Catedrático: Ing. Pren Biba
Auxiliar: Ing. Gladys Rodríguez

PROYECTO 2

Objetivo

Aplicar todos los conceptos y métodos aprendidos durante el curso para resolver un problema de predicción.

Requisitos

1. Seleccionar dataset (Tarea 2)
2. Análisis exploratorio: debe realizarlo para todo el dataset elegido, mostrando información relevante. Mostrar resultados y gráficas
3. Selección de Variables:
    a. Variable a predecir: ***tipo categórico***
    b. Variables predictoras: variables que considere esenciales.
4. Ingeniería de características: deberá desarrollar todo el procedimiento de ingeniería de características que considere necesario para resolver el problema de predicción:
    a. Imputación de variables con data faltante:
        i. Numérica
        ii. Categórica
    b. Codificación de variables categóricas
    c. Transformación de variables numéricas
    d. Tratamiento de outliers
    e. Estandarización de variables
5. Desarrollo de modelo de clasificación: deberá desarrollar y analizar todos los algoritmos vistos en clase con validación cruzada de Kfolds, dedicionalemnte deberá determinar cual es la mejor configuración de hiperparámetros para cada tipo de modelo:
    a. Naive Bayes
    b. LDA
    c. Regresión logística
    d. SVM
    e. Árboles de decisión
    f. Random forest
    g. Análisis de discriminante lineal
    h. Análisis de discriminante cuadrático
    i. AdaBoost
    j. XGBooot
    k. LGBM
6. Modelo Final: para seleccionar el algoritmo que resuelve de forma óptima el problema (basándose en la matriz de confusión)

Entregables

- Link de github:
  - Dataset
  - **Notebook #1:** con Análisis Exploratorio con tablas y gráficas de sus datos.
  - **Notebook #2:** con el análisis, procedimiento y construccion del Pipeline de ingeniería de características.
  - **Notebook #3:** con análisis de modelos, selección de mejor modelo y selección de mejor combinación de parámetros para cada tipo de modelo.
  - **Script:** Código de Python el cual debe incluir dos funicones:
    - Train_model: función para generar el entrenamiento del pipeline de ingeniería de carácterísticas y mejor modelo según lo selección. La salida debe generarse en un archivo de texto donde se epecifique la fecha y hora en la que se corrio la función, asi como el tiempo de entremaiento en segundos, las métricas de entrenamiento, Accuracy, Specificity, Sensitivity, ROC-AUC.
    - Predict: función que debe generar un arhcivo csv con las predicciones de un dataset que recibe como parámetro.
    - La ejecución de estas funciones debe ser por medio de una consola donde usted especifique que operación desea realizar.
  - Link de video:
    - Duración: 5 a 10 minutos (máximo)
    - Todos los participantes deben exponer en el video
    - Mostrar código desarrollado
    - Mostrar graficas resultantes
    - Conclusiones del proyecto