C&T Riqchary
Revista de Investigación en Ciencia y Tecnología
ISSN: 2810-8124 (en línea) / ISSN: 2706-543x

Vol. 5 Núm. 1 (2023)
Publicado: 15/10/23
DOI: https://doi.org/10.57166/riqchary/v5.n1.2023.117

La mejor opción para predecir el precio máximo de las acciones de Intel Corporation: ¿Árbol de regresión o Regresión Lineal Múltiple?

The best choice for predicting Intel Corporation's peak stock price: Regression Tree or Multiple Linear Regression?

Edgar Eloy Carpio Vargas [A], Alicia Roxana Mayda Huanca [B]. D. German Rafael Espinoza Rivas [C], Ecler Mamani Vilca[D] y [E]Betsabe M. Ccolqque Ruiz

ORCID:0000-0001-6457-4597[A], ORCID:0009-0006-4618-8872[B], ORCID: 0000-0002-6547-7761[C], ORCID: 0000-0002-5205-3660[D] ORCID 000-0002-3089-8750[E]

(Recepción: 15/04/2023 y aceptación 08/10/2023)

Resumen— El objetivo del estudio fue comparar el rendimiento del Árbol de regresión frente al Modelo de regresión lineal múltiple en relación al precio de apertura y volumen de ventas diarias de las acciones de Intel Corporation. Se llevó a cabo una investigación descriptiva correlacional de tipo no experimental con diseño transversal, utilizando una muestra por conveniencia. La muestra consistió en 410 registros recopilados desde mayo de 2018 hasta octubre de 2019, obtenidos a través de revisión documental. Los resultados obtenidos mostraron que el Árbol de regresión estableció que la variable más significativa para explicar el precio máximo de las acciones fue el precio de apertura, descartando la variable de volumen. El Error Medio Cuadrático obtenido fue de 1.4480 dólares. Por otro lado, el Modelo de regresión lineal múltiple, utilizando la técnica de eliminación de datos atípicos, presentó un Error Estándar Residual de 0.2257 dólares. En conclusión, se determinó que el modelo más adecuado para predecir el precio máximo de las acciones de Intel Corporation es el Modelo de Regresión Lineal Múltiple con eliminación de puntos atípicos.

Palabras clave: Árboles de regresión, regresión lineal multiple

Abstract— The objective of the study was to compare the performance of the Regression Tree versus the Multiple Linear Regression Model in relation to the opening price and daily sales volume of Intel Corporation shares. A descriptive correlational non-experimental correlational research with cross-sectional design was conducted using a convenience sample. The sample consisted of 410 records collected from May 2018 to October 2019, obtained through documentary review. The results obtained showed that the Regression Tree established that the most significant variable to explain the maximum stock price was the opening price, discarding the volume variable. The Mean Squared Error obtained was $1.4480. On the other hand, the multiple linear regression model, using the outlier elimination technique, presented a Residual Standard Error of 0.2257 dollars. In conclusion, it was determined that the most adequate model to predict the maximum price of Intel Corporation shares is the Multiple Linear Regression Model with the elimination of outlier points.

Keywords: Multiple Linear Regression, Regression Trees

A. *Edgar Eloy Carpio Vargas Departamento de Estadística e Informática – UNA Puno. ecarpio@unap.edu.pe*
B. *Alicia Roxana Mayda Huanca, Escuela de Ingeniería Estadística e Informática – UNA Puno: amaydana@epg.unap.edu.pe*
C. *German Rafael Espinoza Rivas, trabaja en el Departamento de Ingeniería – UNAMBA Perú: gespinoza@unamba.edu.pe*
D. *Ecler Mamani Vilca Departamento de Ingeniería informática y Sistemas UNAMBA, Perú : eclervirtual@unamba.edu.pe*
E. *Betsabe Milagros Ccolqque Ruiz, Departamento de Ingeniería informática y Sistemas UNAMBA bccolque@unamba.edu.pe*

1 INTRODUCCIÓN

La toma de decisiones juega un papel fundamental en cualquier organización o empresa. Por tanto, resulta imperativo buscar técnicas y metodologías estadísticas precisas que nos ayuden a resolver problemas de incertidumbre que surgen al tomar decisiones. Hoy en día, se generan grandes volúmenes de datos a diario, los cuales necesitan ser tratados a través de metodologías que sean capaces de generar información útil para investigar, predecir o tomar decisiones. Además, gracias al avance tecnológico, no solo podemos almacenar la información, sino también procesarla y generar conocimiento. [1].

Intel (Integrated Electronics Corporation) es el mayor fabricante de circuitos integrados del mundo según su cifra de negocio anual. La compañía estadounidense es la creadora de la serie de procesadores x86, los procesadores más comúnmente encontrados en la mayoría de las computadoras personales [2]. Sin embargo, para los accionistas e inversionistas de dicha empresa es muy riesgoso vender acciones sin saber si esta le generará ganancias o pérdidas de dinero al vender su bien, en este sentido, se plantea predecir el Precio máximo que una acción puede tomar durante un determinado tiempo, de esta forma los accionistas e inversionistas tendrán una mejor aproximación, sobre cuanto sería el Precio máximo que alcanzaría una acción. Si bien hoy en día, existe una gran diversidad de técnicas y metodologías estadísticas para predecir el costo de las acciones de las empresas en el mercado, se decidió realizar este estudio comparativo y nos planteamos la siguiente interrogante: ¿Cuál es la mejor opción para predecir el precio máximo de las acciones de Intel Corporation: Árboles de regresión o Regresión Lineal Múltiple

Los trabajos revisados fueron:

Gimenéz [3] indicaron que hacer pronósticos precisos de índices bursátiles es prácticamente imposible. Incluso teniendo modelos que parecieran dar buenos resultados, como los que elegimos finalmente, éstos pueden presentar problemas en la realidad debido al sobreajuste o “overfitting”.

Alvarez [4] indicó que la metodología Ítem a Ítem planteada por Amazon resulto útil para predecir el orden de preferencia de los ligandos que se unen a un determinado blanco sin embargo los resultados predictivos no resultaron suficientemente buenos. Si bien la medida aprendió la tendencia para cada proteína esta no pudo generalizar la información.

En cusco Candia [5] comprobó que, según los resultados encontrados, el algoritmo de árboles de decisión “Random Forest”, fue el algoritmo que tuvo el mejor performance para la predicción del rendimiento académico de los ingresantes en los primeros semestres a la UNSAAC con un 69% de predicción, el segundo algoritmo con mejor performance fue algoritmo de Regresión Logística con un 68% para el presente caso de estudio.

1.1 Árboles de Regresión

Los árboles de clasificación se emplean para asignar sujetos a las clases de una variable dependiente a partir de sus mediciones en uno o más predictores. Modernamente, los árboles de clasificación (AC) constituyen uno de los recursos instrumentales básicos de la llamada "minería de datos" [6]. Heredia [7] define los árboles de clasificación y regresión, conocidos como algoritmos CART (del inglés classification and regression trees), constituyen una aproximación multivariada no paramétrica que permite identificar y dimensionar las variables X de mayor impacto en una variable Y. Los modelos CART particionan los datos en forma recursiva de modo tal de conformar subconjuntos cada vez más homogéneos en base a criterios de partición de las variables explicativas. Cada árbol se obtiene a partir de la clasificación de un nodo parental o raíz que contiene la totalidad de los datos, mediante un algoritmo de partición especificado en función de un criterio de partición referido al tamaño del nodo formado o a la variabilidad contenida en los datos del nodo, también el algoritmo CART. Un árbol de regresión o de clasificación consiste en un conjunto de reglas determinadas por un procedimiento de ajuste mediante particiones binarias recursivas.

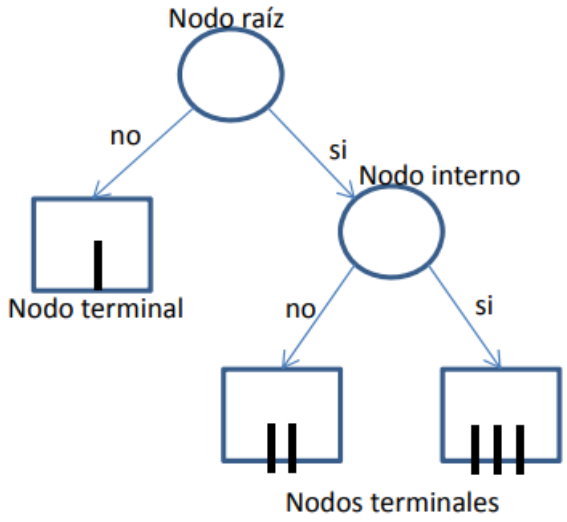

Fig. 1. Elementos del Árbol de regresión

En la fig. 1, el árbol tiene tres niveles de nodos, el primer nivel tiene un único nodo en la cima llamado nodo raíz. Un nodo interno en el segundo nivel, y tres nodos terminales que están respectivamente en el segundo y tercer nivel. El nodo raíz y el nodo interno son particionados cada uno en dos nodos en el siguiente nivel los cuales son llamados nodos hijos (o ramas) izquierdo y derecho.

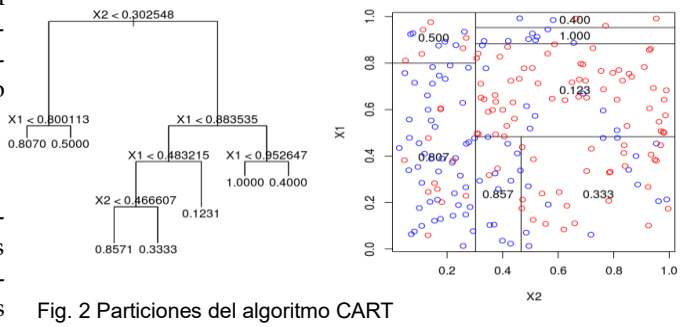

Fig. 2 Particiones del algoritmo CART

los árboles de regresión permiten predecir variables respuestas continuas, las observaciones según umbrales de las variables regresoras, considerando la suma de cuadrados de la respuesta como medida de heterogeneidad dentro de cada partición. Como en el árbol de clasificación, la medida de heterogeneidad entre las observaciones que quedan dentro de un nodo debe ser menor que la calculada entre las observaciones de distintos nodos.

C&T Riqchary
Revista de Investigación en Ciencia y Tecnología
ISSN: 2810-8124 (en línea) / ISSN: 2706-543x

Vol. 5 Núm. 1 (2023)
Publicado: 15/10/23
DOI: https://doi.org/10.57166/riqchary/v5.n1.2023.117

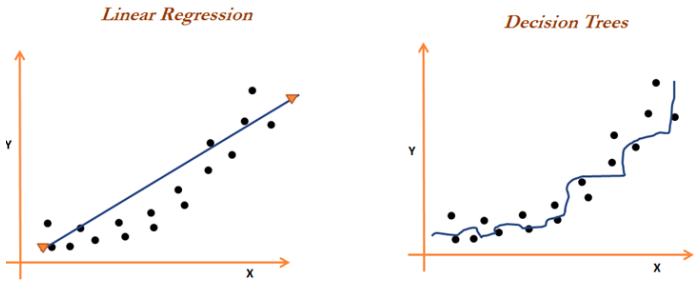

Fig. 3. Regresión lineal y el Árbol de decisión [7]

Podado del árbol, con la finalidad de reducir la varianza del modelo y así disminuir el test error, se somete al árbol a un proceso de poda, intenta encontrar el árbol más sencillo (menor tamaño) que consigue explicar las observaciones, y así para identificar el valor óptimo de penalización α. Por defecto, esta función emplea la desviación para guiar el proceso de poda y manteniendo la estructura robusta que consigue un test error bajo. La selección del sub-árbol optimo puede hacerse mediante la validación cruzada, sin embargo, dado que los árboles se crecen lo máximo posible (tienen muchos nodos terminales) no suele ser viable estimar el test error de todas las posibles sub-estructura que se pueden generan [8].

1.3 Algoritmo para crear un árbol de regresión con pruning

Paso 1:

Se emplea recursive binary splitting para crear un árbol grande y complejo (To) empleando los datos de training y reduciendo al máximo posible las condiciones de parada. Normalmente se emplea como única condición de parada el número mínimo de observaciones por nodo terminal.

Recursive binary splitting

El objetivo del método recursive binary splitting es encontrar en cada iteración el predictor Xj y el punto de corte (umbral) s tal que, si se distribuyen las observaciones en las regiones $\{X|Xj<s\}$ y $\{X|Xj\geq s\}$, se consigue la mayor reducción posible en el RSS. El algoritmo seguido es [9]:

a. El proceso se inicia en lo más alto del árbol, donde todas las observaciones pertenecen a la misma región.

b. Se identifican todos los posibles puntos de corte (umbrales) s para cada uno de los predictores ($X1$,$X2$,…,Xp). En el caso de predictores cualitativos, los posibles puntos de corte son cada uno de sus niveles. Para predictores continuos, se ordenan de menor a mayor sus valores, el punto intermedio entre cada par de valores se emplea como punto de corte.

c. Se calcula el RSS total que se consigue con cada posible división identificada en el paso b en ecuación (1).

$$\sum_{i:x_i \epsilon R_1(j,s)} \left(y_i - \hat{y}_{R_1}\right)^2 + \sum_{i:x_i \epsilon R_2(j,s)} \left(y_i - \hat{y}_{R_2}\right)^2 \quad (1)$$

donde el primer término es el RSS de la región 1 y el segundo término es el RSS de la región 2, siendo cada una de las regiones el resultado de separar las observaciones acordes al predictor j y valor s.

d. Se selecciona el predictor Xj y el punto de corte S que resulta en el menor RSS total, es decir, que da lugar a las divisiones más homogéneas posibles. Si existen dos o más divisiones que consiguen la misma mejora, la elección entre ellas es aleatoria.

e. Se repiten de forma iterativa los pasos 1 a 4 para cada una de las regiones que se han creado en la iteración anterior hasta que se alcanza alguna norma de stop. Algunas de las más empleadas son: que ninguna región contenga un mínimo de n observaciones, que el árbol tenga un máximo de nodos terminales o que la incorporación del nodo reduzca el error en al menos un % mínimo.

Paso 2:

Se aplica el cost complexity pruning al árbol To para obtener el mejor sub-árbol en función de α. Es decir, se obtiene el mejor sub-árbol para un rango de valores de α.

Cost complexity pruning es un método de penalización de tipo Loss + Penalty, similar al empleado en ridge regression o lasso. En este caso, se busca el sub-árbol T que minimiza la ecuación:

$$\sum_{j=1}^{|T|} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j}\right)^2 + \alpha |T| \quad (2)$$

donde |T| es el número de nodos terminales del árbol.

El primer término de la ecuación se corresponde con el sumatorio total de los residuos cuadrados RSS. Por definición, cuantos más nodos terminales tenga el modelo menor será esta parte de la ecuación. El segundo término es la restricción, que penaliza al modelo en función del número de nodos terminales (a mayor número, mayor penalización). El grado de penalización se determina mediante el tuning parameter α. Cuando α=0, la penalización es nula y el árbol resultante es equivalente al árbol original. A medida que se incrementa α la penalización es mayor y, como consecuencia, los árboles resultantes son de menor tamaño. El valor óptimo de α puede identificarse mediante cross validation.

Paso 3:

Identificación del valor óptimo de α mediante k-cross-validation. Se divide el training data set en K grupos. Para k=1, …,

k=K:

- Repetir pasos 1 y 2 empleando todas las observaciones excepto las del grupo ki.
- Evaluar el mean squared error para el rango de valores de α empleando el grupo ki.
- Obtener el promedio de los K mean squared error calculados para cada valor α.

Paso 4:
Seleccionar el sub-árbol del paso 2 que se corresponde con el valor α que ha conseguido el menor cross-validation mean squared error en el paso 3.

1.2 La regresión lineal múltiple

Un modelo de regresión lineal múltiple es un modelo estadístico versátil para evaluar las relaciones entre un destino continuo y los predictores **[10]**. La permite generar un modelo lineal en el que el valor de la variable dependiente o respuesta Y se determina a partir de un conjunto de variables independientes llamadas predictores X ´s. Es una extensión de la regresión lineal simple, por lo que es fundamental comprender esta última. Los modelos de regresión múltiple pueden emplearse para predecir el valor de la variable dependiente o para evaluar la influencia que tienen los predictores sobre ella (esto último se debe que analizar con cautela para no malinterpretar causa-efecto) **[11]**.

Los modelos lineales múltiples siguen la siguiente ecuación:

$$Y_i = (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_{ni}) + e_i$$

$$i = 1,2,3,\ldots,n \qquad (3)$$

$\boldsymbol{Y_i}$**:** Es la variable dependiente o respuesta.
$\boldsymbol{\beta_0}$**:** Es la ordenada en el origen, el valor de la variable dependiente Y cuanto todos predictores son cero.
$\boldsymbol{\beta_i}$**:** Es el efecto promedio que tiene el incremento en una unidad de la variable predictora
$\boldsymbol{Xi}$: Es la variable independiente o explicativas
$\boldsymbol{ei}$**:** Es el residuo o error, la diferencia entre el valor observado y el estimado por el modelo.

Condiciones para la Regresión Lineal Múltiple:

Los modelos de correlación lineal múltiple requieren de las mismas condiciones que los modelos lineales simples más otras adicionales, estas son [11]:

a. **No colinialidad o multicolinialidad:** En los modelos lineales múltiples los predictores deben ser independientes, no debe haber colinialidad entre ellos. La colinialidad ocurre cuando un predictor está linealmente relacionado con uno o varios de los otros predictores del modelo o cuando es la combinación lineal de otros predictores. Como consecuencia de la colinialidad no se puede identificar de forma precisa el efecto individual que tiene cada una de las variables colineales sobre la variable respuesta, lo que se traduce en un incremento de la varianza de los coeficientes de regresión estimados hasta el punto que resulta prácticamente imposible establecer su significancia estadística. Además, pequeños cambios en los datos provocan grandes cambios en las estimaciones de los coeficientes [12].

b. **Normalidad:** Los residuos deben seguir una distribución normal.

c. **Homocedasticidad:** La varianza del error debe ser constante para todos los valores del predictor.
d. **Independencia:** Los errores deben ser independientes entre sí.

Es importante tener en cuenta que estas condiciones son necesarias pero no suficientes para garantizar un buen ajuste del modelo.

Teniendo la variable dependiente (Y), es el precio máximo de venta que la acción ha logrado durante la negociación regular por un determinado periodo de tiempo; durante el día, para esta variable la unidad de medida es el dólar ($).

Variables independientes (X's)
X_1: Precio de apertura en el cual la acción es abierta al mercado (esta varia a lo largo del día). Al igual que la variable anterior, la unidad de medida es el dólar ($).
X_2: Volumen de ventas de un determinado periodo de tiempo de la empresa Intel. Para esta variable la unidad de medida es la cantidad de ventas que se registró, durante el día.

Torres [13] indica que, las fuentes de información, son todos aquellos medios de los cuales procede la información, que satisfacen las necesidades de conocimiento de una situación o problema presentado, que posteriormente será utilizado para lograr los objetivos esperados, para el estudio la obtención de datos fue mediante el internet, los cuales han sido obtenidos de la página https://www.nasdaq.com/es/market-activity/stocks/intc , "Nasdaq" , la técnica utilizada fue documental y el instrumento registro de la Base de Datos.

2 MATERIALES Y MÉTODOS

El diseño de investigación es cuantitativo de tipo descriptivo correlacional con diseño no experimental [14], debido al modo de obtencion de las variables. Por otro lado, se puede indicar que los datos son de corte transversal y fueron obtenidos en un corte de tiempo determinado.

La población para el estudio estuvo conformada por las acciones de la empresa Intel a través del tiempo y la muestra se

Revista de Investigación en Ciencia y Tecnología
ISSN: 2810-8124 (en línea) / ISSN: 2706-543x

Vol. 5 Núm. 1 (2023)
Publicado: 15/10/23
DOI: https://doi.org/10.57166/riqchary/v5.n1.2023.117

define como no probabilística obtenida por conveniencia durante el periodo de estudio, se consideró a todas las acciones de la empresa Intel Corporation las mismas que se vendieron desde mayo del 2018 hasta octubre del 2019, se obtuvieron , 410 registros.
Las técnicas estadísticas aplicadas para el procesamiento y obtención de modelos luego de la recolección de datos fueron: la regresión lineal múltiple y los árboles de regresión, utilizando la metodología machine learnign.

3 Resultados

Para obtener el árbol de regresión se particiono la data en datos de entrenamiento y validación, quedando 80% para entrenamiento y 20% para validación, siendo el total de la muestra 410 registros, de ellos, 328 para entrenamiento y 82 para validación.

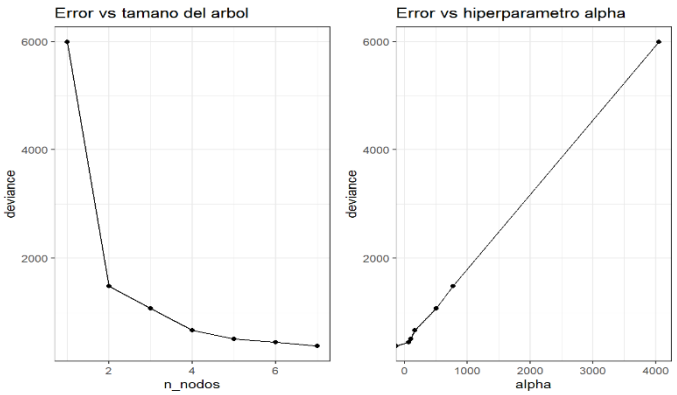

Figura 4. Número de nodos para el entrenamiento

El número de nodos para establecer el Árbol de regresión, en 7, que nos garantiza la no presencia de mucha variación en cuanto al error, observando la desviación deben ser muy pequeños. También la finura nos muestra que podríamos haber tomado 4 nodos.

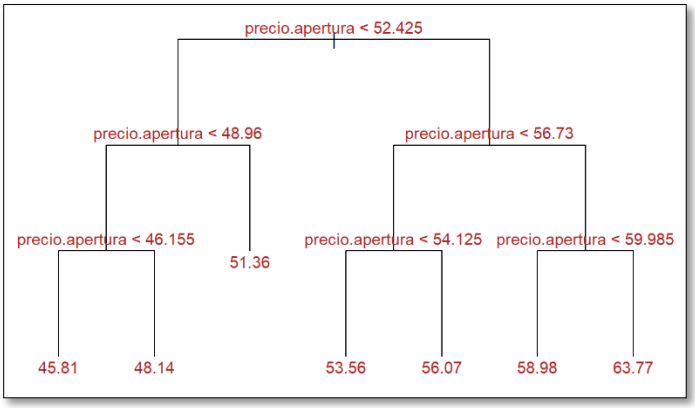

Fig. 5. Árbol de Regresión con 7 nodos terminales.

Se utilizó el software R, que presenta una solución para el Precio de apertura de las acciones de Intel.

Interpretación del modelo.
Intel alcanza el siguiente el precio máximo de acciones:

- $63.77 cuando el precio de apertura es mayor a $59.985, mayor a $56.73 y mayor a $52.425.
- $58.98 cuando el precio de apertura es menor a $59.985, mayor a $56.73 y mayor a $52.425
- $56.07 cuando el precio de apertura es mayor a $54.125, menor a $56.73 y mayor a $52.425
- $53.56 cuando el precio de apertura es menor a $54.125, menor a $56.73 y mayor a $52.425
- $51.36 cuando el precio de apertura es mayor a $48.96 y menor a $52.425
- $48.14 cuando el precio de apertura es mayor a $46.155, menor a $48.96 y menor a $52.425
- $45.81 cuando el precio de apertura es menor a $46.155, menor a $48.96 y menor a $52.425

La fig. 6. muestra a continuación corresponde al algoritmo podado.

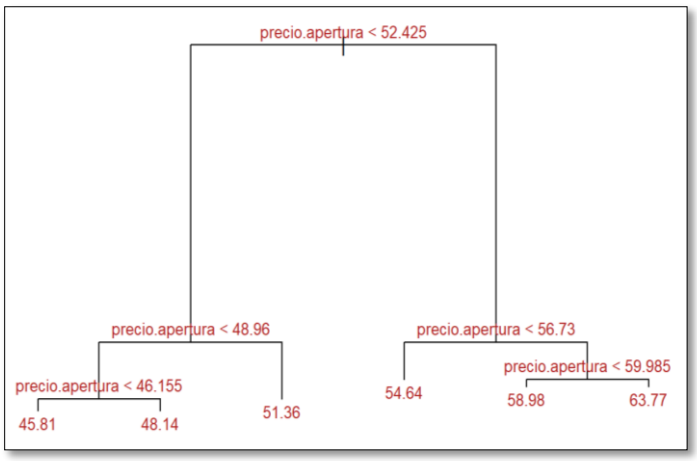

Fig. 6. Representación del Árbol de regresión podado con el algoritmo pruning.

Tabla 1. Comparación de errores para el Árbol de regresión inicial y podado.

Comparación	Árbol de Regresión	
	Sin podar	Podado
MSE	1.448015 $	1.61 $

Se observa en la tabla 1, que el valor de MSE sin podar (1.44) es menor que el árbol podado(1.61), por lo tanto, elegimos la solución sin podar.

Para el modelo de regresión lineal múltiple. A partir de los 410 de registros recopilados a partir del Nasdaq, R nos muestra los coeficientes de β_0, β_1 y β_2 para el primer modelo y las estimaciones queda así:

$$Y = 0.1454 + 1.00100X_1 + 0.00000001492X_2 \quad (4)$$

Las dos variables independientes precio de apertura y volumen resultan ser significativos p<α (0.05) excepto el intercepto. El modelo en su conjunto observando la prueba F es significativa y la bondad de ajuste es bastante alta $R^2 = 0.9934$. lo que indica que, es capaz de explicar el 99,34% de la variabilidad Observada en el precio máximo. El p-valué del modelo es significativo 2.2e-16 por lo que, se puede aceptar que el modelo no es por azar, al menos uno de los coeficientes parciales es distinto de 0.

Comprobación de supuestos:

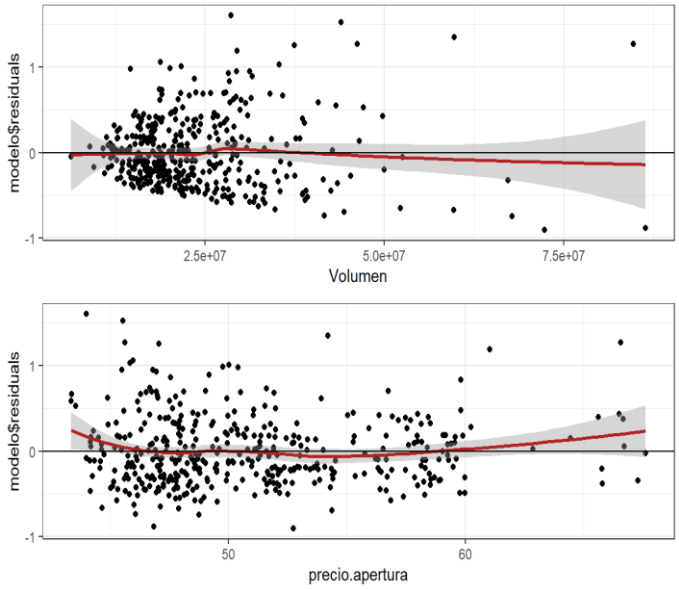

Fig. 7. Linealidad de las variables precio apertura y volumen de las acciones Intel Corporation, se observa el supuesto de linealidad de las variables independientes X's, que evidentemente, al estar la línea dentro de los intervalos se observa linealidad para el precio de apertura y el volumen de ventas de las acciones del año 2019.

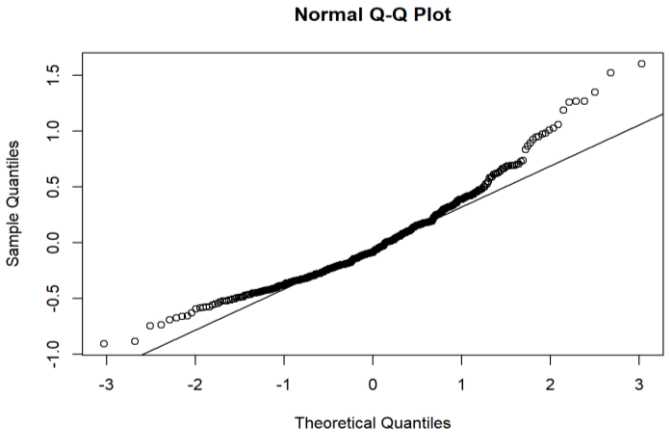

Fig. 8, se observa que los datos presentan normalidad de errores, en el apartado siguiente se realiza el supuesto de normalidad.

Mediante la prueba de Shapiro Wilks, se concluye por p(0.0000000002096)<α(0.05), en este caso se acepta la hipótesis Ha, es decir, no existe normalidad en los errores, y se rechaza la hipótesis H_0.

La prueba de homocedasticidad (test de Breusch Pagan), se concluye, r p(0.00000000000000022)<α(0.05), se acepta la hipótesis Ha, es decir, no existe homocedasticidad en las variables.

Prueba de autocorrelación (Durwin Watson), se concluye p(0.176)>α(0.059), entonces, existe autocorrelación entre las variables, es decir se acepta H0, y se rechaza Ha.
En resumen, se tiene los siguientes resultados: no existe cumplimiento de linealidad, normalidad y Homocedasticidad al contrario de autocorrelación que si cumple.
A pesar del incumplimiento de supuestos, el modelo arroja un R^2 alto 0.9934 y el ANVA significativo.
A continuación, se realiza un análisis de puntos Outliers.

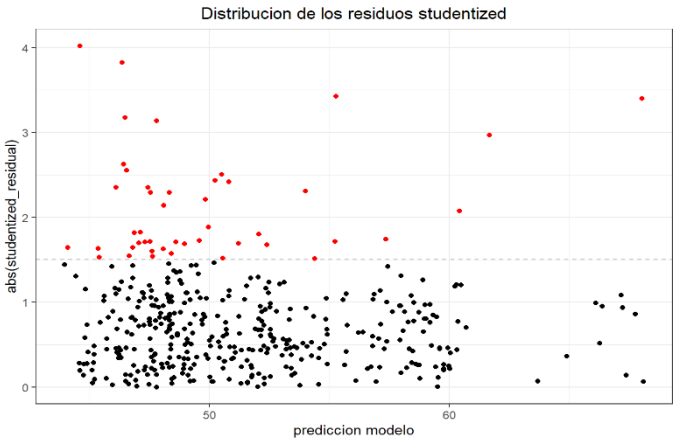

Fig. 9, Se aprecia muchos valores atípicos. Usaremos un corte de 1.5σ. los cuales afectan potencialmente a la estimación de los parámetros, por ende, se realizó la limpieza de datos, con el método de eliminación de puntos Outliers

Obtención del segundo modelo de RLM, para este segundo modelo, luego de haber utilizado el método de eliminación de puntos Outlier queda así:

$$Y = 0.01750 + 1.00300X_1 + 0.00000001199X_2 \quad (5)$$

Las dos variables independientes precio de apertura y volumen resultan ser significativas p<α (0.05) excepto el intercepto. El modelo en su conjunto observando la prueba F es significativa y la bondad de ajuste es bastante alta R^2=0.9968. El error estándar residual este disminuye en comparación al modelo anterior (0.4072 frente a 0.2849 del modelo 2).

Comprobación del supuesto de normalidad:
Mediante la prueba de Kolmogorov Smirnov se determinó, p=0.02752<α=0.05, en este caso se acepta la hipótesis Ha, es decir, no existe normalidad en los errores, y se rechaza la hipótesis H_0.

C&T Riqchary
Revista de Investigación en Ciencia y Tecnología
ISSN: 2810-8124 (en línea) / ISSN: 2706-543x

Vol. 5 Núm. 1 (2023)
Publicado: 15/10/23
DOI: https://doi.org/10.57166/riqchary/v5.n1.2023.117

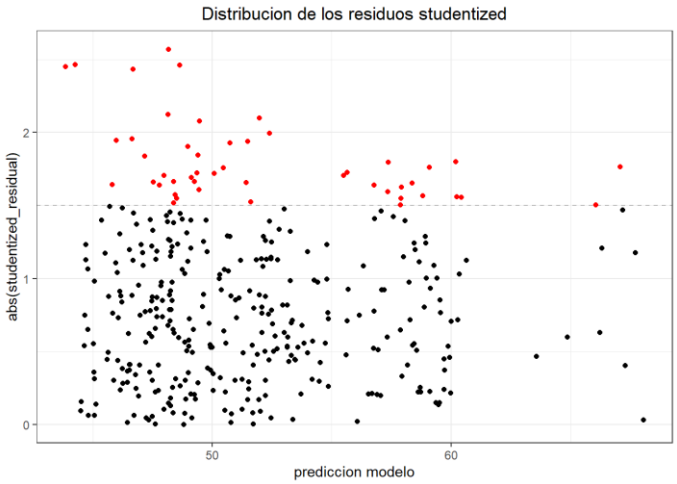

Fig. 10. Gráfico de datos Estandarizados mayores a 1.5 sigma, los valores son excluidos para obtener mejores resultados.

A partir de la limpieza de datos, el modelo de regresión lineal múltiple, queda definido como:

$$Y = 0.02856 + 1.003X_1 + 0.000000009405X_2 \quad (6)$$

Las dos variables independientes precio de apertura y volumen resultan ser significativas p<α (0.05) excepto el intercepto. El modelo en su conjunto observando la prueba F es significativa y la bondad de ajuste es bastante alta R^2=0.9979. El error estándar residual este disminuye en comparación al modelo anterior 0.2849 frente a 0.2257 del modelo 3.

Interpretación del modelo:

$$Y = 0.02856 + 1.003PrecioApert. + 0.000000009405Volum \quad (7)$$

β_1: Por cada unidad monetaria del precio de apertura en el cual la acción es abierta al mercado, el precio máximo de venta en un día, se ve incrementada en 1.003.
β_2: Por cada venta que se realiza el precio máximo de venta de las acciones de Intel, incrementa en 0.000000009405.

Comprobación de los supuestos del tercer modelo de regresión (7).

Linealidad, normalidad, homocedasticidad, autocorrelación análisis de inflación de varianza, cumple todos.

Tabla 2. Resumen de errores de la Regresión lineal múltiple.

Resumen	Regresión Lineal Múltiple		
	Primer modelo	Segundo modelo	Tercer modelo
R^2	0.9934	0.9968	0.9979
Estándar error residual	0.4072	0.2849	0.2257

Determinación de la mejor opción entre Árbol de regresión o Regresión Lineal Múltiple:

Tabla 3, muestra el resumen de los errores MSE y R^2 de Regresión lineal múltiple con la eliminación de puntos atípicos Outlier, obteniendo de esta forma un R^2 muy bueno 0.9979 lo cual indica que el 0.99% de los datos se ajustan al modelo y el Error estándar residual con $ 0.2257, seguidamente se muestra el algoritmo Árbol de regresión, para la validación sin podado se encontró un MSE de $1.4480; y finalmente se concluye que la Regresión lineal múltiple tiene menor error frente al Árbol de regresión.

Tabla 3. Resumen de errores para la Regresión lineal y Árbol de regresión.

Algoritmo	Regresión lineal múltiple	Árbol de regresión
	Tercer Modelo	Validación sin podado
R^2	0.9934	_____
Error Estándar/MSE	$ 0.2257	$ 1.448015

4 Discusiones

Lizares [1] comparo la Regresión logística binaria y Arboles de clasificación (CHAID) para evaluar el rendimiento académico. Concluye que, Según la evaluación de la clasificación de los modelos optamos por la Técnica de Árboles de clasificación, siendo la más óptima por tener mayor Sensibilidad=77,6% AUC=90,1%, Gini =80,2% y Kappa=0,589. A diferencia de esta tesis, en la que se empleó el Árbol de regresión por la presencia de variables cuantitativas, se encuentro mayor MSE $1.448015 comparado al Error Estándar Residual de la RLM.

También Espinosa [15] quien concluye que realizando un análisis técnico de los datos históricos del valor de las acciones durante un plazo determinado, es posible predecir el cambio de precio de una acción, con la finalidad de ayudar al inversor al momento de tomar la decisión de compraventa trading de sus acciones, utilizando para esto regresión múltiple, lo que concuerdan con los resultados del presente trabajo de investigación al tener la regresión lineal múltiple un Error Estándar de 0.2257, se confirma que el modelo de Regresión múltiple es buen candidato al momento de enfrentarse a un problema de predicción de variables continuas.

Confrontando con Sepúlveda [16] que, en su estudio comparo árbol regresión CART y Regresión Lineal, encontró que, el error predictivo de la regresión lineal siempre es menor que el del CART. Aconteció algo similar en esta tesis, en la que adicionalmente se emplearon técnicas estadísticas como la eliminación de puntos Outlier para la Regresión lineal múltiple

y la poda en el árbol de regresión, sin embargo, se concluyó que el modelo de regresión lineal múltiple tiene menor error frente el árbol de regresión para predecir el precio máximo en función al precio de apertura y el volumen de ventas de Intel.

5 CONCLUSIONES

Después de varias etapas desarrolladas finalmente se concluye que el mejor modelo para predecir el precio máximo de acciones de Intel es el modelo de regresión lineal Múltiple con eliminación de Outliers, debido a que este presenta menor error frente a un árbol de regresión. Este modelo ayudara a reducir la incertidumbre en los accionista o Inversionistas de la empresa Intel Corporation.

Referencias

[1] M. Lizares, Comparación de los modelos de clasificación: regresión logística y árboles de clasificación para evaluar el rendimiento académico., 2017.

[2] A. García De Mendoza Ortega, "Intel," Universidad Jesuista de Guadalajara, Guadalajara, 2022.

[3] R. Giménez Fernández and . P. Zamorano, "Modelos predictivos de índices bursátiles relevantes para la economía chilena," Universidad de Chile, Santiago, 2014.

[4] M. L. Alvarez, "Predicción de afinidad de uni´on de ligandos en protéinas," Universidad de Buenos Aires, Buenos Aires, 2016.

[5] D. I. Candia Oviedo, "Predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático," Universidad Nacional San Antonio Abad del Cusco, Cusci, 2019.

[6] J. Bacallao Gallestey and J. M. Parap, "Árboles de regresión y otras opciones metodológicas aplicadas a la predicción del rendimiento académico," *Educación Médica Superior*, 2004.

[7] S. Rosales Heredia, C. Bruno and . M. Balzarini, "Identificación de relaciones entre rendimientos y variables ambientales vía árboles de clasificación y regresión (CART)," *Interciencia*, vol. 35, no. 12, pp. 876-882, 2010.

[8] S. C. S. Pineda, Comparación de árboles de regresión y clasificación regresion logística, 2009.

[9] J. Felipe Díaz and J. Carlos Correa, "Comparación entre árboles de regresión CART," *Comunicaciones en Estadística*, vol. 6, no. 2, pp. 175-195, 2013.https://doi.org/10.15332/s2027-3355.2013.0002.05

[10] IBM, "IBM Cognos Analytics," 2022. [Online]. Available: https://www.ibm.com/docs/es/cognos-analytics/11.1.0?topic=tests-multiple-linear-regression. [Accessed 12 10 2022].

[11] J. Amat Rodrigo, "cienciadedatos.net," https://cienciadedatos.net/documentos/25_regresion_lineal_multiple, 2016. [Online]. Available: https://cienciadedatos.net/documentos/25_regresion_lineal_multiple. [Accessed 2 10 2022].

[12] D. R. Tobergte and S. Curtis, "Introducción a la econometría: Un enfoque moderno," *Journal of Chemical Information and Modeling*, vol. 53, no. 9, 2013.

[13] M. Torres, K. Paz and F. Salazar, "Métodos de recolección de datos para una investigación," no. 2, pp. 2-21, 2019.

[14] R. Hernadez Sampieri, C. Fernández Collado and P. Batista Lucio, Metodología de la Investigación, Cuarta ed., Mexico: MACGRAW-HILL, 2010.

[15] J. I. Espinosa Muñoz, "na aproximación a la predicción del valor de acciones en la bolsa de valores aplicando técnicas de Data Mining," Universidad Politécnica de Madrid, Madrid, 2015.

[16] J. . F. D. Sepúlveda Díaz and J. C. Correa Morales, "Comparación entre árboles de regresión CART y regresión lineal," *Comunicaciones en Estadística 6.2*, pp. 175-195, 2013. https://doi.org/10.15332/s2027-3355.2013.0002.05

Biografías

Edgar Eloy Carpio Vargas, Dr. en estadística e informática, especialista en ciancia de datos, docente de la facultad de Ingenieria Estadistica e Informática, docente RENACYT.

Alicia Roxana Mayda Huanca, Ing Estadistico e informatico de la Universidad Nacional del Altiplano.

Espinoza Rivas German Rafael, especialista en geología, geotecnia y medio ambiente, con estudios de Ingeniería Geológica, Ingeniería Civil, y Topografía en universidades del Perú, maestría en Ingeniería Civil y Medio Ambiente en la Universidad de Utah, USA y segunda maestría en Ingeniería Ambiental en la Universidad Nacional del Altiplano. Doctor en Ciencia, Tecnología y Medio Ambiente.

Ecler Mamani Vilca, Universidad Nacional Micaela Bastidas de Apurímac - Perú, Dr. en Ciencias de la Computación, desarrolla-dor de aplicaciones multimedia y Software Educativo Intercultural.

Betsabe M. Ccolqque Ruiz, docente en la Universidad Nacional Micaela Bastidas de Apurímac, Ingeniero Informático y Sistemas con Magister en SEGURIDAD DE LA INFORMACIÓN Y TECNOLOGÍA y estudiante de doctorado.