UdeA

Probabilidad y Estadística
Proyecto 3: Regresión y Clasificación.
2025-1

Parte 1. Regresión (50%)

Para evaluar el beneficio de implementar prácticas estandarizadas para la medición de la temperatura corporal elevada (EBT) con termógrafos infrarrojos (IRTs), se realizó un estudio clínico con más de mil sujetos. Se midieron las temperaturas orales de los sujetos y se capturaron imágenes térmicas faciales con dos IRTs evaluados. A partir de las imágenes térmicas, se extrajeron temperaturas de diferentes ubicaciones en el rostro, y se construyó una base de datos abierta disponible en este enlace. Toda la información en estos archivos ha sido anonimizada.

La información de este estudió será utilizada en este proyecto para realizar diferentes análisis estadísticos.

Se solicita ingresar al enlace para leer la descripción de la base de datos y entender su organización y el significado de las variables medidas.

El estudio se realizó con dos temperaturas ambientales diferentes, y con dos termógrafos, sin embargo, el análisis se hará para un solo termógrafo. El archivo correspondiente se anexa. Adicionalmente, se cuenta con dos archivos PDF donde se aclara la convención del nombre de las variables. Asimismo, se tiene una figura en la que se ilustra las zonas donde se midió la temperatura facial con los termógrafos.

Se solicita realizar los siguiente con la base de datos:

De los proyectos anteriores tenemos:

- **Inspección del archivo y carga en Python**

- **Análisis Exploratorio de Datos de las siguientes variables:**

  - aveOralM
  - Max1R13_Promedio
  - Max1L13_Promedio

- T_Max_Promedio
- TF_HCC_Promedio

Además, en el proyecto 2 se compararon cada una de las variables con el estándar (temperatura oral) encontrándose que las medidas no son equivalentes, lo cual podría indicar que la medición con el termógrafo se encuentra desajustada. Una forma de calibrar la medición con el termógrafo es encontrar un modelo matemático que logre relacionar la temperatura oral con las diferentes mediciones del termógrafo, y de este modo calcular la temperatura corporal real de la persona.

Procedimiento: (50%)

1. Realice un análisis de regresión para encontrar el modelo que mejor describa la temperatura Oral a partir de máximo las 4 temperaturas del termógrafo analizadas antes. Para esto se espera que se deba intentar diferentes modelos, y que utilice las diferentes herramientas de validación dadas en clase para le selección del mejor modelo. De este modo, los procedimientos a seguir son libres, pero deben estar bien justificados, por lo cual se pide que se describa cada resultado obtenido, presentando la interpretación de cada fase, hasta lograr llegar al modelo que considere más satisfactorio. Esto quiere decir que no entregará solo lo último que intente, sino que deberá mostrar qué cosas descartó en el proceso y por qué.
2. Se espera que se incluya, pero no exclusivamente, construcción y ajuste de modelos, curvas de validación y aprendizaje, y regularizaciones.
3. Presente el modelo matemático final, es decir, muestre la ecuación del modelo.

Nota: Recuerde que el mejor modelo no necesariamente es el que tenga las mejores métricas, se busca siempre lograr un modelo lo más simple posible, que haga un trabajo satisfactorio. Decida cual sería el error admisible.

UdeA

Parte 2: Clasificación (50%)

Contexto

Según la Organización Mundial de la Salud (OMS), el accidente cerebrovascular (ACV) es la segunda causa principal de muerte a nivel mundial, siendo responsable de aproximadamente el 11% del total de muertes.
Se adjunta un conjunto de datos que se puede utilizar para predecir si un paciente es propenso a sufrir un ACV, basándose en parámetros de entrada como el género, la edad, diversas enfermedades y el estado de tabaquismo. Cada fila de los datos proporciona información relevante sobre un paciente.

1. id: identificador único
2. gender: "Male" , "Female" u "Other"
3. age: edad del paciente
4. hypertension: 0 si el paciente no tiene hipertensión, 1 si tiene hipertensión
5. heart_disease: 0 si el paciente no tiene enfermedades cardíacas, 1 si tiene alguna enfermedad cardíaca
6. ever_married: "No" o "Yes"
7. work_type: "children" (niños), "Govt_job" (empleo gubernamental), "Never_worked" (nunca ha trabajado), "Private" (sector privado) o "Self-employed" (trabajador independiente)
8. Residence_type: "Rural" o "Urban"
9. avg_glucose_level: nivel promedio de glucosa en sangre
10. bmi: índice de masa corporal
11. smoking_status: "formerly smoked" (fumó anteriormente), "never smoked" (nunca fumó), "smokes" (fuma actualmente) o "Unknown" (desconocido)*
12. stroke: 1 si el paciente tuvo un ACV, 0 si no

Entonces se busca generar un algoritmo de clasificación que permita determinar a partir de la información que se tiene, si el paciente tiene riesgo de padecer un ACV.

Procedimiento

1. Explore la base de datos, si hay variables que tengan datos NaN deberá imputarlos con la media de la clase correspondiente. Para esto puede usar la siguiente sintaxis:

```
df['medicion'] = df.groupby('clase')['medicion'].transform(lambda x: x.fillna(x.mean()))
```

2. Determine la distribución de las clases (stroke = 1 , Stroke =0). Puede usar gráficos de barras.
3. Los algoritmos de clasificación estudiados requieren de entradas numéricas, por lo cual deberá codificar las variables categóricas que son *str*.

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
data['gender_c'] = le.fit_transform(data['gender'])
```

4. Construya un Modelo Naives-Bayes, otro de tipo de Discriminante Lineal de Fisher, y otro de Regresión Logística. Determine cual de todos se desempeña mejor.

5. Con ayuda de las curvas PR (precisión-Recall) determine un valor de umbral apropiado para evitar al máximo posible la clasificación de pacientes que realmente tienen riesgo de tener un ACV, como pacientes sin riesgo. Tenga en cuenta que eso se lograría con un umbral de 0, pero la idea no que todo sea clasificado como 1.

Entregable

Notebook (archivo .ipynb) con el desarrollo del proyecto. El notebook debe adjuntarse corrido, de modo puedan verse todos los resultados. En este mismo Notebook se debe dar respuesta a todas las preguntas y los análisis solicitados, esto en celdas de texto. El Notebook es el equivalente a un informe, entonces deben estar bien organizado.