Evaluación de la capacidad predictiva de modelos de aprendizaje supervisado para la clasificación de pacientes con cáncer colorrectal

Pablo Román-Naranjo Varela

Máster en Ciencia de Datos
Área 3

Tutor/a de TF
Antonio Ruiz Falcó Rojas (UOC)
Ipek Guler Caamano (AMADIX)

Profesor/a responsable de la asignatura
Ferrán Prados Carrasco

25/06/2023

Universitat Oberta de Catalunya

Ficha del Trabajo Final

Título del trabajo:	Evaluación de la capacidad predictiva de modelos de aprendizaje supervisado para la clasificación de pacientes con cáncer colorrectal
Nombre del autor/a:	Pablo Román-Naranjo Varela
Nombre del Tutor/a de TF:	Antonio Ruiz Falcó Rojas (UOC) Ipek Guler Caamano (AMADIX)
Nombre del/de la PRA:	Ferrán Prados Carrasco
Fecha de entrega:	11/06/2023
Titulación o programa:	Máster en Ciencia de Datos
Área del Trabajo Final:	Área 3
Idioma del trabajo:	Castellano
Palabras clave	cáncer colorrectal, diagnóstico precoz, aprendizaje automático

Resumen del Trabajo

El cáncer colorrectal (CCR) es la segunda causa de muerte por cáncer en todo el mundo, representando un 9,5% de todas las muertes causadas por esta enfermedad. Además de la edad del paciente, existen otros factores que confieren riesgo de desarrollar CCR y, por tanto, deben considerarse para determinar las poblaciones sobre las que se realizan los programas de cribado. La identificación de estos factores de riesgo permitiría un enfoque personalizado y preciso para cada paciente, lo que ayudaría a mejorar la tasa de supervivencia. De esta manera, el objetivo principal de este trabajo fue la identificación de marcadores de riesgo útiles para la detección temprana del CCR haciendo uso de algoritmos de aprendizaje automático.

Para ello, se comparó la capacidad predictiva de diferentes modelos de aprendizaje automático supervisado, como *gradient boosting*, máquinas de vectores de soporte (SVM) o *random forest*, haciendo uso de un conjunto de datos público sobre los niveles de hidroximetilación en los *enhancers* de pacientes con CCR, AAR y controles. Además, se evaluó la idoneidad de K-means para la identificación de subgrupos de pacientes con CCR a partir de estos datos.

Los resultados de este trabajo sugieren que el mejor modelo supervisado para diferenciar los pacientes con CCR y controles a partir de datos de hidroximetilación fue un modelo SVM con kernel lineal, cuya sensibilidad para detectar la enfermedad

fue del 58% tras fijar la especificidad al 95%, mejorando el modelo presentado en el artículo de donde se extrajeron los datos. Además, los enhancers que regulan la expresión de genes como *MYSM1* o *SP1*, o los que regulan genes que codifican proteínas involucradas en rutas como las rutas del TGF-β y las integrinas, fueron identificados como los más relevantes al realizar la clasificación de las muestras en CCR o control. Por otro lado, el uso de K-means identificó 6 clústeres entre las muestras del set de datos de hidroximetilación. Dos de estos clústeres estaban conformados principalmente por muestras con CCR, no obstante, estos no se asociaban a una etapa de desarrollo concreta, y la diferenciación entre clústeres no fue clara, obteniendo clústeres muy próximos.

De esta manera, podemos concluir que los datos de hidroximetilación fueron útiles para la identificación de biomarcadores del CCR, obteniendo resultados prometedores mediante aprendizaje automático supervisado. No obstante, estos resultados deben ser interpretados como preliminares, necesitándose una validación en una cohorte externa y un análisis molecular de los biomarcadores señalados.

Abstract

Colorectal cancer (CRC) is the second most common cause of cancer death, accounting for 9.5% of all cancer deaths. In addition to patient age, other potential risk factors should be considered to correctly identify the target population for CRC screening programmes. The identification of these risk factors would allow a personalised and accurate approach for each patient that would help improve the survival rate. Thus, the main objective of this study was to identify useful risk biomarkers for the early detection of CRC using machine learning algorithms.

For this purpose, we compared the predictive ability of different supervised machine learning models, such as gradient boosting, support vector machines (SVM) or random forest, using a public dataset on hydroxymethylation levels in the enhancer regions in CRC patients, AAR and controls. In addition, we evaluated the suitability of K-means for the identification of CRC patient subgroups using this dataset.

The results of this work suggested that the best supervised model to differentiate CRC patients from controls, using hydroxymethylation data, was a SVM model with linear kernel, whose sensitivity was 58% after setting the specificity to 95%, improving the model presented in the article from which the dataset was extracted. In addition, enhancers that regulate the expression of genes such as *MYSM1* or *SP1*, or those that regulate genes encoding proteins involved in pathways such as TGF-β and integrin pathways, were identified as the most relevant enhancers when classifying

samples into CRC or control. On the other hand, the use of K-means identified 6 clusters among the samples in the hydroxymethylation dataset. Two of these clusters were mainly composed of samples with CCR, however, these clusters were not associated with a specific stage of development, and the differentiation between clusters was not clear, obtaining very close clusters.

Thus, we can conclude that hydroxymethylation data were useful for the identification of CRC biomarkers, obtaining promising results by supervised machine learning approaches. However, these results should be interpreted as preliminary, requiring validation in an external cohort and molecular analysis of the biomarkers identified.

Índice

Lista de Figuras

Capítulo 1: Introducción

1.1. Contexto y justificación del trabajo

El cáncer es un conjunto de enfermedades en las que las células de nuestro cuerpo crecen de manera descontrolada y se multiplican de forma anormal, formando tumores. En el caso del cáncer colorrectal (CCR), es una enfermedad que afecta al colon y al recto debido a la proliferación de las células que forman el tejido o epitelio glandular del colon. El CCR constituye uno de los principales retos del sistema de salud, ya que es el tercer tipo de cáncer más diagnosticado y el segundo más mortal a nivel mundial, solo por detrás del cáncer de pulmón[1]. No obstante, y debido al número de casos identificados en personas mayores a 70 años, se estima que la incidencia del CCR siga al alza, aumentando hasta en un 63% para 2040 cuando se comparan con las cifras obtenidas en 2020 (Figura 1)[2].

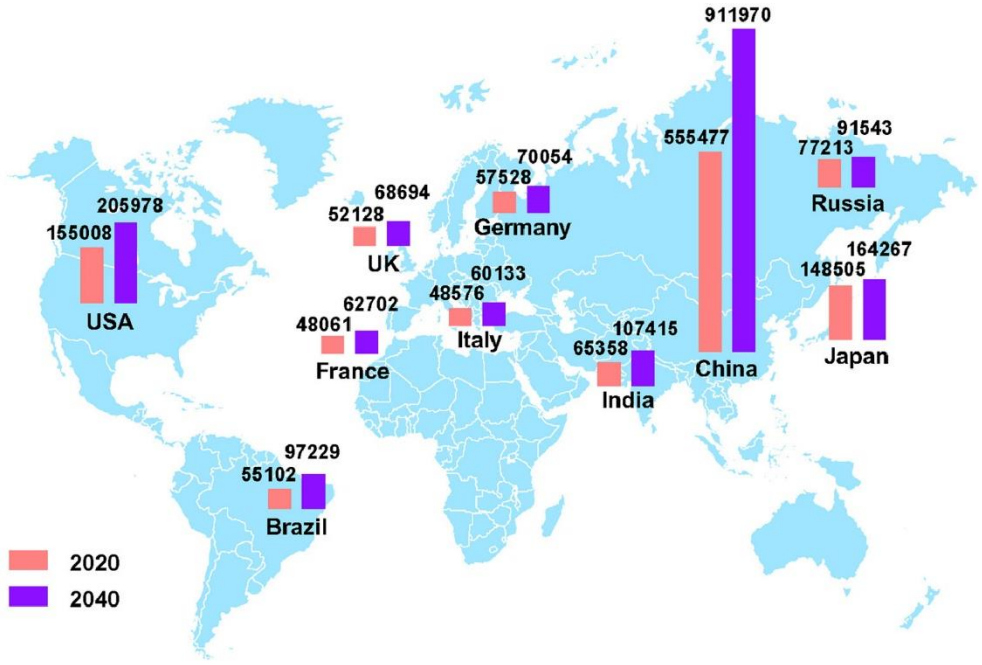


Figura 1: Número de casos identificados con CCR en los 10 países con mayor incidencia en 2020 y la previsión para 2040. Figura extraída y modificada de Xi *et al*[3].

En los últimos años, diferentes estudios han mostrado cómo la incidencia del CCR ha disminuido en la población general de los países desarrollados. Esto se debe principalmente a 1) la detección temprana con programas de cribado, 2) la eliminación de pólipos precancerosos mediante colonoscopia, y 3) la modificación del estilo de vida, disminuyendo

factores de riesgo como el consumo de alcohol y tabaco, dietas poco saludables, sedentarismo o sobrepeso[4,5]. Sin embargo, si nos centramos en la población de los países en vías de desarrollo, o en las personas menores de 50 años (grupo en el que no se suelen hacer cribados) de los países desarrollados, la incidencia del CCR sigue en aumento, señalándose como posible causante la occidentalización de la dieta y el incremento en los índices de sobrepeso y obesidad[6,7].

La supervivencia del paciente con CCR depende, en gran medida, de la etapa o estadio en el que se detecta el cáncer (Figura 2), con tasas de supervivencia a 5 años que oscilan desde aproximadamente el 90% en pacientes diagnosticados en estadio localizado (zona donde comienza el cáncer), a un 10% en pacientes diagnosticados en estadios avanzados donde el tumor se ha diseminado a zonas distantes[8]. Esto hace ver la importancia de una detección temprana del CCR y la necesidad de un mayor conocimiento de los factores ambientales, conductuales y moleculares que puedan influir en la aparición de este tipo de cáncer.

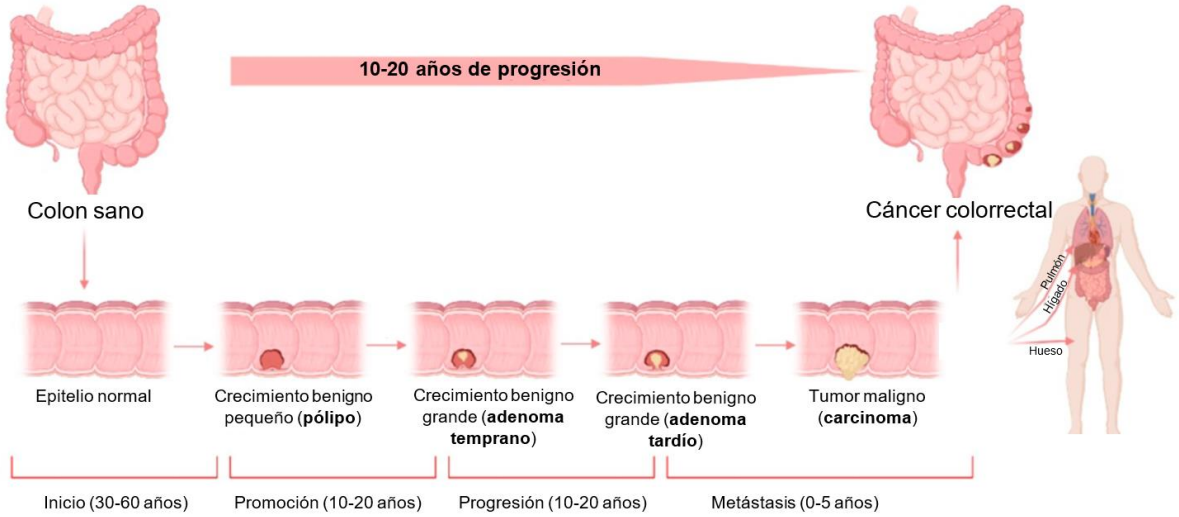


Figura 2: Etapas y desarrollo del CCR. Existen 4 etapas: Inicio, promoción, progresión y metástasis. Figura extraída y modificada de Hossain *et al*[5].

En este contexto, y con el objetivo final de mejorar las posibilidades de supervivencia y el éxito del tratamiento de los pacientes con CCR, en este trabajo se propone explorar y aplicar métodos de aprendizaje automático para analizar qué factores influyen en la aparición del CCR y la forma en que lo hacen. La identificación de estos marcadores, que podrán ser medidos en muestras de sangre, posibilita el desarrollo de una prueba menos invasiva que la colonoscopia, pudiendo mejorar la adherencia al cribado y la detección temprana del CCR.

1.2. Motivación personal

La elección de este trabajo final se basa en mi experiencia profesional previa y el deseo de continuar aportando al desarrollo de soluciones diagnósticas en enfermedades complejas. Durante los últimos 5 años he tenido la oportunidad de investigar una enfermedad rara como es la enfermedad de Ménière, una enfermedad que afecta al oído interno y que parece tener un componente hereditario, tal y como se ha demostrado en diversos estudios de agregación familiar. En esta etapa, junto a mi grupo de investigación en la Universidad de Granada, hemos sido capaces de identificar los que, presumiblemente, son los genes principales de la enfermedad de Ménière familiar: los genes *OTOG*[9], *MYO7A*[10] y *TECTA*[11], codificantes de proteínas clave en la estructura del oído interno y en el proceso de audición.

En este trabajo final de Máster, junto a AMADIX[12], tengo la posibilidad de continuar mi investigación en el campo del cáncer, más en concreto en el diagnóstico precoz del CCR. Aunque la etiología y la patogénesis de las enfermedades oncológicas sean diferentes a las de las enfermedades neurosensoriales, como la enfermedad de Ménière, las metodologías que he venido aplicando estos años en mi puesto de trabajo y el máster en Ciencia de Datos son transferibles a este trabajo final.

El reto de contribuir al conocimiento y mejora de la precisión del diagnóstico precoz del CCR es más que suficiente para afrontar con ganas este trabajo.

1.3. Objetivos del Trabajo

Partiendo del contexto mencionado anteriormente, el objetivo principal que aborda este trabajo final de Máster es la **identificación de marcadores de riesgo clínicos y moleculares útiles para la detección temprana del CCR haciendo uso de algoritmos de aprendizaje automático**.

Además, durante el desarrollo de este proyecto, también se plantean diferentes objetivos secundarios (por orden esperado de consecución):

1. Revisar el estado del arte con el fin de conocer las opciones de preprocesado, análisis y algoritmos adecuados para el análisis de datos de CCR.

2. Realizar un análisis de grupos o *clustering* para identificar subgrupos de pacientes con CCR de acuerdo con sus datos clínicos y moleculares.
3. Evaluar el rendimiento de diferentes algoritmos de aprendizaje automático en la identificación de patrones clínicos y moleculares de la enfermedad. Valorar la necesidad de ingeniería de características.
4. Una vez identificados los factores de riesgo relevantes para el caso de estudio, desarrollar un modelo final para la clasificación de individuos en CCR y controles.
5. Validar el modelo final, comparando sus métricas con marcadores clínicos convencionales y algoritmos previamente publicados.

1.4. Impacto en sostenibilidad, ético-social y de diversidad

Durante el diseño y realización de este trabajo final se trabajará la competencia de Compromiso de Comportamiento Ético y Global (CCEG), contribuyendo a diferentes Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas[13] y abordando tres grandes dimensiones:

1. Sostenibilidad:

El hecho de desarrollar un modelo para predecir el riesgo de desarrollar CCR requiere, *per se,* fomentar la innovación e inversión en infraestructuras tecnológicas. Esto hace que aumente el número de trabajadores en investigación y desarrollo, alineándose con **el ODS 9 – Industria, innovación e infraestructura**.

Por otro lado, el CCR ha sido asociado a diferentes aspectos del estilo de vida de las personas, como la dieta, la actividad física o el consumo de tabaco. Este proyecto tiene como objetivo principal identificar qué factores pueden ser indicativos de riesgo a la hora de desarrollar este tipo de cáncer, por lo que podría ayudar a las personas a tomar decisiones más saludables sobre su estilo de vida. Este punto se ajusta al **ODS 12 – Consumo y producción responsable**.

2. Comportamiento ético y responsabilidad social:

En este apartado, destacamos el papel de este proyecto en el **ODS 3 – Buena salud y bienestar**. El CCR es un problema importante a nivel de salud pública. Por tanto, el desarrollo de un algoritmo que identifique el riesgo de desarrollar este cáncer en personas asintomáticas permitiría un diagnóstico precoz y una intervención preventiva. No obstante, debemos ser cuidadosos a la hora de interpretar los resultados: si el modelo generado no es adecuado, se corre el riesgo de sobreestimar la cifra de casos positivos y llevar a cabo un seguimiento de pacientes costoso e innecesario. Para evitar esto, tendremos que procurar reducir el número de falsos positivos del modelo predictivo.

3. Diversidad y derechos humanos:

Para llevar a cabo este proyecto será necesaria la colaboración entre personal científico, médico y diferentes compañías biotecnológicas, promoviendo el intercambio de conocimientos que permita realizar un modelo predictivo efectivo. Este hecho se alinea con el **ODS 17 – Alianzas para los objetivos**.

Por último, y de acuerdo con los datos epidemiológicos que apuntan que el CCR es más frecuente en personas mayores o personas con historia familiar, este proyecto persigue el objetivo **ODS 10 – Reducción de desigualdades**. Los resultados de este trabajo pueden facilitar el cribado y cuidado de las personas en alto riesgo de desarrollar CCR. Sin embargo, cabe destacar que este proyecto se realizará con muestras provenientes de pacientes europeos y que, por tanto, lo resultados podrían no ser trasladables a individuos de otras regiones debido a diferentes estilos de vida o diferente trasfondo genético. Para solventar esto, se debería validar el modelo generado con muestras de otras poblaciones.

1.5. Enfoque y método seguido

Para la realización de este Trabajo de Fin de Máster se tomará como base la metodología CRISP-DM (del inglés, *Cross-Industry Standard Process for Data Mining*), metodología de referencia para proyectos donde se pretende obtener valor de un conjunto de datos[14]. Esta metodología consta de 6 fases que se adaptarán para alcanzar los objetivos establecidos en el apartado 1.3:

- **Fase 1**, o entendimiento del problema: Como ya se ha mencionado en apartados anteriores, y se ampliará en el Capítulo 2 – Estado del arte, el CCR es el segundo

tipo de cáncer más mortal a nivel mundial. La etapa en la que se produce su diagnóstico, basado principalmente en la colonoscopia, resulta crucial para la supervivencia del paciente, siendo mayor en etapas más tempranas. La mayoría de los pacientes no muestran síntomas hasta etapas avanzadas, cuando ya el cáncer se ha extendido. Por tanto, el desarrollo de un algoritmo predictivo basado en marcadores clínicos y moleculares para detectar el CCR en etapas tempranas puede mejorar significativamente la tasa de supervivencia y la calidad de vida de los pacientes.

- **Fase 2**, o entendimiento y preparación de los datos: Los datos que utilizaremos para desarrollar el modelo predictivo provendrán de pacientes con CCR, AAR y controles. En concreto, este set de datos contendrá los valores de hidroximetilación de diferentes *enhancers* (potenciadores) medidos en sangre, así como covariables como el género, raza o edad de cada muestra. En la fase de entendimiento de los datos deberemos comprender las propiedades de los valores y características con las que estamos trabajando, en este caso. Se desarrollará el análisis exploratorio de los datos (*Exploratory Data Analysis*, EDA), teniendo en cuenta la dimensionalidad de nuestro set de datos.

- **Fase 3**, o preparación de los datos. Tras entender los datos y su distribución, se corrige cualquier error encontrado (valores faltantes, extremos) y se preparan para el proceso de modelado. Para ello deberemos estandarizar o normalizar los datos, comprobar la correlación entre variables (multicolinealidad) y seleccionar aquellas que muestren una mayor importancia para la identificación de la variable objetivo (binaria, CCR o control). Para la identificación de estas variables (marcadores de riesgo) se probarán diferentes metodologías, como la eliminación recursiva de características (*Recursive Feature Elimination*, RFE) o la selección de los k valores más altos (Kbest) de acuerdo con diferentes funciones de puntuación (por ejemplo, Chi2). En este punto se valorará la necesidad de aplicar ingeniería de características (o *feature engineering*) para comprobar si la combinación de ciertas características tiene más valor a la hora de predecir la variable objetivo que ambas por separado.

- **Fase 4**, o modelado: Una vez determinados los factores de riesgo que pueden sernos útiles para diferenciar los controles y los pacientes con CCR, pasamos a entrenar

diferentes modelos de aprendizaje automático supervisado con nuestros datos, como Gradient Boosting (GB), Random Forest (RF), Support Vector Machine (SVM) o modelos de ensemble. El ajuste de hiperparámetros de estos algoritmos se realizará mediante validación cruzada de k iteraciones. Tras esto se evalúan los algoritmos desarrollados, determinando su rendimiento y capacidad de diferenciar las diferentes clases de nuestro set de datos. Para ello, se medirán métricas como la precisión, sensibilidad, especificidad, el valor F1, el área bajo la curva (*Area Under the Curve*, AUC) ROC o la razón de verosimilitud entre clases (class likelihood ratios). De manera paralela, se utilizarán modelos no supervisados, como K-Means o agrupamiento jerárquico, para identificar subgrupos de pacientes con CCR.

- **Fase 5**, o evaluación: Tras seleccionar el modelo que ofrezca mejores métricas, se deben evaluar los resultados. Al estar enmarcado en un ámbito médico, los resultados de este deben ser interpretables. Además, en este punto se deberá comparar los resultados con otros métodos ya publicados, para comprobar que estos son novedosos y útiles. Por último, se revisará detenidamente todo el proceso realizado y se decidirá si se pasa a la fase de despliegue o, por el contrario, se harán nuevas iteraciones de fases anteriores.

- **Fase 6**, o despliegue: Por último, llegamos al resultado final del proceso, los marcadores de riesgo relevantes del CCR y, con ellos, la generación de un modelo predictivo eficaz para la detección precoz de esta enfermedad. En este caso, podríamos considerar como despliegue la entrega final de esta memoria, donde se documenta el proceso seguido y los resultados alcanzados.

La Figura 3 resume estas seis fases de la metodología CRISP-DM, recogiendo a su vez las principales tareas en cada una de ellas. Como podemos ver, este es un proceso cíclico basado en los datos de partida. Si al llegar a la fase de evaluación los resultados no son los esperados, se debería volver a la fase 1 para establecer o modificar los objetivos.

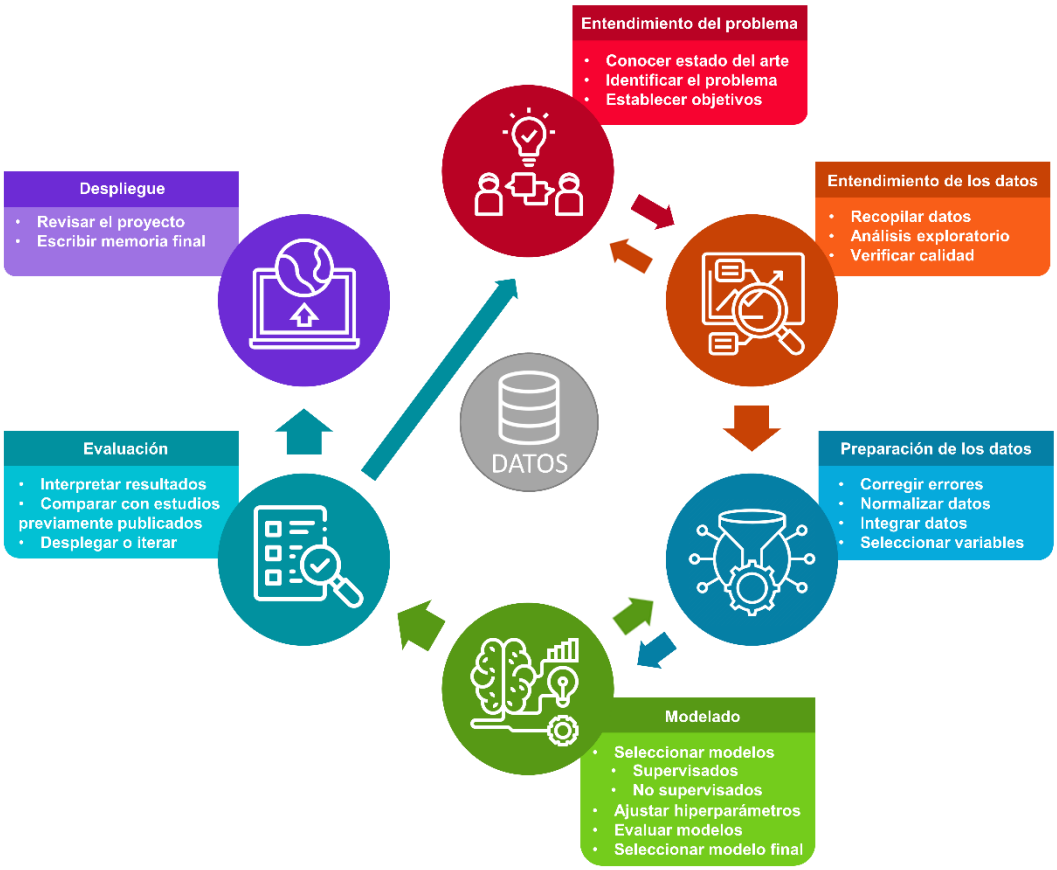


Figura 3: Fases de la metodología CRISP-DM.

1.6. Planificación del trabajo

La planificación de este proyecto se ha realizado teniendo en cuenta las fechas de entrega marcadas por las Pruebas de Evaluación Continua (PEC) y la defensa pública del mismo.

1.6.1. Tareas a realizar

El proyecto tendrá una duración total de 18 semanas, desde marzo a mediados de julio de 2023, y se dividirá en las siguientes tareas:

- **Tarea 1**: Planificación inicial

Esta tarea se corresponde a la entrega de la PEC1, e incluye las siguientes subtareas:

- Propuesta del título del proyecto y definición de palabras clave.
- Resumen de la propuesta.
- Descripción detallada y justificación de la relevancia de la propuesta.

- Motivación personal en el tema del proyecto.
- Definición de objetivos principales y secundarios.
- Descripción de la metodología a seguir.
- Contextualización del proyecto dentro del CCEG.
- Planificación de las fases del proyecto, marcando las fechas clave.

- **Tarea 2**: Análisis del estado del arte

Esta tarea se corresponde a la entrega de la PEC2, incluyendo las siguientes subtareas:

- Búsqueda de referencias que justifiquen los objetivos del proyecto.
- Identificación de trabajos previos con objetivos parecidos.
- Identificación de metodología y técnicas adecuadas para la generación de datos.

- **Tarea 3**: Diseño e implementación del trabajo (DI)

Esta tarea será la más larga del proyecto, ocupando 8 semanas de este, correspondiéndose a la PEC3. En esta tarea se identifican las siguientes subtareas:

- Preprocesamiento de los datos y EDA:
  - Preparación del entorno de trabajo y primer contacto con los sets de datos con los que se llevará a cabo el proyecto.
  - Evaluación de la calidad de los datos (tipo de variables, completitud...)
  - Análisis descriptivo de las variables de nuestro set de datos, calculando media, mediana, desviación estándar y otros parámetros estadísticos.
  - Visualización de datos para explorar patrones y tendencias.
  - Identificación de valores extremos y faltantes.
  - Estandarización o normalización de los datos. Esto dependerá de los modelos y procesos llevados a cabo en fases posteriores.
- Selección de características:
  - Análisis de la correlación entre variables (multicolinealidad).
  - Eliminación de variables con baja varianza (cercana a 0) entre clases.
  - Aplicación de ingeniería de características, si fuese necesario.
  - Evaluación de diferentes metodologías de selección de variables usando modelos de aprendizaje automático: RFE, KBest.

        - Determinación de marcadores candidatos para diferenciar pacientes con CCR y controles.
    - Búsqueda de modelos:
        - Selección de modelos supervisados para la para estratificación de controles y pacientes con CCR.
        - Entrenamiento y validación de los modelos con los conjuntos de entrenamiento y test.
        - Evaluación del rendimiento de los modelos, utilizando validación cruzada, para evaluar la capacidad para generalizar en nuevos datos.
        - Mejora de los modelos mediante ajuste de hiperparámetros.
        - Búsqueda de subgrupos en pacientes con CCR mediante modelos de agrupamiento no supervisados (*clustering*).
    - Interpretación del modelo final:
        - Análisis de la importancia relativa de cada marcador en el modelo final.
        - Interpretación de los resultados del análisis de *clustering*, buscando una explicación clínica.
        - Comparación de los resultados con otros métodos ya publicados.

- **Tarea 4**: Redacción de la memoria

Esta tarea se corresponde a la PEC4.1 y 4.2, los últimos entregables por escrito del proyecto. Se pueden identificar las siguientes subtareas:

- Redacción de la versión final de la memoria del trabajo final de Máster.
    - Corrección de fallos detectados en entregas previas.
    - Identificación de líneas de trabajo futuras.
    - Redactar las conclusiones.

- **Tarea 5**: Defensa del proyecto y defensa pública

Además de la memoria por escrito, se deberán hacer las siguientes subtareas:

- Presentación de un video (15 minutos) con los aspectos clave del proyecto.
- Presentación del proyecto frente al tribunal evaluador, respondiendo sus preguntas.

1.6.2. Cronograma e hitos

La **Tabla 1** muestra el cronograma con las tareas y los entregables que se realizarán durante el desarrollo de este trabajo.

Tareas		Marzo				Abril				Mayo				Junio				Julio			
Planificación inicial			*																		
Análisis del estado del arte					*																
DI1	Preprocesamiento y EDA				*		*														
DI2	Selección de características								*												
DI3	Búsqueda de modelos								*		*										
DI4	Interpretación modelo final												*								
Redacción de la memoria									*		*		*		*		*				
Defensa del proyecto																	*		*		

Tabla 1: Cronograma propuesto para la consecución de objetivos de este proyecto. Los asteriscos (*) marcan las reuniones previstas con cotutores para resolución de dudas. Los bordes de celda marcados con dobles líneas rojas (||) indican las entregas de las PECs. No se incluyen tareas secundarias para facilitar la visualización. DI: Diseño e implementación; EDA: Análisis exploratorio de datos.

En estas tareas y subtareas podemos encontrar algunos hitos que destacan sobre los demás. Estos se presentan de una manera visual en la Figura 4. Los hitos marcados en rojo, que se corresponde a las entregas de las sucesivas PECs y las defensas del proyecto están marcados en rojo, mientras que los hitos extraídos de tareas y subtareas se marcan en negro.

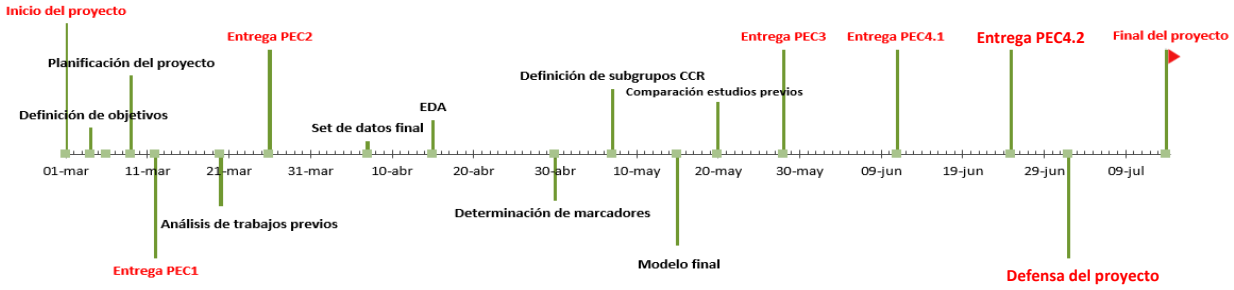


Figura 4: Hitos principales en la escala de tiempo del proyecto.

Los hitos correspondientes a las PECs se presentan con fechas específicas, pudiéndose encontrar en la **Tabla 2**. El resto de los hitos se marcan con una fecha estimada. Cabe destacar que tras la entrega de la PEC3 no se marcan hitos de subtareas ya que, más allá de la propia entrega de la memoria final (PEC4.1 y 4.2) y la defensa del proyecto, no se consideran destacables.

Entregas	Título	Fecha inicio	Fecha entrega
PEC1	Definición y planificación del trabajo final	01/03/2023	12/03/2023
PEC2	Estado del arte	13/03/2023	26/03/2023
PEC3	Diseño e implementación del trabajo	27/03/2023	28/05/2023
PEC4.1	Redacción de la memoria (1ª entrega)	29/05/2023	11/06/2023
PEC4.2	Redacción de la memoria (entrega final)	12/06/2023	25/06/2023
Defensa pública	Defensa pública	03/07/2023	14/07/2023

Tabla 2: Entregas definidas en el Plan Docente de la asignatura.

Por último, en la Tabla 3 se resume cómo se adapta la metodología CRISP-DM detallada en el apartado 1.5 a las PECs propuestas y tareas definidas en este proyecto.

Fase CRISP-DM	Tareas	Entregas
Fase 1	Definición, planificación y estado del arte	PEC1 y PEC2
Fase 2	Preprocesamiento y EDA	PEC3
Fase 3	Selección de características	PEC3
Fase 4	Búsqueda de modelos	PEC3
Fase 5	Interpretación del modelo final	PEC3
Fase 6	Redacción de la memoria	PEC4

Tabla 3: Metodología CRISP-DM. En este planteamiento se considera la redacción y entrega de la memoria como el despliegue final del proyecto.

1.7. Breve sumario de productos obtenidos

Aunque en este proyecto solo se hará una entrega oficial, la memoria final, dentro de esta podemos identificar diferentes productos:

- Metodología o protocolo para el procesamiento de datos clínicos y moleculares enfocado a la generación de modelos de aprendizaje automático.
- Conjunto de datos preprocesado de datos clínicos y moleculares de pacientes con CCR.

- Modelo para la detección precoz del CCR, separando individuos en controles y CCR de acuerdo con los factores de riesgo identificados.
- Modelo para estratificar pacientes con CCR en subgrupos clínicos.

1.8. Breve descripción de otros capítulos de la memoria

En esta sección, se proporciona una visión general de los otros capítulos que se incluyen en la memoria de este trabajo final Máster. Además de este capítulo introductorio, se incluyen otros capítulos que abordan aspectos específicos de la investigación y que contribuyen al logro de los objetivos planteados. A continuación, se describirán brevemente el contenido de estos capítulos y su importancia en el desarrollo del trabajo:

- Capítulo 2: Estado del Arte

En este capítulo se proporciona una revisión completa y actualizada del estado del arte en la detección temprana del CCR, haciendo hincapié en el uso previo de metodologías de aprendizaje automático en este campo. Los conocimientos adquiridos en este capítulo sirven de base para el desarrollo de la metodología y los análisis de los capítulos siguientes.

- Capítulo 3: Diseño e implementación del trabajo

En este capítulo se detalla la metodología utilizada para poder reproducir este proyecto. Se describe la población de estudio, los criterios de inclusión y el tamaño de muestra final, así como el procesado de datos realizado. Se describe el procedimiento de selección de características llevado a cabo para determinar los marcadores de riesgo relevantes de la enfermedad. A partir de estos, se detalla el proceso de generación y evaluación de un modelo para la detección precoz del CCR. Además, se presenta un modelo no supervisado para la estratificación de pacientes con CCR.

- Capítulo 4: Discusión y conclusiones

En este capítulo se presenta una reflexión crítica sobre los resultados obtenidos y se discute su relevancia en el contexto de la detección temprana del CCR. Se presenta una revisión de los objetivos propuestos y alcanzados, así como las posibles implicaciones éticas de la implementación del modelo de predicción en un entorno clínico real. Por último, se resumen las principales conclusiones del trabajo y se proponen posibles líneas de investigación futuras

Capítulo 2: Estado del arte

Tal y como se ha mencionado en el primer capítulo de esta memoria, el CCR es una de las principales causas de muerte por cáncer en todo el mundo. No obstante, los avances en tecnologías como el aprendizaje automático están permitiendo el desarrollo de herramientas de diagnóstico y pronóstico más precisas y eficaces. En este capítulo, profundizaremos en las características más relevantes de este tipo de cáncer, y exploraremos los últimos avances en el uso del aprendizaje automático para mejorar su diagnóstico.

2.1. El cáncer colorrectal

2.1.1. Incidencia y mortalidad

El CCR presenta **prevalencia masculina**, siendo el tercer cáncer más común en hombres tras el cáncer de pulmón y el cáncer de próstata, y con una incidencia de 23,6 casos cada 100.000 hombres y año. En mujeres, el CCR es el segundo más frecuente tras el cáncer de mama, mostrando una incidencia anual de 16,3 casos cada 100.000 mujeres, un 31% más baja que en hombres[15]. El 60% de los casos de CCR ocurren en **países desarrollados**, siendo las zonas del sur (40,6 casos por 100.000 hombres y 24,5 casos por 100.000 mujeres) y norte de Europa (39,2 y 28,8 por cada 100.000, respectivamente) las que presentan una mayor incidencia[16,17]. Esto puede estar provocado por la disponibilidad de mayores recursos para la implementación de programas de cribado, así como por el hecho de que la esperanza de vida es mayor en comparación con países menos desarrollados[18,19].

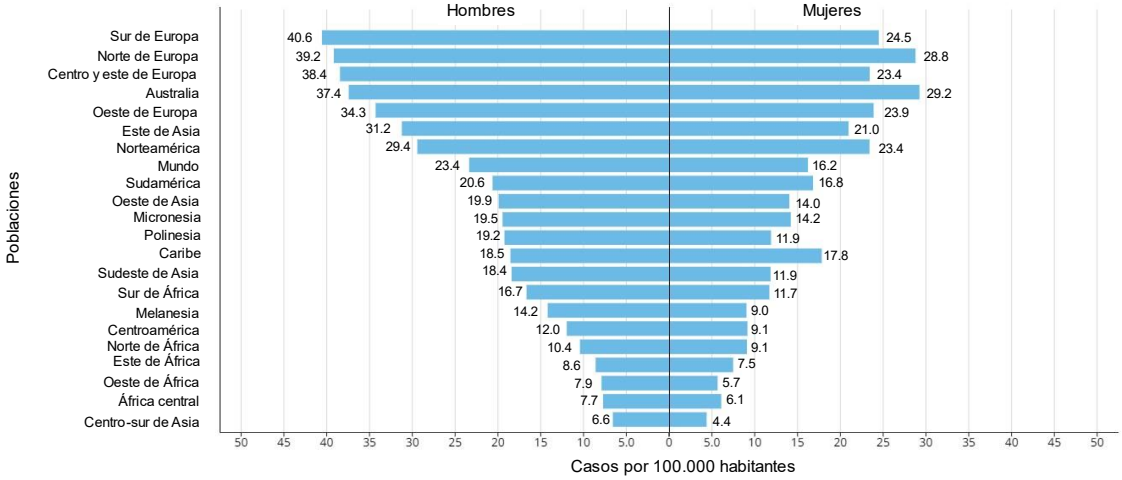


Figura 5: Incidencia ajusta por edad en diferentes regiones del mundo de acuerdo con el género de la población. Figura extraída y modificada de *Global Cancer Observatory*[17].

Centrándonos en la incidencia del CCR en España, en 2020 se identificaron un total 40.441 casos en ambos sexos, siendo el cáncer más diagnosticado por delante del cáncer de próstata (34.613), mama (34.088) o pulmón (29.188). No obstante, si normalizamos esta cifra por cada 100.000 personas, el CCR se posicionaría en tercer lugar con 35,8 casos por cada 100.000 personas, por detrás de los cánceres específicos de género: el cáncer de mama (77,5 casos por cada 100.000) y el cáncer de próstata (70,6 casos por cada 100.000)[17]. Independientemente del género de la persona, esta incidencia es claramente dependiente de la edad, mostrando un incremento acentuado a partir de los 50 años (Figura 6).

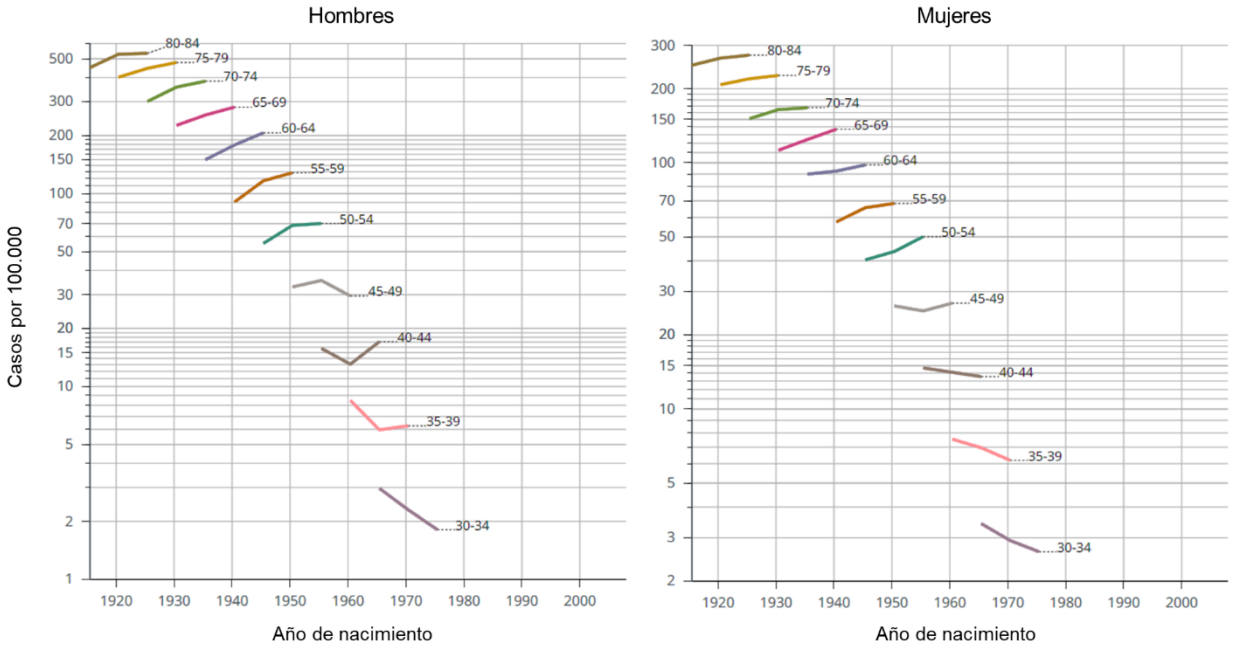


Figura 6: Incidencia ajustada del CCR en mujeres y hombres nacidos en España desde 1914 a 1978. A la derecha de cada línea se muestra la franja de edad de la población a fecha de 2012. Figura extraída y modificada de *Global Cancer Observatory*[17]

La mortalidad del CCR es más homogénea geográficamente que su incidencia. No obstante, existen diferencias entre los países más desarrollados y menos desarrollados. De esta manera, datos de 2020 muestran cómo la mortalidad más alta para ambos sexos se localiza en las regiones central y este de Europa (20,2 fallecimientos por 100.000 hombres y 11,0 casos por 100.000 mujeres), mientras que las zonas central y meridional de Asia muestran el valor más bajo de mortalidad (3,9 y 2,5, respectivamente) (Figura 7). A nivel mundial, en 2020 se registraron un total de 935.173 muertes causadas por CCR, sucediendo 244.824 de ellas en Europa. Esto supone el 9,45% y el 12,52% de todas las muertes causadas por cáncer en el mundo y Europa, respectivamente[17].

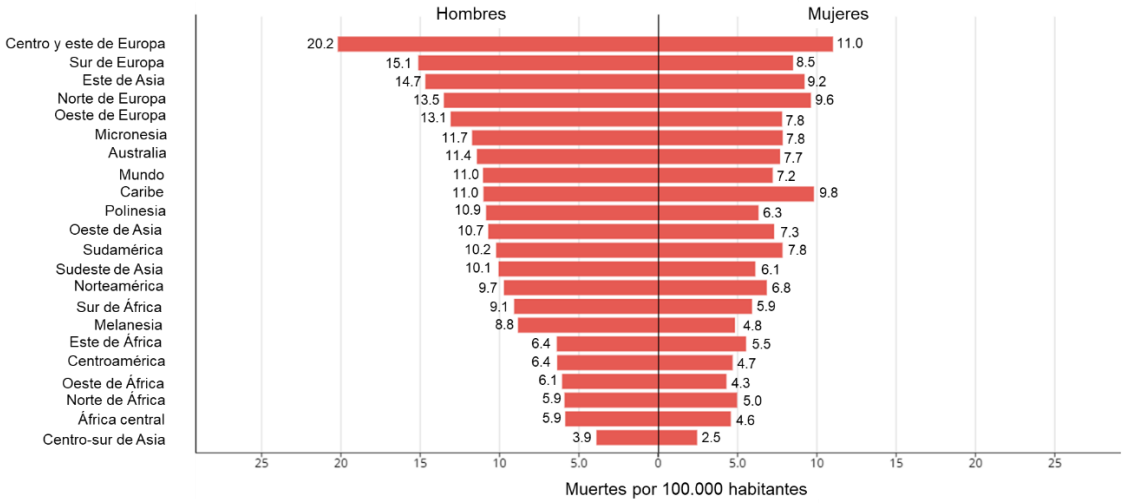


Figura 7: Tasa ajustada de mortalidad en diferentes regiones del mundo de acuerdo con el género de la población. Figura extraída y modificada de *Global Cancer Observatory*[17].

En España, en 2020 se registraron un total de 16.470 muertes (9.640 hombres y 6.830 mujeres) causadas por CCR, suponiendo el 14,57% de todas las muertes asociadas a cáncer[17]. Si observamos los datos desde 1975 a 2021 podemos ver cómo, en ambos sexos, la mortalidad del CCR sigue una tendencia creciente hasta aproximadamente 2010, donde la mortalidad ajustada por edad comienza a disminuir (**Figura 8**)[20]. Aunque es discutible, esto puede ser debido al avance en el conocimiento de los factores de riesgo y la prevención del CCR.

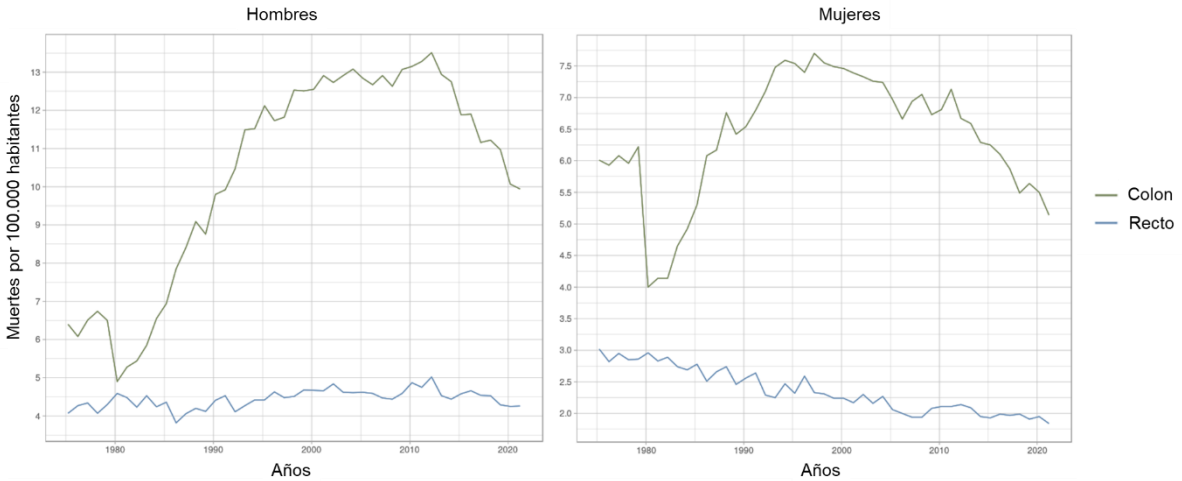


Figura 8: Tendencia de la tasa de mortalidad ajusta por edad del CCR (cáncer de colon y recto) en mujeres y hombres en España desde 1975 a 2021. Imagen extraída y modificada del Servidor Interactivo de Información Epidemiológica (ARIADNA) del Instituto de Salud Carlos III[20].

2.1.2. Factores de riesgo y prevención

Diferentes estudios indican que el origen del CCR se debe a una interacción entre factores genéticos y ambientales. Entre estos factores nos encontramos con **factores no modificables**, como la raza, el sexo o la edad; y **factores modificables** sobre los que se puede intervenir para reducir la incidencia o interrumpir la progresión del CCR (**Figura 9**). Tal y como hemos visto en el punto 2.1.1. de esta memoria, la incidencia del CCR depende en gran medida de la región geográfica en la que esta se mide, siendo mayor en sociedades occidentales. Este aumento en incidencia en países occidentales parece ser debido más al estilo de vida que por aspectos genéticos, tal y como indican diversos estudios llevados a cabo en inmigrantes que adoptan el estilo de vida occidental[21,22].

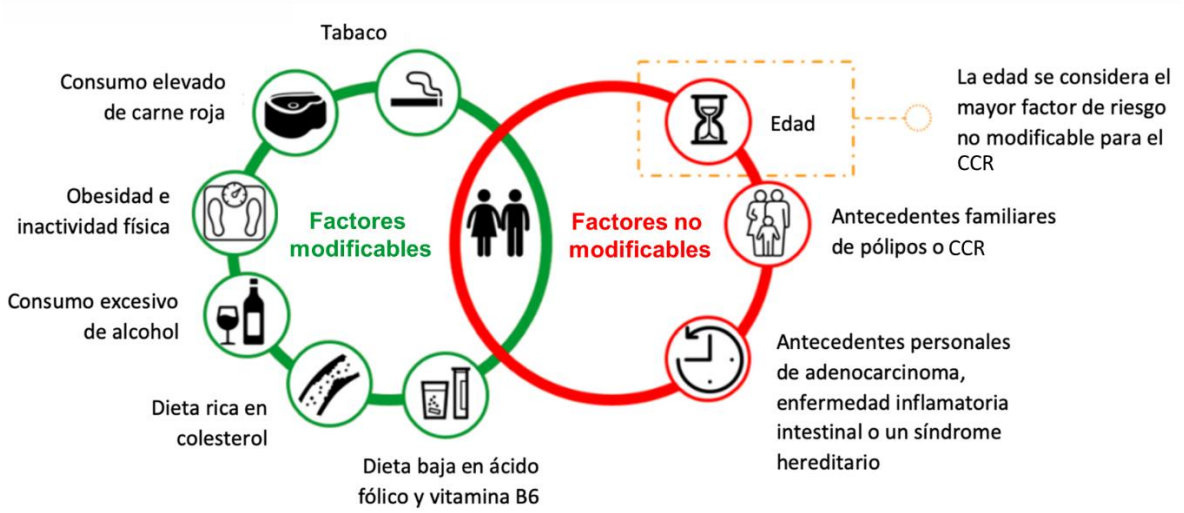


Figura 9: Factores de riesgo modificables y no modificables asociados al CCR. Figura extraída y modificada de https://www.inoncology.es/cancer_CCR.

En este punto es necesario introducir los conceptos del epigenoma y la microbiota intestinal. El **epigenoma** es el conjunto de modificaciones químicas que se producen en nuestro ADN sin producir cambios en su secuencia. Aunque los cambios epigenéticos pueden ser reversibles, la mayoría de ellos se mantienen estables en las diferentes divisiones de nuestras células, alterándolas mediante la modificación de la expresión de ciertos genes. Estos cambios, como la metilación o la hidroximetilación, pueden ser inducidos por factores internos y externos, y pueden llegar a tener efectos similares a las mutaciones patogénicas[23]. Así, la **metilación** es el proceso por el cual se añaden grupos metilos a las citosinas del ADN, y suele tener un **efecto de represión sobre la transcripción genética**. Por su parte, la **hidroximetilación** del ADN, biomarcador en el que nos centraremos en este Trabajo de Fin de Máster, es un **estado intermedio entre el ADN metilado y no metilado, y se**

correlaciona con la activación de la transcripción genética[24]. En la **Figura 10** se presenta un esquema de la metilación y la hidroximetilación para aclarar estos conceptos.

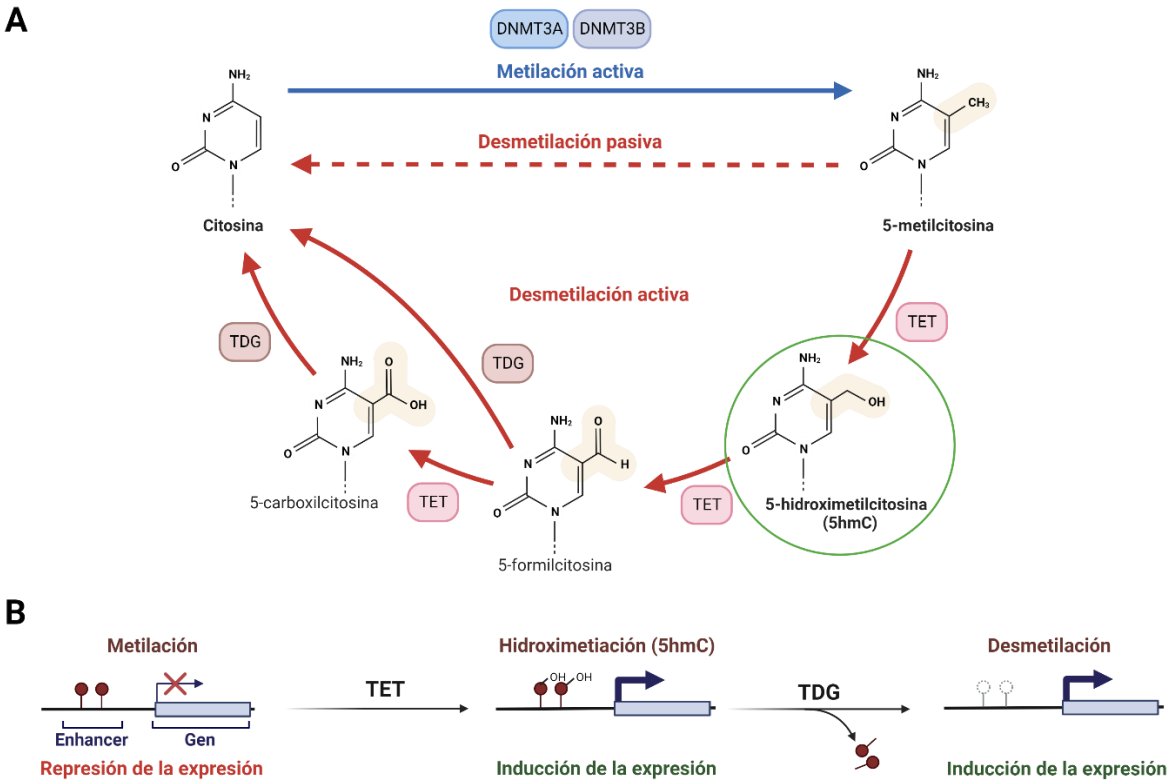


Figura 10: Esquema de la metilación y la desmetilación del ADN. En el **panel A** se observa el proceso completo de metilación activa (llevada a cabo por proteínas DNMT3A y DNMT3B), desmetilación pasiva y desmetilación activa (llevado a cabo por las proteínas TET y TDG) del ADN. Destacado con un círculo verde, el paso intermedio de hidroximetilación que sufre la metilcitosina antes de volver al estado desmetilado. En el **panel B** se esquematiza el efecto de la metilación, la hidroximetilación, y la desmetilación de zonas reguladoras del ADN, como los enhancers, sobre la expresión genética. Figura realizada con Biorender.

Por otro lado, el estilo de vida puede modificar la **microbiota intestinal**, un conjunto de bacterias que se encuentran en nuestro intestino y es esencial para una fisiología normal. Estos cambios en la microbiota pueden llegar a promover la aparición y progresión de enfermedades mediante diferentes procesos de inflamación o metabolitos microbianos[25]. Por tanto, factores asociados al estilo de vida, como la dieta o diferentes factores ambientales, pueden modificar el epigenoma y la microbiota, promoviendo o previniendo el desarrollo del CCR.

De esta manera, existen alimentos que se recomienda **reducir de la dieta** debido a su asociación con el incremento del riesgo de desarrollo del CCR, como las carnes rojas y

procesadas[26]. Por otro lado, alimentos como los productos lácteos[27], productos ricos en fibra y cereales[28], el pescado[29] o el café[30], parecen **reducir el riesgo** de desarrollo de CCR. No obstante, estas recomendaciones se basan en estudios en diferentes poblaciones, y no siempre se encuentra una asociación consistente con el riesgo al CCR. Por tanto, siempre se debe tener en cuenta el contexto individual de cada persona y seguir las indicaciones médicas y nutricionales específicas de cada caso.

Más allá de la dieta, existen **otros comportamientos** del día a día que pueden afectar al desarrollo del CCR. Ejemplos de prácticas poco saludables son:

- Consumo excesivo de **alcohol**: El consumo excesivo de alcohol parece incrementar en un 58% el riesgo de aparición del CCR[31]. No se muestra esta asociación con un consumo moderado[32].
- **Tabaquismo**: Aumentando el riesgo de desarrollo del CCR mediante la modificación de la microbiota[33].
- **Obesidad**: Aunque no está establecido el mecanismo por el cual la obesidad incrementa el riesgo del CCR, la insulinorresistencia y el estado proinflamatorio asociado a la obesidad parecen ser las principales causas[34].

Por otro lado, otros comportamientos parecen estas asociados a un **menor riesgo** de desarrollo del CCR. Entre estos destaca el practicar **ejercicio físico** de manera habitual, con reducciones de hasta el 25% del riesgo de desarrollo de CCR. No obstante, no se ha llegado a una conclusión sobre el tiempo, el tipo, o la intensidad adecuada a la que se debe hacer el ejercicio[35].

2.1.3. Patogénesis

Aunque no se entrará en demasiados detalles en esta memoria, el conocimiento de la etiología y la patogénesis del CCR es esencial para entender los resultados que se puedan obtener de este estudio. Es por ello por lo que en esta sección abordaremos brevemente las principales **vías moleculares** que hacen que la mucosa del colon se transforme en un pólipo

benigno y, posteriormente, en un pólipo maligno (**Figura 11**). Aunque el CCR involucra múltiples y posibles vías moleculares, las **tres principales** son:

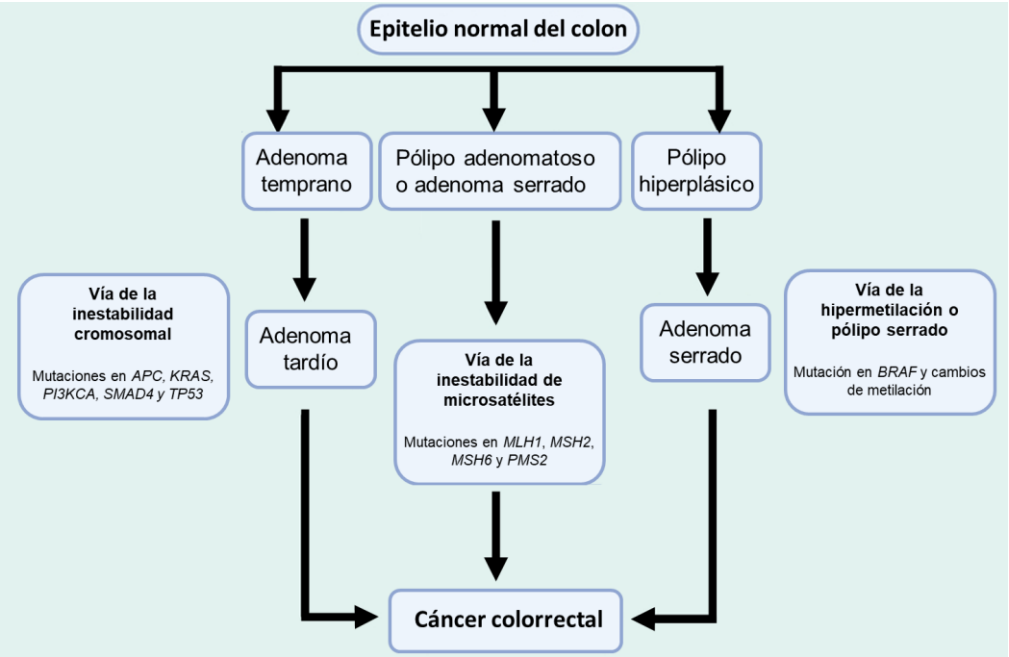


Figura 11: Principales vías moleculares involucradas en el CCR.

- La **vía de la inestabilidad cromosómica**: Esta vía se encuentra en hasta el 85% de los casos con CCR, y se caracteriza por cambios en el número de cromosomas o aberraciones estructurales (ganancia, pérdida o cambio de orden de una parte del cromosoma). Estas alteraciones pueden modificar la expresión de genes que regulan la división celular, pudiendo a su vez iniciar y hacer progresar el CCR[36]. Algunos de los genes que suelen estar afectados por esta inestabilidad y median la tumorogénesis del CCR son *KRAS*, *PI3KCA*, *TP53, SMAD4* o *APC*[37].

- La **vía de la inestabilidad de microsatélites**: Esta vía representa aproximadamente el 15% de los CCR[38]. Los microsatélites son secuencias de ADN repetitivas que se distribuyen por todo el genoma, y que son bastante propensas a acumular mutaciones si el sistema encargado de corregirlas no funciona correctamente. Por tanto, la acumulación de mutaciones en estos microsatélites indica que este sistema de corrección es deficiente. Los genes involucrados en este sistema son los genes *MLH1*, *MSH2*, *MSH6* y *PMS2*, que además están involucrados en la muerte celular y el control del ciclo celular. Modificaciones que conlleven la pérdida de función de cualquiera de estos 4 genes conllevará la acumulación de mutaciones y se asocian

a la enfermedad de Lynch, una enfermedad genética que aumenta el riesgo de padecer diferentes tipos de cáncer, incluyendo el CCR[39].

- La **vía del pólipo serrado o hipermetiladora**: La metilación es una modificación química que utilizan células sobre el ADN para silenciar o inhibir la expresión de los genes. Esta vía se caracteriza por la hipermetilación de una gran cantidad de genes, resultando en una pérdida de la expresión génica. Existen dos tipos de tumores según el grado de metilación: metilación alta, caracterizados por hipermetilación del gen *MLH1*, hipermetilación alta del gen *BRAF*, y una tasa baja de mutaciones en *TP53*; y metilación baja, caracterizados por ausencia de inestabilidad de microsatélites, presencia de mutaciones en *KRAS* y *TP53*, y la ausencia de hipermetilación en *MLH1*[40]. Estos cambios hacen que se inicie y progrese el CCR.

2.1.4. Diagnóstico

Para valorar los posibles resultados que alcancemos en este trabajo, será necesario conocer los principales métodos de diagnósticos del CCR. En la actualidad, la técnica de referencia o *gold standard* para la detección y diagnóstico precoz del CCR es la **colonoscopia**, tanto por su capacidad de diagnóstico como por la posibilidad de eliminar adenomas. No obstante, se trata de una prueba invasiva e incómoda para el paciente, lo que provoca que alguno rechace someterse al proceso[41,42]. La combinación de este y otros factores, como el coste o el bajo rendimiento diagnóstico[43], hacen que **sean necesarias otras técnicas complementarias menos invasivas** que permitan reducir la utilización de la colonoscopia como primera vía.

El CCR, normalmente, se desarrolla lentamente a medida que pasan los años, pudiéndose prevenir si los adenomas se detectan antes de que se vuelvan malignos. Si el CCR se detecta en sus fases iniciales la tasa de supervivencia ronda el 90% de pacientes, no obstante, en fases tardías donde el tumor ha invadido otros órganos y tejidos esta tasa disminuye al 10%[8]. Por tanto, el hecho de que, aproximadamente, el 15-20% de pacientes

con CCR tenga metástasis al efectuar el diagnóstico[44], hace ver la importancia del desarrollo de otras estrategias de cribado y diagnóstico precoz en este tipo de cáncer (**Figura 12**).

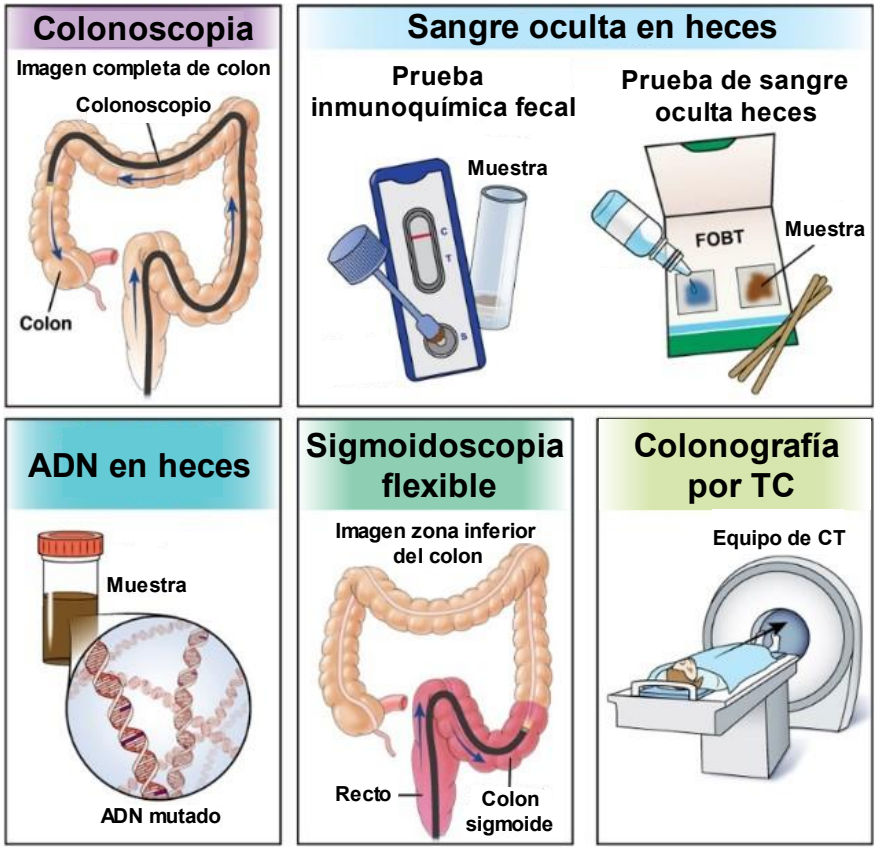


Figura 12: Opciones disponibles para la detección del CCR. Imagen extraída y modifica de patient.gastro.org/colorectal-cancer-screening-options

La **detección de sangre oculta en heces** (SOH) y la **sigmoidoscopia flexible** suelen ser las pruebas realizadas previas a la colonoscopia. El CCR precoz y los adenomas de alto riesgo (AAR) producen pequeñas perdidas de sangre. Existen dos tipos de pruebas de SOH, una de ellas hace uso de una reacción química (**SOH-química**) para detectar los grupos hemo, responsables del color rojo de la hemoglobina y necesarios para transportar el oxígeno; y las otras pruebas detectan específicamente la hemoglobina humana mediante anticuerpos (**SOH-inmunológica**). A lo largo del tiempo, la detección de SOH-química ha mostrado una baja sensibilidad para la detección del CCR precoz, lo que motivó el desarrollo de la detección de SOH-inmunológica[45,46]. Una revisión sistemática sobre diferentes tipos de técnicas de detección de SOH-inmunológica indicó que la sensibilidad y especificidad con un intervalo de confianza (IC) del 95% de estas técnicas eran de 0,79 (IC 95%, 0,69-0,86) y 0,94 (IC 95%, 0,92-0,95), respectivamente[47]. No obstante, su precisión para detectar AAR

es baja[48]. Por otro lado, la **sigmoidoscopia flexible**, una técnica menos invasiva que la colonoscopia que permite examinar el recto y el colon sigmoide, alcanza una sensibilidad del 95% en la detección de AAR y CCR. Su especificidad se estima en un 87%[49].

Existen otras alternativas, como la **colonografía por tomografía computarizada** (TC), que utiliza un sistema especializado para tomar imágenes en 2 y 3 dimensiones del colon. Se suele utilizar cuando la colonoscopia no ha sido efectiva debido a problemas técnicos o anatómicos. Se estima su sensibilidad en un intervalo entre el 67% al 94%, y su especificidad del 86% al 98%[50]. Una técnica más novedosa que las anteriores involucra la utilización de una cápsula con cámaras (**cápsula endoscópica colónica**) que toma fotografías del colon y el recto. La sensibilidad y especificidad para la detección de pólipos de esta prueba es del 85-88% y 97-99%, respectivamente[51].

Por último, de modo complementario a las técnicas anteriores, las pruebas de diagnóstico y pronóstico del CCR se pueden beneficiar del uso de **biomarcadores** derivados de tejido o sangre. De una manera amplia, y aunque existen otros tipos de clasificaciones, estos biomarcadores se pueden clasificar en tres categorías de acuerdo con su utilidad clínica: biomarcadores de **pronóstico**, de **diagnóstico** y **predictivos**. El biomarcador más utilizado es el antígeno carcinoembrionario (CEA, por sus siglas en inglés). Este biomarcador, que se puede detectar en sangre, tiene una buena especificidad para detectar el CCR, pero su especificidad es de solo el 40-60%. Se suele encontrar en fases avanzadas del CCR, no siendo tan común en fases iniciales, por lo que no parece un biomarcador ideal para la detección temprana del CCR[52]. No obstante, los niveles de CEA se suelen utilizar para realizar el seguimiento de pacientes ya diagnosticados, evaluando la progresión del cáncer.

Dado que la lista de biomarcadores es extensa (Anexo 1), en este trabajo nos centraremos en alguno de los posibles biomarcadores útiles para el diagóstico precoz del CCR derivados de tejido y sangre:

- **Citoqueratinas**: Son proteínas expresadas en las células epiteliales, utilizándose especialmente las citoqueratinas 20 (CK20) y 7 (CK7). Las células tumorales son positivas (expresan) para CK20 y negativas (no expresan) para CK7. La detección de expresión de CK20 en células derivadas de tejidos o heces puede ser útil para el diagnóstico precoz del CCR. Además, a medida que el tumor avanza, las células

positivas para ambas citoqueratinas aumentan en número, por lo que puede ser un marcador diferencial del progreso del CCR[53].

- **Mucinas**: Los niveles de otras proteínas epiteliales, las mucinas 2 (MUC2) y 5 (MUC5), están asociados a la progresión del tumor, y podrían ser útiles para el diagóstico precoz y caracterización del CCR[54].
- **ADN tumoral circulante** (ADNtc): Aunque el ADN libre circulante es un biomarcador que se encuentra de manera natural en la sangre, los niveles de ADNtc encontrados en sangre parecen correlacionar positivamente con la fase en la que se encuentra el CCR[55,56]. Además del nivel de ADNtc, parece que el tamaño de los fragmentos de ADN en pacientes con CCR y controles también cambia, siendo más largos en los pacientes con cáncer. Métodos basados en el tamaño del ADNtc ofrecen una sensibilidad del 73,08% y una especificidad del 97,27%[57].
- **micro-ARNs** (miR): Estas pequeñas moléculas de ARN tienen la capacidad de regular la expresión de otros genes, y la expresión de ciertos miR está correlacionada con varios tipos de cáncer. Aunque la lista de miR asociados al CCR no deja de crecer, podemos destacar el miR-21 como biomarcador de diagnóstico del CCR, mostrando una especificidad de 83% y una sensibilidad del 77%[58].

Esto son solo algunos ejemplos de los biomarcadores disponibles para el diagnóstico del CCR. Como se ha comentado anteriormente, la lista completa de biomarcadores es inabordable en este trabajo final, por ello recomendamos y utilizaremos como lectura complementaria a) el capítulo *Biomarkers as Putative Therapeutic Targets in Colorectal Cancer* del libro *Colon Cancer Diagnosis and Therapy*, por Pal *et al*.[59]; y b) la revisión sistemática *Improving diagnosis, prognosis and prediction by using biomarkers in CRC patients* de Nikolouzakis *et al*.[60]

2.2. Aprendizaje automático

Dada la magnitud de los datos que se pueden extraer de un paciente con CCR, el aprendizaje automático se ha convertido en una herramienta prometedora para ayudar en su diagnóstico, tratamiento y pronóstico (**Figura 13**). Estos algoritmos pueden analizar grandes cantidades de datos, identificar patrones y aprender de ellos para detectar los factores de riesgo de cáncer y mejorar la precisión del diagnóstico del CCR. En este apartado se hará hincapié en aquellos algoritmos que hacen uso de datos demográficos (edad, genero, raza), historial médico (comorbilidades), resultados de análisis de sangre, historial

familiar y factores conductuales, ya que será con datos de este tipo con los que se realizará este trabajo. Por el contrario, y aunque las imágenes médicas son una herramienta valiosa para la detección y diagnóstico del CCR[61], no se entrará a discutir aquellos modelos que hagan uso de imágenes, ya que se pretende explorar otros enfoques para mejorar la precisión del diagnóstico en esta enfermedad.

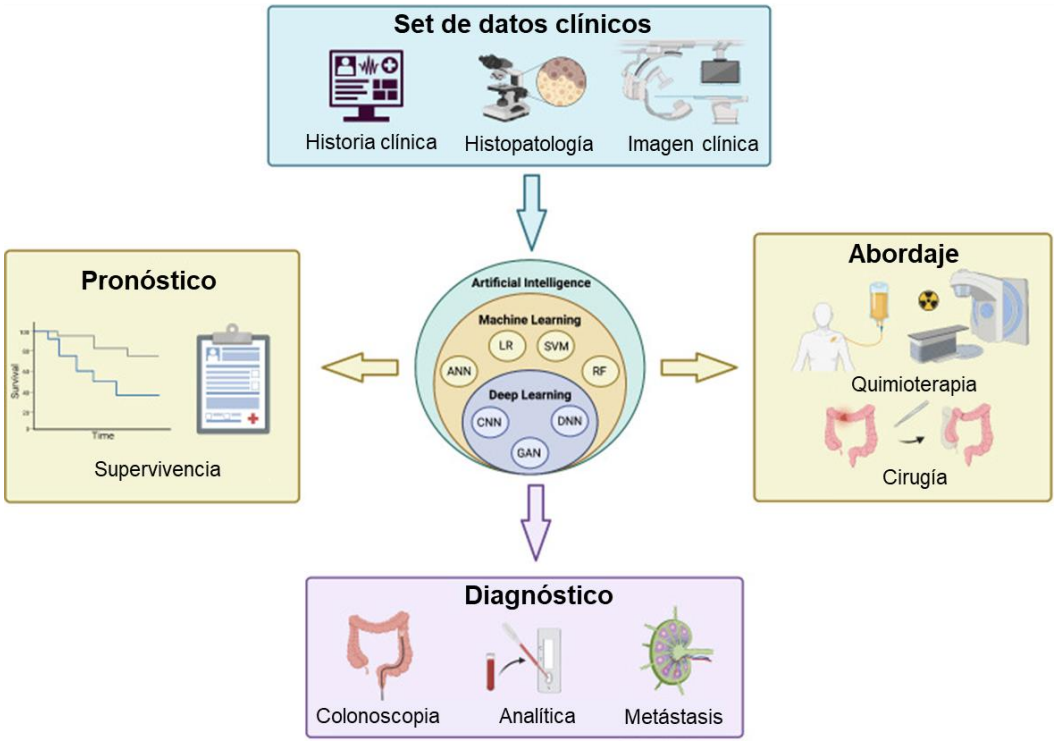


Figura 13: Resumen de las posibles aplicaciones de los algoritmos de inteligencia artificial en el CCR. A partir de un set de datos clínicos, la construcción de un modelo de aprendizaje automático puede ser útil para el diagnóstico, el pronóstico y el abordaje del CCR. Figura extraída y modifica de Mansur *et al*.[62]

2.2.1. Conceptos básicos

Antes de entrar en detalles sobre las aplicaciones específicas en el campo del CCR, debemos refrescar una serie de conceptos básicos relacionados con el aprendizaje automático. Para ello, utilizaremos como referencia el libro de Gironés *et al.* "Minería de datos. Modelos y Algoritmos".[63]

El aprendizaje automático es una rama de la inteligencia artificial cuyo objetivo principal es la generación de algoritmos que sean capaces **de identificar y generalizar patrones a partir de una serie de ejemplos recibidos para realizar pronósticos**. Por tanto, estos algoritmos permiten a las computadoras reconocer estos patrones de manera autónoma, y es por esto por lo que se habla de aprendizaje automático o *machine learning*. Dentro de

esta definición tan amplia, podemos identificar, principalmente, **dos tipos** de algoritmos de aprendizaje automático **de acuerdo con los datos que reciben como entrada**: aprendizaje supervisado y no supervisado.

Los **algoritmos de aprendizaje supervisado** reciben como entrada **datos etiquetados**, es decir, cada ejemplo (registro) que recibe el algoritmo viene asociado a una categoría. Por ejemplo, en un set de datos de pacientes con CCR donde se recogen valores clínicos, las etiquetas de los registros podrían ser “sano” o “cáncer”. El aprendizaje supervisado sigue **tres etapas fundamentales**: 1) el modelo se **entrena** a partir de los datos de entrada, 2) el modelo se **evalúa** y se ajustan sus parámetros (hiperparámetros), 3) el modelo se usa para realizar **predicciones** a partir de nuevos datos (no utilizados en la primera etapa). De esta manera, el algoritmo aprende qué características están asociadas a cada una de las etiquetas. Este tipo de algoritmo cumple **dos funciones principales**:

- **Clasificación:** la variable pronosticada es **categórica**. Una vez que el algoritmo de clasificación aprende las características asociadas a cada categoría, se les asigna la etiqueta correcta a nuevos registros en función de esas características. Los algoritmos de clasificación más frecuentes son SVM, los árboles de decisión, k-vecinos más cercanos (KNN), naïve Bayes (NB) o RF.
- **Regresión:** la variable pronosticada es **numérica**. En este caso las etiquetas del set de datos con el que aprende el algoritmo no son categorías, sino valores numéricos. Por tanto, al recibir nuevos registros, se utiliza el algoritmo de aprendizaje supervisado para predecir el valor numérico asociado en función de las características del nuevo registro. Un ejemplo aplicado al CCR podría ser la predicción del tamaño del tumor (valor numérico) a partir de datos clínicos (características) del paciente. Los algoritmos de regresión más frecuentes son la regresión lineal y la regresión logística (LR). No obstante, algunos algoritmos de clasificación como RF o KNN pueden ser utilizados igualmente para tareas de regresión.

Por su parte, los **algoritmos de aprendizaje no supervisado** reciben **datos no etiquetados**. Estos algoritmos se suelen utilizar para analizar y descubrir patrones dentro de los datos, dando como resultado diferentes **agrupaciones (*clústeres*)** de acuerdo con las similitudes y diferencias de los registros. En el contexto del CCR, un algoritmo de aprendizaje no supervisado sería la agrupación (*clustering*) de pacientes de acuerdo con

datos de expresión de diferentes genes. Ejemplos de algoritmos no supervisados de *clustering* pueden ser K-means o el *clustering* jerárquico. No obstante, además del *clustering*, los algoritmos de aprendizaje automático también son utilizados para:

- **Reglas de asociación:** métodos basados en reglas para encontrar relaciones entre las variables de un conjunto de datos.
- **Reducción de la dimensionalidad:** técnicas que reducen el número de características (dimensiones) del set de datos. Dentro de estas técnicas encontramos el **análisis de componentes principales** (*principal component analysis*, PCA), un método que utiliza una transformación lineal de las características para crear una nueva representación de los datos, lo que genera un conjunto de "componentes principales" no correlacionadas.

La **evaluación** de la calidad o bondad de los algoritmos supervisados de **clasificación** se basa en la comparación de las predicciones que genera el modelo y el valor verdadero del conjunto de datos. Los errores y aciertos del modelo se pueden expresar en una matriz de confusión que permite calcular diferentes métricas, como la exactitud (accuracy), la precisión, la sensibilidad (*recall* o *sensistivity*), la especificidad (*specificity*) y el valor F1. Por otro lado, también se puede caracterizar el rendimiento del modelo midiendo el AUC-ROC, que refleja la tasa de verdaderos positivos (true positive, TP) y falsos positivos (false positive, FP); o el área bajo las curvas de precisión/sensibilidad (AUC-PR), una gráfica que representa la precisión del modelo frente a la sensibilidad o *recall*. Por otro lado, la evaluación de los métodos supervisados de **regresión** sigue la misma lógica que los modelos de clasificación, pero en este caso, en lugar de valores categóricos se compararán los valores numéricos predichos. En este caso, tenemos métricas como el error absoluto medio (MAE), el error cuadrático medio (MSE) y la raíz cuadrada del MSE (RMSE). Por último, en el caso de los modelos no supervisados de ***clustering***, al no disponer de etiquetas en los datos, no podemos saber si la predicción es correcta o no. Por este motivo, la evaluación de este tipo de modelos se hace por métricas de calidad por partición, como el diámetro o la separación entre *clústeres*; métricas de calidad general, como el índice de Dunn o el índice C; o por métricas de calidad externas, que se pueden aplicar cuando disponemos de etiquetas para los datos. Un ejemplo de este tipo de métrica es el índice de Rand.

Las fórmulas de las métricas mencionadas y su interpretación se pueden encontrar en la **Tabla 4**.

Métrica	Fórmula	Interpretación
Exactitud	$\frac{TP + TN}{TP + TN + FP + FN}$	Proporción de predicciones correctas de una clase frente al total.
Precisión	$\frac{TP}{TP + FP}$	Proporción de predicciones positivas correctamente clasificadas.
Sensibilidad	$\frac{TP}{TP + FN}$	Efectividad del modelo para identificar la clase positiva.
Especificidad	$\frac{TN}{TN + FP}$	Efectividad del modelo para identificar la clase negativa.
F1	$\frac{2\ x\ TP}{2\ x\ TP + FP + FN}$	Métrica que combina la sensibilidad y la especificidad de un modelo.
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y_i}\rvert$	Promedio de la diferencia absoluta entre la predicción del modelo y el valor objetivo
MSE	$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$	Media de la diferencia cuadrática entre la predicción del modelo y el valor objetivo
RMSE	$\sqrt{MSE}$	Raíz cuadrada del MSE. Se expresa en la misma escala que los datos originales.
Diámetro	$\max \delta(x_1, x_2) \mid x_1, x_2 \epsilon\ p$	Distancia (δ) máxima entre dos instancias (x_1 y x_2) de un *clúster*.
Separación	$min\ \delta(x_1, x_2) \mid x_1 \epsilon\ d, x_2 \epsilon\ D - d$	Distancia mínima entre un *clúster* (d) y los demás (D – d).
Índice de Dunn	$\frac{\min sep_\delta(d)}{\max diam_\delta(d)} \mid d \epsilon\ D$	Mide la calidad del modelo combinando la separación mínima y diámetro máximo de los *clústeres*.
Índice de Rand	$\frac{TP + TN}{TP + TN + FP + FN}$	Proporción de pares de elementos correctamente clasificados del segundo *clúster* respecto al primero.

Tabla 4: Métricas frecuentes para la evaluación de algoritmos de aprendizaje automático. TP: Verdaderos positivos; TN: Verdaderos negativos; FP: Falsos positivos; TF: Verdaderos negativos.

2.2.2. Aproximaciones al diagnóstico del cáncer colorrectal

El interés del desarrollo de diferentes modelos de aprendizaje automático para el diagnóstico del CCR radica en la **posibilidad de mejorar la adherencia a los cribados** y, por otro lado, **disminuir las posibles complicaciones y el coste de las pruebas actuales** (SOH-inmune y colonoscopia). Además, estos modelos permitirían la identificación de los pacientes en una etapa temprana donde el CCR es altamente tratable.

En la actualidad, existen múltiples artículos publicados donde se describen algoritmos cuyo objetivo principal es la estratificación de pacientes de acuerdo con el riesgo de sufrir CCR o el diagnóstico precoz de los mismos[62]. Por lo tanto, a la hora de diseñar la metodología e interpretar los resultados de nuestro trabajo, debemos conocer el enfoque y el rendimiento alcanzado por estos estudios. Con este objetivo, haremos uso de los resultados de dos revisiones, una de ellas sistemática, publicadas en los años 2022 y 2023, sobre el uso del aprendizaje automático en datos procedentes de pacientes con CCR[62,64].

La revisión sistemática publicada en 2022 por Kennion *et al*. encontró un total de 14 estudios que hacían uso de modelos de aprendizaje automático para la predicción del riesgo de desarrollar CCR y la estratificación de pacientes[64]. El tamaño de muestra utilizado por estos estudios fue muy variable, contando con un mínimo de 70.000 registros hasta más de 2 millones de registros. Cabe destacar que hasta 5 de estos estudios utilizan como referencia para el desarrollo de sus modelos el artículo publicado por Kinar *et al*. en 2016, donde desarrollaron y validaron un modelo predictivo para la detección del CCR a través de datos de hemograma completo[65].

Entre estos 14 estudios, los modelos de aprendizaje automático más utilizados fueron *RF* (9), redes neuronales (6) y LR (5). No obstante, la estrategia más frecuente entre los estudios es la de **comparar diferentes modelos para encontrar el más óptimo**. Por tanto, además de los 3 modelos anteriores, se utilizaron modelos como SVM, NB, árbol de decisión y análisis discriminante lineal. En cuanto a las características (*features*) utilizadas para generar el modelo, el número variaba entre 4 y 50 características, incluyendo características demográficas (como edad y género), historial médico, parámetros obtenidos de muestra de sangre, medicación, y factores conductuales. En la revisión de estos artículos no se encontró una correlación entre el tipo de características utilizadas y el modelo utilizado.

A la hora de evaluar los modelos generados, la métrica más utilizada fue **AUC-ROC**, aunque también se hace uso de otras como sensibilidad, especificidad y el número de predicciones positivas y negativas. Todos los modelos obtuvieron un valor mayor a 0.73 para AUC-ROC, siendo los modelos desarrollados por Hoogendoorn *et al*.[66] y Kop *et al*.[67] los que mostraron un mayor valor para esta métrica, con valores de 0,896 y 0,891, respectivamente. En el caso de Hoogendoorn *et al*. se desarrolló un modelo de LR a partir de 17.095 registros, 900 de ellos de CCR, e incluyendo un total de 23 características sobre edad, sexo y datos de hemograma completo. Por otro lado, Kop *et al.* usaron un modelo LR a partir de 263.879 registros, 1.292 de ellos con CCR, e incluyendo un total de 50 características sobre demografía, historial médico, medicación actual o valores procedentes de analíticas de sangre.

De la revisión realizada por Mansur *et al*.[62] podemos extraer dos estudios adicionales no incluidos en la revisión sistemática anterior. El primero desarrolló un modelo de aprendizaje automático utilizando los niveles de diferentes proteínas encontradas en muestras de sangre para detectar el CCR. Utilizaron una muestra de 263 registros (50 con CCR) para identificar los biomarcadores útiles para esta tarea. El modelo más óptimo, con un valor de 0,86 para AUC-ROC, fue un modelo RF basado en los datos de 5 proteínas medibles en sangre: LRG1, EGFR, ITIH4, hemopexina y la superóxido dismutasa 3[68]. El segundo estudio, que perseguía el mismo objetivo que el primero, contaba con una muestra de 362 registros (89 con AAR y 163 con CCR) para desarrollar un modelo predictivo utilizando los niveles en sangre de diferentes glucanos[69]. El modelo seleccionado en este caso fue un modelo de ensemble (SVM, LR y *Logistic Model Tree*) que obtuvo un valor de 0,8 para AUC-ROC en la diferenciación de controles, pacientes con AAR y pacientes con CCR.

Además, aunque no son discutidos en ninguna de las anteriores revisiones, los modelos no supervisados son de especial interés para identificar subgrupos de pacientes con CCR y factores de riesgo parecidos. Ejemplo de esto lo podemos encontrar en el artículo publicado recientemente por Florensa *et al.*[70], donde detectaron 5 subgrupos en una muestra de 1083 pacientes con CCR haciendo uso de K-means y factores de riesgo como el tabaquismo, aspectos sociodemográficos o el índice de masa corporal de los pacientes.

Los resultados mostrados por estos y otros artículos auguran un futuro prometedor para los test diagnósticos basados en aprendizaje automático y marcadores obtenidos en pruebas menos invasivas que las utilizadas actualmente en los cribados de CCR.

2.3. Retos dentro del contexto del trabajo final

Tras revisar el estado del arte, hemos podido comprobar como el CCR representa uno de los principales desafíos del sistema de salud a nivel mundial en la actualidad. Además, el uso del aprendizaje automático se presenta como una herramienta útil para abordar este problema, permitiendo la detección temprana de factores de riesgo que podrían indicar la presencia de la enfermedad.

Teniendo como base lo anterior, dentro del contexto de este trabajo final se presentan, principalmente, dos retos que deberán ser abordados cuando trabajemos con nuestro set de datos:

- Aplicar algoritmos supervisados para identificar los principales biomarcadores que permiten discernir entre pacientes con CCR y controles. De igual manera, la existencia de modelos predictivos ya publicados nos enfrenta al reto de mejorar las métricas actuales haciendo uso de un modelo interpretable (sin caja negra) para aplicación clínica. Deberemos analizar los resultados y comprobar si los marcadores de riesgo detectados se replican con alguno de los estudios ya publicados.

- Aplicar algoritmos no supervisados para identificar subgrupos de pacientes con CCR de acuerdo con la presencia de diferentes factores de riesgo, y valorar su utilidad para la identificación temprana del CCR.

La aplicación y desarrollo de estos algoritmos implicará la colaboración con otros profesionales, promoviendo el ODS 17. Además, se consiguen superar ambos retos, avanzaremos hacia el desarrollo de un diagnóstico precoz menos invasivo y más económico para el CCR y, por tanto, nos acercaremos al ODS 3 y el ODS 10.

Capítulo 3: Diseño e implementación del trabajo

En este capítulo de la memoria se tratarán, en más profundidad, los aspectos de la metodología esbozados en capítulos anteriores. Así mismo, se presentarán y analizarán los resultados obtenidos usando el protocolo propuesto.

3.1. Metodología

De manera general, este proyecto se ha desarrollado en Python 3.10, utilizando como entorno de desarrollo PyCharm[71]. Además de las librerías estándar de Python, se han usado otras librerías específicas para alcanzar los objetivos de este proyecto, por lo que estas se detallaran en cada uno de los subapartados de la metodología. El equipo con el que se realiza la práctica es un ordenador personal cuyas características principales son:

- Sistema operativo: Ubuntu 22.04.2 LTS (64 bit)
- Procesador: Intel Core i7-12700KF 12th Gen (12 núcleos, 20 hilos)
- Memoria RAM: 32 GB (2x16GB). DDR4 3600MHz.
- Tarjeta gráfica: GeForce RTX 3060 Ti

La metodología la podemos dividir, siguiendo la lógica de un proyecto orientado a datos, en diferentes fases: 1) obtención del set de datos, 2) EDA, 3) preprocesado y limpieza de datos, 4) aplicación de modelos supervisados, 5) aplicación de modelos no supervisados, y 6) evaluación y comparación del rendimiento de los diferentes modelos.

3.1.1. Set de datos

El set de datos utilizado en este proyecto es un set de datos público procedente del estudio FORESEE, llevado a cabo por Cambridge Epigenetix Ltd, y descrito en el artículo publicado en *Scientific Reports* titulado "*Hydroxymethylation profile of cell-free DNA is a biomarker for early colorectal cancer*"[72]. El set de datos es accesible desde Zenodo[73].

Este set de datos contiene registros de **hidroximetilación en diferentes *enhancers*** (potenciadores) para pacientes con CCR en diferentes etapas de desarrollo (N=625), AAR y controles. Aunque se incluyen pacientes con otras patologías, estos serán descartados para adaptar el set de datos al objetivo de este proyecto. La información de la hidroximetilación

de los *enhancers* viene acompañada de diferentes covariables para cada paciente, como género, edad, etapa en la que se encuentra el cáncer, y raza.

Dado que el set de datos original no tenía una estructura adecuada para los objetivos propuestos, se han tenido que llevar a cabo diferentes **procesos de consolidación**:

1. Combinación de conjuntos de entrenamiento y validación: Los datos de hidroximetilación se dividían, por defecto, en un conjunto de muestras de entrenamiento y otro conjunto de validación. Estos dos conjuntos se han combinado para generar un solo archivo con todas las muestras.
2. Transposición del conjunto de datos: El set de datos original se estructuraba de manera que los *enhancers* (características o *features*) se encontraban en las filas, y las diferentes muestras en las columnas. Para estructurarlo de manera correcta, se llevó a cabo una transposición de los datos para tener las *características* en las columnas del set de datos.
3. Adición de covariables: Las covariables de las muestras, como el género, la patología o la raza, se encontraban en un archivo separado. Se ha llevado a cabo un proceso de consolidación para añadir estas covariables al set de datos final.

3.1.2. Análisis exploratorio de los datos

El set de datos, sin filtrar muestras ni características, cuenta con un total de **1392 registros** y **18.443 características**. Se puede encontrar un resumen de las características de este set de datos en la Tabla 5.

Nombre	Interpretación	Tipo	Ejemplo
samples	Código de cada una de las muestras	Texto	CEG99_808_31PC
disease	Clasificación del paciente de acuerdo con su fenotipo	Texto	COLORECTAL CANCER
stage	Si es una muestra de cáncer, se indica la fase	Texto	IV
gender	Género del paciente	Texto	FEMALE
ethnicity	Etnia del paciente	Texto	WHITE

age_at_collection	Edad del paciente a la que se obtuvo la muestra	Número entero	56
enhancers	18.437 columnas que incluyen los valores de metilación de diferentes *enhancers*	Número decimal	1.42

Tabla 5: Características encontradas en el set de datos. Se muestra una interpretación de cada una de ellas, el tipo de valor que puede tomar, y un ejemplo.

Dado que en este primer set de datos se incluyen muestras cuya condición es diferente a CCR, AAR o control, antes de llevar a cabo el EDA descartamos estas muestras. Esto hace que nuestro set de datos esté formado por **1129 registros** y las características descritas en la tabla 6. Estas características las podemos dividir en variable objetivo ("disease"), covariables ("gender", "ethnicity" y "age_at_collection") y variables de metilación (columnas *enhancers*). Por su parte, la variable "stage" puede considerarse una variable objetivo secundaria para la tarea de predicción de la progresión del CCR.

La exploración de las variables categóricas se realizó mediante diferentes gráficos de barras, mientras que la distribución de los valores de hidroximetilación en los *enhancers* se exploró haciendo uso de diagramas de cajas y bigotes (*boxplot*) de acuerdo con la variable objetivo. Esta tarea se realiza con las librerías de Python *seaborn*, *matplotlib* y *plotly*.

3.1.3. Preprocesado y selección de características

Esta etapa de la metodología pretende generar los sets de datos finales para las etapas finales de generación y evaluación de modelos predictivos. Por tanto, el primer paso a realizar es la **eliminación de variables no relevantes** para la predicción de la variable objetivo, como la variable "samples". Esta variable carece de poder predictivo dado que representa el nombre de la muestra. Por otro lado, la variable "ethnicity" se ha decidido omitir a la hora de hacer predicciones para evitar sesgos, centrándonos de esta manera en variables con un mayor impacto clínico, como la edad o el sexo.

Aunque no existen valores nulos, algunos de los valores de las **variables fueron modificados o corregidos**. Los valores de la variable "age_at_collection" normalizaron

dividieron por su mediana. Los valores de la variable “gender” fueron igualmente reemplazados por 0 o 1, según la muestra correspondiera a un hombre o una mujer, respectivamente.

La variable objetivo “disease” también se modificó, convirtiendo las diferentes categorías en valores numéricos. De esta manera, los pacientes con CCR se identifican como 1, mientras que los pacientes control se identifican como 0. Un proceso parecido se realiza con la variable “stage”, codificando las diferentes fases (I, II, III y IV) como 1, 2, 3 y 4. Esta variable se codifica como 0 para los registros de controles.

No se realizan modificaciones en los valores de metilación de los diferentes *enhancers*, ya que, por defecto, están estandarizados (*z-score normalization*) por muestra y se asume una escala comparable. No obstante, y con el objetivo de disminuir la dimensionalidad del set de datos, se realiza un filtrado de estos *enhancers* de manera que:

- Se seleccionan aquellos *enhancers* que regulen la expresión de genes relacionados con cáncer. Los genes regulados por cada uno de los *enhancers* están disponibles en el set de datos original. Los genes asociados con cáncer se obtienen de la colección C6 de la base de datos *Molecular Signatures Database* (MSigDB)[74]. Esta colección, que contiene 189 set de genes, representa un conjunto de genes de diferentes vías celulares que están alteradas en cáncer. La mayoría de estos genes se han obtenido a partir de estudios de microarrays de expresión génica.
- Mediante t de Student (t-test), se encuentran los *enhancers* diferencialmente metilados (EDM) entre las muestras control y CCR. Se seleccionan aquellos *enhancers* cuyo estadístico ajustado por múltiples comparaciones (p-adj) sea < 0.05. Esta corrección se realiza mediante Bonferroni.
- Como alternativa a la aplicación de t-test para la selección de enhancers, se utilizó el método SelectKBest de scikit-learn. De esta manera se seleccionaron los “k” enhancers con mayor capacidad predictiva utilizando como criterio de clasificación la prueba F (f_classif). Se probaron valores de 100, 200 y 300 para “k”.

De manera general, los procesos de selección de características, en lugar de realizarlos sobre todo el conjunto, se realizan solo sobre el conjunto de entrenamiento y se trasladan los resultados al conjunto de prueba. De esta manera evitamos eventos no deseados como *data leakage* y posibles sobreajustes de los modelos generados.

En resumen, y siguiendo la metodología señalada, se generaron **4 sets de datos diferentes dependiendo del número de características** seleccionadas:

- Set de datos **t-test**: con un total de 3127 características.
- Set de datos **K100**: con un total de 100 características.
- Set de datos **K200**: con un total de 200 características.
- Set de datos **K300**: con un total de 300 características.

3.1.4. Aplicación de modelos supervisados

Con el objetivo de desarrollar un método de clasificación capaz de discernir entre pacientes con CCR y controles, se probaron diferentes modelos supervisados partiendo de los sets de datos generados en el proceso de preprocesado y selección de características. Previo a la generación de modelos, los diferentes sets de datos se dividieron en:

- Conjunto de entrenamiento (70%): utilizado para la selección de características, el ajuste de hiperparámetros y entrenamiento del modelo.
- Conjunto de prueba (30%): Utilizado para comprobar la bondad de los modelos generados.

Es importante resaltar que, dado que el tamaño de muestra de pacientes con AAR es pequeño (82) comparado con pacientes con CCR y controles, se decidió descartar esta cohorte y hacer un modelo de clasificación binario. Además, y con el fin de garantizar la reproducibilidad de los procesos aleatorios, se estableció un número de semilla (*seed*). De este modo, garantizamos que procedimientos como la partición del conjunto de datos en conjuntos de entrenamiento y prueba se llevan a cabo de manera consistente en todas las ejecuciones del algoritmo.

Para el desarrollo del algoritmo de clasificación se probaron, en cada uno de los 4 set de datos generados, un total de 6 modelos supervisados diferentes:

- **SVM con kernel lineal**: En este tipo de modelos se busca el hiperplano que maximice la separación de las diferentes clases del set de datos de acuerdo con las características de estas. En el caso del SVM lineal, el hiperplano es lineal. Por tanto, este tipo de modelo es adecuado para aquellas clases que sean linealmente separables.

- **SVM con kernel de base radial (rbf)**: En el caso que las clases no sean linealmente separables, se utilizan funciones de mapeo no lineal (como rbf) para proyectar los datos en un espacio de características de mayor dimensión donde puedan ser separados linealmente. A diferencia del modelo lineal donde no se transforman los datos, al utilizar el modelo rbf no se pueden calcular con facilidad el peso de las características a la hora de clasificar las muestras en las diferentes clases. Por tanto, podemos considerar este modelo como "*blackbox*" en términos de interpretabilidad.
- **Árbol de decisión**: En este modelo, que se basa en la estructura de un árbol (de ahí su nombre), cada nodo interno representa una prueba basada en alguna de las características del set de datos. Dependiendo del resultado de esta prueba, se generan ramas con hojas (clasificación del registro de acuerdo con la prueba). El proceso de generación de nodos, ramas y hojas se repite hasta que se cumple un criterio de parada. Este modelo es la base de otros modelos más complejos como GB o RF.
- **GB**: Es un modelo de ensemble formado por diferentes arboles de decisión simples. A medida que se va ajustando el modelo a los datos de entrada, los nuevos árboles se construyen de manera que reducen el error residual (diferencia entre predicción y valor real). Una vez que se entrenan un número determinado de árboles se agregan sus predicciones para obtener la predicción final. Este modelo puede ser propenso a sobreajuste, por lo que se suelen usar parámetros de regularización para evitarlo.
- **RF**: Al igual que GB, este modelo es un modelo de ensemble formado por un número determinado de árboles de decisión. A diferencia de GB, en este modelo cada árbol se construye y entrena de forma paralela e independiente en una muestra aleatoria del conjunto de datos. Tras esto, las predicciones de cada árbol se combinan para producir una predicción final. Como ventaja sobre GB, suele ser menos propenso al sobreajuste.
- **Modelo de ensemble**: Entrenados y evaluados los modelos anteriores, se elige el modelo con mejores métricas y se construye un modelo de ensemble mediante la combinación de 10 modelos base utilizando subconjuntos aleatorios del conjunto de datos de entrenamiento (*bagging*). La idea detrás de esto es comprobar si la combinación de modelos puede superar las limitaciones de los modelos individuales y mejorar su capacidad de generalización.

Antes de entrenar los modelos, se realizó la búsqueda de los hiperparámetros óptimos para cada uno de los modelos. Para ello se utilizó el método GridSearchCV de scikit-learn, que permite realizar una búsqueda exhaustiva de los hiperparámetros mediante validación cruzada k-fold. En el caso de este proyecto, se dividió el conjunto de entrenamiento en 5 subconjuntos (k=5) y, para cada modelo, se evaluaron los hiperparámetros recogidos en el anexo 2. En primer lugar, y con el objetivo de realizar una valoración más balanceada de los modelos, se escogieron aquellos hiperparámetros que maximizaban el valor F1 del modelo. Comprobada la adecuación y seleccionados aquellos modelos más óptimos, se estudió la posibilidad de hacer modificaciones en estos hiperparámetros para aumentar la especificidad. Haciendo esto, se pretende minimizar el número de controles que el modelo clasifica como casos con CCR (falsos positivos), evitando así posibles tratamientos e intervenciones en individuos sanos.

3.1.5. Evaluación y comparación de modelos supervisados

Tras ajustar los modelos supervisados con el conjunto de entrenamiento, se realizó la evaluación final prediciendo los registros en el conjunto de prueba, no utilizado previamente en ninguna de las tareas de ajuste del modelo.

Para la evaluación y comparación de los modelos, se aplicaron diversas métricas para medir el rendimiento y determinar qué modelo y combinación de características eran los óptimos. En la elección de estas métricas se tuvo en cuenta la naturaleza desbalanceada del conjunto de datos (más muestras de CCR que de controles), seleccionando métricas adecuadas para estos casos, como la precisión, la sensibilidad (*recall* o *sensistivity*), el valor F1, AUC-ROC o AUC-PR.

Para evaluar la generalización de los datos y el posible sobreentrenamiento (*overfitting*) de los modelos, los valores de las métricas al predecir el conjunto de prueba fueron comparados con los valores obtenidos en una validación cruzada 5-fold en el conjunto de entrenamiento. De esta manera, si el rendimiento en el conjunto de prueba es significativamente peor que el promedio de la validación cruzada, puede ser un indicio de sobreentrenamiento y una mala generalización de los datos.

Por último, se comprobó si, tras mezclar aleatoriamente las etiquetas del conjunto de datos, el modelo perdía su capacidad predictiva. Así, si el modelo ha aprendido patrones

verdaderamente relacionados con la condición estudiada, se espera que el rendimiento del modelo disminuya significativamente en comparación con el rendimiento en el conjunto de datos original. Esto se debe a que las etiquetas permutadas ahora están desvinculadas de las características, y el modelo no debería ser capaz de hacer predicciones precisas. Por el contrario, si el modelo mantuviera un rendimiento similar en el conjunto de datos con etiquetas permutadas, sugeriría que no ha aprendido patrones específicos relacionados con la condición estudiada, sino que las predicciones se basan en características generales o ruido en los datos.

3.1.6. Aplicación de modelos no supervisados

Tras comprobar la validez de los métodos supervisados, se llevó a cabo un análisis de *clústeres* utilizando el algoritmo K-means. El objetivo de esto es investigar si controles, pacientes con AAR y pacientes con CCR son separables de manera no supervisada utilizando los patrones de hidroximetilación de los *enhancers* seleccionados.

El algoritmo K-means divide el set de datos en k subgrupos (*clústeres*) no solapantes, haciendo que cada registro (pacientes) pertenezca a un solo subgrupo. Para determinar los *clústeres*, se busca que las diferencias entre los puntos de un mismo *clúster* sean mínimas, mientras que las diferencias entre puntos de diferentes *clústeres* sean máximas. El número óptimo de *clústeres* (k) se obtuvo mediante el método del codo, que consiste en trazar la suma de cuadrados intracluster en función del número de *clústeres*. Esto formará una curva con un “codo”, que se corresponde al número óptimo de *clústeres*.

Dada la alta dimensionalidad de los datos, se ha utilizado una PCA para transformar el conjunto de datos original en un nuevo conjunto de variables no correlacionadas (componentes principales). Este proceso puede eliminar la redundancia y la correlación (colinealidad) entre las características, haciendo más eficiente el uso de K-means. Además, la reducción del set de datos en componentes principales permite una mejor interpretación visual de los *clústeres*, ya que podemos representarlas en espacios de 2 o 3 dimensiones.

3.1.7. Otras herramientas

Además de los métodos mencionados anteriormente, se han utilizado herramientas y bases de datos específicas de cáncer para analizar los resultados obtenidos. La expresión y

metilación de genes en casos de CCR y control se ha consultado utilizando Wanderer[75]. Esta herramienta ha permitido comprender mejor los patrones epigenéticos y los cambios de expresión genética asociados con el CCR. De manera complementaria se ha utilizado *The Human Protein Atlas* que, además de información de expresión genética, proporciona datos sobre la localización de las proteínas codificadas por genes candidatos a nivel celular y tisular[76]. Por último, y con el objetivo de hacer un análisis de enriquecimiento de rutas biológicas (*pathway enrichment analysis*) con los genes regulados por los *enhancers* candidatos, se ha utilizado ToppGene[77].

3.1.8. Accesibilidad de los datos

Los datos y el código utilizado durante el desarrollo de este estudio están disponibles en GitHub (https://github.com/PabloRomanjo/tfm-data-science) y en Zenodo (https://zenodo.org/record/8061669). La referencia a los datos originales se puede encontrar en el apartado 3.1.1.

3.2. Resultados

3.2.1. Exploración de los datos

Antes de poder aplicar algoritmos supervisados o no supervisados sobre nuestros datos, debemos conocer un poco más de ellos. Por ello, el primer paso que se llevó a cabo fue una exploración de los datos para presentar una descripción general de los mismos.

Como se ha mencionado en anteriores apartados, la característica principal (variable objetivo) del set de datos es la condición del paciente, diferenciándose, principalmente, en CCR, AAR y controles. Aunque en el set de datos original se incluían otras condiciones, estas fueron descartadas para los análisis posteriores. El gráfico de barras de la **Figura 14** muestra cómo, del total de registros, existen 625 casos de CCR, 82 casos de AAR, y 422 controles. Por lo tanto, nos enfrentamos a un conjunto de datos desequilibrado en términos de tamaño de muestra para las diferentes clases. Sin embargo, se observó que estas clases están equilibradas en cuanto al género de los individuos.

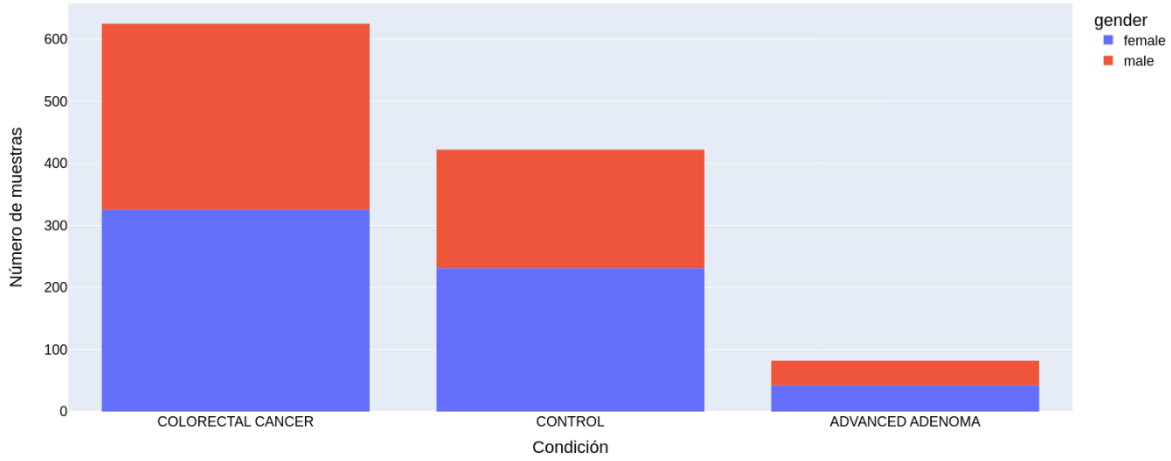


Figura 14: Distribución de las muestras que conforman el set de datos de acuerdo con su condición. La distribución por géneros, marcada en color azul y rojo, es balanceada.

Las muestras con CCR están divididas según la fase en la que se encuentra el cáncer, existiendo 4 fases principales, tal y como se discutió en la introducción de esta memoria. La distribución de las muestras de acuerdo con estas 4 fases parece seguir una distribución normal, tal y como se puede comprobar en las **Figura 15.** Las muestras en cada una de las fases también están balanceadas en cuanto a género.

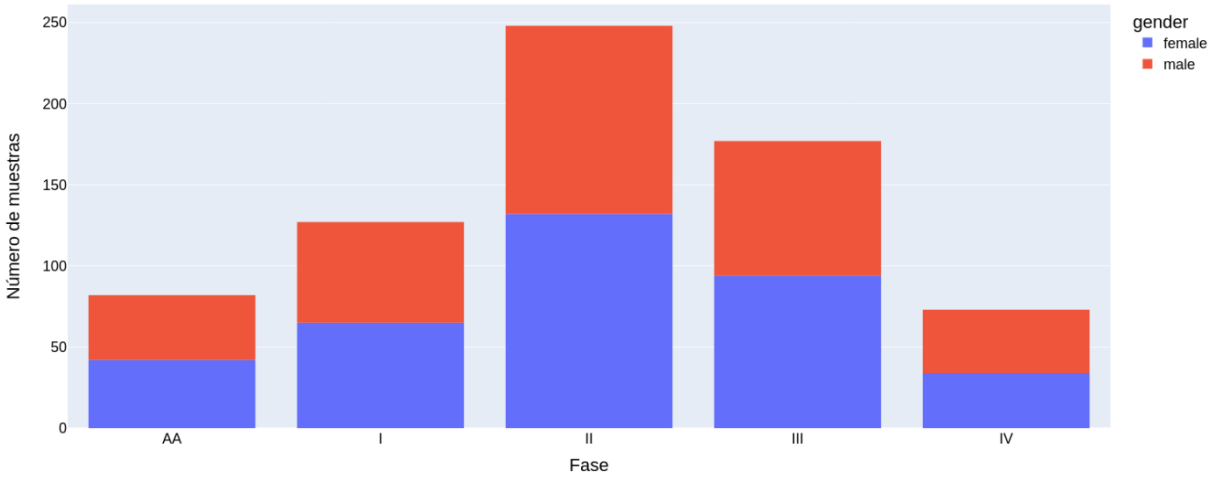


Figura 15: Distribución de las muestras con CCR de acuerdo la fase o etapa del cáncer en la que se encuentra (I, II, III o IV). También se representan las muestras con adenoma avanzado (AA). La distribución por géneros, marcada en color azul y rojo, es balanceada.

En la **Tabla 6** se resumen las diferentes cohortes y subcohortes encontradas en este set de datos.

Cohorte		**Muestra (%)**	**Genero (% mujeres)**
Cáncer colorrectal	Fase I	127 (11,2%)	51,2%
	Fase II	248 (21,9%)	53,2%
	Fase III	177 (15,7%)	53,1%
	Fase IV	73 (6,5%)	46,6%
Adenoma de alto riesgo		82 (7,3%)	51,2%
Controles		422 (37,4%)	54,7%
		Total: 1129	**Total: 51,7%**
Otras patologías		263	50,9%

Tabla 6: Muestras que forman el conjunto de datos final (1129), y otras descartadas por padecer otras patologías (263). Para la muestra final, se muestra el número de registros por grupo, así como su distribución por género.

Por otro lado, se estudiaron las variables demográficas asociadas a estas muestras. La edad de los pacientes, independientemente del grupo, oscilaron entre 45 y 85 años, siguiendo una distribución parecida y próxima a la normalidad en los tres grupos (**Figura 16**).

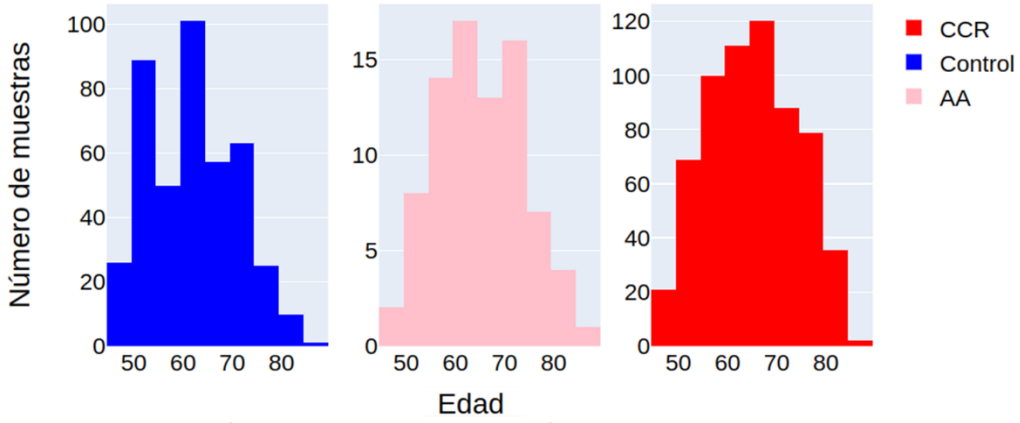


Figura 16: Distribución de los registros de cáncer colorrectal (CCR), adenoma avanzado (AA) y controles según la edad en la que se extrajo la muestra del individuo.

Además, se pudo comprobar como la mayoría de las muestras de los tres grupos eran de raza caucásica (**Figura 17**).

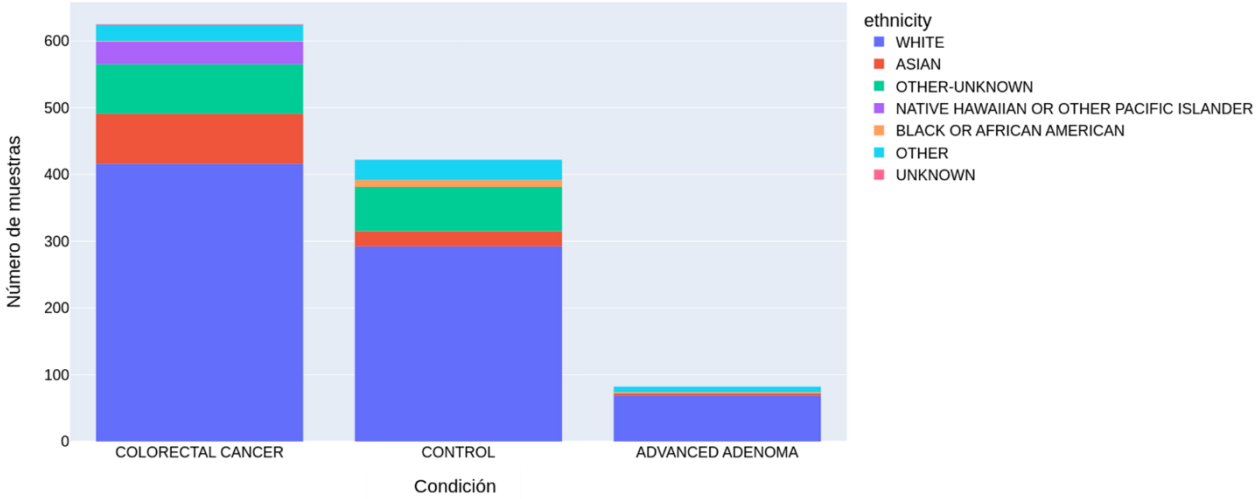


Figura 17: Etnias representadas en los registros de cáncer colorrectal (CCR), adenoma avanzado (AA) y controles.

Por último, se comprobó los valores de hidroximetilación que tomaban los diferentes *enhancers*. En este set de datos existen 18.437 *enhancers* cuyos valores de hidroximetilación, ya normalizados por z-score, oscilan entre -6,87 y 3,47. A pesar de abarcar un rango amplio de valores, la desviación estándar mínima, máxima y media encontrada en

todos los *enhancers* fue de 0,080, 0,90 y 0,27, respectivamente. Esto nos indica que la variabilidad dentro de cada *enhancer* es relativamente baja (**Figura 18**).

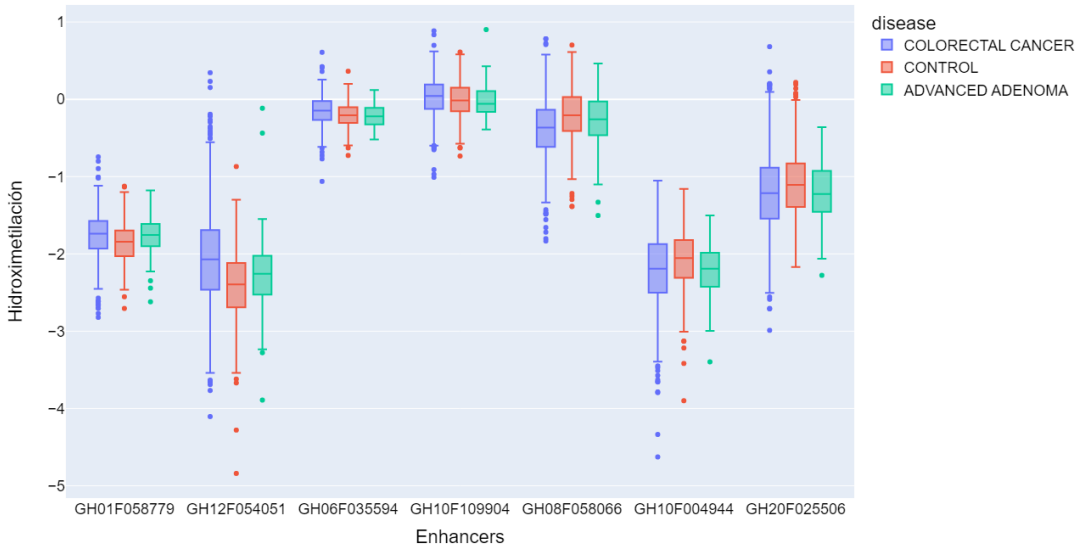


Figura 18: Ejemplos de los valores de hidroximetilación de diferentes *enhancers* en las muestras de cáncer colorrectal, adenoma avanzado y controles.

3.2.2. Modelos supervisados

Tras conocer las dimensiones del conjunto de datos y los valores que toman cada una de sus variables, se valoró la capacidad de diferentes modelos supervisados para clasificar los registros en CCR o controles. El set de datos se dividió en un conjunto de entrenamiento (70%), utilizado para la selección de características, el ajuste de hiperparámetros y entrenamiento del modelo; y un conjunto de prueba (30%) para comprobar el rendimiento real de los modelos.

Tal y como se describió en la metodología de esta memoria, y dada la alta dimensionalidad de los datos, se decidió llevar a cabo diferentes métodos de selección de características previa al entrenamiento de los modelos supervisados. En concreto, se utilizaron dos métodos univariantes, como t-test y la prueba F, generando 4 set de datos de acuerdo con el número de características: t-test, K100, K200 y K300. Cada método fue aplicado de forma iterativa para evaluar su rendimiento en la clasificación de los registros: selección de características, ajuste de hiperparámetros, entrenamiento del modelo y evaluación del rendimiento.

Seleccionadas las características de interés, se procedió al ajuste de hiperparámetros de cada modelo de manera que se maximizara el rendimiento en la clasificación de los registros. Los modelos supervisados probados, descritos previamente en la metodología, fueron SVM con kernel lineal, SVM con kernel rbf, árbol de decisión, GB y Random Forest. Los hiperparámetros seleccionados para cada combinación de modelo y método de selección de características se puede consultar en el **Anexo 3**.

Tras el ajuste de hiperparámetros, se entrenaron los modelos utilizando una validación cruzada de 5 iteraciones en el conjunto de entrenamiento, evaluando el rendimiento en cada iteración. Los modelos que ofrecieron mejores métricas fueron SVM con kernel lineal y SVM con kernel rbf, ambos junto a la selección de características mediante t-test. Entre estos dos modelos, y dado que las métricas eran muy parecidas, se decidió continuar con el modelo SVM con kernel lineal, cuya interpretabilidad es mayor que el modelo con kernel rbf. En el **Anexo 4** se pueden comprobar las métricas para todos los modelos y métodos de selección de características.

Una vez decidido el modelo base, se construyó un modelo de ensemble mediante *bagging* para comprobar si se mejoraban las métricas del modelo individual. En concreto, se crearon 10 modelos SVM independientes, cada uno entrenado con una muestra aleatoria del conjunto de datos de entrenamiento. Ambos modelos se evaluaron con el umbral de decisión por defecto (0.5), además de ajustar el umbral para fijar la especificidad al 90% y 95%, observando como variaban las demás métricas (**Figura 19**).

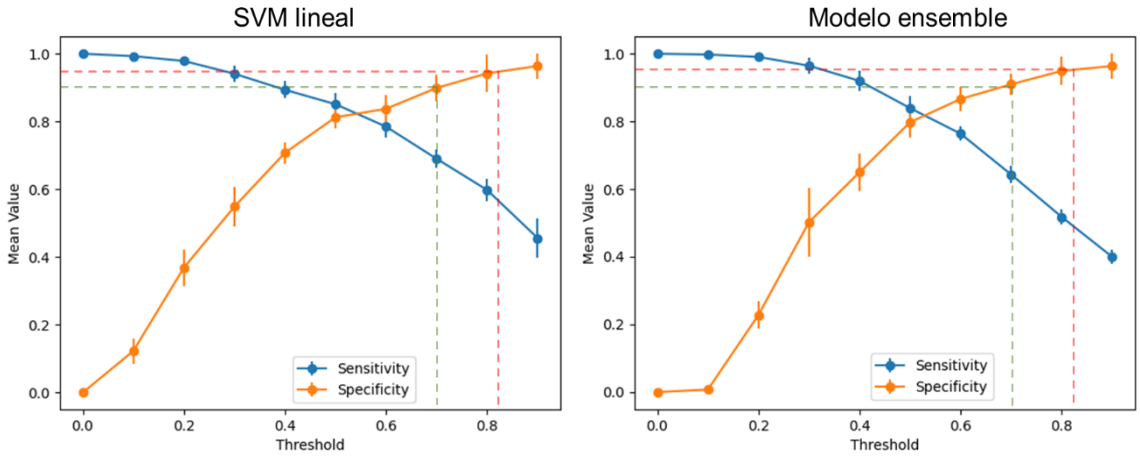


Figura 19: Sensibilidad y especificidad medias durante la validación cruzada de los modelos candidatos para diferentes umbrales de decisión. Las líneas punteadas verdes y rojas indican el umbral donde la especificidad media es del 90% y 95%, respectivamente.

Al comparar el rendimiento de ambos modelos, el modelo de ensemble por *bagging* no mejoró las métricas del modelo individual (**Tabla 7**). Este modelo tampoco mejoraba al modelo individual cuando se estudió su rendimiento en cada una de las fases de la enfermedad de manera separada (**Anexo 5**). Además, la sensibilidad media del modelo SVM lineal (58%) mejoró la sensibilidad del modelo presentado en el artículo original (55%) para una especificidad media del 95%.

Modelo	Umbral de decisión	Especificidad	Sensibilidad
SVM lineal	0.5 (por defecto)	**0.82 ± 0.04**	0,84 ± 0.03
	0.7	**0.90 ± 0.04**	0.69 ± 0.03
	0.82	**0.95 ± 0.05**	0.58 ± 0.04
Ensemble SVM	0.5 (por defecto)	**0.80 ± 0.05**	0,84 ± 0.04
	0.7	**0.90 ± 0.03**	0.64 ± 0.02
	0.82	**0.95 ± 0.05**	0.49 ± 0.02
Modelo *Walker et al.*[72]	0.98	**0.95**	0.55

Tabla 7: Sensibilidad del modelo individual y ensemble con umbral de decisión por defecto, al fijar la especificidad al 90% y al 95%. Se muestra también el modelo del artículo original.

Por tanto, tras la validación cruzada, el modelo finalmente seleccionado fue un modelo SVM con kernel lineal. Las matrices de confusión y la distribución de las probabilidades de predicción para cada una de las iteraciones de la validación cruzada pueden ser consultadas en el **Anexo 6**.

Las importancias relativas de cada una de las características se derivaron de los coeficientes del modelo. Estos coeficientes indican la contribución de cada característica en la separación de las clases, de manera que: A) si el coeficiente es positivo, un aumento del valor de la característica aumenta la probabilidad de que un registro se clasifique en la clase positiva, en este caso, como CCR; y B) si el coeficiente es negativo, un aumento en el valor de la característica disminuye la probabilidad de que un registro se clasifique como positivo. La magnitud de este coeficiente también es importante, siendo los valores más grandes, ya sean positivos o negativos, más relevantes en la clasificación de los registros. En el **Anexo 7** se puede ver una visión general de la distribución de importancias de las características.

Mediante este análisis, se identificaron diferentes *enhancers* cuyos niveles de hidroximetilación estaban correlacionados con la clasificación de los registros como CCR (**Tabla 8**). Estos *enhancers*, cuando presentaban aumentos o disminuciones en sus niveles de hidroximetilación, mostraron una influencia significativa en la predicción del modelo, lo

que sugiere una relación entre la hidroximetilación en estos *enhancers* y la presencia de CCR en los pacientes de nuestro set de datos. Además, los genes regulados por estos *enhancers* podrían desempeñar un papel importante en la patogénesis o progresión del CCR.

Corr.	Enhancer	Coef.	Gen que regula	Efecto esperado	Relacionado con	
					CCR	Otros
Positiva	GH01F058779	0.098	*MYSM1*	Aumento de la expresión del gen	Sí [78,79]	Sí [80]
	GH10F109904	0.086	*ADD3*		No	Sí [81]
	GH06F035594	0.083	*FKBP5*		No	Sí [82]
	GH12F054051	0.083	*SP1*		Sí [83–86]	Sí [87]
	GH03F156786	0.082	*LINC00886*		No	Sí [88]
Negativa	GH08F058066	-0.099	*FAM110B*	Disminución de la expresión del gen	No	Sí [89]
	GH10F004944	-0.089	*AKR1C1*		No	Sí [90]
	GH20F025506	-0.086	*ABHD12*		No	No
	GH10F048574	-0.080	*ARHGAP22*		No	No
	GH16F088381	-0.075	*ZNF469*		Sí [91]	No

Tabla 8: *Enhancers* más relevantes a la hora de clasificar un registro en la clase positiva (cáncer colorrectal) y los genes principales que regulan. Además, se presenta también el efecto esperado en la expresión del gen con respecto a las muestras control. En las últimas columnas se indica si existe alguna relación publicada con cáncer colorrectal u otros cánceres. Corr: Correlación, Coef: Coeficiente.

Por último, y para comprobar el rendimiento real del modelo SVM lineal, este se ajustó con el conjunto de entrenamiento completo y se predijeron las clases del conjunto de prueba, no utilizado en ninguno de los procesos anteriores y, por tanto, no visto previamente por el modelo. Tras comprobar las métricas obtenidas con el modelo en el conjunto de pruebas se observó un rendimiento satisfactorio, no difiriendo demasiado del comportamiento del modelo en la validación cruzada con el conjunto de entrenamiento (**Tabla 9**). Esto sugiere una buena generalización de los datos por parte del modelo.

Métrica	Umbral: 0.5		Umbral: 0.7		Umbral: 0.82	
	Entrenamiento	Prueba	Entrenamiento	Prueba	Entrenamiento	Prueba
Exactitud	0.83+/-0.03	0.81	0.77+/-0.02	0.78	0.73+/-0.01	0.71
Sensibilidad	0.84+/-0.03	0.85	0.69+/-0.03	0.71	0.58+/-0.04	0.57
Precisión	0.88+/-0.03	0.85	0.91+/-0.03	0.9	0.95+/-0.04	0.93
F1	0.86+/-0.02	0.85	0.79+/-0.02	0.79	0.72+/-0.02	0.71
Especificidad	0.82+/-0.04	0.76	0.9+/-0.04	0.87	0.95+/-0.05	0.93
Brier	0.13+/-0.01	0.14	0.13+/-0.01	0.14	0.13+/-0.01	0.14
Log-loss	0.44+/-0.07	0.45	0.44+/-0.07	0.45	0.44+/-0.07	0.45
AUC-ROC	0.89+/-0.03	0.87	0.89+/-0.03	0.87	0.89+/-0.03	0.87
AUC-PR	0.92+/-0.02	0.9	0.92+/-0.02	0.9	0.92+/-0.02	0.9

Tabla 9: Rendimiento del modelo en el conjunto de prueba, con diferentes umbrales de decisión, comparado con el rendimiento del modelo en la validación cruzada con el conjunto de entrenamiento.

De manera alternativa, se comprobó la consistencia del rendimiento del modelo mediante la representación gráfica de las AUC-ROC obtenidas en cada iteración de la validación cruzada con el conjunto de entrenamiento y la AUC-ROC del conjunto de pruebas (**Figura 20A**). Además, también se vio como la calibración del modelo era correcta mediante la representación de las probabilidades pronosticadas por el modelo y la frecuencia real de la clase positiva (**Figura 20B**).

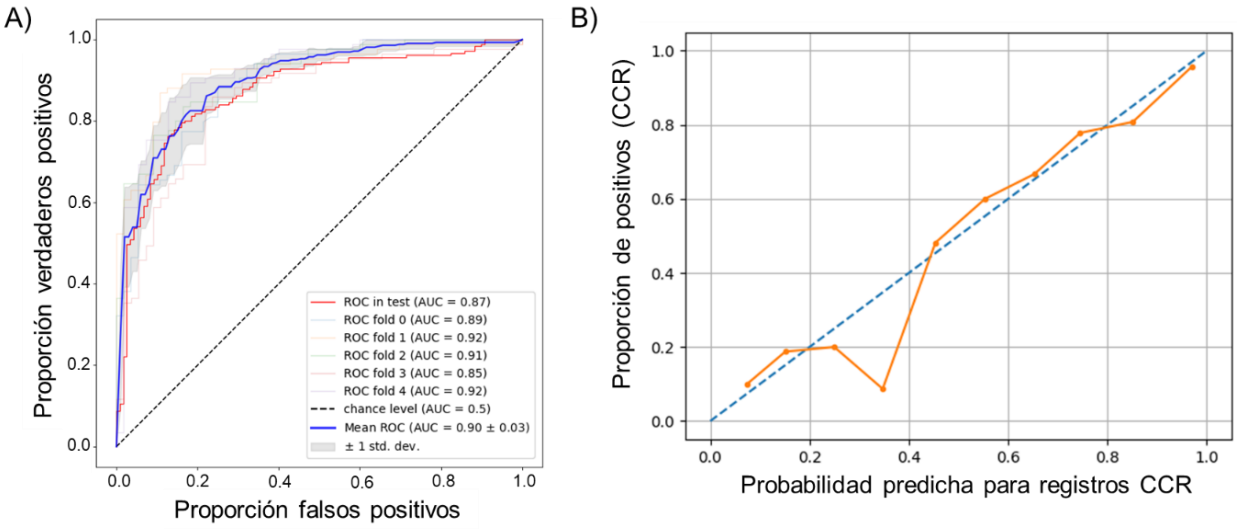


Figura 20: A) AUC-ROC para cada iteración de la validación cruzada en el conjunto de entrenamiento, AUC-ROC media en esta validación cruzada (en azul) y AUC-ROC en conjunto de prueba (en rojo). B) Gráfico de calibración del modelo SVM lineal en conjunto de prueba (en naranja). La línea punteada azul muestra un ajuste perfecto de un modelo hipotético.

3.2.3. Enriquecimiento de rutas biológicas

Además de estudiar los *enhancers* de manera individual, se decidió realizar un enriquecimiento de rutas biológicas con los genes que estos regulan mediante ToppGene[77].

En concreto, para este análisis se seleccionaron aquellos genes regulados por los 400 *enhancers* más relevantes de acuerdo con los coeficientes del modelo de clasificación. Así, y tras corregir los valores por múltiples comparaciones, se detectaron un total de 4 rutas enriquecidas en los genes regulados por los enhancers estudiados. Estas rutas biológicas

están relacionadas con la señalización llevada a cabo por el factor de crecimiento transformante beta (TGF-β) y los procesos llevados a cabo por las integrinas (**Figura 21**).

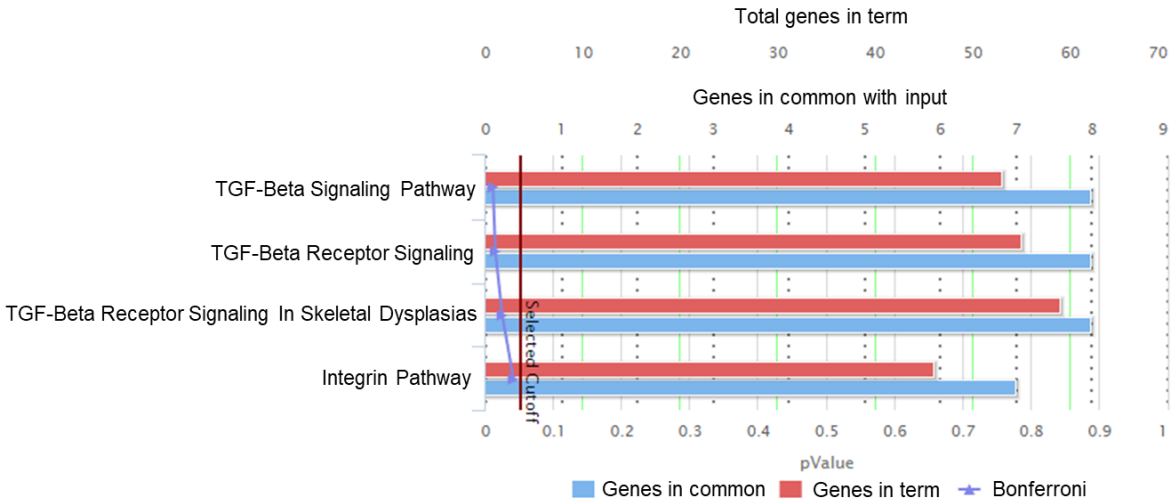


Figura 21: Rutas biológicas enriquecidas en los genes regulados por los 400 *enhancers* más relevantes para el modelo de clasificación. En rojo y en el eje superior se indica el número total de genes en las rutas. En azul, los genes en común entre la consulta y las rutas.

La ruta de señalización del TGF-β está implicada en diferentes procesos biológicos, incluyendo la proliferación celular, la diferenciación celular y la supresión de tumores. En los *enhancers* que regulan genes asociados a esta ruta se han encontrado, principalmente, niveles más bajos de hidroximetilación que en los controles (**Tabla 10**). Estos niveles más bajos de hidroximetilación conllevarían una disminución en la expresión de estos genes.

Enhancer	Coeficiente	Símbolo	Efecto esperado
GH16F049882	-0.060	*ZNF423*	Disminución de la expresión
GH05F180342	-0.058	*MAPK9*	
GH22F041444	-0.055	*EP300*	
GH15F039509	-0.037	*THBS1*	
GH18F050790	-0.035	*SMAD4*	
GH07F041910	-0.035	*INHBA*	
GH22F029631	-0.034	*LIF*	
GH18F048954	0.034	*SMAD7*	Aumento de la expresión

Tabla 10: Enhancers que regulan la expresión de genes involucrados en la ruta del TGF-β.

Por otro lado, las integrinas juegan un papel crucial en la comunicación celular y la adhesión de las células, y están involucradas en procesos biológicos como la proliferación celular, la migración celular y la angiogénesis. En los *enhancers* que regulan genes asociados a integrinas se han encontrado tanto niveles más bajos como más altos de hidroximetilación con respecto a los controles (**Tabla 11**).

Enhancer	Coeficiente	Símbolo	Efecto esperado
GH01F183038	0.052	*LAMC1*	Aumento de la expresión
GH06F112231	0.042	*LAMA4*	
GH17F059286	0.035	*RPS6KB1*	
GH17F066537	0.034	*PRKCA*	
GH07F055017	-0.050	*EGFR*	Disminución de la expresión
GH17F075720	-0.039	*ITGB4*	
GH08F100899	-0.038	*YWHAZ*	

Tabla 11: Enhancers que regulan la expresión de genes involucrados en la ruta de las integrinas.

3.2.4. Modelos no supervisados

Con el objetivo de evaluar el comportamiento del aprendizaje no supervisado en los datos de hidroximetilación, se utilizó el algoritmo K-means junto con una PCA. De esta manera, se redujo la dimensionalidad de los datos y, a su vez, se hizo más eficiente el proceso de K-means.

Tras esto, se aplicó la técnica del codo para determinar el número óptimo de *clústeres*, una estrategia utilizada para evaluar la variabilidad explicada por diferentes números de clústeres (**Figura 22**). Tras analizar la curva, se decidió utilizar 6 *clústeres*, coincidiendo con el número de subgrupos del set de datos: Las cuatro fases del CCR (CCR I, CCR II, CCR III, CCR IV), muestras con AAR y controles.

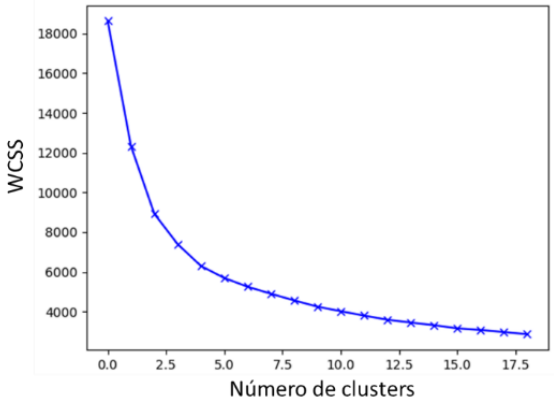


Figura 22: Determinación del número de *clústeres* óptimos de acuerdo con la técnica del codo. WCSS: *Within-Cluster Sum of Squares*.

Mediante K-means, se asignó uno de estos 6 *clústeres* a cada una de las muestras, agrupando así las muestras con patrones similares de hidroximetilación en un mismo *clúster.* Tras comprobar la composición de los *clústeres*, se pudo observar que el *clúster* 1 y el *clúster* 5 estaban formados, principalmente por muestras con CCR. Por el contrario, los clústeres 0, 2, 3 y 4 estaban formados por muestras de las tres condiciones (**Figura 23A**).

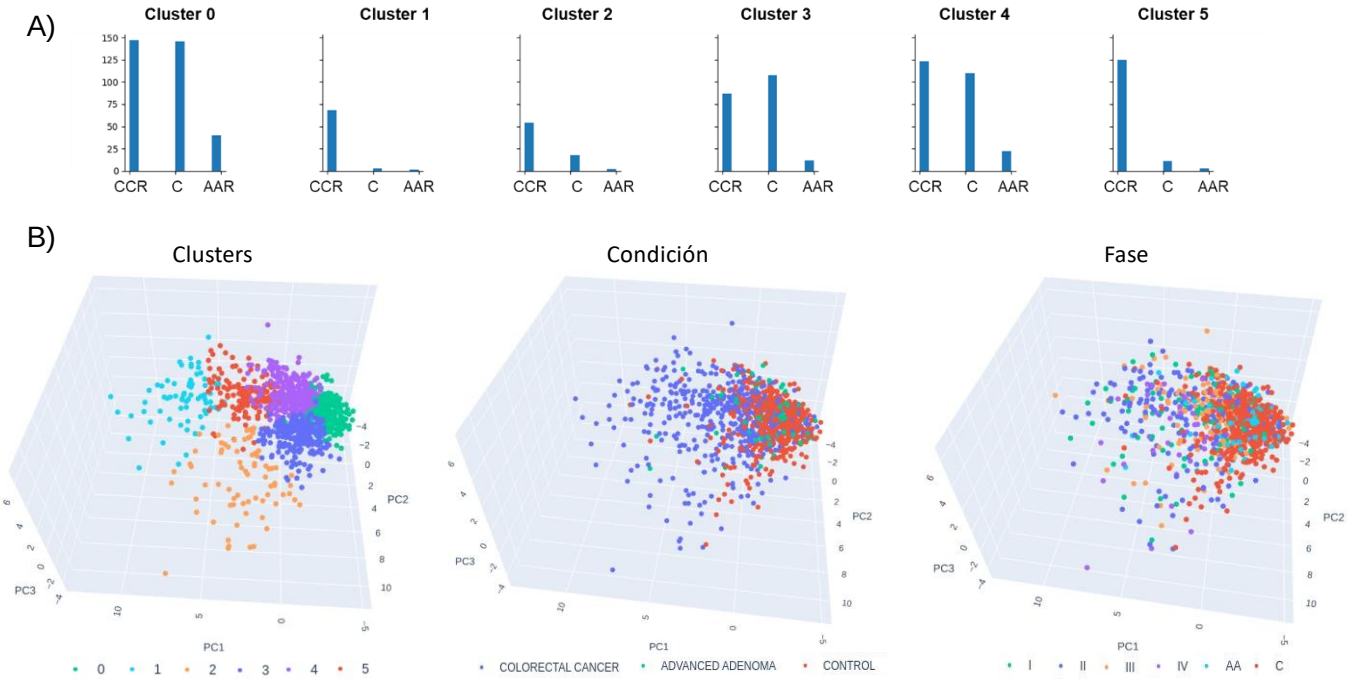


Figura 23: A) Distribución de las muestras según los *clústeres* generados por K-means y su condición. B) Visualización en tres dimensiones de las componentes principales. De izquierda a derecha se representan en diferentes colores a) los clústeres generados por K-means, b) la condición de cada muestra, y c) la subclase (fase) de cada muestra.

Al visualizar en tres dimensiones las tres primeras componentes principales, pudimos ver *clústeres* no muy separados. Al medir el coeficiente de silueta medio, se obtuvo un valor de 0.26, reflejando el hecho de *clústeres* poco diferenciados. Además, aunque el *clúster* 1 y el *clúster* 5 estén formados mayoritariamente por muestras con CCR, estas muestras no pertenecen a una fase específica del mismo (**Figura 23B y Anexo 8**).

Capítulo 4: Discusión, conclusiones y perspectivas

En el último capítulo de esta memoria se presentará una reflexión crítica sobre los resultados obtenidos, y se discutirán sus posibles implicaciones en el contexto de la detección temprana del CCR. Además, se presentará una revisión de los objetivos propuestos y alcanzados, así como las posibles implicaciones éticas de la implementación del modelo de predicción en un entorno clínico real. Por último, se resumirán las principales conclusiones del trabajo, proponiendo posibles líneas de investigación futuras.

4.1. Discusión

El uso de biomarcadores para la detección temprana del CCR es un área de investigación activa y prometedora. Actualmente, los métodos de detección de rutina, como la colonoscopia y las pruebas de SOH, son eficaces pero pueden tener limitaciones en términos de accesibilidad, costes y aceptación por parte de los pacientes. La identificación y validación de biomarcadores específicos de CCR mediante pruebas menos invasivas podrían superar estas limitaciones y mejorar la detección temprana de la enfermedad. En este Trabajo de Fin de Máster se ha abordado la pregunta de si los niveles de hidroximetilación del ADN en sangre podrían ser útiles para detectar el CCR, encontrando resultados prometedores mediante el uso de técnicas de aprendizaje automático.

4.1.1. Hidroximetilación y cáncer colorrectal

Las modificaciones epigenéticas del ADN, como la metilación o la hidroximetilación, juegan un papel importante en la regulación de la expresión génica. En determinados casos estas modificaciones pueden dar lugar al desarrollo de cáncer y, por tanto, pueden funcionar como biomarcadores para la detección de este[92]. Por su parte, las células, incluyendo las cancerosas, liberan ADN al torrente sanguíneo, lo que permite que el análisis de sangre periférica, una prueba poco invasiva que se realiza de rutina en cualquier hospital, sea un buen método para detectar estos cambios epigenéticos en el ADN[93].

En este Trabajo de Fin de Master hemos trabajado con un set de datos público para valorar la utilidad de los niveles de hidroximetilación medidos en sangre para detectar el CCR, abordándolo desde diferentes perspectivas. La primera aproximación fue desarrollar un

modelo supervisado para valorar su capacidad predictiva a la hora de clasificar los registros en CCR o controles. Tras evaluar diferentes modelos, los modelos que ofrecieron mejores métricas fueron un modelo SVM con kernel lineal y un modelo SVM con kernel rbf. Pese a que el modelo con kernel rbf mostró, ligeramente, mejores métricas, decidimos seleccionar el modelo **SVM con kernel lineal** por su mayor interpretabilidad. Esta elección facilitó la interpretación de los resultados y proporcionó una mayor comprensión de cómo se relacionan los niveles de hidroximetilación con la presencia de CCR, ya que el modelo produce límites de decisión lineales en el espacio de características[94]. No obstante, y dado que el modelo con kernel rbf muestra ligeras mejoras en métricas como la sensibilidad, es importante destacar que podrían existir patrones sutiles en los niveles de hidroximetilación que el modelo con kernel lineal no haya sido capaz de capturar de manera óptima. Por tanto, sería interesante realizar un análisis adicional para comprender la naturaleza de las relaciones no lineales presentes en los datos.

Tras evaluar el modelo SVM lineal, se obtuvieron métricas prometedoras en la tarea de clasificación de CCR y controles. El modelo mostró una sensibilidad media del 58% para una especificidad del 95% y un AUC-ROC del 89%, siendo ligeramente superior al rendimiento del modelo del artículo de donde se extrajeron los datos[72]. Las métricas variaron ligeramente cuando se analizaron cada una de las fases del CCR por separado, ofreciendo una sensibilidad media del 53% y del 63% para una especificidad del 95%, y AUC-ROC del 87% y del 91% en las fases I y II del CCR, respectivamente. El cambio de estas métricas podría ser debido a la diferencia de tamaño de muestra entre las diferentes fases (consultar Tabla 6), o al hecho de que los niveles de ADN circulante tumoral en sangre en fases tempranas sean menores a los encontrados en fases más tardías[95].

El rendimiento del modelo presentado en esta memoria se encuentra cerca del desempeño mostrado por otros estudios basados en biomarcadores medidos en sangre de pacientes con CCR. Un estudio llevado a cabo en 2014 ofreció un sensibilidad del 48,2% y una especificidad del 91,5% al medir los niveles de metilación del gen *SEPT9* en plasma de pacientes con CCR y controles[96]. Otro estudio de 2014 comparó el rendimiento de este mismo test de *SEPT9* y la SOH-inmunológica, obteniendo sensibilidades comparables con *SEPT9* (sensibilidad: 73,3%, especificidad: 81,5%) y la SOH-inmunológica (sensibilidad: 68,0%, especificidad: 97,4%), pero con un especificidad bastante menor[97]. En un estudio un poco más reciente, en 2017, haciendo uso de un panel que medía los niveles de metilación

en los genes *SFRP1*, SFRP2, *SDC2* y *PRIMA1* se alcanzaron valores de 91.5% y 97,3% para sensibilidad y especificidad, respectivamente, aunque la muestra con la que se trabajó fue pequeña (47 CCR y 37 controles)[98]. También en 2017, un modelo basado en LR y niveles de hidroximetilación medidos en sangre fue capaz de detectar pacientes con CCR con una sensibilidad del 88% y una especificidad del 89% en una pequeña cohorte de 32 pacientes y 37 controles[99]. En 2019, Wan *et al*. describieron un modelo que alcanzo un 85% de especificidad y sensibilidad a la hora de detectar pacientes con CCR usando como biomarcador el tamaño de los fragmentos de ADN en sangre de pacientes y controles[100].

Al generar un modelo con un rendimiento óptimo en la identificación de muestras con CCR comparado con estudios previos, se cumplió parte del primer reto propuesto dentro del contexto final. Para completar este primer reto, se identificaron los *enhancers* más relevantes en la clasificación de los registros.

4.1.2. *Enhancers* candidatos en el desarrollo o avance del CCR

Al seleccionar un modelo sin caja negra se pudieron identificar las características más relevantes en nuestro set de datos. De esta manera, se encontraron diferentes *enhancers* cuya hidroximetilación era significativamente más alta o baja en pacientes con CCR que en controles. Como se ha discutido en apartados anteriores, estos cambios en los niveles de hidroximetilación en los *enhancers* pueden conllevar un aumento o una disminución de la expresión de los genes que regulan, y por tanto, podrían estar participando en el desarrollo o avance del CCR.

Entre los genes regulados por *enhancers* cuya **hidroximetilación fue mayor en CCR** que en controles nos encontramos con *MYSM1* y *SP1*, relacionados previamente con CCR. El gen ***MYSM1*** ha sido relacionado con el desarrollo del cáncer en diversas ocasiones. En un estudio donde se evaluó su expresión en tejido tumoral de pacientes con CCR, los resultados sugirieron que un aumento de la expresión de este gen estaba correlacionado con la progresión del CCR[78]. No obstante, unos años después, fue publicado un artículo con resultados contradictorios, concluyendo que *MYSM1* actuaba como un supresor de la tumorogénesis, y que su expresión era más baja en CCR que en controles[79]. Los autores de este artículo especularon con la posibilidad de que estos resultados contradictorios se debieran al uso de diferentes anticuerpos para detectar la expresión de la proteína codificada por *MYSM1*. Los resultados obtenidos en esta memoria sugieren que una mayor

hidroximetilación del *enhancer* que regula *MYSM1* podría estar mediando el desarrollo del CCR, y esto podría estar sucediendo por un aumento en su expresión. Por otro lado, la relación entre el gen ***SP1*** y el CCR parece más clara que en el caso de *MYSM1*. La expresión del gen *SP1* es mayor en tejido canceroso que en tejido normal de pacientes con CCR, y estos cambios parecen tener un papel importante en la proliferación celular descontrolada, la angiogénesis (formación de nuevos vasos sanguíneos) y la metástasis del cáncer[85,86]. Por este motivo, este gen y la proteína codifica parecen ser una buena diana terapéutica. Un estudio en un modelo de ratón demostró como la inhibición de la expresión de *SP1* disminuía el crecimiento de las células tumorales de CCR e inducía su muerte por apoptosis[84]. Un estudio publicado en 2023 confirmó este hallazgo, esta vez en líneas celulares humanas de intestino y CCR[101]. Los resultados obtenidos de los datos de hidroximetilación sugieren que la hidroximetilación del *enhancer* que regula la expresión de *SP1* es mayor en casos con CCR que en controles, lo que apunta a una mayor expresión de *SP1* en los casos con CCR, replicando los resultados de los estudios publicados.

Entre los genes regulados por enhancers cuya **hidroximetilación fue menor en CCR** que en controles podemos destacar el gen ***ZNF469***. Aunque la asociación con el CCR no sea tan clara como en los genes expuestos anteriormente, un estudio publicado 2019 encontró una mayor metilación en tejido tumoral de casos esporádicos de CCR frente a tejido normal[91]. Esta mayor metilación conllevaría una menor hidroximetilación, por lo que podríamos considerar que los resultados de este artículo se replican en este Trabajo de Fin de Master. Otro de los genes con menor hidroximetilación fue el gen ***FAM110B***, que aunque no ha sido relacionado directamente con CCR, los resultados de un estudio realizado en una línea celular de cáncer de pulmón sugirieron que su sobreexpresión disminuía la proliferación e invasión del cáncer[89]. En los casos con CCR incluidos en nuestro estudio los niveles de hidroximetilación sugieren una disminución de la expresión de este gen, por lo que *FAM110B* podría estar mediando el desarrollo o progresión del CCR.

Con la identificación de estos *enhancers* se completó el primer reto de manera satisfactoria. No obstante, y aunque los resultados obtenidos son prometedores, cabe destacar que nuestros resultados están basados en datos procedentes de ADN obtenido en sangre, y no en tejido canceroso o células procedentes del mismo, como en los artículos citados anteriormente.

4.1.3. Las rutas del TGF-β y de las integrinas en el CCR

Para obtener una visión más completa de las funciones que pueden estar regulando los *enhancers* identificados, se realizó un análisis de rutas biológicas con los 400 enhancers más relevantes para el modelo de clasificación.

En esta memoria se han identificado, principalmente, dos rutas: la ruta del TGF-β y la ruta de las integrinas. La **ruta o vía de señalización del TGF-β** regula la proliferación y diferenciación celular, la apoptosis y la remodelación de la matriz extracelular, además de estar involucrada en procesos como la angiogénesis y la inflamación[102,103]. Que todas estas funciones dependan de esta ruta hacen que esta juegue un papel importante en la patogénesis del cáncer[104]. Centrándonos en el CCR, un estudio realizado en una línea celular de cáncer CCR encontró que la supresión de la vía del TGF-β aumentaba la angiogénesis y la metástasis[105]. Por otra parte, diferentes estudios sobre *SMAD4*, un gen que codifica una proteína involucrada en esta vía, indican que una menor expresión de este gen es indicativo de un peor pronóstico del CCR[106–108]. Además, otros estudios en *SMAD7* indican que una mayor expresión de este gen están correlacionados con una mayor proliferación de las células cancerosas en pacientes con CCR[109,110]. Nuestros resultados siguen la línea de estos estudios, ya que los niveles de hidroximetilación encontrados en los *enhancers* que regulan *SMAD4* y *SMAD7* sugieren que se encuentran infraexpresado y sobreexpresado, respectivamente, en las muestras con CCR. Por su parte, la **ruta de las integrinas** también juega un papel clave en la iniciación, progresión y la metástasis de tumores[111]. De hecho, existen multitud de estudios que han conseguido identificar diferentes integrinas que podría tener un papel importante en la progresión del cáncer, incluyendo el CCR[112,113]. Los resultados obtenidos en nuestros datos no resultaron fáciles de interpretar en este caso, ya que encontramos *enhancers* con una hidroximetilación tanto mayor como menor en CCR en comparación con controles. No obstante, nuestros resultados sugirieron una mayor expresión de genes como *LAMC1* o *RPS6KBP1*, genes cuya expresión también se ha encontrado aumentada en pacientes con CCR en diferentes estudios[114,115].

Aunque los resultados recogidos en esta memoria son preliminares y, por tanto, serían necesarias diferentes validaciones funcionales y moleculares, estos resultados podrían apoyar la idea ya existente de utilizar estas rutas biológicas como dianas terapéuticas para el CCR[116,117].

4.1.4. Identificación de subgrupos de pacientes con CCR

Por último, se utilizó un algoritmo de aprendizaje automático no supervisado como K-means en los datos de hidroximetilación, con el objetivo de identificar posibles subgrupos de pacientes con CCR y comprobar, a su vez, si se podían diferenciar de esta manera de los casos con AAR o controles.

Mediante este algoritmo se pudieron identificar 6 clústeres de acuerdo con los valores de hidroximetilación en sus *enhancers*, dos de ellos conformados principalmente por muestras con CCR. Los pacientes en estos dos clústeres se encontraron balanceados en cuanto a género, y tanto la distribución de la edad como la etnia siguieron una distribución normal, por lo que la separación de estos grupos no parece ser debida a estas covariables. Además, estos clústeres de pacientes fueron muy heterogéneos en cuanto a la fase en la que se encontraba el cáncer, no consiguiendo separar pacientes en estadios tempranos de pacientes en estadios avanzados. Por otro lado, en los cuatro clústeres restantes pudimos comprobar como la diferencia entre pacientes con AAR y controles no fue la suficiente como para poder separarlos mediante este algoritmo, agrupándolos en los mismos clústeres. Por último, mediante la visualización de la distribución de puntos como mediante el coeficiente de silueta medio, se puede concluir que la diferenciación de clústeres no es muy clara, ya que la distancia entre los mismos fue muy baja.

Analizándolo los resultados en conjunto y en el contexto de la detección temprana del CCR, la utilización de K-means con datos de hidroximetilación no parece lo más adecuado, ya que no conseguimos diferenciar las fases tempranas de las tardías, y la diferenciación entre clústeres fue baja. No obstante, podrían explorarse otros algoritmos de aprendizaje automático no supervisado que permitan identificar patrones más claros y distintivos entre los pacientes con CCR, AAR y controles; o combinar los datos de hidroximetilación con otros biomarcadores para lograr una mejor diferenciación de los subgrupos de pacientes con CCR[70].

4.2. Conclusiones

Con la realización de este Trabajo de Fin de Máster se han alcanzado las siguientes conclusiones:

1. Se han comprendido y procesado datos públicos de hidroximetilación en *enhancers* procedentes de muestras de pacientes con CCR, AAR y controles, generando un set de datos listo para su utilización como entrada de algoritmos de aprendizaje automático.
2. A partir de este set de datos se ajustó un modelo SVM con kernel lineal capaz de diferenciar los pacientes con CCR de controles con una especificidad y sensibilidad del 82% y 84%, respectivamente. Estos resultados sugieren que los niveles de hidroximetilación son un biomarcador útil para la identificación del CCR.
    a. Tras ajustar la especificidad del modelo al 95%, se comprobó que la sensibilidad del modelo generado (58%) mejoraba la sensibilidad alcanzada por el modelo publicado en el artículo original de donde fueron extraídos los datos.
    b. A partir de los coeficientes del modelo se derivaron las importancias relativas de cada *enhancer*, señalando a los enhancers que regulan genes como *MYSM1*, *SP1*, o *ZNF469* como los más relevantes a la hora de realizar la clasificación de las muestras en CCR o control.
3. La ruta del TGF-β y la ruta de las integrinas, que cumplen un papel importante en la carcinogénesis, fueron identificadas como las rutas más afectadas por los cambios en los niveles de hidroximetilación.
4. Tras aplicar K-means sobre los datos de hidroximetilación las muestras se separaron en 6 clústeres, dos de ellos formados principalmente por muestras con CCR.
    a. Tras analizar estos dos clústeres, se comprobó que las muestras que los conformaban pertenecían a diferentes fases del desarrollo del cáncer.
    b. La diferenciación entre clústeres no fue muy clara, obteniendo clústeres muy cercanos unos de otros. Partiendo de datos de hidroximetilación, el uso de aprendizaje no supervisado para la detección temprana del CCR no parece lo más óptimo.

4.3. Líneas de trabajo futuras

Los resultados de este trabajo pueden derivar en posibles líneas de investigación futuras como, por ejemplo:

1. **Confirmación de los resultados obtenidos en una cohorte externa**: Los resultados obtenidos con el set de datos públicos es prometedor. No obstante, el origen de las muestras incluidas en el mismo es, principalmente, Estados Unidos. Sería adecuado evaluar la reproducibilidad y robustez de los resultados alcanzados haciendo uso de un set de datos con muestras de otra población.
2. **Integración con otros datos clínicos**: En el set de datos con el que se ha trabajado solo existían covariables básicas de cada una de las muestras, como género o edad. No obstante, el cáncer es una enfermedad compleja que implica diferentes modificaciones moleculares. Por este motivo, sería interesante integrar a los datos de hidroximetilación otros datos clínicos como los niveles de colesterol, antecedentes familiares o consumo de alcohol o tabaco.
3. **Integración con otros datos ómicos**: Como se ha discutido a lo largo de la memoria, en la actualidad existen diferentes datos ómicos (genómicos, transcriptómicos) que han mostrado su utilidad en la detección del CCR en sus etapas más tempranas. La integración de los datos de hidroximetilación con estos tipos de datos podría potenciar la capacidad de discriminación de los algoritmos de aprendizaje automático.
4. **Utilización de otros algoritmos de aprendizaje automático**: Debido a la limitación temporal de este trabajo, se ha tenido que escoger un conjunto limitado de algoritmos de aprendizaje automático. Por este motivo, se deberían utilizar y comparar los resultados obtenidos con otros algoritmos, supervisados y no supervisados, con los resultados expuestos en esta memoria.
5. **Validación de genes y rutas candidatas**: Sería interesante realizar un análisis funcional detallado de los *enhancers* relevantes para la clasificación de las muestras en CCR y control, como los que regulan los genes *MYSM1*, *SP1* y *ZNF469*. Se debería priorizar la comprobación de la modificación de los niveles de expresión de estos genes en muestras con CCR y control. De la misma manera, se debería analizar y comprender de una manera adecuada cómo afectan los niveles de hidroximetilación a las rutas del TGF-β y las integrinas.

Glosario

En este glosario se reúnen todas las abreviaciones utilizadas durante esta memoria, así como una breve explicación de los términos que se consideren fuera del ámbito del Máster en Ciencia de Datos de la UOC. Los términos relacionados con el CCR se marcan en **azul**, mientras que los términos relacionados con aprendizaje automático se marcan en **verde**. Por orden alfabético según la abreviación:

Término	Abreviación	Definición
Adenoma de alto riesgo	AAR	Tumor epitelial benigno que se asemeja a una glándula. Se considera de alto riesgo cuando su tamaño es mayor a 1 centímetro y presenta células con cambios precancerosos.
ADN tumoral circulante	ADNtc	ADN libre que proviene de células cancerosas y tumores.
Área bajo la curva	AUC	-
Compromiso de Comportamiento Ético y Global	CCEG	-
Cáncer colorrectal	CCR	Cáncer que se genera en el colon o el recto.
Cross-Industry Standard Process for Data Mining	CRISP-DM	Estrategia que proporciona un marco estructurado para la planificación, ejecución y seguimiento de proyectos de minería de datos
Diseño e implementación	DI	-
Análisis exploratorio de datos	EDA	-
Falso negativo	FN	-
Falso positivo	FP	-
Gradient Boosting	GB	-
Selección de los k valores más altos	KBest	Técnica de selección de características para seleccionar las K características más relevantes y útiles de un conjunto de datos, siendo K el un número específico de características final.
K-vecinos más cercanos	KNN	-

Término	Abreviación	Definición
Regresión logística	LR	-
Error absoluto medio	MAE	-
micro-ARNs	miR	Pequeñas moléculas de ARN que pueden modificar la expresión génica.
Error cuadrático medio	MSE	-
Naïve Bayes	NB	-
Objetivos de Desarrollo Sostenible	ODS	Objetivos adoptados por la Asamblea General de las Naciones Unidas que aborda temas como la pobreza, el cambio climático, o la desigualdad.
Análisis de componentes principales	PCA	-
Prueba de Evaluación Continua	PEC	-
Random Forest	RF	-
Eliminación recursiva de características	RFE	-
Raíz cuadrada del error cuadrático medio	RMSE	-
Support Vector Machine	SVM	-
Verdadero negativo	TN	-
Verdadero positivo	TP	-
Glucano	-	Polisacárido formado por unidades de glucosa
Mortalidad ajustada por edad	-	Tasas de mortalidad libre del efecto de la edad para comparar la mortalidad entre regiones con distintas proporciones demográficas.
Colonoscopia	-	Procedimiento en el que un médico examina el recto y el colon haciendo uso de un tubo estrecho y flexible con el objetivo identificar el cáncer colorrectal o los pólipos precancerosos
Países desarrollados	-	Países con niveles altos o muy altos de vida de acuerdo con el índice de desarrollo humano.

Término	Abreviación	Definición
Metilación e hidroximetilación	-	La metilación es la adición de un grupo metilo (-CH3) al ADN. La hidroximetilación es el proceso de añadir un grupo hidroxilo (-OH) sobre el grupo metilo anterior. Estas modificaciones pueden afectar la expresión génica. Se produce en las citosinas del ADN.
Mutación patogénica	-	Variantes encontradas en el ADN que generan una enfermedad.
Metabolito	-	Cualquier sustancia generada durante el metabolismo.
Insulinorresistencia	-	Estado en el que los tejidos responden en menor medida a la insulina. Esto ocurre, por ejemplo, en los pacientes con diabetes tipo II.
Etiología	-	Parte de la medicina que estudia el origen de la enfermedad.
Patogénesis	-	Parte de la medicina que estudia el desarrollo de la enfermedad.
Vía molecular	-	Interacción de moléculas que llevan a cabo una función dentro de la célula.
Colon sigmoide	-	Parte del colon en forma de S que se conecta con el recto.
ADN libre circulante	-	Moléculas de ADN que se encuentran fuera de las células y pueden ser detectadas en sangre u orina. En condiciones normales, los niveles de ADN libre son bajos y provienen principalmente de los procesos apoptóticos (muerte) de las células.
Enhancer	-	También conocidos como potenciadores, es una región de ADN en la que se pueden unir proteínas para aumentar los niveles de expresión de un gen.
Data leakage	-	Filtración de información del conjunto de entrenamiento al modelo de aprendizaje de manera no intencionada, lo que puede provocar una evaluación sesgada en la etapa de pruebas

Término	Abreviación	Definición
Blackbox	-	Modelo que genera predicciones sin que el usuario tenga un conocimiento completo de cómo se llega a las mismas.

Bibliografía

1. Cancer. Accessed March 5, 2023. https://www.who.int/news-room/fact-sheets/detail/cancer

2. Morgan E, Arnold M, Gini A, et al. Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN. *Gut*. 2023;72(2):338-344. doi:10.1136/gutjnl-2022-327736

3. Xi Y, Xu P. Global colorectal cancer burden in 2020 and projections to 2040. *Transl Oncol*. 2021;14(10):101174. doi:10.1016/j.tranon.2021.101174

4. Clinton SK, Giovannucci EL, Hursting SD. The World Cancer Research Fund/American Institute for Cancer Research Third Expert Report on Diet, Nutrition, Physical Activity, and Cancer: Impact and Future Directions. *J Nutr*. 2020;150(4):663-671. doi:10.1093/jn/nxz268

5. Hossain MdS, Karuniawati H, Jairoun AA, et al. Colorectal Cancer: A Review of Carcinogenesis, Global Epidemiology, Current Challenges, Risk Factors, Preventive and Treatment Strategies. *Cancers*. 2022;14(7):1732. doi:10.3390/cancers14071732

6. Li H, Boakye D, Chen X, Hoffmeister M, Brenner H. Association of Body Mass Index With Risk of Early-Onset Colorectal Cancer: Systematic Review and Meta-Analysis. *Am J Gastroenterol*. 2021;116(11):2173-2183. doi:10.14309/ajg.0000000000001393

7. Araghi M, Soerjomataram I, Bardot A, et al. Changes in colorectal cancer incidence in seven high-income countries: a population-based study. *Lancet Gastroenterol Hepatol*. 2019;4(7):511-518. doi:10.1016/S2468-1253(19)30147-5

8. Siegel RL, Miller KD, Goding Sauer A, et al. Colorectal cancer statistics, 2020. *CA Cancer J Clin*. 2020;70(3):145-164. doi:10.3322/caac.21601

9. Roman-Naranjo P, Gallego-Martinez A, Soto-Varela A, et al. Burden of Rare Variants in the OTOG Gene in Familial Meniere's Disease. *Ear Hear*. 2020;41(6):1598-1605. doi:10.1097/AUD.0000000000000878

10. Roman-Naranjo P, Moleon MDC, Aran I, et al. Rare coding variants involving MYO7A and other genes encoding stereocilia link proteins in familial meniere disease. *Hear Res*. 2021;409:108329. doi:10.1016/j.heares.2021.108329

11. Roman-Naranjo P, Parra-Perez AM, Escalera-Balsera A, et al. Defective α-tectorin may involve tectorial membrane in familial Meniere disease. *Clin Transl Med*. 2022;12(6):e829. doi:10.1002/ctm2.829

12. Amadix - Detección temprana de cáncer en sangre. Accessed March 7, 2023. https://amadix.com/es/

13. dpicampaigns. Take Action for the Sustainable Development Goals. United Nations Sustainable Development. Accessed March 10, 2023. https://www.un.org/sustainabledevelopment/sustainable-development-goals/

14. Schröer C, Kruse F, Gómez JM. A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Comput Sci*. 2021;181:526-534. doi:10.1016/j.procs.2021.01.199

15. Abancens M, Bustos V, Harvey H, McBryan J, Harvey BJ. Sexual Dimorphism in Colon Cancer. *Front Oncol*. 2020;10. Accessed March 17, 2023. https://www.frontiersin.org/articles/10.3389/fonc.2020.607909

16. Pillado MTS, Díaz SP, Barreiro VB, Martín CG. Chapter 1 - Incidence and mortality of CRC. In: Sierra AP, ed. *Foundations of Colorectal Cancer*. Academic Press; 2022:3-15. doi:10.1016/B978-0-323-90055-3.00034-X

17. Cancer (IARC) TIA for R on. Global Cancer Observatory. Accessed March 17, 2023. https://gco.iarc.fr/

18. Ghoncheh M, Mohammadian M, Mohammadian-Hafshejani A, Salehiniya H. The Incidence and Mortality of Colorectal Cancer and Its Relationship With the Human Development Index in Asia. *Ann Glob Health*. 2016;82(5):726-737. doi:10.1016/j.aogh.2016.10.004

19. Vabi BW, Gibbs JF, Parker GS. Implications of the growing incidence of global colorectal cancer. *J Gastrointest Oncol*. 2021;12(Suppl 2):S387-S398. doi:10.21037/jgo-2019-gi-06

20. Mortalidad Provincial. Accessed March 18, 2023. http://ariadna.cne.isciii.es/MapaP/

21. Bishehsari F, Mahdavinia M, Vacca M, Malekzadeh R, Mariani-Costantini R. Epidemiological transition of colorectal cancer in developing countries: environmental factors, molecular pathways, and opportunities for prevention. *World J Gastroenterol*. 2014;20(20):6055-6072. doi:10.3748/wjg.v20.i20.6055

22. Vanthomme K, Rosskamp M, De Schutter H, Vandenheede H. Colorectal cancer incidence and survival inequalities among labour immigrants in Belgium during 2004–2013. *Sci Rep*. 2022;12(1):15727. doi:10.1038/s41598-022-19322-1

23. Jung G, Hernández-Illán E, Moreira L, Balaguer F, Goel A. Epigenetics of colorectal cancer: biomarker and therapeutic potential. *Nat Rev Gastroenterol Hepatol*. 2020;17(2):111-130. doi:10.1038/s41575-019-0230-y

24. Lercher L, McDonough MA, El-Sagheer AH, et al. Structural insights into how 5-hydroxymethylation influences transcription factor binding. *Chem Commun*. 2014;50(15):1794-1796. doi:10.1039/C3CC48151D

25. Kim J, Lee HK. Potential Role of the Gut Microbiome In Colorectal Cancer Progression. *Front Immunol*. 2022;12:807648. doi:10.3389/fimmu.2021.807648

26. Bouvard V, Loomis D, Guyton KZ, et al. Carcinogenicity of consumption of red and processed meat. *Lancet Oncol*. 2015;16(16):1599-1600. doi:10.1016/S1470-2045(15)00444-1

27. Barrubés L, Babio N, Becerra-Tomás N, Rosique-Esteban N, Salas-Salvadó J. Association Between Dairy Product Consumption and Colorectal Cancer Risk in Adults: A Systematic Review and Meta-Analysis of Epidemiologic Studies. *Adv Nutr Bethesda Md*. 2019;10(suppl_2):S190-S211. doi:10.1093/advances/nmy114

28. He X, Wu K, Zhang X, et al. Dietary intake of fiber, whole grains and risk of colorectal cancer: An updated analysis according to food sources, tumor location and molecular subtypes in two large US cohorts. *Int J Cancer*. 2019;145(11):3040-3051. doi:10.1002/ijc.32382

29. Caini S, Chioccioli S, Pastore E, et al. Fish Consumption and Colorectal Cancer Risk: Meta-Analysis of Prospective Epidemiological Studies and Review of Evidence from Animal Studies. *Cancers*. 2022;14(3):640. doi:10.3390/cancers14030640

30. Li G, Ma D, Zhang Y, Zheng W, Wang P. Coffee consumption and risk of colorectal cancer: a meta-analysis of observational studies. *Public Health Nutr*. 2013;16(2):346-357. doi:10.1017/S1368980012002601

31. Cai S, Li Y, Ding Y, Chen K, Jin M. Alcohol drinking and the risk of colorectal cancer death: a meta-analysis. *Eur J Cancer Prev Off J Eur Cancer Prev Organ ECP*. 2014;23(6):532-539. doi:10.1097/CEJ.0000000000000076

32. Klarich DS, Brasser SM, Hong MY. Moderate Alcohol Consumption and Colorectal Cancer Risk. *Alcohol Clin Exp Res*. 2015;39(8):1280-1291. doi:10.1111/acer.12778

33. Bai X, Wei H, Liu W, et al. Cigarette smoke promotes colorectal cancer through modulation of gut microbiota and related metabolites. *Gut*. 2022;71(12):2439-2450. doi:10.1136/gutjnl-2021-325021

34. Ning Y, Wang L, Giovannucci EL. A quantitative analysis of body mass index and colorectal cancer: findings from 56 observational studies. *Obes Rev Off J Int Assoc Study Obes*. 2010;11(1):19-30. doi:10.1111/j.1467-789X.2009.00613.x

35. Oruç Z, Kaplan MA. Effect of exercise on colorectal cancer prevention and treatment. *World J Gastrointest Oncol*. 2019;11(5):348-366. doi:10.4251/wjgo.v11.i5.348

36. Grady WM, Markowitz SD. The molecular pathogenesis of colorectal cancer and its potential application to colorectal cancer screening. *Dig Dis Sci*. 2015;60(3):762-772. doi:10.1007/s10620-014-3444-4

37. Nguyen HT, Duong HQ. The molecular characteristics of colorectal cancer: Implications for diagnosis and therapy. *Oncol Lett*. 2018;16(1):9-18. doi:10.3892/ol.2018.8679

38. Poulogiannis G, Frayling IM, Arends MJ. DNA mismatch repair deficiency in sporadic colorectal cancer and Lynch syndrome. *Histopathology*. 2010;56(2):167-179. doi:10.1111/j.1365-2559.2009.03392.x

39. Evrard C, Tachon G, Randrian V, Karayan-Tapon L, Tougeron D. Microsatellite Instability: Diagnosis, Heterogeneity, Discordance, and Clinical Impact in Colorectal Cancer. *Cancers*. 2019;11(10):1567. doi:10.3390/cancers11101567

40. De Palma FDE, D'Argenio V, Pol J, Kroemer G, Maiuri MC, Salvatore F. The Molecular Hallmarks of the Serrated Pathway in Colorectal Cancer. *Cancers*. 2019;11(7):1017. doi:10.3390/cancers11071017

41. Nayor J, Saltzman JR. Colonoscopy quality: measuring the patient experience. *Endoscopy*. 2018;50(1):4-5. doi:10.1055/s-0043-121146

42. Wu W, Huang J, Yang Y, et al. Adherence to colonoscopy in cascade screening of colorectal cancer: A systematic review and meta-analysis. *J Gastroenterol Hepatol*. 2022;37(4):620-631. doi:10.1111/jgh.15762

43. Kamel F, Eltarhoni K, Nisar P, Soloviev M. Colorectal Cancer Diagnosis: The Obstacles We Face in Determining a Non-Invasive Test and Current Advances in Biomarker Detection. *Cancers*. 2022;14(8):1889. doi:10.3390/cancers14081889

44. van der Geest LGM, Lam-Boer J, Koopman M, Verhoef C, Elferink MAG, de Wilt JHW. Nationwide trends in incidence, treatment and survival of colorectal cancer patients with synchronous metastases. *Clin Exp Metastasis*. 2015;32(5):457-465. doi:10.1007/s10585-015-9719-0

45. Quintero E. ¿Test químico o test inmunológico para la detección de sangre oculta en heces en el cribado del cáncer colorrectal? *Gastroenterol Hepatol*. 2009;32(8):565-576. doi:10.1016/j.gastrohep.2009.01.179

46. Kolligs FT. Diagnostics and Epidemiology of Colorectal Cancer. *Visc Med*. 2016;32(3):158-164. doi:10.1159/000446488

47. Lee JK, Liles EG, Bent S, Levin TR, Corley DA. Accuracy of Fecal Immunochemical Tests for Colorectal Cancer: Systematic Review and Meta-analysis. *Ann Intern Med*. 2014;160(3):171. doi:10.7326/M13-1484

48. Mo S, Dai W, Wang H, et al. Early detection and prognosis prediction for colorectal cancer by circulating tumour DNA methylation haplotypes: a multicentre cohort study. *eClinicalMedicine*. 2023;55. doi:10.1016/j.eclinm.2022.101717

49. Knudsen AB, Rutter CM, Peterse EFP, et al. Colorectal Cancer Screening: An Updated Modeling Study for the US Preventive Services Task Force. *JAMA*. 2021;325(19):1998-2011. doi:10.1001/jama.2021.5746

50. Lin JS, Perdue LA, Henrikson NB, Bean SI, Blasi PR. Screening for Colorectal Cancer: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. *JAMA*. 2021;325(19):1978-1998. doi:10.1001/jama.2021.4417

51. Gupta S. Screening for Colorectal Cancer. *Hematol Oncol Clin North Am*. 2022;36(3):393-414. doi:10.1016/j.hoc.2022.02.001

52. Palmqvist R, Engarås B, Lindmark G, et al. Prediagnostic levels of carcinoembryonic antigen and CA 242 in colorectal cancer: a matched case-control study. *Dis Colon Rectum*. 2003;46(11):1538-1544. doi:10.1007/s10350-004-6810-z

53. Hernandez BY, Frierson HF, Moskaluk CA, et al. CK20 and CK7 protein expression in colorectal cancer: demonstration of the utility of a population-based tissue microarray. *Hum Pathol*. 2005;36(3):275-281. doi:10.1016/j.humpath.2005.01.013

54. Wang H, Jin S, Lu H, et al. Expression of survivin, MUC2 and MUC5 in colorectal cancer and their association with clinicopathological characteristics. *Oncol Lett*. 2017;14(1):1011-1016. doi:10.3892/ol.2017.6218

55. Frattini M, Gallino G, Signoroni S, et al. Quantitative and qualitative characterization of plasma DNA identifies primary and recurrent colorectal cancer. *Cancer Lett*. 2008;263(2):170-181. doi:10.1016/j.canlet.2008.03.021

56. El-Gayar D, El-Abd N, Hassan N, Ali R. Increased Free Circulating DNA Integrity Index as a Serum Biomarker in Patients with Colorectal Carcinoma. *Asian Pac J Cancer Prev APJCP*. 2016;17(3):939-944. doi:10.7314/apjcp.2016.17.3.939

57. Hao TB, Shi W, Shen XJ, et al. Circulating cell-free DNA in serum as a biomarker for diagnosis and prognostic prediction of colorectal cancer. *Br J Cancer*. 2014;111(8):1482-1489. doi:10.1038/bjc.2014.470

58. Liu T, Liu D, Guan S, Dong M. Diagnostic role of circulating MiR-21 in colorectal cancer: a update meta-analysis. *Ann Med*. 2021;53(1):87-102. doi:10.1080/07853890.2020.1828617

59. Pal S, Garg M, Pandey AK. Biomarkers as Putative Therapeutic Targets in Colorectal Cancer. In: Nagaraju GP, Shukla D, Vishvakarma NK, eds. *Colon Cancer Diagnosis and Therapy: Volume 1*. Springer International Publishing; 2021:123-177. doi:10.1007/978-3-030-63369-1_8

60. Nikolouzakis TK, Vassilopoulou L, Fragkiadaki P, et al. Improving diagnosis, prognosis and prediction by using biomarkers in CRC patients (Review). *Oncol Rep*. 2018;39(6):2455-2472. doi:10.3892/or.2018.6330

61. Davri A, Birbas E, Kanavos T, et al. Deep Learning on Histopathological Images for Colorectal Cancer Diagnosis: A Systematic Review. *Diagnostics*. 2022;12(4):837. doi:10.3390/diagnostics12040837

62. Mansur A, Saleem Z, Elhakim T, Daye D. Role of artificial intelligence in risk prediction, prognostication, and therapy response assessment in colorectal cancer: current state and future directions. *Front Oncol*. 2023;13:1065402. doi:10.3389/fonc.2023.1065402

63. Gironés J, Casas J, Minguillón J, Caihuleas R. *Minería de Datos*. Editorial UOC; 2017. Accessed March 24, 2023. https://www.editorialuoc.cat/mineria-de-datos

64. Kennion O, Maitland S, Brady R. Machine learning as a new horizon for colorectal cancer risk prediction? A systematic review. *Health Sci Rev*. 2022;4:100041. doi:10.1016/j.hsr.2022.100041

65. Kinar Y, Kalkstein N, Akiva P, et al. Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study. *J Am Med Inform Assoc*. 2016;23(5):879-890. doi:10.1093/jamia/ocv195

66. Hoogendoorn M, Szolovits P, Moons LMG, Numans ME. Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer. *Artif Intell Med*. 2016;69:53-61. doi:10.1016/j.artmed.2016.03.003

67. Kop R, Hoogendoorn M, Teije A ten, et al. Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records. *Comput Biol Med*. 2016;76:30-38. doi:10.1016/j.compbiomed.2016.06.019

68. Ivancic MM, Megna BW, Sverchkov Y, et al. Noninvasive Detection of Colorectal Carcinomas Using Serum Protein Biomarkers. *J Surg Res*. 2020;246:160-169. doi:10.1016/j.jss.2019.08.004

69. Pan Y, Zhang L, Zhang R, et al. Screening and diagnosis of colorectal cancer and advanced adenoma by Bionic Glycome method and machine learning. *Am J Cancer Res*. 2021;11(6):3002.

70. Florensa D, Mateo-Fornés J, Solsona F, et al. Use of Multiple Correspondence Analysis and K-means to Explore Associations Between Risk Factors and Likelihood of Colorectal Cancer: Cross-sectional Study. *J Med Internet Res*. 2022;24(7):e29056. doi:10.2196/29056

71. PyCharm: the Python IDE for Professional Developers by JetBrains. JetBrains. Accessed May 17, 2023. https://www.jetbrains.com/pycharm/

72. Walker NJ, Rashid M, Yu S, et al. Hydroxymethylation profile of cell-free DNA is a biomarker for early colorectal cancer. *Sci Rep*. 2022;12(1):16566. doi:10.1038/s41598-022-20975-1

73. Walker NJ, Rashid M, Yu S, et al. [Dataset] Hydroxymethylation profile of cell free DNA is a biomarker for early colorectal cancer. Accessed May 17, 2023. https://zenodo.org/record/5170265#.ZGSpgHZBxD-

74. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011;27(12):1739-1740. doi:10.1093/bioinformatics/btr260

75. Díez-Villanueva A, Mallona I, Peinado MA. Wanderer, an interactive viewer to explore DNA methylation and gene expression data in human cancer. *Epigenetics Chromatin*. 2015;8(1):22. doi:10.1186/s13072-015-0014-8

76. Sjöstedt E, Zhong W, Fagerberg L, et al. An atlas of the protein-coding genes in the human, pig, and mouse brain. *Science*. 2020;367(6482):eaay5947. doi:10.1126/science.aay5947

77. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res*. 2009;37(Web Server issue):W305-311. doi:10.1093/nar/gkp427

78. Li Y, Li J, Liu H, Liu Y, Cui B. Expression of MYSM1 is associated with tumor progression in colorectal cancer. *PLoS ONE*. 2017;12(5):e0177235. doi:10.1371/journal.pone.0177235

79. Chen X, Wang W, Li Y, et al. MYSM1 inhibits human colorectal cancer tumorigenesis by activating miR-200 family members/CDH1 and blocking PI3K/AKT signaling. *J Exp Clin Cancer Res*. 2021;40(1):341. doi:10.1186/s13046-021-02106-2

80. Lin YH, Wang H, Fiore A, et al. Loss of MYSM1 inhibits the oncogenic activity of cMYC in B cell lymphoma. *J Cell Mol Med*. 2021;25(14):7089-7094. doi:10.1111/jcmm.16554

81. Kiang KMY, Zhang P, Li N, Zhu Z, Jin L, Leung GKK. Loss of cytoskeleton protein ADD3 promotes tumor growth and angiogenesis in glioblastoma multiforme. *Cancer Lett*. 2020;474:118-126. doi:10.1016/j.canlet.2020.01.007

82. Li L, Lou Z, Wang L. The role of FKBP5 in cancer aetiology and chemoresistance. *Br J Cancer*. 2011;104(1):19-23. doi:10.1038/sj.bjc.6606014

83. Takami Y, Russell MB, Gao C, et al. Sp1 regulates osteopontin expression in SW480 human colon adenocarcinoma cells. *Surgery*. 2007;142(2):163-169. doi:10.1016/j.surg.2007.02.015

84. Zhao Y, Zhang W, Guo Z, et al. Inhibition of the transcription factor Sp1 suppresses colon cancer stem cell growth and induces apoptosis in vitro and in nude mouse xenografts. *Oncol Rep*. 2013;30(4):1782-1792. doi:10.3892/or.2013.2627

85. Bajpai R, Nagaraju GP. Specificity protein 1: Its role in colorectal cancer progression and metastasis. *Crit Rev Oncol Hematol*. 2017;113:1-7. doi:10.1016/j.critrevonc.2017.02.024

86. Zhang X, Yao J, Shi H, et al. Hsa_circ_0026628 promotes the development of colorectal cancer by targeting SP1 to activate the Wnt/β-catenin pathway. *Cell Death Dis*. 2021;12(9):802. doi:10.1038/s41419-021-03794-6

87. Beishline K, Azizkhan-Clifford J. Sp1 and the "hallmarks of cancer." *FEBS J*. 2015;282(2):224-258. doi:10.1111/febs.13148

88. Ma B, Luo Y, Xu W, et al. LINC00886 Negatively Regulates Malignancy in Anaplastic Thyroid Cancer. *Endocrinology*. 2023;164(4):bqac204. doi:10.1210/endocr/bqac204

89. Xie M, Cai L, Li J, et al. FAM110B Inhibits Non-Small Cell Lung Cancer Cell Proliferation and Invasion Through Inactivating Wnt/β-Catenin Signaling. *OncoTargets Ther*. 2020;13:4373-4384. doi:10.2147/OTT.S247491

90. Tian H, Li X, Jiang W, et al. High expression of AKR1C1 is associated with proliferation and migration of small-cell lung cancer cells. *Lung Cancer Targets Ther*. 2016;7:53-61. doi:10.2147/LCTT.S90694

91. Pekow J, Hernandez K, Meckel K, et al. IBD-associated Colon Cancers Differ in DNA Methylation and Gene Expression Profiles Compared With Sporadic Colon Cancers. *J Crohns Colitis*. 2019;13(7):884-893. doi:10.1093/ecco-jcc/jjz014

92. Yamashita K, Hosoda K, Nishizawa N, Katoh H, Watanabe M. Epigenetic biomarkers of promoter DNA methylation in the new era of cancer treatment. *Cancer Sci*. 2018;109(12):3695-3706. doi:10.1111/cas.13812

93. Li P, Liu S, Du L, Mohseni G, Zhang Y, Wang C. Liquid biopsies based on DNA methylation as biomarkers for the detection and prognosis of lung cancer. *Clin Epigenetics*. 2022;14:118. doi:10.1186/s13148-022-01337-0

94. Belle VV, Calster BV, Huffel SV, Suykens JAK, Lisboa P. Explaining Support Vector Machines: A Color Based Nomogram. *PLOS ONE*. 2016;11(10):e0164568. doi:10.1371/journal.pone.0164568

95. Wan JCM, Massie C, Garcia-Corbacho J, et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat Rev Cancer*. 2017;17(4):223-238. doi:10.1038/nrc.2017.7

96. Church TR, Wandell M, Lofton-Day C, et al. Prospective evaluation of methylated SEPT9 in plasma for detection of asymptomatic colorectal cancer. *Gut*. 2014;63(2):317-325. doi:10.1136/gutjnl-2012-304149

97. Johnson DA, Barclay RL, Mergener K, et al. Plasma Septin9 versus Fecal Immunochemical Testing for Colorectal Cancer Screening: A Prospective Multicenter Study. *PLoS ONE*. 2014;9(6):e98238. doi:10.1371/journal.pone.0098238

98. Barták BK, Kalmár A, Péterfia B, et al. Colorectal adenoma and cancer detection based on altered methylation pattern of SFRP1, SFRP2, SDC2, and PRIMA1 in plasma samples. *Epigenetics*. 2017;12(9):751-763. doi:10.1080/15592294.2017.1356957

99. Li W, Zhang X, Lu X, et al. 5-Hydroxymethylcytosine signatures in circulating cell-free DNA as diagnostic biomarkers for human cancers. *Cell Res*. 2017;27(10):1243-1257. doi:10.1038/cr.2017.121

100. Wan N, Weinberg D, Liu TY, et al. Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA. *BMC Cancer*. 2019;19(1):832. doi:10.1186/s12885-019-6003-8

101. Zou QT, Lin Y, Luo QY. miR-138-5p inhibits the progression of colorectal cancer via regulating SP1/LGR5 axis. *Cell Biol Int*. 2023;47(1):273-282. doi:10.1002/cbin.11926

102. Gordon KJ, Blobe GC. Role of transforming growth factor-beta superfamily signaling pathways in human disease. *Biochim Biophys Acta*. 2008;1782(4):197-228. doi:10.1016/j.bbadis.2008.01.006

103. Hong S, Lee C, Kim SJ. Smad7 Sensitizes Tumor Necrosis Factor–Induced Apoptosis through the Inhibition of Antiapoptotic Gene Expression by Suppressing Activation of the Nuclear Factor-κB Pathway. *Cancer Res*. 2007;67(19):9577-9583. doi:10.1158/0008-5472.CAN-07-1179

104. Massagué J, Blain SW, Lo RS. TGFbeta signaling in growth control, cancer, and heritable disorders. *Cell*. 2000;103(2):295-309. doi:10.1016/s0092-8674(00)00121-5

105. Geng L, Chaudhuri A, Talmon G, Wisecarver JL, Wang J. TGF-Beta Suppresses VEGFA-Mediated Angiogenesis in Colon Cancer Metastasis. *PLOS ONE*. 2013;8(3):e59918. doi:10.1371/journal.pone.0059918

106. Roth AD, Delorenzi M, Tejpar S, et al. Integrated analysis of molecular and clinical prognostic factors in stage II/III colon cancer. *J Natl Cancer Inst*. 2012;104(21):1635-1646. doi:10.1093/jnci/djs427

107. Voorneveld PW, Jacobs RJ, Kodach LL, Hardwick JCH. A Meta-Analysis of SMAD4 Immunohistochemistry as a Prognostic Marker in Colorectal Cancer. *Transl Oncol*. 2015;8(1):18-24. doi:10.1016/j.tranon.2014.11.003

108. Mizuno T, Cloyd JM, Vicente D, et al. SMAD4 gene mutation predicts poor prognosis in patients undergoing resection for colorectal liver metastases. *Eur J Surg Oncol J Eur Soc Surg Oncol Br Assoc Surg Oncol*. 2018;44(5):684-692. doi:10.1016/j.ejso.2018.02.247

109. Troncone E, Monteleone G. Smad7 and Colorectal Carcinogenesis: A Double-Edged Sword. *Cancers*. 2019;11(5):612. doi:10.3390/cancers11050612

110. Rosic J, Dragicevic S, Miladinov M, et al. SMAD7 and SMAD4 expression in colorectal cancer progression and therapy response. *Exp Mol Pathol*. 2021;123:104714. doi:10.1016/j.yexmp.2021.104714

111. Desgrosellier JS, Cheresh DA. Integrins in cancer: biological implications and therapeutic opportunities. *Nat Rev Cancer*. 2010;10(1):9-22. doi:10.1038/nrc2748

112. Bates RC, Bellovin DI, Brown C, et al. Transcriptional activation of integrin β6 during the epithelial-mesenchymal transition defines a novel prognostic indicator of aggressive colon carcinoma. *J Clin Invest*. 2005;115(2):339-347. doi:10.1172/JCI200523183

113. Chen JR, Zhao JT, Xie ZZ. Integrin-mediated cancer progression as a specific target in clinical therapy. *Biomed Pharmacother*. 2022;155:113745. doi:10.1016/j.biopha.2022.113745

114. Li J, Liu Q, Huang X, et al. Transcriptional Profiling Reveals the Regulatory Role of CXCL8 in Promoting Colorectal Cancer. *Front Genet*. 2020;10:1360. doi:10.3389/fgene.2019.01360

115. J L, J G, J K, et al. A functional polymorphism located at transcription factor binding sites, rs6695837 near LAMC1 gene, confers risk of colorectal cancer in Chinese populations. *Carcinogenesis*. 2017;38(2). doi:10.1093/carcin/bgw204

116. Kim BG, Malek E, Choi SH, Ignatz-Hoover JJ, Driscoll JJ. Novel therapies emerging in oncology to target the TGF-β pathway. *J Hematol OncolJ Hematol Oncol*. 2021;14(1):55. doi:10.1186/s13045-021-01053-x

117. Bergonzini C, Kroese K, Zweemer AJM, Danen EHJ. Targeting Integrins for Cancer Therapy - Disappointments and Opportunities. *Front Cell Dev Biol*. 2022;10. Accessed June 8, 2023. https://www.frontiersin.org/articles/10.3389/fcell.2022.863850

Anexos

Anexo 1: Biomarcadores clasificados por utilidad clínica en el CCR.

Utilidad clínica		Origen	Biomarcador
Biomarcadores de diagnóstico	No invasivo	Heces	SOH-química
			SOH-inmunulógica
			ADN en heces
			miARNs en heces
	Invasivo	Sangre	ctADN
			miARNs
			EV-miRNAs
			lncRNAs
		Tejido	Citoqueratinas
			β-Catenina
			Villina
			CDX2
			SATB2
			Mucina
			Cadherina 17
			Telomerasa
			Glicoproteína A33 (GPA33)
Biomarcadores de pronóstico		Sangre	CEA
		Tejido	BRAF
			MSI
			APC
			p53
			SMAD4
			miRNAs
			lncRNAs
Biomarcadores predictivos		Tejido	KRAS
			BRAF
		Sangre	PI3Ks
			cfDNA

Anexo 2: Hiperparámetros evaluados para cada uno de los modelos supervisados.

Modelos	Hiperparámetros
SVM lineal	**C**: [0.0001, 0.001, 0.01, 0.1, 1, 10] **kernel**: "linear", **gamma**: [0, 1, "scale", “auto”], **random_state**: seed, **probability**: True
SVM rbf	**C**: [0.0001, 0.001, 0.01, 0.1, 1, 10], **kernel**: "rbf", **gamma**: [0, 1, "scale", "auto"], **random_state**: seed, **probability**: True
Árbol de decisión	**min_samples_leaf**: [1, 2, 4, 8, 10], **max_depth**: range(1, 10), **min_samples_split**: range(1, 10), criterion: ['gini', 'entropy'], **random_state**: seed
Gradient Boosting	**learning_rate**: [0.05, 0.075, 0.1, 0.15, 0.2], **min_samples_split**: [2, 5, 10], **min_samples_leaf**: [1, 2, 4, 8, 10], **max_depth**: [3, 5, 8], **max_features**: ["log2"], **criterion**: ["friedman_mse", "squared_error"], **n_estimators**: [100, 500, 1000], **random_state**: seed
Random Forest	**n_estimators**: [200, 300, 500, 1000], **criterion**: ['gini', 'entropy'], **max_depth**: [3, 5, 8], **min_samples_split**: [2, 5, 10], **min_samples_leaf**: [1, 2, 4, 8, 10], **max_features**: ['sqrt', 'log2'], **random_state**: [seed]

SVM: Support Vector Machine; rbf: Función de base radial.

Anexo 3: Hiperparámetros seleccionados para cada uno de los modelos supervisados.

Método de selección de características	Modelo	Hiperparámetros seleccionados
t-test	SVM-lineal	C=0.01, gamma=0, kernel='linear', random_state=23, probability=True
	SVM-rbf	C=10, gamma='scale', kernel='rbf', random_state=23, probability=True
	GB	learning_rate=0.2, max_depth=3, max_features='log2', min_samples_leaf=10, min_samples_split=5, n_estimators=1000, criterion='friedman_mse'
	DT	criterion='entropy', max_depth=7, min_samples_leaf=1, min_samples_split=6, random_state=23
	RF	criterion='gini', max_depth=8, max_features='sqrt', min_samples_leaf=2, min_samples_split=10, n_estimators=300
KBest (100)	SVM-lineal	C=0.01, gamma=0, kernel='linear', random_state=23, probability=True
	SVM-rbf	C=10, gamma='scale', kernel='rbf', random_state=23, probability=True
	GB	GradientBoostingClassifier(learning_rate=0.15, max_depth=8, max_features='log2', min_samples_leaf=1, min_samples_split=5, n_estimators=1000, criterion='squared_error')
	DT	criterion='entropy', max_depth=5, min_samples_leaf=2, min_samples_split=8, random_state=23
	RF	criterion='gini', max_depth=8, max_features='sqrt', min_samples_leaf=2, min_samples_split=2, n_estimators=200
KBest (200)	SVM-lineal	C=0.01, gamma=0, kernel='linear', random_state=23, probability=True
	SVM-rbf	C=10, gamma='scale', kernel='rbf', random_state=23, probability=True
	GB	learning_rate=0.1, max_depth=8, max_features='log2', min_samples_leaf=1, min_samples_split=10, n_estimators=500, criterion='squared_error'
	DT	criterion='entropy', max_depth=6, min_samples_leaf=10, min_samples_split=2, random_state=23
	RF	criterion='entropy', max_depth=8, max_features='sqrt', min_samples_leaf=2, min_samples_split=10, n_estimators=300
KBest (300)	SVM-lineal	C=0.01, gamma=0, kernel='linear', random_state=23, probability=True
	SVM-rbf	C=10, gamma='scale', kernel='rbf', random_state=23, probability=True
	GB	learning_rate=0.05, max_depth=8, max_features='log2', min_samples_leaf=10, min_samples_split=5, n_estimators=500, criterion='friedman_mse', random_state=23
	DT	criterion='gini', max_depth=8, min_samples_leaf=10, min_samples_split=2, random_state=23
	RF	criterion='entropy', max_depth=8, max_features='sqrt', min_samples_leaf=2, min_samples_split=10, n_estimators=300, random_state=23

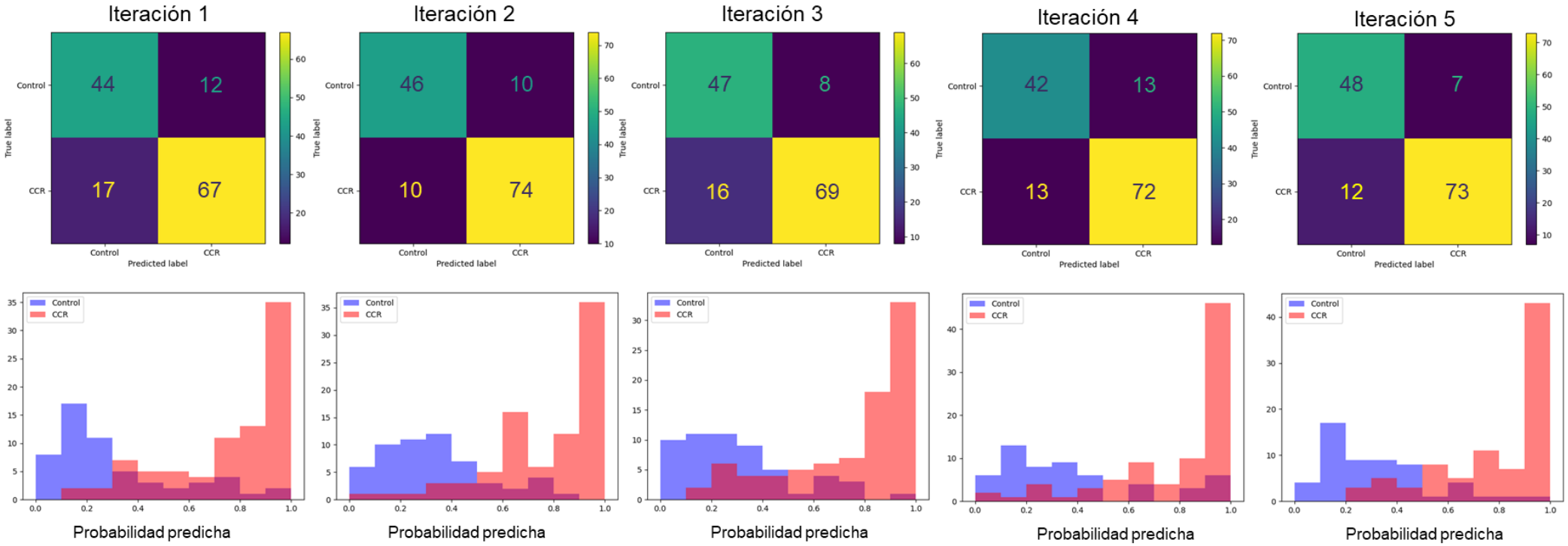
Anexo 4: Métricas de evaluación de modelos supervisados según método de selección de características.

Modelo	Selección	Exactitud	Sensibilidad	Precisión	F1	Especificidad	Brier	Log-loss	AUC-ROC	AUC-PR
SVM lineal	T-test	0.83+/-0.03	0.84+/-0.03	0.88+/-0.03	0.86+/-0.02	0.82+/-0.04	0.13+/-0.01	0.44+/-0.07	0.89+/-0.03	0.92+/-0.02
SVM rbf		0.83+/-0.04	0.86+/-0.04	0.87+/-0.03	0.86+/-0.03	0.80+/-0.05	0.12+/-0.02	0.40+/-0.06	0.90+/-0.03	0.93+/-0.02
Gradient Boosting		0.84+/-0.03	0.87+/-0.04	0.86+/-0.03	0.87+/-0.03	0.79+/-0.06	0.14+/-0.03	0.85+/-0.26	0.90+/-0.03	0.93+/-0.02
Decision Tree		0.65+/-0.06	0.73+/-0.07	0.70+/-0.05	0.71+/-0.05	0.53+/-0.09	0.31+/-0.04	9.81+/-1.24	0.66+/-0.04	0.80+/-0.02
Random Forest		0.79+/-0.03	0.87+/-0.06	0.80+/-0.02	0.84+/-0.03	0.69+/-0.05	0.17+/-0.01	0.51+/-0.01	0.85+/-0.03	0.90+/-0.02
SVM lineal	K-best 100	0.74+/-0.02	0.76+/-0.04	0.81+/-0.02	0.78+/-0.02	0.73+/-0.04	0.17+/-0.01	0.52+/-0.06	0.83+/-0.03	0.88+/-0.02
SVM rbf		0.77+/-0.02	0.80+/-0.03	0.82+/-0.03	0.81+/-0.01	0.73+/-0.05	0.15+/-0.01	0.47+/-0.04	0.85+/-0.02	0.90+/-0.02
Gradient Boosting		0.78+/-0.03	0.83+/-0.06	0.82+/-0.02	0.82+/-0.03	0.71+/-0.06	0.20+/-0.03	1.29+/-0.23	0.84+/-0.02	0.88+/-0.02
Decision Tree		0.72+/-0.02	0.73+/-0.04	0.80+/-0.04	0.76+/-0.02	0.71+/-0.07	0.22+/-0.02	3.29+/-0.96	0.74+/-0.03	0.84+/-0.03
Random Forest		0.77+/-0.03	0.83+/-0.05	0.80+/-0.03	0.81+/-0.02	0.67+/-0.08	0.17+/-0.01	0.50+/-0.02	0.83+/-0.03	0.88+/-0.02
SVM lineal	K-best 200	0.77+/-0.02	0.78+/-0.04	0.84+/-0.02	0.81+/-0.02	0.77+/-0.04	-0.15+/-0.01	0.50+/-0.06	0.86+/-0.02	0.89+/-0.02
SVM rbf		0.79+/-0.02	0.82+/-0.05	0.84+/-0.03	0.83+/-0.02	0.75+/-0.05	-0.14+/-0.01	0.44+/-0.03	0.87+/-0.02	0.91+/-0.02
Gradient Boosting		0.79+/-0.04	0.84+/-0.07	0.82+/-0.03	0.83+/-0.04	0.71+/-0.08	-0.19+/-0.03	1.10+/-0.18	0.86+/-0.02	0.90+/-0.02
Decision Tree		0.71+/-0.04	0.74+/-0.07	0.77+/-0.04	0.75+/-0.04	0.65+/-0.07	-0.23+/-0.03	3.98+/-0.86	0.74+/-0.04	0.83+/-0.03
Random Forest		0.78+/-0.03	0.85+/-0.06	0.81+/-0.03	0.83+/-0.03	0.69+/-0.07	-0.16+/-0.01	0.49+/-0.02	0.85+/-0.02	0.89+/-0.02
SVM lineal	K-best 300	0.79+/-0.03	0.79+/-0.04	0.85+/-0.03	0.82+/-0.02	0.79+/-0.05	0.15+/-0.01	0.49+/-0.07	0.87+/-0.02	0.90+/-0.02
SVM rbf		0.80+/-0.02	0.82+/-0.04	0.85+/-0.01	0.83+/-0.02	0.78+/-0.03	0.14+/-0.01	0.44+/-0.03	0.87+/-0.02	0.91+/-0.02
Gradient Boosting		0.80+/-0.03	0.84+/-0.05	0.83+/-0.04	0.83+/-0.03	0.73+/-0.08	0.16+/-0.03	0.74+/-0.16	0.87+/-0.03	0.91+/-0.02
Decision Tree		0.69+/-0.03	0.74+/-0.07	0.75+/-0.03	0.74+/-0.04	0.61+/-0.07	0.25+/-0.03	4.23+/-1.33	0.71+/-0.05	0.81+/-0.04
Random Forest		0.78+/-0.03	0.84+/-0.06	0.81+/-0.02	0.82+/-0.03	0.70+/-0.04	0.16+/-0.01	0.49+/-0.02	0.86+/-0.02	0.90+/-0.02
Ensemble SVM	T-test	0.82+/-0.03	0.84+/-0.04	0.86+/-0.03	0.85+/-0.02	0.80+/-0.05	0.13+/-0.01	0.45+/-0.07	0.89+/-0.03	0.92+/-0.03

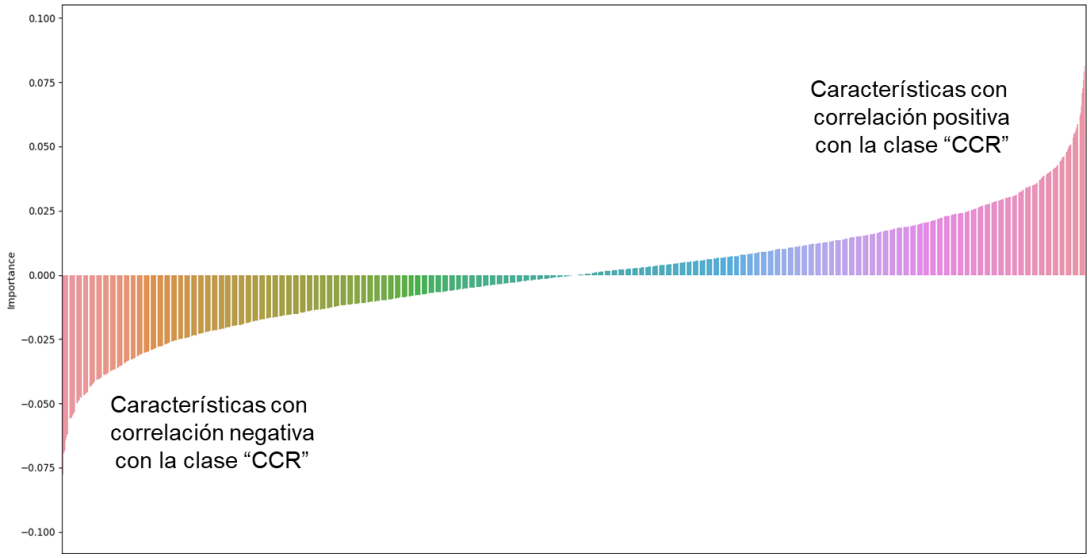
Anexo 5: Rendimiento (validación cruzada) del modelo SVM lineal y el modelo de ensemble con el umbral de decisión por defecto (0.5), aquel que fija la especificidad al 90% (0,70) y el que la fija al 95% (0.82). Se presentan las métricas para todas las fases (I, II, III y IV) de manera conjunta y separada.

Modelo	Umbral de decisión	Fase	Exactitud	Sensibilidad	Precisión	F1	Especificidad	Brier	Log-loss	AUC-ROC	AUC-PR
SVM lineal	0.5	Todas	0.83+/-0.03	0.84+/-0.03	0.88+/-0.03	0.86+/-0.02	**0.82+/-0.04**	0.13+/-0.01	0.44+/-0.07	0.89+/-0.03	0.92+/-0.02
		I	0.8+/-0.03	0.75+/-0.05	0.54+/-0.11	0.62+/-0.08	**0.82+/-0.04**	0.16+/-0.02	0.56+/-0.12	0.87+/-0.03	0.71+/-0.08
		II	0.84+/-0.03	0.87+/-0.04	0.75+/-0.04	0.81+/-0.03	**0.82+/-0.04**	0.14+/-0.02	0.49+/-0.11	0.91+/-0.02	0.85+/-0.05
		III	0.82+/-0.04	0.84+/-0.07	0.68+/-0.05	0.75+/-0.05	**0.82+/-0.04**	0.15+/-0.02	0.54+/-0.11	0.88+/-0.03	0.77+/-0.06
		IV	0.83+/-0.04	0.9+/-0.09	0.46+/-0.13	0.6+/-0.11	**0.82+/-0.04**	0.15+/-0.04	0.56+/-0.17	0.92+/-0.07	0.66+/-0.2
	0.7	Todas	0.77+/-0.02	0.69+/-0.03	0.91+/-0.03	0.79+/-0.02	**0.9+/-0.04**	0.13+/-0.01	0.44+/-0.07	0.89+/-0.03	0.92+/-0.02
		I	0.84+/-0.02	0.63+/-0.06	0.65+/-0.12	0.63+/-0.07	**0.9+/-0.04**	0.16+/-0.02	0.56+/-0.12	0.87+/-0.03	0.71+/-0.08
		II	0.83+/-0.04	0.72+/-0.07	0.82+/-0.06	0.76+/-0.06	**0.9+/-0.04**	0.14+/-0.02	0.49+/-0.11	0.91+/-0.02	0.85+/-0.05
		III	0.82+/-0.04	0.64+/-0.09	0.74+/-0.07	0.68+/-0.07	**0.9+/-0.04**	0.15+/-0.02	0.54+/-0.11	0.88+/-0.03	0.77+/-0.06
		IV	0.89+/-0.03	0.83+/-0.14	0.59+/-0.14	0.67+/-0.1	**0.9+/-0.04**	0.15+/-0.04	0.56+/-0.17	0.92+/-0.07	0.66+/-0.2
	0.82	Todas	0.73+/-0.01	0.58+/-0.04	0.95+/-0.04	0.72+/-0.02	**0.95+/-0.05**	0.13+/-0.01	0.44+/-0.07	0.89+/-0.03	0.92+/-0.02
		I	0.86+/-0.03	0.53+/-0.1	0.79+/-0.17	0.62+/-0.08	**0.95+/-0.05**	0.16+/-0.02	0.56+/-0.12	0.87+/-0.03	0.71+/-0.08
		II	0.83+/-0.03	0.63+/-0.05	0.9+/-0.09	0.73+/-0.04	**0.95+/-0.05**	0.14+/-0.02	0.49+/-0.11	0.91+/-0.02	0.85+/-0.05
		III	0.81+/-0.03	0.51+/-0.08	0.85+/-0.12	0.62+/-0.06	**0.95+/-0.05**	0.15+/-0.02	0.54+/-0.11	0.88+/-0.03	0.77+/-0.06
		IV	0.9+/-0.05	0.64+/-0.17	0.71+/-0.21	0.65+/-0.16	**0.95+/-0.05**	0.15+/-0.04	0.56+/-0.17	0.92+/-0.07	0.66+/-0.2
Ensemble SVM	0.5	Todas	0.82+/-0.03	0.84+/-0.04	0.86+/-0.03	0.85+/-0.02	**0.8+/-0.05**	0.13+/-0.01	0.46+/-0.07	0.89+/-0.03	0.92+/-0.03
		I	0.8+/-0.04	0.8+/-0.04	0.53+/-0.06	0.64+/-0.05	**0.8+/-0.05**	0.16+/-0.02	0.58+/-0.14	0.87+/-0.03	0.71+/-0.08
		II	0.82+/-0.03	0.86+/-0.04	0.72+/-0.06	0.78+/-0.04	**0.8+/-0.05**	0.15+/-0.02	0.52+/-0.12	0.9+/-0.02	0.84+/-0.07
		III	0.8+/-0.03	0.82+/-0.09	0.65+/-0.04	0.72+/-0.04	**0.8+/-0.05**	0.16+/-0.02	0.56+/-0.13	0.89+/-0.03	0.76+/-0.08
		IV	0.81+/-0.05	0.89+/-0.11	0.43+/-0.12	0.57+/-0.12	**0.8+/-0.05**	0.16+/-0.03	0.59+/-0.17	0.91+/-0.05	0.63+/-0.22
	0.7	Todas	0.75+/-0.02	0.64+/-0.02	0.92+/-0.03	0.76+/-0.02	**0.9+/-0.03**	0.13+/-0.01	0.46+/-0.07	0.89+/-0.03	0.92+/-0.03
		I	0.84+/-0.03	0.62+/-0.1	0.67+/-0.1	0.63+/-0.08	**0.9+/-0.03**	0.16+/-0.02	0.58+/-0.14	0.87+/-0.03	0.71+/-0.08
		II	0.81+/-0.02	0.65+/-0.06	0.82+/-0.05	0.72+/-0.04	**0.9+/-0.03**	0.15+/-0.02	0.52+/-0.12	0.9+/-0.02	0.84+/-0.07
		III	0.82+/-0.04	0.62+/-0.08	0.76+/-0.05	0.68+/-0.06	**0.9+/-0.03**	0.16+/-0.02	0.56+/-0.13	0.89+/-0.03	0.76+/-0.08
		IV	0.88+/-0.04	0.71+/-0.17	0.57+/-0.18	0.62+/-0.17	**0.9+/-0.03**	0.16+/-0.03	0.59+/-0.17	0.91+/-0.05	0.63+/-0.22
	0.82	Todas	0.67+/-0.01	0.49+/-0.02	0.95+/-0.04	0.65+/-0.01	**0.95+/-0.04**	0.13+/-0.01	0.45+/-0.07	0.89+/-0.03	0.92+/-0.03
		I	0.84+/-0.04	0.46+/-0.08	0.77+/-0.13	0.56+/-0.07	**0.95+/-0.04**	0.16+/-0.02	0.57+/-0.14	0.87+/-0.03	0.71+/-0.08
		II	0.8+/-0.03	0.55+/-0.08	0.89+/-0.09	0.67+/-0.06	**0.95+/-0.04**	0.15+/-0.02	0.52+/-0.12	0.9+/-0.02	0.84+/-0.07
		III	0.79+/-0.04	0.42+/-0.05	0.83+/-0.13	0.55+/-0.05	**0.95+/-0.04**	0.16+/-0.02	0.56+/-0.12	0.89+/-0.03	0.76+/-0.08
		IV	0.89+/-0.05	0.51+/-0.25	0.63+/-0.24	0.54+/-0.24	**0.95+/-0.04**	0.16+/-0.03	0.58+/-0.17	0.92+/-0.05	0.64+/-0.21

Anexo 6: Matrices de confusión e histogramas de distribución de probabilidades generados durante la validación cruzada del modelo SVM lineal (umbral de decisión 0.5).



Anexo 7: Visión general de la importancia de las características derivadas de los coeficientes obtenidos durante el entrenamiento del modelo SVM lineal. A la derecha, las características con valores positivos muestran una correlación positiva con la clasificación de registros como cáncer colorrectal. A la izquierda, las características con valores negativos se comportante de la manera contraria, mostrando correlación negativa.



Anexo 8: Distribución de las muestras de acuerdo con el *clúster* asignado por K-means. Además, se muestra para las muestras en cada *clúster* la distribución por fase, género y edad.