Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Maestría en Ciencia de Datos

Predicción a largo plazo del movimiento del precio de acciones bursátiles con base en sus estados financieros

---

TRABAJO RECEPCIONAL que para obtener el **GRADO** de
Maestro en Ciencia de Datos

Presenta:
Alan Omar Topete Salazar

Director:
Dr. Jaime Emmanuel Alcalá Temores

Tlaquepaque, Jalisco, 28 de noviembre de 2024

Predicción a largo plazo del movimiento del precio de acciones bursátiles con base en sus estados financieros

Alan Omar Topete Salazar

Resumen

En este trabajo, se replican algunos resultados obtenidos por Milosevic (2016) para predecir el movimiento de precios de acciones a largo plazo. El presente estudio contrasta en que se evalúan a las empresas con mayor capitalización de mercado en Estados Unidos, a comparación del estudio original que estudiaba compañías con capitalización media del mercado americano y europeo. A través de una revisión del trabajo previo, este estudio busca mejorar la precisión de las predicciones del movimiento de precios de acciones a largo plazo utilizando técnicas avanzadas de aprendizaje automático y comparar el rendimiento con métodos previamente propuestos. Se introducen nuevos modelos y técnicas, y se comparan con el enfoque anterior para evaluar su rendimiento. Se analiza la importancia de las características utilizadas en los modelos para entender mejor qué factores influyen más en las predicciones. Finalmente se proponen recomendaciones para futuros investigadores en el área basados en los hallazgos obtenidos

Tabla de Contenidos

Índice de figuras

Página

Índice de tablas

Página

Agradezco a mis padres, Silvia y Ramón, que se esforzaron en apoyarme siempre que lo necesité.

A los amigos que me acompañaron en el camino, sin quienes el viaje que me trajo hasta aquí hubiera sido más difícil y más aburrido.

Finalmente a mis profesores y a mi asesor, de los cuales me llevo grandes aprendizajes y que siempre impulsaron en mí el deseo de aprender y superarme.

1 Introducción

1.1 Contexto

La inversión en el mercado de valores es una actividad de gran interés en la economía global. Prever el movimiento de los precios de las acciones ha sido un área de interés para inversores y académicos durante décadas. El uso del aprendizaje automático en esta área ha mostrado resultados prometedores, como se evidencia en el trabajo de Milosevic [1]. Sin embargo, al tratarse de un área para la cual no existen respuestas definitivas resulta pertinente revisar los métodos con frecuencia. El presente trabajo busca profundizar en los métodos actuales y mejorar su precisión.

1.2 Justificación

Aunque existen numerosos métodos y técnicas para predecir los precios de las acciones, sigue siendo un desafío mejorar la precisión de estas predicciones. Un enfoque mejorado no podría beneficiar a inversores individuales, a instituciones financieras y a investigadores de estos tópicos.

1.3 Problema

A pesar de los avances en técnicas de predicción de precios de acciones usando aprendizaje automático, todavía hay margen de mejora en términos de precisión y eficiencia. Se buscan métodos que aborden cómo se puede mejorar el enfoque actual propuesto por Milosevic y lograr predicciones más precisas del movimiento de precios de acciones a largo plazo.

1.4 Objetivos

1.4.1 *Objetivo general*

Aplicar métodos previamente propuestos por Milosevic para la predicción del movimiento de precios de acciones a largo plazo utilizando técnicas avanzadas de aprendizaje automático a los 500 activos componentes del índice bursátil *S&P 500* y comparar los resultados con los encontrados en el estudio previo.

1.4.2 *Objetivos específicos*

1. Replicar el estudio realizado por Milosevic para los 500 activos integrantes del índice *S&P 500* para tener una base comparativa.
2. Introducir y evaluar nuevos modelos de aprendizaje automático en la predicción de precios de acciones.
3. Comparar el rendimiento de los nuevos modelos con el enfoque propuesto por Milosevic.
4. Analizar la importancia de las características utilizadas en los modelos para entender mejor qué factores influyen más en las predicciones.
5. Proponer recomendaciones para futuros investigadores en el área basados en los hallazgos obtenidos.

2 Metodología

2.1 Descripción de los datos

El conjunto de datos utilizado en este trabajo contiene información financiera clave sobre las empresas listadas en el índice *S&P 500*. Estas variables proporcionan una visión integral de los factores que influyen en el rendimiento financiero de las compañías y su valoración en el mercado bursátil. El conjunto de datos incluye observaciones trimestrales de 2021 y 2022 para cada empresa del índice *S&P 500*. Estas observaciones corresponden a los últimos estados financieros disponibles para cada año. Los datos han sido adquiridos directamente de la *API* de Alpha Vantage [2]. Alpha Vantage es una fuente primaria y confiable en el ámbito financiero, reconocida por su rigurosa actualización y precisión en los datos proporcionados sobre acciones en diversas bolsas de valores internacionales.

El conjunto de datos consta de 18 columnas, que se describen a continuación:

1. *Book value*: Valor contable de la empresa, reflejando la diferencia entre activos y pasivos.
2. *Market capitalization*: Representa el valor de mercado total de una empresa, calculado como el precio de las acciones multiplicado por el número total de acciones en circulación.
3. *Dividend yield*: Proporción del rendimiento obtenido por los dividendos pagados por la acción.
4. *Earnings per share*: Ganancias generadas por cada acción en circulación.
5. *Earnings per share growth*: Porcentaje de aumento o disminución de las ganancias por acción respecto al año anterior.
6. *Net revenue*: Ingresos netos generados por la empresa.
7. *Sales growth*: Tasa de crecimiento de las ventas de la empresa respecto al año anterior.

8. *Price to earnings ratio*: Relación que compara el precio actual de la acción con sus ganancias.

9. *Price to book value*: Relación entre el precio de mercado de una acción y su valor contable.

10. *Price to sales ratio*: Relación entre el precio de la acción y las ventas por acción.

11. *Dividend per share*: Cantidad de dividendos pagados por acción.

12. *Profit margin*: Proporción de las ganancias netas con respecto a los ingresos totales.

13. *Operating margin*: Proporción de las ganancias operativas con respecto a los ingresos totales.

14. *Net revenue growth*: Tasa de crecimiento de los ingresos netos respecto al año anterior.

15. *Current ratio*: Relación entre activos corrientes y pasivos corrientes, indicando la capacidad de una empresa para pagar sus deudas a corto plazo.

16. *Quick ratio*: Medida de la liquidez a corto plazo de una empresa, excluyendo inventarios.

17. *Total debt to equity*: Proporción del total de deudas de una empresa con respecto a su patrimonio neto.

18. *Price*: Precio de la acción en el mercado. Esta columna no fue utilizada como característica predictiva sino que se usó para estimar el rendimiento de la acción año contra año, que posteriormente fue usado como variable objetivo.

Las variables mencionadas serán la base para los modelos predictivos que se desarrollarán en las siguientes secciones, permitiendo una evaluación del impacto del rendimiento financiero sobre el precio de las acciones.

2.2 *Análisis exploratorio*

El conjunto de datos contiene 2012 observaciones, con información sobre 18 variables financieras clave. Sin embargo, algunas columnas presentan valores faltantes, especialmente en los ratios de liquidez.

Para cada una de las variables incluidas en el conjunto de datos, se realizó un análisis de su distribución con el fin de identificar patrones anómalos. Este análisis es esencial, ya que las distribuciones sesgadas o

Variable	Media	Desv.Est.	Mínimo	Mediana	Máximo
Market cap.	9.5E+10	2.8E+11	6.1E+9	3.5E+10	3.2E+12
EPS growth	1.41	14.98	-1.00	0.06	308.28
Net revenue	3.3E+10	6.6E+10	7.6E+8	1.3E+10	6.6E+11
Sales growth	0.06	0.18	-0.57	0.04	2.62
Price to sales ratio	4.19	3.95	0.06	2.90	32.37
Profit margin	0.14	0.14	-1.16	0.13	0.70
Operating margin	0.20	0.31	-5.66	0.19	2.35
Current ratio	5.84	34.31	-1.29	1.35	648.33
Quick ratio	8.81	82.01	-475.64	1.05	1242.24
Debt to equity	0.41	11.47	-220.59	0.60	202.00
Price	164.92	287.68	7.03	101.86	5908.9

Tabla 2.1: Estadísticas Descriptivas de las Variables Clave

asimétricas pueden influir en los resultados de los modelos predictivos si no se manejan adecuadamente.

Ciertas columnas presentaron características particulares que merecen un análisis más detallado debido a su posible impacto en el rendimiento de los modelos predictivos. En particular, algunas variables exhibieron distribuciones altamente sesgadas o valores atípicos, lo que podría generar resultados distorsionados si no se abordan adecuadamente durante la fase de preprocesamiento. A continuación, se describen las métricas más relevantes que requieren una atención cuidadosa por su influencia en las predicciones:

Crecimiento de las ganancias por acción: Esta métrica presenta un sesgo positivo considerable, lo que indica que la mayoría de las empresas del *S&P 500* muestran un crecimiento moderado en las ganancias por acción, mientras que unas pocas empresas presentan crecimientos excepcionalmente altos. Este sesgo hacia valores altos puede impactar los modelos predictivos si no se corrige adecuadamente.

Ingresos netos: Los ingresos netos presentan una distribución sesgada hacia la derecha, con la mayoría de las empresas generando ingresos relativamente bajos y unas pocas compañías con ingresos extremadamente altos. Esto puede deberse a la naturaleza del índice *S&P 500*, que incluye tanto empresas medianas como grandes, afectando la distribución de ingresos.

Precio de las acciones: La variable 'Precio' muestra un sesgo positivo moderado, con un número reducido de empresas que presentan precios de acciones mucho más altos que la mediana. Esto refleja la gran variabilidad en el valor de mercado entre las empresas del índice, donde gigantes como Apple o Microsoft tienen precios de acciones significativamente más altos en comparación con otras empresas.

Crecimiento de las ventas: Similar a las ganancias por acción, el crecimiento de las ventas muestra un sesgo positivo pronunciado, indicando que la mayoría de las empresas experimentan un crecimiento

moderado en sus ventas, pero algunas pocas registran un crecimiento excepcionalmente alto, distorsionando la distribución.

Márgenes de beneficio y márgenes operativos: Ambas métricas presentan distribuciones asimétricas, con sesgos negativos. Un número significativo de empresas reporta márgenes negativos, lo que indica que algunas están operando con pérdidas o enfrentan costos operativos elevados.

Ratios de liquidez (*'Current ratio'* y *'Quick ratio'*): Ambos ratios muestran sesgos positivos extremos, lo que sugiere que, aunque la mayoría de las empresas mantienen una liquidez adecuada, algunas presentan ratios atípicamente altos, lo que podría indicar una sobreacumulación de activos líquidos o una baja deuda a corto plazo.

Deuda total sobre el capital: Esta métrica presenta un sesgo negativo, indicando que algunas empresas tienen una proporción muy baja de deuda en comparación con su capital. Esto es un indicador de solidez financiera, pero puede afectar a las predicciones si no se tiene en cuenta adecuadamente.

Una revisión de los datos reveló la presencia de valores faltantes en algunas variables clave, como los ratios de liquidez y el precio de las acciones. Las variables como el 'Current ratio' y el 'Quick ratio' presentan más del 10 % de valores faltantes, mientras que la columna *'Price'* tiene más de un 25 % de datos ausentes. Estos valores faltantes deberán ser abordados mediante técnicas de imputación o eliminación de filas incompletas, dependiendo de la proporción y relevancia de las variables afectadas.

Dado que varias de las métricas analizadas presentan distribuciones sesgadas, se consideró la aplicación de transformaciones logarítmicas o de raíz cuadrada para reducir estos sesgos y acercar los datos a una distribución normal. Ambas transformaciones son especialmente útiles en datos financieros, donde es común observar grandes diferencias entre los valores de las empresas.

Antes de proceder con el modelado predictivo, se aplicarán transformaciones para normalizar los datos y técnicas de imputación para manejar los valores faltantes. Estos pasos son esenciales para asegurar que los modelos predictivos trabajen con datos representativos y eviten posibles sesgos en los resultados.

2.3 *Descripción de los modelos*

Dada la naturaleza de reproducción y mejora del presente trabajo, inicialmente se propone que los modelos implementados sean justamente los ya propuestos por Milenkovic.

C4.5 Decision Trees: Los árboles de decisión C4.5, propuestos por Quinlan [3] en 1993, son una extensión del algoritmo ID3 y se

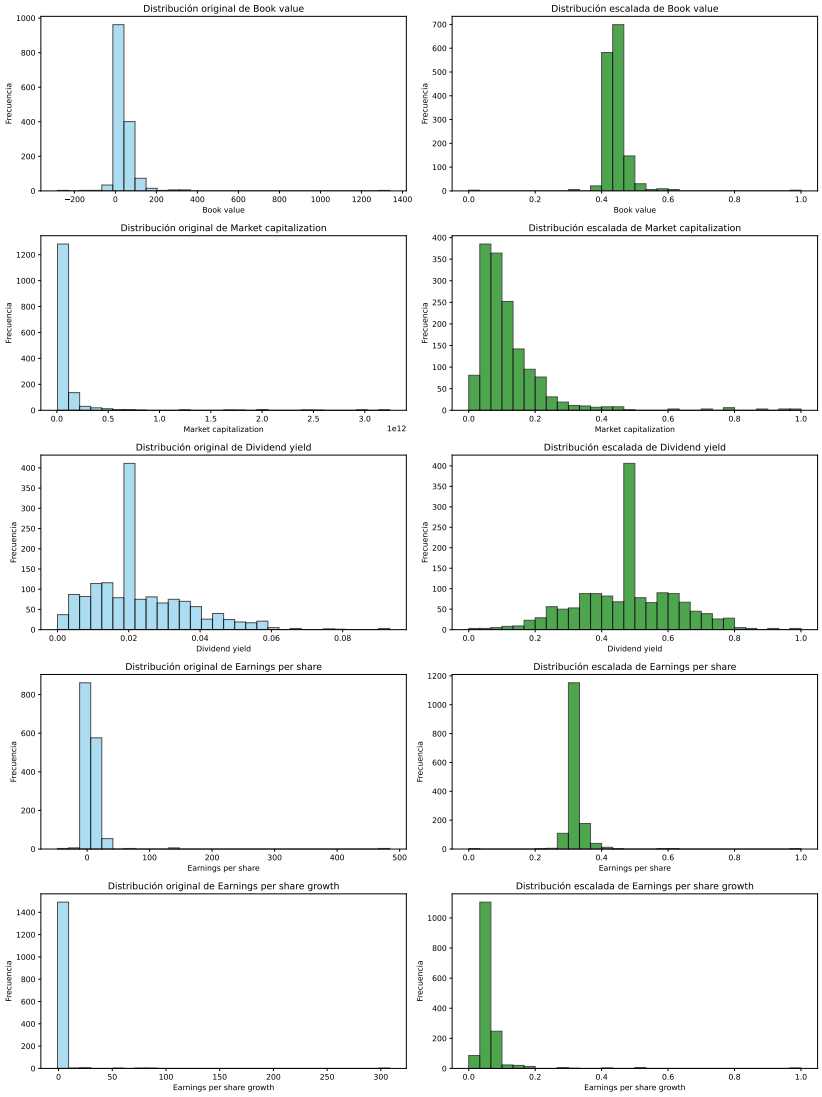


Figura 2.1: Distribución de las variables, Parte 1

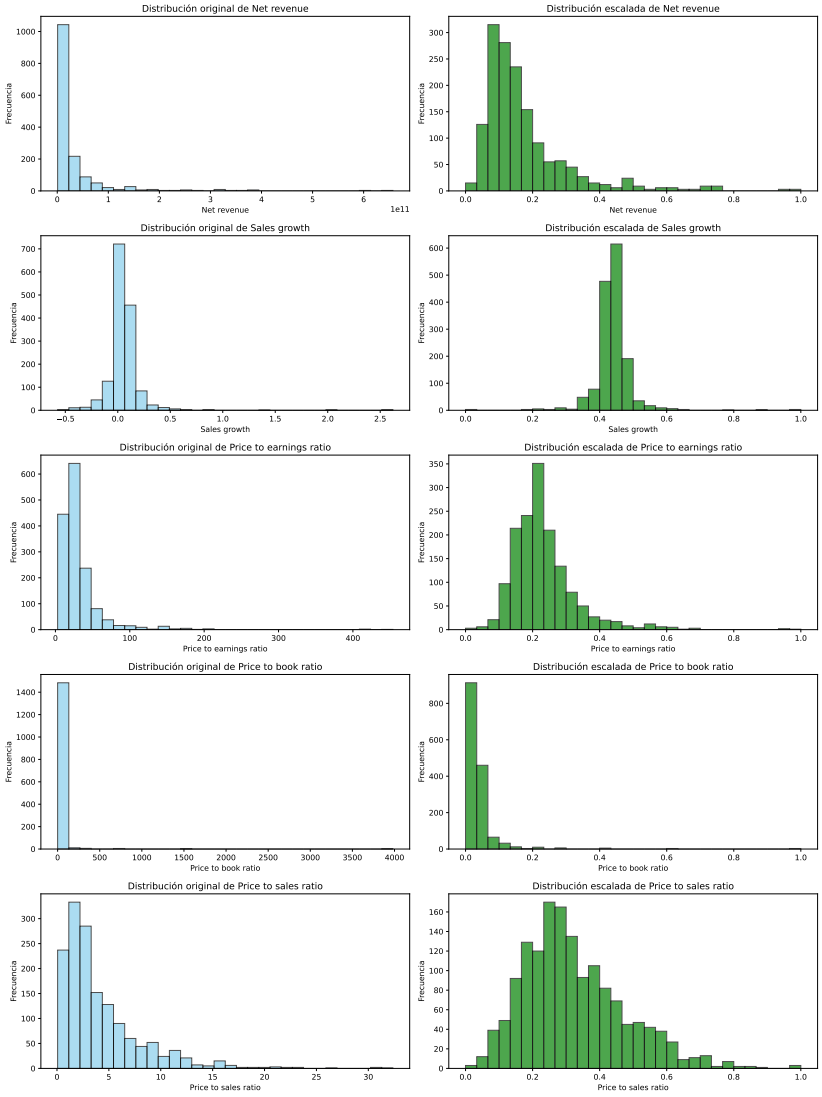


Figura 2.2: Distribución de las variables, Parte 2

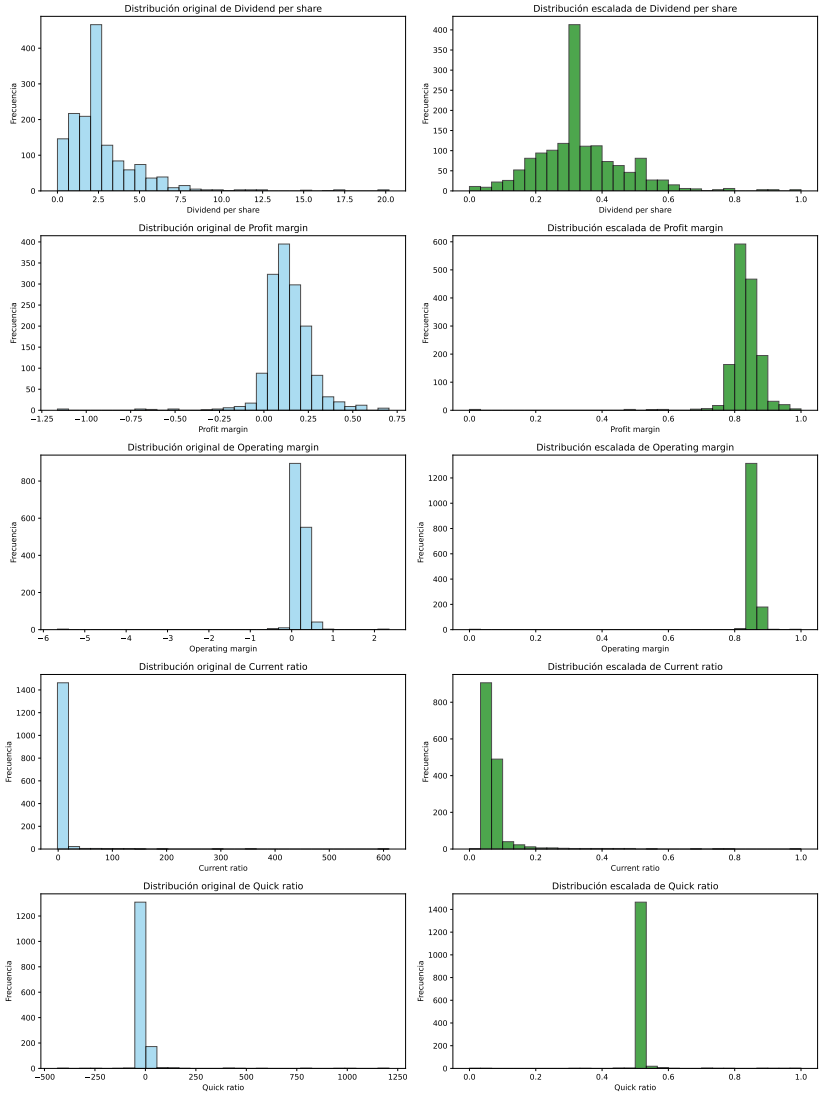


Figura 2.3: Distribución de las variables, Parte 3

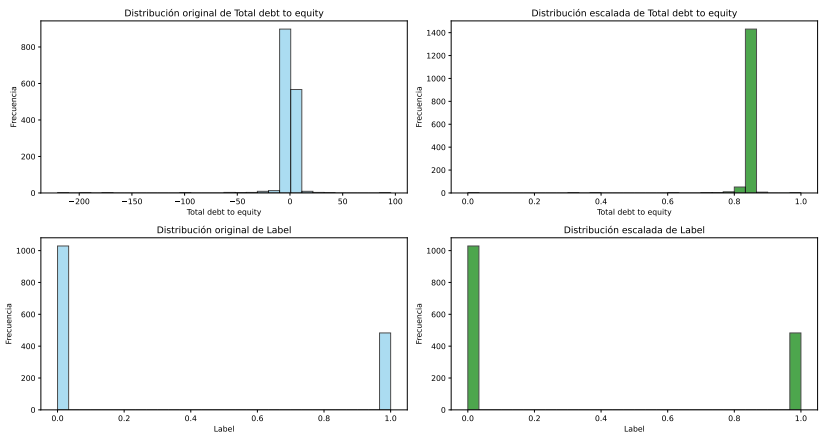


Figura 2.4: Distribución de las variables, Parte 4

utilizan para la clasificación de datos. Se basan en el concepto de dividir el conjunto de datos en subconjuntos más pequeños basados en pruebas de atributos, utilizando el criterio de ganancia de información. Una característica destacada de C4.5 es su capacidad para manejar tanto datos categóricos como numéricos. Además, el modelo maneja eficazmente los valores faltantes y puede podar ramas que no aportan una mejora significativa en la clasificación, reduciendo así el riesgo de sobreajuste.

Máquinas de Soporte Vectorial con *Sequential Minimal Optimization*: Las Máquinas de Soporte Vectorial (SVM) son modelos de clasificación que buscan encontrar un hiperplano óptimo que separe las clases en un espacio de características. Una de las principales fortalezas de SVM es su capacidad para manejar datos de alta dimensión. El algoritmo SMO (*Sequential Minimal Optimization*) es una técnica eficiente para resolver el problema de optimización cuadrática que surge al entrenar una SVM. Al descomponer el problema en subproblemas más pequeños que se resuelven analíticamente, SMO acelera significativamente el proceso de entrenamiento de SVM.

***Random Tree*:** Los árboles aleatorios son una técnica de clasificación que genera un único árbol de decisión durante el entrenamiento, pero introduce aleatoriedad en la selección de características en cada división. A diferencia de otros métodos de árboles de decisión que buscan la mejor característica en cada división, el árbol aleatorio selecciona una característica al azar. Esta aleatoriedad puede ayudar a evitar el sobreajuste y, a menudo, resulta en un árbol más generalizable.

***Random Forest*:** *Random Forest* es un ensamblaje de árboles de decisión que se entrenan con subconjuntos aleatorios de los datos y características. La clasificación final se decide mediante una votación mayoritaria entre todos los árboles del bosque. Al combinar múltiples

árboles, *Random Forest* mitiga los problemas de sobreajuste asociados con un único árbol de decisión y, a menudo, logra un rendimiento superior en términos de precisión y robustez.

Regresión logística: La regresión logística es un modelo de clasificación que estima la probabilidad de que una instancia pertenezca a una clase particular. Es especialmente útil cuando la variable de respuesta es binaria. La regresión logística modela la relación entre una o más características independientes y una variable dependiente categórica, utilizando la función logística. Es fácil de implementar, interpretable y se utiliza ampliamente en aplicaciones médicas, sociales y de marketing.

Redes Bayesianas *Naive*: Las redes bayesianas *Naive Bayes* son un modelo de clasificación basadoclasificación basada en el teorema de Bayes. Asumen que cada característica es independiente de las demás, dado el valor de la clase. A pesar de su simplicidad y la suposición ïngenua"de independencia, estos modelos son sorprendentemente efectivos en muchas situaciones, especialmente en la clasificación de texto. Las redes bayesianas, en general, ofrecen una representación gráfica de las relaciones probabilísticas entre variables, y el modelo *Naive Bayes* es un caso especial con una estructura particularmente simple.

2.4 *Descripción de las métricas*

Las métricas de evaluación son fundamentales para cuantificar la eficacia de los modelos de clasificación en aprendizaje automático.

Como se muestra en la Figura 2.5, la matriz de confusión es una herramienta que permite visualizar los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos, lo cual facilita la interpretación de las métricas de evaluación.

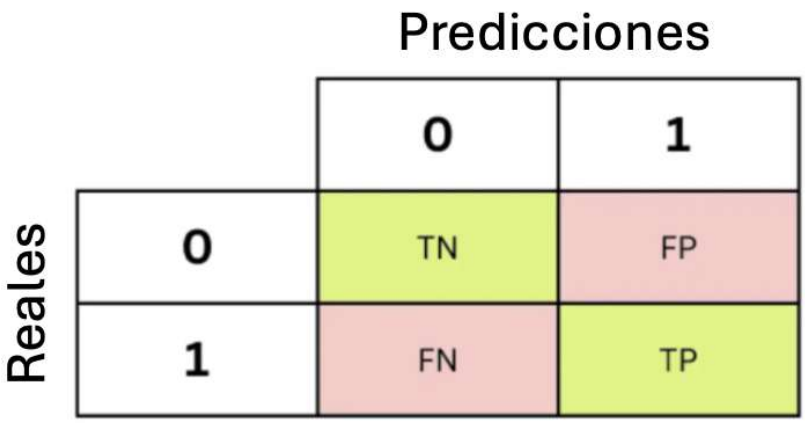
Exactitud. La exactitud, comúnmente conocida como *accuracy*, cuantifica la proporción de predicciones correctas realizadas por el modelo en relación con el número total de observaciones. Matemáticamente, se define como:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2.1}$$

Donde:

- TP representa los Verdaderos Positivos.
- TN representa los Verdaderos Negativos.
- FP representa los Falsos Positivos.
- FN representa los Falsos Negativos.

Figura 2.5: Matriz de confusión



Sensibilidad (*Recall*). La sensibilidad, en inglés denominada *recall*, se refiere a la proporción de observaciones positivas verdaderas predichas correctamente respecto al total de observaciones positivas verdaderas. Matemáticamente, se expresa como:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.2}$$

Esta métrica es particularmente importante en contextos donde los falsos negativos son especialmente relevantes.

Precisión. La Precisión mide la proporción de identificaciones positivas realizadas que fueron efectivamente correctas. Se centra en la calidad de las predicciones positivas del clasificador. Se define formalmente como:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.3}$$

Esta métrica es particularmente importante en contextos donde los falsos positivos son especialmente relevantes.

F1 Score. El *F1-Score* es una métrica armonizada que combina tanto la precisión como el recall en un único valor. Proporciona un balance entre estas dos métricas, siendo particularmente útil en escenarios donde ambas son de interés. Se define matemáticamente como *F1-Score*:

$$F_1 = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}} \tag{2.4}$$

El *F1-Score* fue elegido como la métrica principal de este estudio debido a su capacidad para ofrecer una evaluación balanceada entre precisión y recall, lo cual es crucial en escenarios con clases desbalanceadas, como es el caso en este estudio

2.5 *Descripción de los experimentos*

Los experimentos llevados a cabo en este estudio se enfocaron en la evaluación de diferentes modelos de *machine learning* aplicados a la predicción del rendimiento de acciones pertenecientes al índice *S&P 500*. En particular, enfrentamos la tarea de clasificar empresas basadas en su potencial de rendimiento de inversión, específicamente si se espera que superen o no el 10 % anual.

Se compararon modelos de regresión logística, *Support Vector Machines* (SVM), árboles de decisión (incluido el modelo 4.5), random forests y redes neuronales profundas. Adicionalmente, se implementaron variantes del modelo *XGBoost* y técnicas balanceo de clases como el uso de *Synthetic Minority Over-sampling Technique* (SMOTE).

Este capítulo detalla la configuración de los experimentos, la descripción de los modelos utilizados, las configuraciones específicas de las redes neuronales implementadas, el preprocesamiento de los datos y la evaluación de los modelos. Todos los experimentos fueron realizados con el conjunto de datos del *S&P 500*, utilizando como características fundamentales financieras de las empresas, tales como su capitalización de mercado, razón precio/ganancias, dividendos por acción, y márgenes de beneficio, entre otros.

Para asegurar que estos modelos operen con la máxima eficacia, es esencial implementar técnicas de preprocesamiento y normalización en nuestro conjunto de datos. Por ello, no solo trabajaremos con datos directamente escalados o normalizados, sino que también exploramos el impacto de aplicar transformaciones logarítmicas y de raíz cuadrada a estos datos escalados.

Los datos financieros de las empresas que componen el índice *S&P 500* fueron obtenidos de la *API* de *Alpha Vantage*. La primera fase del proceso experimental implicó la preparación y preprocesamiento de los datos. Esta etapa incluyó la limpieza de datos, imputación de valores faltantes y escalado de características.

Para la limpieza de datos, se identificaron y eliminaron las filas con valores nulos en la columna objetivo, es decir, en la variable "Label", que representa si el valor de las acciones de una empresa ha aumentado más de 10 % o no en un período de un año. Para las demás características numéricas, se procedió a realizar la imputación de valores faltantes utilizando la mediana de cada columna como valor representativo.

Esta estrategia se eligió para evitar el sesgo que podría introducirse al utilizar la media, ya que algunas características financieras presentaban distribuciones sesgadas.

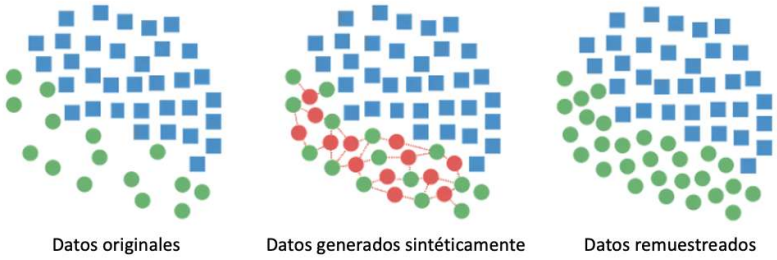
Posteriormente, las características numéricas fueron escaladas utilizando el escalador `MinMaxScaler` de `sklearn`, el cual ajusta los valores en un rango de [0, 1], lo que garantiza que los modelos basados en distancias, como SVM y redes neuronales, no se vean afectados por magnitudes muy distintas entre las características. Además, se aplicó una transformación basada en la raíz cuadrada, la cual ayudó a reducir la varianza en aquellas características que tenían distribuciones sesgadas hacia valores más grandes.

Una vez que los datos fueron preprocesados, se procedió a dividir el conjunto en datos de entrenamiento y prueba. Utilizando la función `train_test_split` de `sklearn`, el conjunto de datos fue dividido en 70 % para entrenamiento y 30 % para prueba, manteniendo la proporción original de las clases en ambas particiones.

Uno de los principales desafíos encontrados en el conjunto de datos fue el desbalanceo de clases. La clase mayoritaria correspondía a los casos donde el valor de las acciones no aumentaba más de 10 % (clase 0), mientras que la clase minoritaria correspondía a los casos en los que el valor de las acciones sí se apreciaba más de un 10 % después de un año (clase 1). Este desbalance afecta significativamente el rendimiento de los modelos, ya que tienden a predecir con mayor precisión la clase mayoritaria mientras ignoran la clase minoritaria. Después de iterar diversas configuraciones en los modelos propuestos, los modelos de redes neuronales presentaban una afectación particularmente alta. Para mitigar este problema, se empleó la técnica SMOTE (por sus siglas en inglés, *Synthetic Minority Over-sampling Technique*), la cual genera nuevas observaciones sintéticas de la clase minoritaria mediante interpolación entre ejemplos existentes. Esta técnica permitió aumentar el número de instancias de la clase minoritaria en el conjunto de entrenamiento sin simplemente duplicar las observaciones, lo que podría haber llevado a sobreajuste. Cabe mencionar que SMOTE fue aplicado exclusivamente al conjunto de entrenamiento, manteniendo el conjunto de prueba intacto para evaluar el rendimiento de los modelos en condiciones realistas.

Regresión logística. El modelo de regresión logística fue el primer modelo base implementado. Este modelo se seleccionó debido a su simplicidad y capacidad para servir como un punto de comparación frente a otros modelos más complejos. Se utilizó el clasificador `LogisticRegression` de `sklearn`, configurado con un máximo de iteraciones de 1000 para asegurar la convergencia. Además, se aplicó el parámetro `class_weight='balanced'`, lo que ajusta el modelo para compensar el desbalance de clases dando más peso a la clase minoritaria.

Figura 2.6: *Synthetic Minority Oversampling Technique* (Dholakiya, 2023)



La métrica de evaluación principal para este modelo fue el F1-score, dado que este combina tanto la precisión como el recall en un solo valor, que es crucial en situaciones de desbalance de clases, lo cual es epecialmente relevante en este modelo que comúnmente tiene limitaciones para manejar este desbalance. [4]

***Support Vector Machine* (SVM)**. El modelo SVM fue seleccionado debido a su capacidad para encontrar límites de decisión complejos y su robustez frente a problemas de clasificación no lineal. Se implementó utilizando un *kernel radial* (RBF) para capturar posibles relaciones no lineales entre las características. Al igual que en el modelo de regresión logística, se utilizó `class_weight='balanced'` para abordar el desbalance de clases.

Redes neuronales. Una de las partes más significativas de los experimentos fue la implementación de varias redes neuronales profundas, con diferentes configuraciones en cuanto a la cantidad de capas ocultas, neuronas y funciones de activación. Las redes neuronales fueron diseñadas utilizando el *API* secuencial de `TensorFlow` y `Keras`. Se implementaron un total de nueve configuraciones de redes neuronales, variando el número de capas y neuronas entre 32 y 256. Algunas configuraciones incluyeron el uso de la función de activación `relu`, mientras que otras utilizaron `tanh`. Adicionalmente, algunas redes incluían capas de `Dropout`, con tasas que oscilaban entre el 0.1 y el 0.3, para prevenir el sobreajuste. Todas las redes neuronales usaron la función de pérdida `binary_crossentropy` y el optimizador `adam`, a excepción de una red que utilizó la función de pérdida `hinge` y el

optimizador `rmsprop`.

***Random Forest* y Árboles de decisión.** El modelo de *Random Forest* fue seleccionado debido a su capacidad para manejar datos desbalanceados y evitar el sobreajuste que suelen experimentar los árboles de decisión individuales. Se probaron múltiples configuraciones del modelo, variando el número de árboles (`n_estimators`) entre 10 y 300, y ajustando parámetros como la profundidad máxima del árbol (`max_depth`) y el número mínimo de muestras por hoja (`min_samples_leaf`). En las configuraciones más avanzadas, se incluyó un ajuste del peso de las clases mediante el parámetro `class_weight={0: 1, 1: 3}`.

Asimismo, se implementaron árboles de decisión utilizando los criterios de entropía y *Gini* para evaluar su rendimiento. Los árboles de decisión individuales proporcionaron una base sólida para la creación de un *Random Forest* optimizado.

XGBoost. Finalmente, se implementaron múltiples configuraciones del modelo *XGBoost*. Este modelo es conocido por su capacidad de manejo de datos desbalanceados mediante el ajuste del parámetro `scale_pos_weight`, que pondera las instancias de la clase minoritaria. Se probaron tres configuraciones, ajustando el parámetro `scale_pos_weight` entre 3 y 5.

3 Resultados y discusión

3.1 Resultados

El modelo de regresión logística se utilizó como un punto de referencia debido a su simplicidad y amplia aplicación en tareas de clasificación binaria. Además, se evaluó el modelo SVM con un kernel radial (*Radial basis function*, RBF) para capturar relaciones no lineales entre las características. Ambos modelos fueron ajustados para manejar el desbalance de clases con class_weight='balanced'.

Modelo	*Precision*	*Recall*	*F1-Score*	*Accuracy*
Regresión Logística	0.66	0.61	0.61	0.63
SVM	0.62	0.57	0.57	0.59

Tabla 3.1: Métricas de desempeño para modelos de Regression Logística y de Máquina de Soporte Vectorial

Como se puede observar en la Tabla 3.3, ambos modelos presentan un rendimiento limitado al intentar manejar el desbalance de clases, con un *F1-score* de 0.63 para la regresión logística y 0.59 para el SVM. Esto refleja las dificultades de estos modelos para predecir correctamente la clase minoritaria. Este comportamiento era esperado dado que los SVM con *kernels* no lineales pueden verse influenciados por datos ruidosos y desbalanceados, lo que puede dificultar la búsqueda de un hiperplano óptimo.

Para contrastar estos resultados ante otra clase de modelos se aplicaron redes neuronales al problema. Se implementaron diversas configuraciones de redes neuronales profundas, variando el número de capas, neuronas y funciones de activación. También se evaluó el uso de dropout para reducir el sobreajuste. La Tabla 3.3 agrupa los resultados obtenidos para nueve configuraciones diferentes de redes neuronales.

En general, las redes neuronales mostraron un rendimiento competitivo, con las configuraciones más profundas (*Neural Network* 9) alcanzando un *F1-score* de 0.65 y una precisión de 0.69. Sin embargo, el problema del desbalance de clases persiste en muchas configuraciones, lo que limita el rendimiento en la clase minoritaria.

Los modelos de *Random Forest* y árboles de decisión (C4.5) mostraron una mejora significativa en comparación con los modelos clásicos. Se

Modelo	*Precision*	*Recall*	*F1-Score*	*Accuracy*
Red Neuronal 1 (64, 32)	0.64	0.69	0.59	0.69
Red Neuronal 2 (128, 64)	0.66	0.69	0.66	0.69
Red Neuronal 3 (32, 32)	0.66	0.70	0.65	0.70
Red Neuronal 4 (64, 32, `Dropout` 0.1)	0.65	0.68	0.65	0.68
Red Neuronal 5 (256, 128)	0.67	0.70	0.64	0.70
Red Neuronal 6 (`tanh`)	0.65	0.69	0.60	0.69
Red Neuronal 7 (`hinge`)	0.65	0.69	0.65	0.69
Red Neuronal 8 (`hinge`)	0.67	0.69	0.67	0.69
Red Neuronal 9 (128, 64)	0.69	0.71	0.65	0.71

Tabla 3.2: Resultados de Redes neuronales (diversas configuraciones).

probaron varias configuraciones de *Random Forest*, variando el número de árboles y ajustando parámetros como el número mínimo de muestras en cada hoja y el número de características consideradas en cada división.

Modelo	*Precision*	*Recall*	*F1-Score*	*Accuracy*
Árbol de Decisión (*Entropy*)	0.63	0.63	0.63	0.63
Árbol de Decisión (Gini)	0.64	0.65	0.64	0.65
Random Forest (100 árboles)	0.69	0.71	0.68	0.71
Random Forest (300 árboles)	0.69	0.71	0.68	0.71

Tabla 3.3: Resultados de Árboles de Decisión y *Random Forest*.

Como se observa en la Tabla 3.3, los modelos de *Random Forest* lograron una precisión y un *F1-score* de 0.68, lo que indica una mejora en la predicción de la clase minoritaria en comparación con los modelos de regresión logística y SVM.

Finalmente, se implementaron tres variantes del modelo *XGBoost*, ajustando el parámetro `scale_pos_weight` para abordar el desbalance de clases. *XGBoost* mostró ser uno de los modelos más robustos, en particular la configuración 2, que usa el parámetro (`scale_pos_weight` = 5, que incluye `use_label_encoder`, y usa `logloss` como métrica de evaluación), alcanzando un *F1-score* de 0.72 y una precisión de 0.71.

Modelo	*Precision*	*Recall*	*F1-Score*	*Accuracy*
XGBoost 1	0.70	0.72	0.72	0.70
XGBoost 2	0.71	0.73	0.73	0.72
XGBoost 3	0.65	0.67	0.67	0.66

Tabla 3.4: Métricas de rendimiento para múltiples configuraciones de *XGBoost*.

3.2 *Discusión*

Los modelos de *Random Forest* y *XGBoost* demostraron ser más efectivos para la tarea de predicción en presencia de desbalance de clases. Los modelos de redes neuronales también obtuvieron buenos resultados, pero requieren ajustes adicionales para mejorar

el rendimiento en la clase minoritaria. Los modelos tradicionales, como la regresión logística y SVM, mostraron limitaciones en este contexto.

En la Tabla 3.5, se presenta una comparación entre los resultados obtenidos en este trabajo y los del estudio realizado por Milosevic. Es importante señalar que, aunque ambos estudios utilizan técnicas de aprendizaje automático para predecir el movimiento de precios de acciones, existen diferencias clave en los conjuntos de datos y las metodologías utilizadas que deben considerarse al interpretar los resultados.

Estudio	Algoritmo	*Precision*	*Recall*	*F1-Score*
Milosevic	*Random Forest*	0.75	0.75	0.75
Presente Estudio	*XGBoost*	0.71	0.73	0.73

Tabla 3.5: Comparación entre los resultados del estudio original de Milosevic y los obtenidos en este trabajo.

El trabajo de Milosevic se basó en datos históricos de los años 2015 y utilizó acciones pertenecientes a los índices *S&P 400*, *S&P 500* y *S&P 600*, lo que abarca un universo más amplio de compañías con diferentes niveles de capitalización bursátil. En cambio, este estudio se centró en los activos del *S&P 500* utilizando datos más recientes, de los años 2021 y 2022, lo que aporta una perspectiva actualizada sobre el comportamiento del mercado de valores de alta capitalización. Aunque el *F1-Score* alcanzado por el modelo *Random Forest* de Milosevic fue ligeramente superior (0.75 frente a 0.72), es relevante destacar que el presente trabajo implementó modelos más avanzados, como *XGBoost*, que fueron capaces de manejar mejor el desbalance de clases. Adicionalmente, la mayor recencia de los datos utilizados en este estudio ofrece una mejor adaptación al entorno financiero actual, marcado por eventos como la pandemia de *COVID-19* y la recuperación económica posterior, lo que introduce nuevas dinámicas de volatilidad y riesgo en el mercado que no estaban presentes en el periodo 2015.

4 Conclusiones y trabajo futuro

4.1 Conclusiones

En función de los resultados obtenidos, se puede concluir que los modelos avanzados como *Random Forest* y *XGBoost* lograron superar a los métodos tradicionales de clasificación binaria, como la regresión logística y el SVM. El uso de *XGBoost* permitió manejar mejor el desbalance de clases, alcanzando un *F1-Score* de 0.72, lo cual es competitivo frente a los 0.75 obtenidos por Milosevic en su estudio de 2016.

Cabe destacar que la diferencia en los conjuntos de datos, tanto en términos de la selección de activos (*S&P 500* frente a una combinación de *S&P 400, 500 y 600*) como de los periodos temporales (2021-2022 frente a 2015), juega un papel clave en la variabilidad de los resultados. Estos cambios en el entorno financiero y las metodologías utilizadas en el preprocesamiento de los datos y en la selección de características muestran que los métodos de predicción deben ajustarse de manera constante a los cambios del mercado.

Además, las redes neuronales, aunque presentaron resultados competitivos en general, continúan enfrentándose al desafío del desbalance de clases. Si bien se observaron mejoras en las configuraciones más profundas, el rendimiento en la clase minoritaria sigue siendo un área de mejora. En resumen, este trabajo ha demostrado que la combinación de técnicas avanzadas de *machine learning*, con el uso de datos financieros actualizados, puede mejorar la precisión en la predicción del movimiento de precios de acciones, aunque se requieren ajustes continuos para optimizar el rendimiento en condiciones de desbalance.

4.2 Trabajo futuro

Este estudio abre nuevas posibilidades para futuras investigaciones en el área de predicción de precios de acciones utilizando aprendizaje automático. Un área clave para explorar es la ampliación del conjunto de datos. Si bien se han utilizado los 500 activos del índice *S&P*

500, extender el análisis a otras bolsas de valores o incluir activos internacionales podría proporcionar una visión más completa de las dinámicas de predicción a nivel global. Además, incorporar datos históricos con mayor granularidad temporal, como precios intradía, podría mejorar la capacidad de los modelos para captar fluctuaciones de corto plazo.

Otro aspecto importante es la optimización de los modelos para abordar mejor el desbalance de clases. A pesar de que en este trabajo se aplicaron técnicas como SMOTE para equilibrar las clases, explorar métodos más sofisticados, como el uso de algoritmos de aprendizaje de costos o técnicas de remuestreo más avanzadas, podría permitir un tratamiento más eficiente del desbalance y mejorar el rendimiento en la clase minoritaria.

Por último, una línea de investigación relevante es el análisis de la interpretabilidad de los modelos. Comprender qué características influyen más en las predicciones es crucial para aumentar la confianza en estos modelos, especialmente en el contexto financiero. Desarrollar enfoques que permitan interpretar los factores más relevantes no solo ayudaría a optimizar los modelos, sino que también aportaría información valiosa para la toma de decisiones financieras, mejorando su aplicabilidad en entornos del mundo real.

Bibliografía

[1] N. Milosevic, "Equity forecast: Predicting long term stock price movement using machine learning," *Journal of Economics Library*, 2016.

[2] AlphaVantage, "Stock market data api for algorithmic trading," 2024.

[3] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, first ed., 1993.

[4] E. Akyildirim, "Forecasting high-frequency stock returns: a comparison of alternative methods," *Annals of Operations Research*, 2022.