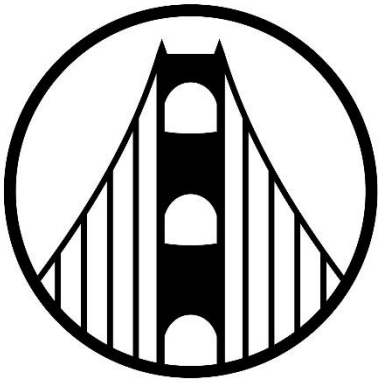
Representación y clasificación según varios modelos de Machine Learning sobre una base de datos



THE BRIDGE

David Torosyan Petrosyan

Objetivos generales

El objetivo general de nuestro análisis consiste en la construcción de modelos de aprendizaje automático sobre la base de datos para predecir la clasificación correcta del cancer de pulmón basándonos en los datos proporcionados.

Se propone además, la comparación entre los diversos modelos ajustados, para extraer conclusiones sobre cuál es más recomendable,proporcionando mayor tasa de aciertos, en esta base de datos.

Si bien los modelos que se van a aplicar ya se han aplicado por otros investigadores,el plantearlo en este trabajo viene justificado como una primera aproximación del estudiante al aprendizaje automático con Python, lenguaje en el que no se ha trabajado en estadística a lo largo del grado en Estadística Empresarial.

Supone pues un reto en sí mismo por el dominio del lenguaje y de técnicas estadísticas que tampoco se han trabajado, al menos desde la perspectiva del aprendizaje automático, en asignatura alguna del grado.

Análisis de los datos proporcionados

El dataset proporcionado por la empresa, contiene información detallada sobre pacientes diagnosticados con cáncer de pulmón, con un total de 890,000 registros. Este conjunto de datos está diseñado para ser utilizado en un modelo de machine learning, con el objetivo de predecir la supervivencia (survived) de los pacientes tras un tratamiento. A continuación, se presenta una documentación exhaustiva del problema basado en los datos disponibles:

Descripción General del Dataset

Tamaño: 890,000 filas y 17 columnas.

Objetivo Principal: Predecir la variable survived (0 = no sobrevivió, 1 = sobrevivió), que indica si el paciente sobrevivió después del tratamiento.

Período Temporal: Los datos abarcan diagnósticos desde el 2 de junio de 2014 hasta el 30 de mayo de 2024, con fechas de finalización del tratamiento extendidas hasta el 30 de mayo de 2026, lo que sugiere un seguimiento a largo plazo.

Ubicación Geográfica: Incluye pacientes de múltiples países europeos (por ejemplo, Suecia, Países Bajos, Hungría, Bélgica, Luxemburgo, etc.), lo que permite un análisis transnacional.

- **Variables del Dataset**

El dataset incluye las siguientes columnas, con sus características y posibles implicaciones:

age (Edad):

Tipo: Float64.

Rango: 4 a 104 años.

Media: 55.01 años.

Desviación estándar: 9.99.

Implicación: La edad es un factor clave en la prognosis del cáncer, con una distribución que abarca desde niños hasta ancianos, lo que sugiere una diversidad significativa en la población estudiada.

gender (Género):

Tipo: Int64 (0 = Masculino, 1 = Femenino tras transformación).

Distribución: 445,134 hombres (50.01%) y 444,866 mujeres (49.99%), casi equilibrada.

Implicación: El género puede influir en la incidencia y respuesta al tratamiento del cáncer de pulmón, con estudios que sugieren diferencias en tasas de supervivencia entre hombres y mujeres.

country (País):

Tipo: Object.

Valores únicos: Múltiples países europeos (ej. Suecia, Países Bajos, Hungría, etc.).

Implicación: Factores como el acceso a la atención sanitaria, estilos de vida y políticas de salud pública pueden variar por país, afectando los resultados.

diagnosis_date (Fecha de diagnóstico):

Tipo: Datetime64[ns].

Rango: 2014-06-02 a 2024-05-30.

Media: 2019-06-02.

Implicación: Permite analizar tendencias temporales en los diagnósticos y su relación con los avances médicos o cambios en la detección.

cancer_stage (Estadio del cáncer):

Tipo: Int64 (1 a 4, probablemente mapeado de "Stage I" a "Stage IV").

Media: 2.50.

Implicación: El estadio del cáncer es un predictor crítico de supervivencia, con estadios avanzados (III y IV) asociándose a peores pronósticos.

family_history (Historial familiar):

Tipo: Int64 (0 = No, 1 = Sí tras transformación).

Media: 0.50 (50% con historial familiar).

Implicación: Un historial familiar positivo puede indicar predisposición genética, influyendo en el riesgo y la gravedad.

smoking_status (Estado de fumador):

Tipo: Int64 (posiblemente mapeado como 0 = No fumador, 1 = Fumador pasivo, 2 = Exfumador, 3 = Fumador actual).

Media: 1.50.

Implicación: El tabaquismo es un factor de riesgo bien conocido para el cáncer de pulmón, y esta variable puede ser un predictor fuerte.

bmi (Índice de masa corporal):

Tipo: Float64.

Rango: 16 a 45.

Media: 30.49.

Implicación: El BMI puede estar relacionado con la salud general y la capacidad de tolerar tratamientos agresivos como la quimioterapia.

cholesterol_level (Nivel de colesterol):

Tipo: Int64.

Rango: 150 a 300.

Media: 233.63.

Implicación: Niveles elevados de colesterol podrían estar asociados con comorbilidades que afectan la supervivencia.

hypertension (Hipertensión), asthma (Asma), cirrhosis (Cirrosis)

other_cancer (Otro cáncer):

Tipo: Int64 (0 = No, 1 = Sí).

Medias: 0.75, 0.47, 0.23, 0.09 respectivamente.

Implicación: Estas comorbilidades pueden complicar el tratamiento y reducir las probabilidades de supervivencia.

treatment_type (Tipo de tratamiento):

Tipo: Int64 (0 = Quimioterapia, 1 = Cirugía, 2 = Combinado, 3 = Radiación tras transformación).

Distribución: ~223,262 (25% cada tipo aproximadamente).

Implicación: La elección del tratamiento depende del estadio del cáncer y la salud general, influyendo en los resultados.

end_treatment_date (Fecha de fin del tratamiento):

Tipo: Datetime64[ns].

Rango: 2014-12-02 a 2026-05-30.

Media: 2020-09-02.

Implicación: La duración del tratamiento (calculada como treatment_duration) puede reflejar la complejidad del caso.

survived (Supervivencia):

Tipo: Int64 (0 = No, 1 = Sí).

Media: 0.22 (22% de supervivencia).

Implicación: Esta es la variable objetivo, indicando un desequilibrio significativo (78% no sobrevivieron), lo que sugiere la necesidad de técnicas como sobremuestreo.

treatment_duration (Duración del tratamiento):

Tipo: Int64. N

Rango: 183 a 730 días.

Media: 458 días.

Implicación: La duración del tratamiento puede estar correlacionada con la gravedad o la respuesta al mismo.

Preprocesado de los datos

El preprocesado de datos engloba todas las tareas relacionadas con el tratamiento de la base de datos en bruto, con la finalidad de convertirla en una base de datos eficiente y fácil de utilizar para su análisis estadístico con modelos de aprendizaje automático, y en particular en nuestro caso, con modelos de Machine Learning.

En nuestro caso, dado que la base de datos ha sido trabajada de modo extensivo y depurada previamente, el preprocesado de datos consistirá en:

- identificar la existencia de valores faltantes, para en su caso, imputar valores o eliminar registros; en nuestro caso los valores faltantes ya habían sido tratados, y la base de datos no contenía valores faltantes;

- estandarización de las variables numéricas (básicamente todas las registradas salvo la respuesta), con el fin de trasladarlas todas a una escala común y así poder abordar de modo apropiado los modelos de aprendizaje;

- creación de variables dummy para las variables categóricas, esto es, para la variable respuesta que identifica el tipo de tumor; - división de datos en muestras de entrenamiento (sobre los que ajustar el modelo) y test (para verificar la calidad del ajuste).

Respecto a la identificación de valores faltantes, utilizamos el comando datos.isnull().sum(). El resultado, como se comentó anteriormente, es nulo: no se encuentra ningún valor faltante. De hecho, la base de datos contiene todos los valores para los 890000 registros de las 17 variables disponibles.

El proceso de estandarización consiste en transformar los datos de tipo numérico, para centrarlos en su media (eliminando así el valor medio de cada característica) y escalarlos dividiendo por su desviación estándar. Este proceso es necesario para aplicar de un modo eficiente las técnicas de clasificación automática, dado que las variables se han medido inicialmente en escalas dispares y no comparables. Esta estandarización se resuelve básicamente con la función StandardScaler() de la librería Scikit-Learn.

Para tratar con técnicas de clasificación automática la variable respuesta que identifica el tipo de tumor (maligno/benigno), hemos de crear una variable dummy numérica, a la que asignamos el valor 1 para los tumores malignos y 0 para los benignos. Este proceso se resuelve con la función OneHotEncoder() de la librería Scikit-Learn.

Una vez transformadas las variables con las que resolver el análisis, las almacenamos en una nueva base de datos que contiene la información en el formato necesario para abordar los análisis posteriores, y a la que accedemos ya directamente para llevarlos a cabo (disponible en datos-estandarizados.csv).

Para ajustar un modelo de Machine Learning basado en clasificación, hemos de dividir la base de datos en muestras de entrenamiento, con la que ajustamos el modelo, y de test, con la que testamos la calidad del mismo. Consideramos una partición aleatoria de los datos a razón de una proporción 80%-20%, respectivamente, para las muestras de entrenamiento y test. Esto se resuelve con la función train_test_split() de la librería Scikit-Learn.

Análisis del Problema

Objetivo: Desarrollar un modelo predictivo para determinar la probabilidad de supervivencia de un paciente con cáncer de pulmón basado en sus características clínicas, demográficas y de tratamiento.

Una vez se ha entrenado un modelo, podemos obtener predicciones sobre unos datos. En el caso de un problema binario, nuestro modelo clasificaría como 0 o 1 un conjunto de datos, y a partir de esto podemos ya crear la matriz de confusión.

	1	0
1	TP	FP
0	FN	TN

Lo que nos han pedido es qué porcentaje de valores que se han clasificado como positivos son realmente positivos.

Para ello utilizaremos la métrica de precisión, la cual se calcula como Precision = TP/(TP+FP)

Desafíos:

- Desequilibrio de clases: Solo el 22% de los pacientes sobrevivieron, lo que puede sesgar los modelos hacia predecir "no sobrevivió" a menos que se apliquen técnicas como sobremuestreo (SMOTE, como se observa en el notebook).
- Variables categóricas: Columnas como country y smoking_status requieren codificación (OneHotEncoder) para su uso en modelos.
- Datos temporales: Las fechas (diagnosis_date, end_treatment_date) pueden ser útiles para calcular duraciones o tendencias, pero deben procesarse adecuadamente.
- Comorbilidades: La presencia de múltiples condiciones (hipertensión, asma, etc.) añade complejidad al modelo, requiriendo un análisis de interacciones.

Contexto Clínico

El cáncer de pulmón es una de las principales causas de muerte por cáncer a nivel mundial, con factores de riesgo como el tabaquismo, la genética y las comorbilidades desempeñando roles cruciales. Este dataset permite explorar cómo las intervenciones médicas y las características del paciente afectan los resultados, lo que podría apoyar decisiones clínicas personalizadas.

Documentación de lo Realizado en el Notebook

El notebook pruebas.ipynb documenta un flujo de trabajo para construir y optimizar un modelo de machine learning para predecir la supervivencia de pacientes con cáncer de pulmón. A continuación, se detalla paso a paso lo realizado:

1. Importación de Librerías

Se importaron bibliotecas esenciales para análisis de datos y machine learning:

pandas y numpy para manipulación de datos.

matplotlib y seaborn para visualización.

sklearn para métricas (accuracy, confusion matrix), división de datos (train_test_split), validación cruzada (cross_val_score), búsqueda de hiperparámetros (GridSearchCV), y modelos como LogisticRegression, RandomForestClassifier, XGBClassifier, LGBMClassifier.

StandardScaler para normalización de características.

Se añadieron módulos personalizados (toolbox_ML, bootcampviztools) desde un directorio específico.

2. Carga y Exploración del Dataset

Se cargó el archivo Lung_Cancer.csv como un DataFrame de pandas, eliminando la columna id como índice.

Se exploraron los primeros registros con df.head() y se obtuvo un resumen estadístico con df.describe() y df.info():

Se confirmaron 890,000 registros y 17 columnas.

Identificamos tipos de datos (float64, int64, object, datetime64[ns]) y la ausencia de valores nulos.

Se convirtieron las columnas diagnosis_date y end_treatment_date a formato datetime para facilitar cálculos temporales.

3. Preprocesamiento de Datos

Codificación de variables categóricas:

gender se mapeó de "Male"/"Female" a 0/1 usando una función personalizada cambiar_indices_gender.

treatment_type se mapeó de texto ("Chemotherapy", "Surgery", "Combined", "Radiation") a valores numéricos (0, 1, 2, 3) con un diccionario tratamiento_map.

4. Análisis Exploratorio

Usamos métodos como value_counts() para analizar la distribución de gender (casi equilibrada) y treatment_type (aproximadamente 25% por categoría).

5. Preparación del Modelo

Se dividieron los datos en conjuntos de entrenamiento y prueba (implícito en X_train_resampled, y_train_resampled, X_test_resampled, y_test), con sobremuestreo aplicado (SMOTE) para manejar el desequilibrio de clases.

Se seleccionaron características importantes (posibles_features_importantes y posibles_features_importantes_2).

6. Entrenamiento y Optimización del Modelo

Se compararon diferentes modelos y con el que mejor resultados se obtuvo se optimizaron sus hiperparametros con un GridSearchCV.

Pudimos observar que el sobremuestreo mejoró las predicciones pero igualmente obtuvimos que la precisión varía entre 0.22 y 0.23, lo que indica que de las predicciones positivas para "sobrevive", solo el 22-23% son correctas. Esto es bajo, sugiriendo muchos falsos positivos. Las predicciones negativas para “sobrevive” son buenas teniendo un 77-78%.

Haría falta contactar con la empresa y que nos dieran más datos para poder llegar a hacer un análisis más exhaustivo y poder hacer un modelo más eficiente.Con los datos que tenemos no es suficiente para hacer un modelo que clasifique si una persona tiene cancer de pulmón.