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

Índice de algoritmos

Índice de listados de código

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 los **modelos de regresión** - haciendo especial énfasis en los modelos de *ensemble* basados en técnicas de ***Gradient Boosting*** - o la búsqueda de hiperparámetros.

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

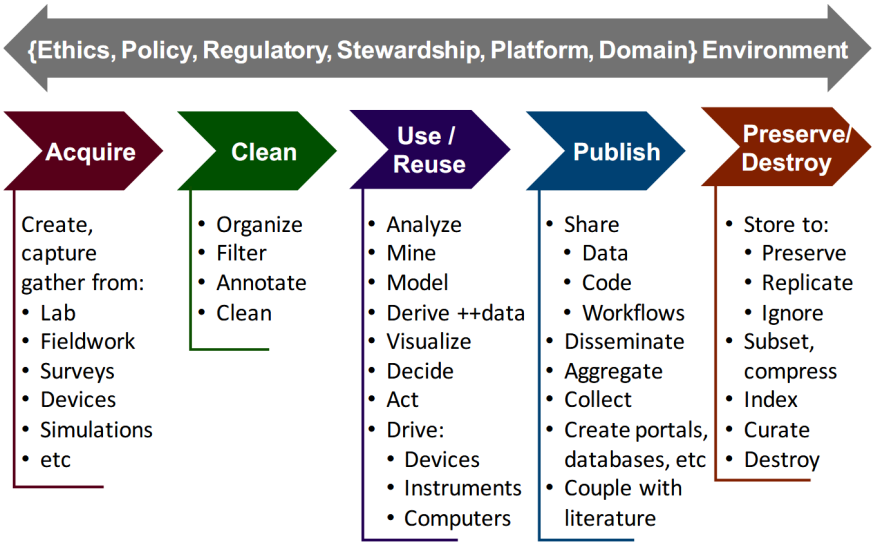

Figura 2.1: Ciclo de vida de los datos [4]

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].

2.1.1. 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] [10].

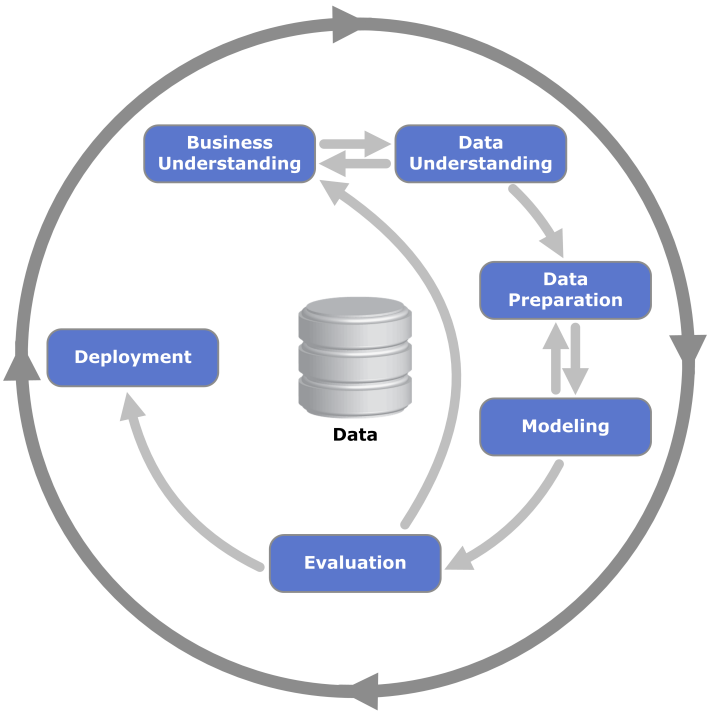

Figura 2.2: Ciclo de vida de CRISP-DM [8]

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 - 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 supervisado y modelos de regresión

2.2.1. Modelos simples

Modelos de regresión lineal

Árboles de decisión

Máquinas de vectores de soporte

2.2.2. Modelos grupales - Ensembles

Bagging

Boosting

Gradient Boosting

2.2.3. Ajuste de hiperparámetros y validación cruzada

3. Estudio del problema

3.1. Definición del problema

3.1.1. Atributos del problema

3.2. Análisis exploratorio de datos

3.2.1. Variable objetivo - distribución y comportamiento

3.2.2. Valores perdidos

3.2.3. Atributos categóricos

3.2.4. Atributos numéricos

3.2.5. Variables geográficas, sociales y económicas

4. Preprocesamiento del conjunto de datos

4.1. Selección de atributos

4.2. Procesamiento de los datos

5. Modelado y experimentación

5.1. Selección de modelos

5.2. Experimentación

5.2.1. Ajuste de hiperparámetros y selección de subconjuntos de atributos

5.2.2. Validación y selección de modelo final

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. 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. dirección: https://kaggle.com/competitions/widsdatathon2024-challenge2.

[2] D. D. and, "50 Years of Data Science", *Journal of Computational and Graphical Statistics*, vol. 26, nº 4, págs. 745-766, 2017. DOI: 10.1080/10618600.2017.1384734. eprint: https://doi.org/10.1080/10618600.2017.1384734. dirección: https://doi.org/10.1080/10618600.2017.1384734.

[3] V. Dhar, "Data science and prediction", *Commun. ACM*, vol. 56, nº 12, págs. 64-73, dic. de 2013, ISSN: 0001-0782. DOI: 10.1145/2500499. dirección: https://doi.org/10.1145/2500499.

[4] F. Berman et al., "Realizing the potential of data science", *Commun. ACM*, vol. 61, nº 4, págs. 67-72, mar. de 2018, ISSN: 0001-0782. DOI: 10.1145/3188721. dirección: https://doi.org/10.1145/3188721.

[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, ISSN: 0001-0782. DOI: 10.1145/3360646. dirección: https://doi.org/10.1145/3360646.

[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, ISBN: 978-3-319-43740-8. DOI: 10.1007/978-3-319-43742-2_15.

[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] G. Mariscal, Ó. Marbán y C. Fernández, "A survey of data mining and knowledge discovery process models and methodologies", *The Knowledge Engineering Review*, vol. 25, nº 2, págs. 137-166, 2010. DOI: 10.1017/S0269888910000032.

[10] 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.

A. Anexo 1