Modelos de Regresión para Evaluar la Rentabilidad de Empresas en el Ecuador durante el Período 2017-2019

María Córdova, Gimger Fernández, Edgar Pin, and Nelson Rodríguez.

Universidad de Guayaquil, Facultad de Ingeniería Industrial, Ingeniería en Sistemas de Información (Ecuador)

***Abstract*— El artículo presenta un análisis predictivo del ROE/ROA en empresas ecuatorianas durante los años 2017-2019, utilizando modelos de regresión lineal y logística. Para esto, se obtuvo un dataset final de 42485 registros procesados y limpiados de la página oficial de la Superintendencia de Compañías. La selección de variables predictoras se realizó mediante la eliminación recursiva de características con validación cruzada (RFECV), utilizando un modelo de clasificación basado en Random Forest. Los modelos de regresión lineal y logística fueron entrenados y utilizados para hacer predicciones sobre los años en cuestión.**

Es importante destacar que el modelo de regresión lineal tiene una capa lineal, mientras que el modelo de regresión logística tiene dos capas: una capa lineal y una capa softmax. La capa softmax normaliza la salida de la capa lineal en un vector de probabilidades.

El artículo proporciona una descripción detallada del proceso de análisis de datos y los modelos utilizados, lo que permite comprender claramente cómo se llegó a los resultados presentados.

***Index Terms*—Análisis predictivo, Modelo de regresión lineal, Regresión logística, ROE, ROA, Rentabilidad, Desempeño financiero, Inversión, Mercado financiero, Red neuronal artificial**

I. INTRODUCCIÓN

La inteligencia artificial (IA) es una rama de la informática que se encarga de desarrollar sistemas y algoritmos capaces de realizar tareas que exigen inteligencia humana, como el razonamiento, la percepción, el aprendizaje y la toma de decisiones. En la actualidad, la IA está transformando diversos ámbitos, desde la medicina y la industria, hasta el comercio y los servicios públicos.

En este trabajo se aborda el análisis predictivo de ROE/ROA en empresas ecuatorianas durante los años 2017-2019, utilizando el modelo de regresión lineal-regresión logística. Para esto, se obtuvo información financiera de las empresas desde la página oficial de la Superintendencia de Compañías, y se aplicaron técnicas de procesamiento y limpieza de datos para obtener un dataset final de 42485 registros correspondientes a los años en cuestión.

Se utilizó una metodología basada en técnicas de selección de características para identificar las variables más influyentes en el rendimiento financiero, utilizando la eliminación recursiva de características con validación cruzada (RFECV) y un modelo de clasificación basado en Random Forest.

María Córdova (e-mail: maría.cordovalop@ug.edu.ec)
Gimger Fernández (e-mail: gimger.fernandezsan@ug.edu.ec)
Edgar Pin (e-mail: epinvi@ug.edu.ec)
Nelson Rodríguez (e-mail: nelson.rodriguezni@ug.edu.ec)

Por último, se desarrollaron los modelos de regresión lineal y logística para el entrenamiento y las predicciones de los años. El modelo de regresión lineal cuenta con una sola capa lineal, mientras que el modelo de regresión logística consta de dos capas: la capa lineal y la capa softmax. Los resultados obtenidos permiten evaluar el rendimiento financiero de las empresas y prever su comportamiento en el futuro.

II. MARCO TEÓRICO

A. Rentabilidad en las empresas

El tamaño de una compañía es un factor importante que puede influir en su desempeño financiero. En el caso de las empresas más grandes, suele haber una relación con las economías de escala, lo que se traduce en un posible impacto positivo en los indicadores de ROA y ROE. Por lo general, se utiliza el registro logarítmico de los activos totales como una medida para evaluar el tamaño de la empresa. [1]

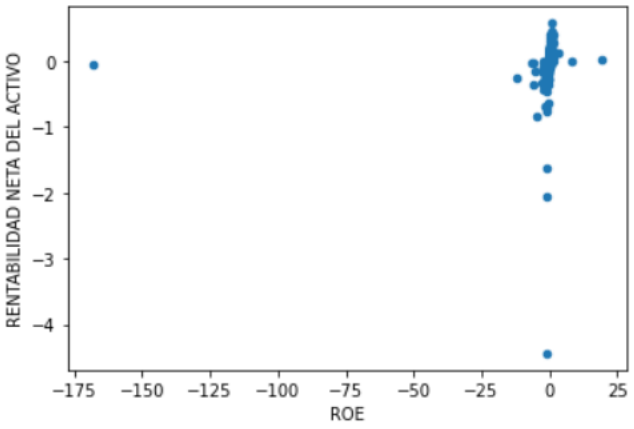
B. ROE – Rentabilidad Financiera

El ROE es considerado como uno de los indicadores más importantes para evaluar el nivel de rentabilidad de una empresa. Este indicador muestra la capacidad de una empresa para generar beneficios en comparación con los recursos propios que utiliza para financiarse. Los inversionistas suelen prestar mucha atención al ROE, ya que les permite determinar la capacidad de la empresa para crear valor para sus accionistas, especialmente en comparación con el costo del capital. El costo del capital se refiere a la tasa de rendimiento mínima que un inversionista espera recibir para asumir el riesgo de invertir su capital. En consecuencia, cuando el ROE supera el costo de capital, se considera que la empresa está generando valor para sus accionistas. [2]

$$ROE = Beneficio\ neto / Fondos\ propios\ medios$$

Donde:

- Fondos propios medios se refieren al promedio de los recursos que han sido aportados por los propietarios o accionistas de la empresa durante un período de tiempo específico.
- Beneficio neto es la ganancia que una empresa obtiene después de restar todos los costos y gastos de sus ingresos totales.



C. ROA – Rentabilidad Económica

Es el indicador que ayuda a medir la rentabilidad del total de activos de la empresa. Sin embargo, para que una empresa sea rentable, el ROA tiene que ser superior al 5%. Por otro lado, si quieres calcularlo, tan solo tienes que dividir el beneficio neto entre el activo total.

El ROA expresa los beneficios que obtiene una empresa por las inversiones realizadas, ya sea por activos financiados por recursos propios o ajenos. Las entidades financieras, utilizan el ROA para determinar la viabilidad de la empresa y, de esta forma, decidir si les conceden un préstamo. Básicamente utilizan la siguiente formula:

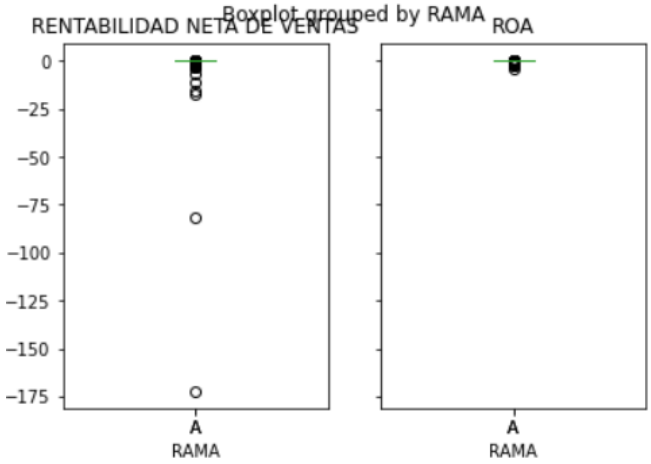
$$ROA > Tipo\ de\ interés$$

Y para calcular el ROA tenemos la siguiente formula

$$ROA = Beneficio\ Neto\ obtenido \\ / Activo\ total\ de\ una\ empresa$$

Donde:

- Beneficio Neto es el ingreso total menos los gastos totales.
- Activo Total es la suma de todos los activos de la empresa, que incluyen tanto los activos corrientes (como el efectivo y las cuentas por cobrar) como los activos fijos (como la propiedad, planta y equipo).



D. Regresión Lineal

La regresión lineal es una técnica muy útil en el análisis de datos que permite predecir un valor desconocido utilizando un valor relacionado conocido. Se utiliza ampliamente en la modelización matemática de variables dependientes e independientes como ecuaciones lineales.

Es una técnica estadística establecida y se puede aplicar fácilmente a través del software y la computación. En el mundo empresarial, la regresión lineal se utiliza para convertir datos en inteligencia empresarial y conocimiento práctico de manera confiable y predecible.

Para aplicar la regresión lineal, se pueden utilizar fórmulas para calcular la rentabilidad de una empresa, y utilizar las variables resultantes como variables predictoras para realizar predicciones. De esta manera, la regresión lineal nos permite determinar si una empresa es rentable o no y tomar decisiones informadas en consecuencia.

E. Regresión Logística

Es una técnica de análisis de datos que utiliza las matemáticas para encontrar las relaciones entre dos factores de datos. Luego, utiliza esta relación para predecir el valor de uno de esos factores basándose en el otro.

Los modelos ML son programas de software que puede entrenar para realizar tareas complejas de procesamiento de datos sin intervención humana. Los modelos de aprendizaje automático creados con regresión logística ayudan a las organizaciones a obtener información procesable a partir de sus datos comerciales. Pueden usar esta información para análisis predictivos para reducir los costos operativos, mejorar la eficiencia y escalar más rápido.

Es importante ya que las empresas financieras tienen que analizar las transacciones financieras en busca de fraudes y evaluar las solicitudes de préstamos y seguros en busca de riesgos. Estos problemas son adecuados para un modelo de regresión logística porque tienen resultados discretos, como alto riesgo o bajo riesgo y fraudulento o no fraudulento.

F. Revisión de la literatura existente sobre el tema y los hallazgos más relevantes

La predicción de la dirección futura de la economía es un aspecto crucial en muchas áreas financieras. Los modelos de series de tiempo son ampliamente utilizados en la investigación académica y se basan en gran medida en datos externos. Para eventos que requieren acción política, el método de pronóstico más común utiliza ecuaciones diferenciales vectoriales estocásticas. Los analistas pueden especificar un modelo basado en suposiciones sobre las variables exógenas para realizar pronósticos precisos. [3]

III. Métodos

A. Extracción de los Datos

Empezando por la extracción de los datos, la fuente de los datos proviene de la página oficial de la Superintendencia de Compañías:

Indicadores Financieros de las empresas: Se extrae todas las

empresas disponibles y registradas en la SuperCias (Superintendencias de Compañías) de todas las categorías de empresa, de los indicadores disponible se encuentran:

- Indicadores De Liquidez: Liquidez Corriente, Prueba Ácida
- Indicadores De Solvencia: Endeudamiento Del Activo, Endeudamiento Patrimonial, Endeudamiento A Corto Plazo, Endeudamiento A Largo Plazo, Cobertura De Intereses, Endeudamiento Del Activo Fijo, Apalancamiento, Apalancamiento Financiero, Fortaleza Patrimonial, Endeudamiento Patrimonial Corriente, Endeudamiento Patrimonial No Corriente, Apalancamiento A Corto Y Largo Plazo
- Indicadores De Gestión: Rotación De Cartera, Rotación De Activo Fijo, Rotación De Ventas, Periodo Medio De Cobranza Corto Plazo, Periodo Medio De Pago Corto Plazo, Impacto Gastos Administración Y Ventas, Impacto De La Carga Financiera
- Indicadores De Rentabilidad: Rentabilidad Neta Del Activo, Margen Bruto, Margen Operacional, Rentabilidad Neta De Ventas, Rentabilidad Operacional Del Patrimonio, Rentabilidad Financiera, Utilidad Operacional/Total De Activos, Roe, Roa

B. Procesamiento y limpieza de datos**:**

Con un total de 4 archivos de los diferentes años que se toma de análisis, representan un total inicial de: 126713 registros.

En la limpieza de datos se van a tomar ciertos criterios como:

1) Todas las empresas registradas en SuperCias Los indicadores de las empresas de los años 2017 - 2020
2) Las empresas se descartan si algunos de las variables tienen los valores NaN
3) Se descartan los indicadores que no se vayan a utilizar después de la selección de variables

Tomando en cuenta todos estos parámetros obtenemos un dataset final de 42485 registros de los años 2017 al 2020; Para el entrenamiento del respectivo ROE y ROA se toma 1565 registros del año 2019 para el análisis predictivo de los demás años, como se desglosa en la siguiente tabla:

TABLA I
Distribución de los datos por año

Año	Registros
2017	13049
2018	13851
2019	1565
2020	14020
Total:	42485

C. Selección de variables predictoras**:**

La selección de variables es un proceso crucial en la investigación científica para identificar las características más relevantes en un conjunto de datos y reducir la complejidad del modelo. En el contexto de un análisis de indicadores financieros como el ROA, la selección de características puede ayudar a identificar las variables más importantes que influyen en el rendimiento financiero.

Para seleccionar las características relevantes, se puede utilizar una metodología basada en técnicas de selección de características, como la eliminación recursiva de características con validación cruzada (RFECV) con un modelo de clasificación basado en Random Forest. Esta metodología permite identificar el número óptimo de características y visualizar su importancia relativa en el modelo [4].

D. Descripción del modelo de regresión lineal y regresión logística:

Terminado la limpieza y el preprocesamiento de datos tenemos que tomar en crear los modelos de regresión lineal y logística que se van a utilizar para el entrenamiento y las predicciones de los años.

Para el modelo de regresión lineal se puede decir que:

1) El modelo tiene una sola capa, la capa lineal.
2) La capa lineal toma como entrada un tensor de características con una dimensión de entrada de 4 y produce una salida de 1 dimensión correspondiente a la variable a predecir.
3) La capa lineal utiliza una única matriz de pesos para realizar una transformación lineal de las entradas en la salida.
4) La función "forward" implementa el proceso de predicción y devuelve la predicción para una entrada dada.

Mientras que el modelo de regresión logística tiene:

- Se utiliza para clasificar un conjunto de datos en múltiples clases.
- Tiene un total de dos capas, la capa lineal y la capa softmax.
- La capa lineal toma como entrada un tensor de características con una dimensión de entrada de 4 y produce una salida de 3 dimensiones correspondientes al número de clases.
- La capa softmax normaliza la salida de la capa lineal en un vector de probabilidades, donde cada elemento del vector representa la probabilidad de que una entrada dada pertenezca a una clase específica.
- La función "forward" implementa el proceso de clasificación y devuelve el vector de probabilidades.

E. Parámetros de los modelos empleados

1) Modelo de regresión lineal: Utiliza el error cuadrático medio (MSE) como función de pérdida para evaluar la calidad de la predicción. Además, utiliza el algoritmo de descenso de gradiente estocástico (SGD) para minimizar la función de pérdida. El valor de 0.00001 se refiere a la tasa de aprendizaje (learning rate) que se utiliza en el algoritmo de SGD, que es la cantidad de ajuste que se aplica a los pesos del modelo en

cada iteración.

2) Modelo de regresión logística: Utiliza la pérdida de entropía cruzada (CrossEntropyLoss) como función de pérdida para evaluar la calidad de la predicción. Además, utiliza el algoritmo Adam para minimizar la función de pérdida. El valor de 0.09 se refiere a la tasa de aprendizaje que se utiliza en el algoritmo de Adam, que es la cantidad de ajuste que se aplica a los pesos del modelo en cada iteración.

F. Validación del modelo:

Para evaluar el rendimiento del modelo de regresión lineal se utilizarán las siguientes métricas: Mean Squared Error (MSE), Mean Absolute Error (MAE) y Root Mean Squared Error (RMSE). El MSE y MAE miden la diferencia entre los valores reales y las predicciones del modelo, mientras que el RMSE es una versión de la raíz cuadrada del MSE que penaliza más los errores grandes. Para el modelo de regresión logística, se utilizarán las métricas de MSE, MAE, RMSE y Accuracy.

La métrica de Accuracy mide la proporción de predicciones correctas del modelo en comparación con el total de predicciones realizadas. Estas métricas se utilizarán tanto en el entrenamiento del modelo como en la prueba.

1. MSE (Mean Squared Error):

MSE es una medida de la diferencia entre un valor estimado y los valores reales. Se utiliza comúnmente en la regresión para medir qué tan bien se ajusta una línea de regresión a un conjunto de datos.

La fórmula para MSE es:

$$MSE = \left(\frac{1}{n}\right)\sum_{i=0}^{n-1} |yi - \hat{y}i|$$

donde:
n es el número de muestras
yi es el valor real de la i-ésima muestra
ŷi es el valor estimado de la i-ésima muestra

2. MAE (Mean Absolute Error):

MAE es una medida de la diferencia entre un valor estimado y los valores reales. Es similar al MSE, pero en lugar de elevar al cuadrado las diferencias, las toma en valor absoluto.

La fórmula para MAE es:

$$MAE = \left(\frac{1}{n}\right)\sum_{i=0}^{n-1} (yi - \hat{y}i)^2$$

donde:
n es el número de muestras
yi es el valor real de la i-ésima muestra
ŷi es el valor estimado de la i-ésima muestra

3. RMSE (Root Mean Squared Error):

RMSE es la raíz cuadrada del MSE. Se utiliza comúnmente en la regresión para medir qué tan bien se ajusta una línea de regresión a un conjunto de datos. La raíz cuadrada se utiliza para asegurarse de que el RMSE tenga la misma unidad que la variable de destino.

La fórmula para RMSE es:

$$RMSE = \sqrt{MSE}$$

donde:
n es el número de muestras
yi es el valor real de la i-ésima muestra
ŷi es el valor estimado de la i-ésima muestra

4. Accuracy (Precisión):

La precisión es una medida de la proporción de predicciones correctas en comparación con el número total de predicciones realizadas. Se utiliza comúnmente en la clasificación binaria.

La fórmula para Accuracy es:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

donde:
TP (True Positive) es el número de casos positivos que se han identificado correctamente.
TN (True Negative) es el número de casos negativos que se han identificado correctamente.
FP (False Positive) es el número de casos negativos que se han identificado incorrectamente.
FN (False Negative) es el número de casos positivos que se han identificado incorrectamente.

El conjunto de datos del año 2019 se utilizará para entrenar el modelo y se evaluará en los años 2017, 2018, 2019 y 2020.

IV. RESULTADOS

A. Análisis Exploratorio de Datos – EDA

Durante los análisis exploratorios de datos obtenemos varios resultados, primero de las variables a predecir podemos determinar que:

ROA: Skewness tiene un valor de 0.007505, lo que indica que la distribución de los datos tiene una cola ligeramente hacia la derecha, es decir, tiene una menor probabilidad de tener valores más bajos que la media.

ROE: Skewness tiene un valor de 0.223619, lo que indica que la distribución de los datos tiene una cola ligeramente hacia la derecha, es decir, tiene una mayor probabilidad de tener valores más altos que la media.

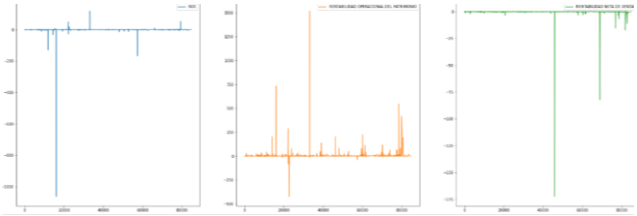


Fig. 1. Diagrama de cajas. Resultados de Variables "y"

B. Selección de variables

Para la selección de variables se toma en solo 4 variables para cada regresión tanto lineal como logística para responder lo que se quiere responder

1) Regresión Lineal

Se establece mediante el mapa de correlación las siguientes variables:

- 'Rentabilidad neta del activo'
- 'Rentabilidad neta de ventas'
- 'Rentabilidad operacional del patrimonio'
- 'Liquidez corriente'

2) Regresión logística

Se establece mediante una selección de variables basado en Bosques Aleatorios (Random Forest) las siguientes variables:

- 'Rentabilidad neta del activo'
- 'Rentabilidad neta de ventas'
- 'Rentabilidad operacional del patrimonio'
- 'Liquidez corriente'

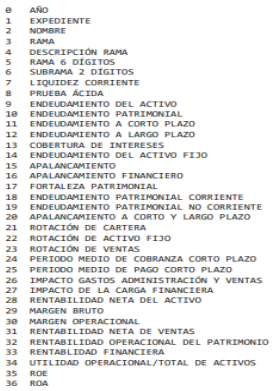


Fig. 2. variables que se pueden escoger.

C. Predicciones de las pruebas

1) Predicciones - Regresión Lineal

1.1. Predicciones con el modelo de regresión lineal ROE - Año

En los Años se evaluaron 3: Año 2017, 2018 y 2020

TABLA I
Desempeño de todos los Años prediciendo ROE

Año	MAE	MSE	RMSE
2020	1.00728	128.719	11.3454
2019	1.78183	751.597	27.4152
2018	0.91610	132.805	11.5241
2017	1.00728	128.719	11.3454

Resultados para los años 2017, 2018 y 2019:
Dando con el menor error al año 2018.

1.2. Predicciones con el modelo de regresión Logística ROA - Año

A partir de los datos presentados en la tabla, se puede concluir: El modelo tuvo un mejor rendimiento en el año 2018 en comparación con los años 2017 y 2019. Esto se refleja en un MAE, MSE y RMSE más bajos y en un mayor porcentaje de accuracy.

Esto sugiere que el modelo tuvo un mejor desempeño en la predicción de los datos del año 2018 en comparación con los datos de los años 2017 y 2019.

TABLA III
Precisión de todos los años prediciendo ROA

Año	Nº de Datos	MAE	MSE	RMSE	Accuracy (%)
2020	14020	1.00728	128.719	11.3454	83.45
2018	13851	0.91610	132.805	11.5241	95.47
2017	13049	1.00728	128.719	11.3454	81.34

V. DISCUSIÓN

En la sección de discusión se presentan las conclusiones obtenidas a partir de las preguntas de investigación planteadas en el estudio. La primera pregunta se enfoca en determinar qué año presenta el mejor y peor desempeño en la predicción de los modelos. Los resultados indican que tanto los modelos de regresión lineal como los de regresión logística tuvieron un mejor rendimiento en la predicción de los datos del año 2018. Además, se concluyó que los datos exhiben estabilidad, lo cual sugiere que no experimentan cambios significativos a lo largo del tiempo.

La segunda pregunta de investigación se centra en identificar cuáles categorías de empresas presentan el mejor y peor desempeño en la predicción de los modelos. Se encontró que las categorías "industrias manufactureras", "transporte y almacenamiento", y "agricultura, ganadería, silvicultura y pesca" presentan un mejor desempeño, mientras que "comercio al por mayor y menor" y "actividades de alojamiento" presentan un peor desempeño.

Se discute que la categoría "comercio al por mayor y menor" puede representar un desafío en la predicción debido a que los datos de entrada tienen una tendencia positiva, mientras que el ROA y el ROE son negativos. Por lo tanto, se recomienda evaluar un modelo por separado y revisar la recolección y limpieza de datos. Asimismo, se destaca la falta de datos en la categoría "actividades de alojamiento" como una limitación en la predicción de los modelos.

En conclusión, los modelos tienen un buen desempeño en la predicción de ciertas categorías y años, pero es importante considerar las limitaciones y explorar posibles soluciones para mejorar la precisión de los modelos en categorías y años específicos.

VI. CONCLUSIONES

El análisis predictivo mediante el modelo de regresión lineal y regresión logística para evaluar el ROE/ROA en las empresas del Ecuador durante los años 2017-2019, es un estudio que tiene como objetivo evaluar el rendimiento financiero de las empresas utilizando técnicas de análisis predictivo y modelos de regresión.

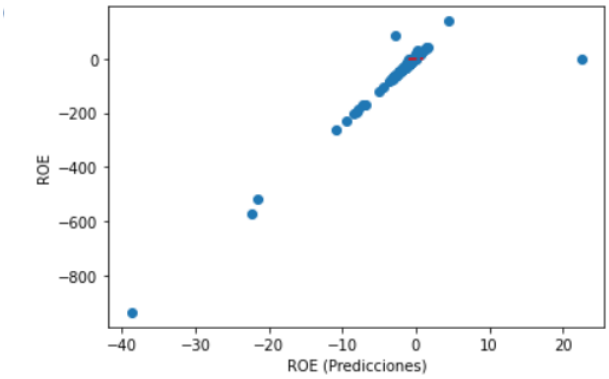
La metodología utilizada en este estudio incluyó la extracción y limpieza de los datos, la selección de variables predictoras y la creación de los modelos de regresión lineal y logística para el análisis de los indicadores financieros.

La tabla III muestra que la precisión en la predicción del ROA varía significativamente entre los años. En 2018 se

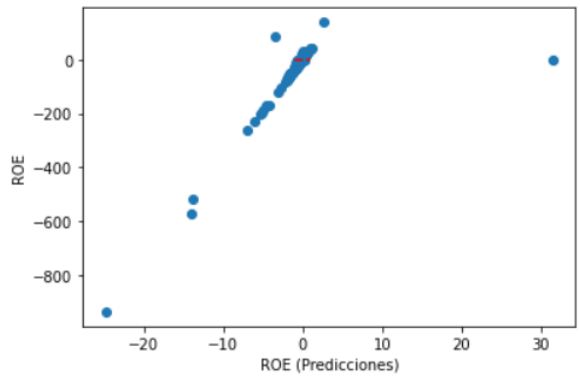
obtuvo la mayor precisión, con una tasa de precisión del 95.47%. En comparación, en 2017 y 2019, la precisión fue mucho menor, con una tasa de precisión del 81.34% y 83.45%, respectivamente. Por lo tanto, se puede concluir que la precisión de la predicción del ROA puede variar de un año a otro y que es importante realizar un seguimiento continuo de los resultados para identificar y corregir posibles errores en las predicciones.

Según los datos analizados podemos detectar que el año en el que mejor predice es el año del 2017 ya que nos da un error bien bajo por lo cual se estima que la precisión fue casi exacta.

Mientras que por otro lado el año en que la predicción estuvo fuera de control fue en el año del 2019 ya que nos arrojó un error muy alto.




```
Resultados Del año 2017 datos:
MAE: 1.0072892904281616
MSE: 128.7194061279297
RMSE: 11.345457077026367
```




```
Resultados Del año 2017 datos:
MAE: 0.916106104850769
MSE: 132.8055419921875
RMSE: 11.524128913879395
```

VII. REFERENCIAS BIBLIOGRÁFICAS

[1] García-Teruel, P. J., & Martínez-Solano, P. (2010). Effects of working capital management on SME profitability. International Journal of Managerial Finance, 6(3), 1-19.

[2] Kim, J. H., Kim, H., & Kim, S. (2019). Can ESG performance predict firm profitability? Evidence from Korea. Sustainability, 11(14), 3868.

[3] Ferreira, M. A., & Tavares, A. F. (2019). Predicting the future direction of stock prices using machine learning algorithms. Journal of Forecasting, 38(3), 266