Regresión Logística

Fecha de entrega: 19 de abril de 2024

Objetivos:

- Comprender los fundamentos de la regresión logística y su aplicación en Python.
- Explorar y preparar el conjunto de datos para el análisis.
- Implementar un modelo de regresión logística para predecir el riesgo de abandono escolar.
- Evaluar la precisión y eficacia del modelo.

Problema a resolver: Desarrollar un modelo predictivo que pueda estimar la probabilidad de abandono escolar en diferentes planteles de México, basándose en características como la entidad, el plantel y el periodo anual.

Instrucciones:

1. **(+1.3 pts)** Carga y exploración datos
   - **(+0.0 pts)** Carga el **conjunto de datos** en un DataFrame de pandas.
   - **(+0.2 pts)** Observa las primeras filas del DataFrame para tener una idea general de los datos y su estructura.
   - **(+0.4 pts)** Comprueba los tipos de datos de cada columna para asegurarte de que son adecuados para el análisis. Por ejemplo, las variables numéricas deberían tener tipos de datos como int o float.
   - **(+0.4 pts)** Realiza un resumen estadístico de las variables numéricas usando `describe()`. Esto te proporcionará una visión rápida de la distribución de los datos, incluyendo medidas como el promedio, la mediana, el mínimo, el máximo y los cuartiles.
   - **(+0.3 pts)** Utiliza `pandas` para agrupar los datos. Primero por *entidad* y luego por *plantel*, calculando el promedio de abondono escolar por entidad.
2. **(+2.95 pts)** Visualización y análisis de datos
   - **(+0.4 pts)** Utiliza un histograma para visualizar cómo se distribuyen los porcentajes de abandono escolar entre los planteles educativos. Esto puede revelar si hay una tendencia central o si existen outliers significativos.
   - **(+0.4 pts)** Utiliza `matplotlib` o `seaborn` para dibujar un gráfico de barras que muestre el promedio de abandono escolar por entidad.
   - **(+0.4 pts)** Selecciona algunos planteles que representen variabilidad en el abandono escolar y dibuja un gráfico similar al de las entidades.
   - **(+1.75 pts)** Analiza los tres gráficos anteriores y responde: ¿puedes identificar algún patrón o cambio significativo en los datos? Cada uno de los integrantes debe dar su análsis.

3. **(+3 pts)** Creación del modelo de regresión logística
   - **(+1.05 pt)** Utiliza el análisis anterior para seleccionar las variables independientes para el modelo. Considera características como la *entidad, plantel* y el *periodo_ anual*.
   - **(+0.2 pts)** Utiliza `train_test_split` para dividir el conjunto de datos en dos: un conjunto de entrenamiento (80%) y un conjunto de prueba (20%).
   - **(+1 pt)** Utiliza `sklearn.linear_model.LogisticRegression` para crear una instancia del modelo de regresión logística de `scikit-learn` y entrena el modelo. Una vez entrenado el modelo, revisa los coeficientes de las variables para entender su influencia en la probabilidad de alto riesgo de abandono escolar.
   - **(+0.75 pts)** Con el modelo entrenado, usa los métodos `.predict()` y `.predict_proba()` para realizar predicciones sobre el conjunto de prueba. Revisa las predicciones y compáralas con los valores reales para obtener una primera impresión sobre el rendimiento del modelo.
4. **(+2.75 pts)** Evaluación del modelo
   - **(+0.5 pts)** Calcula la matriz de confusión
   - **(+0.5 pts)** Calcula otras métricas de evaluación como la presición, la sensibilidad (recall) y el puntaje F1.
   - **(+1.75 pts)** Usando las evaluaciones, escribe tus conclusiones. Todos los integrantes del equipo deben redactar su conclusión.

Entregables:

- Código fuente del proyecto en un archivo `.ipynb`
- Reporte en `pdf` donde se explique la implementación, la experimentación y el análisis de los resultados. Esto también puede ir en el notebook.

Estos dos archivos deben de estar dentro un de `.zip` con el nombre `equipo`-practica07.

Sobre la entrega:

- La práctica se realizará en equipos de exactamente **5 personas**.
- Ante cualquier duda con la práctica, por favor envía un correo a **taniarubi@ciencias.unam.mx** o manda un mensajito por **telegram** :D
- **En caso de no seguir los lineamientos de entrega, no se calificará la práctica en cuestión.**
- En caso de detectarse copias entre equipos, se evaluará a ambas partes con cero sin realizar indagaciones sobre el asunto.
- En caso de detectarse el uso de IA generativa en la mayor parte de la realización de la práctica, se realizará una entrevista a todo el equipo para determinar la calificación final del trabajo.