

UdeA

Probabilidad y Estadística Proyecto 2: Aprendizaje Automático

Trabajo colaborativo, trabajar en equipos estrictamente de 3 personas.

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.

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.

El objetivo del estudio es lograr encontrar una equivalencia entre la medición de temperatura con termografía y la temperatura oral medida con termómetros convencionales. En palabras de nuestro curso, necesitamos encontrar un modelo que relacione la temperatura oral con la temperatura del rostro medida con termografía.

Se solicita realizar los siguiente con la base de datos:

Procedimiento: (50%)

1. Inspección del archivo y carga en Python
2. Seleccione 4 variables de temperatura facial medida con termografía de la base de datos dada. La selección no debe ser aleatoria debe estar justificada.

UdeA

3. Realice un análisis de regresión para encontrar el modelo que mejor describa la temperatura Oral a partir de máximo las 4 temperaturas del termógrafo seleccionadas antes. Para esto se espera que se deba intentar diferentes modelos, y que utilice las diferentes herramientas de validación dadas en clase para la 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.
4. Se espera que se incluya, pero no exclusivamente, construcción y ajuste de modelos, curvas de validación y aprendizaje, y regularizaciones.
5. 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 cuál sería el error admisible.

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

UdeA

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.

```
40
41     data_encoded = pd.get_dummies(data, columns=[ 'gender',...
42
```

UdeA

4. Las técnicas de agrupamiento pueden ser utilizadas como generadores de nuevas características del mismo modo que los algoritmos KNN. Genere dos nuevas características, la primera será las etiquetas entregadas por un algoritmo de agrupamiento k-means, la segunda la probabilidad de ser 1, entregado por un algoritmo de clasificación KNN. Evalúe la utilidad de estas dos características calculando métricas como el recall y la precisión y el F1. Para el caso de K means, asuma que las etiquetas de los grupos generados son lo y_predichos. La utilidad de establece en función de si hay una buena relación entre las etiquetas predicas y reales, es decir si hay buenas métricas el recall y la precisión y el F1, no se espera que sean perfectas, pero que no sean muy malas (menor a 0.5).
5. Usando las características iniciales y las nuevas (si se definió que son útiles), construya un modelo que permita la clasificación de pacientes en riesgo de ACV a partir de las variables usadas. Se espera que pruebe varios modelos de todos los vistos, y que elija uno según su desempeño. Deberá evaluar al menos 1 modelo de aprendizaje automático estadístico, y usar curvas de validación para la optimización de sus hiperparámetros. Nuevamente los procedimientos son libres pero deben estar bien justificados y explicados, debe mostrar los aciertos y desaciertos, no solamente el modelo final, para poder evidenciar el trabajo hecho para la selección del mejor modelo.

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 deben usar celdas de texto para explicar los procedimientos que se hacen, **evite al máximo celdas de código muy largas**, se espera que el **informe sea una narración** de lo que se está haciendo por lo que se espera que use celdas de texto muy frecuentemente para explicar los códigos y sus resultados. El Notebook es el equivalente a un informe, entonces deben estar bien organizado.