INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO

MODELOS ECONOMÉTRICOS Y DE APRENDIZAJE ESTADÍSTICO PARA DETECCIÓN DE CÁNCER CON DATOS DE GENÓMICA COMPUTACIONAL

TESIS

QUE PARA OBTENER EL GRADO DE

LICENCIADO EN ECONOMÍA

PRESENTA

PABLO DE JESÚS CAMPOS VIANA

ASESOR: DR. JOSÉ GERMÁN ROJAS ARREDONDO

MÉXICO, D.F. 2017

Con fundamento en los artículos 21 y 27 de la Ley Federal del Derecho de Autor y como titular de los derechos moral y patrimonial de la obra titulada "MODELOS ECONOMÉTRICOS Y DE APRENDIZAJE ESTADÍSTICO PARA DETECCIÓN DE CÁNCER CON DATOS DE GENÓMICA COMPUTACIONAL", otorgo de manera gratuita y permanente al Instituto Tecnológico Autónomo de México y a la Biblioteca Raúl Bailléres Jr., autorización para que fijen la obra en cualquier medio, incluido el electrónico, y la divulguen entre sus usuarios, profesores, estudiantes o terceras personas, sin que pueda percibir por tal divulgación una contraprestación.

Pablo de Jesús Campos Viana

Fecha

Firma

A mis padres.

Agradecimientos

Agradezco a mis padres, Pablo y Clara, y a mi hermano Felipe, por el apoyo incondicional y por creer en mí.

A Andry, por todo lo que hemos vivido juntos y por lo que nos falta por vivir.

A mi asesor Germán Rojas, por la revisión de mi trabajo.

A mis sinodales Zeferino Parada, Léon Berdichevsky e Ignacio Lobato, por sus comentarios para hacer de éste un mejor trabajo.

A mis profesores en el ITAM. Especialmente a Felipe González por su excepcional clase de aprendizaje estadístico y Adolfo de Unánue por su retadora clase de arquitectura de datos.

A mis compañeros y amigos en BAZ, por hacer del trabajo un lugar ameno e intelectualmente motivante; y a mis jefes, Felipe y Mercedes, por darme la oportunidad de trabajar con ellos y mejorar profesionalmente.

Índice general

Índice de figuras

Índice de cuadros

1 Introducción

1.1. Motivación

Las técnicas estadísticas y econométricas convencionales, como la regresión lineal y los modelos lineales generalizados, históricamente han sido suficientes para el proceso de estimación de modelos, pero existen problemas particulares que aparecen al trabajar con conjuntos de datos masivos que pueden requerir diferentes herramientas, tanto teóricas como computacionales.

En primer lugar, el gran volumen de los datos involucrados puede requerir herramientas de manipulación de datos más potentes que las tradicionales. En segundo lugar, es posible tener un alto número de potenciales variables predictoras para la estimación, por lo que es necesario efectuar algún tipo de selección de variables. En tercer lugar, los grandes conjuntos de datos pueden permitir explorar relaciones más flexibles que los modelos lineales simples. En este sentido, las téc-

nicas del aprendizaje estadístico, que han adquirido bastante importancia en las últimos años debido a su alta capacidad de predicción, pueden permitir formas más efectivas de modelar relaciones complejas.

Un ejemplo claro ocurre en el campo de la economía de la salud, y más concretamente, en la economía médica, la cual utiliza datos de la bioestadística, la bioinformática y la epidemiología para apoyar el desarrollo de políticas públicas relacionadas a la salud pública. En particular, un área de investigación muy activa en la comunidad científica y que motiva el desarrollo de esta tesis es la investigación genómica del cáncer.

De esta forma, los objetivos de esta tesis son los siguientes:

- Mostrar que los datos obtenidos a través de la bioinformática y de la genómica computacional pueden ser utilizados con métodos novedosos de aprendizaje estadístico y métodos econométricos tradicionales para obtener modelos con alto poder predictivo y que también sean interpretables; para esto, se utilizan los datos correspondientes a dos tipos de cáncer ginecológicos: el cáncer cervical (o cervicouterino) y el cáncer de endometrio.

- Ilustrar que en contraste con las fuentes de datos tradicionales utilizadas en la econometría, la utilización de estas nuevas fuentes de datos masivos plantean nuevos retos metodológicos y computacionales que pueden ser solventados a través de métodos del aprendizaje estadístico. En este sentido, se comparan dos estrategias para el proceso de construcción de modelos econométricos: un enfoque tradicional y un enfoque auxiliado por métodos del aprendizaje estadístico.

- Proponer el uso de métodos del aprendizaje estadístico en conjunto con métodos econométricos tradicionales, dentro del contexto de un problema mundial con alto impacto económico y de salud pública.

En particular, en este trabajo se exploran métodos de la literatura reciente del aprendizaje estadístico para el proceso de selección de variables y para la estimación del desempeño de los modelos. Para el problema de predicción de presencia

de los tipos de cáncer, se utiliza el método de **Máquinas de Soporte Vectorial** (**MSV**), el cual es un algoritmo de aprendizaje estadístico ampliamente utilizado para problemas de clasificación; y el método de **regresión logística** (**RL**), el cual en estadística y en econometría, es un método clásico de regresión en donde la variable dependiente es categórica y que en contraste con el método MSV, es un método probabilístico.

Los resultados obtenidos proporcionan evidencia a favor de la utilización de los datos obtenidos a través de la bioinformática y de la incorporación de métodos del aprendizaje estadístico en el análisis econométrico, especialmente en el área de la economía de la salud, para contribuir en conjunto con otras disciplinas en el desarrollo de nuevas técnicas de detección temprana del cáncer y así reducir el impacto económico global del mismo.

La estructura de la tesis es como se describe a continuación. El resto del **capítulo 1** motiva el problema de la detección de dos tipos de cáncer ginecológicos, proporciona un panorama del impacto económico del cáncer, y presenta brevemente conceptos importantes del aprendizaje estadístico. El **capítulo 2** presenta brevemente los métodos de clasificación utilizados en este trabajo: MSV y regresión logística. El **capítulo 3** presenta y contrasta resultados de los modelos obtenidos y finalmente el **capítulo 4** presenta las conclusiones del trabajo.

1.2. El cáncer cervical y el cáncer de endometrio

1.2.1. El cáncer cervical

El cáncer cervical (**CESC** - *Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma*) es una enfermedad en la que el cáncer se desarrolla en los tejidos del cuello uterino. Entre los factores de riesgo para el desarrollo del cáncer cervical se encuentran el tabaquismo y la infección por el virus del papiloma humano (VPH). En el futuro, la vacuna contra el VPH reducirá la tasa de infección. Sin embargo, hasta que se vacune a más población, muchas mujeres seguirán siendo infectadas y el cáncer cervical seguirá siendo un problema mundial de salud pú-

blica. A pesar de que la mayoría del cáncer cervicouterino se descubre a través de pruebas de Papanicolaou anuales, el cáncer no diagnosticado puede ser mortal.

Entre los descubrimientos que la comunidad científica ha realizado sobre este tipo de cáncer utilizando datos de genómica computacional, se encuentran:

- Se han identificado mutaciones en ciertos genes que influyen en el desarrollo del cáncer cervical.
- Se han encontrado cambios genómicos en el cáncer cervical que pueden ser tratados con fármacos disponibles en el mercado.
- Se han caracterizado varios tipos y subtipos de cáncer que no están relacionados con la participación del VPH, lo que ha confirmado que un pequeño porcentaje de cánceres cervicales surgen de otras maneras.

1.2.2. El cáncer de endometrio

El cáncer de endometrio (**UCEC** - *Uterine Corpus Endometrial Carcinoma*) se desarrolla en las células que forman el revestimiento interno del útero, o el endometrio, y es uno de los cánceres más comunes del sistema reproductivo femenino entre las mujeres estadounidenses. Alrededor del 70 % de los cánceres de endometrio se diagnostican en una etapa temprana, y como resultado aproximadamente el 83 % de las mujeres sobrevivirán al menos cinco años después del momento del diagnóstico.

Entre los descubrimientos que la comunidad científica ha realizado sobre este tipo de cáncer utilizando datos de genómica computacional, se encuentran:

- Se han identificado cuatro subtipos de cáncer de endometrio.
- Se han descubierto características genómicas compartidas entre el cáncer de endometrio y el cáncer de ovario seroso, el subtipo Basal de cáncer de mama, así como el cáncer colorrectal.

- Se han caracterizado diferencias marcadas entre los dos tipos de tumores endometriales (endometrioides y serosos), y se encontró que algunos tumores endometrioides han desarrollado un patrón sorprendentemente similar a los tumores serosos, lo que sugiere que pueden beneficiarse de tratamientos comunes.

1.3. Impacto económico del cáncer

1.3.1. Impacto global

De acuerdo al reporte *The global economic cost of cancer* ([32]), llevada a cabo por la organización *American Cancer Society* y la fundación *LiveStrong*, el impacto económico total de la muerte prematura y la discapacidad por cáncer en todo el mundo fue de 895 mil millones de dólares en 2008.

Para estimar el impacto económico, se utiliza una medida denominada *años de vida ajustados por discapacidad* (*DALY*, por sus siglas en inglés), la cual a grandes rasgos representa la suma de los años de vida perdidos por un paciente debido a la muerte prematura, así como los años en los que un paciente vivió con la enfermedad; esta medida es posteriormente multiplicada por una estimación del valor económico de un año de vida sana. Esta cifra, que no incluye los costos directos del tratamiento del cáncer, representa el 1.5 % del Producto Interno Bruto (PIB) mundial. Además, el cáncer causa la mayor pérdida económica de las 15 principales causas de muerte en todo el mundo. El costo económico del cáncer es casi un 20 % más alto que las enfermedades del corazón, las cuales representan la segunda causa principal de pérdida económica. La figura 1.3.1 contiene información al respecto.

La muerte y la discapacidad por cáncer de pulmón, cáncer colorrectal y el cáncer de mama representan los cánceres con mayores costos económicos a escala mundial. En los países de bajo ingreso, los cánceres de boca y garganta, cervicouterino y de mama tienen el mayor impacto.

Como es de esperar, el impacto no se distribuye equitativamente entre las naciones. Por ejemplo, mientras que Estados Unidos tiene la mayor pérdida económica

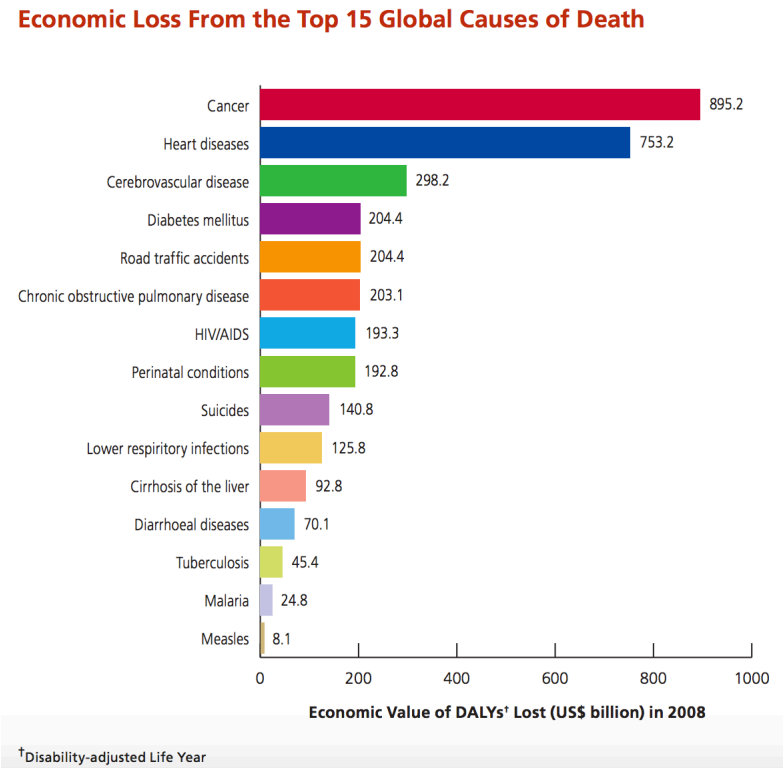

Figura 1.3.1: Pérdida económica estimada de las principales 15 causas de muerte a nivel mundial.

a causa del cáncer, medida en dólares, la enfermedad cuesta al país el 1.73 % de su PIB. Sin embargo, 25 países están perdiendo más del 2 % de su PIB debido a las defunciones y la discapacidad causada por el cáncer. La Organización Mundial de la Salud (OMS) y los expertos mundiales en salud creen que los costos significativos del cáncer podrían ser mitigados por intervenciones focalizadas que han funcionado en países de altos ingresos.

Por otra parte, el impacto económico del cáncer cervicouterino en los países de bajos ingresos es igualmente desproporcionado. Entre las naciones clasificadas por el Banco Mundial como de bajos ingresos, el cáncer cervicouterino representa más del 10 % de la pérdida económica, superada sólo por los cánceres de boca y garganta. La figura 1.3.2 contiene información al respecto.

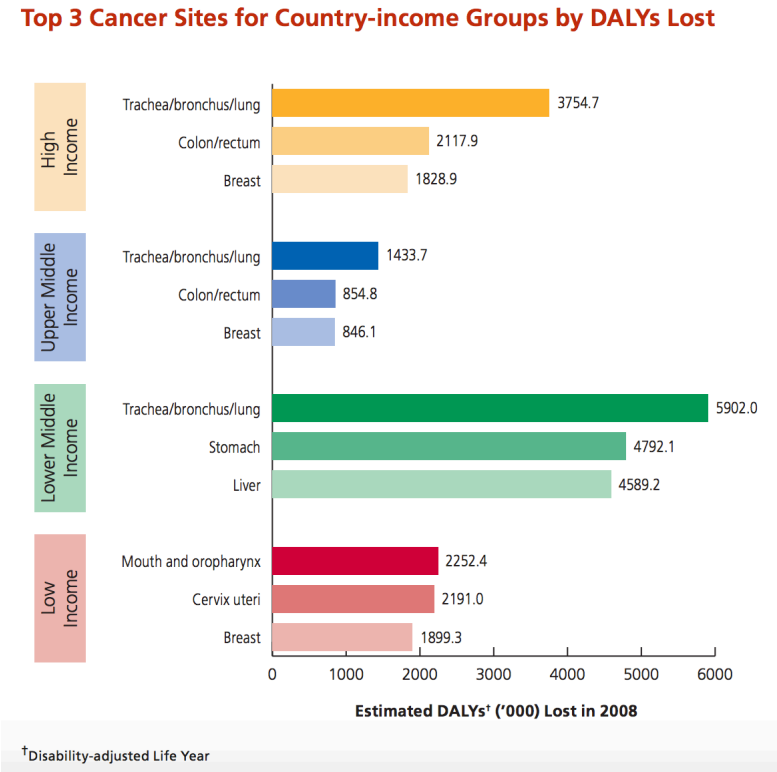

Figura 1.3.2: Principales tipos de cáncer en los diferentes grupos de países por nivel de ingreso.

Desafortunadamente, la mayoría de las mujeres en los países de bajos ingresos no tienen acceso a la atención que pueda prevenir el inicio del cáncer cervicouterino o detectarlo lo suficientemente temprano. Como resultado, muchas mujeres son diagnosticadas demasiado tarde para beneficiarse del tratamiento y salvar su vida. Por el contrario, una gran proporción de mujeres que viven en países de altos ingresos se han beneficiado de las rutinas de detección y modalidades de tratamiento durante más de 50 años, y como resultado, las tasas de cáncer cervical han disminuido drásticamente en esas naciones.

Por otra parte, el cáncer de endometrio es la segunda neoplasia ginecológica más frecuente a nivel mundial. Su mayor frecuencia está condicionada al incremento en la expectativa de vida en la población femenina y al surgimiento de la obesidad

como problema de salud, entre otros factores. La incidencia de este tipo de cáncer es casi seis veces mayor en países desarrollados que en los menos desarrollados, aunque su mortalidad es menor que en aquellos paí- ses con menos desarrollo.

Aunque la detección del cáncer de endometrio en etapas tempranas es favorable con tasas de supervivencia de hasta 80 %, las pacientes con enfermedades de alto riesgo y enfermedad avanzada tienen una supervivencia a largo plazo menor a 50 %. Desafortunadamente, en los últimos 40 años el avance en el tratamiento de esta enfermedad ha sido muy limitado, por un lado debido al escaso número de estudios clínicos realizados y por otro lado, al poco conocimiento de su patología molecular.

1.3.2. El caso de México

De acuerdo a la Unidad de Análisis Económico de la Secretaría de Salud de México ([33]), el impacto económico en el caso de México se puede estimar en dos dimensiones:

- **Sistema Nacional de Salud**: Financiamiento de la atención integral.
- **Nivel macroeconómico**: Impacto en la productividad.

En el caso del nivel macroeconómico, el impacto en la productividad a la vez se puede desglosar en los siguientes ejes, aplicables a la población en edad de trabajar (de 15 a 64 años):

1. **Muerte prematura**.
    - Defunciones por todos los tipos de cáncer.
    - Valor presente del ingreso perdido hasta la edad de jubilación.
2. **Subsidios temporales**.
    - Días de estancia promedio en hospitalización.
    - Días de incapacidad promedio después de la hospitalización.

- Monto de subsidio temporal.

3. **Pensión por invalidez**.
    - Número de supervivientes a 3 años para todos los tipos de cáncer.
    - Valor presente del importe de la pensión por invalidez.
4. **Cuidado no médico**.
    - Período crítico de tratamiento oncológico.
    - Costo de oportunidad para las personas que cuidan a los pacientes.

Mientras que en el caso del financiamiento de la atención integral, el impacto se puede desgosar de la siguiente forma:

1. **Financiamiento público universal de atención médica a:**
    - Cánceres infantiles.
    - Cánceres de la mujer: Mama, cérvix, ovario.
    - Cánceres del hombre: Testículo, próstata.
    - Colorrectal.
2. **Programa Nacional de Detección Oportuna de Cánceres de la Mujer**:
    - Cáncer de mama.
    - Cáncer de cérvix.

El cáncer cervicouterino en México fue la primera causa de muerte por tumores malignos en la mujer hasta el año 2005; a partir de entonces, el cáncer de mama es el que tiene la tasa más elevada. Sin embargo, en 13 estados de la República Mexicana sigue siendo la primera causa de muerte por cáncer en la mujer.

México invierte aproximadamente 6.5 % de su PIB en salud, y destina al IMSS aproximadamente 92 millones de pesos, destinando un gasto de atención hacia el cáncer cervicouterino de 63 mil pesos durante el primer año de tratamiento para

una derechohabiente. Hasta el año 2008, se estimaba que anualmente se diagnostican 68 mil casos de cáncer cervicouterino para lo cual se estima que, considerando el gasto de salud en un año por el número de casos nuevos, el costo es de 4284 millones de pesos en tratamientos (Sánchez, 2012). Por otra parte se estima que el costo promedio del tratamiento de por vida para un derechohabiente es de 335 mil pesos.

Por otra parte, de acuerdo con el Registro Histopatológico de Neoplasias Malignas, en México el cáncer de endometrio ocupa el tercer lugar de los cánceres ginecológicos, después del cáncer cervicouterino y de ovario. En el año 2003 representó 2.16 % del total de los cánceres femeninos y hasta el año 2007 se estima que fue causa de 2.8 % de los egresos hospitalarios por cáncer en todo el país.

1.4. Bioinformática y genómica computacional

La bioinformática es un campo interdisciplinario que desarrolla métodos y herramientas de software para el análisis y la comprensión de datos biológicos. Como un campo interdisciplinario de la ciencia, la bioinformática combina la informática, la estadística, la matemática y la ingeniería para analizar e interpretar los datos biológicos. Además, la bioinformática es también un término más general para la amplia gama de estudios biológicos que utilizan la programación como parte de su metodología, así como una referencia a flujos de análisis específicos que se utilizan repetidamente, particularmente en el campo de la genómica.

La bioinformática se ha convertido en una parte importante de muchas áreas de la biología. En la biología molecular experimental, las técnicas bioinformáticas tales como el procesamiento de imágenes y señales permiten extraer resultados útiles de grandes cantidades de datos crudos. En el campo de la genética y la genómica, ayuda a secuenciar y analizar los genomas y sus mutaciones observadas. Por otra parte, también desempeña un papel crucial en el análisis de la expresión génica.

Más concretamente, la genómica computacional se refiere al uso de análisis computacional y estadístico para descifrar aspectos biológicos a partir de secuencias genómicas, incluyendo secuencias de ADN y ARN, y otros datos derivados a partir de

éstas. De esta forma la genómica computacional puede ser considerada como una subdisciplina de la bioinformática, con enfoque en el uso de los genomas enteros para entender los principios de cómo el ADN de una especie controla su biología a nivel molecular y más allá. En los últimos años ha habido una explosión en la generación y explotación de datos genómicos a nivel mundial; en el caso de México, en el año 2004 fue creado el Instituto Nacional de Medicina Genómica (INMEGEN). Con la abundancia actual de conjuntos de datos biológicos masivos, los estudios computacionales se han convertido en uno de los medios más importantes para el descubrimiento biológico. El área de la genómica computacional incluye tanto aplicaciones de métodos existentes como desarrollo de nuevos algoritmos para el análisis de secuencias genómicas.

En este sentido, una de las aplicaciones de la genómica computacional y que constituye un área de investigación muy activa en la comunidad científica es la investigación genómica del cáncer; en particular, la aplicación de métodos del aprendizaje estadístico para la detección de diferentes tipos de cáncer.

En este trabajo se utilizan datos obtenidos a partir de la genómica computacional, más concretamente, datos que cuantifican la expresión génica. Una descripción más completa sobre la genómica computacional y de los elementos clave para entender los datos utilizados en esta tesis puede ser consultada en el apéndice B.

1.5. Aprendizaje estadístico

El aprendizaje estadístico se refiere al conjunto de métodos computacionales cuyo objetivo es aprender de los datos con el fin de producir reglas para mejorar el desempeño en alguna tarea o toma de decisión.

Las razones usuales para intentar resolver estos problemas computacionalmente son diversas. Por ejemplo:

- Obtener una respuesta barata, rápida, automatizada, y con suficiente precisión.

- Superar el desempeño actual de los expertos o de reglas simples, pero utilizando datos.
- Eentender de manera más completa y sistemática el comportamiento de un fenómeno, identificando variables o patrones importantes.

Las tareas del aprendizaje estadístico se divide en dos grandes gupos: aprendizaje supervisado y aprendizaje no supervisado.

- **Aprendizaje supervisado**: Construir un modelo o algoritmo para predecir o estimar una variable de salida a partir de ciertas variables de entrada. A su vez, los problemas de aprendizaje supervisado se dividen en dos tipos, dependiendo de la variables de salida:
    - **Problemas de regresión**: Cuando la variable de salida es una variable numérica. Por ejemplo, predecir o estimar tasas de mortalidad de cáncer, a nivel país, a partir de variables ambientales o sociodemográficas.
    - **Problemas de clasificación**: Cuando la variable de salida es una variable categórica. Por ejemplo, predecir la presencia de cierto tipo de cáncer, a nivel paciente, a través de variables de historial clínico o bien, de variables obtenidas a partir de la bioinformática.
- **Aprendizaje no supervisado**: En este caso no hay variable de salida. El objetivo es modelar y entender las relaciones entre variables y entre observaciones, o patrones importantes en los datos. Ejemplos de este tipo de aprendizaje son los los métodos de *clustering* y los métodos de reducción de dimensionalidad.

Los problemas supervisados tienen un objetivo claro: hacer las mejores predicciones posibles bajo ciertas restricciones. Los problemas no supervisados tienden a tener objetivos más vagos, y por lo mismo pueden ser más difíciles. Una descripción más formal del aprendizaje supervisado, objeto esencial de estudio en esta tesis, puede ser consultada en el apéndice A.

Por otra parte, en el aprendizaje estadístico se pueden distinguir dos tipos generales de métodos: paramétricos y no paramétricos. Los métodos paramétricos seleccionan un número fijo de parámetros para construir el modelo de predicción, mientras que el número de parámetros de los métodos no paramétricos depende del tamaño del conjunto de datos de entrenamiento. Una introducción detallada al aprendizaje estadístico puede consultarse en [7].

1.6. Econometría y aprendizaje estadístico

Existen diversas similitudes y diferencias entre los métodos econométricos y los métodos de aprendizaje estadístico. Una de las diferencias principales se refiere a la naturalezas paramétrica y no paramétrica de ambas clases de métodos, respectivamente. Por ejemplo, una característica común de muchos métodos del aprendizaje estadístico es que utilizan métodos de selección de variables y métodos de remuestreo para la selección de paramétros, para elegir la complejidad de los modelos; en contraste, en muchos trabajos de econometría aplicada se suelen presentar los resultados de regresiones con varias especificaciones diferentes. El objetivo suele ser mostrar que la estimación de algún parámetro de interés no es muy sensible a la especificación exacta utilizada.

Por otra parte, otra diferencia se refiere a la finalidad de ambas clases de métodos; los métodos de aprendizaje estadístico privilegian el poder predictivo, mientras que los métodos econométricos privilegian la inferencia causal. Es justo en este aspecto en donde mayor potencial de colaboración existe entre ambas disciplinas. En la literatura de la econometría se han desarrollado diversas herramientas para realizar inferencia causal tales como *variables instrumentales*, *regresión discontinua*, *diferencias en diferencias* y varias formas de experimentos naturales y diseñados. Tal como Hal Varian propone en [16], dada la naturaleza creciente en el almacenamiento y el poder de procesamiento de los datos en los últimos años, la econometría puede y debe beneficiarse del uso de los métodos del aprendizaje estadístico, especialmente para trabajar con grandes cantidades de datos. Más concretamente, Varian señala que:

- Tanto en la econometría como en el aprendizaje estadístico suelen preferirse modelos más simples con fines de estimación y predicción. Sin embargo, mientras que en el aprendizaje estadístico se han desarrollado diversas formas para penalizar la complejidad de los modelos, en la econometría no existen formas explícitas para cuantificar la complejidad de los mismos. En particular, el método *Lasso* y sus variantes han ser muy eficientes computacionalmente, logrando buenos resultados en la selección de variables (consultar apéndice C, sección C.2.1).

- Los ciclos iterativos de entrenamiento y prueba, así como las técnicas de remuestreo como la *validación cruzada* (consultar apéndice C, sección C.1.2), son comunmente utilizados en el aprendizaje estadístico para elegir los parámetros de los modelos y para estimar el desempeño de los modelos de forma insesgada y robusta. Durante muchos años, los economistas han reportado las medidas de bondad de ajuste sobre las mismas muestras de entrenamiento con la justificación de que los conjuntos de datos son pequeños. Sin embargo, con la disponibilidad actual de conjuntos de datos más grandes, no existe razón para no realizar la partición de los datos en entrenamiento y prueba como en el aprendizaje estadístico, así como la utilización de las técnicas de remuestreo para la selección de parámetros y estimaciones de desempeño.

Finalmente, Varian concluye que los métodos del aprendizaje estadístico debiesen popularizarse más entre los economistas y utilizarse en la econometría, particularmente al trabajar con conjuntos grandes de datos. Este trabajo va en ese sentido. En esta tesis, se propone el uso de métodos del aprendizaje estadístico, tanto supervisado como no supervisado, en conjunto con métodos econométricos tradicionales para lograr obtener modelos con alto poder de predicción y que permitan estimar el impacto de ciertas variables sobre la presencia de cierta condición, dentro del contexto de un problema mundial con alto impacto económico y de salud pública.

1.7. Trabajo relacionado

1.7.1. Literatura del aprendizaje estadístico

El uso de datos que cuantifican la expresión génica para clasificar el cáncer en sus subtipos, y para discriminar las muestras cancerosas de las normales, es un área de investigación activa y ha sido aplicada a diferentes tipos de cáncer humano, principalmente cáncer de pulmón, cáncer de mama y leucemia. Los trabajos previos en este tipo de literatura han intentado realizar contribuciones en dos sentidos: uno centrada en los métodos de selección de variables correspondientes a los genes; otro en la mejora del desempeño de los algoritmos de clasificación.

Por ejemplo, en [18] varios métodos de selección de variables son utilizados sobre conjuntos de genes que muestran alta correlación, mientras que métodos tales como MSV y *K Nearest Neighbors* (*K-NN*) son utilizados para realizar la clasificación. Por otra parte, en [19]-[25] se proponen ensambles de métodos de selección de variables; sin embargo, debido a a la diversidad de genes, estos ensambles requieren realizar múltiples selecciones de variables en diferentes conjuntos de datos y después agregar los resultados individuales, lo cual resulta muy costoso computacionalmente. En [26], se propone un método de clasificación que utiliza dos enfoques de aprendizaje semi-supervisado para mejorar la calidad de las predicciones. En [27], se utiliza el método *Independent Component Analysis* (*ICA*) para la selección de genes y se utiliza un ensamble de clasificadores de MSV siguiendo un enfoque de comité para realizar predicciones; sin embargo este enfoque también resulta costoso computacionalmente si muchos clasificadores de MSV son construidos para construir el ensamble. Por otra parte, todos estos trabajos han utilizado datos que cuantifican la expresión génica mediante *microarreglos*, la tecnología utilizada hasta antes de los nuevos métodos de secuenciación.

Más recientemente, en [28] se utilizan datos que cuantifican la expresión génica obtenidos de TCGA para la clasificación de los dos suptipos de cáncer de pulmón, construyendo modelos de MSV seleccionando los parámetros con técnicas basadas en el enfoque de comité y que difieren de las tradicionales en la literatura del

aprendizaje estadístico. Por último, [29] es un trabajo aún más reciente en donde también se utilizan los datos de expresión génica para construir clasificadores de todos los tipos de cáncer que forman parte de TCGA; en este trabajo se utilizan algoritmos genéticos junto con KNN para la selección de variables y el método *Extreme Gradient Boosting* ([15]) para realizar la clasificación. En éste último trabajo se reportan resultados ligeramente superiores al 90 % de precisión sobre el conjunto de datos de prueba.

1.7.2. Literatura económica

Un ejemplo relacionado al tema de este trabajo, y que pertenece al ámbito de la economía de la salud, puede consultarse en [17], en donde se estiman los efectos de ciertos factores ambientales y sociodemográficos sobre la prevalencia del cáncer y sus tasas de mortalidad, utilizando datos de 30 países desarrollados de la OCDE hasta el año 2002.

En [17], se afirma que la mayoría de los cánceres resultan de la interacción de los factores genéticos del huésped y la exposición a los peligros ambientales para la salud. El daño del ADN causado por factores ambientales, como el humo del tabaco, puede desencadenar anomalías o mutaciones en los genes, dando como resultado una actividad aumentada y anormal de éstos. Sin embargo, hasta la fecha de la publicación de dicho trabajo (2008), la generación de datos a nivel de genes gracias al desarrollo de la bioinformática apenas se estaba llevando a cabo. Más aún, este tipo de datos no ha sido explotado hasta el momento en la literatura de la economía de la salud.

En particular, la especificación para el modelo del cáncer cervicouterino en dicho trabajo contiene variables relacionadas al tipo de dieta, a niveles de tabaquismo y al índice de desarrollo humano de cada país. El método de regresión *stepwise* hacia atrás es utilizado para elegir el conjunto final de variables. Los resultados de la especificación final utilizando regresión de mínimos cuadrados ordinarios se ilustra en la figura 1.7.1. Las variables explicativas **FRUITVEG**, **HUMDEV** y **SMOKE** se refieren al consumo total de frutas y verduras, al índice de desarrollo

humano y al consumo de tabaco de la población, respectivamente. Estos resultados muestran que el consumo total de frutas y verduras, el índice de desarrollo humano y el consumo de tabaco de la población, son significativos al nivel 1 %. El consumo total de frutas y verduras y el índice de desarrollo humano mantienen una relación negativa con las tasas de incidencia del cáncer cervical, mientras que el consumo de tabaco mantiene una relación positiva.

Dependent Variable : CERVICAL
Observations : 26
Method : Least Squares
R-Squared = 0.869; Adjusted R-Squared = 0.847

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
6	(Constant)	37.680	4.280		8.803	.000
	FRUITVEG	-.009***	.002	-.385	-4.500	.000
	HUMDEV	-39.709***	4.454	-.786	-8.915	.000
	SMOKE	.068**	.026	.231	2.632	.017

*Significant at 10% level; **Significant at 5% level; ***Significant at 1% level.

Figura 1.7.1: Resultados de regresión de mínimos cuadrados ordinarios para el cáncer cervical.

Es relevante señalar algunos aspectos del enfoque seguido en [17] y que motivan el tema de esta tesis.

- Aunque los datos utilizados se encuentran a nivel paciente, el estudio es llevado a cabo a nivel país, lo que limita el análisis.
- El estudio utiliza solamente variables relacionadas a factores ambientales y sociodemográficos, sin incluir datos médicos.
- El tamaño de la muestra es muy pequeño ($n = 26$) por lo que la teoría asintótica no es aplicable. Por otra parte, el pronóstico (o predicción) no es posible.
- El método utilizado para selección de variables sufre de los problemas de sesgo (dado que las pruebas realizadas son llevadas a cabo sobre el mismo

conjunto de datos) y de sobresimplificación de los modelos obtenidos. En este caso particular, se obtiene un modelo con solamente 3 variables.

- Variables como el índice del desarrollo humano pueden tener relaciones ambiguas con las tasas de prevalencia del cáncer, ya que no es claro si una vida más saludable evitará o aumentará el riesgo de desarrollar cáncer. Un país más saludable puede implicar que las personas viven más tiempo; sin embargo, dado que la prevalencia del cáncer aumenta con la edad, una vida más larga podría implicar un mayor riesgo de desarrollar cáncer.

Las aportaciones de esta tesis van en varios sentidos. Por una parte, se explora el problema de predicción de dos tipos de cáncer ginecológicos muy particulares, con datos obtenidos a través de la bioinformática. Por otra parte, se utilizan métodos del aprendizaje estadístico para el proceso de selección de variables y para la estimación del desempeño de los modelos. Además, el método MSV es utilizado con diversas variantes para servir como referencia y punto de comparación para los modelos econométricos obtenidos tanto con un enfoque tradicional como con un enfoque de aprendizaje estadístico.

2

Métodos de clasificación

2.1. Máquinas de soporte vectorial

En esta sección se proporciona una descripción simple e intuitiva del método MSV en el contexto de este trabajo. Una descripción detallada de este método desde el punto de vista matemático y algorítmico puede ser consultada en [8] y [9].

El método MSV es ampliamente utilizado para resolver problemas de clasificación y de regresión. En su formulación más tradicional para problemas de clasificación, dado un conjunto de observaciones en donde cada una pertenece a una de dos clases, el método de MSV construye un modelo que asigna nuevas observaciones a alguna de las dos clases.

Supóngase que se tiene un conjunto de datos de pacientes sanos y pacientes con cáncer, descrito por cantidades relacionadas a dos genes. De esta forma, los pacientes pueden ser representados geométricamente como vectores, tal como ilustra la

figura 2.1.1.

El objetivo del método MSV es encontrar una frontera de decisión lineal que pueda separar ambas clases de pacientes y tenga la máxima distancia posible entre los pacientes que se encuentren más cerca de la misma frontera, tal como se ilustra en la figura 2.1.2.

Sin embargo, si la estructura de los datos no permite encontrar tal frontera de decisión lineal, es posible mapear los datos a través de diversas funciones (*funciones Kernel*) a espacios de dimensión mayor en donde es posible encontrar la frontera de decisión. Esta idea es conocida en la literatura como la **técnica del Kernel** y es ilustrada en la figura 2.1.3.

Un ejemplo es el Kernel radial, definido mediante la expresión:

$$k(x_i, x_j) = \exp(-\sigma \|x_i - x_j\|^2), \quad \sigma > 0 \tag{2.1}$$

Este Kernel forma parte de una amplia familia de funciones llamadas *funciones de base radial*, cuyos valores dependen de una distancia entre dos puntos, i.e. $\varphi(x, c) = \varphi(\|x - c\|)$, y es uno de los más utilizados en aplicaciones reales.

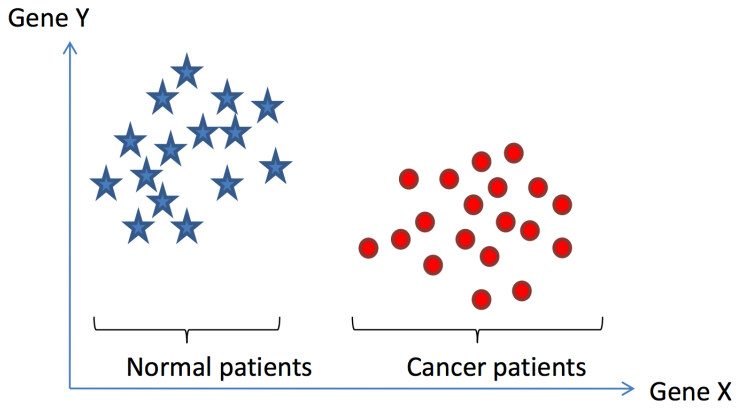

Figura 2.1.1: Representación geométrica del problema de MSV.

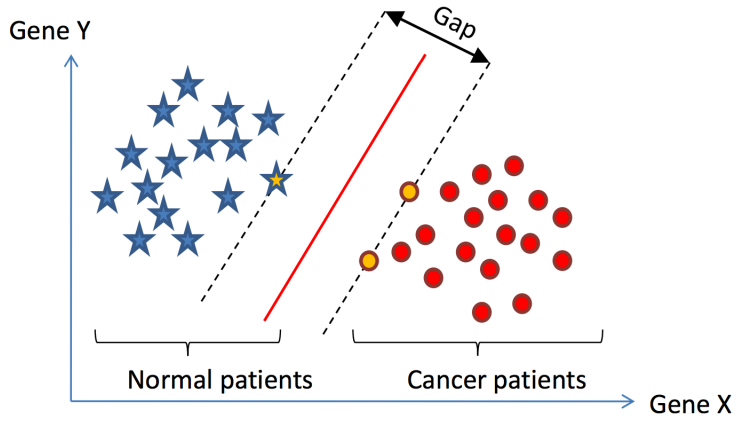

Figura 2.1.2: Ejemplo de solución del método MSV: Kernel lineal

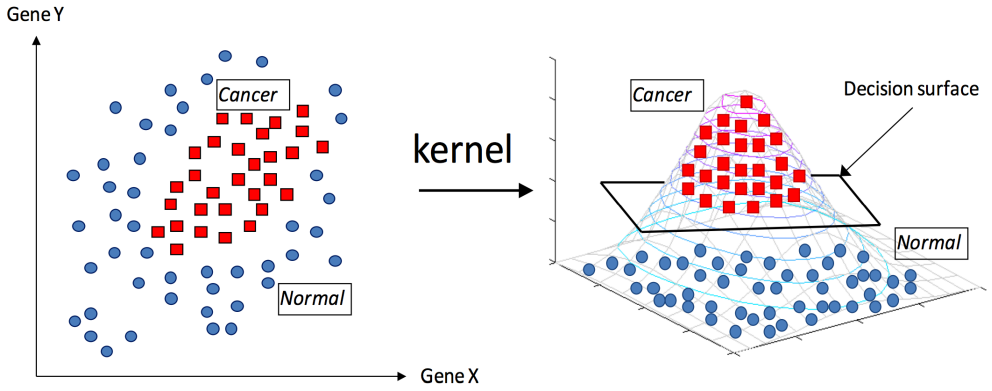

Figura 2.1.3: Ejemplo de solución del método MSV: Kernel no lineal

Más formalmente, dadas dos clases $y \in \{-1, 1\}$, una nueva observación x puede ser clasificada con el método MSV de acuerdo a una función lineal $f(x) = \mathbf{w}^\intercal x + b$ de modo que se tome el signo de esta función para realizar la clasificación. Además, con la técnica del Kernel, se puede realizar una transformación Φ a las nuevas observaciones de modo que la función discriminante sea $f(x) = \mathbf{w}^\intercal \Phi(x) + b$. El plano definido por $f(x) =$ o define una frontera de decisión en el espacio transformado y los parámetros $\mathbf{w}$ y b son determinados a partir de la muestra disponible de los datos.

Dos de las razones principales de la importancia del método MSV y que están relacionadas con este trabajo son:

- Robustez en problemas con bajo número de observaciones y alto número de variables. Esto contrasta con métodos tradicionales como la regresión logística.
- Capacidad de aprender modelos de clasificación tanto simples como complejos.

2.2. Regresión logística

En estadística y en econometría, la regresión logística es un método clásico de regresión en donde la variable dependiente es categórica. El caso más tradicional y el utilizado en este trabajo es cuando la variable dependiente es binaria para distinguir entre presencia o ausencia de cierta condición.

La idea fundamental de la regresión logística es utilizar un mecanismo similar al de la regresión lineal en el sentido de modelar una variable dependiente a través de un predictor lineal, es decir, a través de una combinación lineal de las variables independientes y un conjunto de parámetros asociados a ellas.

En la regresión logística, se utiliza la función logística para modelar la probabilidad de un evento $y_i \in \{0, 1\}$ y en donde $y_i = 1$ y $y_i = 0$ representan usualmente la presencia y ausencia de cierta característica, respesctivamente, mediante la siguiente relación:

$$p_i \equiv \Pr(y_i = 1) = p(\mathbf{x}_i; \mathbf{w}) = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}_i}} \tag{2.2}$$

en donde $\mathbf{x}_i$ es el vector de variables explicativas y $\mathbf{w}$ es el vector de parámetros.

En el contexto de la econometría, es usual formular la regresión logística como un caso particular de los modelos lineales generalizados, los cuales explican y predicen una variable dependiente que puede tener varios tipos de distribuciones de probabilidad, al ajustar un predictor lineal a algún tipo de transformación arbitra-

ria del valor esperado de esa variable. Para el caso de la regresión logística, esta transformación está dada por la función logit:

$$\operatorname{logit}(\mathbb{E}[y_i \mid \mathbf{x}_i]) = \operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{w}^\intercal \mathbf{x}_i \tag{2.3}$$

Los coeficientes de regresión son estimados mediante el método de máxima verosimilitud. Dada la forma funcional del problema, por conveniencia matemática y computacional esta estimación se realiza al maximizar la log-verosimilitud. En el análisis estadístico y econométrico, los coeficientes representan el cambio en el logit por cada unidad de cambio en el predictor lineal. Además, también suele analizarse el efecto de las variables de predicción sobre la función exponencial de los coeficientes de regresión (la razón de momios).

3

Modelos y resultados

En las siguientes secciones, la base de datos se compone de un total de $n = 510$ observaciones con un total de $p = 20501$ variables correspondientes a los genes. De forma seudoaleatoria, el 75 % de las observaciones es asignado al conjunto de entrenamiento y el restante 25 % al conjunto de prueba. En ambos conjuntos, la proporción de casos de cáncer cervicouterino es aproximadamente 60 %, mientras que de cáncer de endometrio es aproximadamente 40 %. El cuadro 3.0.1 muestra la distribución de casos en los conjuntos de entrenamiento y de prueba.

Las métricas de desempeño que se presentan en este capítulo son algunas de las métricas usuales para los modelos de clasificación binaria:

- **Precisión**: Es la proporción de resultados verdaderos (es decir, tanto positivos verdaderos como negativos verdaderos) entre el número total de casos examinados.

Cuadro 3.0.1: Número de casos en los conjuntos de entrenamiento y prueba

Train		**Test**	
383		127	
CESC	**UCEC**	**CESC**	**UCEC**
232	151	77	50

- **Sensibilidad**: Mide la proporción de casos positivos que se identifican correctamente como tales. También es llamada *tasa de verdaderos positivos*.
- **Especificidad**: Mide la proporción de negativos que se identifican correctamente como tales. También es llamada *tasa de verdaderos negativos*.

En el caso de este trabajo, asignamos la clase positiva al cáncer cervicouterino (**CESC**) y la clase negativa al cáncer de endometrio (**UCEC**). La utilización de la sensibilidad y la especificidad como métricas permiten distinguir entre modelos que clasifican mejor a alguno de los dos tipos de cáncer en cuestión. Por otra parte, la precisión como métrica proporciona un panorama general del desempeño de los modelos. Por otra parte, también se presentan *matrices de confusión* en donde cada columna representa las instancias de la clase verdadera y las filas representan las instancias de la clase predicha.

Los experimentos fueron llevados a cabo en una computadora corriendo el sistema operativo macOS Sierra, con memoria RAM de 16GB y un procesador Intel Core i5 con 4 núcleos lógicos. El código utilizado y los resultados pueden ser consultados en `https://github.com/pjcv89/thesis_econ`

Aunque en este trabajo se utilizan estas métricas de desempeño, debe notarse que los resultados de predicción del método MSV pueden llevarse a forma de probabilidades través de un método de calibración propuesto por Platt en [10] y a partir de éstas, calcular métricas tales como el área bajo la curva ROC o el coeficiente de Gini.

3.1. Enfoque tradicional

En econometría, es usual utilizar el método *stepwise* hacia atrás para la selección de variables tanto como para modelos de regresión lineal como para modelos de regresión logística. A pesar de sus debilidades teóricas y computacionales, este método aun es ampliamente utilizado, especialmente en software comercial. Una descripción detallada de este método puede ser consultada en [2].

En la regresión logística, así como en muchos métodos paramétricos, es necesario que se cumpla que el número de variables sea menor al número de observaciones. Bajo un enfoque tradicional, el problema que en este trabajo se trata con $n << p$ es inviable. Sin embargo, es posible seleccionar un conjunto de $\hat{p}$ variables a priori de modo que $n > \hat{p}$; para efectos de realizar este ejercicio, en este trabajo se toma el conjunto de $\hat{p} = 50$ variables con la más alta variabilidad y se filtran aquellas que guardan una correlación absoluta mayor a 0.70 con alguna del resto de variables para realizar el procedimiento *stepwise* hacia atrás utilizando el criterio de información de Akaike (AIC, por sus siglas en inglés). Por otra parte, bajo un enfoque tradicional, tanto la estimación como el cálculo de las métricas de desempeño se realizan sobre el mismo conjunto de datos cuando n no es suficientemente grande; para efectos de realizar este ejercicio, este procedimiento puede simularse sobre el conjunto de datos de entrenamiento y contrastar los resultados obtenidos sobre el conjunto de datos de prueba.

Para la construcción de los modelos de regresión logística, en este trabajo se utilizó la implementación provista por el paquete **glm** en R, mientras que para efectuar la selección de variables mediante *stepwise* hacia atrás, se utilizó la implementación provista por el paquete **MASS**. Después de seleccionar las variables a través del método *stepwise* hacia atrás, se construyó un modelo de regresión logística y se obtuvieron los resultados sobre los conjuntos de datos de entrenamiento y de prueba.

Para realizar la asignación de clases, se utiliza un punto de corte simple sobre la estimación de la probabildiad, establecido en el valor 0.5.

El cuadro 3.1.1 muestra las matrices de confusión para los conjuntos de datos

Cuadro 3.1.1: Resultados RL con *stepwise* hacia atrás

Train			**Test**		
	True			**True**	
Predicted	CESC	UCEC	**Predicted**	CESC	UCEC
CESC	224	3	CESC	72	6
UCEC	8	148	UCEC	5	44

Cuadro 3.1.2: Métricas de desempeño para los modelos de RL con *stepwise* hacia atrás

	Precisión	**Sensibilidad**	**Especificidad**
RL stepwise (Train)	0.9713	0.9655	0.9801
RL stepwise (Test)	0.9134	0.9351	0.8800

de entrenamiento y de prueba. Además, el cuadro 3.1.2 muestra un resumen de las métricas de desempeño (precisión, sensibilidad y especificidad) para el modelo para cada conjunto de datos. Finalmente, los cuadros 3.1.3 muestra un resumen de los parámetros y su significancia para cada uno de los modelos.

De los resultados obtenidos, se obtienen las siguientes observaciones:

- El método *stepwise* hacia atrás para selección de variables es capaz de obtener un modelo con 25 variables, la gran mayoría estadísticamente significativas, que explican la presencia de cada tipo de cáncer en cuestión. Sin embargo, el alto número de variables limita la parsimonia y la interpretabilidad.

- Al comparar entre las métricas calculadas sobre los conjuntos de datos de entrenamiento y prueba, es inmediato observar que el modelo sobreestima todas las métricas.

Cuadro 3.1.3: Modelo RL con variables obtenidas por *stepwise* hacia atrás

	Dependent variable:
	Cancer type
KRT14	24.590
	(17.274)
KRT17	21.032***
	(7.814)
KRT13	62.984
	(45.828)
H19	1.292**
	(0.628)
WFDC2	−4.184**
	(1.987)
GAPDH	−2.149**
	(0.857)
CD74	3.981***
	(1.325)
CD24	−3.638**
	(1.852)
FTH1	−2.806**
	(1.143)
EEF1A1	−3.768***
	(1.027)
HSPB1	−9.952***
	(3.486)
TMSB10	−10.680***
	(3.486)
B2M	−4.748**
	(1.881)
S100A6	6.079***
	(1.928)
KRT15	64.511***
	(24.288)
KRT8	3.457**
	(1.344)
ACTG1	−2.271**
	(0.902)
HP	−11.924*
	(6.119)
VIM	1.558**
	(0.767)
'HLA-A'	1.788
	(1.322)
NDRG1	3.269**
	(1.534)
COL1A2	1.175**
	(0.505)
FBLN1	1.213***
	(0.465)
RPLP1	2.929**
	(1.242)
Intercept	62.999***
	(24.379)
Observations	383
Log Likelihood	-31.866

*p<0.1; **p<0.05; ***p<0.01

3.2. Enfoque de aprendizaje estadístico

3.2.1. Máquinas de soporte vectorial

Para la construcción de los modelos de MSV, en este trabajo se utilizó la implementación provista por el paquete **kernlab** en R. Más detalles sobre esta implementación pueden ser consultados en [4]. Se construyeron los siguientes modelos de MSV, utilizando la técnica RFE para selección de variables y el método t-SNE para reducción de dimensionalidad (consultar apéndice C, secciones C.1.1 y C.1.3, respectivamente) obteniendo sus mejores parámetros a través de *grid search* y *validación cruzada* (consultar apéndice C, sección C.1.2).

- **MSV Kernel lineal**: Modelo utilizando Kernel lineal, con selección de variables a través de RFE.
- **MSV Kernel no lineal**: Modelo utilizando Kernel radial, con selección de variables a través de RFE.
- **MSV Kernel lineal con t-SNE**: Modelo con los datos representados en el espacio de dimensión $d = 2$ obtenido con t-SNE, utilizando Kernel lineal.
- **MSV Kernel no lineal con t-SNE**: Modelo con los datos representados en el espacio de dimensión $d = 2$ obtenido con t-SNE, utilizando Kernel radial.

El cuadro 3.2.1 muestra las matrices de confusión para los mejores modelos de MSV, con Kernel lineal y Kernel no lineal, utilizando las variables elegidas con el método RFE. De forma análoga, el cuadro 3.2.2 muestra las matrices de confusión para los mejores modelos de MSV sobre los datos representados en el espacio de dimensión 2, obtenido después de utilizar el algoritmo t-SNE. Finalmente, el cuadro 3.2.3 muestra un resumen de las métricas de desempeño (precisión, sensibilidad y especificidad) para cada uno de los modelos.

Cuadro 3.2.1: Resultados MSV con variables obtenidas por RFE

Kernel lineal		
	True	
Predicted	CESC	UCEC
CESC	71	2
UCEC	6	48

Kernel no lineal		
	True	
Predicted	CESC	UCEC
CESC	69	0
UCEC	8	50

Cuadro 3.2.2: Resultados MSV con reducción de dimensionalidad

Kernel lineal con t-SNE		
	True	
Predicted	CESC	UCEC
CESC	69	1
UCEC	8	49

Kernel no lineal con t-SNE		
	True	
Predicted	CESC	UCEC
CESC	70	1
UCEC	7	49

Las figuras 3.2.1 y 3.2.2 proporcionan visualizaciones sobre la predicción de los modelos en el conjunto de datos de prueba para el caso del Kernel lineal y Kernel no lineal, respectivamente.

De los resultados obtenidos, se obtienen las siguientes observaciones:

- Estos resultados son consistentes con los obtenidos en [29].
- Los modelos obtenidos para cada variante obtienen resultados muy similares en términos de precisión. Es decir, para el problema particular tratado

Cuadro 3.2.3: Métricas de desempeño para los modelos de MSV

	Precisión	**Sensibilidad**	**Especificidad**
MSV Kernel lineal	0.9370	0.9221	0.9600
MSV Kernel no lineal	0.9370	0.8961	1.0000
MSV Kernel lineal con t-SNE	0.9291	0.8961	0.9800
MSV Kernel no lineal con t-SNE	0.9370	0.9091	0.9800

Figura 3.2.1: Regiones de predicción para el modelo con Kernel lineal en el nuevo espacio de baja dimensión.

en este trabajo, utilizar el método de MSV, tanto en su variante con Kernel lineal y no lineal, con selección de variables mediante el método RFE, es equivalente en términos de precisión a utilizar el método MSV después de realizar reducción de dimensionalidad mediante el método t-SNE.

- De acuerdo a la sensibilidad y especificidad obtenidas para cada variante, el modelo de MSV con Kernel lineal resulta ser superior para detectar la presencia del cáncer cervicouterino, mientras que el modelo de MSV con Kernel no lineal resulta ser superior para detectar la presencia del cáncer de endometrio.

- Sin embargo, debido a la naturaleza del método MSV, en ninguno de los casos es posible dar una interpretación directa en términos de las variables

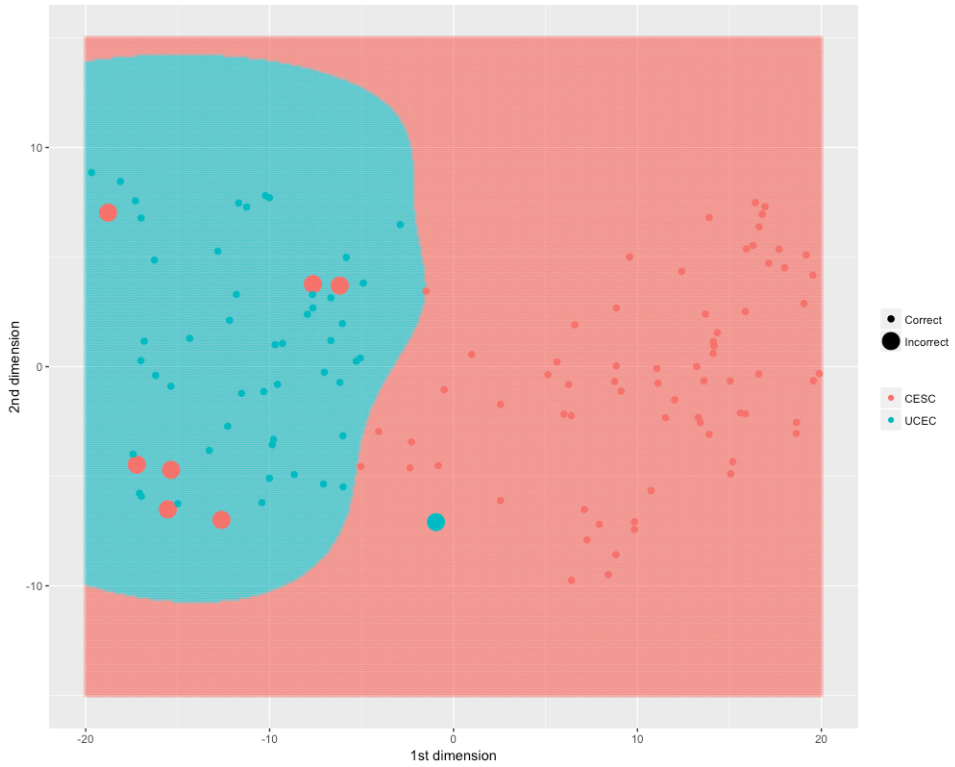

Figura 3.2.2: Regiones de predicción para el modelo con Kernel no lineal en el nuevo espacio de baja dimensión.

utilizadas, es decir, en términos de la expresión de los genes que se incluyen en los modelos.

3.2.2. Regresión logística

Se construyeron los siguientes modelos de regresión logística, utilizando el método Lasso para la selección de variables (consultar apéndice C, sección C.2.1).

- **RL-Lasso completo**: Modelo con las variables obtenidas por Lasso.
- **RL-Lasso reducido**: Modelo con el subconjunto de variables del modelo completo cuyos parámetros son estadísticamente significativos.

Cuadro 3.2.4: Resultados RL con variables obtenidas por Lasso

Modelo completo			**Modelo reducido**		
	True			**True**	
Predicted	CESC	UCEC	**Predicted**	CESC	UCEC
CESC	69	4	CESC	69	5
UCEC	8	46	UCEC	8	45

Cuadro 3.2.5: Métricas de desempeño para los modelos de RL

	Precisión	**Sensibilidad**	**Especificidad**
RL-Lasso completo	0.9055	0.8961	0.9200
RL-Lasso reducido	0.8976	0.8961	0.9000

Para realizar la asignación de clases, se utiliza un punto de corte simple sobre la estimación de la probabildiad, establecido en el valor 0.5.

El cuadro 3.2.4 muestra las matrices de confusión para los modelos de regresión logística: el modelo completo utilizando todas las variables seleccionadas a través de la regresión regularizada y el modelo reducido utilizando el subconjunto de variables del modelo completo cuyos parámetros son estadísticamente significativos, de acuerdo a pruebas de Wald. Además, el cuadro 3.2.5 muestra un resumen de las métricas de desempeño (precisión, sensibilidad y especificidad) para cada uno de los modelos. Finalmente, los cuadros 3.2.6 y 3.2.7 muestran un resumen de los parámetros y su significancia para cada uno de los modelos.

Cuadro 3.2.6: Modelo completo con variables obtenidas por Lasso

	Dependent variable:
	Cancer type
A1BG	-0.402
	(0.294)
ACTR3C	-0.256
	(0.199)
ADORA3	0.157
	(0.185)
CHST11	0.066
	(0.178)
DNAJC3	−0.332*
	(0.201)
FBXO17	-0.219
	(0.165)
GJB4	3.208***
	(0.755)
KPNA1	1.331***
	(0.340)
PAX9	2.813***
	(1.090)
PRIM2	0.516**
	(0.203)
SNTA1	-0.153
	(0.310)
TLE2	0.207
	(0.189)
VPS39	-0.170
	(0.214)
WDR77	−0.638**
	(0.308)
ZNF280C	−0.619***
	(0.214)
ZNF440	-0.282
	(0.263)
ZNF611	-0.183
	(0.219)
ZSCAN1	−0.776**
	(0.367)
Intercept	2.078***
	(0.447)
Observations	383
Log Likelihood	-95.900

*p<0.1; **p<0.05; ***p<0.01

Cuadro 3.2.7: Modelo reducido con variables obtenidas por Lasso

	Dependent variable:
	Cancer type
GJB4	3.270^{***}
	(0.673)
KPNA1	1.216^{***}
	(0.293)
PAX9	2.872^{***}
	(1.094)
PRIM2	0.539^{***}
	(0.172)
WDR77	-0.575^{**}
	(0.268)
ZNF280C	-0.562^{***}
	(0.190)
ZSCAN1	-0.938^{**}
	(0.397)
Intercept	2.050^{***}
	(0.417)
Observations	383
Log Likelihood	-102.685

*p<0.1; **p<0.05; ***p<0.01

De los resultados obtenidos, se obtienen las siguientes observaciones:

- El resumen del modelo completo con variables obtenidas por Lasso en el cuadro 3.2.6 permite distinguir entre aquellas variables que son significativas al nivel 1 %, 5 % o 10 % y aquellas que no lo son. El resumen del modelo reducido con variables obtenidas por Lasso en el cuadro 3.2.7 permite observar que todas las variables son significativas al nivel 1 %.

- La estrategia de utilizar el modelo regularizado para seleccionar las variables a utilizar en el modelo de regresión logística y después seleccionar aquellas que son estadísticamente significativas, permite obtener un modelo parsimonioso con 7 variables, sacrificando muy poca precisión respecto del modelo con todas las variables seleccionadas.

- En comparación con el enfoque tradicional de utilizar la regresión logística con *stepwise* hacia atrás para selección de variables, la estrategia utilizada es capaz de obtener un modelo más parsimonioso. Por otra parte, la estrategia utilizada permite obtener resultados de predicción en el conjunto de datos de prueba que son consistentes con las estimaciones por remuestreo mediante validación cruzada (consultar apéndice C, sección C.2.1). En contraste, la regresión logística con *stepwise* hacia atrás sobreestima los resultados de predicción.

- Los resultados de predicción obtenidos son ligeramente inferiores a cualquiera de las variantes de los modelos de MSV, en términos de precisión, sensibilidad y especificidad.

Cuadro 3.3.1: Resumen de métricas de desempeño y número de variables para todos los modelos obtenidos

	Precisión	Sensibilidad	Especificidad	Variables
RL stepwise (Train)	0.9713	0.9655	0.9801	24
RL stepwise (Test)	0.9134	0.9351	0.8800	24
MSV Kernel lineal	0.9370	0.9221	0.9600	50
MSV Kernel no lineal	0.9370	0.8961	1.0000	30
MSV Kernel lineal con t-SNE	0.9291	0.8961	0.9800	2
MSV Kernel no lineal con t-SNE	0.9370	0.9091	0.9800	2
RL-Lasso completo	0.9055	0.8961	0.9200	18
RL-Lasso reducido	0.8976	0.8961	0.9000	7

- Por otra parte, en contraste con los modelos de MSV, es posible dar interpretación a las variables incluidas en los modelos en términos de sus coeficientes. Además, de acuerdo al signo de sus coeficientes, se concluye que la expresión génica de los genes **GJB4**, **KPNA1**, **PAX9** y **PRIM2** está asociada positivamente al logit, mientras que la expresión génica de los genes **WDR77**, **ZNF280C** y **ZSCAN1** está asociada negativamente al logit. Es interesante notar que en la literatura, las expresiones de los genes PAX9, PRIM2 y WDR77 han sido halladas en diversas manifestaciones del cáncer, incluyendo el cáncer de pulmón y de próstata ([42]).

3.3. Resumen de resultados

El cuadro 3.3.1 muestra un resumen de todos los modelos discutidos en las secciones anteriores. En síntesis, los resultados obtenidos demuestran que el modelo de regresión logística obtenido siguiendo un enfoque tradicional (**RL stepwise**) utiliza un alto número de variables y sobreestima los resultados de predicción; mientras que los modelos de regresión logística obtenidos siguiendo un enfoque de aprendizaje estadístico (**RL-Lasso**) son más simples en términos del número de variables utilizadas y obtienen resultados de predicción ligeramente inferiores y competitivos con los obtenidos por los modelos de MSV.

4

Conclusiones y trabajo futuro

A continuación se presentan algunas conclusiones que se desprenden de este trabajo.

- Los datos obtenidos a través de la bioinformática y de la genómica computacional pueden ser utilizados con métodos novedosos de aprendizaje estadístico y métodos econométricos tradicionales para obtener modelos con alto poder predictivo e interpretables desde el punto de vista probabilístico. En particular, en este trabajo se han explorado métodos de la literatura reciente del aprendizaje estadístico, que se han traducido en modelos con alto poder predictivo y consistentes con la literatura relacionada; por otra parte, la especificación de modelos de regresión logística ha provisto de interpretabilidad en términos del impacto de la expresión génica de ciertos genes sobre la probabilidad de desarrollar los tipos de cáncer que se han tratado en este trabajo.

- Los resultados obtenidos en este trabajo ponen de manifiesto las debilidades existentes en el enfoque tradicional de la econometría; sin embargo, los resultados presentados también proporcionan evidencia de que un enfoque auxiliado por el aprendizaje estadístico puede traducirse en mejores modelos en términos de interpretabilidad y de pronóstico. Más concretamente, en este trabajo se demuestra que utilizar un método econométrico auxiliado de métodos de selección de variables y de estimación de desempeño es mejor elección.

- Si bien los modelos obtenidos mediante el método MSV resultan ser mejores en términos de predicción, la interpretabilidad en términos de las variables utilizadas es limitada o inexistente. En contraste, los modelos obtenidos mediante el método de regresión logística son totalmente interpretables y con resultados de predicción cercanos a los obtenidos por el método MSV.

- En línea con lo que Hal Varian enfatiza [16], los resultados presentados en esta tesis ejemplifican cómo la econometría puede beneficiarse del uso de los métodos del aprendizaje estadístico, especialmente para trabajar con grandes cantidades de datos.

- Las líneas futuras de investigación van en varias direcciones. Por una parte, la metodología propuesta en este trabajo puede ser extendida a un mayor número de variantes del cáncer a través de métodos multiclase. Por otra parte, un mayor número de métodos del aprendizaje estadístico pueden ser incorporados al análisis para obtener mejores modelos predictivos. Finalmente, una amplia gama de métodos econométricos pueden ser incorporados para el análisis de inferencia causal y para la medición del impacto de nuevos tratamientos para las diversas variantes del cáncer. En todos los casos, es posible traducir los hallazgos en nuevas técnicas de detección temprana del cáncer que reduzcan el impacto económico global de la enfermedad.

Apéndice

A

Aprendizaje supervisado y regularización

A.1. Aprendizaje supervisado

Considérese un contexto de **aprendizaje supervisado**. Cada observación $\mathbf{z}$ representa una pareja $(\mathbf{x}, y)$ compuesta de un vector de características $\mathbf{x} \in \mathcal{X}$ y un escalar $y \in \mathcal{Y}$, en donde $\mathcal{X}$ es el espacio de características y $\mathcal{Y}$ el espacio de posibles respuestas.

La teoría del aprendizaje estadístico toma como perspectiva asumir que existe una distribución de probabilidad desconocida sobre el espacio $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, es decir, que existe una distribución de probabilidad $P(\mathbf{z}) = P(\mathbf{x}, y)$. El conjunto de entrenamiento $\mathcal{S}$ se compone de n observaciones de esta distribución:

$$\mathcal{S} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \tag{A.1}$$

El problema de aprendizaje consiste en encontrar una función $f \in \mathcal{F}, f : \mathcal{X} \to \mathcal{Y}$ tal que $f(\mathbf{x}) \approx y$. El espacio de funciones $\mathcal{F}$ es llamado el espacio de hipótesis. Por otra parte, considérese el *funcional de pérdida* $\ell(f(\mathbf{x}), y)$ que define una métrica para el costo de predecir $f(\mathbf{x})$ cuando la respuesta verdadera es y.

Con la finalidad de encontrar la mejor función de aproximación, se define el *riesgo esperado*:

$$E(f) = \int_{\mathcal{Z}} \ell(f(\mathbf{x}), y) dP(\mathbf{z}) \tag{A.2}$$

y dado que la distribución de probabilidad $P(\mathbf{z})$ es desconocida, debe usarse una medida de aproximación al riesgo esperado. Esta medida está basada en el conjunto de entrenamiento $\mathcal{S}$, y se define como el *riesgo empírico*:

$$E_{\mathcal{S}}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i) \tag{A.3}$$

El riesgo empírico $E_{\mathcal{S}}(f)$ mide el desemepeño en el conjunto de entrenamiento, mientras que el riesgo esperado $E(f)$ mide el desempeño de generalización, es decir, el desempeño esperado sobre futuras observaciones.

Teóricamente, la mejor función de aproximación para el caso del riesgo esperado está dada por:

$$f^* = \inf_{f \in \mathcal{F}} \quad E(f) \tag{A.4}$$

mientras que para el caso del riesgo empírico:

$$f_S^* = \inf_{f \in \mathcal{F}} \quad E_{\mathcal{S}}(f) \tag{A.5}$$

El enfoque de elegir una función $f_{\mathcal{S}}^*$ que minimice el riesgo empírico es llamado *minimización del riesgo empírico*.

A.2. Regularización

Supóngase que la familia $\mathcal{F}$ de funciones está parametrizada por un vector de parámetros $\mathbf{w}$. El término *regularización* se refiere a modificar el problema de minimizar el riesgo empírico $E_S(f)$ de modo que el nuevo problema sea minimizar la función:

$$Q(\mathbf{w}) = E_S(f_\mathbf{w}) + \lambda R(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i; \mathbf{w}) + \lambda R(\mathbf{w}) \tag{A.6}$$

en donde λ es un valor fijo tal que $\lambda \geq 0$ y en donde $R(\mathbf{w})$ es una función que penaliza al vector de parámetros $\mathbf{w}$. Este enfoque es conocido como *minimización del riesgo empírico regularizado*. Elecciones típicas del término de regularización son las normas ℓ_1 y ℓ_2 del vector $\mathbf{w}$; y en donde esta última está íntimamente relacionada con el método MSV. En la práctica, la regularización se utiliza para evitar el *sobreajuste*.

B

Genómica computacional y fuentes de datos

B.1. La genética del cáncer

El cáncer es una enfermedad genética, es decir, el cáncer es causado por ciertos cambios en los genes que controlan el funcionamiento de nuestras células, especialmente cómo crecen y se dividen.

Los genes llevan las instrucciones para hacer las proteínas, que hacen mucho del trabajo en nuestras células. Ciertos cambios en los genes pueden hacer que las células evadan los controles normales de crecimiento y se conviertan en cáncer. Por ejemplo, algunos cambios genéticos causantes de cáncer aumentan la producción de una proteína que hace crecer las células. Otros resultan en la producción de una forma anormal, y por lo tanto no funcional, de una proteína que normalmente

repara el daño celular.

Los cambios genéticos que promueven el cáncer pueden ser heredados de nuestros padres si los cambios están presentes en las células germinales, que son las células reproductivas del cuerpo. Tales cambios, llamados cambios de la línea germinal, se encuentran en cada célula de la descendencia.

Los cambios genéticos causantes de cáncer también pueden ser adquiridos durante la vida de una persona, como resultado de errores que ocurren cuando las células se dividen o de la exposición a sustancias, tales como ciertos químicos en el humo del tabaco, y la radiación, como los rayos ultravioleta del sol, los cuales pueden dañar el ADN. Los cambios genéticos que ocurren después de la concepción se llaman cambios somáticos.

En general, las células cancerosas sufren más cambios genéticos que las células normales. Pero el cáncer de cada persona tiene una combinación única de alteraciones genéticas. Algunos de estos cambios pueden ser el resultado del cáncer, en lugar de la causa. A medida que el cáncer continúa creciendo, se producirán cambios adicionales. Incluso dentro del mismo tumor, las células cancerosas pueden tener diferentes cambios genéticos.

Los investigadores trabajando en este tipo de problemas suelen identificar las alteraciones causantes del cáncer comparando la secuencia de todo el ADN en una célula cancerosa con la de las células normales e identificando las diferencias. Este tipo de investigación se llama investigación genómica del cáncer.

B.2. Investigación genómica del cáncer

El estudio de los genomas del cáncer ha revelado anormalidades en los genes que impulsan el desarrollo y el crecimiento de muchos tipos de cáncer. Este conocimiento ha mejorado nuestra comprensión de la biología del cáncer y conducido a nuevos métodos de diagnóstico y tratamiento de la enfermedad.

Por ejemplo, el descubrimiento de los cambios genéticos y causantes de cáncer en los tumores ha permitido el desarrollo de terapias dirigidas a estos cambios, así como las pruebas diagnósticas que identifican a los pacientes que pueden benefi-

ciarse de estas terapias.

Durante la última década, los proyectos de investigación a gran escala han comenzado a examinar y catalogar los cambios genómicos asociados con una serie de tipos de cáncer. Estos esfuerzos han revelado similitudes genéticas inesperadas a través de diferentes tipos de tumores.

Los investigadores en este campo también han demostrado que un determinado tipo de cáncer, como el cáncer mama, de pulmón, o de estómago, puede tener varios subtipos moleculares. Para algunos tipos de cáncer, la existencia de ciertos subtipos no se había conocido hasta que los investigadores comenzaron a perfilar los genomas de las células tumorales.

Los resultados de la investigación en este campo ilustran el diverso panorama de las alteraciones genéticas en el cáncer y proporcionan una base para entender las bases moleculares de este grupo de enfermedades.

B.3. Expresión génica

Una molécula de ADN se divide en unidades funcionales llamadas genes. Cada gen proporciona instrucciones para un producto funcional, es decir, una molécula necesaria para realizar un trabajo en la célula. En muchos casos, el producto funcional de un gen es una proteína. Un ejemplo ocurre en el clásico experimento de flores de Mendel que se ilustra en la figura B.3.1, en donde el gen del color de la flor provee las instrucciones para una proteína que ayuda a crear la pigmentación los pétalos.

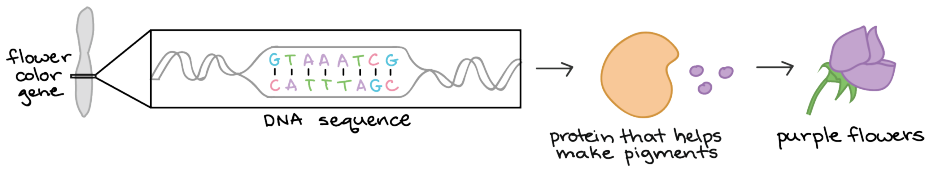

Figura B.3.1: Experimento de flores de Mendel.

Los productos funcionales de la mayoría de los genes conocidos son proteínas,

o, más exactamente, polipéptidos, los cuales se refieren a una cadena de aminoácidos. Aunque muchas proteínas consisten en un único polipéptido, algunas están formadas por múltiples polipéptidos. Los genes que especifican polipéptidos se denominan genes que codifican proteínas.

Muchos genes proporcionan instrucciones para la construcción de polipéptidos. El proceso por el cual el ADN dirige la construcción de un polipéptido implica dos fases principales: transcripción y traducción.

- En la **transcripción**, la secuencia de ADN de un gen se copia para formar una molécula de ARN. Este paso se llama transcripción porque implica reescribir, o transcribir, la secuencia de ADN en un *alfabeto* similar de ARN. En las células eucariotas, es decir, aquellas de los seres vivos, la molécula de ARN debe someterse a procesamiento para convertirse en un ARN mensajero maduro (ARNm).

- En la **traducción**, la secuencia del ARNm se decodifica para especificar la secuencia de aminoácidos de un polipéptido. El nombre de esta fase refleja que la secuencia de nucleótidos de la secuencia del ARNm debe traducirse en el *lenguaje* completamente diferente de los aminoácidos.

Por lo tanto, durante la expresión de un gen codificador de proteínas, la información fluye de ADN $\rightarrow$ ARN $\rightarrow$ proteína. Este flujo direccional de información se conoce como el *dogma central de la biología molecular* y es ilustrado en la figura B.3.2. La esencia del dogma central es que la secuencia de nucleótidos en el ADN (adenina$\rightarrow$A, citosina$\rightarrow$C, guanina$\rightarrow$G, timina$\rightarrow$T) se copia y se transcribe a las secuencias de nucleótidos en ARN (adenina$\rightarrow$A, citosina$\rightarrow$C, guanina$\rightarrow$G, uracilo$\rightarrow$U). Los genes que no codifican proteínas (genes que especifican ARNs funcionales) todavía se transcriben para producir un ARN, pero este ARN no se traduce en un polipéptido. Para cualquier tipo de gen, el proceso de pasar del ADN a un producto funcional se conoce como expresión génica.

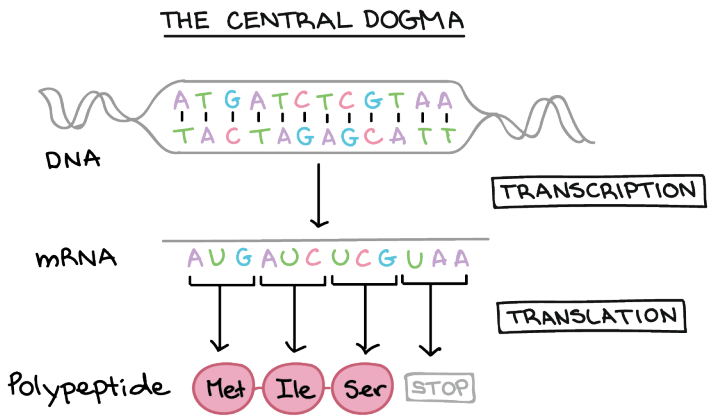

Figura B.3.2: Dogma central de la biología molecular.

B.3.1. Medición

Medir la expresión génica es una parte importante de muchas ciencias, ya que la capacidad de cuantificar el nivel en el que un gen particular se expresa dentro de una célula, tejido u organismo puede proporcionar mucha información valiosa. Por ejemplo, la medición de la expresión génica, entre otras cosas, puede:

- Identificar infección viral de una célula.
- Encontrar si una bacteria es resistente a la penicilina.
- Determinar la susceptibilidad de un individuo a cierto tipo de cáncer (*expresión oncogénica*).

Esta última aplicación de la medición de la expresión génica es la que es de interés en este trabajo.

B.4. Secuenciacíon del ARN

En la genética, la secuenciación del ADN es el proceso de determinar el orden preciso de los nucleótidos dentro de una molécula de ADN. Es decir, hace referencia a cualquier método o tecnología que se utilice para determinar el orden de

las cuatro bases (adenina, citosina, guanina y timina) en una hebra de ADN. El desarrollo de métodos rápidos de secuenciación de ADN ha acelerado considerablemente la investigación biológica y médica.

La secuenciación del ARN (**RNA-Seq**) es una técnica que utiliza métodos de alto rendimiento de secuenciación para revelar la presencia y cuantificar la cantidad de ARN en muestras biológicas en un determinado instante del tiempo. Esta técnica se utiliza para analizar el transcriptoma celular, es decir, el conjunto de todas las moléculas de ARN en una célula o población de células. En particular, la aplicación de esta técnica que es de interés en este trabajo es el análisis de diferencias en la expresión génica en diferentes grupos de muestras.

Dentro de los seres vivos, los genes son transcritos y empalmados para producir transcripciones maduras de ARNm. El ARNm se extrae del organismo, se fragmenta y se copia en ADN complementario (ADNc), el cual se refiere al ADN sintetizado a partir de una plantilla de ARNm en una reacción catalizada por la enzima llamada *transcriptasa inversa*. El ADNc es secuenciado usando métodos de secuenciación de lectura rápida de alto rendimiento. Estas secuencias pueden entonces alinearse con una secuencia de genoma de referencia para reconstruir qué regiones del genoma se estaban transcribiendo. Estos datos pueden usarse para analizar en dónde están expresados los genes o sus niveles relativos de expresión. Este proceso es ilustrado en la figura B.4.1.

Con esta técnica, la expresión génica se cuantifica para estudiar, entre otras cosas, los cambios celulares en respuesta a estímulos externos y las diferencias entre estados de presencia y ausencia de enfermedades, particularmente el cáncer.

De esta forma, la expresión génica se cuantifica contando el número de lecturas que se mapean a cada locus (una posición específica en los cromosomas) en la etapa de ensamble del transcriptoma, el cual es el proceso por el cual se asignan las lecturas crudas a características genómicas.

La expresión génica de diferentes genes puede identificarse utilizando herramientas que cuentan las lecturas de secuenciación por cada gen y las comparan entre muestras. Existen paquetes computacionales disponibles para este tipo de

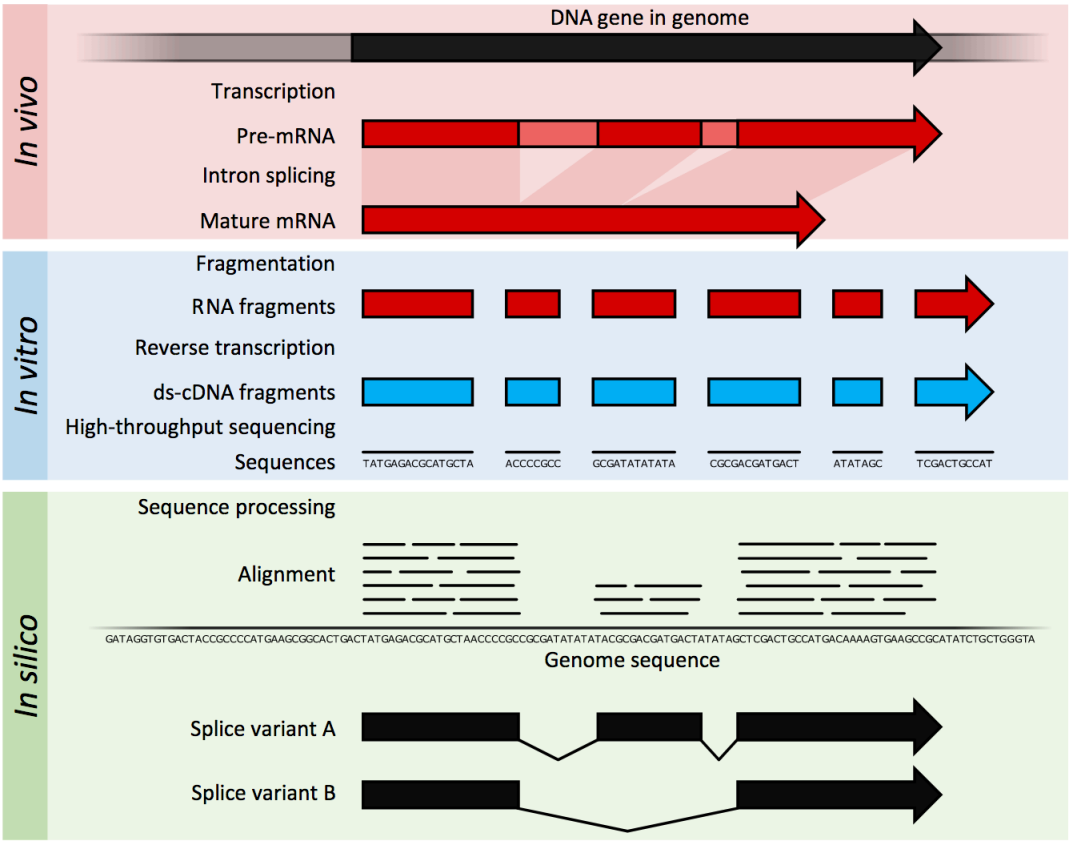

Figura B.4.1: Secuenciación del ARN.

análisis; algunas de las herramientas más utilizadas son los paquetes **DESeq** y **edgeR** del proyecto de código abierto **Bioconductor**, escrito principalmente en el lenguaje R.

B.5. *The Cancer Genoma Atlas*

Existen al menos 200 tipos de cáncer y muchos subtipos de ellos. Cada uno de éstos es causado por errores en el ADN que causan que las células crezcan de forma anormal. Identificar los cambios en el conjunto completo de ADN de cada cáncer - su genoma - y comprender cómo interactúan tales cambios para producir la enfermedad es fundamentar para sentar las bases para mejorar la prevención del cáncer,

la detección temprana y el tratamiento.

El Atlas del Genoma del Cáncer (TCGA por sus siglas en inglés), es una colaboración entre el Instituto Nacional del Cáncer (NCI) y el Instituto Nacional de Investigación del Genoma Humano (NHGRI) iniciado en 2005 en los Estados Unidos y que ha generado amplios mapas multidimensionales de los principales cambios genómicos en 33 tipos de cáncer, haciendo uso de la secuenciación genómica y métodos computacionales de la bioinformática. El conjunto de datos de TCGA comprende más de 2.5 petabytes de datos que describen los tejidos tumorales y los tejidos normales de aproximadamente 14, 000 pacientes y más de 20, 000 genes. Estos datos están disponibles al públigo en general y ha sido ampliamente utilizado por la comunidad científica.

El proyecto TCGA ha creado un flujo de análisis de datos genómicos a gran escala que recopila, selecciona y analiza tejidos humanos para encontrar alteraciones genómicas. El éxito de este proyecto colaborativo entre equipos de investigación y de tecnología sirve de modelo para futuros proyectos científicos alrededor del mundo.

Aunque el proyecto TCGA está llegando a su fin en 2017, las nuevas iniciativas de genómica del NCI, las cuales se ejecutarán a través su Centro de Genómica del Cáncer (CCG), seguirán basándose en el éxito de TCGA utilizando el mismo modelo de colaboración para el análisis genómico a gran escala y haciendo que los datos genómicos estén públicamente disponibles.

B.5.1. Tipos de cáncer seleccionados en TCGA

El proyecto TCGA ha elegido los tipos de cáncer a analizar basado en ciertos criterios específicos, entre los que se encuentran:

- Mal pronóstico y alto impacto en la salud pública.
- Disponibilidad de tumores humanos y muestras de tejidos normales que cumplen los estándares de TCGA para la calidad y cantidad, así como el consentimiento del paciente.

Cuadro B.5.1: Cánceres ginecológicos

Tipo	Casos	Última actualización
Cáncer de ovario	586	31/06/16
Sarcoma uterino	57	29/04/16
Cáncer cervical	307	26/05/16
Cáncer de endometrio	548	02/06/16

La lista completa de los tipos de cáncer incluidos en el proyecto TCGA pueden consultarse en [39]. En este trabajo, son de interés aquellos clasificados como ginecológicos. La tabla B.5.1 muestra un resumen del número de casos y la última fecha de actualización de datos para cada uno.

De éstos, los últimos dos fueron seleccionados para este trabajo debido a que cuentan con un número de casos suficiente para construir modelos de clasificación sin problemas de clases desbalanceadas, por su alto impacto en la salud pública mundial y por formar parte de un tipo de cáncer más general: el cáncer uterino.

B.6. Descripción de los datos

B.6.1. Nivel de los datos

El nivel de los datos es un método de categorización de datos utilizado dentro de la red de TCGA para facilitar a los investigadores la comunicación y localización de sus datos de interés.

Existen cuatro niveles de datos: **nivel 1** (datos crudos), **nivel 2** (datos procesados), **nivel 3** (datos segmentados o interpretados) y **nivel 4** (datos de ciertas regiones genómicas). Un caso particular de datos de nivel 3 son las señales de expresión génica de un conjunto de genes en varias muestras.

Existen dos flujos de análisis utilizadas para crear datos de expresión de nivel 3 a partir de datos de secuencias de ARN. El primer enfoque utilizado en TCGA se basa en el método *RPKM* (*Reads Per Kilobase Million Mappen Reads*), el cual

cuantifica la expresión génica al normalizar la longitud total de la lectura y el número de lecturas de secuenciación. El segundo enfoque utiliza una técnica llamada *MapSplice* para hacer la alineación y una técnica denominada *RSEM* (*RNA-Seq by Expectation Maximization*) para realizar la cuantificación; más detalles sobre estos métodos pueden ser consultados en [30] y [31], respectivamente. Éste último método es denominado **RNA-Seq Version 2** y es de gran importancia ya que en la práctica ha demostrado mejores resultados que su predecesor al incorporar incertidumbre en los procesos de lecturas a través de métodos estadísticos bayesianos.

B.6.2. TCGA2STAT

Los datos de TCGA de cualquier nivel pueden descargarse de portales web o de servicios web, como el portal de datos de TCGA, *cBio*, *canEvolve*, o *Broad Institute GDAC Firehose*. Sin embargo, la descarga e integración manual de estos datos masivos requiere mucho tiempo y algunas llamadas de servicio web requieren la instalación de programas adicionales.

TCGA2STAT es un paquete desarrollado para el lenguaje R y utiliza llamadas de servicio web mediante el protocolo HTTP al sitio del Broad Institute y realiza todo el flujo necesario para descargar e integrar los datos. Esto se logra a través de funciones implementadas a través del paquete **XML** en R.

En este trabajo se utilizan datos de nivel 3 utilizando el método **RNA-Seq Version 2** para la cuantificación de los niveles de expresión génica. Los datos descargados para cada tipo de cáncer son matrices de dimensión (número de muestras) $\times$ (número de genes). Para cada muestra (individuo) y tipo de gen, se tiene el *conteo normalizado*, el cual representa el número estimado de lecturas mapeadas a dicho gen, normalizado al dividir por el percentil 75 de las lecturas en la muestra y multiplicado por 1000.

C

Métodos auxiliares

C.1. Modelos de MSV

C.1.1. Selección de variables: RFE

En muchos métodos de aprendizaje estadístico, incluyendo el método MSV, la selección de variables como parte del preprocesamiento de los datos es fundamental ya que no descartar características irrelevantes puede afectar el desempeño de los métodos en varios aspectos.

En primer lugar, la regularización implícita al seleccionar variables usualmente aumenta la capacidad de generalización de los modelos. En segundo lugar, el uso de variables irrelevantes (es decir, ruidosas o redundantes) aumenta considerablemente el tiempo de entrenamiento. En tercer lugar, la presencia de muchas variables puede dificultar o imposibilitar la convergencia de los algoritmos de op-

timización numérica subyacente.

Por otra parte, la identificación de variables importantes que tienen interpretación intuitiva es otro aspecto importante en muchas aplicaciones. En cuanto a la literatura del método MSV, tradicionalmente la selección de variables se realiza de forma independiente al proceso de entrenamiento de los modelos; sin embargo, en los últimos años han sido propuestos diversos enfoques para realizar la selección de variables y la construcción de los modelos de MSV en un mismo marco conceptual. No obstante, estos enfoques han sido en muchas ocasiones para dominios específicos y no existe ningún enfoque dominante.

En el caso particular de la aplicación del método MSV en este trabajo, se ha utilizado el método *eliminación recursiva de variables con remuestreo* (**RFE**, por sus siglas en inglés), propuesto por Max Kuhn en [12]. Éste es un método de selección de variables parecido al método *stepwise* hacia atrás, con la diferencia de que los métodos de remuestro (validación cruzada, bootstrapping, etc.) son incorporados para tomar en cuenta la variabilidad causada por la selección de variables al momento de calcular métricas de desempeño. Si bien este procedimiento proporciona mejores estimaciones, es más costoso computacionalmente aunque es fácilmente paralelizable. Por otra parte, éste método proporciona un enfoque más probabilístico de la importancia de las variables en comparación a cuando la selección de variables se realiza sobre un solo conjunto de datos de entrenamiento. Al final del algoritmo, se utiliza un ranking por consenso para determinar las mejores variables a mantener en los modelos. El seudoalgoritmo 1 ilustra el procedimiento RFE.

El número final de variables para el caso del Kernel lineal es de $p = 50$ mientras que para el caso del Kernel radial es de $p = 30$. La figura C.1.1 ilustra los resultados del proceso de selección de variables para el Kernel lineal y el Kernel radial, mientras que la figura C.1.2 muestra el conjunto final de variables para el modelo de MSV con Kernel radial, después de realizar el procedimiento RFE. Es interesante notar que en la literatura, las expresiones de algunos de los genes que aparecen en la figura C.1.2, tales como MSX1, SCGB2A1 y MARCKSL1, han sido halladas en

Algorithm 1 Eliminación recursiva de variables con remuestreo

for Cada iteración de remuestreo **do**
  Particionar los datos en muestras de entrenamiento y prueba, a través de remuestreo
  Estimar el modelo en la muestra de entrenamiento con todas las variables
  Realizar predicciones en la muestra de prueba
  Calcular la importancia de las variables
  for Cada tamaño de subconjuntos S_i, $i = \{1, \ldots, S\}$ **do**
    Mantener las S_i variables más importantes
    Estimar el modelo en la muestra de entrenamiento usando S_i variables
    Realizar predicciones en la muestra de prueba
  end for
end for
Calclular el desempeño sobre S_i utilizando la muestra de prueba
Determinar el número de variables a utilizar en el modelo final
Estimar el modelo final utilizando S_i óptimo en la muestra de entrenamiento original

diversas manifestaciones del cáncer ([42]).

C.1.2. Selección de parámetros

Para efectuar la selección de parámetros y obtener el mejor modelo posible en términos de predicción, en este trabajo se ha utilizado **grid search** con **validación cruzada**.

En el aprendizaje estadístico, la forma tradicional de realizar la optimización de los parámetros ha sido mediante *grid search*, que hace referencia simplemente a una búsqueda exhaustiva a través de un subconjunto especificado manualmente del espacio admisible de los parámetros de un algoritmo de aprendizaje. Esta técnica debe guiarse por alguna métrica de desempeño, típicamente medida mediante validación cruzada en el conjunto de entrenamiento. Dado que el espacio de parámetros de los métodos de aprendizaje estadístico puede incluir espacios de valores reales o espacios no acotados para ciertos parámetros, pueden ser necesa-

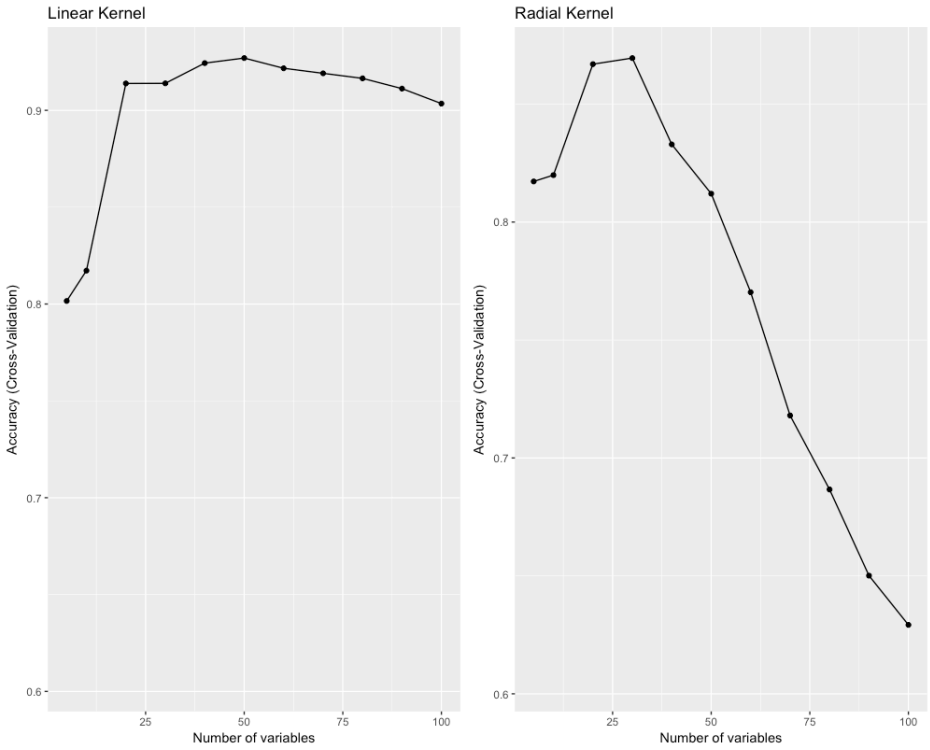

Figura C.1.1: Resultados del proceso de selección de variables para el Kernel lineal y el Kernel radial.

rios establecer límites o realizar discretizaciones manualmente previamente.

Por otra parte, la *validación cruzada* es una técnica de evaluación de modelos acerca de cómo se desempeña un modelo de predicción ante nuevas observaciones. En particular, en este trabajo se ha utilizado la técnica *k-fold cross validation* (validación cruzada de k capas). Adicionalmente, se utiliza esta técnica para calibrar parámetros de modelos; en nuestro caso se ha utilizado para elegir el parámetro C en el caso del Kernel lineal y los parámetros C y σ en el caso del Kernel no lineal. Bajo este enfoque, la muestra total es dividida en k sub-muestras del mismo tamaño. De las k sub-muestras, una de ellas es utilizada como muestra de prueba y las $(k-1)$ sub-muestras restantes son utilizadas como muestra de entrenamiento; este proceso es repetido k veces, utilizando cada sub-muestra como muestra de prueba una única vez.

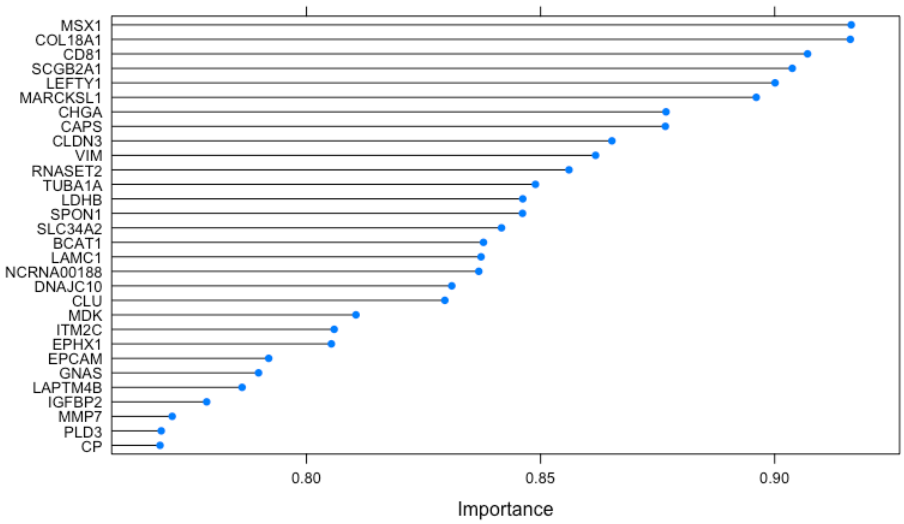

Figura C.1.2: Lista de variables utilizadas en modelo de MSV con Kernel radial y su medida de importancia.

Los k resultados son después promediados para producir una sola estimación:

$$\hat{cv} = \frac{1}{k} \sum_{j=1}^{k} \hat{e}_j \tag{C.1}$$

en donde $\hat{e}_j$ representa el cálculo de la métrica de desempeño en la j-ésima iteración. Además, el proceso permite calcular la desviación estandar de la estimación y por lo tanto tener intervalos de confianza para dicha estimación. Más detalles acerca de esta técnica y sus propiedades estadísticas pueden encontrarse en [2]. Por otra parte, se pueden utilizar métricas de desempeño para realizar este proceso. La figura C.1.3 ilustra este procedimiento.

En este trabajo se ha utilizado la precisión como el criterio a maximizar. El parámetro de validación cruzada fue establecido como $k = 3$ dado que el número n de observaciones en los datos utilizados es pequeño. La búsqueda de parámetros para los modelos de MSV fue realizada sobre los siguientes conjuntos:

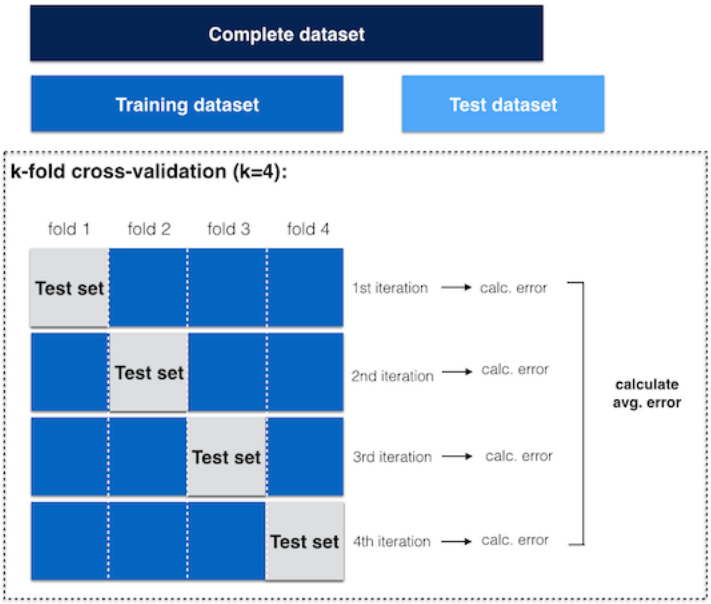

Figura C.1.3: Ilustración del proceso de validación cruzada para el caso particular con $k = 4$ y alguna métrica de error de predicción.

- **Kernel lineal**: $C \in \{0.1, 0.25, 0.5, 1, 10, 100\}$
- **Kernel no lineal**: $(C, \sigma) \in \{0.1, 0.25, 0.5, 1, 10, 100\} \times \{0.1, 0.25, 0.5, 1, 10, 100\}$

Para explotar el paralelismo en el problema de la selección de parámetros, en este trabajo se utilizaron los paquetes **foreach** y **doParallel** en R.

C.1.3. Reducción de dimensionalidad: t-SNE

El algoritmo *t-distributed stochastic neighbor embedding* (**t-SNE**, por sus siglas en inglés) es una técnica de reducción de dimensionalidad utilizada en el aprendizaje estadístico que es particularmente bueno para representar datos en espacios de alta dimensión en un espacio de baja dimensión, para luego ser visualizados. A grandes rasgos, representa cada punto de una dimensionalidad alta en un espacio de dimensión menor de forma que datos similares se representan como puntos cercanos y datos no similares se representan como puntos lejanos.

Dado un conjunto de n puntos $\mathbf{x}_1, \ldots, \mathbf{x}_n$, en donde $\mathbf{x}_i \in \mathbb{R}^p$, el algoritmo t-SNE calcula probabilidades p_{ij} que son proporcionales a la similitud entre los puntos $\mathbf{x}_i$ y $\mathbf{x}_j$ de la siguiente forma:

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2)} \tag{C.2}$$

en donde σ_i es la varianza de la distribución normal centrada en $\mathbf{x}_i$. Esta relación puede interpretarse como sigue: la similitud entre el punto $\mathbf{x}_j$ y el punto $\mathbf{x}_i$ es la probabilidad condicional $p_{j|i}$ de que $\mathbf{x}_i$ elijiera a $\mathbf{x}_j$ como su vecino si es que los vecinos fueran elegidos en proporción a su densidad de probabilidad bajo una distribución normal centrada en $\mathbf{x}_i$. Además, las probabilidades conjuntas p_{ij} se definen de forma que las probabilidades condicionales sean simétricas, es decir, $p_{ij} = \frac{p_{j|i}+p_{i|j}}{2n}$.

El algoritmo *t-SNE* tiene como objetivo aprender un mapeo d-dimensional $\mathbf{y}_1, \ldots, \mathbf{y}_n$, en donde $\mathbf{y}_i \in \mathbb{R}^d$ con $d < n$ que reproduzca p_{ij} tan bien como sea posible. Para esto, se calculan similitudes q_{ij} entre los puntos $\mathbf{y}_i$ y $\mathbf{y}_j$ utilizando un enfoque similar. Concretamente, esta similitud q_{ij} se define de la siguiente forma:

$$q_{ij} = \frac{(1 + \|\mathbf{y_i} - \mathbf{y_j}\|^2)^{-1}}{\sum_{k\neq m}(1 + \|\mathbf{y_k} - \mathbf{y_m}\|^2)^{-1}} \tag{C.3}$$

Aquí se utiliza una distribución t de Student para medir las similitudes entre los puntos de alta dimensión para permitir que los objetos no similares sean representados como distantes en el espacio de baja dimensión.

Finalmente, la localización de los puntos $\mathbf{y}_i$ en el espacio de baja dimensión son determinados al minimizar la divergencia de Kullback-Leibler (KL) entre las distribuciones Q y P, es decir:

$$KL(P\|Q) = \sum_{i\neq j} p_{ij}\log\frac{p_{ij}}{q_{ij}} \tag{C.4}$$

Este proceso de minimización representa un problema de optimización numérica que usualmente es llevado a cabo con el método de descenso de gradiente, o bien, su variante estocástica. Más detalles sobre este método pueden ser consul-

tados en [13] y [14]. En este trabajo se utilizó la implementación provista por el paquete **Rtsne** en R y se eligió el parámetro $d = 2$ para realizar la reducción de dimensionalidad del espacio de dimensión $p = 20501$ correspondiente a los genes.

C.2. Modelos de regresión logística

C.2.1. Selección de variables: modelo regularizado

La regresión logística se puede expresar en el contexto del aprendizaje estadístico mediante una función de pérdida definida como el negativo de la log-verosimilitud de la siguiente forma:

$$\begin{aligned} \ell(\mathbf{w}) &= -\sum_{i=1}^{n} \{y_i \log p(\mathbf{x}_i; \mathbf{w}) + (1 - y_i) \log(1 - p(\mathbf{x}_i; \mathbf{w}))\} \\ &= -\sum_{i=1}^{n} \{y_i \mathbf{w}^\intercal \mathbf{x}_i - \log(1 + \exp(\mathbf{w}^\intercal \mathbf{x}_i)\} \end{aligned} \tag{C.5}$$

En la práctica, métodos iterativos derivados del método de Newton son utilizados para esta estimación.

En el aprendizaje estadístico, el método *Lasso* (*least absolute shrinkage and selection operator*) es una técnica de regresión con regularización que realiza selección de variables con la finalidad de mejorar el poder de predicción del modelo que produce. Este método fue originalmente formulado para el problema de regresión lineal múltiple al modificar la función objetivo añadiendo la norma ℓ_1 de los parámetros del modelo multiplicada por un parámetro λ de regularización, y fue extendido directamente a una gran variedad de métodos estadísticos; en particular a los modelos lineales generalizados. La capacidad de este método para realizar selección de variables reside en la restricción que impone a los parámetros del modelo y posee múltiples interpretaciones en términos geométricos, en términos de la estadística bayesiana y en términos de la optimización convexa.

Como se ha mencionado, la regularización con norma ℓ_1 utilizada en el método

Lasso puede ser extendida a cualquier modelo lineal generalizado. En el caso de la regresión logística, la versión regularizada con norma ℓ_1 de C.5 está dada por:

$$Q(\mathbf{w}) = \sum_{i=1}^{n} - \{y_i \mathbf{w}^\intercal \mathbf{x}_i - \log(1 + \exp(\mathbf{w}^\intercal \mathbf{x}_i)\} + \lambda \|\mathbf{w}\|_1 \qquad \text{(C.6)}$$

Este problema de optimización numérica es resuelto en sus implementaciones más populares con el método de descenso por coordenadas. En este trabajo se utilizó la implementación provista por el paquete **glmnet** en R para la selección de variables sobre el conjunto total de $p = 20501$ genes. Más detalles sobre esta implementación pueden ser consultados en [11].

C.2.2. Selección del parámetro de regularización

El parámetro λ es elegido mediante *grid search* y *validación cruzada*. En la implementación utilizada en este trabajo, una secuencia de valores para el parámetro λ es generada a partir de λ_{max}, el valor de λ más pequeño para el cual todos los coeficientes son cero. La cardinalidad del conjunto de valores de λ es elegida arbitrariamente. En este trabajo, se utilizó una cardinalidad de 100 y la métrica elegida para la elección de λ es el error de clasificación. Además, en la implementación utilizada, ademas de poder obtener el valor λ con el que se obtiene la mejor métrica estimada por validación cruzada, también es posible elegir el valor más grande de λ tal que su métrica estimada se encuentra dentro de una desviación estándar de la métrica estimada del parámetro λ óptimo. En este trabajo, por fines de parsimonia, es éste valor de λ el que ha sido utilizado para elegir las variables a utilizar en el modelo no regularizado.

La figura C.2.1 muestra la relación entre la estimación del error de clasificación por validación cruzada, el número de variables cuyos coeficientes son distintos de cero, y el parámetro de regularización; las líneas verticales corresponden al valor de λ en donde la estimación del error de clasificación por validación cruzada es mínimo y al próximo valor de λ más grande tal que su estimación del error se encuentra a una desviación estándar del mínimo.

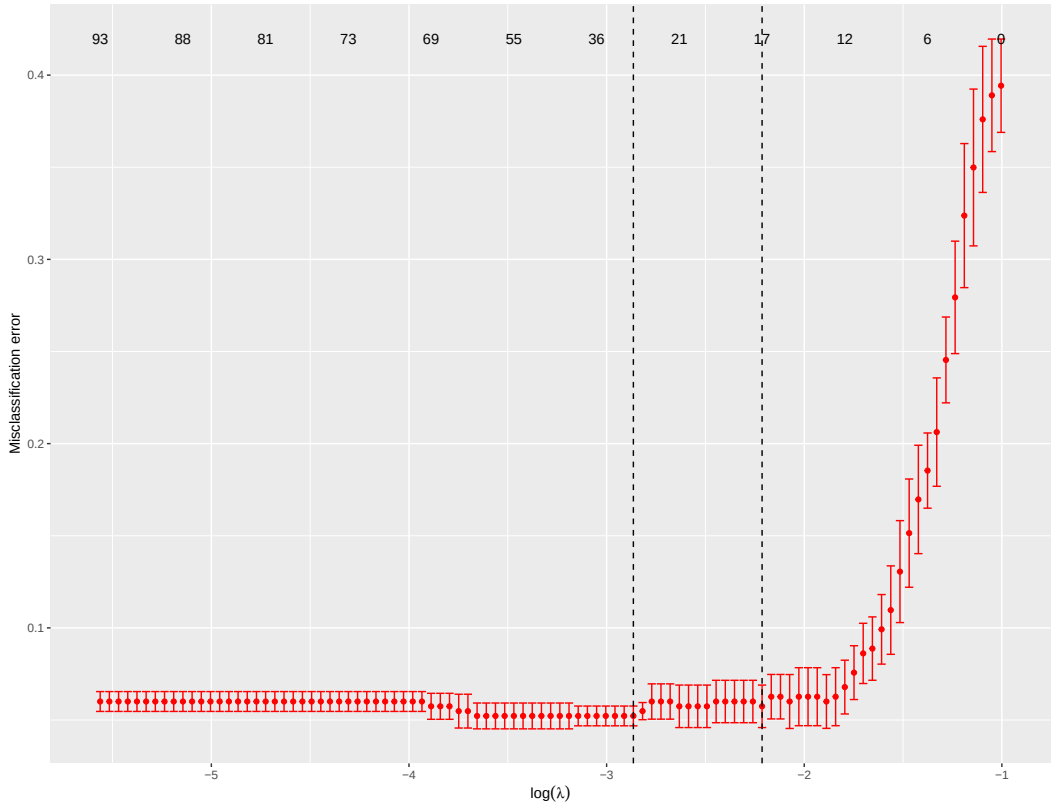

Figura C.2.1: Selección de variables como función del parámetro de regularización.

Bibliografía

[1] Bishop, C. (2010), *Pattern Recognition and Machine Learning*. Springer.

[2] Hastie, T. *et al.* (2008), *The elements of Statistical Learning*. Springer.

[3] Vapnik, V. and Cortes, C. (1995), *Support Vector Networks*. Machine Learning, 20.

[4] Karatzoglou, A. *et al.* (2013), *kernlab: An S4 Package for Kernel Methods in R*. The Comprehensive R Archive Network. Disponible en https://www.jstatsoft.org/article/view/v011i09/v11i09.pdf

[5] Shiloh, R. (2007), *Support Vector Machines for Classification and Regression*. M.Sc. Thesis, McGill University.

[6] Rifkin, R. (2002), *Everything Old Is New Again: A Fresh Look at Historical Approaches in Machine Learning*. Ph.D. Thesis, MIT.

[7] González, F. (2017), *Repositorio de notas, material y datos para el curso de Aprendizaje de Máquina ITAM*. Disponible en https://felipegonzalez.github.io/aprendizaje-maquina-2017/

[8] Campos, P. (2016), *Estudio Numérico de Algoritmos a Gran Escala para Máquinas de Soporte Vectorial*. Tesis de Licenciatura en Matemáticas Aplicadas, ITAM.

[9] Bottou, L. and Lin, C. (2007), *Support Vector Machines Solvers*. MIT Press.

[10] Platt, J. (1999), *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods*. Advances in Large Margin Classifiers. Vol. 10, No.3, pp. 61–74.

[11] Friedman J., Hastie, T. and Tibshirani, R. (2010), *Regularization Paths for Generalized Linear Models via Coordinate Descent*. Journal of Statistical Software. Vol. 33. Disponible en https://www.jstatsoft.org/article/view/v033i01/v33i01.pdf

[12] Kuhn, M. (2010), *Variable Selection Using the caret Package*. Disponible en http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.168.1655&rep=rep1&type=pdf

[13] Van Der Maaten, L. and Hinton, G. (2008), *Visualizing High-Dimensional Data Using t-SNE*. Journal of Machine Learning Research. Vol. 9. Disponible en https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf

[14] Van Der Maaten, L. (2009), *Learning a Parametric Embedding by Preserving Local Structure*. Proceedings of the Twelfth International Conference on Artificial Intelligence Statistics. Vol. 5. Disponible en https://lvdmaaten.github.io/publications/papers/AISTATS_2009.pdf

[15] Chen, T. and Guestrin, C. (2016), *XGBoost: A Scalable Tree Boosting System*. 22nd SIGKDD Conference on Knowledge Discovery and Data Mining. Disponible en https://arxiv.org/abs/1603.02754

[16] Varian, H. (2014), *Big Data: New Tricks for Econometrics*. Journal of Economic Perspectives, Vol. 28, No. 2, pp. 3-28. Disponible en http://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.28.2.3

[17] Stare, S. and Jozefowicz, J. (2008), *The Effects of Environmental Factors on Cancer Prevalence Rates and Specific Cancer Mortality*

Rates in a Sample of OECD Developed Countries. International Journal of Applied Economics. Vol. 5, No. 2, pp. 92-115. Disponible en https://www.researchgate.net/publication/228634693_The_Effects_of_Environmental_Factors_on_Cancer_Prevalence_Rates_and_Specific_Cancer_Mortality_Rates_in_a_Sample_of_OECD_Developed_Countries

[18] Kim, K. and Cho S. (2010), *Exploring Features and Classifiers to Classify MicroRNA Expression Profiles of Human Cancer*. International Conference on Neural Information Processing (ICONIP).

[19] Abeel T., Helleputte, T., Van de Peer Y., Dupont, P., and Saeys, Y. (2010), *Robust Biomarker Identification for Cancer Diagnosis with Ensemble Feature Selection Methods*. Bioinformatics, Vol. 26, No. 3, pp. 392–398.

[20] Haury, C., Gestraud, P., and Vert J. (2011), *The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures*. PLoS ONE, Vol. 6, No. 12.

[21] Liu, H., Liu, L. and Zhang, H. (2010), *Ensemble Gene Selection by Grouping for Microarray Data Classification*. Journal of Biomedical Informatics, Vol. 43, No. 1, pp. 81–87.

[22] Saeys, Y., Abeel, T. and Peer, Y. (2008), *Robust Feature Selection Using Ensemble Feature Selection Techniques*. Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases - Part II. Springer-Verlag, pp. 313–325.

[23] Yang, P., Ho, J. and Zhou, B. (2011), *Gene-Gene Interaction Filtering with Ensemble of Filters*. BMC Bioinformatics, Vol. 12.

[24] Yang, P., Hwa Yang Y., Zhou, B. and Zomaya A. (2010), *A Review of Ensemble Methods in Bioinformatics*. Current Bioinformatics, Vol. 5, No. 4, pp. 296–308.

[25] Yu, L., Han, Y. and Berens, M. (2012), *Stable Gene Selection from Microarray Data via Sample Weighting*. IEEE/ACM Trans. Comput. Biol. Bioinformatics, Vol. 9, No. 1, pp. 262–272.

[26] Ibrahim, R., Yousri, N., Ismail, M. and El-Makky N. (2013), *miRNA and Gene Expression Based Cancer Classification Using Self-Learning and Co-Training Approaches*. IEEE International Conference on Bioinformatics and Biomedicine.

[27] Sun, Z., Zheng, C., Gao Q., Zhang, J. and Zhang, D. (2012), *Tumor Classification Using Eigengene-Based Classifier Committee Learning Algorithm*. IEEE Signal Processing Letters, Vol. 19, No. 8.

[28] Zhao, H. (2014), *Analizying TCGA Genomic and Expression Data Using SVM with Embedded Parameter Tuning*. M.Sc. Thesis, The University of Akron.

[29] Li, Y., Kang, K., Krahn, J., Croutwater, N., Lee, K., Umbach, D. and Li, L. (2017), *A Comprehensive Genomic Pan-Cancer Classification Using The Cancer Genome Atlas Gene Expression Data*. BMC Genomics, Vol. 18, No. 508. Disponible en https://doi.org/10.1186/s12864-017-3906-0

[30] Wang, K. *et al.* (2010), *MapSplice: Accurate Mapping of RNA-Seq Reads for Splice Junction Discovery*. Nucleic Acids Research, Vol. 38, No. 18. Disponible en https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq622

[31] Li, B. and Dewey, C. (2011), *RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome*. BMC Bioinformatics, Vol. 12, No. 323. Disponible en https://doi.org/10.1186/1471-2105-12-323

[32] American Cancer Society and LiveStrong (2010), *The Global Economic Cost of Cancer*. Disponible en http://phrma-docs.phrma.org/sites/default/files/pdf/08-17-2010_economic_impact_study.pdf

[33] Unidad de Análisis Económico de la Secretaría de Salud de México (2015), *Impacto Económico del Cáncer en México*. Disponible en http://www.cefp.gob.mx/difusion/evento/2015/forocancer/presentaciones/po2.pdf

[34] Fierro, V., Ibarra, D., Sánchez J. and Soberanes, A. (2015), *Impacto Económico del Cáncer Cervicouterino en México*. Debate Económico, Vol. 4 (2), No. 11, pp. 87-100. Disponible en http://www.laes.org.mx/wp-content/uploads/2016/03/de-1104impact-eco-cancer.pdf

[35] Ruvalcaba-Limón, E. *et al.* (2015), *Primer Consenso Mexicano de Cáncer de Endometrio. Grupo de Investigación en Cáncer de Ovario y Tumores Ginecológicos de México*. Revista de Investigación Clínica, Vol. 62, No. 6, pp. 585-605. Disponible en http://www.medigraphic.com/pdfs/revinvcli/nn-2010/nn106m.pdf

[36] *The Genetics of Cancer*. Obtenido en agosto 2017 de: https://www.cancer.gov/about-cancer/causes-prevention/genetics

[37] *Cancer Genomics*. Obtenido en agosto 2017 de: https://www.cancer.gov/research/areas/genomics

[38] *The Cancer Genoma Atlas: Program Overview*. Obtenido en agosto 2017 de: https://cancergenome.nih.gov/abouttcga/overview

[39] *The Cancer Genoma Atlas: Cancers Selected for Study*. Obtenido en agosto 2017 de: https://cancergenome.nih.gov/cancersselected

[40] *TCGA2STAT: simple TCGA data access for integrated statistical analysis in R*. Obtenido en agosto 2017 de: https://academic.oup.com/bioinformatics/article/32/6/952/1744407/TCGA2STAT-simple-TCGA-data-access-for-integrated

[41] *Introdution to Gene Expression: The Central Dogma*. Obtenido en agosto 2017 de: https://www.khanacademy.org/

science/biology/gene-expression-central-dogma/central-dogma-transcription/a/intro-to-gene-expression-central-dogma

[42] *GeneCards: Human Gene Database*. Obtenido en agosto 2017 de: http://www.genecards.org/

[43] *The Human Protein Atlas: Cancer Atlas*. Obtenido en agosto 2017 de: http://www.proteinatlas.org/cancer