Universidad Nacional de La Matanza

Departamento de Ingeniería e Investigaciones Tecnológicas

Trabajo Práctico

Aprendizaje de Máquina

Caso Aplicativo Conceptos Regresión Lineal y Logística

Fecha de Entrega: 30/05/2023

Hora de Entrega Límite: 23:00 HS

Profesores:

Dr. Ierache, Jorge

Ing. Becerra Martín

Ing. Sanz Diego

Trabajo Práctico ML - Regresión Lineal - Predicción - Eficiencia Energética

Cuestionario de Regresión Lineal

Considerando Google Colab, como base para el desarrollo del contenido del TP:

Trabajo Práctico ML - Regresión - Predicción - Energy Efficiency

y Dataset correspondiente: Dataset Eficiencia Energética

Se deberán desarrollar las modificaciones necesarias sobre el código inicial y el dataset atravesando los diferentes pasos que componen la **Metodología CRISP DM**, que para este caso nos enfocaremos en: **Selección de Features, Separación de Set de Datos y Normalización de los datos.** A medida que avanzamos en los pasos debemos utilizar las configuraciones realizadas en el paso actual, es decir, se irán encadenando las configuraciones para finalmente obtener el modelo más óptimo. Ejemplo: Una vez encontradas las features más relevantes para el análisis, debemos usar esas features para quedarnos solo con esas columnas al modificar la separación de datos del paso siguiente; y finalmente, con las features del 1) y la proporción de datos para el set de datos prueba del 2) utilizarlos para configurar 3) junto con su correspondiente normalización. **Se deberá ir evolucionando el código en Python aplicando cambios por cada punto.**

1. Selección de Features (correlation matrix & mutual info): Analizamos relación existente entre los diferentes atributos del dataset. Generalmente, al seleccionar features con mayor interrelación entre sí puede indicarnos que son más relevantes para ser analizados por nuestra modelo para llevar a cabo una adecuada predicción, ¿Si reducimos el número de feature cols a analizar por nuestro modelo obtendremos una mayor precisión al evaluar el mismo? Justifique su respuesta. Ver referencias:
* correlationmatrix
* mutual info
* mutual info de Scikit Learn

2. Separación de Set de Datos (train_test_split - test_size): Al momento de entrenar nuestro modelo, previamente debemos separar el set de datos total en: 1) Set de Datos de Entrenamiento, y 2) Set de Datos de Prueba. Generalmente, al aumentar la cantidad de datos para el set de datos de prueba, ¿Aumenta el score y disminuye el error en la etapa de evaluación del modelo? Justifique su respuesta. Ver referencias
train_test_split

3. Normalización (z-score): Para mejorar el procesamiento de las operaciones que lleva a cabo nuestro modelo cuando entrena debemos normalizar los datos con los que vamos a trabajar. Generalmente, al normalizar nuestros set de datos luego de su separación en set de datos de entrenamiento y de prueba, ¿Aumenta el score y disminuye el error en

la etapa de evaluación del modelo? Justifique su respuesta. Ver referencia z-score.

Trabajo Práctico ML - Regresión Logística - Clasificación Multiclase - Frutas

Cuestionario Regresión Logística

Considerando Google Colab, como base para el desarrollo del contenido del TP:

Trabajo Práctico ML - Regresión - Clasificación Multiclase - Frutas

y Dataset correspondiente: Dataset Frutas

Se deberán desarrollar las modificaciones necesarias sobre el código inicial y el dataset atravesando los diferentes pasos que componen la **Metodología CRISP DM**, que para este caso nos enfocaremos en: **Selección de Features, Balanceo de Set de Datos, Separación de Set de Datos y Normalización de los datos**. A medida que avanzamos en los pasos debemos utilizar las configuraciones realizadas en el paso actual, es decir, se irán encadenando las configuraciones para finalmente obtener el modelo más óptimo. Ejemplo: Una vez encontradas las features más relevantes para el análisis, debemos usar esas features para quedarnos solo con esas columnas al modificar la separación de datos del paso siguiente; y finalmente, con las features del 2) , el balanceo de datos de 3) y la proporción de datos para el set de datos prueba del 4) utilizarlos para configurar 5) junto con su correspondiente normalización. **Se deberá ir evolucionando el código en Python aplicando cambios por cada punto.**

1. Observaciones Modelo Inicial de Ejemplo: A partir del análisis inicial realizado como demostración para el desarrollo del contenido del trabajo práctico, contestar las siguientes preguntas observando los resultados obtenidos. Justifique su respuesta considerando balance de instancias del dataset para sus clases, tamaño del set de datos y features seleccionados:

a. ¿Por qué los valores predecidos por nuestro modelo para el set de datos de prueba nos retorna como resultados de predicción una única clase? Justifique su respuesta.

b. ¿Por qué la matriz de confusión solo nos devuelve una columna completa?

c. ¿Por qué el score al evaluar nuestro modelo es tan bajo?

2. Selección de Features (correlation matrix & mutual info): Analizamos relación existente entre los diferentes atributos del dataset. Generalmente, al seleccionar features con mayor interrelación entre sí puede indicarnos que son más relevantes para ser analizados por nuestra modelo para llevar a cabo una adecuada predicción, ¿Si aumentamos el número de feature cols a analizar por

nuestro modelo obtendremos una mayor precisión al evaluar el mismo? Justifique su respuesta. Ver referencia

- Correlation matrix
- mutual info
- Mutual info de Scikit Learn

3. Balancear Set de Datos: Para obtener mejores resultados para la predicción adecuada para todas las clases, generalmente se suele balancear el dataset de manera que haya una cantidad de instancias similar para todas las clases. De esta manera nuestro modelo tendrá información suficiente para poder identificar las características específicas de cada clase y no sesgarse con algunas en particular. Agregar o eliminar instancias de manera de trabajar con un dataset balanceado y obtener mejores predicciones. Ver referencia:

- Imbalanced Dataset - Over Sampling - Random Over Sampler
- Ejemplo

4. Separación de Set de Datos (train_test_split - test_size): Al momento de entrenar nuestro modelo, previamente debemos separar el set de datos total en: 1) Set de Datos de Entrenamiento, y 2) Set de Datos de Prueba. Generalmente, al aumentar la cantidad de datos para el set de dato de prueba, ¿Aumenta el score y disminuye el error en la etapa de evaluación del modelo? Justifique su respuesta. Ver referencia train_test_split

5. Normalización (z-score): Para mejorar el procesamiento de las operaciones que lleva a cabo nuestro modelo cuando entrena debemos normalizar los datos con los que vamos a trabajar. Generalmente, al normalizar nuestros set de datos luego de su separación en set de datos de entrenamiento y de prueba, ¿Aumenta el score y disminuye el error en la etapa de evaluación del modelo? Justifique su respuesta. Ver referencia z-score

Entrega

Elaborar un documento detallando los análisis, código, explicaciones y gráficos requeridos para desarrollar los respectivos cuestionarios con el siguiente formato:

- Nombre del archivo de la documentación: **IA_2023_C1_TP_ML_GRUPO_X.pdf**
- Tamaño: A4.
- Encabezado: Extremo izquierdo: **Inteligencia Artificial 2023 C1** y Extremo derecho: **GRUPO X**.
- Pie de página: debe tener el número de página y total de páginas en el extremo derecho de la hoja.
- Nombre zip: **1127_EIA_2022_C3_TP_ML_GRUPO_X.zip**

Se deberá entregar un zip con el documento en formato PDF.

El documento debe ser entregado de manera **GRUPAL (máximo 7 personas)** según el cronograma del curso con la siguiente estructura:

- Carátula con integrantes: Nombre Apellido + DNI.
- Índice.
- Desarrollo respuesta de cada cuestionario, justificando las decisiones tomadas para obtener la configuración del modelo con la mejor precisión posible al evaluar sus resultados de predicción y clasificación. **Agregar Link compartido de referencia al archivo COLAB con el código para cada ejercicio de regresión.**