Tecnológico de Monterrey

Actividad 4.1: Regresión Logística

Ivanna Maldonado Cervantes

Paula Simonetta Madrid Pérez

Ania Diaz Gonzalez

Miranda Eugenia Colorado Arróniz

Omar Alejandro Quinn

20 de octubre de 2025

Analítica de datos y herramientas de inteligencia artificial II (Gpo 101)

Dr. Alfredo García Suárez

La presente actividad se centra en la aplicación del modelo de regresión logística para explorar las relaciones entre diferentes atributos del conjunto de datos de inventarios de diatomeas (microalgas). El proceso inicia con la importación de la clase LogisticRegression desde sklearn.linear_model, estableciendo el marco metodológico para el análisis de clasificación. La primera etapa crucial consistió en la transformación de variables continuas a variables binarias, un requisito fundamental para la aplicación de regresión logística. Esta conversión se realizó utilizando la mediana como punto de corte, creando una dicotomía que clasifica cada observación como perteneciente a la categoría "alta" (valor 1) o "baja" (valor 0) según si supera o no el valor mediano de la distribución.

Para la variable TaxonName, se calculó la mediana de sus valores numéricos convertidos previamente, estableciendo un umbral que permitió crear la variable binaria TaxonName_bin. Este proceso de binarización asigna el valor 1 a todas las observaciones cuyo TaxonName supera la mediana y 0 a las que se encuentran por debajo. La distribución resultante de valores binarios proporciona información sobre el equilibrio de clases en la variable dependiente, aspecto crucial para evaluar la calidad del modelo de clasificación.

Esta metodología de conversión se aplicó consistentemente a todas las variables categóricas del dataset, incluyendo TaxonCode, SamplingOperations_code, CodeSite_SamplingOperations, y Date_SamplingOperations. Para esta última variable, fue necesario un paso adicional de conversión a formato datetime utilizando pd.to_datetime con el parámetro errors='coerce' para manejar posibles inconsistencias en el formato de fechas.

El análisis de regresión logística se estructuró en cinco modelos independientes, cada uno diseñado para predecir una variable dependiente binaria específica utilizando diferentes conjuntos de variables predictoras. El primer modelo se enfocó en predecir TaxonName_bin utilizando como variables independientes Abundance_nbcell y Abundance_pm. Esta elección de predictores sugiere una hipótesis biológica de que las medidas de abundancia celular pueden ser indicadores efectivos para clasificar los tipos de taxones presentes en las muestras de diatomeas.

El proceso de modelado siguió las mejores prácticas de aprendizaje automático, comenzando con la división del dataset en conjuntos de entrenamiento y prueba mediante train_test_split, asignando el 30% de los datos para validación. La estandarización de las variables predictoras se realizó utilizando StandardScaler, un paso crítico en regresión logística dado que el algoritmo es sensible a las diferencias de escala entre variables. Esta normalización asegura que todas las variables contribuyan equitativamente al proceso de aprendizaje del modelo.

La implementación del algoritmo LogisticRegression se realizó con los parámetros predeterminados de sklearn, seguida del entrenamiento mediante el método fit y la generación de predicciones en el conjunto de prueba. La evaluación del rendimiento del modelo se basó en múltiples métricas complementarias, proporcionando una visión integral de la capacidad predictiva del clasificador.

La matriz de confusión constituye la base fundamental para evaluar el rendimiento del clasificador, mostrando la distribución de predicciones correctas e incorrectas para cada clase. Esta matriz permite identificar patrones específicos de error, como

la tendencia del modelo a clasificar incorrectamente una clase particular. La precisión, calculada como la proporción de predicciones positivas correctas sobre el total de predicciones positivas, indica la confiabilidad del modelo cuando predice la clase positiva. Esta métrica se calculó tanto para la clase 1 como para la clase 0, proporcionando una comprensión balanceada del rendimiento del clasificador.

La exactitud o accuracy representa la proporción total de predicciones correctas y ofrece una medida general del rendimiento del modelo. Sin embargo, en casos de desequilibrio de clases, esta métrica puede ser engañosa, por lo que se complementa con la sensibilidad o recall, que mide la capacidad del modelo para identificar correctamente los casos positivos reales. La sensibilidad es particularmente importante en contextos donde los falsos negativos tienen consecuencias significativas.

El puntaje F1 combina precisión y sensibilidad en una métrica única mediante su media armónica, proporcionando un balance entre ambas medidas. Esta métrica es especialmente valiosa cuando se busca un compromiso entre minimizar tanto los falsos positivos como los falsos negativos. El reporte de clasificación final consolidó todas estas métricas en un formato comprensivo, incluyendo soporte estadístico para cada clase.

El segundo modelo de regresión logística se diseñó para predecir TaxonCode_bin utilizando las mismas variables predictoras que el primer modelo: Abundance_nbcell y Abundance_pm. Esta consistencia en la selección de predictores permite comparaciones directas entre modelos y evalúa si diferentes aspectos de la clasificación taxonómica responden de manera similar a las medidas de abundancia. El proceso de entrenamiento, evaluación y cálculo de métricas siguió la misma metodología rigurosa establecida en el primer modelo.

El tercer modelo introdujo una complejidad adicional al predecir SamplingOperations_code_bin utilizando variables binarias previamente creadas: CodeSite_SamplingOperations_bin y Date_SamplingOperation_bin. Esta aproximación explora las relaciones entre diferentes aspectos del diseño experimental y temporal del muestreo, investigando si la combinación de información espacial y temporal puede predecir efectivamente las operaciones de muestreo. La utilización de variables binarias como predictores representa una estrategia metodológica que puede capturar patrones categoriales subyacentes en los datos.

El cuarto modelo se enfocó en predecir CodeSite_SamplingOperations_bin empleando una combinación diversa de predictores: SamplingOperations_code_bin, Abundance_pm, y TaxonName_bin. Esta selección heterogénea de variables combina información categórica binaria con medidas continuas de abundancia, explorando si la integración de diferentes tipos de información mejora la capacidad predictiva del modelo. La inclusión de TaxonName_bin como predictor sugiere una hipótesis sobre la asociación entre tipos taxonómicos específicos y sitios de muestreo particulares.

El quinto y último modelo se diseñó para predecir Date_SamplingOperation_bin utilizando TotalAbundance_SamplingOperation y Abundance_pm como predictores. Esta configuración investiga si las medidas de abundancia total y específica pueden

indicar patrones temporales en el muestreo, explorando posibles variaciones estacionales o temporales en la composición de las comunidades de diatomeas.

Cada modelo siguió consistentemente el mismo protocolo de evaluación, calculando matrices de confusión, precisión para ambas clases, exactitud, sensibilidad para ambas clases, puntajes F1 para ambas clases, y reportes de clasificación completos. Esta metodología estandarizada permite comparaciones sistemáticas entre modelos y facilita la identificación de patrones de rendimiento consistentes o variables específicas con mayor poder predictivo.

La implementación de múltiples modelos de regresión logística en este contexto biológico proporciona insights valiosos sobre las relaciones entre diferentes variables en el ecosistema de diatomeas. Los modelos exploran desde relaciones básicas entre abundancia y clasificación taxonómica hasta patrones más complejos que involucran dimensiones espaciales y temporales del muestreo. La evaluación comprehensiva mediante múltiples métricas asegura una comprensión robusta del rendimiento de cada modelo y facilita la selección de enfoques más efectivos para futuros análisis.

Finalmente, la aplicación de regresión logística en este contexto demuestra la versatilidad de las técnicas de aprendizaje automático en la investigación ecológica, proporcionando herramientas cuantitativas para explorar patrones complejos en datos de biodiversidad. Los resultados de estos modelos pueden informar decisiones sobre estrategias de muestreo, identificación de factores predictivos clave, y comprensión de la estructura de las comunidades de diatomeas en diferentes contextos ambientales.