Aprendizaje automático
Proyecto sobre regresión
Prof. Carlos B. Ogando M.

PROYECTO INFERENCIA DE INGRESOS

HONESTIDAD ACADÉMICA

Como de costumbre, se aplica el código de honor estándar y la política de probidad académica. Las presentaciones isomórficas a (1) las que existen en cualquier lugar en línea, (2) las enviadas por sus compañeros de clase, o (3) las enviadas por los estudiantes en semestres anteriores serán consideradas plagio.

INSTRUCCIONES

En este proyecto desarrollarán un **modelo de inferencia de ingresos** usando datos de la nómina pública del estado, completando el pipeline de data science explicado en clase y comparando distintos métodos de preparación de datos, EDA, modelamiento y validación.

I. Resumen

II. Detalles de implementación

III. Requerimientos de la entrega

I. Resumen

El objetivo de este proyecto es desarrollar un **modelo de inferencia de ingresos** que estime el ingreso de una persona en base a sus características usando un modelo de Machine Learning supervisado para regresión. Para esto se entrenarán distintos modelos de regresión, comparando su performance, hiperparámetros para finalmente seleccionar el mejor de todos y poder hacer predicciones con el modelo.

Esto se realizará empleando la librería scikit-learn con Python.

II.- Detalles de implementación

El pipeline de un proyecto de Machine Learning supervisado consiste en lo siguiente:

1. Recolección de la data.
    a. Importar el dataset crudo.
2. Preparación de la data/ preprocesamiento la data.
    a. Estandarizar formatos y homigenizar datos.
    b. Ingeniería de características (eliminación de características redundantes o innecesarias)
    c. Limpieza de filas nulas, vacías o con error.
    d. Aplicar un encoder o codificador a las características no numéricas. (Guardar diccionario de codificación)

    e. Normalizar y estandarizar la data con un escalador de datos. (Convertirlos en datos con media cero y desviación estándar uno)
3. Análisis descriptivo de la data (EDA)
    a. Analizar la data con gráficas.
    b. Interpretar las estadísticas de los datos.
    c. Interpretar patrones de los datos con consultas y métodos de visualización.
4. Entrenamiento del modelo.
    a. División del dataset en entradas y salidas/etiquetas (x, y).
    b. División del dataset en entrenamiento y testeo.
    c. Entrenamiento de cada algoritmo con el dataset.
5. Validación y testeo del modelo.
    a. Análisis de performance (matriz de confusión, precisión, recall, accuracy, etc)
    b. Selección de algoritmo óptimo.
6. Despliegue del modelo y comprobación con data recién creada.
    a. Conversión de data nueva cruda a formato de entrada del algoritmo (codificación, escalado, etc)
    b. Predicción de categoría del dato.

La librería scikit-learn con el apoyo de pandas, numpy y matplotlib cuenta con múltiples módulos para realizar estas funciones de forma sencilla en Python.

III.- Requerimientos de la entrega

Recopilarán las nóminas públicas de distintas instituciones y realizarán un EDA (Análisis Exploratorio de la Data) para descubrir insights en la data.

A nivel de recopilación, preparación y análisis de datos debe cumplir los siguientes requerimientos:

- Deben recopilar mínimo cinco nóminas distintas de instituciones diferentes.
- El dataset debe tener al menos 5000 filas después de la limpieza.
- Debe tener al menos 6 características de entrada y una etiqueta de salida después de la limpieza.
- La etiqueta de salida debe ser de tipo real.
- Debe haber al menos una característica de entrada tipo entero, una tipo decimal y una tipo categórica.
- Deben cargar estas nóminas en python usando pandas.
- Deben concatenar estas nóminas en una sola, formatearlos y homogenizarlas usando pandas.
- Deben realizar un análisis de la calidad de datos (cantidad de celdas nulas, cantidad de celdas mal formateadas, entre otros)
- Obtener estadísticas básicas de las columnas (promedio, mediana, mínimo, máximo, desviación estándar)

- Ver las distribuciones de los datos de cada columna.
- Analizar estas distribuciones en conjunto con otras variables (dist. de ingreso por género, por cargo, por institución, entre otros)
- Identificar correlaciones entre las variables.
- Otros hallazgos.

Este proyecto les servirá de introducción a las librerías de procesamiento y análisis de datos: Pandas, Matplotlib y Numpy.

Debe entregar un cuaderno de Jupyter con el código fuente con el se entrenaron los modelos y se hicieron las pruebas.

A nivel de modelamiento debe cumplir los siguientes requerimientos:

- Debe entrenar los siguientes modelos de regresión con el dataset:
  1. Ordinary Least Squares Regression
  2. Ridge Regression
  3. Bayesian Regression
  4. Lasso Regression
  5. Nearest Neighbors Regression
  6. Decision Tree Rregression
  7. Random Forest Regression
  8. SVM (Support Vector Machine) Regression
  9. Neural Network MLP Regression
  10. Ada Boost Regressor
- Debe mostrar las métricas de rendimiento de cada uno de los modelos.
- Se debe poder hacerle pruebas a cada uno de los modelos ingresando un archivo ".csv".
- Las fuentes deben estar debidamente documentadas con docstrings y anotaciones.

Debe entregar un **documento de infraestructura** del proyecto que contenga:

- Gráficas analíticas de modelo (EDA).
- Proporción de testeo/entrenamiento.
- Descripción de algoritmos empleados.
- Hiperparámetros usados en los modelos.
- Estadísticas de modelos.
- Explicación de cómo funciona el sistema de regresión creado.
- Importancia de características.

Debe exponer el proyecto en clase, de lo contrario, no valdrá puntos.