I302 - Aprendizaje Automático y Aprendizaje Profundo

Trabajo Práctico 2: Clasificación y Ensemble Learning

Juan Francisco Lebrero

1 de abril de 2025

Ingeniería en Inteligencia Artificial

Resumen

En este trabajo se realizó un análisis exploratorio y se implementaron diversos modelos de clasificación para diagnosticar cáncer de mama a partir de variables morfológicas y bioquímicas de células. Se identificaron características atípicas en los datos, tales como outliers y valores negativos en variables que, biológicamente, deberían ser positivas. El conjunto de datos presenta un balance adecuado en la variable objetivo (45 % vs 55 %). Asimismo, se observó que ninguna variable individual mostró una correlación fuerte con el diagnóstico, lo que sugiere la necesidad de utilizar modelos que capturen relaciones no lineales e interacciones complejas entre las variables.

1. Introducción

El diagnóstico temprano del cáncer de mama es crucial para mejorar el pronóstico de los pacientes. Este estudio analiza un conjunto de datos obtenido a partir de imágenes histopatológicas de biopsias mamarias, en el cual se han extraído diversas características morfológicas y moleculares de las células. El objetivo es desarrollar un modelo de clasificación que permita predecir si una célula presenta características compatibles con un diagnóstico benigno (0) o maligno (1). Para ello, se dispone de 12 variables numéricas (por ejemplo, tamaño celular, densidad nuclear, tasa de mitosis) y 2 variables categóricas (tipo celular y presencia de mutaciones genéticas).

2. Métodos

En esta sección se describen los algoritmos y métodos empleados para el tratamiento de datos y el modelado predictivo. Se incluyen tanto técnicas de preprocesamiento (detección y tratamiento de outliers, imputación de valores faltantes) como algoritmos de clasificación.

2.1. Detección y Tratamiento de Outliers

El tratamiento de valores atípicos es un paso fundamental en el preprocesamiento, ya que los outliers pueden afectar negativamente el rendimiento de los modelos predictivos. Para la identificación de outliers se empleó el método del Rango Intercuartílico (IQR). En este método, para cada variable se realizan los siguientes pasos:

1. Se calculan el primer cuartil (Q_1) y el tercer cuartil (Q_3).
2. Se determina el Rango Intercuartílico:
$$IQR = Q_3 - Q_1.$$
3. Se establecen los umbrales para la detección de outliers:
$$L_{\text{inf}} = Q_1 - 1{,}5 \times IQR \quad \text{y} \quad L_{\text{sup}} = Q_3 + 1{,}5 \times IQR.$$
4. Se identifican como outliers aquellos valores que se encuentran fuera del intervalo $[L_{\text{inf}}, L_{\text{sup}}]$.

Para el tratamiento de estos outliers se consideraron tres estrategias:

- **Winsorización**: Se reemplazan los valores por debajo de L_{inf} por L_{inf} y los valores por encima de L_{sup} por L_{sup}.
- **Imputación por estadísticos centrales**: Se sustituyen los outliers por la mediana (o la media) de la variable, aprovechando la robustez de la mediana frente a valores extremos.
- **Reemplazo por valores válidos cercanos**: Se sustituyen los outliers por el valor válido más cercano dentro del rango definido.

La elección del método se realizó en función de la distribución de cada variable y su significado biológico.

2.2. Imputación de Valores Faltantes mediante K-Nearest Neighbors (K-NN)

El algoritmo K-Nearest Neighbors (K-NN) es un método no paramétrico que se utiliza para la imputación de valores faltantes, preservando la estructura multivariada de los datos. Para cada observación con valores faltantes, se identifican sus k vecinos más cercanos (en nuestro estudio, $k = 5$), utilizando la siguiente fórmula para la distancia euclidiana:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

Posteriormente, se imputan los valores faltantes utilizando, para las variables numéricas, la media o la mediana de los valores de los vecinos, y para las variables categóricas, el valor modal.

2.3. Modelos de Clasificación

Se implementaron los siguientes algoritmos de clasificación:

2.3.1. Regresión Logística

La regresión logística es un modelo estadístico que utiliza la función logística para predecir la probabilidad de que una observación pertenezca a la clase 1 (maligno). La función de predicción es:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}},$$

donde $\beta_0, \beta_1, \ldots, \beta_p$ son los coeficientes estimados.

2.3.2. Random Forest

El modelo Random Forest es un ensamble de árboles de decisión que utiliza la técnica de bagging para reducir la varianza y mejorar la generalización. Cada árbol se entrena sobre una muestra aleatoria del conjunto de datos y la predicción final se obtiene mediante votación mayoritaria.

2.3.3. Gradient Boosting

El algoritmo de Gradient Boosting construye modelos de forma secuencial, en los que cada nuevo modelo corrige los errores del anterior. La función de pérdida se minimiza mediante el gradiente descendente:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \gamma_m h_m(\mathbf{x}),$$

donde $h_m(\mathbf{x})$ es el nuevo árbol y γ_m el coeficiente de aprendizaje.

3. Desarrollo

En esta sección se presentan los resultados del análisis exploratorio de datos y el desarrollo de los modelos predictivos. Se hace referencia a los métodos descritos en la sección de *Métodos*.

3.1. Análisis Exploratorio de Datos

El análisis exploratorio permitió identificar características relevantes del conjunto de datos. Se analizaron las 12 variables numéricas mediante histogramas y diagramas de caja, tal como se explica en la sección de *Detección y Tratamiento de Outliers*. Se detectaron rangos atípicos, como valores negativos en variables que biológicamente deberían ser positivas

Variable	**Rango Observado**	**Observaciones**
CellSize	-100.0 a 1000.0	Valores negativos atípicos
CellShape	-1.41 a 5.05	Presencia de valores negativos
NucleusDensity	-3.0 a 50.0	Valores negativos atípicos
ChromatinTexture	10.0 a 263.79	Sin valores negativos
CytoplasmSize	5.0 a 299.35	Todos los valores positivos
CellAdhesion	-4.46 a 5.03	Índice que admite negativos
MitosisRate	-50.0 a 100.0	Valores negativos atípicos
NuclearMembrane	-3.44 a 24.87	Valores negativos presentes
GrowthFactor	-50.0 a 563.62	Valores negativos atípicos
OxygenSaturation	-10.0 a 807.18	Saturación negativa atípica
Vascularization	-0.997 a 54.51	Valores mayormente positivos
InflammationMarkers	0.0 a 440.90	Todos los valores positivos

Cuadro 1: Rangos observados en variables numéricas.

Adicionalmente, se identificó un porcentaje significativo de valores faltantes, los cuales fueron imputados mediante el algoritmo K-NN (véase la sección de *Imputación de Valores Faltantes mediante K-Nearest Neighbors*).

3.2. Variables Categóricas y Variable Objetivo

El análisis de las variables categóricas reveló que:

- **CellType**: Con tres categorías, donde una categoría sin etiqueta sugiere la necesidad de una revisión en la fuente de datos.

- **GeneticMutation**: Presenta dos categorías (Present y Absent) con distribución equilibrada.

La variable objetivo **Diagnosis** mostró una distribución equilibrada (aproximadamente 55 % negativos y 45 % positivos), lo que favorece la aplicación de los modelos de clasificación.

3.3. Correlaciones y Relaciones Multivariadas

El análisis de correlación evidenció que ninguna variable numérica mostró una correlación fuerte con la variable *Diagnosis* (coeficientes en el rango |0,05| a |0,10|). Se observaron correlaciones moderadas entre variables como *CellSize* y *CytoplasmSize* ($r \approx 0{,}63$) y entre *ChromatinTexture* y *CytoplasmSize* ($r \approx 0{,}46$). Estos hallazgos sugieren que el diagnóstico depende de interacciones complejas, lo cual justifica el uso de modelos que capturen relaciones no lineales (ver sección de *Métodos*).

3.4. Implicaciones para el Modelado

Los hallazgos del análisis exploratorio indican la necesidad de:

- Aplicar estrategias robustas de preprocesamiento, tales como el tratamiento de outliers y la imputación de valores faltantes mediante K-NN.
- Emplear modelos capaces de capturar relaciones no lineales, como los algoritmos de Gradient Boosting y Random Forest, complementados por la regresión logística para establecer una línea base.

4. Resultados

Se presentan a continuación los resultados más destacados obtenidos tras la implementación de los modelos. Los gráficos y tablas correspondientes incluyen leyendas y etiquetas adecuadas para facilitar su interpretación. En general, se observó que:

- La eliminación y tratamiento de outliers mediante winsorización e imputación mejoró la robustez de los modelos.
- El modelo de Random Forest mostró un desempeño superior en términos de precisión, mientras que la regresión logística presentó limitaciones al capturar relaciones no lineales.

(Aquí se incluirían gráficos de desempeño, curvas ROC y tablas comparativas, con sus correspondientes leyendas y etiquetas.)

5. Conclusiones

El presente estudio demuestra la importancia de un preprocesamiento riguroso en el diagnóstico de cáncer de mama. La detección y tratamiento de outliers, así como la imputación de valores faltantes mediante K-NN, han permitido obtener un conjunto de datos más

homogéneo y representativo. La implementación de diversos modelos de clasificación ha revelado que métodos ensamble, como Random Forest y Gradient Boosting, son los más adecuados para capturar las complejas interacciones entre variables, lo que se traduce en un mejor desempeño predictivo.