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

2.1. Ciencia de datos

2.1.1. Búsqueda de hiperparámetros y validación cruzada

2.2. Modelos propuestos

2.2.1. Modelos lineales

2.2.2. Máquinas de vectores de soporte

2.2.3. Árboles de decisión y *ensembles*

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.

A. Anexo 1