Escola Tècnica Superior
d'Enginyeria Agronòmica i del Medi Natural

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

TRABAJO FINAL DE GRADO

Universitat Politécnica de València – Escola Tècnica Superior

d'Enginyeria Agronòmica i del Medi Natural

Grado en Biotecnología

Curso académico: 2018/2019

Métodos de *machine learning* en estudios biomédicos

Alumno: Arturo González Vilanova

Tutor: Francisco García García

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Valencia, CONVOCATORIA SEPTIEMBRE 2019

Machine learning methods in biomedical studies

Abstract:

The development of high-throughtput technologies in molecular biology and medical imaging has allowed access to large amounts of information of various types, known as big data. This information is so complex that it is very difficult to draw reliable and useful conclusions from it. It requires the use of multivariate statistical methods and a great deal of computing power to glimpse the patterns, models or standards that the data follow. In this context, machine learning is born, a discipline that fuses statistical methods with computing to develop algorithms capable of classifying samples, predicting results and making inferences based on the information previously provided as training. These methods applied to biomedicine can extract the sense of data from genomics, transcriptomics, medical imaging, among others, which would allow the advancement of medicine to a more personalized, accurate and effective form of medical care.

In this work, three of the most popular machine learning models are applied in the context of classification: k-nearest neighbours, support vector machines and random forest. The data used come from the extraction of radiomic features from medical imaging and the extraction of morphological features from cell nuclei.

The objective is to evaluate the performance of these models on potentially relevant information in the clinic. First, an exploratory data analysis was made, consisting of the principal components analysis and clustering analysis. The main body of work consists of six steps: data processing, standardization, data partitioning, feature selection, training and validation. The processing consisted in the elimination of all those samples and variables that for some reason were not suitable for inclusion in subsequent analyses. The data were then transformed by centering and scaling. The data were divided into two subsets, one of which was used for training and the other for validation. During the selection of characteristics, the number of variables to be taken into account for the models was further reduced to only those more relevant. Models were trained and predictions were made about observations that were not used in the training. With the results obtained from the predictions, precision metrics were calculated and analyzed.

The results obtained reveal that the quality and abundance of the data is fundamental for the development of a good predictive model. Different models can be perfectly functional for the same classification problem. Analyses show a clear relationship between some of the characteristics and the clinical outcome.

Keywords: machine learning, radiomics, medical imaging, diagnosis, predictive model.

Métodos de *machine learning* en estudios biomédicos

Resumen:

El desarrollo de las tecnologías de alto rendimiento en biología molecular e imagen médica ha permitido el acceso a grandes cantidades de información de diverso tipo, lo que ahora se conoce como *big data*. Dicha información es de tal complejidad que resulta muy difícil el poder extraer conclusiones fiables y útiles de ella. Se requiere la utilización de métodos de estadística multivariante y un gran poder de computación para vislumbrar los patrones, modelos o normas que siguen los datos. En este contexto nace el *machine learning* o aprendizaje automático, una disciplina que fusiona métodos estadísticos con informática para elaborar algoritmos capaces de clasificar muestras, predecir resultados y realizar inferencias en base a la información que se les proporciona previamente como entrenamiento. Estos métodos aplicados a la biomedicina pueden extraer el sentido de datos de genómica, transcriptómica, imagen médica, entre otros, lo cual permitiría el avance de la medicina a una forma más personalizada, precisa y efectiva de atención médica.

En este trabajo se aplican tres de los modelos de aprendizaje automático más populares en el contexto de la clasificación: *k*-vecinos más próximos, máquinas de soporte vectorial y bosques aleatorios. Los datos utilizados provienen de la extracción de características radiómicas de imagen médica y la extracción de características morfológicas de núcleos celulares.

El objetivo es evaluar el desempeño de estos modelos sobre información potencialmente relevante en la clínica. En primer lugar, se hizo un análisis exploratorio de los datos consistente en el análisis de componentes principales y análisis de agrupamiento. El cuerpo principal del trabajo consta de seis pasos: procesado de los datos, estandarización, partición de los datos, selección de características, entrenamiento y validación. El procesado consistió en la eliminación de todas aquellas muestras y variables que por algún motivo no eran adecuadas para su inclusión en análisis posteriores. A continuación, se transformaron los datos por centrado y escalado. Los datos se dividieron en dos subconjuntos, de los cuales uno sirvió para el entrenamiento y otro para la validación. Durante la selección de características se redujo todavía más el número de variables a tener en cuenta para los modelos hasta tener solo aquellas más relevantes. Se entrenaron los modelos y se realizaron predicciones sobre las observaciones que no se usaron en el entrenamiento. Con los resultados obtenidos de las predicciones, se calcularon y analizaron métricas de precisión.

Los resultados obtenidos revelan que la calidad y abundancia de los datos es fundamental para el desarrollo de un buen modelo predictivo. Diferentes modelos pueden ser perfectamente funcionales para un mismo problema de clasificación. Los análisis demuestran una clara relación entre algunas de las características y el resultado clínico.

Palabras clave: aprendizaje automático, radiómica, imagen médica, diagnóstico, modelo predictivo.

Alumno/a: Arturo González Vilanova

Tutor académico: D. Francisco García García

Valencia, septiembre de 2019

Agradecimientos

La realización de este proyecto ha salido adelante gracias al apoyo de algunas personas a las que quiero agradecer los esfuerzos desinteresados que han hecho por mí a lo largo de estos meses.

Gracias a Francisco García García, director de la Unidad de Bioinformática y Bioestadística del Centro de Investigación Príncipe Felipe, por atender a mis peticiones siempre que le ha sido posible y aconsejarme en cada paso del trayecto. A Adolfo López Cerdán, Javier García Dasí y Sergio Romera Giner por solucionar mis dudas. A todos mis compañeros del grado de biotecnología, en especial a Fabián y Adrià, por estar ahí y reanimarme cuando lo he necesitado (casi siempre). Por último, muchas gracias a Azahara, por hacer más llevadero mi día a día y apoyarme en mis propósitos

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ABREVIATURAS

BD2K: *Big Data to Knowledge*

BISTI: *Biomedical Information Science and Technology Initiative*

CV: *Cross-validation*

EBI: *European Bioinformatics Institute*

ED: *Peritumoral Edema*

ET: *Enhancing Tumor*

GDC: *Genome Data Consortium*

H1CA: *Hierarchical Clustering Analysis*

KNN: *k-Nearest Neighbours*

LGG: *Lower Grade Glioma*

MRI: *Magnetic Resonance Imaging*

NET: *Non-enhancing Tumor*

NIH: *National Institutes of Health*

PCA: *Principal Component Analysis*

RF: *Random Forest*

RFE: *Recursive Feature Elimination*

SVM: *Support Vector Machines*

TCGA: *The Cancer Genome Atlas*

WT: *Whole Tumor*

PPV: *Positive predictive value*

NPV: *Negative predictive value*

1. INTRODUCCIÓN

1.1. TECNOLOGÍAS DE ALTO RENDIMIENTO Y *BIG DATA* EN LAS CIENCIAS DE LA SALUD

De alto rendimiento o *high-throughput* es toda aquella tecnología cuyos sistemas procesan información a gran velocidad, normalmente por la automatización y paralelización del proceso. Como ejemplo, algunas de estas tecnologías son técnicas de secuenciación masiva, *microarrays*, identificación y cuantificación de proteínas y lípidos por espectrometría de masas e imagen médica. Debido al rápido desarrollo y reducción del precio de las herramientas y tecnologías biomédicas de alto rendimiento, los investigadores en las ciencias de la vida y la asistencia sanitaria están produciendo y analizando información biológica compleja en constante crecimiento, con lo cual estas áreas están cada vez más cercanas al *big data*. Incluso en laboratorios pequeños donde no se tenga acceso a la instrumentación necesaria, es posible trabajar con tal tamaño de información, accediendo a ella a través de repositorios públicos como el del European Bioinformatics Institute (EBI, https://www.ebi.ac.uk/) o The Cancer Genome Atlas (TCGA) (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga).

1.1.1. LAS DISCIPLINAS ÓMICAS Y LAS IMÁGENES COMO FUENTE DE INFORMACIÓN BIOLÓGICA

Como resultado de la generalización del uso de tecnologías de alto rendimiento, nacen las disciplinas ómicas, con el fin de medir y estudiar sistemas biológicos complejos y sus interacciones. La genómica, transcriptómica, proteómica, interactómica, metabolómica, fenómica, farmacogenómica y radiómica son algunas de ellas.

La imagen médica es una tecnología importante utilizada en la práctica clínica para asistir en la toma de decisiones. Sin embargo, su potencial es mucho mayor. La imagen tridimensional de un paciente puede contener millones de vóxeles (unidad cúbica que compone un objeto tridimensional), y un tumor (u otra entidad anormal) puede tener cientos de características medibles que describen el tamaño, forma y textura. Teniendo esto en consideración, el concepto de "ómica" puede aplicarse fácilmente al análisis cuantitativo de imagen (Gillies et al., 2016). La radiómica es la extracción de las características de las estructuras fisiológicas presentes en imagen médica, convertidas en datos numéricos de elevada dimensionalidad.

El nacimiento de la radiómica está motivado por la idea de que las imágenes biomédicas de tomografía axial computarizada, resonancia magnética y tomografía de emisión de positrones contienen información que refleja la patofisiología subyacente y que estas relaciones pueden revelarse por análisis de imagen cuantitativos. La cuantificación de características de imagen por parte de expertos radiólogos puede ser subjetiva y sensible a variaciones. Por otra parte, los biomarcadores de imagen obtenidos a través de procesos radiómicos representan y cuantifican las alteraciones presentes de una forma más objetiva y estandarizable que el criterio de una persona. Esta información cuantitativa podría asociarse a parámetros de interés para los clínicos, como el riesgo, supervivencia y gravedad de la condición (Gillies et al., 2016). En la figura 1 se describen esquemáticamente los pasos necesarios para convertir imágenes en datos numéricos.

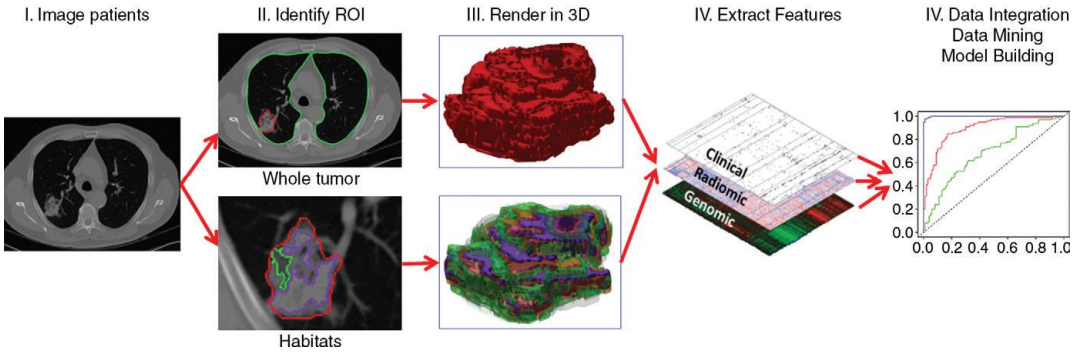

Figura 1: *Flujo de trabajo típico de un proyecto de radiómica. Extraído de Gillies et al., (2016).*

Las características de imagen cuantitativas basadas en intensidad, forma, tamaño y textura ofrecen información sobre el fenotipo del tumor y su microambiente que es distinta a la proporcionada por los informes clínicos, resultados de laboratorio y ensayos genómicos o proteómicos. Estas características pueden integrarse con información de otra naturaleza para correlacionarse con resultados clínicos y con ello, extraer conclusiones que ayuden en la toma de decisiones (Gillies et al., 2016).

Hay investigaciones que ya han demostrado la capacidad de análisis radiómicos para distinguir tejido prostático benigno de cáncer de próstata y además añadir información sobre la agresividad del cáncer (Wibmer et al., 2015). En la evaluación del cáncer de pulmón y en la evaluación de glioblastoma multiforme, la radiómica ha demostrado ser una herramienta con la que valorar el pronóstico del paciente (Coroller et al., 2015).

1.1.2. LOS INCONVENIENTES DE TRABAJAR CON *BIG DATA*

Un solo genoma humano secuenciado tiene un tamaño de 140 gigabytes (Marx, 2013). El EBI tiene una capacidad de almacenamiento de más de 160 petabytes de datos en copias de genes, proteínas y moléculas pequeñas, un número que sigue incrementándose (Cook et al., 2018). El manejo y mantenimiento de este volumen de información es costoso en tiempo y recursos computacionales, así que los investigadores deben estar seguros de que su infraestructura computacional y herramientas informáticas están a la altura de este reto.

Otra desventaja de los datos generados por estas técnicas es que, con frecuencia, las matrices tienen un elevado número de variables, pero poco tamaño muestral. Las características que se pueden medir en una muestra o sujeto suelen ser del orden de cientos o miles, pero la cantidad de muestras o sujetos se ve más limitada por el tamaño del estudio. Este fenómeno se conoce como la "maldición de la dimensión" o "efecto Hughes". Conforme aumenta la dimensionalidad el volumen del espacio de predictores aumenta exponencialmente y con ello aumenta la dispersión de los datos en dicho espacio, lo que hace más difícil obtener resultados estadísticamente significativos (James et al., 2017). Además, analizar todos los datos disponibles a través de análisis estadísticos directos (por ejemplo, test estadísticos entre controles y diferentes condiciones experimentales) es imposible. Una persona no puede realizar, y mucho menos interpretar, los resultados de todos los test posibles.

Sin embargo, el desarrollo de métodos estadísticos orientados a este tipo de problemas y las mejoras tecnológicas en computación permiten el mejor y más rápido manejo de todos estos datos.

El National Institutes of Health (NIH) está desarrollando proyectos como Biomedical Information Science and Technology Initiative (BISTI, https://bisti.nih.gov/) Big Data to Knowledge (BD2K,

https://commonfund.nih.gov/bd2k), iniciativas centradas en apoyar la investigación y el desarrollo de abordajes innovadores y herramientas que maximicen y aceleren la integración del *big data* y la ciencia de datos en la investigación biomédica.

En este contexto, es necesaria la cooperación entre clínicos, investigadores del laboratorio y bioinformáticos para superar los retos que les plantea el *big data*.

1.2. MACHINE LEARNING

Como se ha descrito en el apartado anterior, la cantidad y dimensión de datos que se manejan en las ciencias ómicas hace que la identificación de relaciones y patrones sea difícil o directamente imposible para las personas. Debido a ello, resulta imperativo recurrir a computadoras para la aplicación de cálculos matemáticos complejos que permitan visualizar y entender lo que los datos reflejan de manera que se puedan extraer conclusiones fehacientes.

El *machine learning*, en castellano aprendizaje automático, es una disciplina dentro del campo de la inteligencia artificial centrada en el entrenamiento de modelos matemáticos a través de datos con el fin de que algoritmos que incorporen estos modelos sean capaces de realizar tareas inteligentes, generalmente basadas en el reconocimiento de patrones multiparamétricos (Deo, 2015). Este enfoque hacia los datos es lo que distingue al aprendizaje automático de otras ramas de la inteligencia artificial, donde la conducta de los sistemas informáticos está completamente programada, es decir, restringida al código.

Actualmente se aplica en áreas como el reconocimiento del habla, traducción, navegación automática de vehículos, recomendación de productos y reconocimiento de imagen. Se emplea con dos finalidades: crear sistemas automáticos a los que se puedan relegar tareas repetitivas que normalmente harían las personas y descubrir nuevos patrones imperceptibles para los humanos por la complejidad de los datos.

Los algoritmos de *machine learning* pueden dividirse en supervisados y no supervisados dependiendo de las preguntas a las que responden y el tipo de información disponible para su entrenamiento.

1.2.1. APRENDIZAJE NO SUPERVISADO

En el aprendizaje no supervisado se dispone únicamente de un conjunto variables de entrada. No se dispone de etiquetas preexistentes, es decir, no se tiene un conocimiento *a priori* de la clasificación de los datos, y es por ello que se le llama “no supervisado”. El objetivo es encontrar los patrones que subyacen en los datos y los agrupan de la mejor manera posible. Dos algoritmos son los más representativos:

- *Clustering*: se desea descubrir los grupos naturalmente presentes en los datos. El algoritmo más popular es el *k-means.*
- Asociación: se desea descubrir las reglas que describen los datos. El algoritmo más popular es el *apriori*.

Frecuentemente, para evaluar el desempeño de un método de este tipo se combina con trabajos de aprendizaje supervisado posteriores para comprobar si esos nuevos patrones son útiles de alguna forma.

1.2.2. APRENDIZAJE SUPERVISADO

En el aprendizaje supervisado, en el que se centra este trabajo, se dispone de variables de entrada (predictores, *input*, *X*) y variable de salida (*output*, *Y*) cuyos posibles valores son conocidos. El algoritmo se utiliza para conocer el valor de *Y* en base a *X*. Se centra en dos tipos de problema:

- Clasificación. Se predice la clase a la que pertenece cada observación, aquella que mejor describa los valores de sus predictores, construyendo reglas de decisión. La respuesta es, por lo tanto, una variable cualitativa.

- Problemas de regresión, en los que se estima el valor de un parámetro desconocido perteneciente a un espectro continuo en función del resto de los datos.

Hay métodos diseñados para atender específicamente a uno de estos dos problemas, pero también existen otros que pueden utilizarse en ambos casos. Algunos ejemplos populares de aprendizaje supervisado son regresión lineal (James et al., 2017), bosques aleatorios (Breiman, 2001) y máquinas de soporte vectorial (Cortes and Vapnik, 1995).

La figura 2 representa gráficamente la separación que hacen de las muestras los dos tipos de algoritmos en el espacio de predictores. En el caso del aprendizaje supervisado, un modelo lineal divide el espacio bidimensional en dos subespacios, de forma que los casos que se encuentran a un lado son controles, mientras que al lado contrario se sitúan los casos. Además, cada muestra está etiquetada con su clase correcta. En el caso del aprendizaje no supervisado se conocen los grupos que pueden estar representados en los datos, pero el algoritmo es capaz de rodear dos subespacios, cada uno perteneciente a un grupo de muestras que se parecen entre sí y a su vez se distinguen del otro grupo.

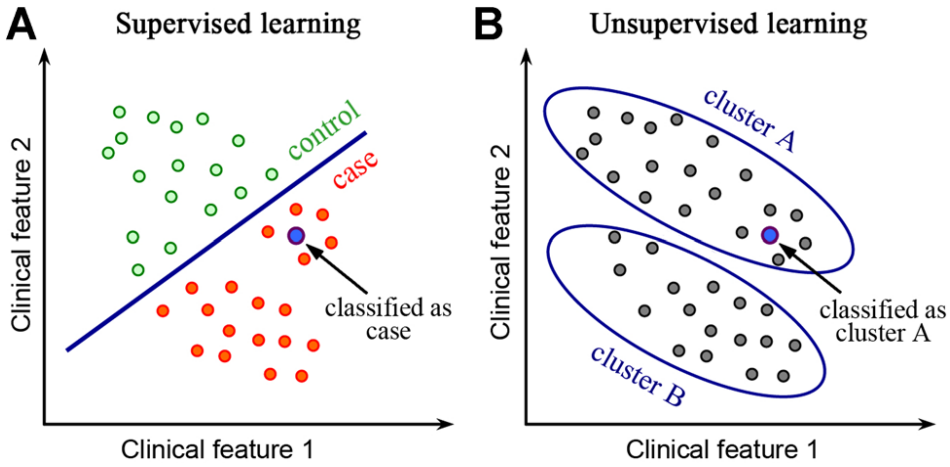

Figura 2: *La diferencia entre aprendizaje supervisado y no supervisado. Extraído de Greene et al. (2014).*

Para la elaboración de cualquier modelo de aprendizaje supervisado son necesarios los siguientes elementos:

- Algoritmo: Los pasos tomados para crear el modelo a partir de las características de los ejemplos de entrenamiento. Los lenguajes de programación más populares para hacer aprendizaje automático son Python y R. En ambos existen paquetes informáticos que agrupan programas y funciones relacionadas con cada paso del proceso de elaboración de un modelo.

Esto hace más simple y accesible el desarrollo de algoritmos a medida de los problemas del usuario.

- Datos etiquetados: Información etiquetada con la clase o valor correcto que sirve como fuente de aprendizaje para el algoritmo. Que la información esté etiquetada es lo que concede al método la característica de "supervisado". Normalmente estos datos se encuentran en forma de tablas, donde las filas son casos, muestras o repeticiones, y las columnas son medidas de variables aleatorias, también conocidas como predictores o características. Antes de crear el modelo será necesario separar los datos en dos grupos:
    - Conjunto de aprendizaje o entrenamiento: son los datos utilizados durante la fase de aprendizaje, en la que el algoritmo recibe tanto la información como su etiqueta y con ello ajusta los pesos o puntos de decisión para optimizar la precisión de las predicciones. En este conjunto se encuentran la mayor parte de los datos.
        - Parte de este conjunto puede separarse para utilizarse en una tarea aparte, la validación, cuya finalidad es afinar algunos parámetros de ajuste específicos para cada modelo y trabajo. Estos parámetros son definidos manualmente por el investigador, normalmente tras probar con diferentes combinaciones y rangos de valores.
    - Conjunto de prueba: otro conjunto de datos que el algoritmo no ha visto nunca y sirve para estimar la capacidad de predicción del modelo generado en el aprendizaje. Las predicciones generadas por el modelo sobre el conjunto de aprendizaje (los datos que ya ha visto) son peligrosamente optimistas, porque el modelo puede haber aprendido características únicas de este conjunto y con ello, su poder predictor no es generalizable a casos que el algoritmo no ha visto antes. Este ajuste a un conjunto de datos concreto es lo que se conoce como *overfitting*. Con el fin de evitarlo, se hacen predicciones con el modelo sobre el conjunto de prueba, para hacer una estimación lo más realista posible de la precisión del modelo.

1.2.3. EL PROCESO DE APRENDIZAJE SUPERVISADO

Habitualmente los datos crudos son sometidos a varios métodos de preprocesamiento, lo que se conoce como *feature engineering*.

- Generar variables *dummies* (valor 1 o 0) a partir de las variables categóricas (si las hubiera).
- Imputación estadística, que consiste en eliminar o sustituir los valores que no estén disponibles.
- Escalado por métodos como normalización, estandarización o *min-max scaling*.
- Reducción de la dimensionalidad. Pueden hacerse muy diversos filtrados a los datos, como por ejemplo eliminar variables con baja varianza (no discriminatorias entre grupos), muy correlacionadas entre sí (redundantes) y poco correlacionadas con la respuesta. En cualquier caso, se desea reducir la dimensionalidad de los conjuntos de datos para hacer que los modelos sean lo más generalizables e interpretables posible. Normalmente este paso es el más laborioso de todo el proceso.

La optimización de los hiperparámetros es otro paso en el que el investigador debe tomar decisiones que afectarán al rendimiento del modelo. En un modelo de *k*-vecinos más próximos (KNN) debe decidirse el valor de k, el número de vecinos que se tendrá en cuenta para calcular las probabilidades de pertenencia a cada clase en un punto determinado del espacio de predictores. La práctica más habitual consiste en probar diferentes combinaciones de hiperparámetros sobre el conjunto de aprendizaje en combinación con métodos de remuestreo como *bootstrapping* y validación cruzada (CV), eligiendo como valores definitivos aquellos que hayan optimizado una métrica del desempeño del modelo, como la precisión, el área bajo la curva o el error cuadrático medio.

En la figura 3 se puede observar que cuanto menor es k, más se ajusta el modelo a los datos utilizados para el entrenamiento. A mayor k, menor flexibilidad del modelo, pero a su vez mayor capacidad para generalizar.

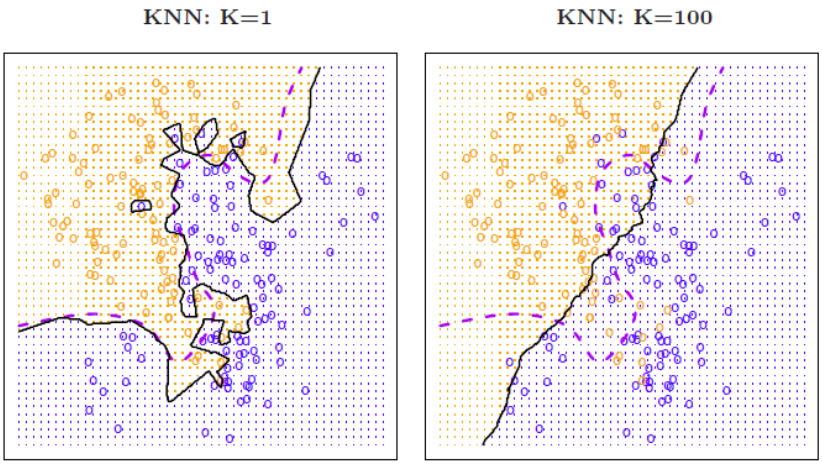

Figura 3: *Efecto de k sobre el umbral de decisión del modelo KNN. Extraído de James et al. (2017).*

Una vez escogidos los hiperparámetros, se entrena el modelo con esos ajustes, utilizando todo el conjunto de aprendizaje.

Con el modelo formado, sólo queda probarlo sobre el conjunto de prueba y con ello estimar su sensibilidad y especificidad. Si los valores obtenidos son aceptables, ya puede utilizarse para predecir en casos nuevos. En caso contrario, los pasos descritos anteriormente pueden modificarse para así obtener un modelo más generalizable. En algunos casos la calidad de los datos utilizados no permite obtener modelos útiles, por más que se refine el resto del proceso.

En la figura 4, un diagrama de flujo describe el proceso de creación de un modelo de aprendizaje automático para su uso posterior en tareas de predicción.

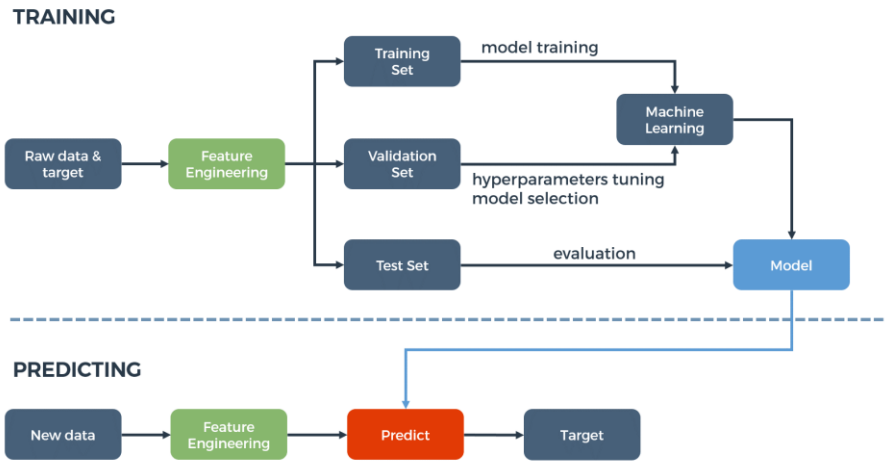

Figura 4: *Proceso habitual de aprendizaje automático. Extraído de Jouganous et al. (2019).*

1.3. MEDICINA PERSONALIZADA Y DE PRECISIÓN

A lo largo de las últimas décadas se ha llegado a conocer que una buena parte de la variabilidad en la respuesta a tratamientos farmacológicos está determinada por factores genéticos, sin ignorar la gran contribución de la edad, sexo, estado de salud, nutrición e interacción con otras terapias. En el cáncer, por ejemplo, es sabido que la mayoría de los tumores clínicamente relevantes son altamente heterogéneos a nivel fenotípico, fisiológico y genómico, y además evolucionan constantemente, adquiriendo resistencias que dificultan el tratamiento. Estos dos factores (la variabilidad en la respuesta a fármacos y la heterogeneidad de las enfermedades) suponen barreras para mejorar la calidad de vida del paciente y prolongarla lo máximo posible (Gillies et al., 2016). Como respuesta a este problema nace la voluntad de ejercer una medicina más personalizada y precisa, dedicada a cada paciente y su situación.

Se define la medicina de precisión como aquellos tratamientos dirigidos a las necesidades de pacientes individuales en base a sus características genéticas, fenotípicas, psicosociales o biomarcadores, que le distinguen a unos pacientes de otros con cuadros clínicos similares. El objetivo de este tipo de medicina es mejorar el resultado clínico para cada paciente, minimizando los riesgos y efectos secundarios innecesarios.

Ya se están realizando esfuerzos para conseguir que la medicina sea cada vez más personalizada. Por ejemplo, en cáncer de pulmón la clasificación tradicional se basaba en criterios histológicos y anatómicos, pero ahora se ha ampliado y también utiliza biomarcadores genéticos de mutaciones en EGFR, MET, RAS y ALK (Kwak et al., 2010). Cada uno de estos grupos responde al tratamiento de un fármaco concreto. Administrar el fármaco equivocado para el tipo de cáncer que padece el paciente no sólo no generará respuesta positiva, sino que además puede provocar efectos secundarios indeseables.

Las tecnologías de alto rendimiento están impulsando una mejor caracterización y clasificación de las enfermedades, lo que permite tanto realizar un mejor diagnóstico del paciente como conocer el grado de beneficio que puede obtener de las terapias disponibles (Jameson and Longo, 2015). La imagen médica no siempre se ha considerado parte de la medicina de precisión, pero ha cambiado profundamente los protocolos de diagnóstico, ya que ahora muchos pueden hacerse con confianza en base a evidencia de imagen, lo cual libra a los pacientes de otros métodos de diagnóstico más invasivos como la cirugía o biopsia sólida (Jameson and Longo, 2015). Los registros de salud electrónicos

contienen información clínica que podrían utilizar algoritmos diseñados para identificar personas con factores de riesgo para enfermedades como diabetes o elevado colesterol en sangre. En resumen, está disponible una gran cantidad de herramientas para la adquisición de información relevante para la clínica, pero la complejidad de esta información requiere que los sistemas de salud tengan a su disposición herramientas informáticas, no para reemplazar el juicio de los profesionales, sino para hacer evidentes los hechos que se ocultan en la información.

Se necesitan mejores biomarcadores para ayudar con la detección de enfermedades y guiar los tratamientos, especialmente para enfermedades comunes. Los esfuerzos por encontrar biomarcadores para concusión (Siman et al., 2013), test de imagen para detectar Alzheimer (Rinne et al., 2010), y biomarcadores circulantes para tumores (Dawson et al., 2013) ejemplifican la necesidad de estas herramientas de diagnóstico.

En este contexto el aprendizaje automático puede ejercer dos papeles: predecir la respuesta de futuras observaciones, y entender la relación entre respuesta y predictores (James et al., 2017).

El primero de estos papeles, la predicción, se basa en la existencia de una variable Y que no puede obtenerse fácilmente, pero puede conocerse si se dispone del conjunto de datos X que la define. Por ejemplo, podría pensarse que $X_1, \ldots, X_p$ son características de la muestra de sangre de un paciente que pueden medirse fácilmente en un laboratorio, mientras que Y es la variable que codifica el efecto que tendrá el fármaco sobre el paciente. Predecir Y usando X permitiría evitar administrar fármacos que no tuvieran efectos positivos sobre el paciente, acelerando el tratamiento por el uso de los tratamientos adecuados y disminuyendo los efectos secundarios.

El segundo papel es la inferencia. En este caso el objetivo no es crear un modelo predictor para el diagnóstico o prevención de enfermedades, sino entender de qué forma están relacionadas X e Y. En este caso debería conocerse exactamente la forma del modelo $f(X)$. Tratar el problema desde este punto de vista nos puede llevar a conocer qué predictores están más asociados con la respuesta, cómo se relacionan los predictores con la respuesta (positivamente, negativamente o dependiente de interacciones con otros predictores), y si la relación es lineal o más compleja. Siguiendo esta metodología podrían encontrarse nuevos biomarcadores para una enfermedad que permitan no sólo un mejor diagnóstico, sino también un mejor entendimiento de sus mecanismos patofisiológicos.

2. OBJETIVOS

El objetivo principal de este trabajo es la aplicación y evaluación del desempeño de tres modelos de aprendizaje automático supervisado utilizando datos de origen biomédico en problemas de clasificación. Estos tres modelos son *k*-vecinos más próximos (KNN), bosques aleatorios (RF) y máquinas de soporte vectorial (SVM).

Se utilizarán dos conjuntos de datos para el entrenamiento de estos modelos. El primero de ellos está formado por medidas volumétricas de gliomas de grado bajo. El segundo está formado por características morfológicas extraídas de imágenes de núcleos de células de cáncer de mama.

Para ello se generarán modelos con diferentes combinaciones de hiperparámetros y procesos de selección de características. Se estudiará el efecto del preprocesado de los datos, número de características e hiperparámetros para cada caso.

3. MATERIALES Y MÉTODOS

3.1. GRUPOS DE DATOS

A continuación, se describen los dos conjuntos de datos utilizados para la aplicación de los métodos de aprendizaje automático.

3.1.1. CARACTERÍSTICAS RADIÓMICAS DE GLIOMAS DE GRADO BAJO

Este conjunto se obtuvo del trabajo realizado en López (2019), en el que se extrajeron las características volumétricas de imágenes tridimensionales de resonancia magnética (MRI), cada una correspondiente a un paciente. Las imágenes provienen de una colección pública del proyecto The Cancer Genome Atlas (CANCER GENOME ATLAS RESEARCH NETWORK et al., 2015).

Se dispone de datos moleculares, clínicos y de imagen médica referentes a pacientes con gliomas de grado bajo antes de ser operados. Los datos clínicos están contenidos en el repositorio Genomic Data Commons (GDC, https://portal.gdc.cancer.gov/), mientras que los datos de imagen médica se encuentran en The Cancer Imaging Archive (TCIA, https://www.cancerimagingarchive.net/).

De los 516 pacientes disponibles, sólo 199 disponen de imágenes de resonancia magnética (MRI). Cada paciente tiene un indicador propio en el que se describe la institución de procedencia y funciona como vínculo entre los dos repositorios descritos.

Tabla 1: *Resumen de los datos de MRI pertenecientes a la colección TCGA-LGG. Extraído y modificado de Bakas et al. (2017).*

Colección	n	Instituciones que contribuyen (n)	TCGA ID
TCGA-LGG	199	St Joseph Hospital/MedicalCenter, Phoenix, AZ (98)	TCGA-HT
		Henry Ford Hospital, Detroit,MI (57)	TCGA-DU
		Case Western ReserveUniversity, Cleveland, OH (22)	TCGA-FG
		Thomas Jefferson University, Philadelphia, PA (20)	TCGA-CS
		University of North Carolina,Chapel Hill, NC (2)	TCGA-EZ

De los 199 pacientes, se trabajó con las imágenes de 108, que eran los que tenían disponibles imágenes de las modalidades de MRI T1, T1c, T2 y T2-FLAIR de los gliomas previas al tratamiento de la enfermedad. En la figura 5 se muestran imágenes 2D de cada una de estas modalidades para un paciente. En cada modalidad se resaltan tejidos distintos, con lo cual todas aportan información complementaria y relevante. T1 y T1c hacen que las zonas más grasas tengan una mayor intensidad y las zonas más ricas en agua sean más oscuras. La diferencia entre las dos es el uso o no del gadolinio como agente de contraste. T2 fomenta el efecto contrario, de forma que las áreas ricas en agua

se ven con mayor intensidad. T2-FLAIR presenta un contraste mayor entre tejidos y lesiones por eliminación de la señal procedente del líquido cefalorraquídeo.

Todas las imágenes están procesadas para normalizar con respecto a la orientación de las cabezas, el tamaño de los vóxeles y la intensidad de señal.

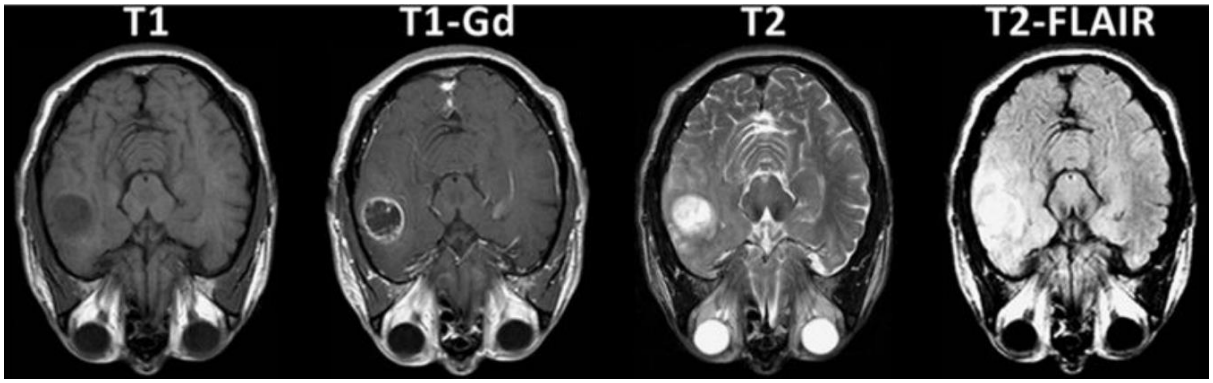

Figura 5: *Corte axial del cráneo de un paciente de glioma bajo las cuatro modalidades de MRI. Extraído y modificado de Bakas et al. (2017).*

La segmentación realizada en López (2019) delimita tres regiones tumorales, ilustradas en la figura 6: la parte realzada en imagen del tumor sólido o *enhancing tumor* (ET), correspondiente a la región de rotura de la barrera hematoencefálica; la parte no realzada del tumor sólido o *non-enhancing tumor* (NET) y el edema consecuencia de la rotura de la barrera hematoencefálica o *peritumoral edema* (ED). La región formada por todo el tumor se denomina *whole tumor* (WT).

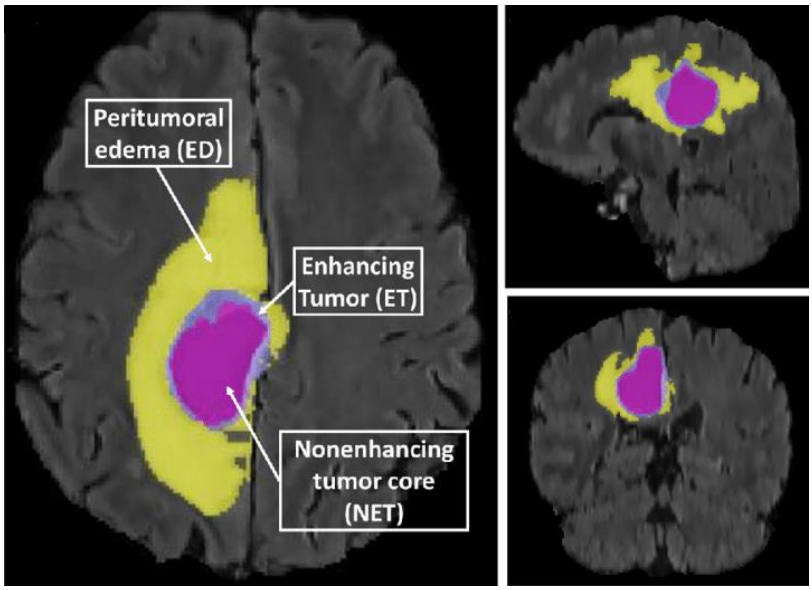

Figura 6: *Segmentación del glioma y sus regiones de interés. Extraído y modificado de Ratore et al. (2018).*

Además, en López (2019) también se aplican varios filtros distintos a las imágenes antes de extraer las características radiómicas lo cual incrementa en gran medida el volumen de información extraída de una sola imagen. En total se dispone de 6316 características. La tabla 2 describe el número de características para cada región y modalidad de MRI.

Tabla 2: *Subconjuntos de características radiómicas en función de la región del tumor y la modalidad de MRI empleada. Extraída de López (2019).*

Código	Región	Mod. MRI	Pacientes	Forma	Primer Orden	Textura
ED	ED	T2-FLAIR	108	13	216	828
ET	ET	T1c	88	13	216	828
NET	NET	T1c	106	13	216	828
T1c	WT	T1c	108	-	216	828
T2	WT	T2	108	-	216	828
FLAIR	WT	T2-FLAIR	108	-	216	828
WShape	WT	T2-FLAIR	108	13	-	-

Las características extraídas están predefinidas en la plataforma *PyRadiomics* (https://pyradiomics.readthedocs.io/en/latest/features.html).

Con respecto a los datos clínicos, se obtuvieron del trabajo de Ceccarelli et al. (2016), de la tabla de información suplementaria S1, donde se describen las características de sexo, edad, diagnóstico histológico, número de mutaciones, entre otros, todo ello asociado al identificador de TCGA. De las variables clínicas disponibles, dos de ellas se utilizaron para el siguiente trabajo:

- Grado de tumor, para el cual existen dos clases: grado 2 (G2) y grado 3 (G3). La división en grados es una organización propuesta por la Organización Mundial de la Salud (WHO). Los grados 2 y 3 muestran niveles intermedios de malignidad, a diferencia del grado 1, generalmente benignos, y el grado 4 (glioblastoma multiforme), que es el más agresivo de los tumores. El pronóstico del paciente está altamente asociado con el grado del tumor. Se cree que tumores de grado bajo pueden progresar a grados superiores por la adquisición de ciertas alteraciones genéticas (Mendelsohn et al., 2015).

- Diagnóstico histológico, para el cual existen tres clases:

    - Astrocitomas. Son los más frecuentes. Se les llama así porque muestran características morfológicas y bioquímicas que indican diferenciación astroglial.

    - Oligodendrogliomas. Son los segundos más frecuentes. Ocurren con más frecuencia en la materia blanca. Pueden tener un amplio espectro de diferenciación. Su origen celular no es del todo comprendido, aunque se cree que provienen de células precursoras de oligodendrocitos. Histopatológicamente, estos tumores muestran células redondas, compactas, con un núcleo denso y citoplasma claro, se dice que tienen apariencia de

"huevo frito". Se distinguen de los astrocitomas sobre todo por sus perfiles moleculares y citogenéticos.

- Oligoastrocitomas. Se caracterizan por presentar características histopatológicas y moleculares de los dos anteriores.

3.1.1. CARACTERÍSTICAS MORFOLÓGICAS DE NÚCLEOS DE CÉLULAS DE CÁNCER DE MAMA

Este conjunto de datos se obtuvo del trabajo realizado en Street et al. (1993), descargado del UC Irvin Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php). Consta de 569 muestras de pacientes con cáncer de mama. La variable clínica de interés sobre la que se trabaja es la malignidad o no del tumor.

Las características de este conjunto de datos se computan a partir de una imagen digitalizada de una punción aspiración con aguja fina (PAAF) de masa tumoral mamaria. En cada imagen se presentan varias células, cuyos núcleos han sido delimitados a través de un procedimiento que combina el trazo manual de expertos con un procesado posterior de dicho trazo. A partir de estas imágenes, se describen estas características del núcleo:

1. Radio: media de las distancias desde el centro del núcleo hacia puntos de su perímetro.
2. Textura: desviación estándar de los valores en escala de gris de las imágenes.
3. Perímetro.
4. Área.
5. Suavidad: variación local de la longitud de los radios.
6. Compactibilidad: perímetro^2/área - 1.0
7. Concavidad: gravedad de las porciones cóncavas del contorno.
8. Puntos cóncavos: número de porciones cóncavas del contorno.
9. Simetría.
10. Dimensión fractal.

En la matriz de datos se encuentran la media, el error estándar y el "peor" (media de los tres valores más altos) para cada una de estas características, haciendo un total de 30 predictores.

3.2. HERRAMIENTAS Y RECURSOS INFORMÁTICOS

3.2.1. R Y RSTUDIO

R es un lenguaje de programación para la computación estadística y gráfica (https://www.r-project.org/foundation/). Es usado ampliamente por estadísticos y científicos de datos para el desarrollo de software de análisis de datos. R y sus librerías implementar una amplia variedad de técnicas gráficas y estadísticas, incluyendo modelado, test estadísticos clásicos, clasificación, *clustering* y otros. La versión utilizada para la realización de este trabajo es la 3.6.1.

RStudio (https://www.rstudio.com/) es el entorno de desarrollo integrado diseñado para R. Consta del editor de código fuente y otras herramientas que facilitan trabajar con R, como pueden ser el autocompletado inteligente, un depurador y un buscador de objetos. La versión del programa utilizada es la 1.2.1335.

Estos dos recursos fueron utilizados para crear los programas para llevar a cabo tanto el análisis de los dos conjuntos de datos descritos anteriormente, como la generación de los modelos predictores.

3.2.2. CARET

El paquete informático caret, abreviación de *Classification And Regression Training* (https://topepo.github.io/caret/), creado por Max Kuhn, agrupa funciones relacionadas con operaciones en el ámbito del aprendizaje supervisado, con el fin de agilizar el proceso de confeccionar programas que se sirvan de estas operaciones. Así pues, dispone de herramientas para:

- Separación de los datos en subconjuntos.
- Preprocesado.
- Selección de características.
- Ajuste de hiperparámetros por remuestreo.
- Estimación de la importancia de las variables.

3.3. PREPROCESADO DE LOS DATOS

3.3.1. ELIMINACIÓN DE VALORES FALTANTES

Peculiaridades del conjunto de datos de gliomas de grado bajo hicieron necesaria la eliminación de algunos pacientes y variables del estudio. En dos casos no pudo realizarse la extracción para el conjunto NET, de forma que también se eliminaron de la matriz de datos. Debido a que sólo 88 pacientes presentaban la región ET previamente a la extracción de características, se descartaron para su uso las características de esta región. También se eliminaron todas aquellas variables que tenían valores faltantes para algún caso. Estos procedimientos redujeron las dimensiones de la matriz a 106 filas y 5214 columnas.

A continuación, se realizó el filtrado de las características clínicas para utilizar únicamente aquellos casos que todavía estaban reflejados en la matriz de características radiómicas, y se extrajeron las características de grado de tumor y diagnóstico histológico.

Uno de los pacientes presentaba discrepancia en cuanto al grado del tumor, de forma que no fue tenido en cuenta para análisis posteriores. De esta forma, quedaron 105 casos para su estudio.

El conjunto de datos de cáncer de mama tenía información completa, tanto la clínica como la morfológica de los núcleos, de forma que no fue necesaria la eliminación de ningún caso o variable.

3.3.2. ANÁLISIS EXPLORATORIO DE LOS DATOS

El análisis exploratorio constó del análisis de componentes principales (PCA) y análisis de grupos por *clustering* jerárquico (HCA). Con esto se pretendía conocer la distribución de los casos según la clase a la que pertenecen y detectar valores atípicos o *outliers*.

Como ya se ha descrito en la introducción, el *clustering* es una forma de aprendizaje no supervisado que puede utilizarse para descubrir los grupos presentes en los datos, aunque estos se desconozcan. En este caso, se utilizará como herramienta de análisis exploratorio. En primer lugar, se realiza un cálculo de distancias entre muestras. La ecuación de distancia puede tener diversas formas dependiendo de qué relaciones se quieran identificar. En este caso se utilizará la distancia euclídea como métrica. Una vez

calculadas las distancias, se agrupan las observaciones en *clusters*, se empieza con cada observación como un *cluster* propio e iterativamente se van juntado los *clusters* en *clusters* mayores hasta que todos los *clusters* convergen en uno solo que agrupa todas las observaciones. En este caso, el tipo de *clustering* utilizada es *complete-linkage*, donde la distancia viene determinada por la distancia máxima entre los grupos. El resultado final se muestra como un dendrograma donde la altura de las ramas equivale a la distancia que separa los dos nodos que se unen.

El PCA simplifica la complejidad de datos de elevada dimensionalidad manteniendo sus tendencias y patrones. Esto se consigue transformando los datos a dimensiones menores que funcionan como un resumen de las características. Para ello, se proyectan geométricamente los datos sobre dimensiones menores llamadas componentes principales. La primera componente principal (PC1) es la dirección que recoge mayor varianza en los datos, que es la misma que aquella que minimiza la distancia total de los puntos en el espacio de predictores y su proyección sobre la componente. Las siguientes componentes principales se seleccionan con el mismo criterio, pero una restricción adicional, y es que estas deben ser ortogonales entre sí, de forma que las componentes principales no están correlacionadas unas con otras. Teniendo en cuenta esto, se sabe que el máximo número de componentes principales es igual al número de observaciones o de variables, el menor de estos dos. Al final pueden utilizarse dos o tres componentes principales en un gráfico de dispersión que resuma la distribución de los datos.

3.3.3. FILTRADO DE PREDICTORES DE BAJA VARIANZA

En algunas situaciones, los mecanismos generadores de datos pueden crear predictores que tienen valores únicos (*zero-variance*) o muy baja varianza (*near-zero variance*). Debido a ello estos predictores no tienen capacidad discriminatoria entre las clases y utilizarlos en el modelo no tendría sentido.

Para detectar y eliminar estos predictores se calculan dos métricas:

- La frecuencia del valor más prevalente sobre la frecuencia del segundo valor más prevalente, llamado ratio de frecuencia. Debería ser cercano a 1 para predictores útiles y ser muy grande para datos desequilibrados.
- El porcentaje de valores únicos, que se acerca a 0 conforme la granularidad de los datos se incrementa.

Los valores umbrales utilizados en este trabajo fueron 95/5 para la ratio de frecuencia 10% para el porcentaje de valores únicos. Aquellos predictores que superaban la ratio de frecuencia y tenían menor porcentaje de valores únicos, fueron eliminados.

3.3.4. FILTRADO DE PREDICTORES CORRELACIONADOS

Reducir el nivel de correlación entre los predictores puede ser beneficioso tanto para el poder de clasificación como para la interpretabilidad del modelo. Además, la eliminación de predictores innecesarios puede acelerar pasos de procesamiento posteriores y el aprendizaje del algoritmo.

En un conjunto de datos como el de características radiómicas, es esperable encontrar elevada correlación entre muchas variables, ya que son medidas volumétricas. Estas nos dan información similar y por lo tanto redundante. En consecuencia, se calculó la matriz de coeficientes de correlación de Pearson (figura 7) para las matrices de datos empleadas en el trabajo.

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Figura 7: *Ecuación de correlación donde $\boldsymbol{\rho_{X,Y}}$ es el coeficiente de correlación de Pearson, $\boldsymbol{\sigma_{XY}}$ es la covarianza entre las variables X e Y, $\boldsymbol{\sigma_X}$ es la desviación típica de X y $\boldsymbol{\sigma_Y}$ es la desviación típica de Y.*

Se consideró que dos variables estaban altamente correlacionadas si tenían un coeficiente de Pearson mayor de 0,9. Del par de variables se eliminó aquella que tuviera una mayor correlación absoluta media con el resto de las variables.

3.3.5. ESTANDARIZACIÓN

La estandarización es una transformación de los datos que permite la comparación independientemente de la unidad de medida y la escala. Es un requisito para algunos métodos de aprendizaje automático.

En primer lugar, se centran las variables restando la media de cada una de ellas a todos los valores. Con esto, la nueva media es de 0 y en los valores solo se observa la variación con respecto a la media. A continuación, se divide cada valor por la desviación estándar de la variable, obteniendo el *z-score*. Así las variables tienen una desviación estándar de 1 y la escala deja de ser importante.

3.3.6. SEPARACIÓN EN SUBCONJUNTOS

Como ya se ha descrito en el apartado 1.2.2, es imprescindible tener un conjunto de entrenamiento y un conjunto de prueba independientes para poder estimar con la mayor confianza posible el poder predictivo de los modelos.

En los datos radiómicos se hicieron dos divisiones distintas, una para cada tipo de clasificación. En los dos casos, 81 observaciones formaron parte del conjunto de entrenamiento, mientras que las 19 restantes se utilizaron para la validación.

Para el conjunto de datos de cáncer de mama 456 observaciones formaron parte del conjunto de entrenamiento y 113 del conjunto de prueba.

Las divisiones se realizaron de tal manera que se mantuvo equilibrio de representación de las clases.

3.4. SELECCIÓN DE CARACTERÍSTICAS

En espacios de elevada dimensionalidad, como los que se tratan en este trabajo, es deseable reducir el número de características utilizadas para el entrenamiento con el fin de mejorar la interpretabilidad de los modelos resultantes y la capacidad predictiva de los clasificadores.

Con tal de minimizar el sobreajuste de la selección al conjunto de datos del que se dispone, este paso se realiza teniendo en cuenta únicamente las observaciones que forman parte del conjunto de entrenamiento.

En este trabajo se utiliza el método de eliminación recursiva de características (RFE) para la selección. Es una forma de *wrapper*, es decir, un método de selección que evalúa múltiples modelos y añade o elimina características para optimizar el desempeño.

El primer paso en RFE es ajustar un modelo con todos los predictores. Luego estos se ordenan en un ranking tomando como criterio de ordenación la importancia sobre el modelo. A continuación, se repiten varias iteraciones en cada una de las cuales se entrena de nuevo el modelo, pero reteniendo un número limitado de predictores (S_i), priorizando aquellos que están más arriba en el ranking. De todos los subconjuntos probados, se retienen las características utilizadas para entrenar a aquel que ha tenido un mejor desempeño.

Utilizar métodos de remuestreo se hace necesario en esta etapa para evitar el posible sobreajuste a un subconjunto de predictores que pueda resultar útil por azar. En este trabajo se utiliza validación cruzada para disminuir la probabilidad de tener predictores no informativos en el subconjunto final.

Consiste en dividir el total de los datos de entrenamiento en *k* partes, de las cuales *k*-1 se utilizan para entrenar el modelo y la parte restante se utiliza como conjunto de validación. Este proceso se repite *k* veces, cambiando el conjunto de entrenamiento y el de validación en cada iteración. En este caso, al estar tratando un problema de clasificación, en cada iteración se genera una matriz de precisión (tabla 3), a partir de la cual se calcula la precisión del modelo (figura 8) en esa iteración y con un valor concreto para un parámetro de ajuste. Una vez han transcurrido todas las iteraciones para dicho valor de parámetro, se toman las *k* precisiones calculadas y se promedian (figura 9). Esto da una estimación bastante cercana de la precisión que se obtendría con un conjunto de prueba, sin necesidad de disponer de más observaciones. Así puede observarse el efecto de parámetros de ajuste como es *S* en este caso. Una vez se conoce el valor óptimo de dicho parámetro, se construye el modelo con este ajuste y utilizando el conjunto de entrenamiento al completo.

Tabla 3: *Ejemplo de matriz de confusión 2x2.*

		Predicciones	
		Positivo	**Negativo**
Observaciones	**Positivo**	Verdaderos positivos (VP)	Falsos negativos (FN)
	Negativo	Falsos positivos (FP)	Verdaderos negativos (VN)

$$Acc = \frac{VP + VN}{T}$$

Figura 8: *Ecuación de precisión donde Acc es la precisión, VP es el número de verdaderos positivos, VN es el número de verdaderos negativos y T es el total de las predicciones.*

$$CV = \frac{1}{k}\sum_{i=1}^{k} Acc_i$$

Figura 9: *Ecuación de validación cruzada donde CV es la precisión media en el total de las iteraciones, k es el número de iteraciones y Acc_i es la precisión para la iteración i.*

El algoritmo sigue el siguiente esquema de trabajo:

1. Para un total de 10 iteraciones de validación cruzada:
   1.1. Se divide el conjunto de datos (entrenamiento) en dos subconjuntos, uno de los cuales contiene el 90% de las muestras y el otro el 10% restante.
   1.2. Se entrena el modelo con el subconjunto mayor.
   1.3. Se realizan predicciones con el modelo sobre los datos no utilizados para el entrenamiento.
   1.4. Se calcula la importancia y se elabora el ranking.
      1.4.1. Para cada uno de los tamaños *S* de subconjunto de características definidos:
         1.4.1.1. Se utilizan las S_i mejores características según el ranking para volver a entrenar el modelo.
         1.4.1.2. Se mide el desempeño por predicción sobre las observaciones no utilizadas en el entrenamiento.
2. Se calcula el desempeño para cada tamaño S_i sobre el total de las iteraciones de validación cruzada.
3. Se determina el número apropiado de predictores. En este caso, será el subconjunto de predictores de menor tamaño que tenga como máximo un 2% menos de precisión asociada que el mejor de todos los subconjuntos.
4. Entrenamiento del modelo final utilizando los S_i predictores seleccionados.

La precisión se calcula con los valores de la matriz de confusión obtenida de las predicciones del modelo sobre las observaciones de validación, según se muestra en la tabla 3 y en la figura 8.

Como ya se ha descrito, es necesario ajustar un modelo para hallar el subconjunto de características óptimo. Las características seleccionadas serán óptimas para el tipo de modelo en cuestión. En este trabajo se pretende comparar cuatro modelos distintos, por lo tanto, se llevarán a cabo cuatro procesos de selección de características independientes con el objetivo de optimizar cada uno de los modelos. Los modelos utilizados y los detalles del funcionamiento de los mismos se describirán en el apartado 3.5. Teniendo en cuenta que se van a realizar 3 análisis distinto (grado y tipo para gliomas, malignidad para cáncer de mama), esto hace un total de 12 conjuntos de características seleccionados.

Para el cálculo de la importancia pueden tomarse diversos abordajes dependiendo del modelo que se quiera ajustar. Algunos, como bosques aleatorios, tienen incorporado un método de cálculo de la importancia, mientras que otros deben recurrir a abordajes basados en “filtros”.

Los tamaños de subconjunto *S* probados fueron 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20, 25 y 30 para el conjunto de datos de gliomas, y todos los valores del 3 al 20 para el conjunto de datos de cáncer de mama.

3.5. MODELOS DE CLASIFICACIÓN

3.5.1. *K*-VECINOS MÁS PRÓXIMOS

Conocido en inglés como *k-nearest neighbours*, es un método que se basa en la estimación de la distribución de las clases en función de los datos de entrenamiento y luego clasifica las nuevas observaciones en base a la probabilidad estimada más alta. KNN identifica los *k* puntos más cercanos a nuestra observación x_0, representados por N_0, y después estima la probabilidad condicional para la clase *j* como la fracción de puntos en N_0 que pertenecen a *j*:

$$\Pr(Y = j|X = x_0) = \frac{1}{K}\sum_{i \in N_0} I(y_i = j)$$

Figura 10: *Ecuación de probabilidad condicionada de la clase j, donde Y es la clase real de la observación X, K es el número de vecinos cercanos a X, N_0 es el conjunto de vecinos cercanos e I es un valor que depende de la pertenencia del vecino i a la clase j.*

Esto se repite para cada clase y finalmente se aplica la regla de Bayes, clasificando la observación x_0 en la clase con la mayor probabilidad. Es un abordaje simple, pero en muchas ocasiones efectivo y no requiere hacer asunciones sobre la relación entre *X* e *Y*. Por otra parte, esto tiene la desventaja de que los modelos generados son más difíciles de interpretar.

El clasificador obtenido puede variar mucho dependiendo de qué valor de *k* se escoja. Valores bajos de *k* están asociados con modelos muy flexibles que probablemente no se ajusten a la realidad. Conforme *k* se incrementa, el modelo se vuelve menos flexible, lo cual lo hace más generalizable a otros datos, pero puede cometer errores en la clasificación de muestras más confusas (más cercanas al umbral de decisión). Elegir el nivel adecuado de flexibilidad es crítico para el éxito de cualquier método de aprendizaje automático.

3.5.2. BOSQUES ALEATORIOS

El método de bosque aleatorio se basa en la agregación de las predicciones de múltiples árboles de decisión diseñados de una forma específica. Así pues, es necesario entender en qué consiste un modelo de árbol de decisión.

El árbol de decisión es realmente una analogía utilizada para hablar de modelos basados en la segmentación del espacio de predictores en diversas subregiones, llamadas nodos terminales. En cada una de estas subregiones, se sitúa un subconjunto del total de los datos de entrenamiento. La clase más frecuente de entre este subconjunto será aquella utilizada para predecir futuras observaciones en este nodo terminal.

La forma en que se delimitan las subregiones utiliza como criterio de división los valores del índice de Gini:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

Figura 11: *Ecuación del índice de Gini, donde G es el índice o coeficiente de Gini, K es el número de clases representadas en una región generada por división binaria recursiva y $\hat{p}_{mk}$ es la fracción de observaciones pertenecientes a la clase k en la región m.*

En esta ecuación, *G* es el índice o coeficiente de Gini, $\hat{p}_{mk}$ es la fracción de observaciones de entrenamiento pertenecientes a la clase *k* en la región *m*. El objetivo del algoritmo de entrenamiento debe ser minimizar este índice, o lo que es lo mismo, maximizar la pureza de las regiones resultantes en cada división.

Debido a que no es computacionalmente factible considerar toda posible partición del espacio de predictores en regiones, se toma un abordaje "de arriba a abajo" (*top-down*), conocido como "división binaria recursiva" (*recursive binary splitting*). El abordaje es *top-down* porque empieza en lo alto del árbol, punto en el cual todas las observaciones pertenecen a una misma región, y sucesivamente se divide el espacio de predictores. Este es un abordaje realizable computacionalmente pero no óptimo, porque se considera que la mejor división es aquella que separe mejor las clases sólo en el paso del proceso en el que se realiza, en lugar de realizar divisiones que, aunque no minimicen el índice de Gini, permitan que el árbol sea mejor clasificador por otras divisiones posteriores.

En primer lugar, se selecciona el predictor X_j y el punto de corte *s* tal que se separen dos subespacios que minimicen *G*. Se le llama "recursiva" porque para cada subespacio generado, se repite el proceso de búsqueda del mejor predictor y el mejor punto de corte hasta que se alcanza un criterio de finalización como, por ejemplo, que cada región recoja un número mínimo de observaciones.

Los árboles de decisión por si mismos se quedan cortos en poder de predicción frente a otros modelos de aprendizaje automático. Uno de sus mayores problemas es que la división binaria recursiva tiende a sobreajustar el modelo a los datos de entrenamiento. Se puede solucionar este problema si se combinan los árboles con el *bootstrap*.

Dado un conjunto independiente de observaciones, la varianza de la media de las observaciones viene dada por σ^2/n. En otras palabras, promediar un conjunto de observaciones reduce la varianza. Teniendo esto en cuenta, se puede reducir la varianza e incrementar la capacidad predictiva de un modelo estadístico extrayendo conjuntos de entrenamientos de la población repetidamente y construyendo modelos de predicción independientes (por estar basados en diferentes conjuntos de entrenamiento) y promediando las predicciones resultantes para finalmente dar una predicción única de consenso, como se muestra en la figura 12. El problema de este abordaje es que lo habitual es no tener acceso a múltiples conjuntos de entrenamiento. En su lugar, puede hacerse *bootstrap*, extraer muestras repetidamente de un conjunto de entrenamiento hasta generar tantos conjuntos derivados como se desee.

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x)$$

Figura 12: *Ecuación de bagging con árboles de decisión, donde $\hat{f}_{bag}$ es la clase estimada por este método, B es el número de iteraciones de bootstrap y $\hat{f}^{*b}$ es la estimación de clase en la iteración b.*

Esto es lo que se conoce como empaquetado o *bagging*. En este caso, los modelos de predicción independientes mencionados en el párrafo anterior son árboles. El voto de la mayoría de los árboles es el elegido como predicción consenso definitiva. El número de árboles a entrenar es uno de los hiperparámetros a tener en cuenta, aunque en este caso elegir un valor muy alto no se traducirá en sobreajuste, de forma que se suele dejar en un valor lo suficientemente elevado como para minimizar el error de clasificación.

Los bosques aleatorios suponen un añadido sobre el *bagging* con árboles de decisión. El proceso es el mismo, pero en el momento de construir los árboles, cada vez que se considera una división, se elige aleatoriamente una muestra de *m* predictores como candidatos, en lugar de tener en cuenta todos los *p* predictores. En cada división se toman diferentes *m* predictores como candidatos, normalmente $m = \sqrt{p}$.

Esto favorece más todavía diferencias en las divisiones de diferentes árboles y consecuentemente se evita que estén correlacionados. Promediar predicciones correlacionadas no lleva a una reducción de la varianza tan grande como promediar predicciones no correlacionadas. Así pues, tener en consideración *m* predictores en lugar de *p*, hace que el clasificador final esté más cerca del modelo real.

Debido a que este modelo tiene un método de selección de características integrado (la optimización del índice de Gini), se han utilizado todas las características disponibles en la etapa de entrenamiento, en lugar de sólo aquellas seleccionadas por los métodos descritos en el apartado 3.4.

3.5.3. MÁQUINAS DE SOPORTE VECTORIAL

El modelo SVM es una expansión del modelo *support vector classifier* (SVC), que a su vez es una expansión sobre el modelo *maximal margin classifier* (MMC). Es por ello que se describirán antes estos dos últimos y luego se incidirá en las adiciones que aporta SVM.

El MMC se basa en la determinación del hiperplano que mejor separa las clases. Un hiperplano es un subespacio plano de dimensión *p - 1*. En dos dimensiones un hiperplano es una línea, mientras que, en tres dimensiones, es un plano. La figura 13 representa un hiperplano de dimensión *p*.

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0$$

Figura 13: *Ecuación general de un hiperplano en p dimensiones.*

Si una observación satisface la ecuación 4, esta estará sobre el plano, mientras que, si no la satisface, estará a uno u otro lado del plano, dependiendo del signo.

En el caso de un problema de clasificación binaria, se etiquetan las observaciones con 1 (las pertenecientes a una clase) y -1 (las pertenecientes a la otra clase). La propiedad clave de un hiperplano separador perfecto se describe matemáticamente en la figura 14.

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} > 0 \text{ si } y_i = 1$$

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} < 0 \text{ si } y_i = -1$$

Equivalentemente:

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) > 0$$

Para $i = 1, \ldots, n$

Figura 14: *Propiedades de un hiperplano separador.*

Si nuestros datos pueden separarse por un hiperplano, existirán infinitos hiperplanos que sirvan para esta tarea. Esto se debe a que un hiperplano puede rotarse y ser desplazado ligeramente sin entrar en contacto con las observaciones. Es por ello que se debe imponer alguna restricción que nos permita seleccionar el mejor de todos los hiperplanos posibles.

Queremos que el hiperplano separe nuestras observaciones de entrenamiento, pero además queremos que esta separación sea generalizable a observaciones cuya clase desconocemos. Un hiperplano clasificador generalizable debería situarse entre los dos grandes espacios de clases de tal forma que esté a una distancia equitativa de los dos grupos. En caso contrario, estaría favoreciendo la clasificación en una de las dos clases sobre la otra. Otra forma de expresar esto es buscar el hiperplano que guarda una mayor distancia con las observaciones de entrenamiento, lo que en este contexto se conoce como "margen".

En la figura 15, una representación gráfica de un hiperplano y su margen en dos dimensiones. Solo tres de las observaciones se utilizan para seleccionar el hiperplano separador óptimo. Estas observaciones son las llamadas vectores de soporte, porque son vectores en el espacio *p*-dimensional y soportan al hiperplano en el sentido de que definen las características del hiperplano.

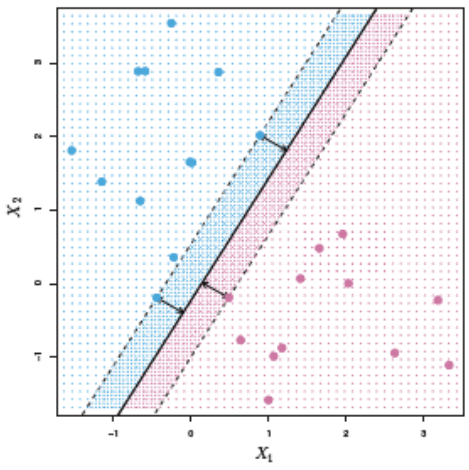

Figura 15: *Ejemplo en dos dimensiones de un hiperplano separador óptimo. Extraído de James et al. (2017).*

Construir el hiperplano es un problema de optimización que sigue las restricciones expuestas en la figura 16.

$$\text{maximizar M modificando } \beta_0, \beta_1, \dots, \beta_p \text{ sujeto a } \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right) \geq M \; \forall \; i = 1, \dots, n$$

Figura 16: *Restricciones de un hiperplano separador óptimo. β son los coeficientes, coordenadas del vector perpendicular al plano.*

El conjunto de estas restricciones garantiza que el hiperplano esté situado de tal manera que todas las observaciones estén en el lado que les corresponde y además a una distancia mínima *M*, el margen. Los detalles del problema de optimización están fuera del objetivo de este trabajo.

Por desgracia, este clasificador solo puede aplicarse en casos en los que las clases puedan separarse perfectamente por una *delimitación* lineal. El SVC es la generalización del MMC que permite su aplicación en los casos en los que las clases no son perfectamente separables por una *delimitación* lineal y además es menos sensible al sobreajuste provocado por observaciones individuales.

Como se observa en la figura 17, la adición de una observación al conjunto de entrenamiento hace que el hiperplano rote pronunciadamente y disminuye su margen.

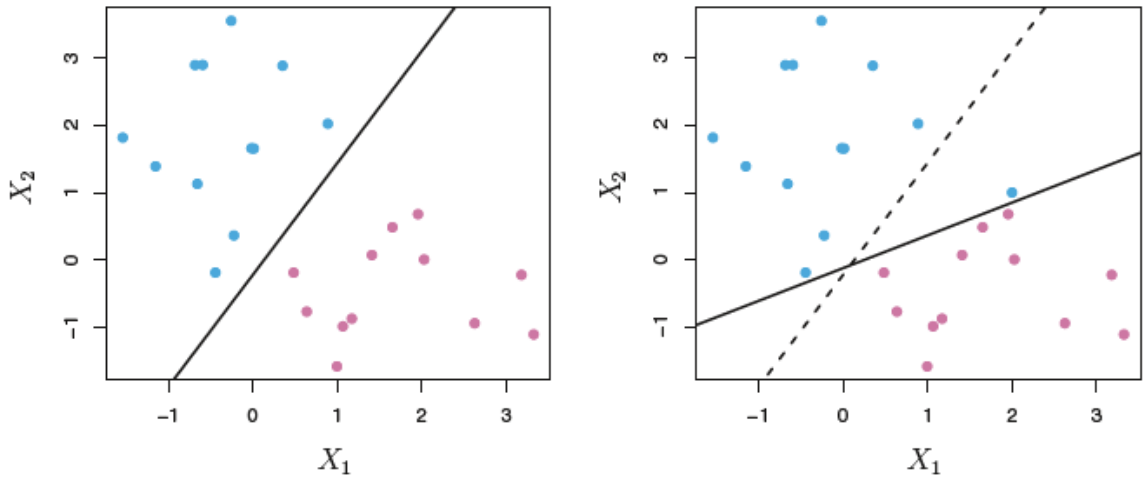

Figura 17: *Efecto de observaciones individuales sobre la configuración del hiperplano separador en MMC. Extraída de James et al. (2017).*

Puede ser beneficioso clasificar incorrectamente esa observación con el fin de separar mejor el resto del conjunto. Esto es inevitable cuando no hay un hiperplano separador. El *support vector classifier*, conocido también como *soft margin classifier*, responde a estos problemas permitiendo que algunas muestras estén dentro del margen o incluso en el lado incorrecto del hiperplano.

El problema de optimización tiene algunos añadidos sobre el de MMC (figura 18):

$$\text{maximizar M modificando } \beta_0, \beta_1, \dots, \beta_p, \epsilon_1, \dots, \epsilon_n \text{ sujeto a } \sum_{j=1}^{p} \beta_j^2 = 1$$

$$y_i\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \ \beta_p x_{ip}\right) \geq M(1 - \epsilon_i)$$

$$\epsilon_i \ \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C$$

Figura 18: *Restricciones para el hiperplano separador óptimo de margen suave. ϵ son variables slack, C es el coste y M es el grosor del margen.*

Donde ϵ son variables que permiten a las observaciones estar en una posición incorrecta con respecto al hiperplano o el margen. Si $\epsilon = 0$, la observación se encuentra en el lado correcto del margen, si $\epsilon > 0$, se dice que la observación ha infringido el margen, mientras que $\epsilon > 1$indica que la observación está en el lado incorrecto del hiperplano.

C es el coste, un hiperparámetro de ajuste que limita el valor de la suma de ϵ. Así determina el número y severidad de las infracciones al margen y al hiperplano que se está dispuesto a tolerar. Conforme *C* se incrementa, también lo hará el margen, porque habrá mayor tolerancia a las infracciones sobre este. Funciona de una forma similar a *k* en *k*-vecinos más próximos, regulando la flexibilidad del modelo.

El mayor problema de este método de clasificación es que solo funciona bien si las muestras son separables por un hiperplano. Las máquinas de soporte vectorial pretenden generalizar este modelo para poder aplicarlo en una mayor variedad de casos. Para ello, es necesario expandir el espacio de predictores de tal manera que la forma de la frontera de decisión en el espacio original (no expandido) deja de ser lineal.

Para resolver el problema del SVC se calculan los productos escalares de las observaciones como se muestra en las figuras 18 y 19.

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i \langle x, x_i \rangle = \sum_{i \in S} \alpha_i \sum_{j=1}^{p} x_{ij} x_{i'j}$$

Figura 19: *Ecuación de SVC, donde $f(x)$ es la función que separa las observaciones en el espacio, β_0 es la ordenada en el origen, S es la colección de índices de los vectores de soporte y α_i son coeficientes de peso sobre los productos escalares entre las observaciones x y x_i.*

Para estimar el parámetro β y todos los parámetros α, sólo es necesario el cálculo de los $n(n-1)/2$ productos escalares. Se tiene que calcular el producto escalar para nuestra observación *x* con todas las observaciones del conjunto de entrenamiento, sin embargo, el parámetro $\alpha = 0$ para todas las observaciones de entrenamiento que no son vectores de soporte.

Para transformar el delimitador lineal en uno no lineal, el SVM añade una transformación al producto escalar basada en *kernelización*. Un *kernel* es una función que cuantifica la similitud entre dos observaciones.

El *kernel* es una función de las observaciones y puede tomar diversas formas. Cada *kernel* da lugar a un modelo distinto. Los más conocidos son el lineal (el equivalente a SVC) y el radial (figura 20).

$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2)$$

Figura 20: *Ecuación de kernel radial, donde K es el kernel, γ es un hiperparámetro positivo, x_i y $x_{i'}$ son dos observaciones para las que se está haciendo el cálculo, p es el total de dimensiones y j es una de las dimensiones.*

Si una observación de prueba está lejos de una observación de entrenamiento concreta en términos de distancia euclídea, entonces la suma de cuadrados de la diferencia será alta, y por lo tanto la *K* será pequeña. Esto quiere decir que la observación de entrenamiento no tendrá prácticamente ningún efecto sobre la predicción final. El *kernel* radial tiene un carácter muy local, en el sentido de que solo las observaciones de entrenamiento cercanas tienen un efecto en la clasificación de una observación de prueba. Para el uso de cualquier *kernel*, la función del clasificador quedaría como se muestra en la figura 21.

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i)$$

Figura 21: *Ecuación de SVM, donde $f(x)$ es la función que separa las observaciones en el espacio, β_0 es la ordenada en el origen, x_i y $x_{i'}$ son dos observaciones para las que se está haciendo el cálculo, S es la colección de índices de los vectores de soporte y K es el kernel aplicado.*

Para realizar clasificación teniendo en cuenta más de dos clases, se pueden tomar abordajes alternativos como *One-Versus-All*, que consiste en un SVM para cada clase. Cada uno compara una clase con el conjunto de todas las demás.

3.6. ENTRENAMIENTO Y VALIDACIÓN DE LOS MODELOS

Con las características seleccionadas, se procede a entrenar los modelos. En esta etapa hay que decidir cuáles serán los valores para los hiperparámetros de ajuste de cada modelo. El abordaje habitual implica la utilización de métodos de remuestreo para la prueba de varios valores distintos para cada hiperparámetro. De todos los valores posibles, se escoge aquel que optimiza alguna métrica. En este caso se utilizó validación cruzada con 10 iteraciones y precisión como métrica a optimizar.

Una vez obtenidos los modelos finales para cada tarea de clasificación (12 en total), se utiliza por primera vez el conjunto de prueba. Igual que se hacía en la validación cruzada, se hacen predicciones para cada observación del conjunto de prueba con los modelos finales y se calculan las métricas deseadas, que en este caso son la precisión y el coeficiente kappa de Cohen.

Kappa se parece a la precisión en cuanto a que es mayor cuanto mejor es la clasificación, pero está normalizada por probabilidades aleatorias en el conjunto de datos. Es una medida útil en problemas con

clases desequilibradas. Básicamente informa de qué tan bueno es el clasificador con respecto a un clasificador que simplemente asigna clases al azar dependiendo de la frecuencia de las clases.

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Figura 22: *Ecuación del coeficiente Kappa de Cohen, donde p_o es la precisión y p_e es la probabilidad hipotética de acierto por azar.*

4. RESULTADOS Y DISCUSIÓN

4.1. ANÁLISIS EXPLORATORIO

En el conjunto de gliomas de grado bajo, el PCA y el HCA (figura 25) revelaron la presencia de 5 casos atípicos. En el PCA estas observaciones son responsables de la variabilidad recogida por la PC1 (43,36%). Según muestran los colores de las figuras 23 y 24, los tumores no pertenecen a ningún grupo concreto en cuanto al grado o tipo histológico. Debido a las grandes diferencias de estas observaciones con el resto, se decidió no tenerlos en cuenta en pasos posteriores.

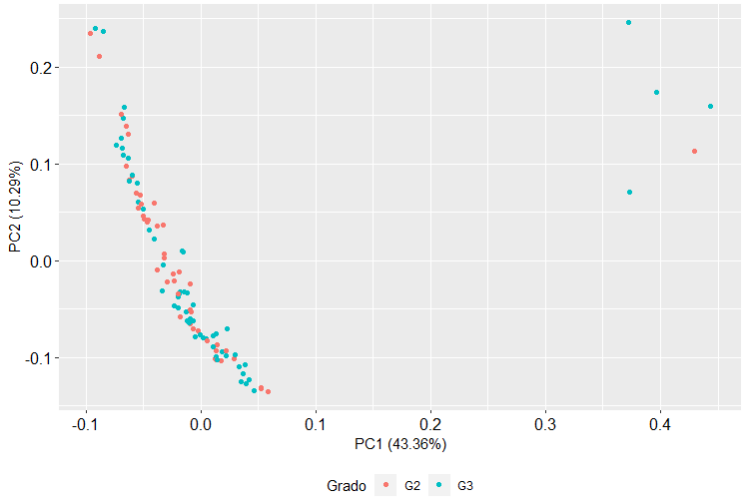

Figura 23: PCA de los datos de gliomas de grado bajo. Entre paréntesis, el porcentaje de la variabilidad total explicada por cada una de las componentes. Las observaciones se colorean según el grado. En la leyenda, G2 indica grado 2 y G3 indica grado 3.

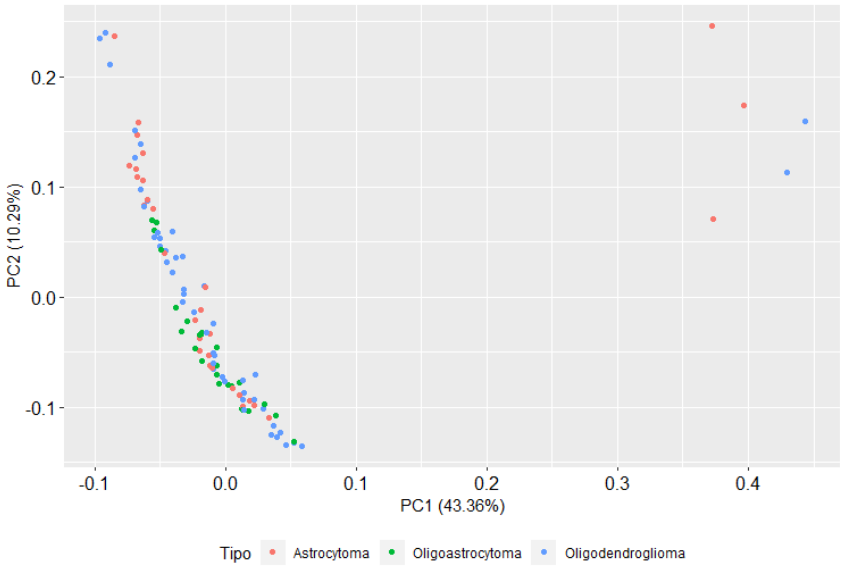

Figura 24: PCA de los datos de gliomas de grado bajo. Entre paréntesis, el porcentaje de la variabilidad total explicada por cada una de las componentes. Las observaciones se colorean según el tipo histológico.

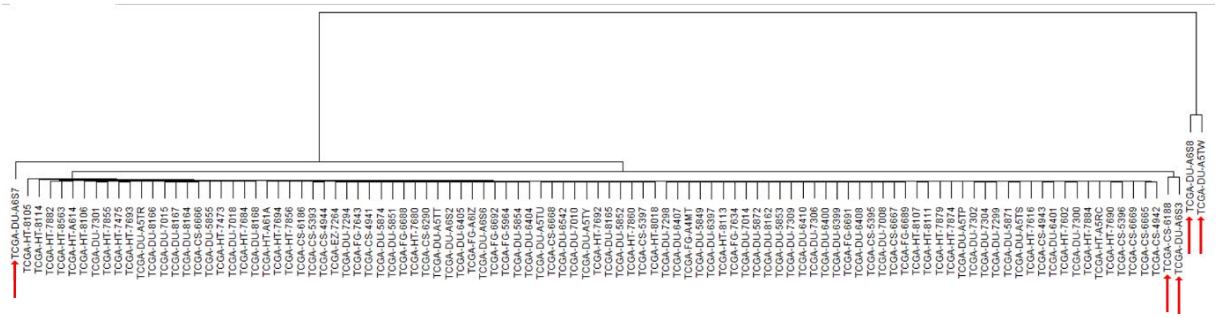

Figura 25: HCA por distancia euclídea de los datos de gliomas de grado bajo. Los casos atípicos están señalados con flechas rojas.

En las figuras 26 y 27 se muestra PCA una vez retirados los casos atípicos. No parece haber una distribución clara de las clases teniendo en cuenta todas las características, lo que nos anticipa que la tarea de clasificación será difícil porque estas características no separan las clases adecuadamente.

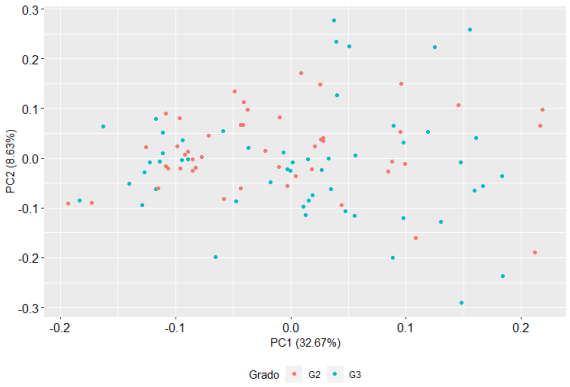

Figura 26: PCA de los datos de gliomas de grado bajo tras retirar los casos atípicos. Entre paréntesis, el porcentaje de la variabilidad total explicada por cada una de las componentes. Las observaciones se colorean según el grado. En la leyenda, G2 indica grado 2 y G3 indica grado 3.

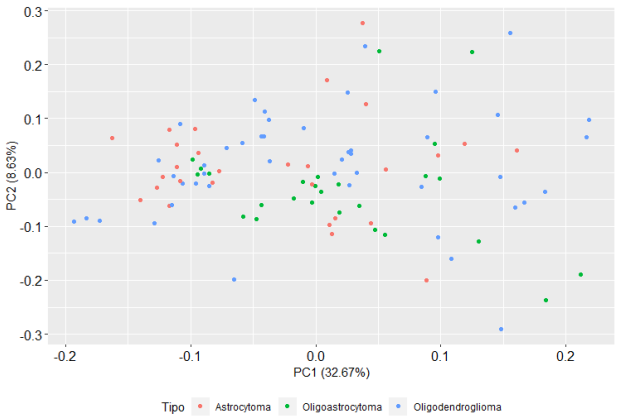

Figura 27: PCA de los datos de gliomas de grado bajo tras retirar los casos atípicos. Entre paréntesis, el porcentaje de la variabilidad total explicada por cada una de las componentes. Las observaciones se colorean según el tipo histológico de tumor.

En el conjunto de datos de núcleos de células de cáncer de mama, se ve una división clara de los dos grupos por una diagonal (figura 28). Gracias a esto se puede esperar que los clasificadores sepan distinguir la clase en la mayoría de las observaciones. Además, no se observa ningún valor atípico, de forma que se conservan las 536 observaciones para análisis posteriores.

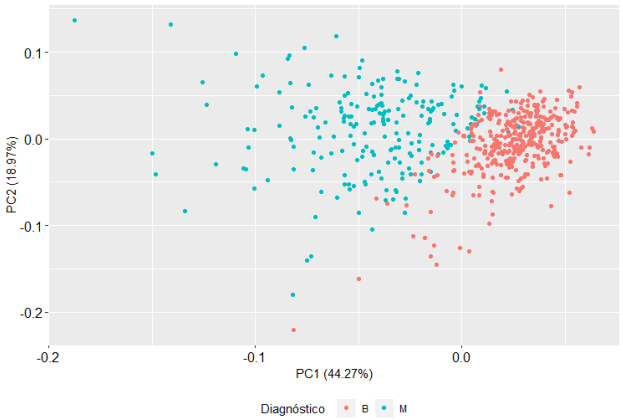

Figura 28: PCA de los datos de células de cáncer de mama. Entre paréntesis, el porcentaje de la variabilidad total explicada por cada una de las componentes. En la leyenda, B equivale a "benigno" y M equivale a "maligno".

4.2. FILTRADO DE PREDICTORES

El filtrado por varianza detectó y descartó 1112 características en el conjunto de datos gliomas, dejando un total de 5214 características. Este procesado no tuvo efecto sobre los datos de cáncer de mama.

Por otra parte, el filtro de correlación aplicado al conjunto de gliomas de grado bajo redujo el número de variables de 5214 a 1102, como era de esperar por la naturaleza de los datos. En el caso de los datos de cáncer de mama, 9 de las variables fueron descartadas, dejando un total de 21 disponibles para el siguiente paso del proceso.

4.3. SELECCIÓN DE CARACTERÍSTICAS

4.3.1. GLIOMAS DE GRADO BAJO

La figura 29 muestra cómo una mayor cantidad de características no tiene por qué estar relacionada con mejor precisión de los clasificadores. En la clasificación por grado de tumor la máxima precisión se alcanza con tan solo 3 predictores en el caso de KNN, SVML y SVMR. El clasificador RF se beneficia de un número intermedio de variables.

Cientos de variables han sido consideradas irrelevantes para la clasificación por el método RFE, ya que su adición al modelo no mejora su capacidad predictiva e incluso en algunos casos la empeora. Por otra parte, en los clasificadores según el tipo de glioma, añadir más variables tiene un efecto positivo, excepto en el modelo RF (figura 30). Esto puede significar que para la distinción del tipo de glioma es necesario tener en cuenta una mayor cantidad de variables. Los modelos con más variables son más flexibles y por lo tanto más útiles en el modelado de patrones complejos en los que pueden participar interacciones entre variables.

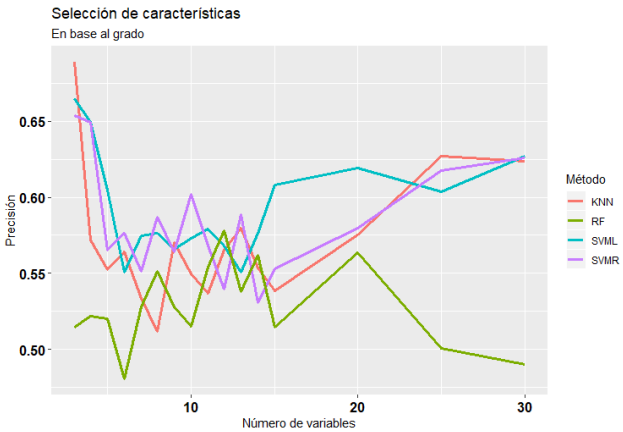

Figura 29: *Resultado de las iteraciones de validación cruzada en el proceso de RFE. Se enfrentan la precisión con el número de variables utilizadas para entrenar los modelos. La variable objetivo es el grado de glioma.*

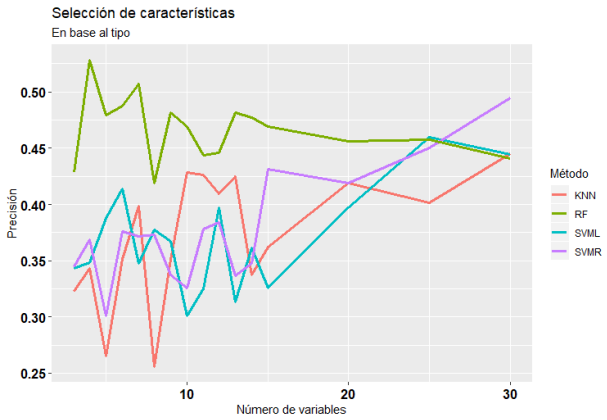

Figura 30: *Resultado de las iteraciones de validación cruzada en el proceso de RFE. Se enfrentan la precisión con el número de variables utilizadas para entrenar los modelos. La variable objetivo es el tipo histológico de glioma.*

En cualquier caso, las oscilaciones en precisión que experimentan los modelos con la adición de variables hacen patente que la calidad de los datos no es lo suficientemente buena como para utilizarse para entrenar un clasificador con aprendizaje automático, ya que se alcanzan picos de precisión de 0,7 en grado de tumor y de 0,56 en tipo de tumor.

Las características seleccionadas para cada modelo se enumeran en la tabla 4.

Tabla 4: *Número e identificador de las variables seleccionadas para cada modelo en los problemas de clasificación de gliomas. Si el número de variables es mayor de 6, solo se muestran los primeros 5 identificadores. KNN = k vecinos más próximos, SVML = máquinas de soporte vectorial con kernel lineal, SVMR = máquinas de soporte vectorial con kernel radial, RF = bosques aleatorios.*

Clasificación	Modelo	Nº Variables	ID Variables
Grado	KNN	3	NET.original_shape_Sphericity, NET.log.sigma.5.0.mm.3D_firstorder_Median, T1c.wavelet.LLL_firstorder_Kurtosis
	SVML	3	NET.original_shape_Sphericity, NET.log.sigma.5.0.mm.3D_firstorder_Median, T1c.wavelet.LLL_firstorder_Kurtosis
	SVMR	3	NET.original_shape_Sphericity, NET.log.sigma.5.0.mm.3D_firstorder_Median, T1c.wavelet.LLL_firstorder_Kurtosis
	RF	12	NET.original_shape_Sphericity, NET.original_shape_SurfaceVolumeRatio, NET.log.sigma.5.0.mm.3D_firstorder_Median, NET.wavelet.HLL_glcm_Imc2, T1c.wavelet.LHH_glszm_ZoneEntropy, …
Tipo	KNN	30	T2.wavelet.LHH_glcm_ClusterShade, NET.wavelet.LHH_firstorder_Skewness, FLAIR.wavelet.LHH_glcm_ClusterShade, T2.wavelet.LHH_firstorder_Mean, T1c.log.sigma.1.0.mm.3D_glcm_InverseVariance, …
	SVML	25	T2.wavelet.LHH_glcm_ClusterShade, NET.wavelet.LHH_firstorder_Skewness, FLAIR.wavelet.LHH_glcm_ClusterShade, T2.wavelet.LHH_firstorder_Mean, T1c.log.sigma.1.0.mm.3D_glcm_InverseVariance, …
	SVMR	30	T2.wavelet.LHH_glcm_ClusterShade, NET.wavelet.LHH_firstorder_Skewness, FLAIR.wavelet.LHH_glcm_ClusterShade, T2.wavelet.LHH_firstorder_Mean, T1c.log.sigma.1.0.mm.3D_glcm_InverseVariance, …
	RF	4	T2.wavelet.LHH_glcm_ClusterShade, FLAIR.wavelet.HLL_glcm_Imc1, ED.wavelet.LLH_glrlm_RunVariance, T1c.wavelet.HLH_glcm_Imc1

En la clasificación por grado, las tres variables más importantes y frecuentes en los conjuntos seleccionados son:

- NET.original_shape_Sphericity. Medida de la esfericidad de la región NET. Es una medida adimensional, independiente de escala y orientación. El rango de valores está entre 0 y 1 porque es relativa a una esfera. Un valor de 1 indica una esfera perfecta.
- NET.log.sigma.5.0.mm.3D_firstorder_Median. La mediana de intensidad de los vóxeles en la región NET.
- T1c.wavelet.LLL_firstorder_Kurtosis. La curtosis es una medida estadística de la distribución de los valores en la región de interés de la imagen, en este caso es todo el tumor y bajo la modalidad T1 de MRI con gadolinio (agente de contraste). Una curtosis más elevada implica que la masa de la distribución está concentrada en las colas más que entorno a la media. Curtosis baja implica justo lo contrario.

En la clasificación por tipo, las variables más importantes son T2.wavelet.LHH_glcm_ClusterShade, FLAIR.wavelet.LHH_glcm_ClusterShade, NET.wavelet.LHH_firstorder_Skewness y T2.wavelet.LHH_firstorder_Mean. Estas cuatro características están relacionadas con la textura y la distribución de intensidades de la imagen cuando se le aplica un filtro wavelet, cuyo principal objetivo es para disminuir el ruido de fondo.

La variedad de características seleccionadas y la precisión obtenida en las iteraciones de RFE no permiten afirmar con seguridad que estas características sean determinantes en la clasificación del grado o tipo de un glioma o, por lo menos, no son determinantes por si mismas, sino que necesitan combinarse para, en el mejor de los casos, hacer clasificaciones ligeramente mejores que las obtenidas por azar.

4.3.2. CÁNCER DE MAMA

En este caso la RFE también fue efectiva para reducir la cantidad de variables de interés. A pesar de que en la mayoría de los modelos el mejor resultado se obtiene utilizando todas las variables (figura 31), se escogió un subconjunto reducido con el fin de mejorar la interpretabilidad y hacer que los modelos finales fueran más generalizables.

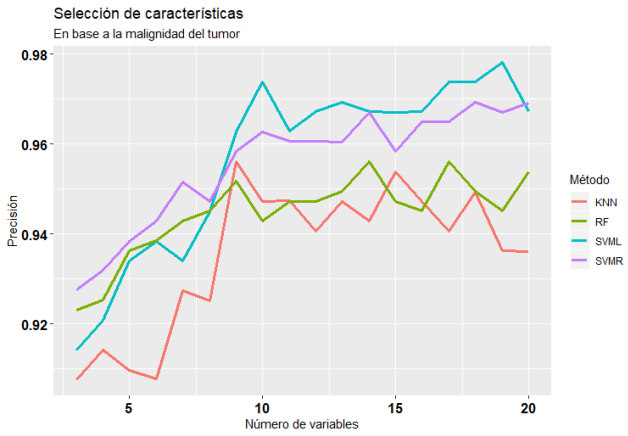

Figura 31: *Resultado de las iteraciones de validación cruzada en el proceso de RFE. Se enfrentan la precisión con el número de variables utilizadas para entrenar los modelos. La variable objetivo la malignidad o no del cáncer de mama.*

Las oscilaciones no son tan grandes como se observaba en el otro conjunto de datos y pocas variables son suficientes para alcanzar precisiones de más de 0.9. Esto es señal de la buena calidad de las características seleccionadas y del total del conjunto.

En la tabla 5 se enumeran las características más importantes. Las propiedades morfológicas que tienen mayor representación en los conjuntos de características seleccionados son el área y las concavidades o irregularidades en la membrana nuclear. La selección de estas características indica que los núcleos de las células de tumores malignos tienen un mayor tamaño e irregularidades en su membrana nuclear.

Tabla 5: *Número e identificador de las variables seleccionadas para cada modelo en el problema de clasificación de cáncer de mama. Si el número de variables es mayor de 6, solo se muestran los primeros 5 identificadores. KNN = k-vecinos más próximos, SVML = máquinas de soporte vectorial con kernel lineal, SVMR = máquinas de soporte vectorial con kernel radial, RF = bosques aleatorios.*

Modelo	Nº Variables	ID Variables
KNN	9	Worst.Concave.Points, Mean.Area, Worst.Concavity, Mean.Compactness, SE.Radius, …
SVML	9	Worst.Concave.Points, Mean.Area, Worst.Concavity, Mean.Compactness, SE.Radius, …
SVMR	7	Worst.Concave.Points, Mean.Area, Worst.Concavity, Mean.Compactness, SE.Radius, …
RF	6	Mean.Area, Worst.Concave.Points, SE.Radius, Worst.Concavity, Worst.Texture, Worst.Compactness

Según la bibliografía esto tiene sentido biológico, ya que en las células malignas son habituales los cambios en la superficie, volumen, ratio núcleo/citoplasma, forma, densidad, organización y homogeneidad. Estas características están relacionadas con la segmentación del núcleo, formación de invaginaciones y cambios en la cromatina como reducción de la heterocromatina, aumento en gránulos de intercromatina y pericromatina, incremento en el número de poros nucleares y la formación de inclusiones (Baba et al., 2007).

4.4. ENTRENAMIENTO Y VALIDACIÓN

4.4.1. GLIOMAS DE GRADO BAJO

En los dos problemas de clasificación de gliomas, el modelo con mejor desempeño es el SVML (tabla 6), lo cual implica que las clases son mejor separadas en el espacio de predictores por un hiperplano. Aun así, la precisión de SVML sobre el conjunto de prueba es de tan solo 0,737 (en clasificación por grado) y 0,632 (en clasificación por tipo), números demasiado bajos como para considerar la aplicación de los modelos en un contexto biomédico.

Tabla 6: *Clasificadores de gliomas resultantes de la selección de hiperparámetros por CV. Se muestran la precisión y el coeficiente Kappa de Cohen tanto para el proceso de validación cruzada como para las predicciones sobre el conjunto de prueba.*

Clasificación	Modelo	Hiperparámetros seleccionados	Precisión		Kappa	
			CV	Validación	CV	Validación
Grado	KNN	k = 13	0.721	0.632	0.442	0.265
	SVML	C = 0.03125	0.721	0.737	0.448	0.475
	SVMR	sigma = 0.25, C = 16	0.734	0.684	0.467	0.367
	RF	mtry = 2	0.759	0.684	0.511	0.360
Tipo	KNN	k = 14	0.588	0.579	0.366	0.350
	SVML	C = 0.125	0.579	0.632	0.346	0.422
	SVMR	sigma = 0.03125, C = 1	0.589	0.526	0.342	0.216
	RF	mtry = 1	0.618	0.579	0.387	0.350

Dada la escasez de muestras y vista la distribución de los datos en las figuras 26 y 27, estos resultados eran de esperar y señalan la importancia de un buen diseño experimental y del análisis exploratorio que nos describirá el punto de partida que disponemos y anticipará los resultados y dificultades que puedan aparecer.

4.4.2. CÁNCER DE MAMA

En el problema de clasificación de cáncer de mama, todos los clasificadores tienen un elevado coeficiente Kappa, siendo KNN y SVMR ligeramente mejores.

Tabla 7: *Clasificadores de células de cáncer de mama resultantes de la selección de hiperparámetros por CV. Se muestran la precisión y el coeficiente Kappa de Cohen tanto para el proceso de validación cruzada como para las predicciones sobre el conjunto de prueba.*

Modelo	Hiperparámetros seleccionados	Precisión		Kappa	
		CV	Validación	CV	Validación
KNN	k = 5	0.965	0.974	0.924	0.942
SVML	C = 0.5	0.967	0.965	0.929	0.924
SVMR	sigma = 0.125, C = 4	0.952	0.974	0.895	0.942
RF	mtry = 1	0.961	0.956	0.915	0.904

En el trabajo de Street et al. (1993) se obtuvieron resultados parecidos con el método *multi-surface*, un 97% de precisión estimada por validación cruzada.

En este punto podría considerarse la aplicación de los modelos como método de diagnóstico, pero antes conviene evaluar otras métricas de predicción: la sensibilidad y especificidad. La sensibilidad es la fracción de casos positivos clasificados correctamente, mientras que la especificidad es la fracción de casos negativos clasificados correctamente. Ambas existen en un equilibrio y modificar el método para incrementar una, reducirá la otra.

Estas métricas nos ayudan a entender la magnitud de falsos positivos y/o falsos negativos que podrían diagnosticarse en caso de la aplicación masiva de los modelos de aprendizaje automático como método de diagnóstico. En el caso de diagnosticar un falso negativo, el enfermo, desconocedor de su condición, podría empeorar y los beneficios de un diagnóstico temprano se perderían. En caso de diagnosticar un falso positivo, personas sanas se verían sometidas a métodos de diagnóstico más invasivos o tomarían un tratamiento que no necesitan, con todos los efectos perjudiciales que pueda acarrear.

Por estos motivos, en clínica se utilizan métricas derivadas de estas dos que además dependen de la prevalencia de la enfermedad en la población: valores predictivos positivo (PPV) y negativo (NPV). El PPV es la probabilidad de que el paciente sea verdadero positivo dado que el diagnóstico ha sido positivo. El NPV es la probabilidad de que el paciente sea verdadero negativo dado que el diagnóstico ha sido negativo.

Para el caso concreto de KNN y SVMR, el PPV tiene un valor de 1 y el NPV tiene un valor de 0,9595, tomando el tumor maligno como el caso positivo. Como prueba de diagnóstico primario todavía no sería completamente fiable ya que 4 de cada 100 tumores malignos serían erróneamente diagnosticados como benignos. En trabajos posteriores sería conveniente adaptar los hiperparámetros *k*, *sigma* y *C* con

tal de obtener mayor sensibilidad, aunque menor especificidad. Otra opción sería aumentar el número de observaciones utilizadas para el entrenamiento con tal de refinar las dos métricas.

Al ser esta una prueba poco invasiva, su utilidad sería mayor como primera prueba de diagnóstico y, en el caso de un resultado positivo, podría realizarse alguna prueba posterior, tal vez más invasiva o compleja con la que confirmar o rechazar la clasificación realizada por el modelo. En caso de un resultado negativo, un NPV de 1 nos garantizaría que la predicción ha sido correcta. Así esta prueba sería útil para descartar todos los negativos y someter a mayor análisis a los positivos.

No hay que olvidar que este trabajo tiene sus limitaciones, ya que hay que tener en cuenta que la muestra poblacional con la que se han entrenado los modelos puede no ser representativa de cualquier otra población, es decir, estos modelos deberían aplicarse únicamente a la población de la cual proceden los datos de entrenamiento porque se desconoce si hay factores dependientes de población que puedan afectar al rendimiento de los clasificadores. Con tal de descubrir si se pueden aplicar estos clasificadores a otras poblaciones, sería necesario obtener muestras de dichas poblaciones, tanto las variables seleccionadas como la variable objetivo, y utilizar dichas muestras como conjunto de prueba. De las matrices de confusión obtenidas se calcularían nuevos valores de PPV y NPV que servirían como criterio de evaluación de los modelos en las nuevas poblaciones.

5. CONCLUSIONES

1. El análisis exploratorio de los datos mediante métodos de aprendizaje automático no supervisados, como PCA o *clustering*, es una forma eficaz de anticipar la utilidad que tendrán los datos en la elaboración de un modelo de clasificación.
2. La eliminación recursiva de características es un método apto para la reducción de la dimensionalidad de los datos y al mismo tiempo la identificación y conservación de las características influyentes en resultados clínicos. Además, nos permite conocer cuál es el efecto del uso de más variables en los modelos a nivel de métricas de precisión, lo cual también es indicativo de la calidad de dichas variables.
3. En los problemas de clasificación, la calidad de los datos viene determinada por la cantidad de observaciones y la relevancia de las características en la clasificación. A pesar de que se disponga de miles de variables, si estas no están asociadas con la variable objetivo, no serán de utilidad.
4. Cualquiera de los cuatro métodos utilizados en este trabajo (*k*-vecinos más próximos, máquinas de soporte vectorial lineal y radial y bosques aleatorios) puede tener un buen desempeño dada la buena calidad de los datos.
5. El efecto de la elección del modelo sobre el resultado final depende del problema que se quiera atender. No existe un modelo perfecto para cualquier situación. Siempre que sea computacionalmente factible, es recomendable probar varios modelos clasificadores con tal de evaluar cuál de ellos se ajusta mejor al problema de clasificación que se está tratando.
6. Los métodos de aprendizaje supervisado pueden ser aplicados con éxito en estudios biomédicos, pero es necesaria una extracción de características dedicada a encontrar aquellas variables que puedan estar relacionadas con el resultado clínico que quiere predecirse o clasificarse, además de analizar las métricas pertinentes.

6. REFERENCIAS

- BABA, A. and CATOI, C. (2007). *Comparative oncology*. Bucharest: The Publishing House of the Romanian Academy.

- BAKAS, S.; AKBARI, H.; SOTIRAS, A.; BILELLO, M.; ROZYCKI, KIRBY, J.S., … DAVATZIKOS, C. (2017). Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, 4, p.170117.

- BREIMAN, L. (2001). Random Forests. *Machine Learning*, 45(1), pp.5-32.

- CANCER GENOME ATLAS RESEARCH NETWORK; BRAT, D.J.; VERHAAK, R.G.; ADAPE, K.D.; YUNG, W.K.; SALAMA, S.R., … ZHANG J. (2015). Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *New England Journal of Medicine*, 372(26), pp.2481-2498.

- CECCARELLI, M.; BARTHEL, F.; MALTA, T.; SABEDOT, T.; SALAMA, S.; MURRAY, B., … ZMUDA, E. (2016). Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell*, 164(3), pp.550-563.

- COLLINS, F.S.; VARMUS H. (2015). A new initiative on precision medicine. *The New England Journal of Medicine*, 372(9), pp.793–795.

- COOK, C.; LOPEZ, R.; STROE, O.; COCHRANE, G.; BROOKSBANK, C.; BIRNEY, E. and APWEILER, R. (2018). The European Bioinformatics Institute in 2018: tools, infrastructure and training. *Nucleic Acids Research*, 47(D1), pp.D15-D22.

- COROLLER, T.; GROSSMANN, P.; HOU, Y.; RIOS VELAZQUEZ, E.; LEIJENAAR, R.; HERMANN, G.; LAMBIN, P.; HAIBE-KAINS, B.; MAK, R. and AERTS, H. (2015). CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiotherapy and Oncology*, 114(3), pp.345-350.

- CORTES, C. and VAPNIK, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3), pp.273-297.

- DAWSON, S.; TSUI, D.; MURTAZA, M.; BIGGS, H.; RUEDA, O.; CHIN, S., … ROSENFELD, N. (2013). Analysis of Circulating Tumor DNA to Monitor Metastatic Breast Cancer. *New England Journal of Medicine*, 368(13), pp.1199-1209.

- DEO, R. (2015). Machine Learning in Medicine. *Circulation*, 132(20), pp.1920-1930.

- GILLIES, R.; KINAHAN, P. and HRICAK, H. (2016). Radiomics: Images Are More than Pictures, They Are Data. *Radiology*, 278(2), pp.563-577.

- GREENE, C.; TAN, J.; UNG, M.; MOORE, J. and CHENG, C. (2014). Big Data Bioinformatics. *Journal of Cellular Physiology*, 229(12), pp.1896-1900.

- JAMES, G.; WITTEN, D.; HASTIE, T. and TIBSHIRANI, R. (2017). *An introduction to statistical learning*. Ed. 7 Springer. Nueva York.

- JAMESON, J. and LONGO, D. (2015). Precision Medicine — Personalized, Problematic, and Promising. *New England Journal of Medicine*, 372(23), pp.2229-2234.

- JOUGANOUS, J.; SAVIDAN, R. and BELLEC, A. (2019). *A brief overview of Automatic Machine Learning solutions (AutoML)*. [online] Hacker Noon. Disponible en: https://hackernoon.com/a-brief-overview-of-automatic-machine-learning-solutions-automl-2826c7807a2a [Accedido 30 de junio de 2019].

- KWAK, E.; BANG, Y.; CAMIDGE, D.; SHAW, A.; SOLOMON, B.; MAKI, R., … IAFRATE, A. (2010). Anaplastic Lymphoma Kinase Inhibition in Non–Small-Cell Lung Cancer. *New England Journal of Medicine*, 363(18), pp.1693-1703.

- LÓPEZ CERDÁN, A. (2019). Aplicación del análisis radiómico en el estudio de características biomédicas de interés en gliomas de grado bajo (Trabajo de final de máster). Centro de Investigación Príncipe Felipe, Valencia.

- MARX, V. (2013). The big challenges of big data. *Nature*, 498(7453), pp.255-260.

- MENDELSOHN, J.; GRAY, J.; HOWLEY, P.; ISRAEL, M., and THOMPSON, C. (2015). *The molecular basis of cancer*. Philadelphia, PA: Elsevier, Saunders.

- RATHORE, S.; AKBARI, H. and DOSHI, J. (2018). Radiomic signature of infiltration in peritumoral edema predicts subsequent recurrence in glioblastoma: implications for personalized radiotherapy planning. *Journal of Medical Imaging*, 5(02), p.1.

- RINNE, J.; BROOKS, D.; ROSSOR, M.; FOX, N.; BULLOCK, R.; KLUNK, W., … GRUNDMAN, M. (2010). 11C-PiB PET assessment of change in fibrillar amyloid-β load in patients with Alzheimer's disease treated with bapineuzumab: a phase 2, double-blind, placebo-controlled, ascending-dose study. *The Lancet Neurology*, 9(4), pp.363-372.

- SIMAN, R.; GIOVANNONE, N.; HANTEN, G.; WILDE, E.; MCCAULEY, S., HUNTER, … SMITH, D. (2013). Evidence That the Blood Biomarker SNTF Predicts Brain Imaging Changes and Persistent Cognitive Dysfunction in Mild TBI Patients. *Frontiers in Neurology*, 4.

- STREET, W.; WOLBERG, W. and MANGASARIAN, O. (1993). Nuclear feature extraction for breast tumor diagnosis. *Biomedical Image Processing and Biomedical Visualization*.

- WIBMER, A.; HRICAK, H.; GONDO, T.; MATUSMOTO, K.; VEERARAGHAVAN, H.; FEHR, D.; ZHENG, J.; GOLDMAN, D.; MOSKOWITZ, C.; FINE, S.; REUTER, V.; EASTHAM, J.; SALA, E. and VARGAS H. (2015). Haralick texture analysis of prostate MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores. *European Radiology*, 25(10), pp.2840-2850.