido a nivel mundial de 2000 a 2019

https://www.kaggle.com/datasets/ankitpranay/global-emissions-from-agriculture-and-forest-land
Emisiones globales de la agricultura

Proyecto

Tendrás que hacer un proyecto de analítica de datos para el cual deberás:

1. **Definir un problema predictivo**
2. **Obtener un dataset para resolverlo**
3. **Realizar el preprocesado y limpieza de datos**
4. **Encontrar los mejores hiperparámetros para DOS algoritmos predictivos**
5. **Encontrar los mejores hiperparámetros para DOS combinaciones de algoritmo no supervisado + algoritmo predictivo**
6. **Realizar las curvas de aprendizaje para cada uno de los cuatro casos anteriores**
7. **Realizar una evaluación diagnóstica que contenga:**
    1. **Para cada uno de los cuatro casos anteriores un diagnóstico de overfitting/bias etc.**
    2. **Una recomendación justificada de qué pasos a seguir si tuvieras que intentar mejorar el desempeño obtenido.**
    3. **Una evaluación sobre los retos y condiciones para desplegar en producción un modelo (como establecerías el nivel de desempeño mínimo para desplegar en producción, cómo se desplegaría en producción, cómo serían los procesos de monitoreo del desempeño en producción)**

Reglamentación del proyecto¶

- **Podrá hacerse individual o formarse grupos de 2 o 3 estudiantes.**
- **Tendrás que hacer CADA entrega en un repositorio github propio. TODOS LOS MIEMBROS DE CADA EQUIPO HAN DE DEPOSITAR UNA COPIA DE CADA ENTREGA EN UN REPOSITORIO GITHUB PROPIO.**
- **Usa los nombres y formatos indicados en el ejemplo más abajo. SI USAS OTRO NOMBRE O FORMATO LA ENTREGA NO SERÁ VÁLIDA. Para los informes, sólo se aceptan documentos en PDF.**
- **Cada miembro del grupo tiene que tener su propio repositorio github con una copia de las distintas entregas.**
- **Los videos de las distintas entregas han de subirse a YouTube. No es necesario que todos los miembros suban todos los videos. Una copia de cada video en YouTube es suficiente.**
- **Tienes que incluir un fichero README.md que contenga:**
    - **Nombre, cédula y programa matriculado de cada integrante del proyecto.**
    - **En enlace a la fuente de los datos usados.**
    - **Cómo obtener los datos y hacerlos disponibles en los notebooks cuando se ejecutan en Colab. No es suficiente con el enlace anterior, es necesario incluir estas instrucciones.**
    - **Los enlaces a los videos en YouTube de cada entrega.**

- **Se pondrá a disposición un formulario en el que cada estudiante indicará cual es su repositorio de github.**