Contrastación de Técnicas Econométricas Tradicionales y Aprendizaje Automático en la Predicción de los Precios de los Apartamentos de Santiago de Cali en el 2019

Sebastián Dow Valenzuela
Fabián Andrés Salazar Jaramillo

Bajo la dirección del profesor
Luís Eduardo Girón Cruz

Pontificia Universidad Javeriana Cali
Facultad de Ciencias Económicas
y Administrativas

Cali - Valle del Cauca
2023

Resumen

En el presente trabajo se pretenden contrastar las predicciones de los precios obtenidos por técnicas tradicionales de econometría y técnicas computacionales basadas en el aprendizaje automático. A partir de datos de 5074 apartamentos en Cali en el 2019 con sus características obtenidos de las páginas de ventas de inmuebles y utilizando regresión múltiple, K-NN, regresión LASSO y bosques aleatorios, encontrando que, en general, las técnicas de Machine Learning arrojan predicciones más precisas que el método de pronóstico fundamentado en regresión múltiple pero no por un margen muy amplio.

Palabras clave:

Econometría, Aprendizaje Automático, Regresión Múltiple, K-NN, Máquinas de Soporte Vectorial, Bosques Aleatorios, Pronóstico, LASSO

Dedicatoria

Dedicamos este trabajo a aquellos que han sido nuestra fuente de inspiración, apoyo y motivación a lo largo de este arduo camino. A nuestros familiares y seres queridos, quienes han sido testigos incansables de nuestras largas horas de estudio, de nuestras preocupaciones y triunfos. Su amor incondicional y aliento constante nos han impulsado a alcanzar este logro.

A nuestros amigos y compañeros, quienes han compartido con nosotros risas, desafíos y momentos inolvidables. Su amistad y camaradería nos han brindado el equilibrio perfecto entre el estudio y la diversión, haciendo que este proceso sea mucho más enriquecedor.

A nuestros profesores y mentores, cuya sabiduría, guía y conocimiento nos han iluminado en cada paso de esta travesía académica. Sus enseñanzas han moldeado nuestro pensamiento crítico y nos han impulsado a superar obstáculos y alcanzar nuevas metas.

Este trabajo es el resultado de nuestra dedicación conjunta y trabajo en equipo. Agradecemos mutuamente el esfuerzo compartido, las largas horas de discusión y el compromiso constante para lograr los mejores resultados. Juntos hemos enfrentado desafíos, celebrado victorias y aprendido valiosas lecciones.

Por último, agradecemos a todas las personas que de alguna manera contribuyeron a nuestro crecimiento académico y personal. Vuestra confianza en nosotros nos ha dado la fortaleza para enfrentar este reto y nos ha recordado que no estamos solos en este viaje.

A todos ustedes, dedicamos esta tesis como un humilde reconocimiento a su invaluable influencia en nuestras vidas. Sin su apoyo, este logro no habría sido posible. Gracias por ser parte de nuestro camino.

Agradecimientos

En primer lugar, queremos expresar nuestro profundo agradecimiento a Dios y a la Santísima Virgen María, fuente de toda sabiduría y fortaleza, por guiar nuestros pasos en este arduo camino. Agradecemos por las oportunidades brindadas, por las bendiciones recibidas y por sostenernos en cada desafío y obstáculo. Reconocemos que sin su amor y gracia, este logro no habría sido posible.

Nos gustaría expresar nuestros más sinceros agradecimientos a todas las personas que contribuyeron de manera significativa a la realización de esta tesis de grado. Este logro no habría sido posible sin su apoyo, orientación y aliento constante a lo largo de este emocionante viaje.

Queremos agradecer especialmente a nuestro director de tesis, Luis Eduardo Girón Cruz, por su sabiduría, orientación y dedicación. Su experiencia y conocimiento nos han guiado en cada etapa del proceso de investigación. Agradecemos su paciencia, apoyo incondicional y la confianza depositada en nosotros.

Extendemos nuestro agradecimiento a nuestros profesores por su valiosa retroalimentación y sugerencias constructivas que han mejorado significativamente nuestro trabajo, especialmente al profesor David Arango Londoño quien además nos proporciono la base de datos con la cual se realizo esta tesis, al profesor Jaime Rafael Ahcar Olmos, a la profesora Leidi Johana Rojas y al profesor Yoan José Pinzón Ardila por Su asesoria y comentarios que nos han ayudaron a perfeccionar nuestra investigación y ampliar nuestro conocimiento en el campo.

A nuestros amigos y compañeros de clase, en especial a Isabella Rebolledo, Juan Martín Giraldo, Carolina Arboleda y Andrés Camilo Neira, gracias por estar a nuestro lado durante todo el proceso. Sus conversaciones estimulantes, discusiones y apoyo mutuo han enriquecido nuestra experiencia y han hecho que esta travesía sea memorable. Valoramos su amistad y el tiempo compartido juntos.

Nuestros mas profundos agradecimientos a la Pontificia Universidad Javeriana Cali, a las becas John Boris Rincón y Reina de la Paz por su confianza en nuestras capacidades y el reconocimiento de nuestro potencial académico han sido un gran estímulo para nuestra motivación y dedicación, su apoyo ha sido crucial.

Por último, pero no menos importante, queremos agradecernos mutuamente. Esta tesis es el resultado de nuestra colaboración y trabajo. Nos hemos apoyado mutuamente, hemos compartido ideas y hemos superado desafíos juntos. Nuestra dedicación conjunta y compromiso han sido la clave.

A todas las personas mencionadas anteriormente y a aquellas que no pudieron ser mencionadas, les agradecemos sinceramente por su contribución y apoyo en este importante logro. Sus palabras de aliento, amistad y consejos han dejado una huella imborrable en nuestras vidas.

Nuestro más profundo agradecimiento a todos por ser parte de este emocionante viaje académico, que este logro sea también un testimonio de nuestra gratitud y reconocimiento hacia ustedes.

Con aprecio,
Sebastián Dow Valenzuela
Fabian Andrés Salalazar Jaramillo

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Frente a la creciente necesidad humana de refinar las técnicas existentes para la minimización de la incertidumbre, en la actualidad las instituciones académicas y corporativas interesadas en mantenerse a la vanguardia del desarrollo tecnológico y científico buscan implementar métodos cada vez más precisos para la predicción de eventos en diversos campos del conocimiento. Susan Athey, profesora de la Universidad de Harvard y Universidad de Stanford, afirma que el aprendizaje automático tendrá un efecto dramático el campo de la economía en un corto periodo de tiempo, y que debido a que este es un proceso que ya está en marcha, no es demasiado difícil predecir algunos de sus efectos. (Athey, 2018)

La econometría y el aprendizaje automático son campos distintos y complementarios que se utilizan para analizar y hacer predicciones a partir de datos. La econometría, es una rama de la economía que utiliza métodos estadísticos para probar y estimar teorías económicas. Los economistas usan modelos matemáticos para representar relaciones entre variables económicas, y los econometristas, estiman dichas relaciones para hacer predicciones sobre tendencias y resultados futuros. Estos modelos a menudo se basan en la suposición de una relación causal entre las variables y se estiman utilizando datos de observaciones anteriores. (Varian, 2014)

Por otro lado, el aprendizaje automático, según Tom Mitchell, es un subcampo de la inteligencia artificial que se enfoca en el desarrollo de algoritmos que pueden ”aprender automáticamente” patrones en los datos y hacer predicciones. En contraste con la econometría, los algoritmos de aprendizaje automático no se basan en suposiciones a priori sobre las relaciones entre variables. En su lugar, utilizan datos para capturar las relaciones por sí mismos, lo que los hace ideales para aplicaciones donde hay una gran cantidad de observaciones y donde los modelos tradicionales pueden no proporcionar predicciones precisas.

Uno de los grandes desafíos que enfrenta la econometría en la actualidad son los desarrollos recientes que buscan instaurar las técnicas basadas en el aprendizaje automático para la realización de pronósticos más precisos, más aún, en vista del acelerado crecimiento del mercado de big data, la literatura reciente sobre aprendizaje automático promueve la introducción de métodos más sofisticados para el apoyo y la realización de predicciones en múltiples áreas (Varian, 2014).

Debido a esto, se hace necesario evaluar el alcance que estas herramientas han logrado en la actualidad, y la pertinencia de prescindir o no de las técnicas tradicionales ante los avances recientes, además de la posibilidad de la integración de los métodos más precisos y eficientes tanto en

econometría como en aprendizaje automático en cuanto sea factible.

Tanto la econometría como el aprendizaje automático tienen sus propias contribuciones únicas a los campos de la economía y la ciencia de datos. La econometría proporciona un enfoque estructurado e interpretable para comprender las relaciones económicas, con la capacidad de realizar pruebas de hipótesis, pruebas de significancia y análisis de sensibilidad. El aprendizaje automático proporciona un enfoque flexible y basado en datos para hacer predicciones, sin necesidad de suposiciones a priori sobre las relaciones entre variables.

Teniendo presente lo anteriormente expuesto, es de esperarse que un modelo econométrico proporcione una validación a la teoría en las cual se fundamenta, ayudando a comprender las relaciones existentes en el sistema que se representa y brindando una aproximación general al panorama económico, mientras que uno desarrollado a partir de técnicas de aprendizaje automático sea más práctico a la hora de pronosticar.

No obstante, una verificación empírica de esto puede ayudar a cuantificar la brecha de la precisión en el pronóstico entre ambas disciplinas. En última instancia, la elección entre econometría y aprendizaje automático depende del problema específico en cuestión, la disponibilidad y calidad de los datos, y el nivel deseado de interpretabilidad y precisión de las predicciones.

En el presente trabajo de investigación se pretenden desarrollar modelos predictivos sobre el comportamiento de los precios de los apartamentos en el mercado inmobiliario de Santiago de Cali, en aras de contrastar los resultados obtenidos con las técnicas econométricas y los algoritmos de aprendizaje automático, haciendo uso de los recursos y las librerías con las que cuentan los lenguajes de programación de R, para determinar el alcance en términos de pronóstico.

Capítulo 2

Estado del arte y planteamiento del problema

En el marco puntual de la predicción de los precios en los mercados inmobiliarios de distintas naciones, se han desarrollado algunos estudios que han hecho uso de estas metodologías para determinar su alcance y las ventajas de sus aplicaciones para la predicción en base a diferentes estructuras de datos.

En 2017 se desarrolló un estudio en la universidad de Sevilla, el cual tenía como objetivo estimar los precios de la viviendas en la ciudad de Sevilla mediante un modelo hedónico de precios, obteniendo como principales resultados que el precio de las vivienda depende sobre todo de los metros cuadrados construidos (del Simeón;2017).

En 2022 se desarrolló una investigación para estimar los precios de las viviendas en el mercado inmobiliario de algunas ciudades de Francia utilizando siete técnicas computacionales basadas en el aprendizaje automático y la georreferenciación. En este trabajo se expone que debido a la complejidad y la no linealidad del problema de estimación de precios en los mercados inmobiliarios, en lugar de métodos de regresiones hedónicas de estimación de la demanda en base a preferencias reveladas, se hace uso de varios métodos computacionales como análisis envolvente de datos, lógica difusa y algoritmos genéticos (Tchuente & Nyawa, 2021).

En 2019 se realizó un análisis sobre la relevancia de la reciente literatura en la economía y la econometría, haciendo énfasis en los objetivos, métodos y contextos entre la literatura tradicional sobre econometría y estadística y la literatura sobre machine learning que consideran importantes para la realización de investigaciones empíricas en ambas áreas (Athey & Imbens, 2019a).

En 2019 se realizó la comparación entre el aprendizaje automático (machine learning) y la econometría en la predicción de la taquilla de películas. Se enfoca en el análisis de la capacidad predictiva de ambos enfoques en el contexto del rendimiento económico de las películas en taquilla. El estudio examina la eficacia de los modelos de aprendizaje automático en comparación con los enfoques tradicionales de la econometría en términos de precisión y capacidad para capturar las complejidades del mercado cinematográfico. El objetivo es proporcionar una visión general del estado del arte en este campo y resaltar las ventajas y desafíos de cada enfoque en la predicción de la taquilla de películas. (Liu & Xie, 2019)

Como se observa en los trabajos anteriores, existen pocas investigaciones que traten los pronós-

ticos en un área determinada como la predicción de precios en el mercado inmobiliario utilizando y comparando ambas metodologías (econometría tradicional y aprendizaje automático), por lo tanto, la pregunta de investigación que rige este trabajo es: ¿Qué enfoque tiene una mayor capacidad predictiva para predecir los precios de los apartamentos?

Capítulo 3

Objetivo

En el presente trabajo se pretende contrastar las predicciones de los precios de los apartamentos en Cali en el año 2019 obtenidas mediante el uso de técnicas tradicionales de econometría y técnicas computacionales basadas en el aprendizaje automático.

3.1. Objetivos específicos

- Describir el funcionamiento de las técnicas econométricas y de machine learning utilizadas.
- Definir el alcance de los métodos econométricos frente a las técnicas computacionales para la predicción del comportamiento de los precios en el mercado inmobiliario de Santiago de Cali.

Capítulo 4

Metodología

En esta sección se describen detalladamente los pasos seguidos en el presente trabajo de investigación, así como el tipo de estudio, las fuentes y técnicas de recolección de datos y las técnicas para el tratamiento de la información.

4.1. Tipo de estudio y método de investigación

Se pretende realizar un análisis cuantitativo para evaluar la capacidad predictiva de la técnica de regresión lineal múltiple frente a los métodos algorítmicos k-Nearest Neighbors, LASSO y Bosques Aleatorios. En este estudio se describen y aplican las técnicas y los modelos que se pretenden contrastar, se realiza un análisis de correlación entre las variables cuantitativas seleccionadas para la predicción, la confrontación de la eficacia de los modelos con base en criterios determinados, y a la explicación de las razones por las cuales un método puede ser más eficiente que otro en el ámbito específico de predicción.

Para la presente investigación, se escogieron los métodos de k-NN, LASSO y Bosques Aleatorios debido a su popularidad, facilidad de aplicación e interpretación y eficacia probada para afrontar problemas de regresión en economía. La regresión LASSO permite además, la selección de las variables que capturan una mayor variación para el modelo. No se utilizaron métodos como SVM debido a que su utilidad se limita a problemas de clasificación.

4.2. Fuentes y técnicas de recolección de datos

Para el desarrollo de los modelos, se realizará un proceso de investigación en fuentes secundarias (las paginas web, metrocuadrado.com y fincaraiz.com.co). Se trabajará con una muestra compuesta de un conjunto de 5074 observaciones con datos de apartamentos a la venta en la zona urbana de la ciudad de Santiago de Cali para el año 2019, recolectados mediante un algoritmo de scraping, el cual, permitirá recoger datos sobre los apartamentos de varias plataformas digitales de compra, venta y registro de estos bienes. Posteriormente, se realizará un proceso de estructuración de los datos obtenidos, en el que se filtrarán y depurarán datos atípicos, datos incompletos para la modelación mediante las técnicas ya mencionadas.

4.3. Técnicas para tratamiento de la información

Se realizará un breve análisis descriptivo del conjunto de datos estructurado para obtener la información más relevante para la modelación. Posteriormente, se formulará el modelo de regresión múltiple con base en la teoría de precios hedónicos el cual será estimado en R, y se obtendrán los modelos de aprendizaje automático utilizando las distintas técnicas mencionadas.

Para calcular el error cuadrático medio de predicción y el coeficiente de desigualdad de Theil en los modelos, en aras de obtener métricas comparables, se dividirá la muestra en dos submuestras aleatorias: una submuestra de estimación o entrenamiento con el 70 % de los datos y una de prueba con el 30 % restante que se utilizará para medir la capacidad de predicción de todos los modelos.

Inicialmente, para el modelo econométrico se realizará una regresión lineal múltiple utilizando todas las observaciones y se desarrollará una serie de pruebas para verificar que el modelo esté correctamente especificado. Posteriormente, se realizará un modelo con la misma estructura utilizando los datos del conjunto de entrenamiento para comparar su poder predictivo frente a los métodos de ML.

Luego, se desarrollarán los modelos computacionales y se definirán los parámetros óptimos (tanto los coeficientes del modelo como el parámetro de penalización) de la regresión LASSO en función de los resultados obtenidos mediante k-Fold Cross Validation con los datos del conjunto entrenamiento.

En este punto se podrá realizar la contrastación entre las diferentes técnicas con base en el error cuadrático medio de predicción y la desigualdad de Theil obtenidos con los datos de prueba para evaluar la exactitud de las predicciones obtenidas con los métodos.

Posteriormente, se expondrá una serie de conclusiones sobre los resultados obtenidos en el proceso de contrastación de las técnicas puestas en práctica para la realización de los modelos predictivos, que brinden luces sobre las ventajas y desventajas de estos métodos.

Capítulo 5

Marco conceptual

5.1. Econometría

Según varios autores como Gujarati, Wooldridge y Samuelson, la econometría es una disciplina interdisciplinaria que se basa en el análisis estadístico para estimar empíricamente las relaciones cuantitativas entre variables económicas. Integra la economía, la estadística, el análisis matemático y herramientas computacionales, bajo un supuesto de relación causal entre un conjunto de variables dentro del sistema económico que se estudia, para explicar el impacto que tiene un grupo de ellas sobre otra.(Gujarati et al., 2010)

Los modelos desarrollados con econometría se utilizan para hacer predicciones sobre tendencias y resultados futuros, proporcionando un enfoque estructurado, sistemático y científico para comprender y pronosticar fenómenos económicos. La validez de las predicciones desarrolladas en econometría depende de la precisión de las suposiciones hechas sobre las relaciones entre variables y la calidad de los datos usados para estimar el modelo.(Pindyck & Rubinfeld, 1998)

5.2. Aprendizaje Automático

En contraste con la econometría, el aprendizaje automático se apoya en los métodos algorítmicos para la determinación de patrones con base en el análisis computacional de datos empíricos (James et al., 2013). En otras palabras, simula procesos de aprendizaje para dilucidar las posibles relaciones existentes entre las variables de un sistema pero sin la definición a priori de un marco teórico en el cual se fundamentarán dichas relaciones.

Aunque tanto la econometría como el aprendizaje automático utilizan herramientas computacionales para la automatización de los procesos algorítmicos (Generalmente de optimización matemática), necesarios para la estimación del modelo, el aprendizaje automático hace uso de métodos más sofisticados de estadística computacional orientados a la aplicación práctica del modelo para predicción (Varian, 2014).

5.3. Predicción y pronóstico

Una predicción es una declaración sobre cierto evento o fenómeno desconocido en el presente. En el caso de un pronóstico, se busca hacer una predicción basándose en experiencias previas relacionadas con el fenómeno en cuestión.

Un pronóstico, según la definición de (Pindyck & Rubinfeld, 1998), es una estimación cuantitativa de la verosimilitud de un efecto futuro que se elabora con base en información pasada y presente. Cuando se habla de modelos predictivos, un pronóstico es el valor predicho de una variable dependiente dada cierta configuración de las variables independientes para una población o muestra para un momento determinado del tiempo.

La principal motivación para el desarrollo de un modelo uniecuacional en econometría o un modelo de aprendizaje automático es el pronóstico, cuyo resultado puede depender de factores observables, que se incorporan en el modelo de forma directa, y no observables, que son incorporados como parte del proceso aleatorio que sigue la variable dependiente.

5.3.1. Predicción media e individual

En el campo de la econometría, hay dos enfoques fundamentales para la elaboración de modelos predictivos: la predicción media y la predicción individual. La predicción media se refiere al valor promedio de una variable en una muestra, mientras que la predicción individual se refiere al valor de una observación individual en la muestra.

La predicción promedio se obtiene al estimar los parámetros de un modelo utilizando datos de la muestra y luego usarlos para estimar o predecir puntualmente el valor medio de la variable dependiente para un conjunto de valores dado para las variables independientes. (Pindyck & Rubinfeld, 1998).

Por otro lado, la predicción individual se enfoca en predecir un valor individual de una variable dependiente dado un valor de una variable independiente. Esta predicción se obtiene utilizando los parámetros estimados del modelo y los valores de las variables independientes para la observación individual.(Gujarati et al., 2010)

Es importante tener en cuenta que tanto la predicción individual como la de media pueden ser obtenida en forma puntual o por intervalo para un nivel de confianza dado. La predicción de punto predice un valor puntual o la media para la variable explicada en el modelo en una instancia de las variables regresoras, mientras que la predicción de intervalo define un intervalo en el que se espera que se encuentre el valor real de la variable explicada. (Wooldridge, 2016).

5.4. Pronóstico en econometría

En econometría existen pronósticos uniecuacionales o multiecuacionales dependiendo la estructura del modelo. Actualmente se utilizan los modelos VAR que han demostrado tener una mayor potencia predictiva que los modelos uniecuacionales. No obstante, en este documento nos enfocaremos en la utilización de pronósticos uniecuacionales.

Para estimar un modelo uniecuacional, el enfoque de mínimos cuadrados ordinarios (MCO) es ampliamente utilizado en la literatura económica y, mientras se cumplan los supuestos para que el estimador sea MELI, produce el pronóstico con la menor varianza entre todos los estimadores lineales insesgados. Una vez que los parámetros de un modelo lineal uniecuacional han sido estimados, se puede calcular el error de pronóstico como la diferencia entre el valor esperado de la variable dependiente dada una configuración de las variables explicativas y la estimación de dicho valor esperado dada la misma configuración de las variables del modelo.

Capítulo 6

Marco teórico

En el proceso de elección racional, los individuos buscan alcanzar cierto grado de comprensión del sistema en el que interactúan los elementos que determinan el resultado de su decisión, con el fin de tener el mayor control posible sobre él. Por esta razón, se han desarrollado ciencias para modelar los sistemas naturales y sociales que interactúan directa o indirectamente con nosotros, y que nos garantizan ciertos márgenes de ganancia o pérdida en el resultado de nuestras elecciones.

Como se mencionó en el marco conceptual, este proceso de aproximación a un resultado desconocido a través de un modelo de la realidad se conoce como pronóstico. Sin embargo, debido a la complejidad de los sistemas que se estudian, siempre existen elementos que no se tienen en cuenta en el proceso de modelación, lo cual genera una probabilidad de error inevitable en el pronóstico.

Para acotar los valores en los que se estima que existirá una probabilidad significativa de que se produzca un resultado, se utiliza la estadística. La estadística es una disciplina que se desarrolló para estimar el resultado de eventos aleatorios con base a anteriores observaciones del mismo. La observación del experimento que se busca modelar será aleatoria debido a la información desconocida presente, ya sea por el límite en las capacidades de los investigadores o por la simplicidad del modelo, lo que afectará su resultado.

Ahora bien una de las grandes limitantes de la ciencia económica es la presencia de una gran cantidad de variables no observables en los sistemas que modela, las cuales potencialmente afectan el resultado del proceso de elección racional de los individuos. Por esta razón, los modelos económicos se basan en una gran cantidad de supuestos que restringen las condiciones involucradas en este proceso de decisión. La econometría es la rama de la economía que, con base en el análisis estadístico, se encarga de la estimación de modelos económicos.

6.1. Análisis estadístico

Dependiendo de las características de las variables que se pretenden estudiar a través de un modelo, la forma en la que se formula el modelo y se interpretan los resultados puede variar, así como medidas asociadas al comportamiento de los datos.

Las variables incluidas en un modelo estadístico pueden ser:

- **Variables cualitativas:** Estas variables pueden tomar una cantidad finita de valores pertenecientes a un conjunto de categorías definidas. Estas categorías pueden ser dos, en el caso

de las variables binomiales, como por ejemplo el éxito o fracaso en un experimento, o multinomiales, si hay más de dos categorías en el conjunto de posibilidades de la variable. Adicionalmente, si se ha establecido una relación de orden entre las categorías de la variable, se puede clasificar la misma como ordinal, como en el caso de rangos socio-económicos bajo, medio-bajo, intermedio, medio-alto o alto, o nominal en caso contrario, como en el caso de países o etiquetas similares.

- **Variables cuantitativas:** Son aquellas que adoptan valores escalares pertenecientes a conjuntos numéricos, tales como temperatura o tiempo. Estas variables pueden ser continuas, como en el caso de una magnitud física como longitud o peso, o discretas, como en el caso de cantidades de productos industriales.

6.1.1. Medidas de tendencia central y dispersión

Las medidas de tendencia central y dispersión son conceptos fundamentales en estadística ya que nos permiten resumir y comprender las características de un conjunto de datos. En el contexto de la regresión, estas medidas son especialmente relevantes, ya que nos proporcionan información sobre la distribución y la variabilidad de las variables involucradas.

La media es una de las medidas de tendencia central más importantes en estadística, representa el valor promedio de un conjunto de datos. Se calcula sumando todos los valores y dividiendo la suma por el número total de observaciones. Formalmente, la media se denota con la letra griega μ y se calcula de la siguiente manera:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \tag{6.1}$$

donde x_i representa cada valor individual en el conjunto de datos y n es el número total de observaciones. La media es particularmente útil en la regresión, ya que proporciona una medida representativa del comportamiento medio de la variable dependiente dado un conjunto de valores asignados a las variables independientes.

Además de las medidas de tendencia central, las medidas de dispersión son igualmente importantes en la regresión, ya que nos permiten entender la variabilidad de los datos y evaluar la precisión de nuestros modelos. Una medida común de dispersión es la varianza, que mide la dispersión de los valores individuales con respecto a la media. Formalmente, la varianza se denota como σ^2 y se calcula de la siguiente manera:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n} \tag{6.2}$$

donde x_i representa cada valor individual en el conjunto de datos, μ es la media y n es el número

total de observaciones. Sin embargo, la varianza se expresa en unidades cuadradas de los datos, lo cual puede dificultar la interpretación directa. Por lo tanto, es común utilizar la desviación estándar, que es la raíz cuadrada de la varianza, para tener una medida de dispersión en las mismas unidades que los datos originales. La desviación estándar se denota como σ y se calcula de la siguiente manera:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}} \tag{6.3}$$

Otra medida de dispersión que se utiliza es el coeficiente de variación, que es una medida relativa de la dispersión. El coeficiente de variación se calcula dividiendo la desviación estándar por la media y multiplicando el resultado por 100 para obtener un porcentaje. Esta medida es útil cuando se desea comparar la variabilidad de dos conjuntos de datos que tienen escalas diferentes. El coeficiente de variación se denota como CV y se calcula de la siguiente manera:

$$CV = \left(\frac{\sigma}{\mu}\right) \times 100 \tag{6.4}$$

Las medidas de tendencia central y dispersión, como la media, la mediana, la moda, la varianza, la desviación estándar y el coeficiente de variación, son fundamentales en la regresión estadística. Estas medidas nos permiten entender el comportamiento central de los datos, así como su variabilidad. Además, son útiles para evaluar la calidad de los modelos de regresión y la precisión de las predicciones. El uso adecuado de estas medidas contribuye a una interpretación sólida y rigurosa de los resultados obtenidos en el análisis de regresión.

6.1.2. Distribuciones de probabilidad y probabilidad acumulada

La probabilidad juega un papel fundamental en el análisis de regresión ya que se relaciona con la verosimilitud de la ocurrencia de un evento puntual en el fenómeno modelado. La probabilidad es una medida numérica que asigna un valor entre 0 y 1 a un evento o conjunto de eventos. Representa la posibilidad de que un evento en particular ocurra. En el contexto de la regresión, la probabilidad se utiliza para cuantificar la incertidumbre asociada a los resultados y estimaciones del modelo.

Una distribución de probabilidad es una función matemática que describe la frecuencia relativa de los diferentes posibles valores de una variable aleatoria. En la regresión, estas distribuciones son importantes porque permiten modelar y analizar los errores o residuos del modelo, que son las diferencias entre los valores observados y los valores predichos por el modelo de regresión. Al suponer que los errores siguen una distribución de probabilidad específica, se pueden realizar inferencias y cálculos estadísticos más precisos.

Existen varias distribuciones de probabilidad comúnmente utilizadas en la regresión estadís-

tica, las cuales, describen el comportamiento de la tendencia central y la dispersión de los datos dependiendo de su naturaleza. Las distribuciones de probabilidad se clasifican en dos categorías: las discretas y las continuas:

Las distribuciones de probabilidad discretas se utilizan para modelar variables aleatorias que pueden tomar un número finito o infinito numerable de valores. Un ejemplo común de distribución de probabilidad discreta es la distribución binomial, que se utiliza para modelar el número de éxitos en un número fijo de ensayos independientes.

Las distribuciones de probabilidad continuas se utilizan para modelar variables aleatorias que pueden tomar cualquier valor dentro de un intervalo continuo. Un ejemplo común de distribución de probabilidad continua es la distribución normal.

La distribución normal, también conocida como distribución de Gauss, es una de las distribuciones más importantes y ampliamente utilizadas. Se caracteriza por su forma de campana simétrica y está completamente determinada por su media (μ) y desviación estándar (σ). Se denota como $X \sim N(\mu, \sigma^2)$ y su función de densidad de probabilidad está dada por:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

donde x es una variable aleatoria, μ es la media y σ es la desviación estándar de la distribución.

La distribución T-Student es una distribución utilizada cuando el tamaño de la muestra es pequeño y se desconoce la desviación estándar de la población. Es similar a la distribución normal, pero tiene colas más pesadas. Se utiliza en pruebas de hipótesis y construcción de intervalos de confianza. Se denota como $X \sim t(n)$, donde n es el número de grados de libertad.

la distribución Chi-cuadrado es una distribución utilizada para analizar la varianza y la independencia en los modelos de regresión. Se denota como $X \sim \chi^2(k)$, donde k es el número de grados de libertad.

La distribución F-Snedecor, o simplemente F, es una distribución utilizada en el análisis de varianza y en pruebas de hipótesis sobre la igualdad de varianzas. Se denota como $X \sim F(d_1, d_2)$, donde d_1 y d_2 son los grados de libertad.

Estas son solo algunas de las distribuciones de probabilidad utilizadas en la regresión estadística. Cada una tiene propiedades específicas que se ajustan a diferentes escenarios y problemas. Al comprender y utilizar adecuadamente estas distribuciones, se puede realizar un análisis más preciso y robusto en el contexto de la regresión.

La función de probabilidad acumulada representa la probabilidad de que una variable aleatoria tome un valor menor o igual a un valor dado.

La función de probabilidad acumulada se denota por $F(x)$ y se calcula como la suma acumulada de las probabilidades de todos los valores menores o iguales a x. En el caso de una distribución

continua, la probabilidad acumulada se puede calcular como el área bajo la curva de densidad de probabilidad hasta el valor x. Es decir:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du$$

La probabilidad acumulada es una herramienta importante para la regresión en estadística porque permite calcular la probabilidad de que una variable aleatoria tome un valor dentro de un intervalo específico. Esta información es útil para tomar decisiones basadas en datos y para evaluar la calidad de los modelos de regresión ya que es necesaria para calcular intervalos de confianza y realizar pruebas de hipótesis en los modelos de regresión.

6.1.3. Análisis estadístico en la actualidad

En la actualidad, la ciencia de datos ha permitido desarrollar técnicas que han facilitado las tareas necesarias para la predicción y el resumen de grandes bases de datos, favoreciendo además, el adelanto de aplicaciones para su manipulación y visualización.

La implementación y la integración de nuevas herramientas computacionales para la realización de análisis estadístico sobre Big Data, así como la creación de nuevas aplicaciones basadas en métodos algorítmicos para el procesamiento, depuración y gestión de grandes volúmenes de información, han permitido el uso de submuestras de mayor tamaño para el desarrollo de predicciones tanto en el campo de aprendizaje automático como en el de la econometría.

Tanto los métodos de aprendizaje automático (*ML*, por su nombre en inglés Machine Learning) como los métodos de econometría se emplean para abordar problemas de clasificación y regresión para la toma de decisiones apoyándose en el análisis estadístico, sin embargo, estudios recientes indican que los modelos de aprendizaje automático en términos de pronóstico son más efectivos que los métodos de econometría tradicionales (Charpentier et al., 2019). Esto se observa particularmente en la obtención de resultados puntuales en predicciones a corto plazo, gracias a su alto rendimiento para el desarrollo de pronósticos con submuestras homogéneas y su capacidad para capturar no linealidades e irregularidades en los datos (Liu & Xie, 2019).

El aprendizaje automático brinda una serie de técnicas computacionales orientadas a favorecer el tratamiento de la información presente en bases de datos que en la actualidad, presentan cifras elevadas con respecto al número de observaciones y variables implicadas, permitiendo así la realización de pronósticos más precisos y disminuyendo la variación en el modelo computacional, surgiendo como una alternativa a los métodos de predicción tradicionales, los cuales presentan limitaciones a la hora de trabajar con grandes volúmenes de información.

En el caso de bases de datos grandes o con múltiples variables correlacionadas, la estimación mediante técnicas de econometría tradicional puede resultar insuficiente para la obtención de un

modelo óptimo. No obstante, las técnicas de econometría son más apropiadas para la identificación de tendencias a largo plazo, ya que se basan en la inferencia causal para la estimación de parámetros en un modelo predictivo (Mullainathan & Spiess, 2017)).

El uso frecuente de técnicas de aprendizaje automático en la predicción de fenómenos, sin la definición de un marco teórico adecuado, a menudo resulta en modelos que carecen de un poder explicativo suficientemente alto. Las hipótesis formuladas en torno al comportamiento de los modelos computacionales pueden ser difícilmente comprobadas, lo que dificulta la identificación de relaciones espurias.

Una relación espuria se produce cuando dos variables presentan una correlación aparente sin un transfondo de causalidad entre ellas. Este es un fenómeno común en modelos de series de tiempo dado que, sobre todo en el contexto económico, es normal que variables que no tienen una influencia directa entre sí, presenten una tendencia común en el tiempo debido a fenómenos que pueden o no estar relacionados.

Un ejemplo de relación espuria puede ser el incremento en la temperatura de la superficie terrestre producida por el calentamiento global, y el incremento en la producción mundial, debido a que, aunque las actividades de ciertas industrias pueden generar contaminación atmosférica, no existe una relación directa entre ambas variables y el calentamiento global puede depender en mayor medida de otros factores.

En contraste, un amplio espectro de las aplicaciones relativas a la econometría implementa modelos basados en procesos para la identificación y el análisis de relaciones significativas en aras de la explicación del comportamiento de una variable determinada a partir de ciertas variables regresoras, cuyo impacto relativo es medido a través de la formulación y contrastación de hipótesis mediante el método de inferencia causal. “Econometristas han desarrollado varias herramientas para la inferencia causal como: variables instrumentales, regresión discontinua, de diferencias en diferencias y diversas formas de experimentos naturales y diseñados.” (Angrist & Krueger, 2001).

Gracias al enfoque científico para la formulación de modelos econométricos, es una herramienta apropiada para medir el impacto de decisiones en el ámbito de la formulación de políticas públicas o estrategias empresariales, ya que proporciona conocimiento sobre el porqué de un resultado y permite evaluar la efectividad de una decisión desde ——

Pese a que se han desarrollado algunos modelos que integran la metodología de la econometría con modelos predictivos computacionales en aras de la comprensión de los fenómenos observados y la obtención de mejores resultados, no se han formalizado métodos estandarizados para incorporar las técnicas de aprendizaje automático al análisis econométrico, además, dado que la promoción e implementación de estos métodos es algo relativamente reciente, no se ha estandarizado una terminología formal y sus denominaciones se basan en descripciones que indican su aplicación general (Varian, 2014).

A continuación, se hará mención de las técnicas más convencionales utilizadas comúnmente para el desarrollo de pronósticos en las dos disciplinas.

6.2. Análisis de regresión en econometría

El análisis de regresión constituye un conjunto de métodos utilizados para la estimación de las relaciones entre una serie de variables explicativas y una variable explicada, logrando así, acercarse a un parámetro poblacional a partir de la observación de datos adquiridos por medio de experimentos naturales (Gujarati et al., 2010).

El análisis de regresión resulta conveniente para el esclarecimiento de las conexiones en la construcción de modelos resumidos en base al supuesto de causalidad unidireccional, no obstante, en vista de la dificultad existente para la realización de experimentos controlados, el proceso de observación puede estar sujeto al sesgo de variables omitidas entre otros problemas.

En la realización de análisis estadísticos en el campo de la econometría se desarrollan trabajos que pretenden predecir, resumir, estimar y poner a prueba hipótesis planteadas. (Varian, 2014) señala que la herramienta más común con la que cuenta la econometría para el desarrollo de modelos resumidos es el análisis de regresión lineal, mientras que las aplicaciones computacionales soportadas por técnicas de aprendizaje automático ofrecen un conjunto de herramientas que permiten estimar modelos resumidos sobre relaciones no lineales en los datos.

6.2.1. Modelos econométricos

Los métodos de regresión utilizados en la formulación de modelos económicos comprenden tres paradigmas de modelación paramétrica definidos a partir de la forma funcional del mismo que corresponden a:

- El modelo de regresión lineal simple, el cual presenta una relación lineal entre una variable aleatoria dependiente y una única variable explicativa, la cual, se asume que es determinística en el enfoque clásico de la econometría y aleatoria en el enfoque neoclásico. Se adiciona un término de error para representar la desviación del valor observado con respecto al valor esperado, causada por factores omitidos en la modelización.

- El modelo de regresión lineal múltiple, el cual generaliza el modelo simple para un conjunto de variables explicativas asociadas a parámetros linealmente relacionados.

- Por último, existen modelos que pueden ser lineales o no lineales. En los lineales los coeficientes de regresión del modelo se asocian en relaciones lineales, en los que sus exponentes

son uno y no están multiplicados o divididos entre sí independientemente de la forma funcional de las variables.

En el libro de (Rao & Toutenburg, 1995) sobre Modelos Lineales (1995), se explican algunos de estos paradigmas de regresión.

Los modelos de ecuaciones simultáneas representan una alternativa al método de regresión lineal, ya que permiten capturar el efecto multiplicador a través de la inclusión de ecuaciones que correlacionan las variables dependientes e independientes. Esto posibilita el análisis de fenómenos que involucran causalidad bidireccional, es decir, la interacción recíproca entre dos o más variables. De esta manera, es posible considerar cómo la variación de una variable influye sobre otra y viceversa, lo que permite una comprensión más completa de los procesos que se estudian. Estos modelos son de gran utilidad en disciplinas como la economía, la sociología y la psicología, entre otras.

Adicionalmente, la clasificación de un modelo puede depender de la naturaleza temporal y espacial de los datos del estudio. Por ejemplo, un modelo de corte transversal se estima utilizando una muestra de observaciones que capturan el estado de n individuos de la misma población en un momento determinado, o donde se asume que el tiempo no genera variación en los datos.

Los modelos de regresión que utilizan datos de corte transversal son particularmente efectivos en la predicción a corto plazo y es útil cuando se desea comparar el estado actual de las explicativas entre individuos debido a que se espera que factores no observables no hayan generado un impacto significativo sobre el estado de los individuos y la variable que se desea explicar dependa únicamente del estado actual de las explicativas, haciendo que los datos sean comparables. Sin embargo, este tipo de estudios puede presentar problemas de heterocedasticidad debido a una mala especificación o a la naturaleza misma de las variables.

Por otro lado, los modelos de regresión que utilizan datos de series de tiempo son útiles cuando se desea explicar una característica en función de un conjunto de variables dadas para una única unidad de observación cuyo estado cambia a lo largo del tiempo, y se cuenta con t observaciones del mismo en un periodo determinado y con una frecuencia constante. Este tipo de estudio comúnmente presenta problemas de autocorrelación en los datos, lo que puede permitir la formulación de modelos dinámicos en donde el estado de una entidad en el tiempo t dependa de su estado anterior con cierto margen de rezago.

Finalmente, el análisis de modelos de regresión que utilizan datos panel integra ambos tipos de datos. Se emplea cuando se desea estimar un modelo que explique una variable para n individuos de la misma naturaleza observados durante t periodos. Es natural la presencia de heterocedasticidad y autocorrelación en los datos debido a su estructura multidimensional. Aunque los datos de panel suministran mucha información para el desarrollo de un modelo, su modelamiento suele ser más complejo que el de corte transversal o series de tiempo.

Tal y como se ha indicado, los métodos de regresión lineal convencionales que utiliza la econometría para la estimación de modelos de predicción se basan en un conjunto de suposiciones estadísticas sobre los datos, como la normalidad, la linealidad, la homocedasticidad y la independencia de los errores. Estas suposiciones deben verificarse mediante pruebas estadísticas para garantizar que la estimación del modelo de regresión sea confiable.

En aras de poder plantear hipótesis sobre las relaciones existentes en un sistema de interés, se debe obtener un modelo que permita explicar con cierto grado de bondad dichas relaciones, reduciendo el sesgo producido por problemas en el muestreo o la especificación del modelo.

En estadística, el sesgo se define como la diferencia entre el valor esperado del estimador de un parámetro poblacional y el valor real de dicho parámetro. Debido a la incapacidad de los investigadores para obtener todos los datos de la población que se quiere estudiar, se obtiene un subconjunto aleatorio (muestra) de esta población aspirando que este sea estadísticamente representativo y resulte ser útil para una estimación confiable de las condiciones de la población.

Se dice que un estimador es consistente si el sesgo tiende a reducirse conforme aumenta el tamaño de la muestra con la que se realiza la estimación. Este concepto está estrechamente relacionado con la ley de los grandes números, en la cual, se postula que conforme crece el número de observaciones obtenidos mediante un experimento aleatorio (Conforme el tamaño de la muestra aumenta) el promedio de los resultados obtenidos converge al valor esperado del parámetro poblacional.

No obstante, existen ciertas limitaciones en esta propiedad, por ejemplo, hay casos en los que debido a que la distribución de la variable que se estudia presenta colas pesadas (Es decir, una alta frecuencia de datos alejados a la media), puede que bajo un supuesto de normalidad, la estimación del modelo no sea consistente.

Si el estimador es consistente, significa que conforme el tamaño de la muestra aumenta, la función de ECM tiende a minimizarse cada vez más cerca del parámetro poblacional, lo que permitiría obtener una mejor estimación mediante métodos de regresión convencionales.

A la hora de modelar, nos encontramos con cuatro expresiones diferentes que capturan la variación de la variable que se quiere explicar:

- El modelo de regresión poblacional formulado para describir la forma en la que en la realidad se comportan las variables del mismo:

$$Y_i = f(x) = \beta_0 + \sum_{j=1}^{P} X_j \beta_j + \epsilon_i \tag{6.5}$$

Donde P=k-1

- Al suprimir el término de error en dicho modelo, obtenemos el hiperplano de regresión poblacional, el cual, captura la variación promedio entre la variable y y las variables x y permite obtener el valor esperado de la variable de respuesta dada una configuración en las regresoras:

$$\mathrm{E}[y \mid x_1, x_2, x_3 ...] = f(x) = \beta_0 + \sum_{j=1}^{P} X_j \beta_j \tag{6.6}$$

- Al realizar la estimación de los coeficiente de regresión del modelo poblacional, se obtiene el modelo de regresión muestral que, al adicionar los residuos de la regresión, captura los valores de la muestra. Dado que estos valores provienen de observaciones de la población que se desea estudiar, son los valores reales de y explicados mediante el modelo muestral.

$$Y_i = f(x) = \hat{\beta}_0 + \sum_{j=1}^{P} X_j \hat{\beta}_j + \hat{\epsilon}_i \tag{6.7}$$

Nótese que, dado que la muestra a partir de la cual se estima el modelo es aleatoria, la naturaleza de los coeficientes estimados de regresión del vector $\hat{\beta}$ también será aleatoria, ya que tanto el valor del intercepto como de las pendientes del modelo variarán dependiendo de los datos a partir de los cuales se obtenga la estimación, haciendo que $\hat{\beta}$ sea un vector aleatorio.

- Finalmente, para realizar interpretaciones sobre las relaciones entre las variables del modelo, conviene considerar los coeficientes obtenidos mediante MCO del hiperplano de regresión estimado con los datos de la muestra:

$$\hat{y} = \widehat{\mathrm{E}[y \mid x]} = f(x) = \hat{\beta}_0 + \sum_{j=1}^{P} X_j \hat{\beta}_j \tag{6.8}$$

6.2.2. Estimación mediante mínimos cuadrados

El método de mínimos cuadrados (MC) es ampliamente utilizado para estimar los coeficientes del modelo de regresión lineal. Propuesto por Gauss, este método busca encontrar las estimaciones de los coeficientes del modelo que minimizan la suma de los cuadrados de los errores, permitiendo obtener la recta que mejor se ajuste a una serie de puntos. En el caso de modelos con múltiples variables explicativas, se obtiene un hiperplano de regresión que representa la combinación lineal de las variables explicativas ponderadas por los coeficientes estimados.

La estimación mediante mínimos cuadrados permite determinar los coeficientes del modelo que describen la relación entre la variable de respuesta y las variables explicativas. Adicionalmente,

debido a la naturaleza aleatoria de la variable dependiente, el modelo incluye un término de error, que captura la variabilidad aleatoria no explicada por las variables explicativas y refleja la diferencia entre el valor esperado y el valor observado de una observación.

El método de mínimos cuadrados ordinarios (MCO) busca, a través de algún método de optimización matemática, minimizar la función de Error Cuadrático Medio (ECM) en la estimación de modelos de regresión. El ECM es una medida de riesgo que relaciona el valor esperado del cuadrado de los errores con los estimadores del modelo. Cuando el estimador es insesgado, el valor de la función ECM coincide con el valor de la varianza del estimador, lo que implica que al minimizar el ECM se obtiene el estimador con la menor varianza.

El ECM se define como la media de los cuadrados de los residuos entre los valores observados (y_i) y los valores predichos ($\hat{y}_i$) por el modelo. Matemáticamente, se expresa de la siguiente manera:

$$ECM = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Al obtener el valor esperado del cuadrado de los errores en un modelo de regresión, se realiza una operación en la cual se multiplica la suma de los errores al cuadrado por el inverso multiplicativo del tamaño de la muestra. Esta multiplicación por una constante no afecta los valores críticos de la función y, por lo tanto, es común que los métodos de optimización utilizados para minimizar el ECM no consideren este factor. En cambio, se busca minimizar la suma de cuadrados del error (SCE) del modelo, que es una función cuadrática de los parámetros estimados.

Además, el error cuadrático medio puede expresarse como la suma de cuadrados del error (SCE), que es una función cuadrática de los parámetros del modelo. Dado que el error al cuadrado siempre es no negativo, la SCE es estrictamente positiva, lo que implica que existe un mínimo absoluto en la función SCE:

$$SCE(\beta) = \sum_{i=1}^{N}(y_i - f(x_i))^2$$

$$= \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{N} x_i j\beta_j)^2$$

Donde β es el vector de coeficientes que se busca estimar, β_0 es el intercepto y x_{ij} son los valores de las variables explicativas para la i-ésima observación y j-ésima variable.

La relación entre la función de ECM y la función de SCE radica en el hecho de que la función de SCE es una función cuadrática de los parámetros del modelo y, por lo tanto, siempre es positiva para todos los puntos. Al minimizar la SCE, también se minimiza el ECM, lo que garantiza que se obtenga el estimador con la menor varianza.

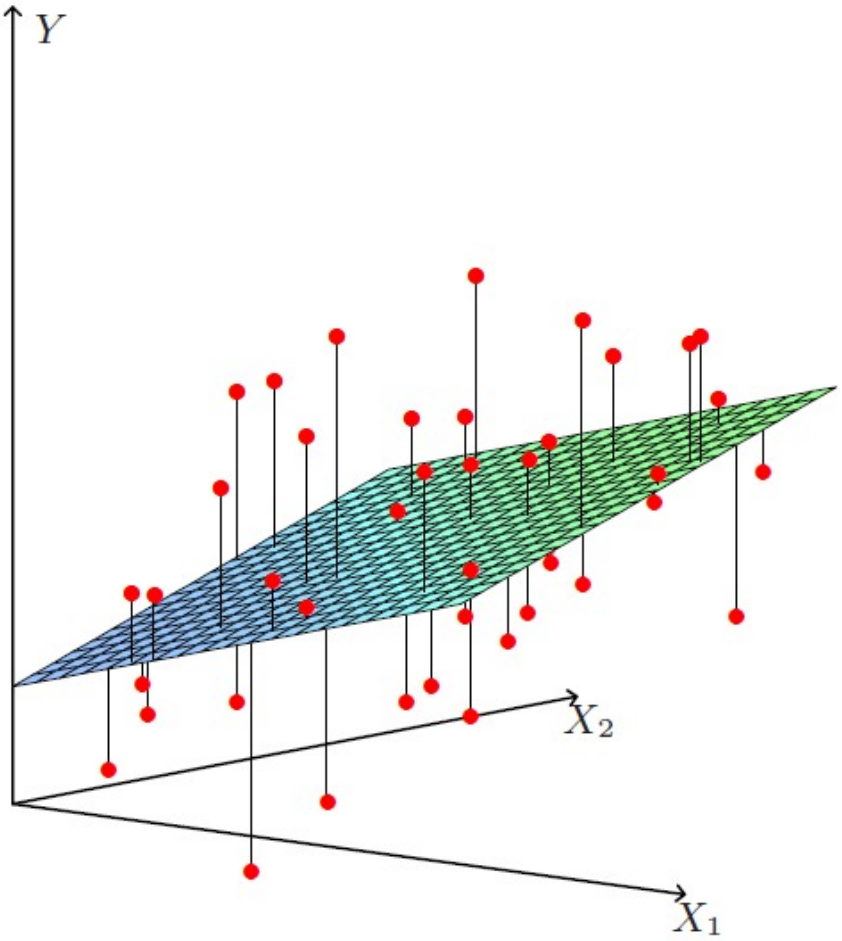

Figura 6.1: Regresión lineal en $\mathbb{R}^2$
(Hastie et al., 2009)

El Teorema de Gauss-Markov establece que el estimador obtenido mediante el método de MCO tiene la varianza más baja entre los estimadores lineales insesgados si se cumplen los supuestos de homocedasticidad, no autocorrelación y valor esperado de los errores igual a cero (Gujarati et al., 2010).

Estos supuestos no requieren que los errores sigan una distribución normal ni que sean independientes e idénticamente distribuidos. Sin embargo, es fundamental que el estimador sea insesgado, ya que existen estimaciones sesgadas con una varianza menor. Asimismo, es importante tener en cuenta que un estimador que minimiza el sesgo puede no minimizar el ECM, tal es el caso de estimadores como el de la regresión ridge o LASSO.

6.2.3. Estimación mediante Máxima Verosimilitud

El método de estimación por Máxima Verosimilitud (MV) se utiliza también para obtener estimaciones de los parámetros de un modelo estadístico. Este método busca maximizar la función de verosimilitud de las observaciones, lo que implica encontrar los valores de los parámetros que hacen que los datos observados sean más probables.

La función de verosimilitud representa la función de probabilidad conjunta desconocida de la muestra aleatoria en función de los parámetros del modelo. Matemáticamente, si tenemos una muestra de observaciones independientes e idénticamente distribuidas (i.i.d.) $y_1, y_2, \ldots, y_n$ con una función de densidad de probabilidad o función de masa de probabilidad $f(y_i|\theta)$ que depende de un vector de parámetros desconocidos θ, la función de verosimilitud se expresa como:

$$L(\theta) = f(y_1|\theta) \cdot f(y_2|\theta) \cdot \ldots \cdot f(y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

El estimador de Máxima Verosimilitud $\hat{\theta}_{\text{MV}}$ es aquel valor del vector de parámetros θ que maximiza la función de verosimilitud. En términos matemáticos, esto se puede expresar como:

$$\hat{\theta}_{\text{MV}} = \arg\max_{\theta} L(\theta)$$

El estimador de MV tiene propiedades deseables, como la consistencia y la eficiencia. Sin embargo, el estimador de MV presenta un sesgo hacia abajo en la estimación de la varianza, lo que significa que el valor estimado de la varianza puede ser menor que el verdadero valor de la varianza.

Si asumimos que las observaciones siguen una distribución normal y queremos estimar los parámetros de una regresión lineal, la función de verosimilitud se puede expresar de la siguiente manera:

Sea y el vector de observaciones, X la matriz de variables explicativas y β el vector de coeficientes que queremos estimar. Supongamos que los errores siguen una distribución normal con media cero y varianza σ^2. Entonces, la función de verosimilitud $L(\beta, \sigma^2)$ se define como:

$$L(\beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - X_i\beta)^2}{2\sigma^2}\right)$$

Donde X_i es la i-ésima fila de la matriz de variables explicativas X. Tomando el logaritmo de la función de verosimilitud, obtenemos la verosimilitud logarítmica ($l(\beta, \sigma^2)$):

$$l(\beta, \sigma^2) = \sum_{i=1}^{n} \left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - X_i\beta)^2}{2\sigma^2}\right)$$

Para estimar los parámetros β y σ^2 por máxima verosimilitud, maximizamos la verosimilitud logarítmica respecto a estos parámetros. Esto se puede hacer de manera iterativa utilizando técnicas de optimización numérica, como el método de Newton-Raphson o el descenso de gradiente.

La estimación de Máxima Verosimilitud en el contexto de la regresión lineal nos permite obtener los valores de los coeficientes $\hat{\beta}$ que maximizan la probabilidad conjunta de observar los datos de la muestra, asumiendo que siguen una distribución normal. Esta estimación considera la relación lineal entre las variables explicativas y la variable respuesta, y busca encontrar los coeficientes que mejor ajustan los datos en términos de verosimilitud.

En este caso, la estimación de los coeficientes mediante el método de MV es equivalente a la obtenida mediante el método de Mínimos Cuadrados Ordinarios (MCO). Además, a medida que aumenta el tamaño de la muestra, el sesgo en la estimación de la varianza disminuye, ya que el estimador de MV de la varianza se vuelve asintóticamente igual al estimador insesgado de la varianza obtenido mediante MCO. Esto implica que el estimador de MV de la varianza es consistente a medida que aumenta el tamaño de la muestra. (Gujarati et al., 2010).

6.3. Ajuste de un modelo econométrico

Para evaluar la bondad relativa de un modelo econométrico a la hora de explicar la variable dependiente, se cuantifica el grado de ajuste del modelo a los datos observados, es decir, su capacidad para capturar la variabilidad de la variable explicada y reducir el error en aras de la obtención de un mejor pronóstico.

La suma de cuadrados esperada, misma antes espresada como suma de los cuadrados de error SCE, también se expresa comúnmente como SSE (del inglés, Sum of Squares for Errors), y como se definió, es una medida de cuánto se desvían los datos observados de la media ajustada por un modelo, es decir, la suma de los errores al cuadrado o las distancias cuadráticas entre los valores reales y predichos por la recta (o hiperplano) de regresión que se define como la esperanza condicional de y dada una configuración en las variables independientes. Recordemos su expresión matemática como:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Donde n es el número de observaciones, y_i es el valor observado de la variable dependiente en la i-ésima observación, y $\hat{y}_i$ es el valor ajustado por el modelo.

La suma de cuadrados residual, SSR (del inglés, Sum of Squares for Regression), es una medida de cuánto se puede explicar de la variabilidad de los datos a través del modelo. Matemáticamente se puede expresar como:

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Donde $\bar{y}$ es la media de los valores observados de la variable dependiente.

La suma total de cuadrados, SST (del inglés, Total Sum of Squares), es la variabilidad total de los datos observados. Matemáticamente la SST se puede expresar como:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

El coeficiente de correlación de Pearson, r, es una medida del grado de relación lineal entre las variables dependientes y la independiente de un modelo. En el caso de un modelo de regresión lineal simple, se define como la covarianza entre las dos variables dividida por el producto de sus desviaciones estándar y se puede expresar matemáticamente como:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Donde x_i e y_i son los valores observados de las dos variables en la i-ésima observación, y $\bar{x}$ y $\bar{y}$ son sus medias.

Cuando se el coeficiente de correlación de Pearson se eleva al cuadrado, se obtiene el coeficiente de determinación, R^2, el cual, es una medida de la proporción de la variabilidad total de la variable dependiente que se puede explicar por el modelo y se puede expresar como una proporción entre las medidas mencionadas:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Debido a que la SSR y la SSE son funciones cuadráticas, siempre serán positivas y ambos valores serán menor que la SST, por lo que el valor de R^2 siempre está entre 0 y 1, y cuanto más cercano a 1, mayor será la proporción de la variabilidad de la variable dependiente explicada por el modelo.

6.3.1. Problemas en los modelos

Como se mencionó anteriormente, dependiendo de la naturaleza de las variables que se desean incluir en el estudio, un modelo puede presentar problemas de heterocedasticidad si se trata con datos de corte transversal, o de autocorrelación si se trata con datos de series de tiempo.

La heterocedasticidad se define como la heterogeneidad en la varianza de los errores. Si la varianza del error en el modelo es constante para los valores observados de la variable regresada

en función de cierta conjunto de variables regresoras , se dice que el modelo es homocedástico, en caso contrario, este será heterocedástico. (Wooldridge, 2016).

Este problema surge frecuentemente cuando: existen valores atípicos, asimetría en las regresoras, procesos de enseñanza y aprendizaje, o cuando se pretende ajustar un conjunto de datos a un modelo con una forma funcional que no corresponde en realidad con la forma en la que se relacionan, y puede ser común al estimar un modelo lineal sobre las observaciones sin la definición de un marco teórico apropiado o una comprensión adecuada sobre las relaciones que presentan las variables en el modelo.

En el caso de los estudios de corte transversal, es natural considerar que el error en una variable aleatoria puede depender del valor de una relacionada.

Normalmente esto se produce porque las mediciones realizadas sobre ciertas variables pueden ser menos exactas en ciertos rangos de otras como el tiempo o distancia, o debido a que el factor que se espera que explique la variable dependiente deja de ser determinante en ciertos rangos, por lo que aunque la mala especificación de un modelo puede producir problemas de heterocedasticidad, también puede presentarse debido a la naturaleza misma de las variables.

La heterocedasticidad puede conducir a una estimación insesgada y consiste pero ineficiente de los coeficientes de regresión del modelo, adicionalmente, puede hacer que se sobreestime la bondad del ajuste al obtener un coeficiente de correlación alto.

La autocorrelación es una característica que se presenta con modelos de serie de tiempo y que se refiere a una semejanza o diferencia significativa en los valores cercanos en el tiempo o el espacio con respecto a otros valores más alejados, y se presenta fundamentalmente cuando: Existe inercia en las variables debido al fenómeno de la telaraña, una mala forma funcional, la omisión de variables importantes y no estacionalidad en las variables que participan en el modelo. Los procesos de raíces unitarias, los procesos estacionarios de tendencia, los procesos autorregresivos (AR) y los procesos de media móvil (MA) constituyen algunas de las formas más comunes de autocorrelación.

La presencia de autocorrelación en un modelo genera un estimador de los parámetros insesgado y consistente pero ineficiente.

Cuando una predictora en un modelo de regresión lineal múltiple presenta una relación lineal con las demás con un nivel significativo de precisión se dice que el modelo presenta alta multicolinealidad. Aunque la presencia de alta multicolinealidad no altera las propiedades del estimador, bajo el cumplimiento de los supuestos del modelo, genera implicaciones de orden práctico, dado que no se rechazan las hipótesis de no significancia con frecuencia.

Existen diferentes formas de detectar alta multicolinealidad: Cuando se obtiene un coeficiente de determinación alto, un F significativo pero pocas razones t significativas en el modelo y un VIF (Factor Inflacionario de la Varianza) mayor a 10. Este rasgo puede causar que los coeficientes estimados mediante MCO cambien abruptamente ante variaciones en los datos (Wooldridge,

2016).

Aunque la presencia de alta multicolinealidad no reduce la capacidad predictiva de un modelo o su confiabilidad para la predicción dentro del conjunto de datos de la muestra, tiene implicaciones teóricas importantes al alterar el impacto de los predictores individuales en el modelo, pudiendo incluso cambiar los signos de las pendientes.

6.3.2. Alternativas a mínimos cuadrados ordinarios

Los métodos de regresión robustos son alternativas a la regresión lineal tradicional que buscan superar algunas de sus limitaciones. Uno de los métodos más conocidos es el de desviaciones mínimas absolutas. A diferencia del método de mínimos cuadrados ordinarios (MCO), el método de desviaciones mínimas absolutas minimiza las distancias absolutas entre los puntos y el hiperplano de la regresión sin potenciar los errores, como se muestra en la ecuación:

$$DMA = \frac{1}{N}\sum_{i=1}^{N} |y_i - f(x_i)| \tag{6.9}$$

La principal ventaja del método de desviaciones mínimas absolutas es que es más robusto que el método de MCO frente a valores atípicos, ya que los errores no crecen de forma cuadrática. Esto significa que los valores atípicos tienen una menor influencia en la estimación final. Sin embargo, es importante tener en cuenta que este método no garantiza una solución única debido a que la distancia puede minimizarse al pivotar la recta entre puntos simétricos.

Además del método de desviaciones mínimas absolutas, existen otros métodos de regresión robustos que asumen distribuciones de los errores distintas a la normal, como aquellas que tienen más peso en las colas, disminuyendo la influencia de los errores atípicos en el estimador. Estos métodos también pueden ayudar a superar las limitaciones de la regresión lineal tradicional, especialmente cuando el modelo presenta heterocedasticidad que no puede ser corregida mediante un cambio en su especificación.

Existen otros métodos de regresión robustos que buscan superar las limitaciones de la regresión lineal tradicional y reducir la influencia de valores atípicos en la estimación final. Estos incluyen el método de Huber y el método del estimador M.

El método de Huber busca minimizar una función que combina la suma de los errores cuadráticos cuando la magnitud de los errores es pequeña y la suma de los errores absolutos cuando la magnitud de los errores es grande. Esta función es conocida como función de Huber y está definida como:

$$\psi(z) = \begin{cases} \frac{1}{2}z^2, & \text{si } |z| \leq k \\ k(|z| - \frac{1}{2}k), & \text{si } |z| > k \end{cases} \tag{6.10}$$

donde k es una constante que define el punto de transición entre la función cuadrática y la función absoluta. De esta manera, el método de Huber tiene la capacidad de ajustar adecuadamente a los datos que siguen una distribución normal, pero también es capaz de tolerar ciertos niveles de datos atípicos.

Por otro lado, el método del estimador M es una generalización presentada por Huber del método de máxima verosimilitud que permite al estimador no tener que suponer una distribución específica de los errores, lo que puede resultar ventajoso cuando no se conoce exactamente la distribución de los errores o cuando la distribución no se ajusta bien a la normalidad. El método del estimador M busca minimizar una función llamada función de influencia, que mide la contribución de cada punto de datos a la estimación final. Esta función se define como la derivada de la función de los parámetros del modelo en relación con el punto de datos correspondiente.

Un tercer método de regresión robusto es el de regresión por cuantiles, el cual, se basa en estimar los parámetros de la regresión para distintos percentiles de la distribución de los errores, en lugar de estimar el modelo para toda la distribución. Esto permite obtener estimaciones robustas para los percentiles extremos, los cuales son más sensibles a los valores atípicos.

6.3.3. Elección binaria con mínimos cuadrados

Para afrontar problemas de clasificación binaria, el enfoque de mínimos cuadrados puede ser utilizado para pronosticar una variable con respuesta 0/1, separando en el espacio vectorial las observaciones en dos grupos en función de su valor para dicha variable, esto se logra mediante una regresión en la que se buscan los valores de y iguales de probabilidad a 0.5 para cada configuración de los regresores. Este método proporciona un ajuste para el hiperplano que separa con mayor bondad los grupos de observaciones, pero puede no ser muy eficiente a la hora de agrupar conjuntos no convexos.

Otro enfoque clásico para el planteamiento de un modelo de elección binaria en econometría es el del modelo de probabilidad lineal, en el cuál, la variable dependiente es la probabilidad de que un ensayo de Bernoulli (un experimento en donde el resultado observado pueda tomar uno de dos valores) sea exitoso.

Al ser estimada la recta mediante mínimos cuadrados, el ajuste es ineficiente, y puede ser mejorado mediante un método iterativo de mínimos cuadrados ponderados en el que se realizan iteraciones de la estimación del modelo para estimar la varianza condicional que variaría entre las observaciones, de forma análoga a la estimación por máxima verosimilitud.

Una limitación del modelo de probabilidad lineal es que las predicciones pueden estar por fuera del intervalo unitario de (0, 1) en el cual operan las probabilidades. Una alternativa a estos modelos que corregiría este problema es el del modelo de regresión logit (o regresión logística) y regresión probit.

6.3.4. Modelos de regresión logit y probit

El modelo logit, o modelo de regresión logística, representa las probabilidades logarítmicas de que un evento aleatorio determinado ocurra como una combinación lineal de los los predictores y los coeficientes del modelo. Dado que la función de probabilidad acumulada del evento aleatorio es logística, el rango de la función estará acotado en el intervalo unitario (0, 1), lo que permitiría una mejor modelación de las probabilidades que el proporcionado por el enfoque de probabilidad lineal.

El modelo probit, o modelo de probabilidad unitaria, tiene un enfoque similar, pero al formularlo se asume que las variables siguen una función de distribución normal. Al igual que para el modelo logit, el propósito de este modelo es estimar la probabilidad de que un individuo con ciertas características dadas pertenezca a una categoría específica.

Estos modelos son estimados mediante el método de máxima verosimilitud y tanto el modelo de probabilidad lineal, como el modelo logit, como el probit, se interceptan en la configuración de los regresores que generan una probabilidad igual a 0.5 en la variable de respuesta.

6.3.5. Desigualdad de Theil

La desigualdad de Theil es una medida utilizada para evaluar la precisión de los modelos de pronóstico en comparación con los valores reales observados. Esta herramienta permite comparar diferentes modelos, incluso si tienen distintos niveles de complejidad o utilizan diferentes métodos de pronóstico.

La desigualdad de Theil es particularmente útil para comparar la precisión de los pronósticos hechos con modelos econométricos en relación con la precisión de otros tipos de pronósticos, como el juicio de expertos o modelos simples de series de tiempo. Esto se debe a que la desigualdad de Theil compara los valores pronosticados de una variable dependiente con los valores reales observados de la misma variable.

La desigualdad de Theil se expresa como una relación entre el error cuadrático medio del pronóstico y el error cuadrático medio de las observaciones reales. Esta medida se utiliza para evaluar la precisión relativa de diferentes modelos de pronóstico. La ecuación de la desigualdad de Theil se define como:

$$U = \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(Y_t^s - Y_t^a)^2}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(Y_t^s)^2} + \sqrt{\frac{1}{T}\sum_{t=1}^{T}(Y_t^a)^2}} \tag{6.11}$$

Donde Y_t^s es el valor pronosticado en el período t, y Y_t^a es el valor real observado en el mismo período. T es el número total de períodos en el conjunto de datos.

La ecuación se divide en dos partes. La primera parte, en el numerador

$$\sqrt{\frac{1}{T}\sum_{t=1}^{T}(Y_t^s - Y_t^a)^2}$$

es el error estándar de predicción y mide el error medio del pronóstico, es decir, la diferencia entre los valores pronosticados y los valores reales observados. La segunda parte, en el denominador

$$\sqrt{\frac{1}{T}\sum_{t=1}^{T}(Y_t^s)^2} + \sqrt{\frac{1}{T}\sum_{t=1}^{T}(Y_t^a)^2}$$

mide el nivel de las observaciones reales y pronosticadas. Conforme se minimiza el error cuadrático medio de pronóstico el numerador tiende a 0, haciendo que el pronóstico sea mejor.

El índice U adoptará valores entre el intervalo unitario (0, 1), acercándose a 0 conforme el error cuadrático medio de pronóstico disminuye y a 1 conforme aumente y la capacidad predictiva del modelo sea peor.

La desigualdad de Theil proporciona una forma de comparar diferentes modelos de pronóstico, ya que cuanto menor sea el valor de U, mayor será la precisión relativa del modelo en cuestión. Además, es una medida útil para evaluar la precisión de los modelos econométricos y otros tipos de pronósticos, y permite evaluar la precisión relativa de diferentes modelos de pronóstico en relación con las observaciones reales.

El coeficiente de desigualdad de Theil puede descomponer en proporciones de sesgo (U^S), varianza(U^V) y covarianza (U^C) así:

$$U^S = \frac{(\bar{Y^s} - \bar{Y^a})^2}{(\frac{1}{T})\sum(Y_t^s - Y_t^a)^2} \tag{6.12}$$

La proporción de sesgo mide la extensión que se desvían entre sí los valores promedio pronosti-

cados y los valores promedio reales, cualquiera que sea el valor de U se espera que U^M sea cercano a 0, valores cercanos a 1 indicarían la presencia de un sesgo sistémico.(Pindyck & Rubinfeld, 1998)

$$U^V = \frac{(\sigma_s - \sigma_a)^2}{(\frac{1}{T})\sum(Y_t^s - Y_t^a)^2} \tag{6.13}$$

La proporción de varianza muestra la capacidad que tiene el modelo para capturar el grado de variabilidad de la variable dependiente, si U^V es grande significaría que el valor real a fluctuado considerablemente mientras que el valor pronosticado a cambiado poco, o viceversa, advirtiendo que el modelo debe revisarse.

$$U^C = \frac{2(1-\rho)\sigma_s\sigma_a}{(\frac{1}{T})\sum(Y_t^s - Y_t^a)^2} \tag{6.14}$$

La proporción de covarianza mide el error restante después de que se han explicado las desviaciones de los valores promedio, en este punto como no se espera que los valores pronosticados tengan una correlación perfecta con los valores reales, hace que este componente sea el menos relevante.

Cabe resaltar que $U^S + U^V + U^C = 1$. y que los valores de $U > 0$. Adicionalmente la distribución de la desigualdad más ideal seria $U^S = U^V = 0$ y $U^C = 1$.

6.4. Análisis de regresión en aprendizaje automático

En comparación con las técnicas tradicionales de econometría, los métodos de pronóstico de aprendizaje automático tienen la capacidad de aprender patrones y relaciones directamente de los datos sin hacer suposiciones fuertes sobre la distribución de las variables o su forma funcional. Estos métodos tienen la capacidad de manejar relaciones complejas y no lineales entre los predictores y el resultado, y también, pueden capturar interacciones y efectos no aditivos que no son considerados a la hora de formular modelos lineales desde una perspectiva tradicional (Hastie et al., 2009).

En el campo del aprendizaje automático, los modelos analíticos se centran principalmente en tareas de predicción. Estos modelos pueden ser estimados mediante técnicas de aprendizaje supervisado y no supervisado.

El aprendizaje supervisado trabaja con muestras bajo instancias iniciales de relacionamiento de datos de entrenamiento, que constan normalmente de un vector de datos regresores de entrada, y un valor de salida para la definición inicial de parámetros en las relaciones del modelo computacional.

Como lo plantea (Rojas, 2020), los métodos de aprendizaje supervisado se fundamentan en enseñar al algoritmo como debe realizar su trabajo, definiendo qué variables serán utilizadas como

entradas en el modelo y cuál será regresada, basado en un conjunto de datos que giran en torno a una idea y cuya finalidad es la búsqueda de patrones para el análisis(Hurwitz & Kirsch, 2018).

Por otro lado, existen las tareas de aprendizaje no supervisado, que se diferencian principalmente por la falta de clasificación o etiquetado de los datos como predeterminados y de respuesta (Bishop & Nasrabadi, 2006).Aunque no se profundizará en las técnicas de aprendizaje no supervisado en este texto.

Las tareas de aprendizaje no supervisado, por su parte, se diferencian principalmente de las de aprendizaje supervisado por la no clasificación o etiquetado de los datos y a pesar de lo anterior sigue descubriendo patrones sin realizar un entrenamiento previo (Bishop & Nasrabadi, 2006).

Entre las técnicas más utilizadas por lo algoritmos en el campo del aprendizaje supervisado, se cuentan, k-NN, métodos de regresión lineal penalizada (regresión LASSO, ridge, y red elástica), los árboles de decisión, máquinas de soporte vectorial, redes neuronales y deep learning que permiten desarrollar modelos más eficaces a través del análisis de relaciones complejas.

Los métodos de aprendizaje automático se enfocan en el ajuste de un modelo flexible y complejo que puede adaptarse a los datos, en lugar de enfocarse en hacer inferencias estadísticas acerca de los coeficientes del modelo. Por lo tanto, los modelos de aprendizaje automático se evalúan en función de su desempeño predictivo en un conjunto de prueba retenido en lugar de la importancia estadística de sus coeficientes. (Hastie et al., 2009) señalan que las pruebas estadísticas no son necesarias para evaluar la validez de los modelos de aprendizaje automático.

Sin embargo, debido a la complejidad del modelo obtenido mediante algoritmos de aprendizaje automático, no se garantiza la insesgades del estimador del modelo. Este fenómeno puede resultar perjudicial en procesos de entrenamiento iterativos, afectando sistemáticamente al modelo y haciendo que su capacidad de predicción se vea reducida únicamente a datos conocidos.

Al igual que en econometría, algunas técnicas para afrontar problemas de clasificación y de regresión en aprendizaje automático trabajan con el método de mínimos cuadrados de regresión lineal para estimar hiperplanos que se utilizan al determinar relaciones entre variables cuantitativas o agrupar observaciones homogéneas en categorías, no obstante, otros métodos que buscan un mayor ajuste pueden trabajar con métodos de interpolación polinomial o regresiones no paramétricas, tales como la regresión kernel.

En los métodos de regresión no paramétricos el modelo no se ajusta sobre una forma funcional predeterminada (como una recta en el caso de la regresión lineal), sino que se construye con base en los valores de los datos de la muestra que se utiliza para la estimación.

Para evitar el sobreajuste, estos métodos trabajan mejor con una gran cantidad de datos, y no arrojan estimaciones precisas lejos de los rangos de la estimación, pero pueden ser especialmente eficientes en la predicción puntual debido a su alto desempeño en la minimización del error. Como es de esperarse, estos métodos suelen carecer de una base teórica fuerte más allá del proceso de

selección de variables.

Adicionalmente, los métodos de boosting y bootstrap son técnicas computacionales iterativas que mejoran el rendimiento y la estabilidad de los algoritmos de clasificación y regresión, evitando así, los problemas de sobreajuste de las observaciones y aumentando la precisión en el pronóstico del modelo.

El boosting permite reforzar los criterios de clasificación de observaciones mediante la definición de nuevos criterios en el proceso de aprendizaje asistido, implementando nuevas condiciones identificadas por la máquina en muestreos repetidos, y anexándolos a las originales gracias a la determinación de su nivel de interacción.

El método de agregación bootstrap, también conocido como bagging, al igual que el boosting trabaja realizando un proceso de muestreo repetido para la determinación de distintos parámetros en múltiples modelos construidos a partir de submuestras del conjunto inicial (o de arranque) de entrenamiento.

Este método permite obtener algoritmos más eficientes computacionalmente mediante un proceso de promediación o votación de los diferentes modelos desarrollados para la determinación de los parámetros que se utilizarán en el proceso de análisis y predicción (Varian, 2014).

El perceptrón es un algoritmo de aprendizaje supervisado que se utiliza para clasificar observaciones en categorías distintas. Su formulación matemática básica fue propuesta por Rosenblatt en 1958 y ha sido el precursor de las máquinas de soporte vectorial (SVM). Aunque el perceptrón es un modelo lineal, similar al modelo de mínimos cudrados para clasificación binaria visto anteriormente, su evolución posterior dio lugar a técnicas más avanzadas, como las SVM, que pueden manejar problemas de clasificación no lineales de manera más efectiva.

El perceptrón se fundamenta en la búsqueda de un hiperplano de separación que pueda distinguir entre las dos categorías de observaciones. Dado un conjunto de datos de entrada x y sus respectivas etiquetas de clase y, el objetivo del perceptrón es encontrar un vector de pesos W y un término de sesgo b de tal manera que la función de decisión lineal $W^T x + b$ clasifique correctamente las observaciones. Si $W^T x + b$ es mayor o igual a cero, se clasifica como una categoría, y si es menor que cero, se clasifica como la otra categoría.

La actualización de los pesos y el sesgo en el perceptrón se realiza mediante la regla de aprendizaje de Rosenblatt, que ajusta los parámetros en función de los errores de clasificación cometidos en cada iteración. La regla de aprendizaje actualiza los pesos y el sesgo de la siguiente manera:

$$W \leftarrow W + \eta \cdot y \cdot x \; b \leftarrow b + \eta \cdot y \tag{6.15}$$

donde η es la tasa de aprendizaje, y es la etiqueta de clase y x es el vector de características de

la observación. Estos ajustes se realizan iterativamente hasta que se alcanza la convergencia o se alcanza un número máximo de iteraciones.

A pesar de su simplicidad, el perceptrón tiene limitaciones en términos de su capacidad para manejar problemas de clasificación no lineales. No puede aprender fronteras de decisión complejas que no sean lineales. Para superar esta limitación, se desarrollaron las máquinas de soporte vectorial, técnica que permitió un avance extraordinario en el campo de la inteligencia artificial.

Las máquinas de soporte vectorial amplían el enfoque del perceptrón al utilizar hiperplanos de separación más sofisticados y al permitir clasificaciones no lineales a través del uso de kernels. En lugar de intentar ajustar un hiperplano directamente, las SVM buscan encontrar el hiperplano que maximice el margen entre las clases, es decir, el espacio vacío más grande posible entre las observaciones de cada clase. Esto se conoce como la formulación de margen máximo.

En el caso de Linear de SVM, la técnica consiste consiste en encontrar un hiperplano de separación que divida el espacio de observaciones en dos conjuntos de puntos lo más homogéneos posible (es decir, que contengan etiquetas idénticas). En la dimensión dos, el algoritmo consiste en determinar una línea que separa el espacio en dos zonas lo más homogéneas posible (Charpentier et al., 2019).

El hiperplano de separación de la SVM lineal tiene la forma $W^T x - b = 0$, donde W es un vector de pesos, x es el vector de características de entrada y b es el término de sesgo. La función objetivo para encontrar el hiperplano óptimo se expresa como:

$$[\frac{1}{n}\sum_{i=1}^{n} max(0, 1 - y_i(W^T x_i - b))] + \lambda||W||^2 \tag{6.16}$$

donde n es el número de observaciones, y_i es la etiqueta de la observación x_i, y λ es un parámetro de regularización que controla la complejidad del modelo. La primera parte de la función objetivo representa la pérdida de bisagra, que penaliza las observaciones mal clasificadas. La segunda parte es un término de regularización que busca minimizar la norma del vector de pesos W para evitar un sobreajuste.

Además del enfoque lineal, existen variantes no lineales de SVM, conocidas como SVM kernel. En estos métodos, se utiliza una función de transformación no lineal, también llamada kernel, para mapear el espacio de características original a uno de mayor multiplicidad geométrica donde los datos sean linealmente separables. Esto permite un mayor ajuste y flexibilidad en la separación de los grupos de observaciones.

El uso de kernels en las SVM permite trabajar con funciones no lineales en el espacio transformado. La formulación matemática general de SVM kernel se define como:

$$\min_{W,b} \left[\frac{1}{n} \sum_{i=1}^{n} \text{máx}(0, 1 - y_i(W^T \phi(x_i) - b)) \right] + \lambda ||W||^2 \tag{6.17}$$

donde $\phi(\cdot)$ es la función de transformación no lineal que mapea las observaciones originales a un espacio de características transformado. Esto permite que las SVM kernel sean capaces de encontrar separaciones más complejas en los datos.

Cuando se realiza un modelo incorporando múltiples perceptrones en capas para afrontar un problema de clasificación, se obtiene una red neuronal, una técnica de aprendizaje automático basada en grafos e inspirada en redes neuronales biológicas y que puede utilizarse tanto en aprendizaje supervisado como no supervisado.

6.4.1. Método de K-Nearest Neighbors

El método de K vecinos más cercanos (k-NN), desarrollado por Evelyn Fix y Joseph Hodges en 1951, brinda un enfoque popular en el aprendizaje supervisado utilizado ampliamente en el análisis de datos y la minería de datos por su superioridad en términos de eficiencia frente a MCO (James et al., 2013). Este método se utiliza tanto para la clasificación como para la regresión y se basa en la idea de que los observaciones con características similares pertenecen a la misma clase. La clasificación mediante k-NN consiste en minimizar la distancia entre un individuo sin clasificar y los individuos del modelo en el hiperplano de las variables regresoras.

La idea detrás de k-NN es encontrar los K vecinos más cercanos a una nueva muestra en el espacio de entrada y asignar una etiqueta en base a la mayoría de las etiquetas de los vecinos seleccionados (Géron, 2019). La cantidad de vecinos K se puede especificar previamente y afecta directamente la precisión del modelo.

$$Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j) \tag{6.18}$$

Donde N_0 son los k individuos de entrenamiento cercanos a x_0.

El algoritmo k-NN se ejecuta en dos fases: en la primera fase, se almacenan todas las muestras de entrenamiento y sus etiquetas correspondientes, y en la segunda fase, se utiliza una medida de similitud (comúnmente la distancia Euclidiana):

$$d(x, y) = \sqrt{\sum (xi - yi)^2} \tag{6.19}$$

La idea detrás de k-NN es encontrar los K vecinos más cercanos a una nueva muestra en el espacio de entrada y asignar una etiqueta en base a la mayoría de las etiquetas de los vecinos seleccionados. La cantidad de k vecinos se puede especificar previamente y afecta directamente la precisión del modelo (Hastie et al., 2009).

$$\widehat{Y(x)} = \frac{1}{k} \sum_{i \in N_k(x)} y_i \tag{6.20}$$

En el caso de un problema de regresión, en vez de asignarle el valor que más se repite entre los k vecinos más cercanos, se obtiene una media ponderada del valor de la variable de dichos individuos que se pondera en función de la distancia a la nueva observación.

El método de k-NN presenta resultados altamente consistentes. A medida que la cantidad de observaciones en el conjunto de entrenamiento se acerca al infinito, se garantiza que el algoritmo k-NN de dos clases producirá una tasa de error no peor que el doble de la tasa de error de Bayes, la cual es la tasa de error mínima alcanzable dada la distribución de los datos.

6.4.2. Árboles de Elección Binaria

Los árboles de elección binaria, también conocidos como árboles de decisión, son una técnica algorítmica que permite representar estructuras ramificadas para la clasificación y regresión de observaciones. En esta técnica, se emplean pruebas lógicas en cada nodo del árbol para dividir las observaciones en submuestras de acuerdo con sus características.

Una de las principales ventajas de los árboles de decisión es su capacidad para incorporar los costos de decisión en el modelo. Esto significa que se puede asignar un valor numérico a cada posible decisión, lo que permite una mejor toma de decisiones considerando aspectos económicos o de utilidad.

Los árboles de regresión y clasificación son ampliamente utilizados en la predicción efectiva de observaciones, incluso en muestras diferentes a aquella utilizada para su construcción. Los primeros trabajos sobre el crecimiento automático de árboles de regresión se remontan a 1963, cuando James A. Morgan y John A. Sonquist de la Universidad de Michigan desarrollaron una técnica para abordar los problemas en el análisis de datos de encuestas.

Sin embargo, la utilización de árboles de regresión puede presentar inconvenientes en términos de sobreajuste a los datos de entrenamiento. Es decir, los árboles pueden adaptarse demasiado a los datos utilizados para su construcción y tener dificultades para generalizar correctamente a nuevas muestras. Para mitigar este problema, se puede implementar un costo de complejidad computacional que permita podar los árboles, lo que da lugar a lo que se denomina un árbol de inferencia

condicional.

(Varian, 2014) sostiene que la poda de árboles es una técnica fundamental para mejorar la generalización de los resultados a nuevas muestras de datos. El costo de complejidad computacional se utiliza como un parámetro de control que regula la complejidad del árbol, evitando que crezca excesivamente y se ajuste demasiado a los datos de entrenamiento. Al introducir este costo, se logra un equilibrio entre la precisión en la muestra de entrenamiento y la capacidad de predicción en nuevas observaciones.

La implementación del costo de complejidad computacional puede resultar en un árbol más simple, con menos divisiones y reglas, pero que tiene una mayor capacidad de generalización y predicción en datos no vistos anteriormente. Es decir, se busca encontrar un equilibrio entre la complejidad del árbol y su capacidad para representar las relaciones subyacentes en los datos de manera más general.

6.4.3. Métodos de Combinación de Árboles de Decisión

Los métodos de árboles de decisión son ampliamente empleados en el aprendizaje automático debido a su simplicidad y facilidad de interpretación. Sin embargo, para mejorar su precisión y rendimiento predictivo, se han desarrollado técnicas de combinación de árboles de decisión, como bagging, random forests y boosting.

La técnica de bagging consiste en crear múltiples muestras de entrenamiento mediante muestreo con reemplazo del conjunto de datos original. A partir de estas muestras, se construyen árboles de decisión independientes, y las predicciones de cada árbol se promedian para formar un modelo combinado.

La técnica de random forests es una variante del bagging donde, además del muestreo con reemplazo, se realiza una selección aleatoria de características en cada árbol. Esto reduce la correlación entre los árboles y mejora aún más la capacidad predictiva del modelo combinado.

El boosting es otra técnica de combinación de árboles de decisión que se basa en la construcción de árboles secuenciales, donde cada nuevo árbol se enfoca en corregir los errores cometidos por el modelo anterior. En cada iteración, se da mayor peso a las observaciones mal clasificadas, lo que permite que el modelo se enfoque en aprender de los errores.

En la construcción y evaluación de modelos de árbol de decisión mediante estos métodos de combinación, la teoría de probabilidades juega un papel fundamental. El cálculo de probabilidades en cada nodo de decisión y la definición de una función de pérdida permiten evaluar el rendimiento del modelo y seleccionar las variables más relevantes en la clasificación.

Es importante tener en cuenta que al utilizar técnicas de combinación de árboles, como bagging, random forests y boosting, se puede perder la transparencia y simplicidad de un solo árbol, lo

que dificulta la interpretación del modelo. Sin embargo, la mejora en la precisión de la predicción generalmente compensa esta pérdida, por lo que estos métodos son ampliamente utilizados en el aprendizaje automático (James et al., 2013).

6.4.4. Bosques Aleatorios

Los bosques aleatorios son una técnica de aprendizaje utilizada tanto en problemas de clasificación como en regresión. Consiste en la construcción de múltiples árboles de decisión durante el proceso de entrenamiento, donde cada árbol se desarrolla de forma independiente y no se somete a un proceso de poda. Esta característica de no podar los árboles permite que capturen una mayor cantidad de información y evita el problema del sobreajuste que puede presentarse en los árboles individuales.

El algoritmo de construcción de los bosques aleatorios se basa en la división consecutiva de los datos de entrenamiento utilizando un criterio de asignación específico. Cada árbol se construye mediante un proceso de división recursiva, donde se selecciona una variable y un punto de corte que maximice la homogeneidad o la reducción de la impureza en los nodos resultantes. Esta división se repite hasta alcanzar un criterio de parada, como por ejemplo, el tamaño mínimo de los nodos o la profundidad máxima del árbol.

Un aspecto importante de los bosques aleatorios es el uso de técnicas de bootstrap y boosting para corregir posibles problemas de sobreajuste y mejorar la precisión del modelo. El bootstrap se emplea en la selección de las muestras de entrenamiento utilizadas para construir cada árbol. En cada iteración, se selecciona una muestra de tamaño igual al conjunto de entrenamiento, pero con reemplazo. Esto permite tener conjuntos de datos diferentes en cada árbol y capturar diferentes perspectivas del problema. Por otro lado, el boosting se utiliza en la combinación de los árboles para formar el modelo final. Cada árbol se construye de forma secuencial, enfocándose en corregir los errores del modelo anterior.

Aunque los bosques aleatorios son una técnica poderosa en términos de precisión y capacidad de manejo de datos, presentan la limitación de ser considerados çajas negras". Esto significa que no proporcionan información resumida explícita sobre las relaciones y las interacciones presentes en el modelo computacional. A diferencia de otros métodos, como la regresión lineal, los bosques aleatorios no brindan coeficientes o estimaciones directas de la importancia de las variables en el proceso de toma de decisiones.

6.4.5. Regresión regularizada

Los métodos de regularización matemática son comúnmente utilizados para penalizar un modelo en función de su complejidad. Un modelo más simple puede reducir el sobreajuste, colaborar

para el reformulamiento de un problema mal planteado y mejorar la interpretabilidad estadística de los resultados.

Algunos de los modelos de regresión más recientes integran una gran cantidad de predictores para la explicación de la variable dependiente, lo que puede generar problemas de multicolinealidad y requerir de criterios de regularización para la selección de las variables más determinantes en el pronóstico en aras de la obtención de un modelo más eficiente para la predicción de datos por fuera de la muestra.

Estos métodos de regularización permiten estimar los coeficientes mediante una regresión penalizada, minimizando una SCE sujeta a una restricción, normalmente, descartando aquellas variables que tienen un menor impacto sobre la variación de y, o limitando el valor de los parámetros. Los criterios de regularización permiten incrementar el número de grados de libertad en un modelo de regresión y reducir el error (Hastie et al., 2009).

Los métodos de regresión Ridge y LASSO son técnicas de regularización utilizadas para reducir el sobreajuste en modelos de regresión lineal. El sobreajuste ocurre cuando el modelo se ajusta demasiado a los datos de entrenamiento y no generaliza bien a nuevos datos. Estos métodos agregan una penalización a la función de pérdida de la regresión para limitar el valor de los coeficientes de regresión.

El método de regresión Ridge se basa en la minimización de la suma residual de cuadrados penalizada, donde se agrega un término de penalización que depende del valor al cuadrado de los coeficientes de regresión. La función objetivo a minimizar es:

$$\hat{\beta^{ridge}} = armin \sum_{i=1}^{N} (y_t - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 \tag{6.21}$$

$$sujeto\ a \sum_{j=1}^{P} \beta_j^2 \leq t$$

Teniendo así, una función de perdida con restricción a minimizar de la siguiente forma:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{6.22}$$

donde β_0 es la intersección, β_j son los coeficientes de regresión y x_{ij} es el valor de la variable independiente j para el i-ésimo observación. El término de penalización $\lambda \sum_{j=1}^{p} \beta_j^2$ se agrega para penalizar los coeficientes grandes de regresión, con λ como un parámetro de ajuste que controla la

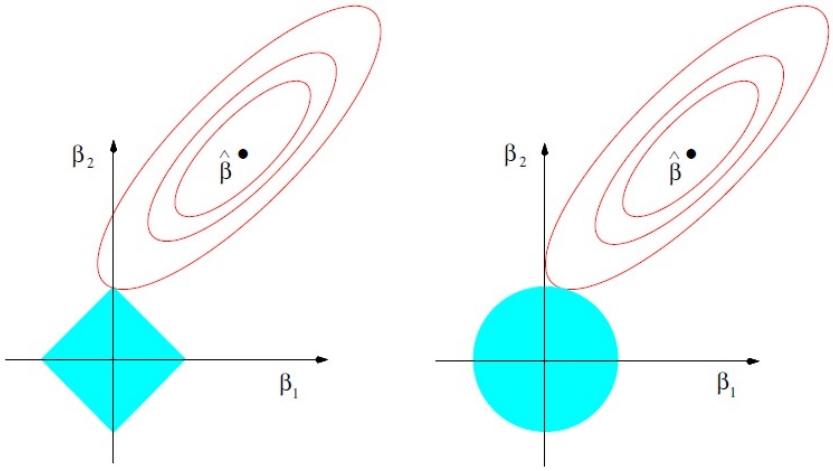

Figura 6.2: Estimaciones Lasso y Ridge.
(Hastie et al., 2009)

magnitud de la penalización.

El método de regresión LASSO, por otro lado, utiliza una penalización diferente que es proporcional al valor absoluto de los coeficientes de regresión. La función objetivo a minimizar es:

$$\hat{\beta^{lasso}} = argmin \sum_{i=1}^{N} (y_t - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 \tag{6.23}$$

$$sujeto\ a \sum_{j=1}^{P} |\beta| \leq t$$

Obteniendo una función de pérdida penalizada de la siguiente forma:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{6.24}$$

De manera similar a Ridge, el término de penalización $\lambda \sum_{j=1}^{p} |\beta_j|$ se agrega para penalizar los coeficientes grandes de regresión. La principal diferencia con Ridge es que LASSO tiene la capacidad de reducir algunos coeficientes de regresión a cero, lo que puede resultar en un modelo más simple y fácil de interpretar.

6.5. Pronóstico en econometría con modelos uniecuacionales

Al formular un modelo uniecuacional se asume una relación de causalidad unidireccional entre una serie de variables regresoras que tienen incidencia sobre una regresada. Este tipo de modelo se construye con el objetivo principal de pronosticar el comportamiento de dicha variable para distintas observaciones en el tiempo o el espacio.

Conociendo las relaciones existentes entre las variables de un sistema puede predecirse el comportamiento de las mismas a lo largo de su observación, lo que permite, por ejemplo, la aplicación de políticas adecuadas previendo los efectos futuros de una configuración dada de las circunstancias que constituyen el panorama macroeconómico de la nación, o anticipar la conducta de producción y consumo en un mercado teniendo en cuenta los incentivos existentes que afectan las decisiones de los agentes económicos en este, y tomar decisiones que resulten en un mayor rédito.

(Pindyck & Rubinfeld, 1998) sostiene que la predicción es el principal objetivo de la econometría. Los modelos de econometría buscan captar relaciones causales entre variables para hacer predicciones precisas. Sin embargo, los modelos de econometría convencionales, como el modelo de regresión lineal, se basan en supuestos estadísticos sobre los datos y requieren pruebas estadísticas para validar su uso. Los métodos de ML para pronóstico, por otro lado, pueden aprender patrones complejos y no lineales directamente de los datos y no requieren supuestos estadísticos fuertes para su uso.

(Wooldridge, 2016) Explica que los modelos de regresión lineal tienen supuestos específicos sobre la forma funcional de la relación entre los predictores y el resultado. Si estos supuestos no se cumplen, los modelos de regresión lineal pueden producir resultados inexactos. Por otro lado, los modelos de *ML* para pronóstico no hacen supuestos fuertes sobre la forma funcional de la relación entre los predictores y el resultado, lo que les permite capturar patrones más complejos y no lineales.

Por último, (Gujarati et al., 2010) mencionan que los modelos de regresión lineal son útiles en situaciones en las que se conocen las variables que influyen en un resultado y se desea cuantificar su impacto. Sin embargo, estos modelos son limitados en situaciones en las que no se conocen todas las variables que influyen en el resultado. Los modelos de *ML* para pronóstico pueden manejar estas situaciones, ya que pueden aprender patrones y relaciones complejas directamente de los datos, sin la necesidad de conocer todas las variables que influyen en el resultado.

Es preciso mencionar que para la realización del estudio expuesto en el presente proyecto se hace uso de una metodología enfocada en el análisis de una base de datos de corte transversal, es decir, recopilada mediante la observación de múltiples datos en un mismo periodo de tiempo o sin marcas de tiempo definidas, puesto que la mayoría de aplicaciones de aprendizaje automático trabajan sin considerar diferencias en los periodos de observación de los individuos de la muestra, y adicionalmente, se asume una correlación nula e igual distribución para los datos.

6.6. Error en un modelo estadístico

Cuando se realiza la estimación de un modelo utilizando datos de una muestra aleatoria, la calidad del estimador de la relación entre las variables dependerá de la representatividad de los datos de la muestra en relación con la población, entre otros factores.

Es importante tener en cuenta que los parámetros estimados en un modelo de regresión son variables aleatorias. Por lo tanto, al calcular el error de pronóstico como la diferencia entre la regresión del modelo estimado y el modelo poblacional, se obtiene una ecuación con dos fuentes de error: el término de error del modelo poblacional debido a la varianza de la variable dependiente, y la aleatoriedad de los parámetros de regresión estimados, que dependerá del número de grados de libertad en el modelo.

El tamaño de la muestra utilizada en el proceso de estimación del modelo y la varianza de las variables regresoras son factores que influyen en la magnitud del error de pronóstico. A medida que aumenta el tamaño de la muestra y la varianza de las variables regresoras, el error de pronóstico disminuye. Además, los pronósticos realizados cerca de la media de las variables regresoras tienen un menor error, ya que se reducirá la incidencia que tiene la aleatoriedad de los parámetros en el pronóstico

6.6.1. Error estándar como medida de la precisión en el pronóstico

El error estándar de pronóstico es una medida comúnmente utilizada para evaluar la precisión de un modelo predictivo. Mide la variabilidad de los valores pronosticados de una variable dependiente en torno a su valor esperado. El error estándar se utiliza para cuantificar la incertidumbre asociada con un pronóstico y es una herramienta importante para evaluar la precisión de los modelos econométricos.

Adicionalmente, el error cuadrático medio de pronóstico de un modelo se utiliza también para medir su capacidad predictiva, al igual que la desviación media absoluta de pronóstico.

6.6.2. Cross Validation

En el campo de la estadística computacional, tanto el cross validation como el K-fold cross validation son técnicas utilizadas para evaluar y validar modelos estadísticos, especialmente en el contexto del aprendizaje automático y la regresión.

La validación cruzada es una técnica esencial en la evaluación de modelos de aprendizaje automático, que permite medir su capacidad de generalización a datos desconocidos (James et al., 2013). Esta técnica consiste en dividir el conjunto de datos en dos subconjuntos: uno para entrenamiento y otro para evaluación (Géron, 2019). La validación cruzada puede aplicarse a cualquier

modelo predictivo.

Sin embargo, el enfoque de validación cruzada presenta una limitación: se basa en una única división de los datos en conjuntos de entrenamiento y prueba, lo que puede generar una estimación sesgada y poco robusta del rendimiento del modelo, especialmente cuando se tienen conjuntos de datos pequeños o cuando la división introduce cierta aleatoriedad. Para superar esta limitación, surge la técnica de validación cruzada K-fold.

6.6.3. K-fold Cross Validation

En el método K-fold de validación cruzada, el conjunto de datos se divide en K subconjuntos o pliegues de tamaño aproximadamente igual. Luego, se realiza un proceso iterativo en el que cada subconjunto se selecciona una vez como conjunto de prueba, mientras que los K-1 subconjuntos restantes se utilizan como conjunto de entrenamiento. El modelo se entrena y evalúa K veces, utilizando cada uno de los subconjuntos como conjunto de prueba en una iteración diferente. Finalmente, se promedian los resultados obtenidos en las K iteraciones para obtener una estimación más robusta y precisa del rendimiento del modelo (James et al., 2013).

La diferencia principal entre la validación cruzada y la validación cruzada K-fold radica en la forma en que se realiza la evaluación del modelo. Mientras que la validación cruzada utiliza una única división de los datos en conjuntos de entrenamiento y prueba, la validación cruzada K-fold realiza múltiples divisiones del conjunto de datos y promedia los resultados obtenidos. El valor de K es un parámetro que se elige de antemano y puede variar según el tamaño del conjunto de datos y la cantidad de información disponible.

$$CV_k = \frac{1}{k} \sum_{i=1}^{k} ECM_i \tag{6.25}$$

Donde ECM_i es el error cuadrático medio

El método de validación cruzada K-fold es ampliamente utilizado en la estadística computacional debido a su capacidad para proporcionar una estimación más precisa y confiable del rendimiento del modelo. Al realizar múltiples evaluaciones con diferentes divisiones de los datos, se reduce la dependencia de una única división particular y se obtiene una visión más general del rendimiento del modelo en diferentes configuraciones de entrenamiento y prueba.

El método de validación cruzada K-fold se utiliza, por ejemplo, para encontrar los parámetros de penalización utilizados en métodos de regresión penalizada, de los cuales se hablará en el marco teórico. Además, este método permite encontrar el nivel óptimo de flexibilidad o ajuste que mi-

nimiza el error cuadrático medio, al probar la capacidad predictiva del modelo con datos que no fueron utilizados en su estimación, evitando así el sobreajuste.

6.7. Teoría de la Demanda Hedónica

Cuando se pretende estimar un modelo que explique o prediga el comportamiento del precio o de la demanda de un bien, se deben analizar las observaciones de las elecciones realizadas por los individuos en torno a dicho bien. La teoría de preferencias reveladas, introducida en la ciencia económica por el economista estadounidense Paul Anthony Samuelson en 1938, es una rama de la microeconomía que desarrolla una serie de modelos en los que se asume que las preferencias de los consumidores pueden ser obtenidas mediante el análisis de sus hábitos de consumo.

Debido a que los modelos fundamentados sobre la teoría de elección del consumidor requieren conocer la estructura de las preferencias de los individuos para comprender su proceso de decisión, la teoría de preferencias reveladas brinda un enfoque empírico para la estimación de las funciones de utilidad. En econometría, uno de los métodos de preferencia revelada utilizados frecuentemente a la hora de estimar los precios o las cantidades demandadas de un bien es el de la regresión hedónica, basada en la teoría de precios hedónicos.

La teoría de los precios hedónicos está formulada con base en la hipótesis de que el valor de un bien se determina por los precios implícitos de sus atributos, en otras palabras, que en un mercado competitivo son aquellas características que le otorgan cierto grado de diferenciación en relación con otros bienes de su misma naturaleza las que determinan su precio. Adicionalmente, los precios inherentes a los atributos de un bien pueden ser revelados observando los precios de los productos diferenciados y las cantidades específicas de los atributos asociados a ellos (Rosen, 1974) por ello, estos precios pueden ser estimados mediante mínimos cuadrados.

Debido a que en el caso puntual de los mercados inmobiliarios, los bienes como casas o apartamentos suelen ser bastante heterogéneos, para analizar la elección del consumidor y predecir los precios de dichos bienes, se tienen en cuenta características que son fácilmente observables, como las dimensiones de la edificación, del predio que constituye la propiedad, la cantidad de habitaciones y el uso que estas tienen, los medios de acceso al interior y al exterior de la propiedad, y características circunstanciales tales como la ubicación de la vivienda con respecto a determinados puntos de interés, el estrato socio-económico del sector, la estética del mismo entre otras.

Asumiendo que las características mencionadas en bienes inmuebles tienen por separado un aporte cuantificable al precio de la propiedad, se puede estimar, por medio de una regresión lineal, un modelo que prediga el precio de un bien inmueble dada cierta configuración de características.

Capítulo 7

Resultados

7.1. Análisis exploratorio

Para validar la correlación entre las variables explicativas y el precio de los apartamentos se realizó una serie de gráficos en R.

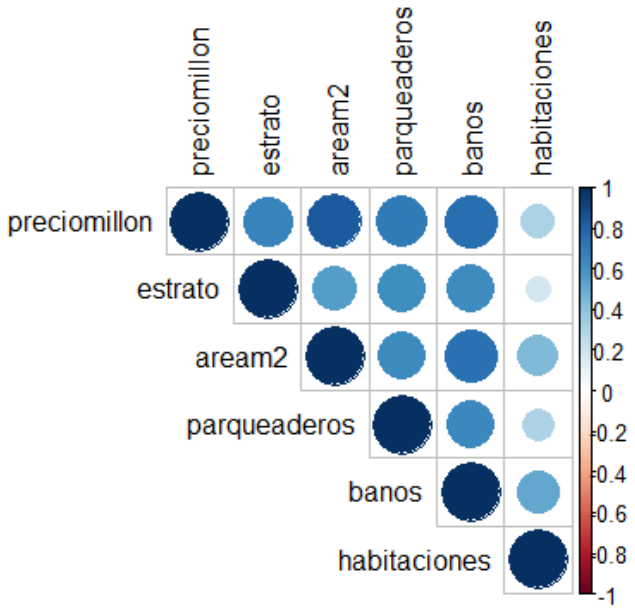

Figura 7.1: Correlaciones, Elaboración propia

La Figura **??** ilustra los niveles de correlación presentes entre las variables cuantitativas que serán utilizadas en los modelos, mostrando que existe una fuerte correlación entre las variables independientes, área en m^2, parqueaderos y baños específicamente con la variable dependiente precio.

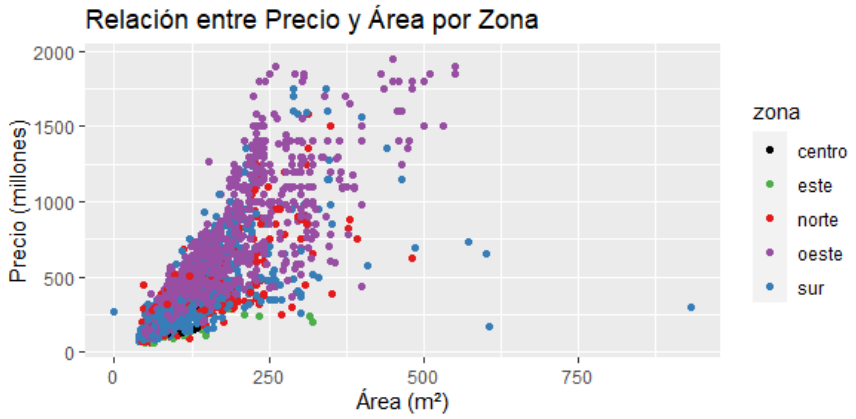

Figura 7.2: Precio y área, Elaboración propia

En la Figura 7.2 es notoria la relación positiva existente entre la variable precio y el área en m^2, y debido a la observación de una relación cuadrática entre el área y el precio, (Lo que se ha evidenciado antes en otros estudios similares) se incluirá esta variable en un polinomio cuadrático en el modelo econométrico.

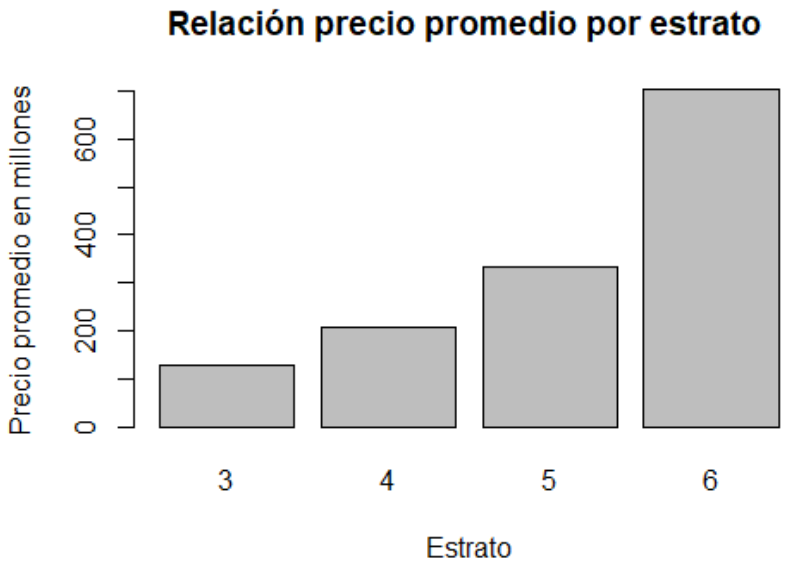

Figura 7.3: Precio y estratos, Elaboración propia

En cuanto a la Figura 7.3 es posible intuir de cierta manera la existencia de una relación positiva entre el precio promedio y el estrato.

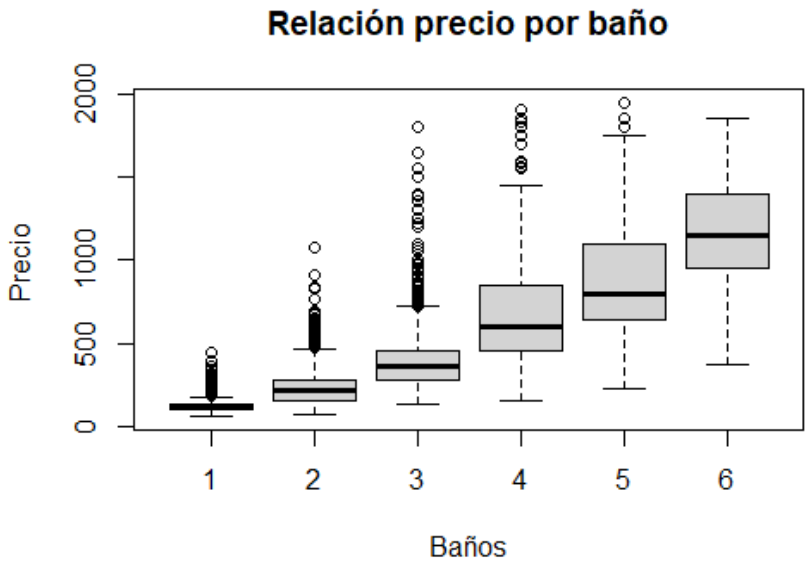

Figura 7.4: Precio y baños, Elaboración propia

La figura 7.4 muestra una relación positiva entre el número de baños en un apartamento y el precio del mismo.

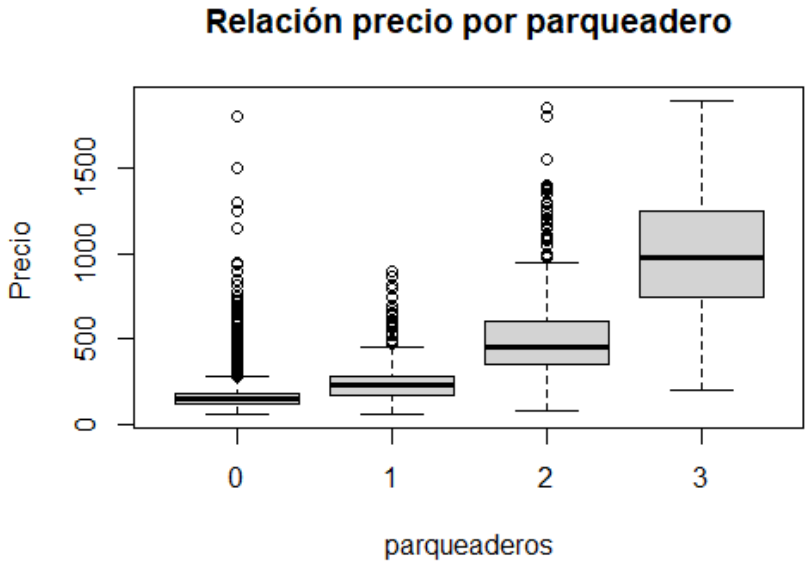

Figura 7.5: Precio y parqueaderos, Elaboración propia

Se encontró también que el precio promedio de los apartamentos es de 366,713 millones de pesos y una varianza de 83.773,09 millones de pesos

7.2. Regresión lineal múltiple

Modelo econométrico propuesto con base en otros formulados en estudios anteriores de la misma naturaleza:

$$Precio_i = \beta_1 + \beta_2 aream2_i + \beta_3 aream2_i^2 + \beta_4 zona_i + \beta_5 habitaciones_i + \beta_6 banos_i + \beta_7 parqueaderos_i + \beta_8 estrato_i + \beta_9 aream2_i * estrato_i + \beta_{10} aream2_i^2 * estrato_i + U_i \tag{7.1}$$

Cuadro 7.1: Coeficientes regresión con el 100 %

Variable	Coeficientes	Error estándar
(Intercept)	-8.279e+00	4.051e+01
aream2	2.316e+00	5.741e-01
aream2sq	-5.470e-03	1.874e-03
zonaeste	9.494e+00	2.818e+01
zonanorte	1.673e+01	2.432e+01
zonaoeste	6.061e+01	2.464e+01
zonasur	6.164e+00	2.427e+01
habitaciones	-3.157e+01	3.125e+00
banos	3.973e+01	2.820e+00
parqueaderos	4.871e+01	2.761e+00
estrato4	3.012e+01	3.317e+01
estrato5	2.553e+01	3.073e+01
estrato6	-1.642e+02	3.356e+01
aream2:estrato4	-3.806e-01	6.309e-01
aream2:estrato5	-6.816e-03	5.828e-01
aream2:estrato6	2.509e+00	5.936e-01
aream2sq:estrato4	3.571e-03	2.055e-03
aream2sq:estrato5	3.619e-03	1.884e-03
aream2sq:estrato6	1.177e-03	1.902e-03

Adicionalmente, aunque las pruebas gráficas parecen Q-Q, y de la distribución de los errores parecen indicar que los errores se comportan normalmente, los resultados de las pruebas de Shapiro Wilk y Kurtosis y Oblicuidad indican que los residuos del modelo no se distribuyen de esta forma.

Además, se verificó la homocedasticidad de los errores a través de la prueba de Breusch-Pagan, una baja multicolinealidad (para los predictores sin interacción) con el índice inflacionario de la varianza VIF, y la una correcta especificación mediante el test RESET de Ramsey.

A continuación, se presentan los resultados obtenidos de la estimación del modelo que será utilizado para pronóstico:

Cuadro 7.2: Coeficientes regresión con el 70 % de las observaciones

Variable	Coeficiente	Error estándar
(Intercept)	6.540e-01	4.597e+01
aream2	2.196e+00	6.671e-01
aream2sq	-5.247e-03	2.307e-03
zonaZona Norte	1.009e+01	2.757e+01
zonaZona Oeste	5.538e+01	2.792e+01
zonaZona Oriente	4.140e+00	3.197e+01
zonaZona Sur	9.552e-01	2.749e+01
habitaciones	-3.018e+01	3.558e+00
banos	3.824e+01	3.187e+00
parqueaderos	5.130e+01	3.149e+00
estrato4	2.370e+01	3.781e+01
estrato5	2.806e+01	3.528e+01
estrato6	-1.653e+02	3.853e+01
aream2:estrato4	-2.488e-01	7.247e-01
aream2:estrato5	4.737e-03	6.753e-01
aream2:estrato6	2.567e+00	6.878e-01
aream2sq:estrato4	3.341e-03	2.473e-03
aream2sq:estrato5	3.583e-03	2.315e-03
aream2sq:estrato6	9.941e-04	2.335e-03

Al realizar la prueba con la submuestra de testeo se obtuvo un ECM de 12.874,1 y un índice de Theil bajo de 0.1223.

Además se encontró que las variables más significativas en orden para este modelo fueron, área en metros cuadrados, número de baños, estrato socioeconómico, número de parqueaderos, seguidas de las distintas zonas, algo llamativo es que la zonas más importantes son oeste y sur y la menos importante es oriente.

7.3. Random forest

Cuadro 7.3: Importancia de las variables

Variable	Reducción promedio de la impureza
aream2	75586572
aream2sq	**79257522**
zona	11358502
habitaciones	3666131
banos	37324042
parqueaderos	36661257
estrato	38873026

Para el caso del modelo de Random forest se obtuvieron los mejores resultados en términos de desempeño predictivo, con un ECM (error cuadrático medio) de 10.865,87 y un índice de Theil de 0.1179139 siendo ambos valores los más bajos entre los modelos realizados y probados.

Adicionalmente, se encontró que las variables que tienen un mayor impacto en orden para este modelo fueron, área en metros cuadrados, número de baños, estrato socioeconómico, número de parqueaderos, seguidas de las distintas zonas, algo llamativo es que la zonas más importantes son oeste y sur y la menos importante es oriente.

7.4. LASSO

Cuadro 7.4: Coeficientes LASSO

Variable	**Coeficientes**
(Intercept)	2.037931e+01
aream2	1.813303e+00
aream2sq	-3.289811e-03
zonaeste	.
zonanorte	9.573848e+00
zonaoeste	5.354325e+01
zonasur	-9.835256e-01
habitaciones	-3.148487e+01

Cuadro 7.4 – Continuación

Variable	**Coeficiente**
banos	3.996478e+01
parqueaderos	4.869142e+01
estrato4	1.122937e+01
estrato5	3.764239e+00
estrato6	-1.729389e+02
aream2:estrato4	6.379231e-02
aream2:estrato5	4.887538e-01
aream2:estrato6	2.877123e+00
aream2sq:estrato4	1.539696e-03
aream2sq:estrato5	1.443142e-03
aream2sq:estrato6	-7.379200e-04

Inicialmente en cuanto a la capacidad predidctiva del modelo LASSO se puede observar que el ECM fue relativamente bajo 13.400,96, en comparación con la varianza del precio, indicando que el modelo tiene la precisión suficiente para predecir el precio del apartamento.

El índice de Theil fue de 0.1258 indicando que el modelo tiene una buena capacidad de generalización y puede predecir con precisión nuevos datos que no se pueden encontrar durante el entrenamiento del modelo.

Además, este modelo es útil para identificar las variables más importantes en la predicción del precio de los apartamentos las cuales fueron, área en metros cuadrados, zona en la que se ubica el apartamento, el número de habitaciones, baños y parqueaderos y estrato socio económico de la vivienda. Se hace importante enunciar que previamente a la estimación del modelo se implementó validación cruzada para encontrar el mejor valor de lambda posible para la penalización, siendo este 0.07371026.

Es importante tener en cuenta que estos coeficientes están sujetos a la penalización de la regularización LASSO y, por lo tanto, son menores en magnitud que los coeficientes estimados por una regresión lineal. Además, como se mencionó anteriormente, el modelo LASSO selecciona las variables más importantes para la predicción y descarta aquellas con un efecto más débil.

Ahora bien en lo particular, se observo que las zonas de ubicación Zona Norte y Zona Oeste tienen un efecto positivo significativo en el precio.

Se observó que la interacción entre área en metros cuadrados y estrato socio-económico es una variable importante para predecir el precio. Esto indica que el impacto de los metros cuadrados en el precio puede diferir dependiendo del estrato del apartamento.

7.5. k-NN

Finalmente dentro de los modelos de aprendizaje automático realizados en el presente proyecto, está el modelo de k-NN, utilizado para regresión por votación, con un ECM de 13.507,06 y un índice de Theil de 0.1234, teniendo como variables mas influyentes el área en m^2, el estrato socioeconómico, el número de parqueaderos y el número de baños.

Debido a que este modelo normalmente se realiza para afrontar problemas de clasificación, en los que ha demostrado tener un buen desempeño, era de esperarse que no tuviese un desempeño tan alto al enfrentar un problema de regresión, no obstante, aunque su capacidad predictiva estuvo por debajo de las demás técnicas, esto no fue por un margen muy grande, lo que indica que puede ser utilizado también como una técnica para regresión de variables cuantitativas.

7.6. Resultados generales

Cuadro 7.5: Resultados generales

Método	ECM	Índice de Theil
Regresión lineal múltiple	12,874.1	0.1223
Random Forest	10,865.87	0.1179
LASSO	13,400.96	0.1258
k-NN	13,507.06	0.1234

Capítulo 8

Conclusiones

- El modelo con el mejor desempeño utilizando como instrumento de medición del error el ECM y el índice de Theil fue el de Random Forest, lo cual, era un resultado esperado debido a la sofisticación del método y su efectividad en predicción en corte transversal ya demostrada en otros estudios.

- Se observo una correlación negativa entre el número de habitaciones y su precio. Este fenómeno es común en los modelos econométricos de este tipo de estudios y puede deberse a factores como la disponibilidad de espacio en el inmueble, la ubicación, las comodidades y las preferencias o los costos de construcción. Adicionalmente, la variable que captura la valorización del inmueble en función del espacio del mismo es el área, la cual, presentó un comportamiento creciente pero cóncavo, indicando que un mayor área para un apartamento tendrá el mismo efecto positivo en el precio del mismo cuando el espacio excede ciertas dimensiones.

- El modelo de regresión lineal múltiple mostró resultados no muy alejados de las técnicas de aprendizaje automático utilizadas e incluso superó al método LASSO, lo que puede explicarse debido a la poca heterogeneidad de los datos, su número relativamente pequeño, su buen comportamiento en relación a su linealidad y normalidad, la inclusión de no muchas variables en los modelos y la buena especificación del modelo econométrico.

- Aunque, siendo la técnica más comúnmente utilizada en modelos de aprendizaje automático, k-NN se utiliza normalmente para afrontar problemas de clasificación, ha demostrado ser útil para problemas de regresión como el del experimento realizado obteniendo resultados similares a los del modelo de regresión lineal múltiple.

- Se recomienda realizar más estudios para validar empíricamente el desempeño de modelos econométricos y de aprendizaje automático en predicción para poder comparar el rendimiento de otras técnicas y en escenarios en los que las condiciones de los datos puedan variar.

Bibliografía

Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, *15*(4), 69-85.

Athey, S. (2018). The Impact of Machine Learning on Economics. *The Economics of Artificial Intelligence An Agenda*. https://www.gsb.stanford.edu/faculty-research/publications/impact-machine-learning-economics

Athey, S., & Imbens, G. W. (2019a). Machine Learning Methods That Economists Should Know About. *Annual Review of Economics*, *11*(1), 685-725. https://doi.org/10.1146/annurev-economics-080217-053433

Athey, S., & Imbens, G. W. (2019b). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685-725.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.

Boelaert, J., & Ollion, '. (2018). The Great Regression: Machine Learning, Econometrics, and the Future of Quantitative Social Sciences. *Revue Française de Sociologie*, *59*(3), 475-506.

Charpentier, A., Flachaire, E., & Ly, A. (2019). Econometrics and Machine Learning. *Economie et Statistique / Economics and Statistics*, *505d*, 147-169. https://doi.org/10.24187/ecostat.2018.505d.1970

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc.

Greene, W. H. (2018). *Econometric analysis* (8th). Pearson Education Limited.

Gujarati, D. N., Porter, D. C., & Gunasekar, S. (2010). *Basic econometrics*. Tata McGraw-Hill Education.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer.

Hurwitz, J., & Kirsch, D. (2018). *Machine Learning IBM Limited Edition*. IBM. https://www.ibm.com/downloads/cas/GB8ZMQZ3

Iskhakov, F., Rust, J., & Schjerning, B. (2020). Machine learning and structural econometrics: contrasts and synergies. *The Econometrics Journal*, *23*(3), S81-S124.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112).

Liu, Y., & Xie, T. (2019). Machine learning versus econometrics: prediction of box office. *Applied Economics Letters*, *26*(2), 124-130.

López de Prado, M. (2019). Más allá de la econometría: una hoja de ruta hacia el aprendizaje automático financiero [Disponible en SSRN 3365282].

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: Methods and Applications*. John Wiley & Sons.

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, *31*(2), 87-106.

Pindyck, R. S., & Rubinfeld, D. L. (1998). *Econometric models and economic forecasts* (4th). McGraw-Hill/Irwin.

Rao, C. R., & Toutenburg, H. (1995). Linear models. En *Linear models* (pp. 3-18). Springer.

Rojas, E. M. (2020). Machine Learning: análisis de lenguajes de programación y herramientas para desarrollo. *Revista Ibérica de Sistemas e Tecnologias de Informação*, (E28), 586-599.

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, *82*(1), 34-55. http://www.jstor.org/stable/1830899

Shobana, G., & Umamaheswari, K. (2021). Forecasting by machine learning techniques and econometrics: A review. *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, 1010-1016.

Tchuente, D., & Nyawa, S. (2021). Real estate price estimation in French cities using geocoding and machine learning. *Annals of Operations Research*, *308*(1-2), 571-608. https://doi.org/10.1007/s10479-021-03932-5

Tse, R. Y. C. (1997). Una aplicación del modelo ARIMA a los precios inmobiliarios en Hong Kong. *Journal of Property Finance*, *8*(2), 152-163. https://doi.org/10.1108/09588689710167843

Varian, H. R. (2014). Big data: new tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3-28.

Wooldridge, J. M. (2016). *Introducción a la econometría: Un enfoque moderno* (4a). Cengage Learning.

Anexos

A continuación, se adjunta el enlace de acceso al repositorio de GitHub en el que se encontraran los archivos correspondientes a las pruebas realizadas y los resultados obtenidos en los diferentes métodos y el conjunto de datos utilizado.

https://github.com/Salazar1019/Machine-learning-Vs-Econometria.git