TRABAJO FIN DE MÁSTER

Máster en Ciencia de Datos e Ingeniería de Datos en la Nube

WiDS Dathaton 2024 - Challenge 2: Modelos de regresión para estimación del periodo de diagnóstico metastático

Autor: Luna Jiménez Fernández

Tutor: Juan Carlos Alfaro Jiménez

Junio, 2025

Dedicado a la gente que, pese a todo,
sigue persiguiendo sus sueños.
Nunca os rindáis.

Declaración de autoría

Yo, **Luna Jiménez Fernández**, con DNI **47092045M**, declaro que soy la única autora del Trabajo Fin de Master titulado ***"WiDS Dathaton 2024 - Challenge 2: Modelos de regresión para estimación del periodo de diagnóstico metastático"***, que el citado trabajo no infringe las leyes en vigor sobre propiedad intelectual, y que todo el material no original contenido en dicho trabajo está apropiadamente atribuido a sus legítimos autores.

Albacete, a … de **Junio** de **2025**

Fdo.: **Luna Jiménez Fernández**

Resumen

TODO RESUMEN AQUI

Abstract

TODO ABSTRACT HERE

Agradecimientos

En primer lugar, quiero agradecer a todos mis compañeros y amigos del grupo de **Sistemas Informáticos y Minería de Datos (SIMD)** - y, especialmente, a mi amigo y director **Juan Carlos Alfaro Jiménez** - por su apoyo, recursos y consejos durante la realización de este trabajo. Aunque ya no sea formalmente parte de este grupo, siempre me sentiré vinculada a él.

Además, quiero a gradecer a mis amigos y familia del **Curso de Comic Online de la Escola Joso** - **Arai, Aina, Arkaitz, Clara, Irene, Martín, Pau, Rafi...** -, con los que compartí un proyecto de gran importancia personal, mi primer comic publicado, y en los que he encontrado un grupo al que pertenecer. Muchas gracias por todo.

Finalmente, quiero agradecer a **mi familia y seres queridos** - tanto los que me acompañan presencialmente como los que se encuentran a distancia. Vuestro apoyo y cariño continuo me ha ayudado a seguir adelante y acabar este trabajo a pesar de todas las dificultades.

Índice general

Índice de figuras

Índice de tablas

1. Introducción

El acceso equitativo a una **atención sanitaria de calidad** es un problema de gran interés a nivel global, existiendo desigualdades sustanciales en la **calidad y acceso** a dicho servicio entre distintas poblaciones. Estos problemas, además, se pueden llegar a exacerbar por distintos factores: geográficos, socioeconómicos y climáticos.

Con el fin de estudiar la influencia de dichos factores en la atención sanitaria, la iniciativa *Women in Data Science* propuso en el año 2024 una **competición** [1] con el objetivo de **estimar el tiempo necesario para realizar un diagnóstico de metástasis para cáncer de mama** a partir de un conjunto de datos médico ampliado con información geográfica, socioeconómica y climática - y, a su vez, estudiar como dichos factores pueden influir al tiempo necesario para realizar un diagnóstico.

Por tanto, la meta de este trabajo es la creación de **modelos de regresión** capaces de estimar dicho tiempo de diagnóstico con el menor error posible - utilizando, para ello, el proceso completo de **ciencia de datos**.

1.1. Objetivos

El principal objetivo de este trabajo es el **desarrollo de un modelo de regresión** capaz de resolver el problema propuesto por la competición: la predicción del tiempo necesario para realizar un diagnóstico de metástasis de cáncer de mama, evaluando su rendimiento y dejando disponible el modelo para ser accesible por los hipotéticos usuarios finales.

Para alcanzar dicho objetivo, es necesario llevar a cabo los siguientes pasos, siguiendo el **ciclo de vida de la ciencia de datos**:

1. Análisis exploratorio de los datos disponibles en la competición, para comprender su comportamiento y características.
2. Pre-procesamiento de los datos para la propuesta de subconjuntos de atributos reducidos y preparación posterior para el uso con modelos.
3. Estudio, selección y caracterización de los modelos y sus hiperparámetros a estudiar durante el proceso.

4. Experimentación y estudio de los resultados para seleccionar un modelo definitivo a ser utilizado.
5. Creación de una aplicación web para desplegar el modelo final entrenado, con el fin de ser utilizado por expertos en el campo de la medicina sin experiencia previa en ciencia de datos.

A su vez, este trabajo aborda el segundo objetivo planteado por la propia competición: el **estudio de la influencia de los factores geográficos, socioeconómicos y climáticos** en la calidad de la atención sanitaria.

1.2. Estructura de la memoria

La memoria está dividida en un total de **7** capítulos, como se describen a continuación:

- **Capítulo 1:** En este capítulo se introduce el problema a resolver, los objetivos que se busca cumplir con el trabajo y la estructura general de la memoria.
- **Capítulo 2:** En este capítulo se realiza una breve revisión de las principales técnicas a utilizar durante la memoria: tanto el proceso de ciencia de datos y sus etapas como los modelos a utilizar durante la experimentación - desde los modelos simples como las regresiones lineales y los árboles de decisiones hasta los *ensembles* de modelos simples.
- **Capítulo 3:** En este capítulo se realiza un estudio más exhaustivo del problema: tanto su definición como un análisis exploratorio de los datos disponibles, estudiando el comportamiento de la variable objetivo y la relevancia y correlación de los atributos respecto al tiempo de diagnóstico.
- **Capítulo 4:** En este capítulo se introduce el pre-procesamiento a realizar sobre el conjunto de datos, obteniendo varios subconjuntos de atributos reducidos a ser estudiado posteriormente y preparando *pipelines* automáticos para realizar todas las transformaciones necesarias para el uso de los datos por parte de los modelos.
- **Capítulo 5:** En este capítulo se detalla la experimentación a realizar. Se proponen varios modelos sobre los que se realizará un proceso de ajuste de hiperparámetros y selección de modelos, con el fin de obtener un modelo definitivo a ser utilizado para resolver el problema. Además, se presentan y estudian los resultados de dicha experimentación.
- **Capítulo 6:** En este capítulo se presenta una aplicación web a través de la cual se hace disponible a los usuarios expertos el modelo obtenido en el capítulo anterior - detallando la interfaz gráfica y las distintas funcionalidades ofrecidas.
- **Capítulo 7:** Finalmente, en este capítulo se muestran las conclusiones alcanzadas tras el desarrollo del trabajo, proponiendo posibles lineas de trabajo futuro para ampliarlo.

2. Revisión de técnicas

En este capítulo se describen los procesos y algoritmos utilizados a lo largo del trabajo descrito en esta memoria. Concretamente, se comienza explicando el concepto de la **ciencia de datos** y su ciclo de vida, haciendo énfasis en **CRISP-DM** como metodología utilizada a lo largo del proyecto para resolver el problema propuesto. Tras esto, se estudian conceptos de aprendizaje automático como la **selección de modelos** o los **modelos de regresión** - haciendo especial énfasis en los modelos de *ensemble* basados en técnicas de ***Gradient Boosting***

2.1. Ciencia de datos y el ciclo de vida de los datos

2.1.1. Ciencia de datos

La **ciencia de datos** es el estudio de la extracción de conocimiento útil a partir de datos, y de la generalización de dicho proceso a cualquier problema [2]. Dicho proceso incluye la recolección y almacenamiento, mantenimiento, procesamiento, análisis y visualización de enormes cantidades de datos heterogéneos - asociados a un gran abanico de aplicaciones y dominios en muchas ocasiones multidisciplinarios [3].

Desde su origen, la ciencia de datos ha evolucionado como un campo interdisciplinar que integra conocimientos y técnicas de otras disciplinas afines como el análisis de datos, la estadística o la minería de datos [4]. Ahora bien, la principal diferencia con estos campos se encuentra en el fin: el aprendizaje a partir de los datos [2] y la capacidad de adquirir nuevo conocimiento capaz de ser utilizado para la toma de decisiones y la predicción [3].

Por definición, la ciencia de datos depende de los datos sobre los que se está trabajando. Por esto, el proceso de trabajo de la ciencia de datos depende generalmente del **ciclo de vida de los datos**: las distintas etapas por las que pasa un conjunto de datos desde su recolección e investigación hasta su uso final [5]. Como se observa en la **Figura 2.1**, este ciclo está tradicionalmente dividido en **cinco** apartados [4]:

1. **Adquisición:** En la actualidad, los datos se generan en cantidades masivas - del orden de **exabytes por hora** [6]. Por tanto, el primer paso del ciclo consiste en la adquisición y almacenamiento eficiente de los datos necesarios para el proceso.

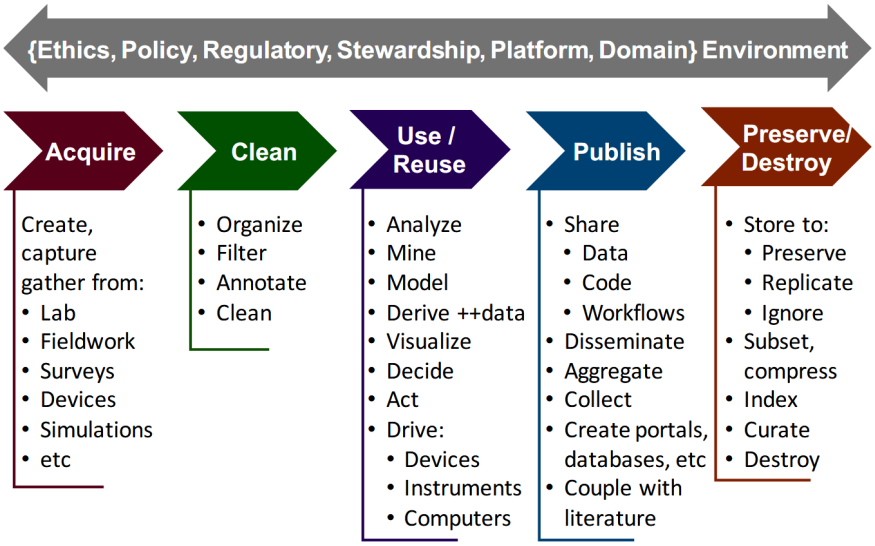


Figura 2.1: Ciclo de vida de los datos [4]

2. **Limpieza:** Tras la adquisición, el segundo paso del ciclo consiste en la transformación de los datos originales en datos utilizables posteriormente - a través de procesos de limpieza, imputación, formateo...
3. **Uso y re-uso:** El tercer paso del ciclo consiste en el uso de los datos procesados con el fin de adquirir conocimiento y tomar decisiones a partir de éstos. Éste apartado se puede dividir, a su vez, en tres subapartados [6]:

   (a) **Análisis exploratorio:** El estudio del comportamiento de los datos con el fin de plantear hipótesis para guiar el resto del ciclo de datos [7].

   (b) **Modelado:** El uso de técnicas computacionales y estadísticas para extraer conocimiento y predicciones a partir del conjunto de datos.

   (c) **Visualización, interpretación y actuación:** La representación gráfica de los resultados del uso de los datos, con el fin de facilitar la toma de decisiones posterior a las personas.

4. **Publicación:** El cuarto paso del ciclo consiste en la diseminación de los resultados del proceso - con el fin de que el conocimiento creado pueda ser conocido y reutilizado por el mayor número de personas posible.
5. **Preservación o destrucción:** El quinto y último paso del ciclo consiste en la preservación o destrucción de los datos utilizados - cumpliendo con otros factores como pueden ser las consideraciones éticas o regulatorias.

Con el fin de regularizar, estandarizar y hacer reproducible el proceso completo de la ciencia de datos - desde la adquisición de los conjuntos de datos hasta la distribución de los resultados -, se han propuesto varias ampliaciones y adaptaciones del ciclo de datos estudiado, conocidas como **ciclos de vida de la ciencia de datos** [5].

Aunque actualmente no existe un ciclo estandarizado, uno de los procesos más utilizados para ciencia de datos es el **Cross-Industry Standard Process for Data Mining (*CRISP-DM*)**, propuesto originalmente para el campo de la minería de datos pero adaptado a las necesidades de la ciencia de datos [8] - siendo el proceso utilizado a lo largo del trabajo descrito en esta memoria.

2.1.2. Cross-Industry Standard Process for Data Mining - CRISP-DM

Cross-Industry Standard Process for Data Mining (abreviado como *CRISP-DM*) es una metodología desarrollada con el fin de ofrecer un proceso de trabajo completo de principio a fin para la minería de datos; independientemente del campo, las herramientas o la aplicación final de los datos [8]. Si bien fue propuesto originalmente en el año 2000, en la actualidad sigue siendo uno de los procesos más utilizados tanto en minería de datos como en ciencia de datos [9].

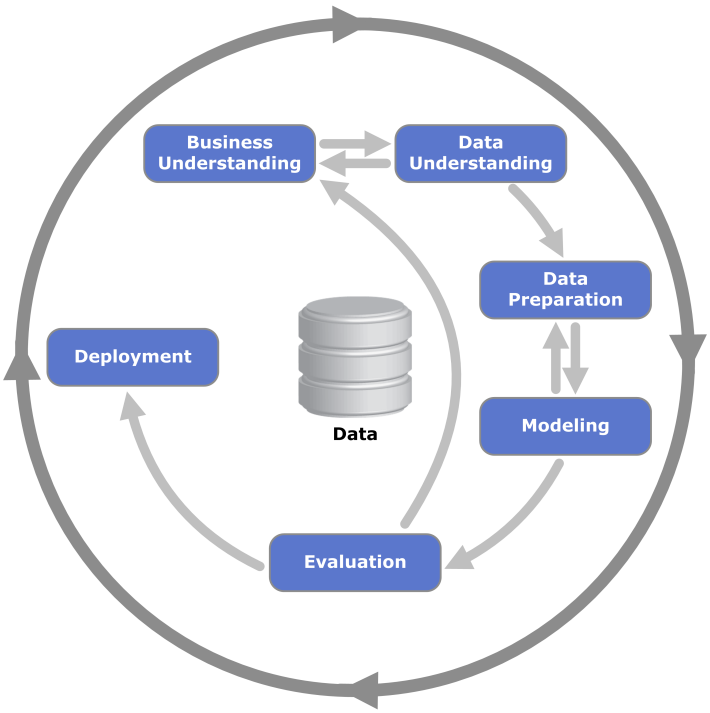


Figura 2.2: Ciclo de vida de CRISP-DM [8]

Como se observa en la **Figura 2.2**, el ciclo de CRISP-DM está dividido en **seis fases** [8], similares al ciclo de datos estudiado:

1. **Conocimiento del campo (*Business Understanding*):** El primer paso del ciclo consiste en entender el problema y los objetivos a resolver - estudiando la situación actual y estableciendo los pasos para alcanzar las metas propuestas.

2. **Conocimiento de los datos (*Data Understanding*):** El segundo paso del ciclo consiste en adquirir y estudiar los datos - tanto de forma superficial como en un análisis exploratorio más profundo -, además de verificar que los datos disponibles son útiles para los objetivos propuestos.
3. **Preparación de los datos (*Data Preparation*):** Tras la adquisición de conocimiento, el tercer paso consiste en preparar los datos obtenidos para su uso posterior - seleccionando las instancias relevantes, limpiando los datos para eliminar valores perdidos, enriqueciendo los datos con información externa...
4. **Modelado (*Modeling*):** Con los datos preparados, la cuarta fase del ciclo consiste en el uso y calibración de modelos de aprendizaje automático a aplicar sobre los datos - definiendo los estudios y experimentos a realizar sobre los modelos, y evaluando el rendimiento final de éstos.
5. **Evaluación (*Evaluation*):** Antes de desplegar el modelo final, la quinta fase del ciclo consiste en evaluar si los resultados obtenidos satisfacen los objetivos propuestos y si el proceso de ciencia de datos se ha aplicado de forma adecuada.
6. **Despliegue (*Deployment*):** La última fase del ciclo es el despliegue y diseminación de los resultados obtenidos - haciendo disponible el modelo y los resultados a los usuarios finales.

Es importante destacar que, como indican las flechas de la **Figura 2.2**, la metodología propuesta no es linear, sino que el flujo entre los distintos pasos se puede ver alterado:

- Las fases tienen dependencias entre sí - los descubrimientos en algunas fases pueden producir que sea necesario volver a fases anteriores para perfeccionar el proceso.
- El proceso es **cíclico** - los conocimientos adquiridos durante las distintas fases se aplican para refinar futuros procesos, ya sean sobre el mismo conjunto de datos o datos nuevos.

2.2. Aprendizaje automático y ajuste de modelos

2.2.1. Aprendizaje automático

El **aprendizaje automático** (también conocido en inglés como *Machine Learning*) es una rama de la inteligencia artificial que consiste en la creación de programas capaces de **aprender** - es decir, de mejorar su rendimiento en una tarea - a través de la experiencia y de la información que se les aporta [10].

También se puede definir el término como el conjunto de métodos capaces de detectar patrones en los datos de forma autónoma, y de utilizar dichos patrones para predecir datos futuros [11] - siendo este último de mayor interés a los principios de la ciencia de datos.

Generalmente, los algoritmos de aprendizaje automático se dividen en dos grandes familias, en función del tipo de datos e información que se aporta a los algoritmos [12] [11]:

- **Aprendizaje supervisado:** El objetivo del algoritmo es aprender una función capaz de, dados unos datos de entrada X, predecir una salida Y. Esta función se aprende a partir de un conjunto de datos $D = (x_i, y_i)_{i=1}^{N}$ donde a cada instancia x_i del conjunto de datos D se le asocia un valor esperado y_i.

  Este tipo de aprendizaje se puede dividir a su vez en dos categorías dependiendo del tipo de salida Y que se espera [12]:

  - **Clasificación**: El algoritmo busca obtener para cada entrada x_i un valor concreto **dentro de un conjunto finito de posibles valores**.
  - **Regresión**: El algoritmo busca obtener, para cada entrada x_i, un **valor numérico continuo**.

- **Aprendizaje no supervisado:** El objetivo del algoritmo es aprender patrones subyacentes de los datos de entrada X ofrecidos, sin buscar predecir una salida. Esta función se aprende a partir de un conjunto de datos $D = x_{i}{}_{i=1}^{N}$ donde no se ofrece ningún tipo de etiqueta a cada instancia x_i.

Un **modelo** es el resultado del proceso de aprendizaje automático: una función capaz de predecir una salida para una entrada dada, y cuyos parámetros e hiperparámetros han sido ajustados a través de un entrenamiento sobre un conjunto de datos para **minimizar un error** [13].

De cara a cumplir el objetivo propuesto por el trabajo descrito en esta memoria - la creación de un modelo capaz de **predecir el tiempo de diagnóstico** -, se van a trabajar con modelos supervisados de **regresión**. Por esto, resulta de interés describir los principales modelos a utilizar y el ajuste que se va a realizar sobre ellos.

2.2.2. Selección de modelos

Durante el entrenamiento de los modelos, se pueden encontrar algunos problemas:

- Se han entrenado varios modelos, y es necesario elegir de forma objetiva uno de ellos en base a su rendimiento.
- Se necesitan elegir los hiperparámetros de un modelo que mejor rendimiento ofrecen sobre el conjunto de datos, de forma objetiva.
- Los modelos entrenados han aprendido variaciones insignificantes y patrones falsos - ruido interpretado como información real - a partir de los datos de entrada, llevando a un problema de **sobreajuste** [11] que puede afectar de forma negativa al rendimiento del modelo.

En estos casos, nos interesa seleccionar de entre todos los modelos a evaluar el modelo más **generalizable** - es decir, el modelo que tendría el **menor error esperado** si se evaluase con un conjunto de datos distinto al utilizado durante el entrenamiento [11].

Ahora bien, a la hora de la verdad no es común tener acceso a dicho hipotético conjunto de test - o si se tiene, solo debería ser utilizado para evaluar el rendimiento del modelo seleccionado finalmente para evitar que se sesgue la selección de modelos [12].

Para solucionar este problema y realizar una selección de modelos **honesta**, se pueden utilizar las siguientes opciones [11]:

- **Conjunto de validación:** Se particiona el conjunto de datos inicial en entrenamiento y validación - utilizando el primero para entrenar los modelos, y el segundo para evaluar su rendimiento honesto y seleccionar el modelo.
- **K-Validación cruzada:** En caso de que el conjunto de datos no sea suficientemente grande para particionarse, se puede optar por particionar el conjunto de datos en **K trozos** de igual tamaño. Para cada uno de estas particiones, se entrenan los modelos con el resto de particiones y se evalúan los rendimientos sobre la partición seleccionada. Tras realizar este proceso K veces, se puede utilizar el error promedio de los K entrenamientos como una aproximación al rendimiento honesto del modelo para realizar la selección.

2.2.3. Modelos de regresión

En el caso de la **regresión**, el objetivo del modelo es aprender una función capaz de predecir un valor numérico continuo para cada instancia de datos de entrada [11]. Dicha función se ajusta buscando encontrar el conjunto de parámetros que **minimiza** la diferencia entre los valores predichos por la función y los valores reales asociados a los datos de entrada [13].

Se han propuesto y estudiado un gran número de modelos de regresión, con parametrizaciones y funcionamientos diversos, en la bibliografía [14]. Pese a esta variedad, es posible dividir todas estas familias de modelos en dos grandes grupos: modelos **tradicionales** y modelos de **conjuntos o *ensembles*** [12]

Modelos tradicionales - regresión lineal, árboles de decisiones y máquinas de vectores de soporte

Si bien no hay una definición consensuada sobre su definición, se puede entender como **modelo tradicional** a un modelo que entrena una única función con el fin de realizar predicciones sobre la salida esperada para cada entrada de datos [12].

Existe una gran cantidad de familias de modelos con una larga trayectoria en la bibliografía existente [13]. Ahora bien, el estudio realizado en la memoria se centra en las siguientes tres familias de modelos utilizadas en el trabajo:

- **Regresión lineal:**

  Los modelos más simples, trabajando con la suposición de que **existe una correlación lineal** entre los atributos de entrada y la salida del modelo [11]. Por lo general, la salida y para una entrada x se predice utilizando la siguiente fórmula:

$$y(x) = \sum_{j=1}^{D} w_j x_j$$

Donde x_j representa cada atributo de la entrada y w_j el peso asignado a cada atributo, siendo el objetivo de estos modelos ajustar los pesos asignados a cada atributo para minimizar el error cuadrado [12]. Ahora bien, cuando se trabaja con conjuntos de datos de gran dimensionalidad, el gran número de atributos puede afectar de forma negativa al rendimiento del modelo, causando un sobreajuste al conjunto de entrenamiento [15].

Para evitar este problema, se proponen técnicas de **regularización** - penalizaciones aplicadas a la fórmula del error con el objetivo de conseguir modelos menos complejos y más generalizables [16]. Los tres modelos de regularización más utilizados son los siguientes:

- **Ridge (L2) [17]:** Como factor de penalización, se utiliza $\sum_{j=1}^{D}(w_j)^2$ - la suma de los pesos cuadrados del modelo, buscando reducir de forma generalizada la influencia de los atributos para evitar sobreajustes y correlaciones.
- **Lasso (L1) [18]:** Como factor de penalización, se utiliza $\sum_{j=1}^{D}|w_j|$ - la suma del valor absoluto de los pesos del modelo, buscando eliminar los atributos irrelevantes reduciendo su peso a 0.
- **Elastic-Net [16]:** Como factor de penalización, se utiliza $\lambda\left(\sum_{j=1}^{D}(w_j)^2\right) + (1-\lambda)\left(\sum_{j=1}^{D}|w_j|\right)$ - utilizando a la vez las regularizaciones L1 y L2 de forma ponderada, buscando aunar los beneficios de ambas aproximaciones.

- **Máquinas de vectores de soporte (SVM):**

  Las máquinas de vectores de soporte se pueden entender como una evolución de los modelos de regresión lineal donde, en vez de buscar la linea que mejor se ajusta al conjunto de datos, se busca el **hiperplano** capaz de ajustarse al conjunto de datos con el **mayor margen** [12].

  En regresión, esto se traduce en la búsqueda de la función representando al mejor hiperplano capaz de ajustarse a todas las instancias del conjunto de datos a la vez que es capaz de mantener una distancia inferior a un margen ϵ con todos los puntos [19].

  La principal utilidad de estos modelos radica en las dos siguientes características [12]:

  - **Funciones kernel:** Un problema de los modelos lineales es que los conjuntos de datos no siempre son linealmente separables. Para solventar este problema, las máquinas de vectores de soporte son capaces de utilizar **funciones *kernel*** para transformar los datos a una mayor dimensionalidad - donde si es posible ajustar un hiperplano con mayor margen.
  - **Vectores de soporte:** Para definir el modelo no es necesario almacenar información sobre el conjunto de datos completo, sino que es suficiente con almacenar información sobre los **puntos que definen la frontera entre el hiperplano y el margen** - conocidos como los vectores de soporte.

- **Árboles de decisión:**

  Los árboles de decisión son modelos de reglas representando su función a través de grafos dirigidos acíclicos [13] donde la predicción se obtiene realizando una serie de comprobaciones secuenciales empezando desde la raíz, ramificando hasta llegar a una hoja final [12]. Estos árboles se dividen en los siguientes componentes:

  - **Nodos:** Nodos internos del árbol donde se realiza una comprobación sobre el valor de un atributo. Dependiendo del resultado de la comprobación, el nodo se **ramifica** a otros nodos u hojas.
  - **Hojas:** El valor final predicho para una entrada, alcanzado tras una serie de comprobaciones en nodos.

  El objetivo del modelo es, por tanto, aprender el conjunto de reglas que minimiza el error del modelo para el conjunto de datos dado. Ahora bien, estos modelos tienden a **sobreajustar** creando árboles de gran profundidad [12].

Modelos de ensemble - bagging y boosting

Como contraste a los modelos tradicionales, un **modelo de conjunto, meta-modelo o *ensemble*** es un modelo que, durante su entrenamiento, ha aprendido un **conjunto de funciones o modelos sencillos** - por lo general, una agrupación de modelos tradicionales -, agrupando las predicciones de todos éstos para obtener una predicción general de la salida esperada para cada entrada de datos [12].

Los algoritmos de *ensemble* buscan aprender un gran número de modelos simples con rendimiento ligeramente mejor que un modelo aleatorio - conocidos como **modelos de aprendizaje débil**, generalmente **árboles de decisión** [13]. Suponiendo que cada uno de estos modelos es completamente independiente al resto, la unión de sus resultados lleva a una predicción final **más precisa y generalizable** que un único modelo entrenado [13].

Dependiendo de la metodología utilizada para entrenar los modelos simples - ya sea de forma secuencial o paralela -, se pueden dividir los algoritmos en dos familias [12]:

- **Bootstrap Aggregating (Bagging):**

  Los modelos de *ensemble* trabajan con la suposición de que cada uno de los modelos individuales que lo componen es totalmente independiente al resto de modelos. Ahora bien, en la práctica no suele ser factible entrenar a cada modelo individual sobre un conjunto de datos independiente [20].

  Para solventar este problema, los modelos de ***bagging*** entrenan cada modelo sobre una **muestra uniformemente aleatoria con reemplazo** (*bootstrap* en inglés) del conjunto de datos original [20] - obteniendo como resultados modelos sencillos e independientes, y siendo la predicción final el **promedio** de las predicciones de cada modelo.

  Algunos de los modelos de *bagging* más importantes son los siguientes:

  - ***Random Forest* [21]:** Un modelo de *ensemble bagging* de **árboles de decisión profundos** en el que se añade un segundo proceso de muestreo aleatorio para aumentar la independencia entre los predictores simples entrenados:

  - **Muestreo de instancias:** Cada árbol se entrena con un subconjunto aleatorio de instancias del conjunto de datos.
  - **Muestreo de atributos:** Cada árbol se entrena con un subconjunto aleatorio de atributos del conjunto de datos.

- ***Extremely Randomized Trees [22]:*** Una evolución del modelo de *Random Forest* en el que se añade un proceso de muestreo adicional, con los siguientes cambios:
  - Cada árbol pasa a entrenarse sobre el **conjunto de datos completo**, sin muestreo - aunque se sigue realizando un muestreo de los atributos a considerar por cada árbol.
  - Durante la construcción del árbol, se generan de forma aleatoria varias **particiones del conjunto de datos** para cada atributo - en vez de calcular la partición óptima. A la hora de construir cada nodo, se elige la partición que mejor puntuación obtiene de todas las generadas.

- **Boosting:**

  Para garantizar la independencia entre los modelos entrenados, los modelos de *ensemble* de tipo **boosting** optan por entrenar sus modelos de forma **secuencial** sobre un **conjunto de datos con pesos** - donde, para cada modelo, se da más peso a las instancias del conjunto de datos que se han predicho erróneamente en los modelos anteriores [12].

  Durante el trabajo, se ha considerado el siguiente modelo de *boosting*:

  - ***Adaptive Boosting [23]:*** Un modelo básico de *boosting* que sigue estrictamente el proceso descrito. Concretamente, se comienza con un conjunto de datos de pesos uniformes sobre el que se entrena el primer modelo, ajustando los pesos de las instancias - aumentando el peso de las instancias con mayor error, y reduciendo el peso de las instancias con menor error. Tras esto, se repite el proceso de entrenamiento de modelos y ajuste de pesos hasta entrenar todos los predictores [12].

    La predicción final es una **media ponderada por el error** de la predicción de todos los modelos - donde los modelos con menor error tienen un mayor peso en la ponderación.

  Una subfamilia dentro de estos algoritmos son los modelos de **Gradient Boosting**. Estos modelos también utilizan la metodología de *boosting* - entrenar modelos secuencialmente ajustándose a los errores del modelo anterior -, con la diferencia de que el entrenamiento no se hace sobre el conjunto de datos directamente, sino sobre los **errores residuales** (la diferencia entre la predicción y el valor real) de cada instancia [24].

  Este comportamiento es similar al **gradiente descendiente** utilizado para entrenar otros modelos como la regresión lineal o las redes neuronales [13]. En concreto, se llevan a cabo los siguientes pasos:

  1. Se comienza realizando una predicción inicial, generalmente el valor promedio de todas las instancias.

2. Utilizando el valor estimado, se calculan los **errores pseudo-residuales** de cada instancia - el **error** entre la predicción y el valor real. Estas pseudo-residuales dependen de la función de error que se elija.
3. Repitiéndose para cada modelo que se tiene que entrenar:
   (a) A partir de los valores pseudo-residuales, se entrena un modelo simple que **predice el valor residual de cada instancia**.
   (b) Utilizando los nuevos residuales estimados, se recalculan los valores pseudo-residuales para que el siguiente modelo ajuste mejor a las instancias clasificadas erroneamente.

Para obtener una predicción final, se suma al valor promedio inicial los residuales calculados por cada uno de los modelos - ponderados por un **factor de aprendizaje** para evitar el sobreajuste [24].

Actualmente, los modelos de *Gradient Boosting* son considerados el estado del arte para la mayoría de problemas de predicción estructurada [13], siendo algunos de los modelos más populares los siguientes:

- ***Extreme Gradient Boosting [25]:*** Un algoritmo de *Gradient Boosting* utilizando el método de Newton-Raphson - en vez de calcular los errores pseudo-residuales, se calcula una función de la segunda y la primera derivada de la función de error. Además, el modelo está diseñado para permitir el entrenamiento en paralelo de los árboles.
- ***Categorical Boosting [26]:*** Un algoritmo de *Gradient Boosting* diseñado para trabajar de forma nativa con atributos categóricos sin necesidad de convertirlos previamente a valores numéricos.

  El modelo se entrena utilizando ***Ordered Boosting*** - utilizando una permutación aleatoria del conjunto de entrenamiento para cada modelo, donde para calcular las pseudo-residuales de cada instancia se consideran solo las instancias anteriores en la permutación [27] - para evitar introducir sesgos.
- ***Light Gradient-Boosting Model [28]:*** Un algoritmo de *Gradient Boosting* con las siguientes características:
  - **Histogramas:** Para optimizar el rendimiento, los valores de los atributos continuos se agrupan en histogramas.
  - **Crecimiento del arbol por hojas:** Frente a otros algoritmos que entrenan los árboles nivel a nivel, el modelo ramifica siempre por la hoja que minimizaría el error - llevando a árboles más ajustados.

  Existe otra implementación de este algoritmo, conocida como **Histogram-Based Gradient Boosting**, ofrecida por la librería de ciencia de datos *Scikit-Learn* [29] - aunque ambas se basan en el mismo modelo y no presentan diferencias significativas.

3. Estudio exploratorio del problema

En este capítulo se estudia en detalle el problema a resolver a través del proceso de ciencia de datos. Se comienza realizando una definición del problema y sus objetivos, seguido por un **análisis exploratorio de los datos** que lo definen. En este análisis se estudian los atributos que describen los datos junto a sus distribuciones y comportamientos - haciendo hincapié en los atributos de carácter geográfico, social y económico.

3.1. Definición y objetivos del problema

El **cáncer de mama triple negativo** es uno de los cánceres de mama más agresivos y difíciles de tratar. En el caso de que además se agravase con una **metástasis**, se necesita un tratamiento rápido y urgente, sin retrasos innecesarios, para aumentar al máximo las posibilidades de éxito. Ahora bien, el tiempo de espera para acceder a dicho tratamiento no es igual para todos los pacientes, y existe la posibilidad de que hayan sesgos influyendo en el tiempo de diagnóstico - como pueden ser algunos factores geográficos, socioeconómicos o incluso climáticos [1].

El principal objetivo del problema - y, por tanto, el objetivo que guía el proceso de ciencia de datos - es **crear un modelo de regresión** capaz de predecir el tiempo de diagnóstico de metástasis en base a la información dada de un paciente. Además, se busca estudiar la **influencia de factores geográficos, socioeconómicos y climáticos** en dicho tiempo de diagnóstico, para comprobar si existe un sesgo real en el trato a los pacientes. El problema a resolver se planteó originalmente como el segundo de los desafíos ofrecidos por la institución **Women in Data Science** como parte de su *Datathon* de 2024 [1].

3.2. Análisis exploratorio de datos

El primer paso en el proceso de ciencia de datos es realizar un estudio exhaustivo del conjunto de datos con el fin de comprender mejor su comportamiento y la distribución de sus datos.

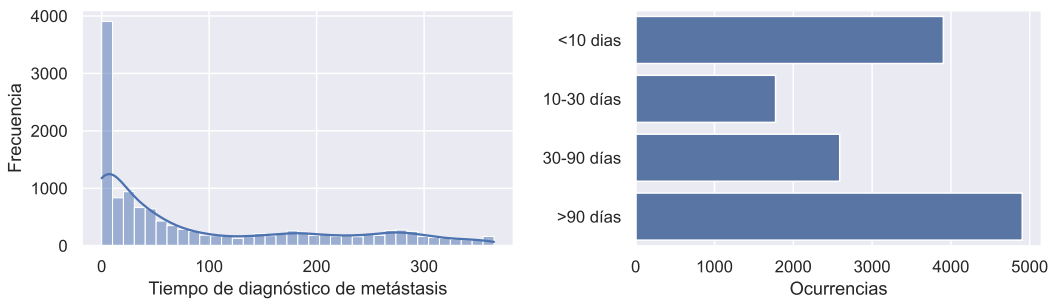
3.2.1. Distribución del conjunto de datos

El conjunto de datos contiene un total de **13173 instancias**, cada una de ellas descrita por **150 atributos** - divididos en **11 atributos categóricos** y **139 atributos numéricos** - y una **variable objetivo numérica**. Describir individualmente todos los atributos en la memoria no sería factible, por lo que se describen los principales grupos de atributos:

- **Atributos médicos (13 atributos - 11 categóricos y 2 numéricos)**: Datos identificativos e información sobre el diagnóstico, tratamiento y seguro del paciente.
- **Atributos socioeconómicos (65 atributos - 2 categóricos y 63 numéricos)**: Por lo general, datos **estadísticos** reflejando información socioeconómica relacionada con la población del paciente. Estos estadísticos se pueden dividir, a su vez, en:
  - **Porcentajes (49 atributos numéricos):** Porcentajes en el rango $[0, 100]$ representando valores estadísticos - matrimonios, educación, demografía... - de la ubicación del paciente.
  - **Medianas (10 atributos numéricos):** Valores representando la mediana de algunos estadísticos - edad, ingresos, alquileres... - de la ubicación del paciente.
  - **Información geográfica (6 atributos - 2 categóricos y 4 numéricos):** Valores concretos - población, densidad... - de la ubicación del paciente, sin ser representados a través de un porcentaje o una mediana.
- **Atributos climáticos (72 atributos numéricos)**: Temperatura promedio (en grados **Farenheit**) de la población del paciente - representada de forma mensual entre los años 2013 y 2018.

Variable objetivo - tiempo de diagnóstico

La variable objetivo - el **tiempo de diagnóstico de una metástasis** - es una variable **numérica entera** con valores en el conjunto de entrenamiento en el rango $[0 - 365]$, cuya distribución se puede observar en la **Figura 3.1**.



(a) Distribución general **(b)** Distribución agrupada

Figura 3.1: Distribución del tiempo de diagnóstico

Se puede ver que los tiempos de diagnóstico siguen una **distribución de Poisson** - con la mayoría de casos diagnosticados en un rango de $[0 - 10]$ días. Ahora bien, como se observa en la **Figura 3.1b**, si se agrupan los valores en rangos **la mayoría de casos tardan más de 90 días en ser diagnosticados**.

Valores perdidos

Antes de realizar un estudio más exhaustivo de los atributos, es de interés estudiar el comportamiento de los **valores perdidos** en el conjunto de datos - para comprobar si hay un gran número de éstos, si existen atributos irrelevantes por tener un alto grado de información perdida y si sería necesario realizar algún tipo de tratamiento sobre éstos valores.

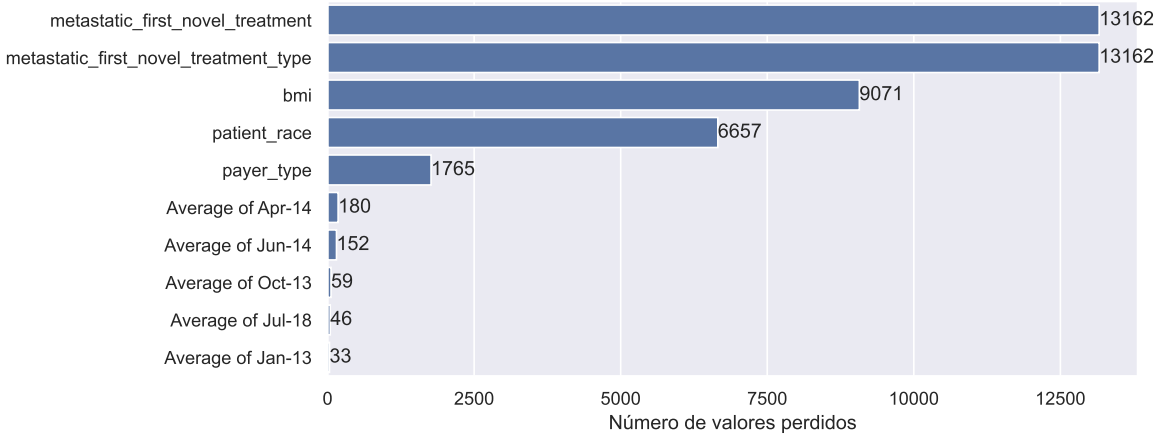


Figura 3.2: Distribución de valores perdidos en el conjunto de datos

Estudiando la distribución, **72** de los 150 atributos disponibles presentan valores perdidos - con un promedio de **624 instancias perdidas** por atributo. En primera instancia puede parecer un número muy elevado de valores perdidos, si se observa cómo se distribuyen los valores perdidos - como se representa en la **Figura 3.2** - se puede observar un **sesgo** claro, donde la amplia mayoría de valores perdidos se agrupan alrededor de cinco atributos:

- **Tratamiento:** Debido al número tan elevado de valores perdidos en ambos atributos, **solo se tiene información sobre el tratamiento de 11 pacientes** - lo que significa que no sería relevante el atributo debido a la falta de información.
- **Índice de masa corporal del paciente:** Se conoce el índice de masa corporal de **menos de la mitad de los pacientes.** Además, al ser información numérica **no existe un valor por defecto** por el que se puedan reemplazar los valores perdidos - por lo que sería razonable no estudiar en más detalle el atributo.
- **Raza y tipo de seguro médico del paciente:** En ambos casos hay un número considerable de instancias con valores desconocidos. Ahora bien y a diferencia del IMC, al ser atributos categóricos puede considerar que **es significativo para el estudio que no se conozcan estos valores** - tratándolos como una categoría adicional, *"Desconocido"*.

En el resto de atributos el número de valores perdidos es más reducido - en el orden de **100 instancias** o menor -, por lo que el tratamiento es más simple, pudiendo descartar las instancias con valores perdidos o realizando una imputación simple con el valor promedio.

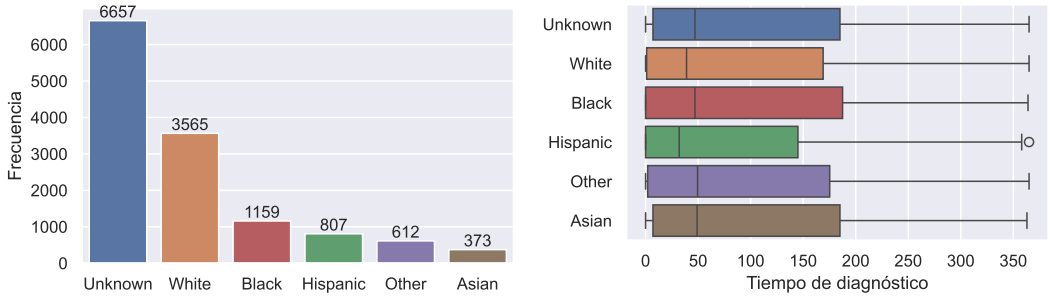
3.2.2. Estudio de atributos categóricos

Tras un primer análisis - en el que se ha estudiado la distribución de la variable objetivo, los atributos y sus valores perdidos -, la segunda parte del análisis exploratorio es realizar un **estudio exhaustivo individualizado** de los atributos de interés y su relación con la variable predictora. Al ser reducido el número de atributos categóricos (con un total de **11**) es posible realizar un análisis individual de cada uno de estos atributos:

Datos personales - raza, género y tipo de seguro del paciente

- **Raza del paciente:**

  Como se ha comentado durante el estudio de los valores perdidos, hay **un número significativo de valores perdidos** de este atributo - que serán tratados como una categoría adicional, *"Unknown"*.



(a) General **(b)** Respecto al tiempo de diagnóstico

Figura 3.3: Distribución de la raza del paciente

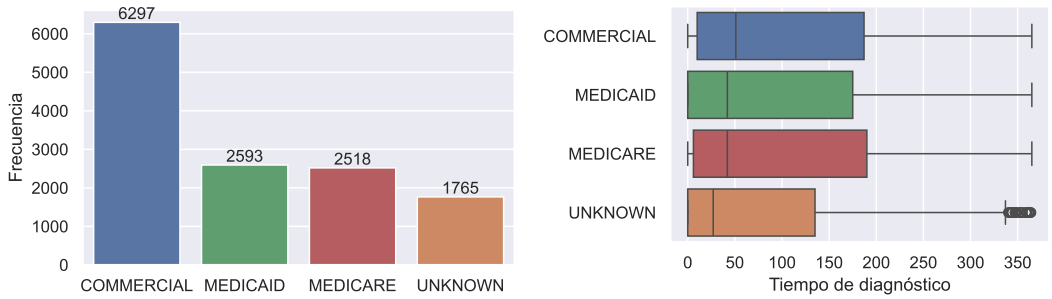
En la **Figura 3.3** se puede observar que:

- **Distribución general:** Como se podía esperar, la mayoría de pacientes presentan una raza desconocida - algo que se puede interpretar como que **la mayoría de los pacientes no se sienten cómodos especificando su raza**. Tras esto, la raza **blanca** es la más frecuente - siendo tres veces más frecuente que la raza negra -, siendo la raza asiática la menos frecuente.
- **Relación con el tiempo de diagnóstico:** Si bien todas las razas tienen un rango de tiempos de diagnóstico amplio, **las razas blanca e hispánica tienen una mediana ligeramente inferior al resto** - sugiriendo que **la raza puede influir en el tiempo de diagnóstico**.

Dada estas observaciones, puede resultar de interés considerar la raza del paciente a la hora de hacer una selección de atributos.

- **Tipo de seguro médico del paciente:**

  Igual que con la raza, hay una cantidad significativa de valores perdidos de este atributo - que serán categorizados como un nuevo valor, *"UNKNOWN"*.



(a) General **(b)** Respecto al tiempo de diagnóstico

Figura 3.4: Distribución del tipo de seguro médico del paciente

En la **Figura 3.4** se puede estudiar que:

- **Distribución general:** El seguro más frecuente - siendo la mitad del conjunto de datos - es el **seguro comercial privado**. Los dos seguros públicos - **Medicaid** y **Medicare Advanced** - tienen proporciones similares entre sí, siendo en conjunto algo inferior al número de seguros privados. Finalmente, hay una cantidad ligeramente inferior de seguros desconocidos - que podría referirse a **pacientes sin seguro médico**.
- **Relación con el tiempo de diagnóstico:** En contra de lo que se podría esperar, los seguros desconocidos presentan **un tiempo de diagnóstico mediano y un rango sustancialmente inferior** al del resto de seguros. Aunque los tres tipos de seguros restantes tienen distribuciones similares, parece que **los seguros privados tienen un tiempo de diagnóstico ligeramente superior**.

Dadas estas diferencias, es posible que **el tipo de seguro del paciente influya en el tiempo de diagnóstico** - por lo que será considerado posteriormente a la hora de realizar una selección de atributos.

- **Género del paciente:**

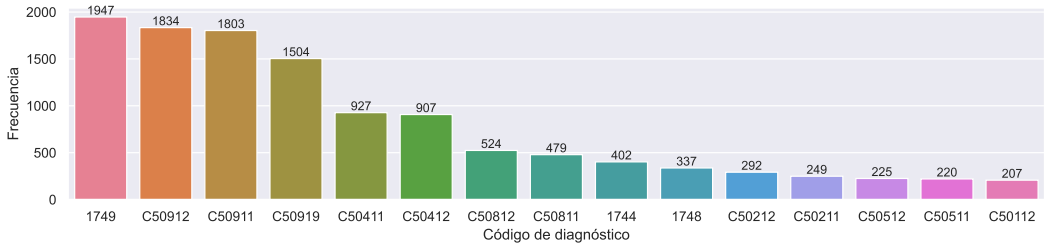
  Pese a no haber ningún valor perdido, el atributo presenta un problema de cara a su uso posterior: **todas las instancias del conjunto de datos tienen el mismo valor (mujer)**. Por tanto, el uso de este atributo no ofrece ninguna capacidad discriminatoria y puede ser descartado sin problema.

Datos médicos - códigos de diagnóstico y tipos de tratamiento

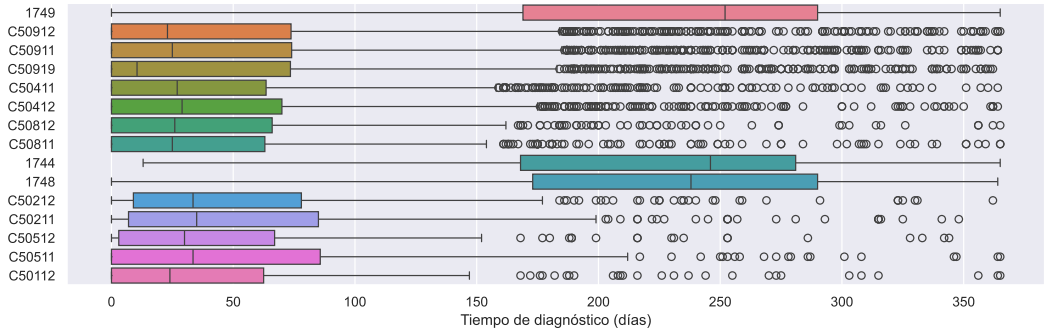
- **Código de diagnóstico de cáncer de mama:**

  Existen dos atributos en el conjunto de datos clasificando la misma información: **código de diagnóstico** y **descripción del diagnóstico**. Al representar la misma información - y tras comprobar que hay una correlación directa entre ambos atributos -, es suficiente con estudiar **uno de los dos atributos**, eligiendo estudiar el **código de diagnóstico de cáncer de mama**.

  El principal problema a la hora de estudiar este atributo es que se tienen **47 valores únicos** para el atributo, siendo un número demasiado elevado para estudiar en detalle. Además, no hay garantía de que **estos valores sean exhaustivos** - es decir, es posible que **existan códigos de diagnóstico no incluidos en el conjunto de entrenamiento**. Por esto, se realizará el estudio sobre los **15 códigos más frecuentes**.



(a) General



(b) Respecto al tiempo de diagnóstico

Figura 3.5: Distribución de los 15 valores más frecuentes del código de diagnóstico de cáncer de mama

En la **Figura 3.5a** se estudia la distribución general, y como se puede observar, **existen dos tipos de codificaciones** - las codificaciones que empiezan por la letra **C** (ICD10, más moderno) y las que empiezan por un número (ICD9). Además, se puede ver que la mayoría de diagnósticos se encuentran agrupados en cuatro códigos - con la frecuencia del resto de atributos bajando rápidamente hasta llegar a los diagnósticos con una o dos instancias no representados. Estos diagnósticos, si se comprueba su código descriptivo, hacen referencia a **cánceres en sitios sin especificar** - es decir, los códigos más genéricos y por tanto

aplicables a un mayor número de casos.

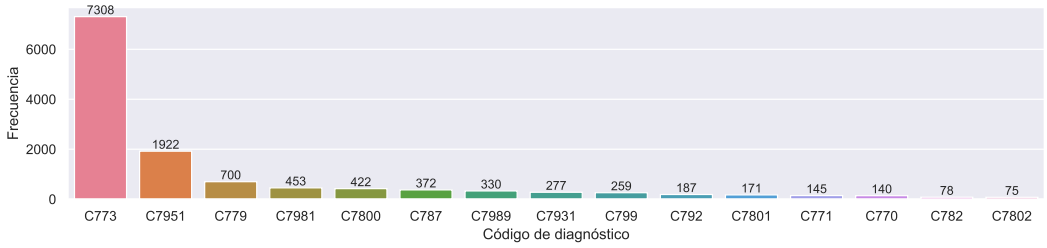
En la **Figura 3.5b** se puede apreciar una diferencia clara en tiempos de diagnóstico dependiendo de la **codificación utilizada**, con los diagnósticos con codificación de tipo **ICD9** teniendo un tiempo de diagnóstico promedio notablemente superior. Si bien no hay una explicación clara para ésto, puede deberse a que hagan referencia a casos más antiguos -y, por tanto, casos con menor conocimiento y recursos.

Por esto, resulta evidente que el código de diagnóstico de cáncer de mama ofrece información discriminatoria muy relevante de cara a ser utilizada en el modelo posterior.

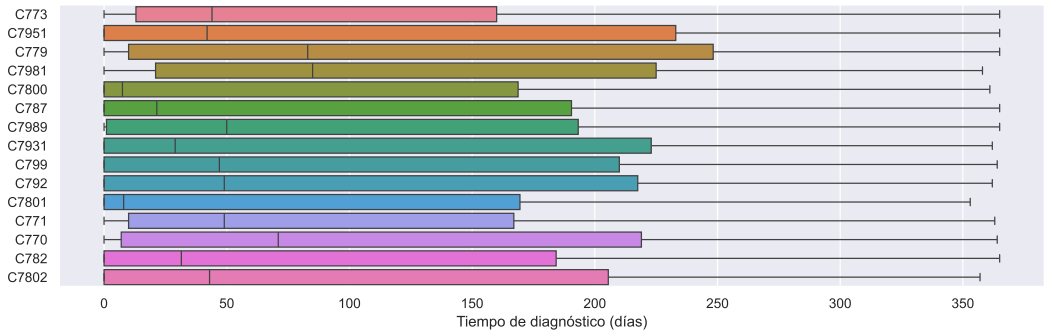
- **Código de diagnóstico de cáncer metastásico:**

  A diferencia del código de diagnóstico para cáncer de mama, para el **diagnóstico de cáncer metastásico** solo se tiene un atributo.

  Ahora bien, se sigue teniendo el mismo problema de dimensionalidad: el conjunto de datos contiene **43 valores únicos** para este atributo, sin garantía de que sea un conjunto **exhaustivo**. Por esto, se realizará el estudio sobre los **15 códigos más frecuentes**.



(a) General



(b) Respecto al tiempo de diagnóstico

Figura 3.6: Distribución de los 15 valores más frecuentes del código de diagnóstico de cáncer de mama

Como se observa en la **Figura 3.6a** y a diferencia del diagnóstico de cáncer de mama, en este caso **la mayoría de diagnósticos se concentran alrededor de un único diagnóstico** - $C773$, metástasis en nodos linfáticos auxiliares y de las extremidades superiores -, con el resto de diagnósticos reduciendo su frecuencia rápidamente.

Respecto a la relación con el tiempo de diagnóstico, en la **Figura 3.6b** se puede ver que, si bien no hay una diferencia tan pronunciada como en el caso del código de diagnóstico de cáncer de mama, **el tipo de metástasis diagnosticado parece tener influencia sobre el tiempo necesario para su diagnóstico**. Ahora bien, si se estudia la localización de la metástasis de cada código estudiado **no se observa correlación entre la localización y el tiempo de diagnóstico**.

Por esto, se puede considerar al **código de diagnóstico del cáncer metastásico** otra variable de gran interés para los modelos desarrollados posteriormente - pudiendo ofrecer una gran capacidad discriminatoria.

- **Tipo de tratamiento:**

  Como se mencionó durante el estudio de los valores perdidos, **solo se tienen 11 valores** para este atributo - de **13173** instancias totales. Por tanto, no tiene ningún sentido estudiar este atributo, al no contener suficiente información para ser significativo.

Datos geográficos - estado de residencia y ubicación geográfica

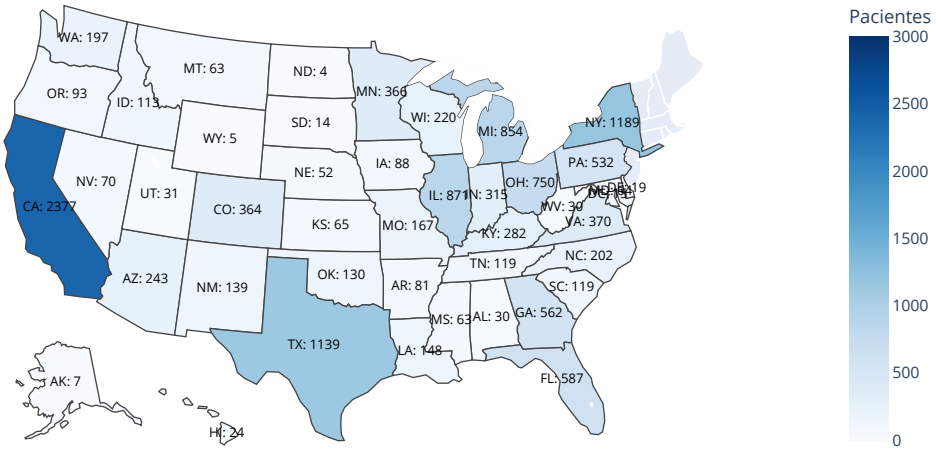
El estudio de la información geográfica del conjunto de datos es de especial importancia al ser uno de los objetivos planteados por el problema a resolver. Además, toda la información socio-económica y climática del conjunto de datos está **asociada al código zip de los pacientes** - por lo que estos atributos codifican de forma innata todos estos factores de sesgo.

Ahora bien, la información se encuentra representada en el conjunto de datos a través de **4 atributos jerárquicos**, donde cada atributo inferior describe con más granularidad el atributo superior. Al codificar la misma información, es de interés seleccionar **un único atributo** sobre el que realizar el estudio.

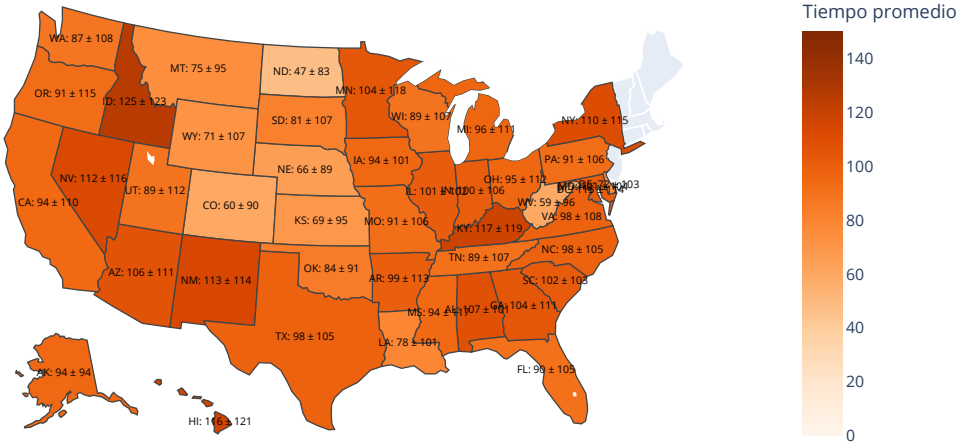
Atributo	Valores totales	p-valores (Tests de hipótesis)	
		Paramétrico (ANOVA)	No paramétrico (Kruskal)
Región	4	$2,80 \times 10^{-3}$	$1,93 \times 10^{-6}$
División	8	$6,13 \times 10^{-3}$	$1,68 \times 10^{-5}$
Estado	44	5.86×10^{-10}	1.06×10^{-16}
Código zip	751	—	—

Tabla 3.1: p-valores de los atributos geográficos

Para realizar la selección se han realizado **tests de hipótesis** - tanto paramétricos (**ANOVA**, para estudiar desviaciones en la media) como no paramétricos (**Kruskal-Wallis**, para estudiar desviaciones en la mediana) - sobre todos los atributos excepto el código zip, debido a su alta dimensionalidad. Los resultados se pueden observar en la **Tabla 3.1**, siendo el atributo a seleccionar el **estado**, al tener el p-valor más bajo - y, por tanto, tener la mayor certeza de que **su valor influye sobre el promedio del tiempo de diagnóstico**.



(a) Frecuencia



(b) Tiempo de diagnóstico promedio

Figura 3.7: Distribución geográfica de los pacientes

En el mapa de la **Figura 3.7a** se representa la distribución de los pacientes en el mapa de los Estados Unidos, donde se observa que los pacientes están **agrupados en estados concretos** - en general, los estados de mayor población -, sin haber una correlación geográfica clara en su ubicación. También se observa que **existen algunos estados sin pacientes** - por lo que, igual que con los códigos de diagnóstico, **los valores del atributo no son exhaustivos**.

Estudiando el valor promedio del tiempo de diagnóstico, la **Figura 3.7b** muestra que el **tiempo de diagnóstico promedio es similar entre todos los estados** - rondando alrededor de los **80 días**, pero ubicado en el rango de los **60 a 120 días**. También es llamativo el hecho de que **la desviación estándar es muy elevada** - en la práctica totalidad de los estados se trabaja con desviaciones estándar de alrededor de **100 días**.

El test de hipótesis indica que el estado del paciente **influye de forma significativa sobre el tiempo de diagnóstico**, por lo que es un atributo de interés de cara a la creación posterior de modelos. Ahora bien, el estudio gráfico de la distribución también muestra que existe un **error sustancial** en dicha diferencia, al haber un rango muy elevado de posibles valores dentro de cada estado.

3.2.3. Estudio de atributos numéricos

Tras el estudio de los atributos categóricos, el siguiente paso es realizar un **estudio exhaustivo** de los atributos numéricos más significativos. Sin embargo, el elevado número de atributos, con un total de **138 variables numéricas**, hace imposible el estudio individualizado. Por tanto, el objetivo es seleccionar los **principales atributos numéricos** - entendiendo como tales los **atributos con mayor influencia sobre el tiempo de diagnóstico**.

Una forma de realizar esta selección es mediante el **coeficiente de correlación de Pearson** - un valor numérico en el rango $[-1, 1]$ indicando la **relación lineal entre dos atributos**, donde los valores cercanos a los extremos indican una correlación fuerte y un valor cercano a 0 indica independencia entre los atributos.

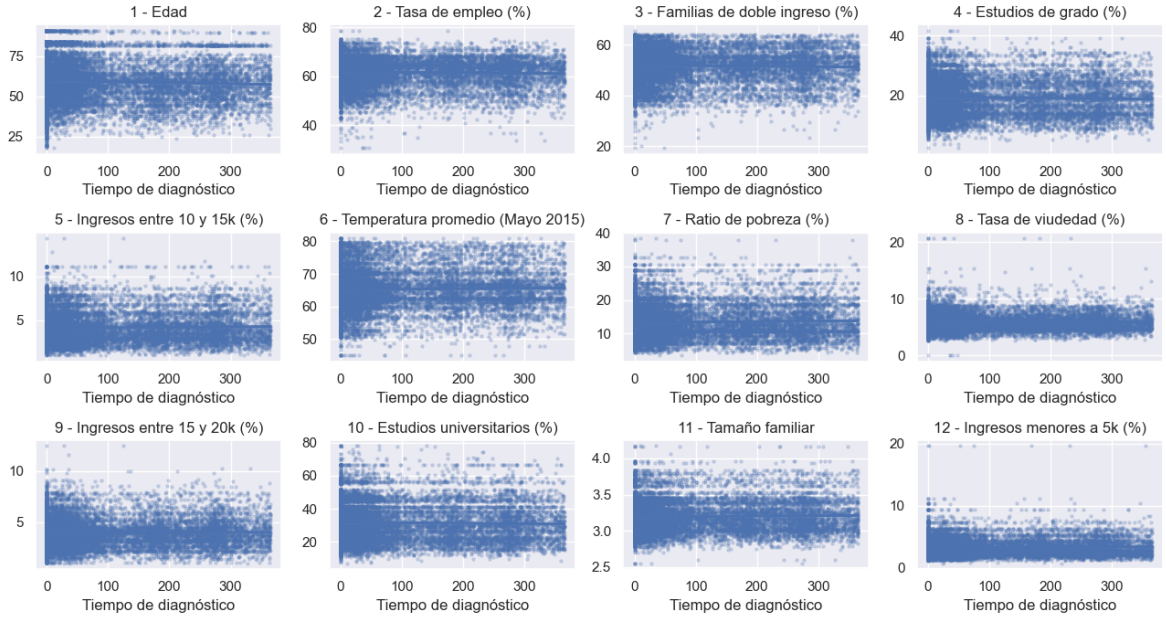


Figura 3.8: Relación entre valor y tiempo de diagnósticos para los 12 atributos de mayor correlación

Ahora bien, como se refleja en la **Figura 3.8**, cuando se calcula la correlación entre los **138 atributos** y el tiempo de diagnóstico se observa claramente que **los valores de correlación de Pearson son muy bajos** - siendo el valor más alto de 0.055. Estos valores se traducen en que **los atributos numéricos no tienen apenas influencia sobre el tiempo de diagnóstico** - y, por tanto, **pueden ser descartados** en las siguientes etapas del proceso de ciencia de datos.

Al ser la práctica totalidad de atributos socioeconómicos y climáticos de tipo numérico, esto también se traduce en una respuesta al segundo objetivo del problema: identificar que, en contra de lo que se podría esperar, **los factores socioeconómicos y climáticos no parecen tener influencia sobre el tiempo de diagnóstico** - al menos, para el conjunto de datos proporcionado.

4. Preparación del conjunto de datos

En este capítulo se describe el conjunto de transformaciones y técnicas aplicadas para preparar el conjunto de datos para su uso posterior durante el entrenamiento y experimentación.

En primer lugar se propone un número de **subconjuntos de atributos** de cara a reducir la dimensionalidad del conjunto de datos. Tras esto, se plantean y describen las **transformaciones** aplicadas al conjunto de datos previo al entrenamiento para estandarizar los datos y mejorar el rendimiento de los modelos.

4.1. Selección de atributos

Durante el análisis exploratorio se realizó un estudio exhaustivo de los atributos contenidos en el conjunto de datos, en el que se han identificado los siguientes problemas:

- **Alta dimensionalidad:** El conjunto de datos tiene **150 atributos** en total, donde la mayoría de atributos categóricos tienen **40 o más valores únicos**. Esto, unido al número de instancias bajo para dicha complejidad, puede significar que el modelo acabaría **sobreajustándose** al no poder aprender generalizaciones de forma adecuada.
- **Irrelevancia de los atributos:** De los 150 atributos estudiados, **la amplia mayoría no presentan correlación con la variable objetivo** - por lo que mantenerlos puede implicar una disminución del rendimiento final del modelo y un aumento del tiempo de entrenamiento.

Debido a esto, resulta necesario realizar una **selección de atributos** - proponiendo varios **subconjuntos de atributos** a evaluar durante la experimentación y selección de modelos, con el objetivo de optimizar el rendimiento del modelo reduciendo la dimensionalidad.

4.1.1. Selección manual de atributos

El primer subconjunto de atributos propuesto se realizar a partir de las observaciones obtenidas a través del análisis exploratorio de datos realizado en el capítulo anterior - estando éste formado por las **variables con mayor significancia para la predicción del tiempo**

de diagnóstico, según las gráficas y tests realizados. Tras este estudio, se han seleccionado los siguientes **5 atributos**:

- Código de diagnóstico del cáncer de mama.
- Código de diagnóstico del cáncer metastásico.
- Estado de residencia del paciente.
- Raza del paciente.
- Tipo de seguro médico del paciente.

Como se puede observar, **todos los atributos seleccionados son categóricos**. Esto se debe a la correlación prácticamente nula entre los atributos numéricos y el tiempo de diagnóstico. Además, se ha optado por representar la información geográfica a través del **estado de residencia** - al ser el atributo geográfico con menor valor en los tests de hipótesis.

A través de esta selección se ha reducido el conjunto de atributos de **150 a 5 atributos**, reduciendo sustancialmente la dimensionalidad. Ahora bien, los atributos elegidos siguen siendo complejos debido al gran número de valores posibles, por lo que será necesario un procesamiento posterior para **agrupar los valores menos frecuentes**.

4.1.2. Selección automática de atributos

El subconjunto de atributos propuesto en la sección anterior está basado en el análisis exploratorio realizado. Ahora bien, al basar la decisión únicamente en tests estadísticos y gráficas - sin evaluar el rendimiento -, existe la posibilidad de que se haya introducido un **sesgo personal** o que existan otros subconjuntos de atributos que ofrezcan mejor rendimiento.

Con el fin de solucionar estos problemas y de acercar el proceso de selección de atributos al funcionamiento real de los modelos, se proponen **dos subconjuntos adicionales** obtenidos a través de **algoritmos de selección automática de variables** [30] - basados en técnicas estadísticas y en entrenamiento de modelos.

Filter - Selección mediante tests estadísticos

Los **métodos de filtrado** (también conocidos como *filter*) son algoritmos que evalúan la **relevancia de cada atributo** a través de tests estadísticos, sin necesidad de entrenar ningún modelo - lo que los hace más ágiles que otros métodos, pero más genéricos e incapaces de encontrar todas las correlaciones entre grupos de atributos [30].

Para este problema se ha utilizado un **test estadístico F** - una medida de la **dependencia lineal** entre atributos -, calculando la correlación entre cada atributo numérico y el tiempo de diagnóstico. A partir de estas puntuaciones, se eligen los **10 atributos** con mayor dependencia lineal:

- **Atributos categóricos (5):**
  - **Código de diagnóstico:** Cáncer de mama y cáncer metastásico.
  - **Atributos del paciente:** Tipo de seguro médico, raza y estado de residencia del paciente.

- **Atributos numéricos (5):**
  - **Atributos del paciente:** Edad del paciente.
  - **Estadísticos socioeconómicos (porcentajes):** Tasa de empleo, habitantes con estudios de grado, familias con dos o mas ingresos y habitantes con estudios universitarios o superiores.

El subconjunto de atributos obtenido reafirma la selección manual realizada, al tener ambos conjuntos los **mismos atributos categóricos** - siendo el **código de diagnóstico del cáncer de mama** el atributo con mayor relevancia con diferencia. La principal diferencia se encuentra en que se han seleccionado además **atributos numéricos**, algunos de ellos teniendo incluso mayor relevancia que otros atributos categóricos - como la **edad del paciente**.

Wrapper - Selección mediante entrenamiento de modelos

Los **métodos de envoltura** (también conocidos como *wrapper*) son algoritmos que realizan su selección de atributos a través del **entrenamiento de un modelo de aprendizaje automático** y la selección de las variables más relevantes en base a los parámetros y pesos aprendidos por el modelo. A diferencia de los métodos de *filter* el proceso de selección suele ser más lento, pero los resultados suelen ser más fiables al trabajar con modelos reales [30].

Para este problema se ha utilizado un modelo de **Random Forest**, entrenado con los hiperparámetros por defecto de su implementación en *Scikit-Learn* - **100 árboles** sin profundidad máxima. A partir de este modelo entrenado se extraen los **10 atributos** con mayor peso sobre el modelo entrenado:

- **Atributos categóricos (4):**
  - **Código de diagnóstico:** Cáncer de mama y cáncer metastásico.
  - **Atributos del paciente:** Tipo de seguro médico y raza del paciente.
- **Atributos numéricos (6):**
  - **Atributos del paciente:** Edad e índice de masa corporal del paciente.
  - **Estadísticos socioeconómicos**: Tiempo de viaje al trabajo promedio, porcentaje de personas de raza nativa, porcentaje de habitantes con estudios STEM, porcentaje de habitantes con edades entre 40 y 49 años.

Frente a las selecciones manuales y de filtrado, el **subconjunto *wrapper* no incluye información geográfica en su selección** - optando, en su lugar, por incluir un mayor número de atributos numéricos tanto médicos como socioeconómicos.

4.2. Pre-procesamiento de los datos

A la par que se propone una selección de atributos para reducir la dimensionalidad del conjunto de datos, resulta también necesario realizar un **pre-procesamiento** - una serie de transformaciones secuenciales sobre los datos - para reducir la complejidad y paliar posibles

problemas como los valores perdidos o la codificación de los atributos categóricos. De esta forma, se busca mejorar el rendimiento de los modelos entrenados.

Ahora bien, los atributos necesitan **transformaciones distintas** dependiendo del tipo de datos que representen - siendo necesario distinguir entre atributos numéricos y categóricos. Para el trabajo descrito en la memoria, se han propuesto las siguientes transformaciones dependiendo del tipo de datos del atributo:

- **Atributos categóricos:**

  1. **Imputación de valores perdidos:** Como se estudió durante el análisis exploratorio de datos, para la mayoría de atributos categóricos **resulta de interés tratar los valores perdidos como categorías separadas**, al ser relevante para el estudio que faltasen valores.

     Por esto, se opta por reemplazar todos los valores perdidos por un **valor constante, *"Unknown"***.
  2. **Codificación:** De forma inherente, la mayoría de modelos propuestos son incapaces de trabajar con atributos categóricos - necesitando transformar estos atributos en algún tipo de codificación numérica.

     Para estos casos, lo estándar es utilizar una codificación de tipo ***One-Hot*** [31]: dividiendo el atributo x en **tantos atributos** at_i **como valores tiene la variable**, donde para cada atributo se cumple que $at_i = 1$ si $x = i$ y $at_i = 0$ si $x \neq i$ - representando de esta manera los valores categóricos en un formato numérico.

     Es importante destacar las siguientes particularidades para el problema actual:

     - Debido a que los atributos categóricos del conjunto de datos tienen una **alta dimensionalidad** y **valores no exhaustivos**, es necesario realizar una **agrupación de los atributos menos frecuentes y desconocidos** bajo un único atributo at_{other}. El umbral para considerar un atributo como poco frecuente será determinado durante el proceso de experimentación y selección de modelos.
     - La implementación de los modelos de **Gradient Boosting** codifican de forma inherente los atributos categóricos, por lo que no es necesario aplicar este paso para ellos.

- **Atributos numéricos:**

  1. **Imputación de valores perdidos:** Debido a la presencia de valores extremos en la mayoría de atributos numéricos, se reemplazan los valores perdidos por el **valor mediano del atributo al que pertenece** - ofreciendo de esta forma un valor promedio resistente a sesgos y *outliers*.
  2. **Escalado:** Por lo general, es necesario **escalar los datos** - transformar los valores de todos los atributos para que se encuentren en el mismo rango - de los atributos numéricos para el funcionamiento adecuado de los modelos.

     Para el problema descrito, se intenta evitar los problemas introducidos por los valores extremos utilizando un escalado alrededor de la **mediana y el rango intercuartil**, donde cada valor se transforma utilizando la fórmula $z(x) = \frac{x - mediana}{IQR}$.

5. Modelado y evaluación

5.1. Selección de modelos e hiperparámetros

5.2. Experimentación

5.3. Análisis de resultados

5.3.1. Rendimiento de los subconjuntos de hiperparámetros

5.3.2. Rendimiento de los modelos entrenados

5.3.3. Rendimiento del modelo final

6. Despliegue y aplicación web

6.1. Aplicación para usuario - predicción individual

6.2. Aplicación *batch* - predicción en grupo

7. Conclusiones

7.1. Trabajo futuro

Bibliografía

[1] Women in Data Science, *WiDS Datathon 2024 Challenge 2*, 2024.

[2] D. D. and, "50 Years of Data Science", *Journal of Computational and Graphical Statistics*, vol. 26, nº 4, págs. 745-766, 2017.

[3] V. Dhar, "Data science and prediction", *Commun. ACM*, vol. 56, nº 12, págs. 64-73, dic. de 2013.

[4] F. Berman et al., "Realizing the potential of data science", *Commun. ACM*, vol. 61, nº 4, págs. 67-72, mar. de 2018.

[5] V. Stodden, "The data science life cycle: a disciplined approach to advancing data science as a science", *Commun. ACM*, vol. 63, nº 7, págs. 58-66, jun. de 2020.

[6] J. M. Wing, "The Data Life Cycle", *Harvard Data Science Review*, vol. 1, nº 1, jul. de 2019, https://hdsr.mitpress.mit.edu/pub/577rq08d.

[7] M. Komorowski, D. Marshall, J. Salciccioli e Y. Crutain, "Exploratory Data Analysis", en sep. de 2016, págs. 185-203.

[8] C. Shearer, "The CRISP-DM model: the new blueprint for data mining", *Journal of data warehousing*, vol. 5, nº 4, págs. 13-22, 2000.

[9] J. Saltz, *CRISP-DM is Still the Most Popular Framework for Executing Data Science Projects - Data Science PM — datascience-pm.com*, `https://www.datascience-pm.com/crisp-dm-still-most-popular/`, [Accessed 28-05-2025], 2024.

[10] T. M. Mitchell, *Machine learning*. McGraw-hill New York, 1997, vol. 1.

[11] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[12] S. Russell y P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd. USA: Prentice Hall Press, 2009.

[13] A. Burkov, *The Hundred-Page Machine Learning Book*. 2019.

[14] Y. Tai, *A Survey Of Regression Algorithms And Connections With Deep Learning*, 2021.

[15] A. Y. Ng, “Feature selection, L1 vs. L2 regularization, and rotational invariance”, en *Proceedings of the Twenty-First International Conference on Machine Learning*, ép. ICML ’04, Banff, Alberta, Canada: Association for Computing Machinery, 2004, pág. 78.

[16] H. Zou y T. Hastie, “Regularization and Variable Selection Via the Elastic Net”, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, nº 2, págs. 301-320, mar. de 2005.

[17] A. E. Hoerl y R. W. Kennard, “Ridge Regression: Biased Estimation for Nonorthogonal Problems”, *Technometrics*, vol. 42, nº 1, págs. 80-86, 2000.

[18] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso”, *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, nº 1, págs. 267-288, 1996.

[19] A. Smola y B. Schölkopf, “A tutorial on support vector regression”, *Statistics and Computing*, vol. 14, págs. 199-222, ago. de 2004.

[20] L. Breiman, “Bagging predictors”, *Mach. Learn.*, vol. 24, nº 2, págs. 123-140, ago. de 1996.

[21] L. Breiman, “Random Forests”, *Mach. Learn.*, vol. 45, nº 1, págs. 5-32, oct. de 2001.

[22] P. Geurts, D. Ernst y L. Wehenkel, “Extremely randomized trees”, *Machine Learning*, vol. 63, nº 1, págs. 3-42, abr. de 2006.

[23] Y. Freund y R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting”, en *Proceedings of the Second European Conference on Computational Learning Theory*, ép. EuroCOLT ’95, 1995, págs. 23-37.

[24] J. H. Friedman, “Greedy function approximation: A gradient boosting machine.”, *The Annals of Statistics*, vol. 29, nº 5, págs. 1189-1232, 2001.

[25] T. Chen y C. Guestrin, “XGBoost: A Scalable Tree Boosting System”, en *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ép. KDD ’16, ACM, ago. de 2016, págs. 785-794.

[26] A. V. Dorogush, V. Ershov y A. Gulin, *CatBoost: gradient boosting with categorical features support*, 2018.

[27] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush y A. Gulin, *CatBoost: unbiased boosting with categorical features*, 2019.

[28] G. Ke et al., “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”, en *Advances in Neural Information Processing Systems*, I. Guyon et al., eds., vol. 30, Curran Associates, Inc., 2017.

[29] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python”, *Journal of Machine Learning Research*, vol. 12, págs. 2825-2830, 2011.

[30] I. Guyon y A. Elisseeff, “An Introduction of Variable and Feature Selection”, *J. Machine Learning Research Special Issue on Variable and Feature Selection*, vol. 3, págs. 1157-1182, ene. de 2003.

[31] N. Draper y H. Smith, *Applied regression analysis* (Wiley series in probability and mathematical statistics). New York [u.a.]: Wiley, 1966, IX, 407.

A. Contenidos del repositorio