Modelo supervisado de clasificación binaria para el diagnóstico de cáncer de seno con la base de datos de la Universiad de Wisconsin.

Nicolás Andrés Yaya Tomases

Universidad del Norte
Departamento de Matemáticas y Estadística

Barranquilla, Colombia.

Agosto de 2020.

Índice general

Índice de figuras

1 Generalidades

1.1 Descripción del Problema

Según la Sociedad Americana de Cancer (ACS por sus siglas en inglés) el cáncer se puede originar en cualquier parte del cuerpo. Este tiene su origen cuando las células empiezan a crecer descontroladamente logrando que su tamaño supere al de las células normales, dificultando así el funcionamiento habitual del organismo.

El *cáncer de seno* se origina cuando las células mamarias empiezan a aumentar su tamaño de manera descontrolada, llegando generalmente a formar tumores que se pueden percibir al tacto o mediante una radiografía. Este tipo de cáncer es uno de lo más comunes, con una alta tasa de mortalidad en las mujeres [2] (aunque también puede ser padecido por hombres). En este sentido, su oportuno diagnóstico juega un papel fundamental en la salud del paciente, convirtiendose en un elemento determinante a la hora de seguir un tratamiento o intervención.

Por tal razón, existen numerosos modelos de clasificación en Minería de Datos que están siendo aplicados en el diagnóstico y prevención del cáncer, no siendo el cáncer de seno la excepción. Estos modelos están basados en los registros clínicos históricos de los pacientes. De allí se extraen características que permitan comparar los casos nuevos de cáncer de seno con los casos registrados, para servir de apoyo al momento de elaborar un diagnóstico o también para definir ciertos parámetros que permitan desarrollar un modelo de predicción.

El Dr. William H. Wolberg, del departamento de cirugía general de la Universidad de Wisconsin; W. Nick Street y Olvi L. Mangasarian, ambos del departamento de Ciencias de la Computación de la Universidad de Wisconsin, han elaborado un repositorio de 569 casos de cáncer de seno en mujeres, 357 benignos y 212 malignos. Se encargaron de analizar imágenes computarizadas de núcleos celulares extraidos por aguja de las masas mamarias, destacando ciertas carácterísticas importantes o relevantes que pueden servir para un posterior análisis estadístico.

Así, se pretende realizar un análisis exploratorio de los datos y posteriormente comparar la eficacia de los distintos métodos de clasificación en aprendizaje supervisado, que permitan decidir si un tumor es benigno o maligno a partir de sus características celulares. Además se pretende mostrar los elementos matemáticos de fondo que justifican el funcionamiento de dichos métodos, así como su eficacia.

1.2 Justificación

La elaboración de un modelo de clasificación binaria para el diagnóstico de cáncer de seno basado en los datos antes mencionados constituye un medio de aplicación para las técnicas de estadística, probabilidad, optimización y ciencias de la computación que han podido ser adquiridos a lo largo del pregrado en matemáticas.

A nivel académico, se desarrollará un documento que contará con la exploración de conocimientos en el área de aprendizaje automático, sin descuidar el sustento matemático, sirviendo así de guía para futuros estudiantes de la carrera que estén interesados en elaborar proyectos basados en datos.

1.3 Estado del arte

Es evidente que el desarrollo tecnológico de las ultimas décadas ha facilitado la realización de ciertas tareas, pero también ha creado nuevas necesidades. Los avances en procesamiento, almacenamiento, memoria y redes de transmisión de información han abierto la posibilidad a otras disciplinas de resolver los problemas propios de su área sirviendose de las nuevas tecnologías. El aprendizaje automático ha venido siendo utilizado por la medicina de diversas maneras, estas van desde el análisis de los tiempos de espera en las salas de urgencia hasta el diagnóstico y tratamiento de enfermedades. Debido a que los resultados de estas investigaciones han sido bastante productivos, el sector médico ha aumentado su interés y confianza en los modelos basados en datos [5].

En el año 2006, Cruz, J. A. & Wishart, D. S. realizaron una revisión bibliográfica sobre las aplicaciones de modelos de Aprendizaje automático para el diagnóstico de cáncer, encontrando en los estudios más prestigiosos y relevantes que la implementación de estos modelos puede mejorar significativamente (de 10 a 25 %) la exactitud de predecir susceptibilidad al cáncer [1]. Además concluyen que el aprendizaje automático no solo está ayudando a realizar mejores pronósticos, sino que también está brindando un mejor entendimiento de la enfermedad a los especialistas.

Adicionalmente, un grupo de investigadores publicaron en 2012 un artículo [4] en la *Journal of Medical Systems*, este se enfocaba hacia el diagnóstico de cáncer de seno con métodos de clasificación de aprendizaje automático aplicado a mamografías. Evaluaron masivamente distintas configuraciones de estos modelos para clasificar vectores con ciertas características que habían sido extraídas de regiones segmentadas de una mamografía desde dos proyecciones, craneocaudal y medio-lateral oblicua. En total evaluaron 286 casos de un repositorio de la Facultad de Medicina de la Universidad de Porto (FMUP), ejecutando alrededor de 20.000 modelos de clasificación, encontrando que estos clasificadores alcanzan un 99.6 % de exactitud cuando se combinan vectores de características de los dos tipos de mamografías, antes mencionadas, sobre un mismo caso.

Por otro lado, investigadores de la Universidad estatal de Nueva York en Binghamton y de la Universidad del sur de Illinois en Edwardsville, en lugar de estudiar los métodos de toma de datos que mejoran los resultados de los modelos, se centraron en estudiar la efectividad de 12 tipos de Máquinas de Soporte Vectorial (SVM por sus siglas en inglés) para el diagnóstico de cáncer de seno [6]. Utilizaron distintos repositorios de datos con el objetivo de reducir la varianza y aumentar la exactitud en el diagnóstico del cáncer. Lograron proponer un modelo SVM que reduce la varianza en un 97.98 % y aumenta la exactitud del diagnóstico en un 33.34% en comparación con el mejor modelo SVM que ya operaba con el conjunto de datos. Además manifiestan que la mejora que ellos

implementaron no solamente puede aplicarse a este tipo de diagnósticos sino que puede extenderse a procesos más robustos de detección de enfermedades.

Un estudio adicional, de Montazeri et. al., publicado en la revista *Technology and Health Care* en el año 2016, se orientó hacia la implementación de técnicas de Aprendizaje automático para la predicción de supervivencia en casos de cáncer de seno [3]. Utilizaron una base de datos de 900 pacientes, 876 mujeres y 24 hombres. El modelo propuesto es una combinación de distintas técnicas de clasificación, dentro de ellas Navie Bayes (NB), Bosques de decisión aleatorio (TRF), Vecino Más Cercano (1NN), Boosting Adaptativo (AD), Máquinas de Soporte Vectorial (SVM), Redes de Base Radial (RBFN) y Perceptrón Multicapa (MLP). Como resultado del estudio se obtuvo que el método de Bosques de decisión aleatorio mostró mejores resultados (96 % de exactitud) en comparación a las otras técnicas. Además, el método del Vecino Más Cercano mostró el peor rendimiento, con una exactitud del 91 %.

1.4 Objetivos

1.4.1 Objetivo General

Elaborar un modelo supervisado de clasificación binaria para el diagnóstico de Cáncer de seno mediante la implementación de técnicas de Aprendizaje automático a la base de datos de la Universidad de Wisconsin.

1.4.2 Objetivos Específicos

- Explicar la relevancia de las características extraídas de una imagen computarizada de núcleos celulares de masas mamarias que servirán de soporte para el modelo de clasificación.

- Realizar un proceso de filtro y limpieza de datos para el posterior entrenamiento y validación del modelo de Aprendizaje automático.

- Describir y establecer cuáles son los distintos métodos de clasificación en aprendizaje supervisado, identificados en la revisión literaria, que se pueden usar para el diagnóstico de cáncer de seno.

- Comparar la eficacia de los distintos métodos de clasificación al implementarlos en la base de datos de Cáncer de seno de la Universidad de Wisconsin.

1.5 Metodología

Se pretende desarrollar los siguientes pasos para la implementación del modelo.

[1] Definición de las clases objetivo[1].

[2] Se filtran y normalizan los datos, retirando las características que no presenten una variación significativa.

[3] Se realiza el análisis descriptivo y la visualización de los datos.

[4] Se selecciona y construye el clasificador con el 70 % de los datos.

[5] Se utiliza el restante 30 % de los datos para validar la exactitud del modelo.

[6] Se ajustan los parámetros del clasificador y se repiten los dos pasos anteriores hasta minimizar el error.

[7] Se evalúa el resultado final del clasificador.

[1] Las categorías: maligno y benigno.

1.6 Organización y cronograma

El siguiente cronograma es la propuesta inicial de desarrollo de actividades. Está sujeto a modificación según en discurrir del curso.

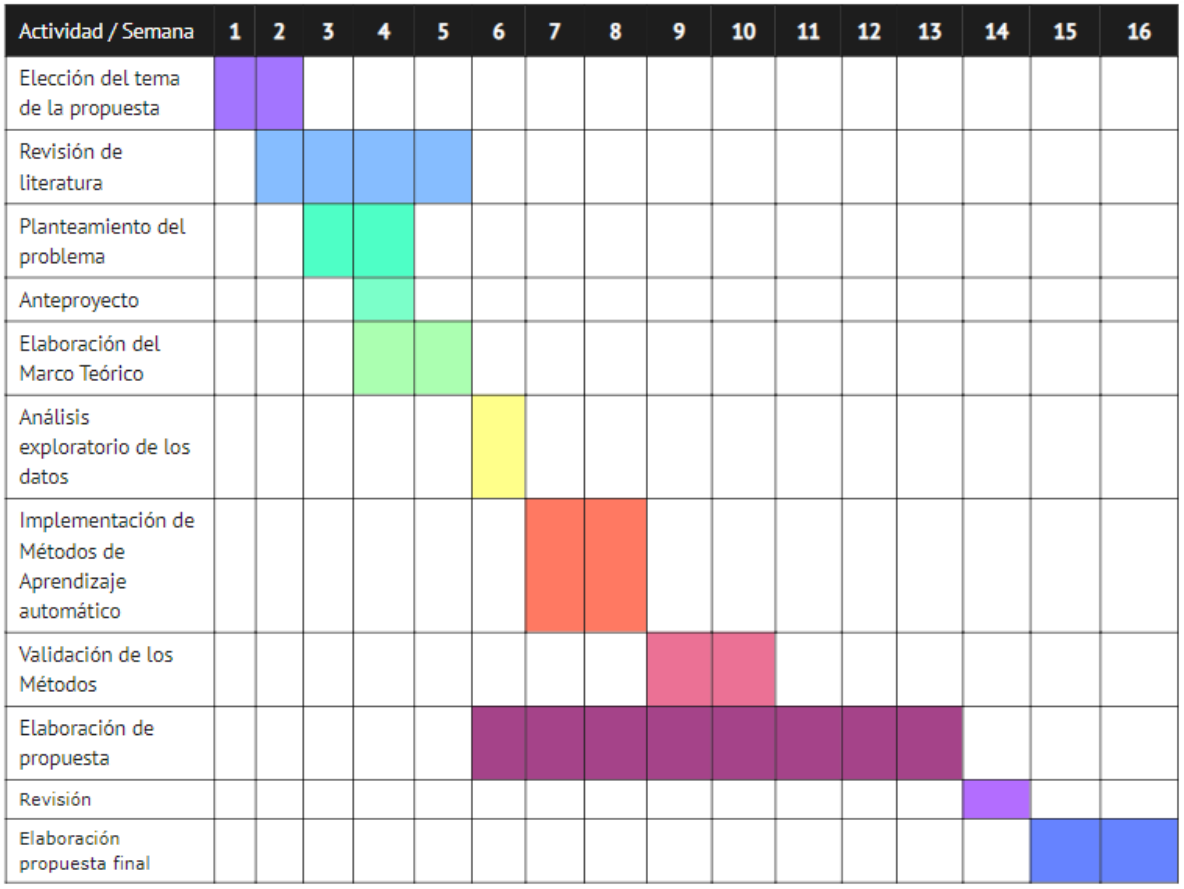

Figura 1: Cronograma de trabajo.

Referencias

[1] Cruz, J. A., & Wishart, D. S.. *Applications of machine learning in cancer prediction and prognosis*, 2006. Cancer informatics, 2, 117693510600200030.

[2] DeSantis, C., Ma, J., Bryan, L., & Jemal, A. (2014). *Breast cancer statistics*, 2013. CA: a cancer journal for clinicians, 64(1), 52-62.

[3] Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A. *Machine learning models in breast cancer survival prediction*, 2016. Technology and Health Care, 24(1), 31-42.

[4] Ramos-Pollán, R., Guevara-López, M.A., Suárez-Ortega, C. et al. *Discovering Mammography-based Machine Learning Classifiers for Breast Cancer Diagnosis*, 2012. J Med Syst 36, 2259–2269 .

https://doi.org/10.1007/s10916-011-9693-2

[5] Sánchez Gómez, C. *Desarrollo de soluciones software mediante aprendizaje automático en el ámbito de la salud: situación tecnológica y perspectivas*, 2019.

[6] Wang, H., Zheng, B., Yoon, S. W., & Ko, H. S. *A support vector machine-based ensemble algorithm for breast cancer diagnosis*, 2018. European Journal of Operational Research, 267(2), 687-699.