UNIVERSIDAD DE MÁLAGA

E.T.S. INGENIERÍA INFORMÁTICA
UNIVERSIDAD DE MÁLAGA

Grado en Ingeniería de la Salud

Mención en Bioinformática

Métodos computacionales para la interpretabilidad de los resultados bioinformáticos en el ámbito clínico

Computational methods for the interpretability of bioinformatic results in the clinical setting

Realizado por
Soledad del Castillo Carrera

Tutorizado por
Prof. José Manuel Jerez Aragonés

Co-tutorizado por
Julio Montes Torres

Departamento
Lenguajes y Ciencias de la Computación
UNIVERSIDAD DE MÁLAGA

Málaga, junio 2022

UNIVERSIDAD DE MÁLAGA

E.T.S. INGENIERÍA INFORMÁTICA
UNIVERSIDAD DE MÁLAGA

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA
GRADUADA EN INGENIERÍA DE LA SALUD CON MENCIÓN EN BIOINFORMÁTICA

MÉTODOS COMPUTACIONALES PARA LA INTERPRETABILIDAD DE LOS RESULTADOS BIOINFORMÁTICOS EN EL ÁMBITO CLÍNICO

COMPUTATIONAL METHODS FOR THE INTERPRETABILITY OF BIOINFORMATIC RESULTS IN THE CLINICAL SETTING

Realizado por
Soledad del Castillo Carrera

Tutorizado por
Prof. José Manuel Jerez Aragonés

Co-tutorizado por
Julio Montes Torres

Departamento
Lenguajes y Ciencias de la Computación

UNIVERSIDAD DE MÁLAGA
MÁLAGA, JUNIO DE 2022

Fecha defensa: julio 2022

Resumen

Actualmente, uno de los mayores desafíos de las técnicas de aprendizaje computacional es la predicción en el dominio sanitario. La Inteligencia Artificial se usa para el apoyo a la toma de decisiones médicas, por lo cual los algoritmos se enfrentan a conjuntos de datos inestables o incompletos, así como a la incertidumbre y/o ambigüedad en los resultados.

Es por ello por lo que ha surgido la necesidad de modelar algoritmos de aprendizaje computacional para la ayuda de la toma de decisiones médicas. Los algoritmos modelados en este proyecto para dicha ayuda son Redes Neuronales y K-Vecinos más cercanos (KNN). Por otro lado, también se puede usar el Razonamiento basado en casos (CBR) para realizar estas predicciones.

Se van a implementar tres scripts en R [1], desarrollando en cada uno de ellos los métodos anteriormente nombrados. El objetivo de estos scripts va a ser el de hacer el preprocesamiento de los datos clínicos, para, posteriormente, entrenarlos, validarlos y testearlos con la finalidad de obtener una predicción para la clasificación.

Finalmente, se hará una comparación de los resultados obtenidos con los distintos métodos, con objeto de evaluar la semejanza de las predicciones obtenidas con CBR y los demás.

Palabras clave: Inteligencia Artificial, Redes Neuronales, K-Vecinos más cercanos, CBR, Ayuda al diagnóstico.

Abstract

Nowadays, one of the biggest challenges of machine learning techniques is prediction in the healthcare domain. Artificial intelligence is used to support medical decision-making, so algorithms face unstable or incomplete data sets, as well as uncertainty and/or ambiguity in results.

That is why the need to model machine learning algorithms for medical diagnostic support has arisen. The algorithms modeled in this project for such help are Neural Networks and K-Nearest Neighbors. On the other hand, Case-Based Reasoning (CBR) can also be used to make these predictions.

Two R-scripts are to be implemented, each developing the models. The objective of these scripts will be to preprocess clinical data, then train, validate and test them in order to obtain a prediction for classification.

Finally, we will compare the results obtained using these methods, aiming to assess the similarity of the predictions obtained with CBR and the others.

Keywords: Artificial Intelligence, Neural Networks, K-Nearest Neighbors, CBR, Diagnostic support.

Índice

Índice de figuras

Índice de tablas

1

Introducción

La **Inteligencia Artificial (IA)** es un conjunto de tecnologías que hace que las máquinas aprendan de la experiencia, así como que realicen tareas como seres humanos. Empleando las técnicas de aprendizaje computacional, aprendizaje profundo y al procesamiento del lenguaje natural, las computadoras son entrenadas para realizar tareas específicas procesando grandes cantidades de datos y reconociendo patrones en dichos datos [2].

La IA en la actualidad podría clasificarse según dos estrategias de aplicación: Aprendizaje Automático (Machine Learning) y Aprendizaje Profundo (Deep Learning).

El **Aprendizaje Automático (Machine Learning)** es una rama de la inteligencia artificial que se basa principalmente en la idea de que los sistemas pueden aprender de datos, identificar en ellos patrones y tomar decisiones con una mínima intervención humana. Es decir, los algoritmos de Machine Learning aprenden con la experiencia [3].

El **Aprendizaje Profundo (Deep Learning)** es una especialización del Aprendizaje Automático que configura parámetros básicos acerca de los datos recibidos y entrena a

la máquina para que aprenda por sí misma reconociendo patrones mediante el uso de muchas capas de procesamiento [4].

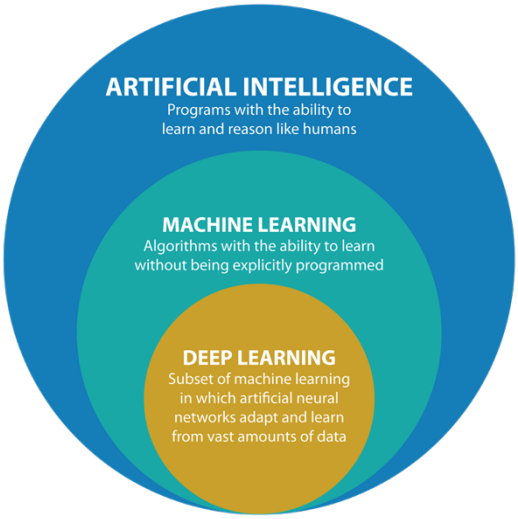

Figura 1. La Inteligencia Artificial abarca los términos de Machine Learning y Deep Learning.

Imagen obtenida en [5].

Otra parte importante de la Inteligencia Artificial es el Razonamiento Basado en Casos (CBR). Al inicio de su nacimiento, este algoritmo solo se utilizaba para un área de investigación muy reducida. Hoy en día, se ha convertido en una materia de amplio interés y multidisciplinar. Esto se debe a que el CBR, en lugar de confiar solamente en el conocimiento general del problema, es capaz de utilizar conocimiento específico de experiencias previas, de casos concretos. Es decir, cuando se presenta un problema nuevo se resuelve recordando un problema pasado similar y reutilizando el conocimiento y la información[6].

1.1 Motivación

Como se ha comentado anteriormente, el dominio clínico (tal como la medicina) es uno de los mayores desafíos de la Inteligencia Artificial. Cada vez hay más profesionales de este dominio que confían en la predicción de los algoritmos de inteligencia artificial,

por lo que nuestro deber es refinarlos hasta que la predicción sea la más exacta posible para poder servir de base y ayudar a las toma de decisión en el diagnóstico.

Es por ello que la principal motivación de este proyecto es la de implementar métodos de Inteligencia Artificial que sirvan de ayuda al diagnóstico para un paciente de cáncer, así como que en el futuro estas técnicas se puedan adaptar a otras especialidades del dominio clínico.

Esto puede ser beneficioso tanto para médicos como para pacientes, por lo que es indispensable que la ayuda sea lo más certera posible y le facilite la decisión al médico, ya que una predicción errónea podría ser muy perjudicial para los pacientes.

1.2 Objetivos

El objetivo principal de este proyecto es el de comparar los resultados obtenidos con las técnicas tradiciones de Aprendizaje Automático y los obtenidos con la técnica de Razonamiento Basado en Casos, para ver cuál es el que mejor se adaptaría para la ayuda a la toma del diagnóstico médico.

Se van a implementar un algoritmo de Redes Neuronales, otro de K-Nearest Neighbors y otro de CBR para cada uno de los conjuntos de datos seleccionados.

Todos los datos seleccionados recogen información sobre pacientes con cáncer de mama en tratamiento, teniendo que predecir en cada uno de ellos una variable distinta: en el primero, se va a predecir la clase del tumor (si es benigno o maligno); en el segundo, la severidad del tumor (si es benigno o maligno), y en el tercero si la paciente ha sufrido una recaída o no.

Una vez implementados y probados con cada conjunto de datos, se compararán para elegir el mejor modelo que se adapte al conjunto y ese modelo será el que se le mostrará

gráficamente a los profesionales clínicos para que tomen la decisión final en el diagnóstico.

1.3 Estructura de la memoria

El proyecto se va a dividir en 4 capítulos más, los cuales van a ser: estado del arte, en el que se va a hablar de dos artículos de estudios similares al realizado en este proyecto y en los que nos hemos basado para algunas partes del proyecto; metodología, en el cual se van a describir todas las metodologías usadas; desarrollo, en el que se va a contar como ha sido el desarrollo del proyecto; resultados y discusión, en el que se van a mostrar los resultados obtenidos y se van a discutir, y conclusión y líneas futuras, en el que se va a hablar de las posibles líneas futuras del proyecto y sobre las conclusiones que se han obtenido de él.

2

Estado del arte

Para la realización de este proyecto se han tomado como base tres artículos científicos que abarcan el tema propuesto. Cada uno de ellos abarca una parte diferente del proyecto, por lo que se van a plantear por separado.

2.1 Inteligencia artificial en medicina

En este artículo se propone un método CBR que se puede ejecutar como un algoritmo y presentarse posteriormente de manera visual para proporcionar una explicación visual de los resultados.

En el caso de aplicar el método CBR en la medicina, los casos se corresponden con los pacientes y el problema a resolver es la clasificación de un paciente nuevo. Es, por ello, que hay dos clases bases para predecir en la medicina: si sufre el trastorno o si no lo sufre.

Es, debido a esto, que el algoritmo tiene una base de datos de casos que contiene a los pacientes anteriormente clasificados, de los cuales se conoce el diagnóstico y el tratamiento, a la cual accede cuando tiene que hacer una nueva clasificación.

En este artículo proponen el siguiente enfoque:

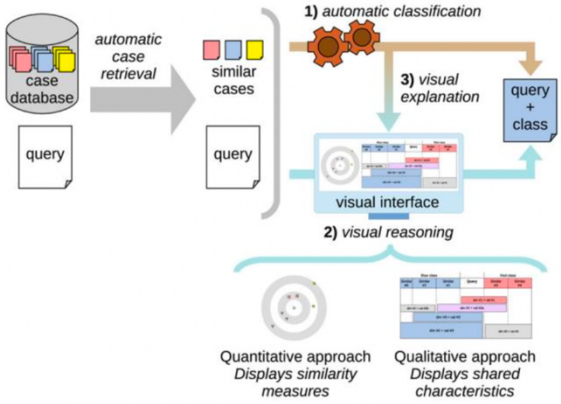

Figura 2. Enfoque CBR.

Una vez que el algoritmo recibe el caso nuevo y recupera los casos similares a este, una interfaz visual muestra las similitudes cuantitativas y cualitativas entre la consulta del caso nuevo y el caso similar, de modo que el caso se puede clasificar de forma visual.

Al combinar el enfoque cuantitativo y el enfoque cualitativo, lo que se espera es que la predicción obtenida del método sea más precisa y ayude a clasificar al nuevo paciente de manera más eficaz que la que se usa en los métodos básicos de la Inteligencia Artificial.

El artículo que se ha tomado como referencia en esta sección es: *'Explainable artificial intelligence for breast cancer: A visual case-based T reasoning approach'* [7].

2.2 Causabilidad y explicabilidad de la IA en medicina

Tal y como se dice en este artículo, la inteligencia artificial explicable atrae mucho interés en la medicina. Esta se ocupa de la implementación y trazabilidad de los métodos de aprendizaje automático de caja negra, centrándose en los de aprendizaje profundo.

No obstante, se afirma que, además de la explicabilidad, es necesario tener en cuenta la causabilidad.

Se enfatiza que los sistemas de la Inteligencia Artificial deberían poder construir modelos causales, los cuales respaldan la explicación, en lugar de resolver simplemente un problema de reconocimiento de patrones. Sin embargo, esto es algo difícil de alcanzar porque no solo se necesita aprender de los datos previos, extraer conocimiento, generalizar y tratar con la dimensionalidad, sino que se necesita desentrañar los factores explicativos subyacentes de los datos para poder entender el contexto del dominio de aplicación.

Además, hay que tener en cuenta que estos estudios en la medicina se enfrentan a conjuntos de datos inestables, ya sea por datos desconocidos, incompletos, ruidosos, erróneos, inexactos… Por eso, para poder explicar la Inteligencia Artificial en este contexto es necesario tener en cuenta que muchos datos pueden contribuir a un resultado relevante, lo cual requiere que los profesionales de este ámbito sean capaces de comprender cómo y por qué se ha tomado esa decisión.

En el artículo distinguen dos tipos de IA explicables, las cuales son:

- Explicabilidad posthoc: ocurre después del evento.
- Explicabilidad ante-hoc: ocurre antes del evento.

La diferencia de estas dos es que la primera predice el modelo basándose en lo que es fácilmente interpretable, mientras que la segunda incorpora la explicabilidad en la estructura del modelo, siendo la explicabilidad parte del diseño.

Uno de los métodos de ejemplo que se proporciona en este artículo es el de las redes neuronales artificiales. Afirman que estas son aplicables a muchos problemas prácticos, en los cuales, además, se obtienen buenos resultados.

El artículo que se ha tomado como referencia en esta sección es: *'Causability and explainability of artificial intelligence in medicine'* [8].

2.3 Artificial Feeding Birds (AFB)

En este proyecto, para el cálculo de los ángulos de los casos CBR se ha tomado como base la metaheurística descrita en el artículo *'Artificial Feeding Birds (AFB): a new metaheuristic inspired by the behavior of pigeons'* [30].

El AFB es una nueva metaheurística inspirada en el comportamiento de las palomas a la hora de buscar alimentos. Este artículo afirma que las palomas, cuando no tienen comida a simple vista, la buscan en su entorno moviéndose de la única forma que saben: caminando y volando. Es, por ello, que el algoritmo va a tener que simular cómo las palomas caminan y vuelan.

El algoritmo, teóricamente, funciona de la siguiente manera:

1. La paloma camina lentamente hacia una nueva posición cercana a la actual.
2. Vuela y aterriza en una posición aleatoria.
3. Vuela y regresa a una posición conocida rica en comida.
4. Vuela y aterriza junto a otra paloma.

Esto optimiza la búsqueda de los alimentos, ya que el primer paso permite una búsqueda local y cercana al individuo; el segundo, permite la explorar el espacio de forma aleatoria; el tercero, permite recuperar comida de una posición conocida o seguir buscándola por los alrededores, y el cuarto, el cual puede permitirle al individuo beneficiarse de la comida que otro individuo hubiese podido encontrar.

Esta metaheurística realiza varios ciclos en los cuales cada ave realiza uno de los movimientos descritos anteriormente. El algoritmo se define como un problema de optimización que utiliza tres funciones principalmente: costo, volar y caminar. Con estas funciones se pretende encontrar la solución para el problema.

Cuando el proceso termina, la mejor solución encontrada es la posición memorizada por las aves.

Tomando como referencia esta información, en este proyecto se va a usar esta metaheurística para construir el modelo de Razonamiento Basado en Casos.

3

Metodología

Como se ha comentado anteriormente, se van a implementar algoritmos de Redes Neuronales, K-Nearest Neighbors y Razonamiento Basado en Casos.

Los dos primeros algoritmos son parte de lo que se conoce como algoritmos de Aprendizaje Automático. A pesar de que Redes Neuronales y KNN son algoritmos muy diferentes entre sí, comparten la base común de los algoritmos de ML. Dado un conjunto de datos de entrada, se hace una partición con el esquema deseado, en nuestro caso 60-20-20: el primer subconjunto de datos, llamado conjunto de entrenamiento, se va a entrenar con el 60% de los datos iniciales; el segundo, llamado conjunto de validación, va a validar un 20% de los datos restantes con el modelo de entrenamiento obtenido anteriormente, y el tercero, llamado conjunto de prueba, va a probar el último 20% de los datos con los modelos obtenidos de la validación de datos.

Lo que difiere en cada algoritmo es la forma en la que este entrena, valida y prueba los datos.

En este capítulo, se va a profundizar en la metodología que siguen los algoritmos elegidos, pero para ello previamente también hay que describir los conjuntos de datos elegidos.

3.1 Conjuntos de datos de entrada

Para este proyecto se van a utilizar tres conjuntos de datos, todos ellos relacionados, de forma general, con datos de pacientes con cáncer. Estos datos se han obtenido del repositorio UC Irvine Machine Learning [9].

De forma más específica, se va a detallar la información contenida en cada conjunto de datos.

3.1.1 Conjunto de datos 1

La información contenida en este conjunto de datos se corresponde con los resultados obtenidos en el análisis de una **citología** a pacientes con cáncer de mama[10]. Las características de las que se han tomado nota en la citología han sido las siguientes: el grado en que los agregados de células epiteliales eran monocapa o multicapa (grosor del grupo); la cohesión de las células periféricas de los agregados de células epiteliales (adhesión marginal); el diámetro de la población de las células epiteliales más grandes en relación con los eritrocitos, la proporción de núcleos epiteliales únicos desprovistos de citoplasma circundante (núcleos desnudos), la blancura de la cromatina nuclear, los nucleolos normales, las mitosis poco frecuentes, la uniformidad del tamaño de las células epiteliales y la uniformidad de la forma celular.

Estas variables se han agrupado en 11 columnas, las cuales son:

1. Id: Número de identificación de la muestra.
2. Grosor del grupo (*cthickness*): toma valores en el intervalo 1-10.
3. Uniformidad del tamaño de celda (*csize*): toma valores en el intervalo 1-10.

4. Uniformidad de la forma celular (*cshape*): toma valores en el intervalo 1-10.
5. Adhesión marginal (*adhesion*): toma valores en el intervalo 1-10.
6. Tamaño de la célula epitelial única (*epi*): toma valores en el intervalo 1-10.
7. Núcleos desnudos(*bnuc*): toma valores en el intervalo 1-10.
8. Cromatina suave(*bchrom*): toma valores en el intervalo 1-10.
9. Nucléolos normales(*nnuc*): toma valores en el intervalo 1-10.
10. Mitosis(*mit*): toma valores en el intervalo 1-10.
11. Clase(*clase*): es la variable a predecir. Toma el valor 0 para cáncer benigno y valor 1 para cáncer maligno.

3.1.2 Conjunto de datos 2

La información contenida en este conjunto de datos se corresponde con los resultados obtenidos en el análisis en el cual se ha realizado una **mamografía** a una masa en pacientes con cáncer de mama[11].

Los datos de este conjunto muestran una discriminación de masas mamográficas benignas y malignas en función de los atributos BI-RADS y la edad de la paciente.

Estos se han distribuido en 6 columnas, las cuales contienen:

1. Evaluación BI-RADS (*birads*): toma valores en el intervalo 1-5.
2. Edad (*age*): recoge la edad de la paciente en años.
3. Forma (*shape*): representa la forma de la masa. Toma el valor 1 si es redonda, el 2 si es oval, el 3 si es lobular y el 4 si es irregular.
4. Margen (*margin*): representa el margen de la masa. Toma el valor 1 si es circunscrito, el 2 si es microlobulado, el 3 si es oculto, el 4 si es mal definido y el 5 si es espiculado.
5. Densidad (*density*): representa la densidad de la masa. Toma el valor 1 si es alta, el 2 si es media, el 3 si es baja y el 4 si contiene grasa.

6. Gravedad (*severity*): es la variable a predecir. Toma el valor 0 si es benigno y el valor 1 si es maligno.

3.1.3 Conjunto de datos 3

Este conjunto de datos contiene información de pacientes de cáncer de mama, proporcionados por el Centro Médico Universitario, Instituto de Oncología, Ljubljana, Yugoslavia. Los doctores M. Zwitter y M. Soklic proporcionaron los datos [12]. En él, se recoge información realizada en un **examen histológico** a las pacientes.

Este conjunto de datos está formado por 10 columnas, las cuales son:

1. Clase (*class*): es la variable e predecir. Toma los valores de eventos-recurrentes y eventos-no-recurrentes, es decir, quiere predecir si la paciente ha sufrido una recaída o no.
2. Edad (*age*): recoge la edad de la paciente en el momento del diagnóstico. La muestra en intervalos, los cuales son 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.
3. Menopausia (*menopause*): recoge la información de si la paciente es pre o postmenopáusica en el momento del diagnóstico Los valores que toma son lt40, ge40, premeno.
4. Tamaño del tumor (*tumorsize*): recoge la información del tamaño del tumor extirpado en milímetros en intervalos, los cuales son 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.
5. Ganglios linfáticos (*invnodes*): recoge el número de ganglios linfáticos axilares que contienen cáncer de mama metastásico visible en el examen histológico. Se muestra en los intervalos 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.

6. Cápsulas ganglionares (*nodecaps*): si el cáncer hace metástasis en un ganglio linfático, aunque esté fuera del lugar original del tumor, puede permanecer contenido por la cápsula del ganglio. Toma los valores de sí o no.
7. Grado de malignidad (*degmalig*): muestra el grado histológico del tumor. Los tumores de grado 1 están formados predominantemente por células que, aunque son neoplásicas, conservan muchas de sus características habituales. Los tumores de grado 3 están formados predominantemente por células muy anormales. Toma valor en el intervalo 1-3.
8. Mama (*breast*): el cáncer de mama puede producirse obviamente en cualquiera de las dos mamas. Esta columna recoge la información de en qué mama se encuentra. Es por ello que toma el valor de izquierda o derecha.
9. Cuadrante mamario (*breastquad*): la mama puede dividirse en cuatro cuadrantes, utilizando el pezón como punto central. Esta columna recoge el cuadrante en el que se encuentra el tumor. Toma los valores izquierda-arriba, izquierda-abajo, derecha-arriba, derecha-abajo, central.
10. Irradiación (*irradiat*): la radioterapia es un tratamiento que utiliza rayos X de alta energía para destruir las células cancerosas. Esta columna recoge la información de si la paciente tuvo radioterapia o no. Es por ello que toma los valores de sí o no.

Esta descripción del conjunto de datos se ha tomado de [13].

3.2 Redes Neuronales Artificiales (RNA)

Las Redes Neuronales Artificiales (RNAs) emulan el comportamiento del cerebro humano, el cual está caracterizado por el aprendizaje mediante la experiencia, así como por extraer conocimiento genérico a partir de un conjunto de datos [14].

Las RNAs imitan esquemáticamente la estructura neuronal del cerebro, ya sea mediante simulaciones con programas de ordenador, mediante emulaciones, a través de estructuras de procesamiento con capacidad de cálculo paralelo o bien mediante construcción física de sistemas con arquitectura similar a la de una red neuronal [14].

El clasificador se basa en el mecanismo de activación de las neuronas como su base de funcionamiento: cada neurona es receptora de un conjunto de impulsos electrónicos que propaga en forma de impulso eléctrico hacia las neuronas con las que está conectado[15].

Estas redes se organizan en capas, las cuales son: capa de entrada, la cual está formada por unidades que representan campos de entrada; capas ocultas, que son las capas intermedias, y capa de salida, formada por la unidad de destino [15]. Cada capa contiene una o más neuronas, por lo que cuantas más capas tenga el modelo más neuronas artificiales tendrá el sistema, lo que conlleva que la efectividad a la hora de medir y proporcionar resultados sea mayor[16].

Los datos de entrada se presentan en la capa de entrada y los valores se propagan desde esas neuronas hasta todas las neuronas de la capa siguiente. Finalmente, la capa de salida es la que envía el resultado del modelo [17].

Es por ello que la red aprende examinando los datos individualmente, generando una predicción para cada uno de ellos y realizando ajustes a las ponderaciones cuando falla en la predicción. Dado que el proceso se repite en bucle, la red mejora sus predicciones hasta que alcanza el criterio de parada. Una vez que la red ha sido entrenada, se puede aplicar a futuros casos en los que el resultado no es conocido [17].

Las Redes Neuronales Artificiales son, esencialmente, un modelo de Aprendizaje Automático. No obstante, ha evolucionado mucho a lo largo de los años y se ha convertido en un sistema muy complejo, el cuál es la base principal del Aprendizaje Profundo (Deep Learning).

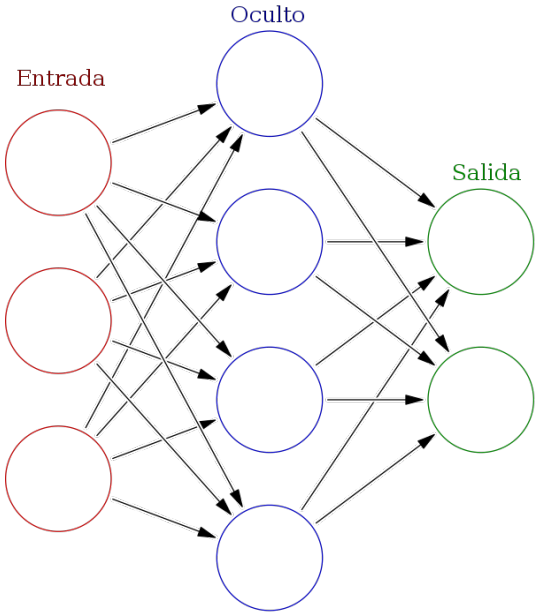

Figura 3. Estructura de las Redes Neuronales Artificiales. Imagen obtenida de Wikipedia.

3.3 K-Vecinos más cercanos (KNN)

Para entrar en contexto, en aprendizaje supervisado el algoritmo recibe un conjunto de datos etiquetados con los valores de salida sobre los que se puede entrenar y definir un modelo de predicción. Dicho algoritmo podrá ser utilizado posteriormente sobre nuevos datos para predecir los datos de salida [18].

El algoritmo de KNN es un clasificador de aprendizaje supervisado no paramétrico. Este algoritmo utiliza la proximidad para hacer clasificaciones y/o predicciones sobre la agrupación de un punto de datos individual. KNN se usa como un algoritmo de clasificación, suponiendo inicialmente que se pueden encontrar puntos similares cerca el uno del otro [19].

A diferencia de otros algoritmos de aprendizaje automático, KNN no incluye ninguna fase de aprendizaje previa, al igual que tampoco calcula el método predictivo como la regresión lineal. KNN lo que hace es encontrar similitudes entre las variables. Para ello,

miden la distancia que hay entre las variables almacenadas y las variables de entrada. A menos distancia, más similitud [20].

A la hora de implementar este algoritmo hay que prestar especial atención al cálculo de la distancia entre puntos y a la elección del número k:

3.3.1 Distancias

Las métricas de distancia ayudan a formar límites de decisión, los cuales dividen los puntos a consultar en diferentes regiones [19]. Para calcular estas medidas de distancia hay varios métodos, pero nos vamos a centrar en los más usados:

3.3.1.1 Distancia Euclidiana

La distancia Euclidiana indica la separación de dos puntos en un espacio donde se cumple el teorema de la geometría de Euclides [21].

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Fórmula de distancia Euclídea

3.3.1.2 Distancia Manhattan

La distancia Manhattan indica que la distancia entre dos puntos es la suma de las diferencias absolutas de sus coordenadas [22].

$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i|$$

Fórmula de distancia de Manhattan

3.3.2 Elección de K

Elegir un buen valor de K es muy importante en este algoritmo, ya que es el que va a determinar el número de valores seleccionados con los que se va a relacionar el dato a clasificar.

Inicialmente, se le suele fijar un valor de k=1, y de ahí se va incrementando el número hasta encontrar la solución más óptima.

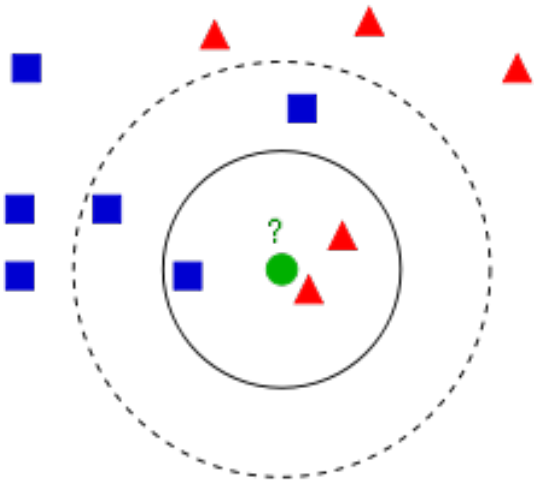

Figura 4. Ejemplo de clasificación con KNN. Imagen obtenida de Wikipedia.

3.4 Razonamiento Basado en Casos (CBR)

El Razonamiento Basado en Casos, como se ha comentado previamente en el capítulo de Introducción, es una metodología de Inteligencia Artificial en auge. Esto se debe a que ha tenido un gran éxito en muchos campos de aplicación.

El CBR es un método para resolver problemas recordando situaciones similares que han ocurrido con anterioridad, de la cual se va a reutilizar la información y el conocimiento sobre ese caso. Este proceso consta de cuatro pasos fundamentales:

- **Recuperar** el caso o casos con más similitud.

- **Reutilizar** la información y el conocimiento del caso seleccionado para resolver el problema.
- **Revisar** la solución propuesta.
- **Almacenar** las partes de esta experiencia que se consideran útiles para resolver futuros problemas.

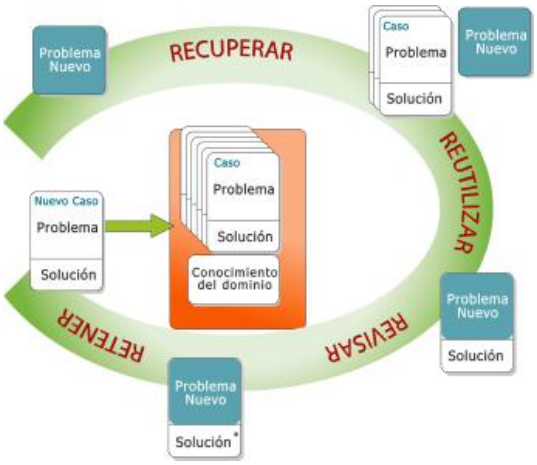

Figura 5. Ciclo CBR. Imagen obtenida de ResearchGate.

3.4.1 Recuperación de casos

La recuperación es la primera fase del ciclo CBR, ya que al llegar un caso nuevo lo primero que hay que hacer es recuperar uno o más casos similares de la base de casos, en la cual se encuentran almacenadas las experiencias previas en forma de casos.

Para la tarea de recuperación es necesario poseer un algoritmo de recuperación y una medida de similitud, los cuales serán usados para obtener el conjunto de datos similar. En este proyecto se ha elegido el algoritmo de KNN como algoritmo de recuperación y la distancia Euclídea como medida de similitud.

3.4.2 Reutilización de casos

La reutilización de casos es el segundo paso en el ciclo CBR. En ella, el algoritmo se centra en dos aspectos:

- Marcar las diferencias entre el caso pasado y el nuevo.

- Determinar qué partes del caso recuperado pueden ser transferidas al nuevo caso.

Normalmente, la tarea de reutilización se basa principalmente en copiar la solución pasada al nuevo caso, mientras que en otros casos la solución reutilizada no puede ser aplicada directamente, si no que tiene que adaptarse al caso. Hay dos tipos de adaptaciones:

- *Adaptación estructural:* se aplican reglas de adaptación a la solución recuperada.
- *Adaptación Derivacional:* se reutilizan las fórmulas que generaros la solución recuperada para construir una nueva solución para el problema actual.

Como se comentó en el apartado 3.3.1, el método de reutilización que se ha usado en este proyecto ha sido el de KNN. Es por ello que se ha podido elegir el número de vecinos a usar (número k) y adaptar la solución del nuevo caso mediante votación de la solución de todos los casos recuperados.

3.4.3 Revisión de casos

La revisión de casos es el tercer paso del ciclo CBR. En ella, la solución generada en la tarea de reutilización de casos es evaluada. Si el resultado es satisfactorio, el nuevo caso y la nueva solución del caso se almacenan en la base de casos (en el último paso del ciclo). Si no, habría que volver al paso 2 y recuperar otro caso pasado similar.

3.4.4 Almacenamiento de casos

El almacenamiento de casos es el cuarto y último paso del ciclo CBR. En él, el caso nuevo y su solución asociada son almacenados en la base de casos para un posible uso futuro.

Durante este proceso, el sistema tiene que seleccionar la información del caso a almacenar, la forma en la que se almacena dicha información y cómo indexar el caso en la estructura de la memoria para una posible recuperación.

Para la información relacionada con el algoritmo de Razonamiento Basado en Casos, se ha tomado como guía el documento pdf que contiene la referencia [23].

4

Desarrollo

En este capítulo se va a describir cómo se ha desarrollado cada algoritmo utilizado y para qué sirve, además de describir el preprocesamiento que se le ha aplicado a los datos.

4.1 Preprocesamiento de los datos

4.1.1 Mamografía

Este conjunto de datos no se le puede pasar directamente a las funciones que se han desarrollado, ya que hay algunas muestras incompletas. Además de esto, hay que asegurarse de que todas las variables sean de tipo numéricas. Para ello, se utiliza la función *as.numeric(conjuntodatos$variable)* [24] para cambiar la variable a tipo numérica.

Para ello, se ha utilizado la función *na.omit()* [25] para descartar todos los valores *NA*. Una vez realizado esto, los datos ya están listos para poder ser usados durante la implementación.

4.1.2 Citología

En este caso, además de eliminar los valores NA se ha eliminado los valores de la variable BIRADS ya que, como se comentó en la metodología, no es una variable predictiva.

Para el algoritmo KNN, además de esto, ha sido necesario aplicarle una normalización al conjunto de datos. Se ha hecho mediante la función *scale()* [26]. Son estos datos normalizados los que han sido pasados al método.

4.1.3 Examen histológico

Este conjunto de datos ha sido al que más preprocesamiento ha habido que aplicarle. Esto se debe a que todas sus variables eran de tipo categórico, por lo que se han tenido que pasar a tipo numérico.

Se ha realizado un *bucle for* que recorre todo el conjunto de datos y va cambiando los valores de categórico a numérico según las siguientes condiciones:

- Edad: los datos de esta variable se recogen en los siguientes intervalos: 20-29, 30-39, 40-49, 50-59, 60-69 y 70-79. Cada intervalo ha sido sustituido por los valores 0, 1, 2, 3, 4 y 5 respectivamente.
- Menopausia: los datos se recogen en las variables premeno, ge40 e it40. Cada valor ha sido sustituido por los valores 0, 1 y 2 respectivamente.
- Tamaño del tumor: los datos de esta variable se recogen en los siguientes intervalos: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49 y 50-54. Cada valor ha sido sustituido por los valores 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 y 10 respectivamente.
- Ganglios linfáticos: los datos de esta variable se recogen en los siguientes intervalos: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17 y 23-26. Cada valor ha sido sustituido por los valores 0, 1, 2, 3, 4, 5 y 6 respectivamente.

- Cápsulas ganglionares: los datos de esta variable son de sí o no, y se han sustituido por los valores 1 y 0 respectivamente.
- Radiación: los datos de esta variable son de sí o no, y se han sustituido por los valores 1 y 0 respectivamente.
- Mama: los datos de esta variable son de izquierda o derecha, y se han sustituido por los valores 1 y 0 respectivamente.
- Cuadrante mamario: los datos de esta variable pueden ser izquierda-arriba, izquierda-abajo, derecha-arriba, derecha abajo y central. Se han sustituido por los valores 3, 2, 1, 0 y 4 respectivamente.

4.2 Redes Neuronales Artificiales

Como se ha explicado en el capítulo de Metodología, este método divide los datos en un esquema 60-20-20: el 60% de los datos se entrena, un 20% se valida y el último 20% se prueba. Es por ello que para este algoritmo se han implementado tres funciones: entrenar, validar y probar. Los datos deben ser previamente preprocesados para su correcta aplicación.

4.2.2 Entrenar

En la función de entrenamiento se construye la red neuronal desde cero probando con todas las combinaciones posibles de selección de variables de entrada. La red ha sido construida con hasta 5 capas (de 1 a 5) utilizando la función *nnet()* [27] del paquete *nnet*, la cual estima un modelo de red neuronal artificial.

Esta función tiene como parámetro de entrada los datos de entrenamiento obtenidos de la partición, y devuelve el modelo construido y entrenado con dichos datos.

4.2.3 Validar

Esta función obtiene como parámetros de entrada el modelo construido durante el entrenamiento y los datos de validación obtenidos de la partición.

En ella, se le tiene que especificar al algoritmo que variable debe predecir en cada conjunto de datos, por lo que la función de validación es única para los diferentes conjuntos de datos.

En esta función, el algoritmo hace las predicciones de los datos que le llegan. Una vez hechas las predicciones, obtiene la precisión de estas y el área bajo la curva.

La precisión se obtiene calculando el número de datos predichos correctamente entre el número total de datos que hay.

$$Acc = \frac{(datos\$varpred == prediccion)}{\dim(datos)}$$

Fórmula para calcular la precisión

Siendo ***Acc*** la precisión, ***datos*** el conjunto de datos de entrada de la función, ***varpred*** la variable a predecir y ***prediccion*** la predicción hecha por el modelo.

Una vez obtenida la precisión (acc), se obtiene el modelo que mejor acc devuelve.

El área bajo la curva (auc) se obtiene con la función *roc()* [28] del paquete *pROC.*

Una vez obtenida la auc, se obtiene el modelo que mejor auc devuelve.

Es, por ello, que la función de validación tiene como parámetros de salida los dos mejores modelos construidos, el de la acc y el de la auc.

4.2.4 Probar

Por último, la función de probar lo que hace es, dándole como parámetros de entrada el modelo con mejor acc o el modelo con mejor auc y los datos de prueba obtenidos de la partición, obtener el mejor valor posible de acc y de auc.

La función devuelve, por tanto, el mejor valor obtenido para cada caso.

Una vez que todas las funciones han sido ejecutadas, el algoritmo devuelve el valor final obtenido de precisión y el del área bajo la curva.
Además, se muestra una gráfica representando dicha área, facilitando así su estudio visual.

4.3 KNN

Para este método también hay que hacer una partición previa de los datos. No obstante, el esquema elegido aquí ha sido el de 50-25-25: el 50% pertenece a los datos de entrenamiento, un 25% a los de validación y el último 25% a los de prueba.
Una vez hecha dicha partición, el algoritmo KNN se realiza con la función *knn()* [29] del paquete class.
A esta función se le pasan los datos de entrenamiento y los de prueba para las variables deseadas, y se le dice que variable debe predecir. Con lo que devuelve el algoritmo se construye la matriz de confusión, la cual muestra el número de pacientes que han sido bien clasificados y los que no.

4.4 CBR

El desarrollo del modelo CBR para este proyecto no será completo ya que no se necesita realizar el ciclo completo, si no obtener los valores de precisión con los que el modelo está clasificando.
Por ello, el modelo desarrollado se va a centrar en calcular la distancia entre los casos, así como el ángulo para poder representarlo gráficamente a través de coordenadas polares. Es, por ello, que se va a dividir el desarrollo en:

4.4.1 Calcular distancias

Se han implementado dos funciones:

- *calcular.distancia():* esta función calcula la distancia euclídea normalizada entre dos casos que se le pasan como parámetro de entrada. Tiene tres parámetros de entrada: el caso p, el caso q, y los datos, de los cuáles se obtendrá el valor máximo y el mínimo de la variable deseada. Esta función va a devolver la distancia en valor numérico.
- *calcular.distancias():* esta función construye una matriz con las distancias entre todos los patrones de los datos. Devuelve la matriz construida.

4.4.2 Calcular ángulos

Para poder calcular los ángulos, se ha implementado la función principal *calcular.angulos(),* la cual va a recurrir a las funciones secundarias *volar()*, *caminar()* y *coste()*.

- *calcular.angulos():* esta función implementa el algoritmo AFB para calcular los ángulos de las coordenadas polares.
- *volar():* esta función inicializa los ángulos con valores aleatorios.
- *caminar():* esta función asigna el valor de un ángulo cercano.
- *coste():* esta función calcula el coste mediante la expresión de la pág. 5 del artículo [7].

4.4.3 Generar diagrama

Una vez calculadas las distancias y los ángulos, se generarán los diagramas. Para ello, hay que pasar las coordenadas polares a cuadráticas.

Se han implementado las siguientes funciones:

- *polares.a.cuad():* esta función cambia las coordenadas polares a coordenadas cuadrangulares.

- generar.diagrama(): esta función genera el diagrama con las coordenadas cuadrangulares. Dependiendo de si el paciente presenta o no el fenómeno a predecir, se pintará de rojo o verde. En este caso, los pacientes cuyo cáncer es benigno o no han sufrido recaída se representan de color rojo y los cuales el cáncer es maligno o sí han sufrido recaída se representan de color verde.

4.4.4 Establecer clase

Por último, se establece la clase a la que pertenece el caso nuevo tomando como referencia los 5 patrones más cercanos.

5

Resultados y discusión

En este capítulo se van a mostrar los resultados obtenidos con cada algoritmo para cada conjunto de datos.

5.1 Redes Neuronales Artificiales

5.1.1 Mamografía

Los resultados obtenidos para el conjunto de datos referente a la mamografía son:

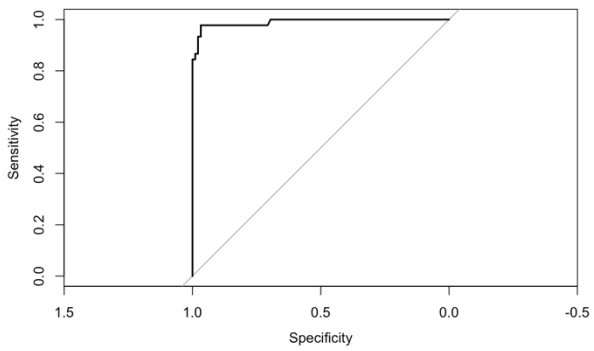

Figura 6. Representación AUC datos mamografía en RNA.

Acc	Auc
0,9474	0,9781

Tabla 1. Resultados ACC y AUC para mamografía en RNA.

Tanto el valor obtenido en la precisión como el obtenido en el comprendido en el área bajo la curva son mayores que 0.94, por lo que se puede afirmar que el algoritmo implementado está clasificando a los pacientes de manera muy eficaz y precisa.
Esto se ve reflejado en la figura 6, la cual muestra una curva ROC casi perfecta,
Este algoritmo sería confiable a la hora de ayudar a la toma de decisiones, siempre contando con la última palabra del profesional clínico.

5.1.2 Citología

Los resultados obtenidos para el conjunto de datos referente a la citología son:

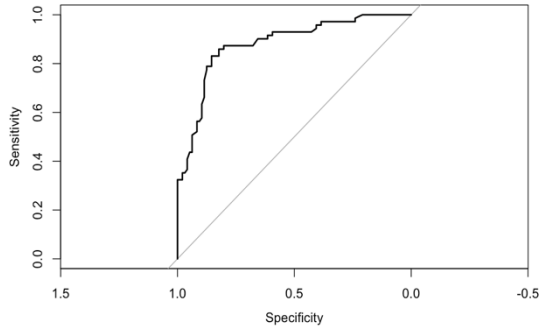

Figura 7. Representación AUC datos citología en RNA.

Acc	Auc
0,7910	0,8624

Tabla 2. Resultados ACC y AUC para citología en RNA.

El valor obtenido para la precisión es de 0.79, por lo que el algoritmo tiene una precisión buena pero no tan confiable como la anterior. No obstante, tiene un valor del área bajo la curva de 0.86, lo que nos dice que está clasificando bien al 86% de los pacientes. Esto se ve reflejado en la figura 7.
En este caso en concreto, el algoritmo podría ser válido para la ayuda al diagnóstico.

5.1.3 Examen histológico

Los resultados obtenidos para el conjunto de datos referente al examen histológico son:

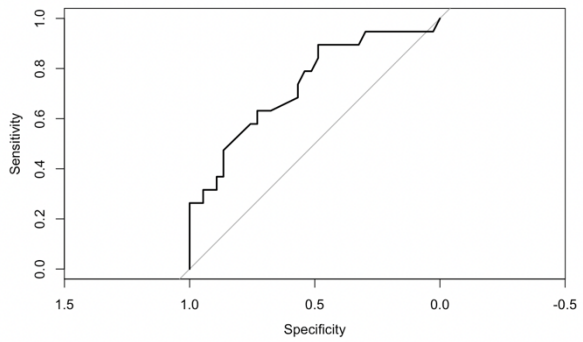

Figura 8. Representación AUC datos examen histológico en RNA.

Acc	Auc
0,7125	0,5896

Tabla 3. Resultados ACC y AUC para examen histológico en RNA.

El valor obtenido para el área bajo la curva es muy bajo, siendo este de 0.59. Esto se ve reflejado en la figura 8. Este resultado lo que nos quiere decir es que el algoritmo no está clasificando bien casi al 50% de los pacientes.

Tras hacer algunos cambios y algunas consultas, se ha llegado a la conclusión de que esto se puede deber a la forma en la que se han transformado los datos de su valor categórico a su valor numérico, por lo que se ha planteado utilizar RandomForest para ello.

5.2 KNN

5.2.1 Mamografía

Los resultados obtenidos en la matriz de confusión para el conjunto de datos referente a la mamografía son:

	Datos predichos → 0, benigno; 1, maligno.		
Datos originales → 0, benigno; 1, maligno.		**0**	**1**
	0	110	1
	1	4	55

Tabla 4. Matriz de confusión para mamografía.

Los valores obtenidos de Acc y Auc son:

Acc	Auc
0,9706	0,9616

Tabla 5. Resultados ACC y AUC para mamografía en KNN.

Además, se va a representar gráficamente el área bajo la curva:

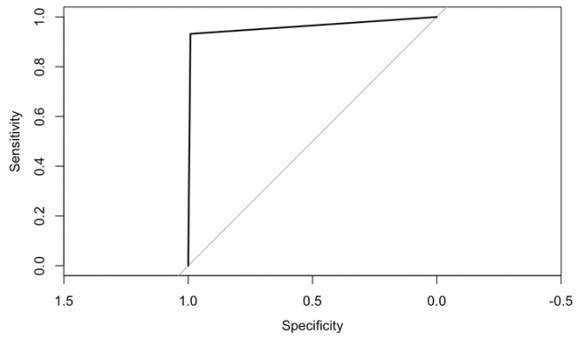

Figura 9. Representación AUC datos mamografía en KNN.

Se puede observar claramente que el algoritmo está obteniendo muy buenos resultados. Tenía que clasificar a 111 pacientes cuyo tumor esa benigno y ha clasificado correctamente a 110 de ellos. Solamente ha clasificado de forma errónea a 1.

Por otro lado, eran 59 los pacientes cuyos tumores eran malignos, de los cuales ha clasificado correctamente a 55.

Es por ello que se puede afirmar que este algoritmo es un buen clasificador y se podría tomar como base de apoyo para la toma del diagnóstico.

5.2.2 Citología

Los resultados obtenidos en la matriz de confusión para el conjunto de datos referente a la citología son:

	Datos predichos → 0, benigno; 1, maligno.		
Datos originales → 0, benigno; 1, maligno.		**0**	**1**
	0	88	18
	1	26	74

Tabla 6. Matriz de confusión para citología.

Los valores obtenidos de Acc y Auc son:

Acc	Auc
0,7864	0,7851

Tabla 7. Resultados ACC y AUC para citología en KNN.

Además, se va a representar gráficamente el área bajo la curva:

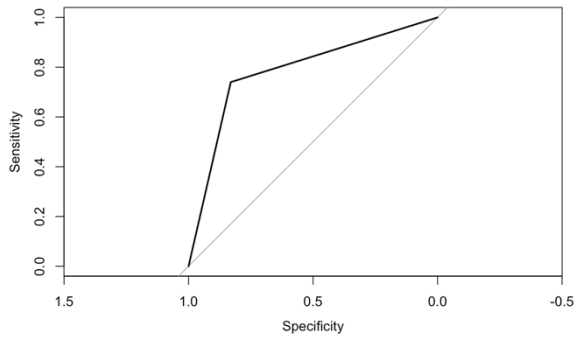

Figura 10. Representación AUC datos citología en KNN.

Había que clasificar a 106 pacientes cuyo cáncer era benigno y a 100 pacientes cuyo cáncer era maligno.

De los 106 se ha clasificado correctamente a 88 y de los 100 a 74, por lo que el algoritmo tiene una buena clasificación aunque se podría mejorar.

No obstante, los profesionales clínicos también lo podrían tomar de ayuda.

5.2.3 Examen Histológico

Los resultados obtenidos en la matriz de confusión para el conjunto de datos referente al examen histológico son:

	Datos predichos → 0, no recaída; 1, recaída.		
Datos originales → 0, no recaída; 1, recaída.		**0**	**1**
	0	44	7
	1	11	7

Tabla 8. Matriz de confusión para examen histológico.

Los valores obtenidos de Acc y Auc son:

Acc	Auc
0,7391	0,6258

Tabla 9. Resultados ACC y AUC para examen histológico en KNN.

Además, se va a representar gráficamente el área bajo la curva:

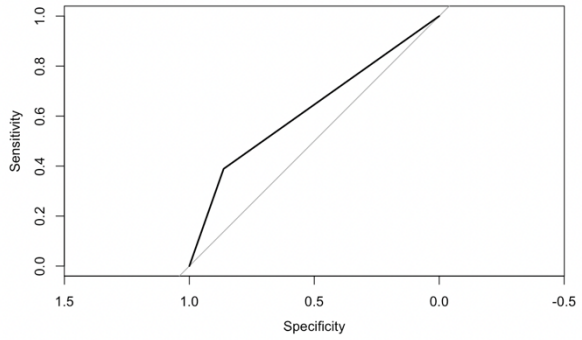

Figura 11. Representación AUC datos examen histológico en KNN.

Había que clasificar a 51 pacientes que no habían sufrido recaída y a 18 que sí. El algoritmo ha clasificado de manera correcta a 44 de los primeros y a 7 de los segundos, por lo que se podría decir que su clasificación no es mala, pero no sería recomendable que el profesional clínico en este caso tomara en cuenta para su clasificación final la opinión del algoritmo.

5.3 CBR

5.3.1 Mamografía

Los resultados obtenidos para el conjunto de datos referente a la mamografía son:

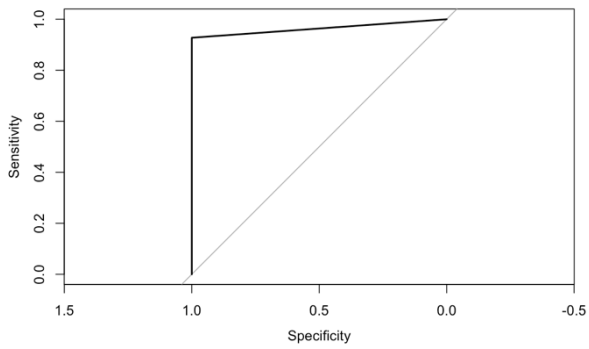

Figura 12. Representación AUC datos mamografía en CBR.

Acc	Auc
0,9708	0,9636

Tabla 10. Resultados AUC y ACC para mamografía en CBR.

Tal y como se observa, los valores obtenidos tanto para acc como auc son muy buenos. Esto se ve reflejado en la curva ROC, la cual es bastante buena también.

Estos resultados lo que reflejan es que el modelo está realizando buenas predicciones y que estas podrían servir de ayuda para la toma de decisiones en el diagnóstico de las pacientes.

Además de esto, como se ha comentado en el capítulo de desarrollo, se genera un diagrama de dispersión en el que se representa el caso nuevo frente a los casos conocidos. Se han generado de dos formas, usando el algoritmo AFB para calcular los ángulos y usando la selección aleatoria de los ángulos:

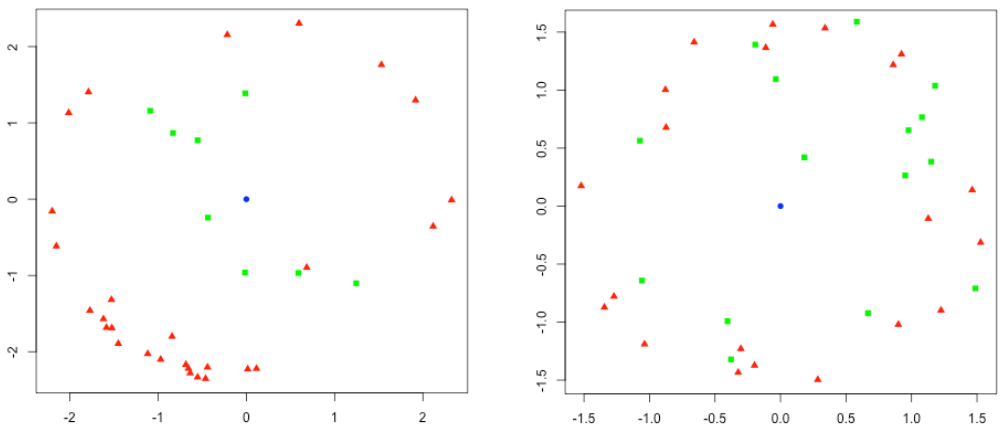

Figura 13. Diagramas generados con AFB para la mamografía.

Como se puede observar en ambas gráficas, la mayor parte de los 5 casos más cercanos se corresponden a pacientes cuyo cáncer es maligno, por lo que el personal clínico al ver estos diagramas podrían clasificar al paciente nuevo como un paciente con cáncer maligno.

5.3.2 Citología

Los resultados obtenidos para el conjunto de datos referente a la citología son:

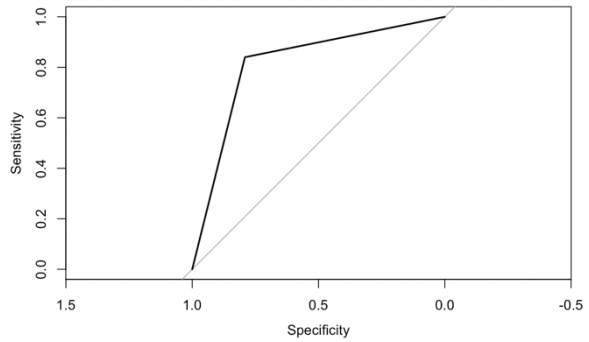

Figura 14. Representación AUC datos citología en CBR.

Acc	Auc
0,8132	0,8156

Tabla 11. Resultados AUC y ACC para citología en CBR.

Los valores obtenidos de Acc y Auc son valores buenos y fiables. De igual forma, el profesional clínico podría tomar como base la clasificación del modelo aunque sería decisión suya el diagnóstico final.

Los diagramas generados para este conjunto de datos son:

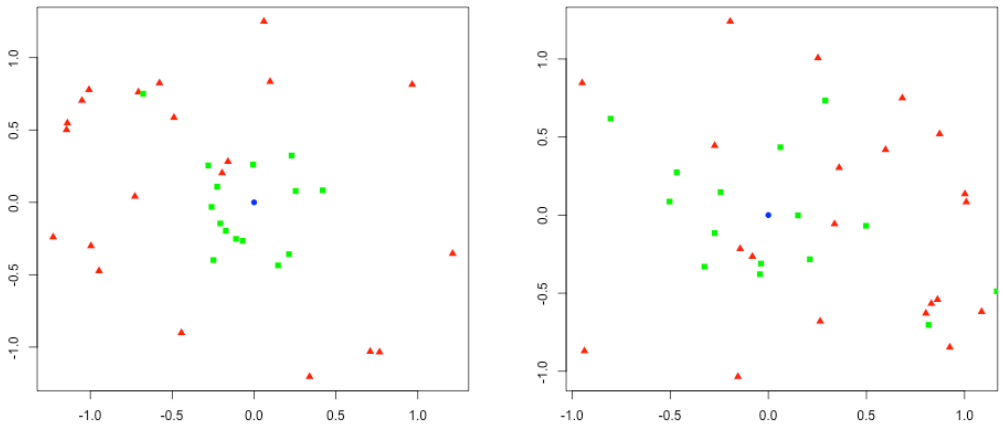

Figura 15. Diagramas generados con AFB para la citología.

Observando ambos diagramas, entre los 5 casos más cercanos vuelve a predominar los pacientes con cáncer maligno, por lo que la clasificación de este paciente por parte del personal clínico sería el de paciente con cáncer maligno.

5.3.3 Examen histológico

Los resultados obtenidos para el conjunto de datos referente al examen histológico son:

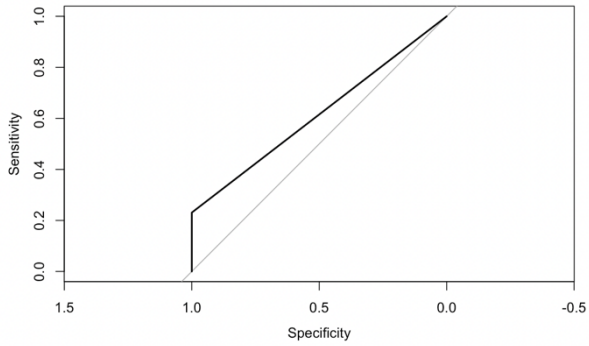

Figura 16. Representación AUC datos examen histológico en CBR.

Acc	Auc
0,8214	0,6154

Tabla 12. Resultados AUC y ACC para examen histológico en CBR.

En este caso, aunque el valor de precisión es mayor que 0.8, el área bajo la curva es de 0.61, lo que nos llevaría a inclinarnos por no tener muy en cuenta las predicciones realizadas por el modelo a la hora de diagnosticar a los pacientes.

Los diagramas generados para este conjunto de datos son:

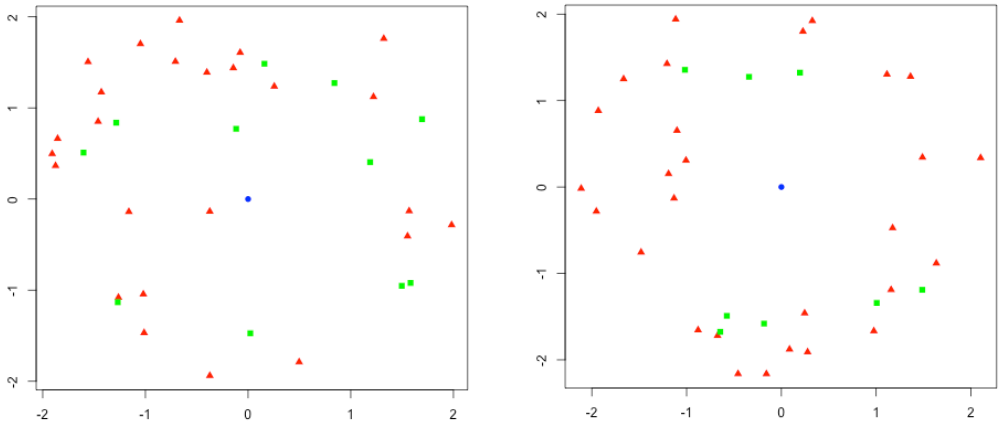

Figura 17. Diagramas generados con AFB para el examen histológico.

En este caso, al analizar los diagramas se obtiene que entre los 5 casos más cercanos predominan los pacientes que no han sufrido una recaída en el cáncer al realizarle el examen histológico, por lo que el personal clínico clasificaría el caso nuevo como paciente que no tiene cáncer.

6

Conclusiones y líneas futuras

Tras haber realizado un estudio de la metodología propuesta y sus resultados obtenidos, se ha llegado a la conclusión de que el modelo CBR permite obtener predicciones estadísticamente similares a los algoritmos tradicionales propuestos, siendo, además, una alternativa de bajo coste computacional. Esto se debe a que es un algoritmo sencillo el cual se basa en la reducción de dimensionalidad de los patrones, lo que evita tener que entrenar modelos más costosos computacionalmente.

Además de esto, el modelo CBR tiene la ventaja de que con él se obtiene una representación visual de las clasificaciones realizadas, lo que permite que el personal clínico, el cual no suele tener un amplio conocimiento sobre las métricas obtenidas, sea capaz de entender la clasificación que le está mostrando el diagrama y así obtener un diagnóstico final basándose en lo observado.

Tras varías ejecuciones se ha observado que, a pesar de que las métricas obtenidas con los 3 modelos propuestos son muy similares, siempre se obtiene una mayor precisión para cada conjunto de datos con el modelo CBR.

Es, por todo esto, que se concluye que el modelo CBR es una opción igual de válida que los algoritmos tradicionales, teniendo incluso algunas ventajas su utilización.

No obstante, las relaciones obtenidas entre el nuevo patrón y los patrones incluidos en la base de datos es, en su totalidad, de carácter estadístico, lo que quiere decir que sigue sin establecerse una relación causal entre ellos. Es, por ello, que una de las líneas futuras que se proponen es la de completar el modelo CBR para incorporarlo a un sistema en funcionamiento, como puede ser un módulo en Galén [31], un sistema de información hospitalaria en funcionamiento de los servicios de oncología de los diferentes hospitales de Málaga, y añadir información adicional que establezca relaciones casuales con el conocimiento específico de la medicina.

Durante la realización de este proyecto también ha surgido la necesidad de transformar un conjunto de datos compuesto por variables categóricas a variables numéricas. En concreto, ese conjunto de datos es en el que peor resultados se han obtenido, por lo que otra de las posibles líneas futuras de este proyecto es la de usar *Random Forest* para dicha transformación y ver si mejoran así los resultados obtenidos.

Por último, se propone mejorar el modelo KNN desarrollando un algoritmo que optimice los resultados obtenidos probando con distintos valores de k y devolviendo la mejor solución.

Referencias bibliográficas

[1] "The R project for statistical computing" *R*. [Online]. Available: https://www.r-project.org/

[2] "Inteligencia artificial: Qué es y por qué importa" *SAS*. [Online]. Available: https://www.sas.com/es_es/insights/analytics/what-is-artificial-intelligence.html

[3] "Aprendizaje Automático: Qué es y por qué importa" *Qué es y por qué importa | SAS ES*. [Online]. Available: https://www.sas.com/es_es/insights/analytics/machine-learning.html

[4] "¿Qué es deep learning?" *SAS*. [Online]. Available: https://www.sas.com/es_es/insights/analytics/deep-learning.html

[5] E. I. A. Oracle, "Diferencias entre la inteligencia artificial y el machine learning" *Medium*, 14-Sep-2018. [Online]. Available: https://medium.com/@experiencIA18/diferencias-entre-la-inteligencia-artificial-y-el-machine-learning-f0448c503cd4

[6] L. Lozano and J. Fernández, "Razonamiento Basado en Casos: Una Visión general" *Universidad de Valladolid*. [Online]. Available: https://www.infor.uva.es/~calonso/IAI/TrabajoAlumnos/Razonamiento%20basado%20en%20casos.pdf

[7] J.-B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, and B. Séroussi, "Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach" *Artificial Intelligence in Medicine*, 14-Jan-2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365718304846

[8] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine" *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. [Online]. Available: https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.1312

[9] *UCI Machine Learning Repository*. [Online]. Available: https://archive.ics.uci.edu/ml/index.php

[10] *UCI Machine Learning Repository: Breast Cancer wisconsin (original) data set*. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)

[11] *UCI Machine Learning Repository: Mammographic mass data set*. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/mammographic+mass

[12] *UCI Machine Learning Repository: Breast Cancer Data Set.* [Online]. Available: https://archive.ics.uci.edu/ml/datasets/breast+cancer

[13] N. Alva, "Using machine learning techniques to predict the recurrence of breast cancer" *LinkedIn*, 06-Feb-2021. [Online]. Available: https://www.linkedin.com/pulse/using-machine-learning-techniques-predict-recurrence-breast-alva/

[14] López Raquel Flórez and M. Fernández Fernández José, Las Redes neuronales artificiales: Fundamentos Teóricos y aplicaciones prácticas. Oleiros, La Coruña: Netbiblo, 2008. Pags 10-11 https://books.google.es/books?hl=es&lr=&id=X0uLwi1Ap4QC&oi=fnd&pg=PA11&dq=redes+neuronales+artificiales&ots=gOMAfrrq3j&sig=eXSlebMm-fPeUxB-azqv2P16BMI#v=onepage&q&f=false

[15] M. Merino, "Machine Learning para la predicción de interacciones entre microARN y ARN mensajeros" *Univesitat Oberta de Catalunya*, 2018. [Online]. Available: http://openaccess.uoc.edu/webapps/o2/bitstream/10609/73785/3/manmermonTFM0118memoria.pdf

[16] "Redes Neuronales y machine learning" *Tokio School*, 30-Aug-2021. [Online]. Available: https://www.tokioschool.com/noticias/redes-neuronales-machine-mearning/

[17] "El modelo de redes neuronales" *IBM*, Aug-2021. [Online]. Available: https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=networks-neural-model

[18] B. L and T. B, "¿Qué es el algoritmo knn?" *Formación en ciencia de datos | DataScientest.com*, 04-Mar-2022. [Online]. Available: https://datascientest.com/es/que-es-el-algoritmo-knn

[19] "¿Qué es el algoritmo de k Vecinos Más cercanos?," *IBM*, 2022. [Online]. Available: https://www.ibm.com/mx-es/topics/knn

[20] D. S. Team, "K Vecinos Más Cercano" *DATA SCIENCE*, 30-Nov-2020. [Online]. Available: https://datascience.eu/es/programacion/k-vecinos-mas-cercanos-un-poderoso-algoritmo-de-aprendizaje-automatico-con-implementacion-en-python-r/

[21] A. Cajal, "Distancia Euclidiana: Concepto, Fórmula, cálculo, ejemplo" *Lifeder*, 04-Dec-2019. [Online]. Available: https://www.lifeder.com/distancia-euclidiana/

[22] Eloviparo, "La Distancia manhattan O La Distancia Euclidea" *Todos no somos mamiferos*, 13-Mar-2018. [Online]. Available: https://eloviparo.wordpress.com/2018/03/13/la-distancia-manhattan-o-la-distancia-euclidea/

[23] A. B. Bregón, A. S. Hurtado, C. J. A. González, J. B. P. Junquera, Q. I. M. Sancho, and J. J. R. Diez, "Un sistema de razonamiento basado en casos para la clasificación de fallos en Sistemas Dinámicos" *Dialnet*, 01-Jan-1970. [Online]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=7975170

[24] "Numeric: Numeric vectors" *RDocumentation*, 2022. [Online]. Available: https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/numeric

[25] "Na.omit: Handle missing values in objects" *RDocumentation*, 2022. [Online]. Available: https://www.rdocumentation.org/packages/photobiology/versions/0.10.10/topics/na.omit

[26] "Scale: Scaling and centering of matrix-like objects" *RDocumentation*, 2022. [Online]. Available: https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/scale

[27] "Nnet: Fit Neural Networks," *RDocumentation*, 2022. [Online]. Available: https://www.rdocumentation.org/packages/nnet/versions/7.3-17/topics/nnet

[28] "ROC: Build a ROC curve," *RDocumentation*, 2022. [Online]. Available: https://www.rdocumentation.org/packages/pROC/versions/1.18.0/topics/roc

[29] "KNN: K-nearest neighbour classification," *RDocumentation*, 2022. [Online]. Available: https://www.rdocumentation.org/packages/DMwR/versions/0.4.1/topics/kNN

[30] J.-B. Lamy, "Artificial Feeding Birds (AFB): A new metaheuristic inspired by the behavior of pigeons," *HAL Open Science*, 06-Aug-2019. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02264232

[31] N. Ribelles, J. M. Jerez, D. Urda, J. L. Subirats, A. Márquez, C. Quero, E. Torres, L. Franco, and E. Alba, "Galén: Sistema de Información para la Gestión y coordinación de procesos en un servicio de oncología," *RevistaeSalud.com*, 2010. [Online]. Available: https://investigacion.ubu.es/documentos/5f9929242999525e73a05cf6

UNIVERSIDAD DE MÁLAGA | uma.es

E.T.S de Ingeniería Informática
Bulevar Louis Pasteur, 35
Campus de Teatinos
29071 Málaga

E.T.S. DE INGENIERÍA INFORMÁTICA