

Pontificia Universidad Javeriana Cali



Pontificia Universidad
JAVERIANA
Cali

**Facultad de Ingeniería
y Ciencias**
Ingeniería Electrónica

**Comparación de Métodos de Aprendizaje de Máquina Aplicados a la Predicción de
Movimientos a Corto Plazo de Acciones del DJIA Durante Periodos de Reportes
Financieros Trimestrales**

Juan David Vargas Mazuera

Septiembre, 2022

Pontificia Universidad Javeriana Cali



Pontificia Universidad
JAVERIANA
Cali

Facultad de Ingeniería
y Ciencias
Ingeniería Electrónica

**Comparación de Métodos de Aprendizaje de Máquina Aplicados a la Predicción de
Movimientos a Corto Plazo de Acciones del DJIA Durante Periodos de Reportes
Financieros Trimestrales**

Juan David Vargas Mazuera

**Trabajo de Grado realizado como prerrequisito para optar
por el título de ingeniero electrónico**

Directora:

Ingeniera María Constanza Pabón Burbano, PhD.

Septiembre, 2022

Santiago de Cali, septiembre 28 del 2022.

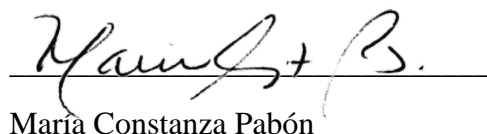
Señores
Pontifica Universidad Javeriana Cali
Dr. Luis Eduardo Tobón Llano
Director de Carrera de Ingeniería Electrónica
Cali.

Cordial Saludo.

Me complace presentar ante usted el trabajo de grado titulado “Comparación de Métodos de Aprendizaje de Máquina Aplicados a la Predicción de Movimientos a Corto Plazo de Acciones del DJIA Durante Periodos de Reportes Financieros Trimestrales”.

Manifiesto con mi firma, que, el presente trabajo, se encuentra finalizado y está listo para su sustentación.

Atentamente,



María Constanza Pabón

Santiago de Cali, septiembre 28 del 2022.

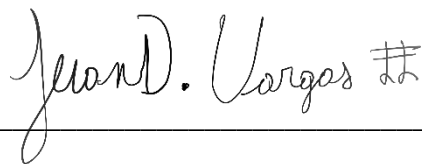
Señores
Pontifica Universidad Javeriana Cali
Dr. Luis Eduardo Tobón Llano
Director de Carrera de Ingeniería Electrónica
Cali.

Cordial Saludo.

Me complace presentar ante usted el anteproyecto de grado titulado “Comparación de Métodos de Aprendizaje de Máquina Aplicados a la Predicción de Movimientos a Corto Plazo de Acciones del DJIA Durante Periodos de Reportes Financieros Trimestrales” cuya finalidad es cumplir con los requisitos establecidos por la Universidad para realizar el proyecto de grado y culminar mi proceso de formación como Ingeniero Electrónico.

Manifiesto con mi firma, que, comprendo las directrices para la presentación de trabajos de grado de la Facultad de Ingeniería y Ciencias, las cuales fueron aprobadas el 26 de noviembre de 2009, donde se establecen la normatividad y los plazos correspondientes al desarrollo del anteproyecto y proyecto de grado. Además, manifiesto que el proyecto se encuentra finalizado y está listo para sustentación.

Atentamente,

A handwritten signature in dark ink, reading "Juan D. Vargas" followed by a stylized symbol resembling a hash or a double cross. The signature is written over a horizontal line.

Juan David Vargas
Código 8940230

DEDICATORIA

El presente trabajo de grado se lo dedico a dios, mi familia más cercana, y mis mejores amigos.

El amor y apoyo casi incondicional de mi abuela, papá, y mamá han hecho que pueda lograr muchas de mis metas a pesar de las limitaciones. Agradezco profundamente a ellos. Además, quiero agradecer a dios debido a todo lo bendecido que he logrado ser en mi vida. Finalmente, ha sido un gran apoyo para mí el tener grandes mejores amigos. Es de las mejores cosas de mi vida tener amigos tan increíbles como Alejandra, Laura, Fabián, y Brianda. A ellos les agradezco profundamente por ser una parte muy importante de mi vida.

AGRADECIMIENTOS

Agradezco especialmente a mi directora de grado, quién ha estado muy pendiente del progreso, me ha dirigido, y ha aportado ideas que se aplicaron a la realización de este proyecto. Agradezco especialmente su constante apoyo durante los casi 7 meses de desarrollo/implementación del proyecto.

Además, agradezco a mi mamá, quién me ha apoyado en este proceso con sus consejos, quién, además, ha permitido que pueda seguir un buen camino para su realización y ha estado constantemente pendiente del desarrollo de mi trabajo de grado. Igualmente, agradezco a mi amigo Julián, con quién inicié el proyecto, y con quién trabajé en la etapa del anteproyecto.

ÍNDICE GENERAL

ÍNDICE DE TABLAS	9
ÍNDICE DE FIGURAS	12
RESUMEN	19
ABSTRACT	20
1 INTRODUCCIÓN	21
2 JUSTIFICACIÓN.....	24
3 PLANTEAMIENTO DEL PROBLEMA.....	25
4 OBJETIVOS.....	27
4.1 OBJETIVO GENERAL.....	27
4.2 OBJETIVOS ESPECÍFICOS	27
5 MARCO TEÓRICO.....	28
5.1 TÉRMINOS Y CONCEPTOS DE LA BOLSA DE VALORES.....	28
5.1.1 Mercados financieros	28
5.1.2 Índice Bursátil	28
5.1.3 Acción	29
5.1.4 Promedio Industrial Dow Jones	29
5.1.5 Reportes financieros.....	29
5.1.6 Beneficios por acción.....	29
5.1.7 EPS estimadas (Pre- resultados financieros oficiales)	30
5.1.8 Análisis Técnico.....	30
5.1.9 Análisis Fundamental.....	30
5.1.10 Trader	30
5.1.11 Trading algorítmico.....	30
5.1.12 Trading a corto vs. largo plazo.....	31
5.1.13 Horarios de la bolsa de valores	31
5.1.14 “Sell-off”	31
5.1.15 “Rally”	32
5.1.16 “Crash”.....	32
5.1.17 Potencial de ganancias	32
5.1.18 Cobertura financiera.....	32

5.1.19	Volatilidad.....	32
5.2	CONCEPTOS Y MODELOS DE APRENDIZAJE DE MÁQUINA	33
5.2.1	Modelos de regresión lineal y polinómico.....	33
5.2.2	Modelo de aprendizaje de máquina de clasificación.....	33
5.2.3	Hiperparámetros.....	33
5.2.4	Optimización de hiperparámetros	34
5.3	MÉTRICAS DE DESEMPEÑO DE APRENDIZAJE DE MÁQUINA.....	34
5.3.1	R2-Square	34
5.3.2	TP, TN, FP, Y FN	34
5.3.3	Certeza o exactitud.....	34
5.3.4	Precisión.....	35
5.3.5	Recall	35
5.3.6	F1-score.....	35
5.3.7	Matriz de confusión	35
5.4	ESTADO DEL ARTE	36
5.4.1	Stock Price Movements Classification Using Machine Learning and Deep Learning Techniques-The Case Study of Indian Market.....	36
5.4.2	A Regression Model to Predict Stock Market Mega Movements and/or Volatility Using Both Macroeconomic Indicators & Fed Bank Variables.....	36
5.4.3	Algorithmic Financial Trading with Deep Convolutional Neural Networks: Time Series to Image Conversion	37
5.4.4	A Deep Learning Model for Predicting BUY and SELL Recommendations in Stock Exchange of Thailand using Long Short-Term Memory	37
5.4.5	On Stock Market Movement Prediction Via Stacking Ensemble Learning Model ..	38
5.4.6	The impact of Public Information on the Stock Market.....	38
5.4.7	Fear and Greed: a Returns-Based Trading Strategy around Earnings Announcements	38
5.4.8	Volatility Skew, Earnings Announcements, and the Predictability of Crashes	39
5.4.9	Rationality of analysts' earnings forecasts: evidence from dow 30 companies.....	39
6	RECOPIACIÓN DE DATOS Y CONSTRUCCIÓN DE SETS DE DATOS	40
6.1	ELECCIÓN DE EMPRESAS PARA LA IMPLEMENTACIÓN DE APRENDIZAJE DE MÁQUINA ...	40
6.2	DESCRIPCIÓN DE DATOS REQUERIDOS PARA LA IMPLEMENTACIÓN	42
6.3	RECOPIACIÓN DE DATOS.....	42
6.4	DEFINICIÓN DE SETS DE DATOS.....	44

7	ADECUACIÓN DE DATOS, Y DEFINICIÓN DE PARÁMETROS Y MÉTRICAS.....	47
7.1	SORPRESA DE BENEFICIOS POR ACCIÓN	47
7.2	CAMBIO DE PRECIO DE LAS ACCIONES	51
7.3	REDEFINICIÓN DE LOS NOMBRES DE LOS SETS DE DATOS	54
7.4	IMPLEMENTACIÓN CON ALGORITMOS DE REGRESIÓN	54
7.5	ETIQUETAS PARA ALGORITMOS DE CLASIFICACIÓN	59
7.6	MÉTRICAS DE OPTIMIZACIÓN EN LOS MÉTODOS DE CLASIFICACIÓN	63
7.7	PARÁMETROS ADICIONALES PARA CONSIDERAR PARA LA IMPLEMENTACIÓN DE LOS MÉTODOS DE APRENDIZAJE DE MÁQUINA.....	64
8	IMPLEMENTACIÓN DE MÉTODOS DE APRENDIZAJE DE MÁQUINA DE CLASIFICACIÓN	65
8.1	PRUEBAS INICIALES DE MÉTODOS DE CLASIFICACIÓN.....	65
8.2	ELECCIÓN DE MÉTODOS DE CLASIFICACIÓN	75
8.3	PRUEBAS INICIALES DE OPTIMIZACIÓN DE HIPERPARÁMETROS.....	77
8.4	OPTIMIZACIÓN GENERAL DE HIPERPARÁMETROS	85
8.5	OPTIMIZACIÓN EXTENSIVA DE HIPERPARÁMETROS	98
9	CONCLUSIONES	107
10	DIFICULTADES	110
11	TRABAJOS FUTUROS Y RECOMENDACIONES.....	111
12	BIBLIOGRAFÍA.....	113

ÍNDICE DE TABLAS

Tabla 1.	10 empresas más antiguas del DJIA a mayo del 2011	41
Tabla 2.	5 empresas más antiguas del DJIA a julio del 2022.....	41
Tabla 3.	Comparación fecha mínima de recopilación para cada tipo de dato.....	43
Tabla 4.	Información general de las empresas seleccionadas.....	45
Tabla 5.	Horario de publicación de resultados financieros de cada empresa seleccionada	51
Tabla 6.	Redefinición de los nombres de los sets de datos definidos.....	54
Tabla 7.	Desviación estándar periodo de 5 y 30 días para cada una de las empresas seleccionadas.....	59
Tabla 8.	Límites Q1 y Q3 usados para definir las etiquetas BUY/HOLD/SELL en cada set de datos	61
Tabla 9.	Hiperparámetros de los métodos KNN y árbol de decisión	65
Tabla 10.	Tiempos de ejecución para cada tipo de modelo de aprendizaje de máquina sin optimización de hiperparámetros.....	74
Tabla 11.	Descripción de métodos de aprendizaje de máquina usados[56], [57], [59]–[62]	76
Tabla 12.	Periodos de días de variación porcentual seleccionados para la optimización de hiperparámetros en la prueba inicial	77
Tabla 13.	Rangos de hiperparámetros usados para la optimización del método KNN en la etapa de prueba inicial[54]	78
Tabla 14.	Rangos de hiperparámetros usados para la optimización del método árbol de decisión en la etapa de prueba inicial[55]	78
Tabla 15.	Rangos de hiperparámetros usados para la optimización del método Random Forest en la etapa de prueba inicial[64]	79
Tabla 16.	Rangos de hiperparámetros usados para la optimización del método “Logistic Regression” en la etapa de prueba inicial[65].....	79
Tabla 17.	Rangos de hiperparámetros usados para la optimización del método SVC en la etapa de prueba inicial[66].....	79
Tabla 18.	Rangos de hiperparámetros usados para la optimización del método Light GBM en la etapa de prueba inicial[67]	80
Tabla 19.	Hiperparámetros que lograron los mejores resultados para el método KNN	82

Tabla 20. Hiperparámetros que lograron los mejores resultados para el método Decision Tree	83
Tabla 21. Hiperparámetros que lograron los mejores resultados para el método Random Forest	83
Tabla 22. Hiperparámetros que lograron los mejores resultados para el método Logistic Regression	83
Tabla 23. Hiperparámetros que lograron los mejores resultados para el método SVC	84
Tabla 24. Hiperparámetros que lograron los mejores resultados para el método Light GBM	84
Tabla 25. Tiempos de ejecución para la optimización de cada método de aprendizaje de máquina en la etapa de prueba inicial	84
Tabla 26. Rangos reducidos de hiperparámetros para cada método de aprendizaje de máquina	85
Tabla 27. Recopilación de mejores métodos de aprendizaje de máquina y tipos de datasets para el dataset de IBM (izquierda) y AMEX (derecha).....	89
Tabla 28. Recopilación de mejores métodos de aprendizaje de máquina y tipos de datasets para el dataset de 3M (izquierda) y MERCK (derecha).....	90
Tabla 29. Recopilación de mejores métodos de aprendizaje de máquina y tipos de datasets para el dataset de PG (izquierda) y TECH (derecha).....	91
Tabla 30. Recopilación de mejores métodos de aprendizaje de máquina y tipos de datasets para el dataset de NONTECH (izquierda) y ALL (derecha)	92
Tabla 31. Tiempos de ejecución de cada set de datos para una optimización general de hiperparámetros y considerando todos los tipos de sets de datos y métodos de aprendizaje de máquina	97
Tabla 32. Periodos de variación porcentual seleccionados para la optimización extensiva de hiperparámetros	98
Tabla 33. Rangos de hiperparámetros usados para optimización extensiva.....	99
Tabla 34. Resultados de los modelos de KNN después de optimización extensiva de hiperparámetros.....	103
Tabla 35. Resultados de los modelos de Random Forest después de optimización extensiva de hiperparámetros.	105
Tabla 36. Tiempos de ejecución de optimización extensiva de hiperparámetros para cada método de aprendizaje de máquina.....	105

**Tabla 37. Hiperparámetros de cada modelo de KNN y Random Forest que
generaron los mejores resultados de F1-Score SELL y de certeza..... 106**

ÍNDICE DE FIGURAS

Figura 1.	Evolución de la proporción de adultos estadounidenses con acciones en la bolsa[1].	21
Figura 2.	Evolución del precio de la acción GME desde octubre 1 del 2020[4].....	22
Figura 3.	Estructura y componentes de los mercados financieros [12].....	28
Figura 4.	TP, TN, FP, y FN en matriz de confusión con una variable objetivo de 2 clases	36
Figura 5.	Ejemplo de formato de datos históricos del precio de acciones	43
Figura 6.	Gráficos de sorpresa de beneficios por acción de IBM	48
Figura 7.	Gráficos de sorpresa de beneficios por acción de American Express	48
Figura 8.	Gráficos de sorpresa de beneficios por acción de 3M	48
Figura 9.	Gráficos de sorpresa de beneficios por acción de Merck	49
Figura 10.	Gráficos de sorpresa de beneficios por acción de P&G	49
Figura 11.	Gráficos de sorpresa de beneficios por acción del grupo IBM + American Express	49
Figura 12.	Gráficos de sorpresa de beneficios por acción del grupo Merck + P&G + 3M	50
Figura 13.	Gráficos de sorpresa de beneficios por acción del grupo de todas las acciones	50
Figura 14.	Coeficiente de correlación entre sorpresa de beneficios por acción y variación porcentual del precio de la acción para cada set de datos.....	53
Figura 15.	Sorpresas en series de tiempo (izquierda) y resultados de R2 square (derecha) para los diferentes periodos del set de IBM en métodos de regresión polinómica	55
Figura 16.	Sorpresas en series de tiempo (izquierda) y resultados de R2 square (derecha) para los diferentes periodos del set de AMEX en métodos de regresión polinómica	55
Figura 17.	Sorpresas en series de tiempo (izquierda) y resultados de R2 square (derecha) para los diferentes periodos del set de 3M en métodos de regresión polinómica	55

Figura 18. Sorpresa en series de tiempo (izquierda) y resultados de R2 square (derecha) para los diferentes periodos del set de MERCK en métodos de regresión polinómica	56
Figura 19. Sorpresa en series de tiempo (izquierda) y resultados de R2 square (derecha) para los diferentes periodos del set de PG en métodos de regresión polinómica	56
Figura 20. Sorpresa en series de tiempo (izquierda) y resultados de R2 square (derecha) para los diferentes periodos del set de TECH en métodos de regresión polinómica	56
Figura 21. Sorpresa en series de tiempo (izquierda) y resultados de R2 square (derecha) para los diferentes periodos del set de NONTECH en métodos de regresión polinómica	57
Figura 22. Sorpresa en series de tiempo (izquierda) y resultados de R2 square (derecha) para los diferentes periodos del set de ALL en métodos de regresión polinómica	57
Figura 23. Mejores valores de R2 square para cada set de datos en diferentes métodos de regresión polinómica	58
Figura 24. Valores promedio de R2 square para cada set de datos en diferentes métodos de regresión polinómica	58
Figura 25. Cantidad de etiquetas BUY, HOLD, y SELL de IBM (izquierda) y AMEX (derecha)	62
Figura 26. Cantidad de etiquetas BUY, HOLD, y SELL de 3M (izquierda) y MERCK (derecha)	62
Figura 27. Cantidad de etiquetas BUY, HOLD, y SELL de P&G (izquierda) y del grupo de AMEX - IBM (derecha).....	62
Figura 28. Cantidad de etiquetas BUY, HOLD, y SELL del grupo de 3M - MERCK - P&G (izquierda) y del grupo de todas las acciones (derecha)	63
Figura 29. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos IBM al entrenar el modelo KNN sin optimización de parámetros	66
Figura 30. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos AMEX al entrenar el modelo KNN sin optimización de parámetros	66

Figura 31. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos 3M al entrenar el modelo KNN sin optimización de parámetros 67

Figura 32. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos MERCK al entrenar el modelo KNN sin optimización de parámetros 67

Figura 33. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos PG al entrenar el modelo KNN sin optimización de parámetros 67

Figura 34. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos TECH al entrenar el modelo KNN sin optimización de parámetros 68

Figura 35. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos NONTECH al entrenar el modelo KNN sin optimización de parámetros 68

Figura 36. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos ALL al entrenar el modelo KNN sin optimización de parámetros 68

Figura 37. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos IBM al entrenar el modelo árbol de decisión sin optimización de parámetros 69

Figura 38. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos AMEX al entrenar el modelo árbol de decisión sin optimización de parámetros 69

Figura 39. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos 3M al entrenar el modelo árbol de decisión sin optimización de parámetros 70

Figura 40. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos MERCK al entrenar el modelo árbol de decisión sin optimización de parámetros 70

Figura 41. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos PG al entrenar el modelo árbol de decisión sin optimización de parámetros 70

Figura 42. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos TECH al entrenar el modelo árbol de decisión sin optimización de parámetros	71
Figura 43. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos NONTECH al entrenar el modelo árbol de decisión sin optimización de parámetros	71
Figura 44. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos ALL al entrenar el modelo árbol de decisión sin optimización de parámetros	71
Figura 45. Mejores resultados de cada uno de los sets de datos para el método KNN	72
Figura 46. Resultado promedio de cada uno de los sets de datos para el método KNN	72
Figura 47. Mejores resultados de cada uno de los sets de datos para el método árbol de decisión	73
Figura 48. Resultado promedio de cada uno de los sets de datos para el método árbol de decisión	73
Figura 49. Mejores métodos de aprendizaje de máquina de clasificación para sets de datos con pocos elementos[59]	75
Figura 50. Mejores resultados de F1-score SELL (izquierda) y de certeza (derecha) para el método de KNN	80
Figura 51. Mejores resultados de F1-score SELL (izquierda) y de certeza (derecha) para el método de árbol de decisión	81
Figura 52. Mejores resultados de F1-score SELL (izquierda) y de certeza (derecha) para el método de Random Forest	81
Figura 53. Mejores resultados de F1-score SELL (izquierda) y de certeza (derecha) para el método de Logistic Regression	81
Figura 54. Mejores resultados de F1-score SELL (izquierda) y de certeza (derecha) para el método de SVC	82
Figura 55. Mejores resultados de F1-score SELL (izquierda) y de certeza (derecha) para el método de Light GBM	82
Figura 56. Mejores resultados de F1-score y certeza para el set de datos IBM con optimización reducida de hiperparámetros	85

Figura 57. Mejores resultados de F1-score y certeza para el set de datos AMEX con optimización reducida de hiperparámetros	86
Figura 58. Mejores resultados de F1-score y certeza para el set de datos 3M con optimización reducida de hiperparámetros	86
Figura 59. Mejores resultados de F1-score y certeza para el set de datos MERCK con optimización reducida de hiperparámetros	86
Figura 60. Mejores resultados de F1-score y certeza para el set de datos PG con optimización reducida de hiperparámetros	87
Figura 61. Mejores resultados de F1-score y certeza para el set de datos TECH con optimización reducida de hiperparámetros	87
Figura 62. Mejores resultados de F1-score y certeza para el set de datos NONTech con optimización reducida de hiperparámetros.....	87
Figura 63. Mejores resultados de F1-score y certeza para el set de datos ALL con optimización reducida de hiperparámetros	88
Figura 64. Cantidad porcentual de veces que cada modelo de aprendizaje de máquina generó mejores resultados	93
Figura 65. Cantidad porcentual de veces que cada tipo de dataset generó mejores resultados	94
Figura 66. Cantidad porcentual de veces que cada modelo de aprendizaje de máquina generó resultados promedios entre F1-score SELL y certeza mayores o iguales a 0.7.....	95
Figura 67. Cantidad porcentual de veces que cada tipo de dataset generó resultados promedios entre F1-score SELL y certeza mayores o iguales a 0.7	95
Figura 68. Cantidad porcentual de veces que cada dataset generó resultados promedios entre F1-score SELL y certeza mayores o iguales a 0.7	96
Figura 69. Cantidad porcentual de veces que cada dataset generó resultados promedios entre F1-score SELL y certeza mayores o iguales a 0.8	96
Figura 70. Cantidad porcentual de veces que cada modelo de aprendizaje de máquina generó resultados promedios entre F1-score SELL y certeza mayores o iguales a 0.7 en el set de datos MERCK.....	99
Figura 71. Resultados de F1-score SELL y Certeza después de optimización extensiva de hiperparámetros usando el modelo KNN.....	100
Figura 72. Resultados de F1-score SELL y Certeza después de optimización extensiva de hiperparámetros usando el modelo Random Forest	100

Figura 73. Matriz de confusión para el modelo KNN con optimización extensiva de hiperparámetros, para el periodo de 10 días de variación porcentual.....	101
Figura 74. Matriz de confusión para el modelo KNN con optimización extensiva de hiperparámetros, para el periodo de 17 días de variación porcentual.....	101
Figura 75. Matriz de confusión para el modelo KNN con optimización extensiva de hiperparámetros, para el periodo de 21 días de variación porcentual.....	102
Figura 76. Matriz de confusión para el modelo Random Forest con optimización extensiva de hiperparámetros, para el periodo de 10 días de variación porcentual.....	103
Figura 77. Matriz de confusión para el modelo Random Forest con optimización extensiva de hiperparámetros, para el periodo de 17 días de variación porcentual.....	104
Figura 78. Matriz de confusión para el modelo Random Forest con optimización extensiva de hiperparámetros, para el periodo de 21 días de variación porcentual.....	104

ÍNDICE DE FÓRMULAS

Fórmula 1.	Fórmula de beneficios por acción	30
Fórmula 2.	Fórmula de certeza.....	34
Fórmula 3.	Fórmula de precisión	35
Fórmula 4.	Fórmula de recall	35
Fórmula 5.	Fórmula de F1-score	35
Fórmula 6.	Fórmula de sorpresa de beneficios por acción.....	44
Fórmula 7.	Fórmula de variación porcentual del precio de las acciones para reportes BFO	52
Fórmula 8.	Fórmula de variación porcentual del precio de las acciones para reportes AMC	52

RESUMEN

Los mercados financieros y la bolsa han sido desde el siglo pasado un poderoso instrumento tanto de construcción de grandes riquezas, cómo de especulación y pérdidas irrecuperables de dinero. Actualmente, y junto a las facilidades del comercio de activos financieros desde plataformas en línea, es cada vez más común la inclusión de nuevos inversores con poca experiencia y conocimiento, los cuales están muy expuestos a la volatilidad inherente a los mercados financieros. De hecho, la bolsa tiene periodos de especial volatilidad, siendo uno de ellos los periodos cuando las compañías están obligadas a publicar sus beneficios, deuda, y otras métricas financieras para lograr transparencia frente a sus inversores. Estos resultados suelen ser un catalizador para compras o ventas de las acciones de dichas compañías, dado que se suele pensar que las expectativas de los inversores pesan en el precio de las acciones, por lo que, de no cumplirse, o sobrepasarse, esto puede generar reacciones de compra o venta. Este tipo de situaciones son especialmente peligrosas para los inversores nuevos, sin experiencia, y con poco o nulo conocimiento de las mecánicas de la bolsa de valores, por lo que puede generar pérdidas irreparables al tomar malas decisiones. Se busca, de esta manera, desarrollar modelos, en este caso de comprar/vender/mantener las acciones de una empresa, que apoyen la toma de decisiones durante estos periodos de alta volatilidad, con un enfoque en los periodos de reportes financieros de las empresas del DJIA (promedio industrial Dow Jones) que cotizan en las bolsas de valores de estados unidos.

En este trabajo, se relacionó las expectativas de los inversores, y los beneficios de las empresas, y se generó una única variable llamada sorpresa de beneficios por acción, siendo esta variable usada para la predicción de los movimientos (para un periodo determinado) del precio de las acciones de cada compañía, además, se usó las etiquetas BUY/HOLD/SELL como variable objetivo(para modelos de clasificación), y se usó el porcentaje de variación del precio de la acción, para modelos de regresión. Para este fin, se crearon diferentes sets de datos, agrupando las empresas del DJIA, o usándolas como conjuntos separados, obteniendo un total de 8 conjuntos de datos para generar modelos con los diferentes métodos de aprendizaje de máquina. Además, se realizó la predicción usando técnicas de aprendizaje de máquina, tanto de regresión, cómo de clasificación, buscando este tipo de modelo predecir la decisión de BUY/HOLD/SELL para diferentes periodos de tiempo desde 1 día hasta 1 mes después de la publicación de reportes financieros trimestrales de la empresa o conjunto relacionado. Se realizaron pruebas preliminares tanto para los modelos de clasificación, como para algunos de regresión, y se acotaron los rangos de hiperparámetros usados para cada método de clasificación elegido: Random Forest, Decision tree, KNN, LightGBM, SVC, Logistic regression classifier. Así, se logró obtener la mejor combinación de método de aprendizaje de máquina y parámetros de entrada para los modelos que predicen las decisiones de BUY/HOLD/SELL de cada periodo de tiempo. Finalmente, se analizaron los resultados obtenidos y se eligió un set de datos para una optimización extensiva de hiperparámetros de tres de sus periodos con mejores resultados, para la posterior visualización de resultados por medio de matrices de confusión y las diferentes métricas relevantes al trabajo definidas, tales como F1-score, y certeza.

ABSTRACT

Financial markets and the stock market have been so far, this century, powerful capitalist tools for growing wealth, as well as for speculation, and irreversible monetary losses. Currently, and along an always improving easy way to handle assets and financial instruments given the currently available electronic trading platforms, it is usually more and more normal for the ever-growing number of inexperienced and, knowledge lacking, investors, who are exposed to the constant intrinsic volatility of the financial markets. Furthermore, the stock market (mainly), has especially volatile periods, where one of these is during financial reports publications given that stock market companies are legally bound to state openly values such as debt, income, net earnings, liabilities, as well as other metrics to appeal to their investors as transparent. These results are a catalyst to stock “rallies” and “Sell-offs”, given that it is a common thought that investor expectations are related to stock price, so that, given that unexpected results appear, financial reports could cause successive BUY or SELL reactions. This kind of situation is especially dangerous for new investors without enough knowledge, so it can, not strangely, cause irreversible monetary losses were they to decide on impulse. It is intended to develop models, of stock movement expectations, that support decision-making during especially high volatility periods, focusing on quarterly financial reports periods of DJIA (Dow Jones Industrial Average) companies that make part of America’s stock markets.

In this project, investor expectations and companies’ earnings are linked, and a single variable named earnings surprise was calculated to be used in the stock’s movement prediction (for a to-be-defined period) for every company. To achieve this, various datasets were created grouping companies, as well as using them as standalone datasets, so that 8 datasets were created to train and compare various machine learning methods. Furthermore, the prediction was done using both regression and classification machine learning methods, so that the former gave a BUY/HOLD/SELL prediction for different periods from 1 day up to 1 month after a company’s earnings report public release. Preliminary tests were done for classification models, as well as some regression models, and hyperparameter ranges were reduced for every of the used classification methods: Random Forest, Decision tree, KNN, LightGBM, SVC, and Logistic regression classifier. Thus, it was obtained the best combination of machine learning methods and input parameters for the BUY/HOLD/SELL models of every period. Finally, the results were analyzed, and a single dataset was picked for an extensive hyperparameter optimization of three periods, whichever achieved the best results. Results were visualized using confusion matrixes as well as several relevant metrics such as F1-score and accuracy.

1 INTRODUCCIÓN

Los mercados financieros y la bolsa han sido desde el siglo pasado un poderoso instrumento tanto de construcción de grandes riquezas, cómo de especulación y pérdidas irrecuperables de dinero. Esto explica tanto la considerable porción de la población que hoy invierten de alguna manera en el mercado bursátil, aproximadamente 58% de adultos en Estados Unidos, como la lenta recuperación del número relativo de adultos que invierten en la bolsa después de escenarios tan deprimentes como el estallido de la “burbuja de las .com” y la crisis financiera del 2008[1]. En la figura 1 se puede ver la evolución de la proporción de adultos estadounidenses que poseen algún tipo de acción, desde 1998 hasta 2022.

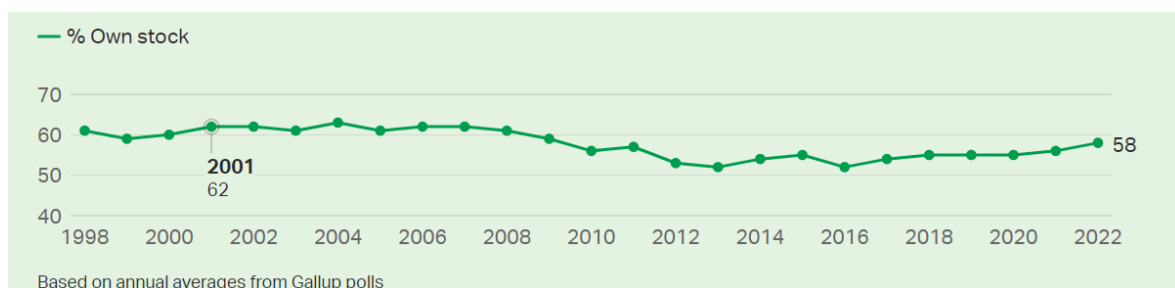


Figura 1. Evolución de la proporción de adultos estadounidenses con acciones en la bolsa[1].

Además de esto, la introducción de la tecnología moderna y la internet, en los mercados bursátiles, ha permitido cambios sustanciales. Desde la facilidad para consultar información financiera, hasta menores requisitos para abrir una cuenta de inversión, y la disminución de las comisiones por transacción[2]. Ahora los mercados financieros son más “democratizados” que cualquier otro momento en la historia. De hecho, tal es el cambio que, de costar cientos de dólares por transacción en la bolsa de valores de Estados Unidos antes de 1975, hoy se paga inclusive \$0,01 por acción (costo de transacción por acción), y sin costos fijos (abolidos en 1975) [3].

Así, la democratización financiera ha traído muchos beneficios para pequeños inversores. Sin embargo, igualmente se ha introducido mucha incertidumbre y volatilidad a los mercados bursátiles. Cómo es de esperarse, esto viene principalmente de mano de pequeños inversionistas, los cuales son las llamadas “manos débiles” del mercado bursátil, esto debido a que vienen por el objetivo, generalmente, a corto/medio plazo de ganar dinero, lo cual puede causar fácilmente los llamados “sell-offs” o “rallys”, en el mercado bursátil o con una acción en especial.

Solo en el año 2020 y 2021, hubo múltiples eventos promovidos por grupos de pequeños inversores que causaron grandes pérdidas a grandes entidades financieras (caso Game Stop y AMC). Así, son en muchas ocasiones eventos de “rallys” y “sell offs”, como los mencionados, los que han llevado a muchos pequeños inversores a retirarse definitivamente por una pérdida que posiblemente pudo haber sido recuperada en el largo plazo. La figura 2 muestra una gráfica del caso Game Stop.



Figura 2. Evolución del precio de la acción GME desde octubre 1 del 2020[4]

La figura 2 muestra cómo el furor del momento logró que pequeños inversores llevarán el precio de una acción de 17.08 dólares a un máximo de 483 en cuestión de días, lo que hizo que cientos de personas sufrieran pérdidas irrecuperables durante la subsecuente caída[4].

Un caso muy diferente a este, cuando la volatilidad de los mercados bursátiles puede ser muy alta, pero de manera relativamente predecible, es en publicaciones trimestrales financieras de algunas de las empresas más importantes. Este acontecimiento es investigado a mayor profundidad en el artículo “Volatility Skew, Earnings Announcements, and the Predictability of Crashes” de Andrew Van Buskirk[5]. Durante esta época hay tres posibilidades: las expectativas de los inversionistas están alineadas con los resultados financieros, son sobrepasadas, o son aplastadas por resultados financieros negativos. Estos resultados hipotéticamente afectan significativamente el movimiento de las acciones de las empresas con las que están relacionadas en el corto y medio plazo.

En este trabajo se compararon diversas implementaciones de modelos de aprendizaje de máquina tanto de regresión como, principalmente, de clasificación. Estos modelos usaron datos históricos de la sorpresa de beneficios por acción y la variación porcentual después de la presentación de cada reporte financiero trimestral. Dicha predicción de variación porcentual se realizó en periodos desde 1 día hasta 30 días posteriores a la presentación de resultados financieros, consecuente a esto se obtuvieron los mejores resultados después de una optimización general de hiperparámetros y de una selección tanto de métodos de aprendizaje de máquina, cantidad de periodos de sorpresa de beneficios por acción a usar en cada modelo, y periodo en días de predicción de la variación porcentual del precio de las acciones. Este modelo usó etiquetas BUY/HOLD/SELL para la clasificación, y se realizó una optimización extensiva de hiperparámetros para obtener los mejores resultados posibles de los modelos. Este modelo es una aproximación básica a un sistema que podría permitir, en un futuro, predecir la viabilidad de comprar o vender acciones, así como servir de herramienta en la toma de decisiones durante periodos de reportes financieros trimestrales

de, en este caso, la empresa Merck para la cual se consiguieron modelos de aprendizaje de máquina que lograron buenos resultados en la predicción de etiquetas BUY/HOLD/SELL.

Un modelo como el que se propone en este trabajo puede ser útil para proteger los intereses de los pequeños inversores que pudiesen sentir miedo durante grandes periodos de volatilidad, en este caso los causados cuando las empresas públicas (empresas que cotizan en bolsa) fallan en cumplir con las expectativas de sus inversores.

2 JUSTIFICACIÓN

La forma en que las personas realizan operaciones en el mercado bursátil ha cambiado drásticamente, ahora estas operaciones se realizan principalmente en línea. Inclusive, J.P. Morgan estima que cerca del 80% de transacciones son realizadas por computadores por complejos algoritmos preestablecidos[6]. Es una necesidad de los inversores en el mercado bursátil el informarse, apoyándose en noticias, informes financieros, rumores, análisis históricos, y otros medios. Además, tradicionalmente ha habido índices para medir la volatilidad del mercado, sin embargo, existe la necesidad de generar mecanismos para ayuda en la toma de decisiones, especialmente en la predicción del comportamiento de acciones durante periodos donde es usual que haya comportamientos bruscos en el precio de las acciones. Siendo estos momentos de alta tensión para pequeños inversionistas que temen perder su capital, llevándolos en muchos casos a quedar “afuera del juego”, que significa salir de forma casi completa y por un periodo indefinido de tiempo de las inversiones en los diferentes mercados bursátiles[7].

Por esta razón, tomar decisiones acertadas en el mercado financiero es vital. Así, a pesar de que invertir en el Dow Jones tiene una rentabilidad histórica anualizada (7,69%) mayor a cualquier otro instrumento financiero de renta fija (bonos del tesoro, CDT, ahorro en un banco), estos periodos y malas decisiones pueden llevar a grandes pérdidas y en muchos casos a retirarse permanentemente del mercado bursátil[8].

Esto se da en gran medida debido a los efectos no deseados de la democratización de los mercados financieros, ya que esta ha venido de la mano de nuevos pequeños inversores que son analfabetas financieramente, e inclusive han generado eventos tan absurdos como la carrera por los “meme stocks”, generando así en muchos casos pérdidas financieras irreparables[9].

Este modelo servirá para un análisis pre-trade, el cual proporciona a los inversores los datos necesarios para realizar operaciones informadas. Este modelo proporciona a los inversores estimaciones de riesgos y la magnitud de subidas o caídas en el corto plazo de las acciones del Dow Jones de las que se tenga información, en este caso solamente se hará con una acción. Por este motivo, un modelo como este sería de gran utilidad para este proceso de toma de decisiones durante momentos de tensión en el mercado bursátil, en este caso durante la presentación de resultados financieros trimestrales.

Una herramienta que predice una tendencia de volatilidad (BUY/HOLD/SELL) del precio de acciones particulares puede ser de mucha ayuda para pequeños inversores, los cuales podrían tomar mejores decisiones en algunos de estos periodos, principalmente cuando son eventos relativamente predecibles con información histórica y actual, en este caso haciendo referencia a las presentaciones trimestrales de resultados financieros. Dicha herramienta podría hacer que el mercado bursátil sea un lugar más seguro para inversores con poca información concisa y nuevos a la volatilidad de los mercados bursátiles.

3 PLANTEAMIENTO DEL PROBLEMA

El encontrar patrones en el comportamiento histórico del mercado bursátil es útil para la toma de decisiones. Se ha desarrollado un campo de estudio completo conocido como análisis técnico, el cual estudia patrones en los movimientos de las acciones teniendo en cuenta movimientos históricos en su precio. Es importante notar que estos movimientos se suelen relacionar con las reacciones de inversionistas ante distintas noticias y eventos. En la actualidad, esto puede ir desde especulaciones en las redes sociales, noticias macroeconómicas y, el tema a abordar en este trabajo, la presentación de resultados financieros de las empresas que cotizan en bolsa.

Hay un viejo dicho en Wall Street que dice: “los mercados financieros están dirigidos por dos poderosas emociones: miedo, y codicia” [10]. Uno de los momentos más importantes que representa esto es, justamente, durante la presentación de resultados financieros, dónde las expectativas y pesimismo llega a los pequeños inversores, causando los que denominamos “rallies” y “Sell-offs”. Este fenómeno humano suele causar grandes pérdidas para muchos inversores que inclusive pueden “salirse del juego” al simplemente tener una pérdida lo suficientemente grande que desmotive futuras inversiones en el mercado bursátil. Sin embargo, aún este tipo de eventos puede ser usado para aplicar estrategias de trading a corto plazo[7].

El conocer este hecho, es útil para el desarrollo de sistemas para apoyar en la toma de decisiones durante este tipo de eventos. Por lo tanto, se busca en este trabajo encontrar patrones históricos que puedan estimar posibles movimientos volátiles durante la época de presentaciones financieras de las empresas, esto para brindarles una herramienta a los nuevos inversores que les ayuden a evitar pérdidas financieras, así como se quiere brindar información útil que les ayude conservar o inclusive aprovechar la volatilidad para tener ganancias de capital, a discreción del inversor y sus objetivos particulares de inversión.

Además, es necesario comprender el estado del arte en las estrategias de toma de decisiones. En el mercado bursátil existen cuatro tipos de información para la toma de decisiones. Estos tipos son: análisis fundamental(análisis basado en aspectos económicos de la empresa como deuda, crecimiento anualizado, presencia de marca, porcentaje de mercado), análisis técnico(análisis basado en el comportamiento histórico del precio de las acciones de la empresa, incluye análisis de gráficos y análisis de indicadores técnicos como la media móvil), y análisis de sentimiento (análisis basado en información cualitativa procesada para determinar “estados de ánimo” de los mercados que pueden ocasionar “rallies” y “Sell-offs”). Los traders examinan los datos técnicos para comprender las tendencias de los precios, la presión de los precios y los niveles de resistencia, otras herramientas son el consenso del mercado, las posiciones de los traders y las reacciones ante las noticias para determinar el sentimiento del mercado. El comportamiento del mercado y el impacto de las instituciones en la liquidez y la volatilidad, como los bancos centrales, los grandes fondos y los

comerciantes de alta frecuencia, también cambian la dinámica de los precios del mercado, por lo que también se les realiza un seguimiento[11]. En consecuencia, se trabajó en entender los mercados financieros y su dinámica, porque se quiere saber su funcionamiento y métodos actuales de toma de decisiones para una mejor implementación de un algoritmo de aprendizaje de máquina y la elección de los parámetros, relacionados a reportes financieros, que tomaremos en el modelo supervisado.

Se compararon diversas implementaciones de modelos de aprendizaje de máquina tanto de regresión como, principalmente, de clasificación. Estos modelos usaron datos históricos de la sorpresa de beneficios por acción y la variación porcentual después de la presentación de cada reporte financiero trimestral. Dicha predicción de variación porcentual se realizó en periodos desde 1 día hasta para 30 días posteriores a la presentación de resultados financieros, consecuente a esto se obtuvieron los mejores resultados después de una optimización general de hiperparámetros y de una selección tanto de métodos de aprendizaje de máquina, cantidad de periodos de sorpresa de beneficios por acción a usar en cada modelo, y periodo en días de predicción de la variación porcentual del precio de las acciones. Este modelo usó etiquetas BUY/HOLD/SELL para la clasificación, y se realizó una optimización extensiva de hiperparámetros para obtener los mejores resultados posibles de los modelos. Este modelo es una aproximación básica a un sistema que podría permitir, en un futuro, predecir la viabilidad de comprar o vender acciones, así como servir de herramienta en la toma de decisiones durante periodos de reportes financieros trimestrales de, en este caso, la empresa Merck para la cual se consiguieron modelos de aprendizaje de máquina que lograron buenos resultados en la predicción de etiquetas BUY/HOLD/SELL.

4 OBJETIVOS

4.1 OBJETIVO GENERAL

Generar y comparar al menos dos modelos de aprendizaje de máquina, que encuentren patrones históricos de una compañía del DJIA (promedio industrial Dow Jones), y nos permitan con variables financieras reales (reporte financiero trimestral), y las expectativas de los pequeños inversionistas (estimados de analistas especializados), estimar si se espera en el corto plazo (al menos tres periodos a definir) subidas o caídas en el precio de la acción de la que se tienen estos datos.

4.2 OBJETIVOS ESPECÍFICOS

Estos objetivos se definen en pasos como sigue:

1. Elegir las compañías del Dow Jones a utilizar, y elegir las variables financieras más relevantes para el modelo de aprendizaje de máquina.
2. Obtener los datos de las variables financieras elegidas, recopilar los datos históricos requeridos, y definir los periodos a evaluar.
3. Elegir los métodos de aprendizaje de máquina a comparar.
4. Preprocesar los sets de datos e implementar los métodos de aprendizaje de máquina.
5. Validar los métodos de aprendizaje de máquina con métricas definidas, usando fuentes como el curso de IBM “Aprendizaje de máquina con Python: una introducción práctica”.
6. Analizar y comparar los resultados obtenidos.

5 MARCO TEÓRICO

5.1 TÉRMINOS Y CONCEPTOS DE LA BOLSA DE VALORES.

Hay muchos términos técnicos en economía vinculados a este trabajo, los mercados financieros requieren una comprensión suficiente de su dinámica, por lo tanto, el marco teórico ayuda a establecer estos conceptos, además de conceptos relacionados con el denominado “trading” y estrategias financieras, que en muchos casos usan inteligencia artificial.

5.1.1 Mercados financieros

Los mercados financieros se refieren en términos generales a cualquier mercado donde se negocian valores, incluido el mercado de valores, el mercado de bonos, el mercado de divisas y el mercado de derivados, entre otros. Los mercados financieros son vitales para el buen funcionamiento de las economías capitalistas. Los mercados financieros se crean mediante la compra y venta de numerosos tipos de instrumentos financieros, incluidas acciones, bonos, divisas y derivados (consulte la Figura 3 como referencia) [12].

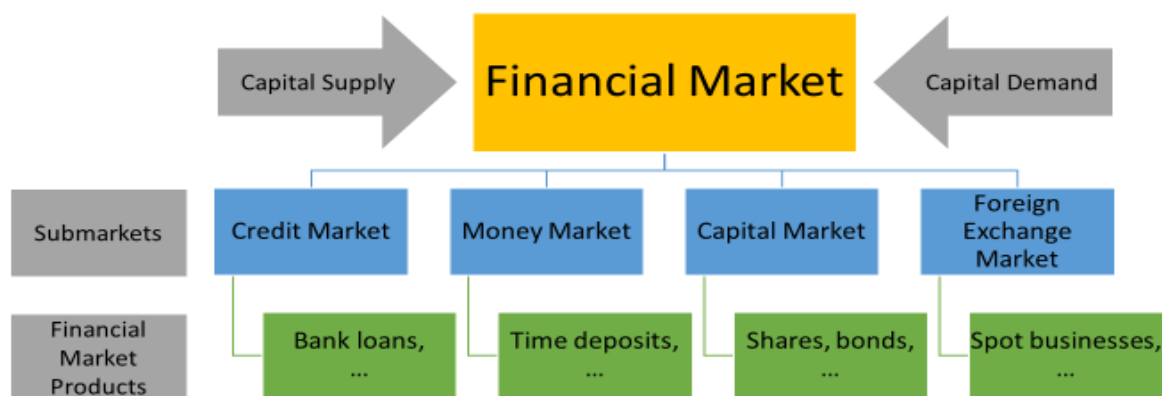


Figura 3. Estructura y componentes de los mercados financieros [12]

5.1.2 Índice Bursátil

Al referirse a un índice bursátil se hace referencia a un indicador que representa el sentimiento de mercado, el cual puede referirse a un sector específico, como el de empresas tecnológicas en crecimiento, tanto al conjunto de empresas de pequeña, media o gran capitalización de un país en específico. Este indicador usa diferentes parámetros como el precio de algunas acciones de compañías previamente elegidas para formar parte de este índice. Este indicador suele ser usado para medir el riesgo de un mercado por parte de pequeños y grandes inversores. Existen índices de todo tipo, y la forma en que cada uno se calcula es enteramente diferente. En general, los dos tipos de índices más conocidos son: índice de precios ponderados, y los índices de capitalización ponderada[13].

Los índices de precio ponderado son una media aritmética, mientras que los índices de capitalización ponderada tienen en cuenta el peso de la capitalización de mercado de cada empresa que hace parte del índice.

5.1.3 Acción

Una acción es un instrumento financiero que representa una fracción de propiedad de una empresa o corporación que cotiza en la bolsa de valores. Las acciones se clasifican en acciones ordinarias, preferenciales, y privilegiadas. Las acciones ordinarias son las que serán de importancia para la realización de este trabajo. Las acciones ordinarias dan titularidad a una persona o entidad de una fracción de las empresas que se compra la acción, por tanto, derechos de votación, activos, y beneficios. Las acciones se pueden obtener mediante un bróker o intermediario con la bolsa de valores, un ejemplo de bróker es eToro [14].

5.1.4 Promedio Industrial Dow Jones

El Promedio Industrial Dow Jones es un índice de precios ponderados constituido por las 30 empresas con mayor capitalización bursátil de la bolsa de valores de Nueva York. Es el índice de mayor prestigio en el mundo de inversiones en el mercado bursátil. Este grupo de acciones está conformado por compañías especialmente sólidas financieramente, ya que inicialmente hacen parte de un selecto grupo llamado S and P 500, las cuales tienen que, inclusive, cumplir con características de capitalización de mercado, oferta al público, liquidez y rentabilidad. El Dow Jones solo ha cambiado su estructura 55 veces desde su comienzo en mayo 26 de 1896[15].

5.1.5 Reportes financieros

Las compañías que cotizan en bolsa están legalmente obligadas a reportar públicamente sus actividades financieras más relevantes, esto, impuesto por la comisión de bolsa y valores correspondiente al país donde cotiza cada compañía. Estos reportes dan cuenta del rendimiento por periodos de la compañía, usualmente por años y por trimestres. Estos reportes son sujeto de auditorías debido a su uso para determinar impuestos, financiamiento, o inversión. Existen diferentes tipos de reportes: para compañías con ánimos de lucro, y para compañías sin ánimo de lucro. Los reportes para compañías con ánimo de lucro incluyen: balance general, estado de resultados, estado de flujo de efectivo, y estado de cambios en el patrimonio. Para la inversión, estos reportes, principalmente el balance general, es usado para decisiones de compra o de venta de acciones. Algunas de las métricas usadas para estas decisiones son: nivel de deuda, beneficios por acción, y crecimiento interanual[16].

5.1.6 Beneficios por acción

Es la porción de ganancias que corresponde a cada una de las acciones de una empresa. Este indicador sirve para medir de forma precisa la rentabilidad de una empresa desde el punto de vista de un inversor. En inglés es EPS (earnings per-share). Este indicador es relativamente preciso debido a que usa las ganancias brutas de una compañía y le reduce los gastos operativos, salarios, EBITDA (Interés, impuestos, depreciación, y amortización) y otros

componentes que pueden nublar la visión acerca de la rentabilidad de una empresa[17]. La fórmula 1 muestra cómo se calculan los beneficios por acción o eps.

$$\text{eps} = (\text{beneficio neto}) / (\text{número total de acciones})$$

Fórmula 1. Fórmula de beneficios por acción

5.1.7 EPS estimadas (Pre- resultados financieros oficiales)

Se refiere al valor estimado de beneficios por acción, antes de las publicaciones financieras oficiales. Estos estimados son publicados por analistas especializados que trabajan en compañías de investigación en inversiones. Usando estos estimados, los analistas pueden evaluar el flujo de caja y encontrar un valor presente aproximado de la compañía.

En este tipo de estimados se tienen en cuenta las operaciones de una compañía, además se evalúa la administración de la gerencia de la compañía, y otra información relevante para calcular de forma aproximada las futuras ganancias por acción de la compañía.

Los inversores suelen usar esta información para hacer estimaciones del potencial de crecimiento de una compañía, así como para tomar decisiones de trading[18].

5.1.8 Análisis Técnico

Es un tipo de análisis que usa los movimientos de los precios de las acciones, indicadores y gráficos para predecir el comportamiento futuro del mercado bursátil[19].

5.1.9 Análisis Fundamental

El análisis fundamental es un tipo de análisis aplicado al mercado de valores que busca establecer el precio de una acción o de un instrumento financiero a través de variables financieras que afecten su valor. Este tipo de análisis estudia la viabilidad de compra o venta de un instrumento financiero a través de un estudio de la situación macroeconómica, la situación del país donde tiene su sede la compañía la compañía de la que se está pensando adquirir un instrumento financiero, la situación del sector de la compañía, y métricas y parámetros financieros importantes de la compañía como beneficios por acción, deuda, y crecimiento interanual[20].

5.1.10 Trader

La definición comúnmente utilizada de negociación es: un “trader” intenta obtener ganancias al realizar transacciones o apostar en los movimientos de precios, anticipando la tendencia futura de los precios de la manera más correcta posible [21].

5.1.11 Trading algorítmico

Se refiere a la ejecución computarizada de instrumentos financieros: ej. opciones, acciones. Además, es implícito el uso de un set de reglas e instrucciones que sirvan para llevar a cabo la ejecución de la transacción. En este caso las órdenes se colocan en el algoritmo para ejecución. En el algoritmo luego se divide la orden grande en órdenes pequeñas para su ejecución completa en un periodo de tiempo definido. El objetivo de esta implementación es

asegurar que las decisiones de inversión sean consistentes con los objetivos previamente propuestos por el trader, mientras se optimizan los costos generales de las transacciones[21].

5.1.12 Trading a corto vs. largo plazo

El horizonte de inversión es un tema importante que determina fundamentalmente el tipo de “trading”. Normalmente, hay dos grupos: day trading y swing trading.

En el caso de operaciones a corto plazo, el “trader” abre y cierra sus operaciones individuales en unos pocos minutos u horas. Dado que el período especulativo generalmente se limita a un día, estas operaciones se denominan transacciones diarias.

Si el período de preservación de una posición es de unos pocos días a semanas o incluso meses, se denomina swing trading.

Las posibilidades de aplicación de estos dos tipos de trading y los requisitos respectivos para los traders son fundamentalmente diferentes.

El trading intradiario suele ser menos adecuado para las personas empleadas debido a las limitaciones de tiempo, ya que a menudo es necesario estar atento a los gráficos de precios durante el horario de apertura del mercado. Por ejemplo, Cuando la Bolsa de Fráncfort abre a las 9:00 a.m., la mayoría de las personas en Alemania probablemente estén en el trabajo y no puedan seguir los movimientos de precios activamente. Sin embargo, un day trader alemán podría, alternativamente, cambiar a otros mercados bursátiles y negociar activamente en las bolsas americanas o asiáticas después de su horario laboral [21].

5.1.13 Horarios de la bolsa de valores

Una bolsa de valores es un mercado dónde se pueden comprar y vender activos financieros durante determinadas horas del día. Además, asegura la correcta marcación del precio de los activos y el intercambio ordenado de activos financieros. Sin embargo, cada bolsa de valores tiene horarios de apertura y cierre determinados. La bolsa dónde se encuentran las acciones del Dow Jones, por ejemplo, es la bolsa de Nueva York, cuyos horarios van desde las 9:30 am hasta las 4:00 pm, hora del este. Caso contrario a, por ejemplo, la bolsa de Frankfurt, cuyos horarios van desde las 9:00 am hasta las 5:30 pm de la hora central europea de verano[22].

5.1.14 “Sell-off”

Un “Sell-off” se da cuando un número muy alto de instrumentos financieros (acciones, opciones, ETFs) se venden en un periodo muy corto de tiempo. Esto hace que el precio del instrumento baje de forma escalonada o, en casos extremos, exponencial. Además, tan pronto como estos instrumentos sean ofrecidos en mayor medida de lo que otras personas están dispuestas a comprar, esto ocasiona un rápido declive en el precio del instrumento. Esto es causado en gran medida por factores psicológicos de las personas.

Estos eventos llamados “Sell-off” pueden ser causados por distintos factores. Algunos de estos son: problemas macroeconómicos, noticias de la empresa, crisis financieras, rumores, o malos resultados financieros[23].

5.1.15 “Rally”

Este periodo es opuesto a un “Sell-off”. En este caso hay un aumento escalonado o exponencial en el precio de los instrumentos financieros, tales como acciones, ETFs, futuros y opciones. A diferencia del otro caso, la demanda es más alta que la oferta y el precio del instrumento aumenta. Esto igualmente puede ser causado por sucesos macroeconómicos, noticias de la empresa, rumores, o buenos resultados financieros[24].

5.1.16 “Crash”

Un “crash” es una caída considerable en el valor del mercado bursátil o de la acción de una compañía en particular. Un “crash” es usualmente asociado con expectativas infladas de los participantes del mercado. Sin embargo, puede darse por un rango grande de causas cómo: factores macroeconómicos, resultados financieros negativos, o escándalos[25].

5.1.17 Potencial de ganancias

En el trading, es posible apostar tanto a la subida como a la baja de los precios y, por lo tanto, obtener ganancias incluso si el mercado de valores u otro mercado financiero cae.

Si el trader cree, con base en los resultados de su análisis, que las acciones de Apple o el tipo de cambio entre el EUR y el USD se apreciarán, puede comprar las acciones o las monedas correspondientes hoy y venderlas a un precio más alto más tarde para obtener ganancias.

Por otro lado, si el trader asume que los precios bajarán, puede tener esto en cuenta en su plan de trading e ingresar a una llamada operación de venta (corta) con la que se beneficia cuando los precios de los activos subyacentes se deprecian. Por otro lado, si se tienen buenas perspectivas a largo plazo, el trader podría usar la caída momentánea a su beneficio a largo plazo.

5.1.18 Cobertura financiera

Hablamos de cobertura cuando se tiene una estrategia para contrarrestar efectos adversos (pérdida de capital) de una operación financiera de inversión, realizando una operación opuesta (si se compró, se vende y viceversa). Este tipo de operaciones se realizan y van de mano con estrategias para disminuir el riesgo y limitar pérdidas sobre operaciones financieras[26].

5.1.19 Volatilidad

La volatilidad es una medida estadística de la dispersión de los rendimientos de un valor o índice de mercado determinado. En la mayoría de los casos, cuanto mayor sea la volatilidad, más riesgoso será el valor. La volatilidad a menudo se mide como la desviación estándar o la varianza entre los rendimientos de ese mismo valor o índice de mercado[27].

5.2 CONCEPTOS Y MODELOS DE APRENDIZAJE DE MÁQUINA

El objetivo de esta investigación requiere una mirada extensa en diversas técnicas utilizadas en la ciencia de datos, y algunas de las métricas que se usan para su evaluación. Así, las principales se resumieron para establecer los conocimientos necesarios para este proyecto.

5.2.1 Modelos de regresión lineal y polinómico

En este tipo de modelo de aprendizaje automático se usa el valor de una o más variables independientes para predecir el valor de otra. Para este método se estiman coeficientes para así formar una ecuación lineal que calcule de forma aproximada el comportamiento del sistema a predecir. Siendo el caso polinómico muy similar, pero con el uso de una ecuación polinómica para calcular de forma aproximada el comportamiento del sistema. Usando este tipo de métodos se disminuye las discrepancias entre los valores de salida previstos y los reales.

Según IBM, un caso de éxito de este tipo de aplicación es para predecir las ventas anuales totales de un vendedor. Sin embargo, en este caso se usan múltiples parámetros como edad, educación y años de experiencia[28].

5.2.2 Modelo de aprendizaje de máquina de clasificación

Este tipo de modelo permite identificar y caracterizar objetos o datos de naturaleza similar o con características definidas similares. Además de esto, también puede usarse para clasificar eventos, por ejemplo, una anomalía en el funcionamiento del internet, o inclusive una fractura de un hueso. Para realizar este tipo de modelos se tiene que buscar estructuras y atributos de clasificación relevantes y con sentido, de forma que se pueda clasificar el rango definido de posibles objetos o datos dentro de los atributos o características previamente definidas. Al realizar este tipo de modelo es relevante proporcionar nombres que identifiquen de manera significativa los atributos[29]. En el caso de esta aplicación en el mercado bursátil, es común encontrar modelos que clasifican una acción en tres categorías: BUY, HOLD, SELL, dependiendo de las características actuales y movimientos pasados que puedan implicar movimientos en los precios de las acciones.

5.2.3 Hiperparámetros

Los hiperparámetros contienen información que puede afectar el proceso de entrenamiento de un modelo de aprendizaje de máquina. Por ejemplo, al realizar una red neuronal de aprendizaje profundo, se debe decidir cuántas capas ocultas de nodos se incluirán entre las capas de entrada y salida. Adicionalmente, se debe determinar el número de nodos a usar en cada nivel. Estas variables no están relacionadas con los datos de entrenamiento. Estas variables son variables de configuración, afectan la forma en que se lleva a cabo el entrenamiento de los modelos de aprendizaje de máquina, y es a lo que nos referimos con hiperparámetros. Estos hiperparámetros, además, difieren para cada método de aprendizaje de máquina y se pueden consultar en la documentación de cada uno[30].

5.2.4 Optimización de hiperparámetros

La optimización de hiperparámetros ejecuta múltiples pruebas en un único trabajo de entrenamiento de datos. Cada prueba consiste en un entrenamiento completo de un modelo de aprendizaje de máquina con valores diferentes de hiperparámetros en límites o rangos que sea especificada. La optimización de hiperparámetros optimiza una única variable objetivo que se especifique. Esta variable es maximizada durante el entrenamiento del modelo de aprendizaje de máquina y las pruebas con diferentes combinaciones de hiperparámetros. Una métrica común a maximizar es la certeza del modelo, aunque igualmente se pueden usar otras métricas como F1-score[30].

5.3 MÉTRICAS DE DESEMPEÑO DE APRENDIZAJE DE MÁQUINA

5.3.1 R2-Square

La métrica R2 square es usada para medir la dispersión de los datos de la línea de regresión obtenida después de implementar el método de aprendizaje de máquina de regresión. Esta métrica es típicamente usada para medir el rendimiento de los modelos de aprendizaje de máquina de regresión y típicamente tiene un valor entre -1 y 1, dónde 1 significa una relación directa muy fuerte, y 0 una relación inversa muy fuerte [31]. Sin embargo, este valor puede ser superior a 1 o inferior a -1 cuando el modelo encontrado, después de aplicar el método de regresión, es muy malo para acoplarse a la dispersión de los datos.

5.3.2 TP, TN, FP, Y FN

Estos son términos usados para describir la correcta o no predicción de la variable objetivo que genera un modelo de aprendizaje de máquina y entrenado. Esto se define dependiendo de los valores reales. A continuación, se da una breve explicación de cada término[32]:

- Verdadero Positivo (TP): Predicho verdadero y verdadero en realidad.
- Verdadero Negativo (TN): Predicho falso y falso en realidad.
- Falso Positivo (FP): Predicho verdadero y falso en realidad.
- Falso Negativo (FN): Predicho falso y verdadero en realidad.

5.3.3 Certeza o exactitud

La certeza es en muchos casos la métrica más importante al evaluar modelos de aprendizaje de máquina de clasificación, ya que se puede saber la frecuencia con que el clasificador es correcto. Su fórmula consiste en dividir el total datos de prueba entre la suma de los verdaderos positivos (TP) y los verdaderos negativos (TN). La fórmula de certeza se ve en la fórmula 2[32].

$$Certeza = \frac{(TP + TN)}{total}$$

Fórmula 2. Fórmula de certeza

5.3.4 Precisión

Este valor es especialmente útil para evaluar clases particulares cuando hay un desequilibrio entre clases. Además, en este caso se mide la frecuencia con qué la predicción es correcta para una determinada clase. La fórmula 3 muestra el cálculo del valor de precisión[32].

$$Precisión = \frac{TP}{TP + FP}$$

Fórmula 3. Fórmula de precisión

5.3.5 Recall

Esta métrica, por otra parte, nos da la tasa positiva verdadera, la cual es la proporción de los verdaderos positivos a todo lo positivo. La fórmula 4 muestra el cálculo del valor de recall[32].

$$Recall = \frac{TP}{TP + FN}$$

Fórmula 4. Fórmula de recall

5.3.6 F1-score

La métrica F1-score o puntuación F1 es la media armónica de la precisión y el recall, siendo 1 su mejor valor y 0 el peor. Este valor tiene la ventaja de mitigar los valores atípicos extremos y a agravar el impacto de los pequeños. Es muy útil debido a que permite evaluar tanto la métrica precisión como recall en un único valor. La fórmula 5 muestra el cálculo del valor de F1-score[32].

$$F1 - score = 2 \times \frac{Precisión \times Recall}{Precisión + Recall}$$

Fórmula 5. Fórmula de F1-score

5.3.7 Matriz de confusión

Es una representación matricial de los resultados de las predicciones en la etapa de pruebas y validación del entrenamiento de los métodos de aprendizaje de máquina. Esta matriz consiste en los valores reales de cada una de las clases, así como los valores predichos por el modelo. Esta matriz es muy útil para entender el comportamiento en términos de predicciones de las diferentes clases respecto a los valores reales, y puede dar indicios de cómo se pueden optimizar los modelos de aprendizaje de máquina para mejorar el comportamiento de, por ejemplo, la predicción de una clase, en especial con las métricas F1-score, recall, o precisión. En la figura 4 se puede ver TP, TN, FP, y FN en una matriz de confusión con una variable objetivo de 2 clases[32].

		Predicted Values	
		Negative	Positive
Actual Values	Negative	TN True Negative	FP False positive
	Positive	FN False Negative	TP True Positive

Figura 4. TP, TN, FP, y FN en matriz de confusión con una variable objetivo de 2 clases

5.4 ESTADO DEL ARTE

Para la realización de este trabajo se consultaron diferentes trabajos enfocados en la predicción aplicada al mercado de valores, especialmente acciones. Además, se enfocó en la búsqueda de trabajos que usaron métodos de clasificación, no excluyendo otro tipo de trabajos. Estos trabajos se resumen a continuación.

5.4.1 Stock Price Movements Classification Using Machine Learning and Deep Learning Techniques-The Case Study of Indian Market

En este artículo se introducen dos de los grandes problemas de grandes fondos de inversión y de cobertura, estos son: identificar parámetros técnicos relevantes para determinar futuros movimientos de los precios de acciones, por otra parte, está identificar un modelo de predicción preciso para movimientos del precio de acciones. Se proponen en este artículo el uso de técnicas de aprendizaje de máquina y aprendizaje profundo. En este trabajo se encuentra que el comportamiento del modelo de aprendizaje profundo es mejor que las técnicas de aprendizaje de máquina. Este trabajo tomó información del mercado de acciones de la India y mostró mejoras significativas en la precisión de clasificación de acciones respecto a métodos tradicionales e introducidos en el trabajo. Se usaron las métricas F-Measure y Certeza. En el caso de la métrica F-Measure, se obtuvieron resultados para los distintos sets de datos entre 0.69 y 0.84, mientras que con la métrica de certeza se obtuvieron resultados entre 0.68 y 0.84[33].

5.4.2 A Regression Model to Predict Stock Market Mega Movements and/or Volatility Using Both Macroeconomic Indicators & Fed Bank Variables

En este trabajo se actualiza un método de regresión lineal multivariable pasado para predecir el valor del S & P 500, el cuál es un índice del mercado de valores estadounidense, el cual

está compuesto por 500 compañías líderes que cotizan en bolsa, principalmente las compañías con mayor capitalización del mercado(mayor valor)[34]. En este trabajo se consigue una mejora estadística en este modelo, además de que varios parámetros son mejorados tales como el coeficiente de determinación y el “F Stat” del modelo. Además, en este trabajo se muestra como esto es logrado de igual manera al simplificar el modelo, pues se reducen el número de variables independientes. Este modelo predice grandes desviaciones antes de un “crash” días antes de que el evento llegue a suceder, además de predecir cuándo se llega al “market bottom” con días de anterioridad. De hecho, aún al ser un modelo lineal, este modelo consigue un r-square de 0.95[35].

5.4.3 Algorithmic Financial Trading with Deep Convolutional Neural Networks: Time Series to Image Conversion

En este trabajo se calculan 15 indicadores técnicos con 15 amplitudes de periodos diferentes. Posteriormente, se convierte las nuevas características en imágenes de 15x15 y se nombra la información como BUY/HOLD/SELL. Finalmente, se entrena el modelo de redes neuronales convolucionales como cualquier otro problema de clasificación de imagen. Este modelo usa conceptos como el promedio simple móvil y se definen en diferentes periodos de tiempo. Además, se consiguieron resultados de certeza entre 0.58 y 0.62, mientras que la métrica F1-score de la clase SELL consiguió resultados entre 0.24 y 0.29[36]. Este concepto se usó en el presente trabajo “Comparación de Métodos de Aprendizaje de Máquina Aplicados a la Predicción de Movimientos a Corto Plazo de Acciones del DJIA Durante Periodos de Reportes Financieros Trimestrales” para tomar diferentes periodos de tiempo para las comparaciones con diferentes periodos de variación porcentual.

5.4.4 A Deep Learning Model for Predicting BUY and SELL Recommendations in Stock Exchange of Thailand using Long Short-Term Memory

Este trabajo usa técnicas de aprendizaje profunda para predecir recomendaciones de compra y venta de acciones de la bolsa de Tailandia con el uso de memoria de largo-corto plazo. El modelo captura dependencias de largo plazo de forma que pueda mejorar la certeza del modelo. La certeza del modelo propuesto es puesta a prueba en 5 acciones de la bolsa de Tailandia y sus resultados son comparados con otros métodos de aprendizaje de máquina como SVM, “multilayer perceptron”, árbol de decisión, Random Forest, regresión logística y k vecinos más cercanos. Este trabajo usó sets de datos con 730 elementos, y se usaron dos clases (BUY, SELL), 5 sets de datos correspondientes a diferentes empresas en tres periodos de tiempo (10, 30, 60, y 90 días). La mejor combinación de estas variaciones generó una certeza de 1, lo cual significa que acertó todas las predicciones. Esto se produjo con la empresa Airports of Thailand PCL para casi todos los métodos de aprendizaje de máquina usados excepto MLP(multilayer perceptron), pero para el periodo de 90 días exclusivamente [37]. Este trabajo tiene como similitud con el presente trabajo “Comparación de Métodos de Aprendizaje de Máquina Aplicados a la Predicción de Movimientos a Corto Plazo de Acciones del DJIA Durante Periodos de Reportes Financieros Trimestrales” el uso de los datos de 5 compañías para la prueba de los modelos de aprendizaje de máquina, además del

uso con diferentes métodos de aprendizaje de máquina, y de diferentes periodos. Además, los métodos de aprendizaje de máquina mencionados en este trabajo son igualmente usados en el presente trabajo, a excepción del modelo de aprendizaje profundo y el método “multilayer perceptron”.

5.4.5 On Stock Market Movement Prediction Via Stacking Ensemble Learning Model

En este trabajo, se comparan diferentes técnicas de ML para predecir movimientos bursátiles en el mercado de valores de Kenia. Las acciones se clasifican según su precio de cierre y el precio de apertura, dependiendo de esto la decisión de compra o venta. Se utilizan cuatro características: la diferencia entre el precio mínimo y el precio máximo, la diferencia entre el precio de cierre y de apertura, la capitalización de mercado y el volumen negociado. Los resultados se comparan utilizando: Adaptive Boosting, k vecinos más cercanos, y Stacking ensemble classifier. En este trabajo fueron evaluados los resultados con la métrica AUC, la cual mide la capacidad de un clasificador de distinguir entre clases, al aplicar esta métrica los mejores resultados fueron de 0.8238 y de 0.8139 por parte del método KNN [38]. Este trabajo tiene como similitud con el presente trabajo “Comparación de Métodos de Aprendizaje de Máquina Aplicados a la Predicción de Movimientos a Corto Plazo de Acciones del DJIA Durante Periodos de Reportes Financieros Trimestrales” el uso parcial del concepto de diferencia de precio de apertura y precio de cierre, concepto que es usado de manera parcial en este trabajo, pues se toma como un porcentaje y, además, está condicionado a la hora de publicación de resultados financieros de una empresa. Esto se puede ver en la sección 7.2 de este trabajo.

5.4.6 The impact of Public Information on the Stock Market

En este trabajo, se estudia la relación entre el número de nuevos anuncios reportados por Dow Jones and Company y métricas que miden el grado de actividad del mercado, como el volumen de transacciones y los retornos del mercado. Se encuentra en este trabajo que la actividad en el mercado está relacionada con factores como “day of the week dummy variables”, noticias en primera página de grandes periódicos de Nueva York, y grandes anuncios macroeconómicos. Sin embargo, la relación encontrada no es muy fuerte. Los analistas de Dow Jones confirman, igualmente, la dificultad de relacionar el volumen y la volatilidad a métricas de la información observada[39]. Este trabajo tiene como similitud con el presente trabajo “Comparación de Métodos de Aprendizaje de Máquina Aplicados a la Predicción de Movimientos a Corto Plazo de Acciones del DJIA Durante Periodos de Reportes Financieros Trimestrales” el uso de variables públicas, en el caso de este trabajo usando la sorpresa de beneficios por acción cómo es definido en el capítulo 7.

5.4.7 Fear and Greed: a Returns-Based Trading Strategy around Earnings Announcements

Este estudio muestra la utilidad de los reportes financieros de las empresas como una “muestra de realidad” en el corto plazo. Se muestra cómo acciones con ganancias anormales

semanas antes de los reportes financieros de las empresas tienden a sufrir fuertes reversiones en los precios de las acciones. Se muestra una estrategia de trading que logra ser rentable en 40 de los 43 años antes del 2016, obteniendo beneficios 1,3% superiores a lo normal en un periodo de 2 días posterior a la publicación de resultados financieros[7]. Este estudio muestra la posible ventana de oportunidad en torno a los periodos de reportes financieros trimestrales y confirma hasta cierto punto los movimientos anormales de los precios de acciones alrededor de estos periodos, los cuales pueden traducirse en ganancias o pérdidas extraordinarias por parte de inversionistas.

5.4.8 Volatility Skew, Earnings Announcements, and the Predictability of Crashes

En este trabajo se investiga la relación entre la volatilidad implícita de una empresa y la probabilidad de un posible “crash” posterior a los resultados financieros de dicha empresa. Este trabajo encuentra que la posibilidad de una correcta predicción incrementa con la inclusión de un mayor número de información de volatilidad histórica, opacidad de los informes financieros, y la sorpresa de ganancias del periodo actual[5]. Este trabajo fue de importancia para entender la relación que puede haber de la ventana de tiempo referente a resultados financieros y la posibilidad de una caída en bolsa. Además, se introduce el concepto de “sorpresa de ganancias”, concepto que es usado como variable para el entrenamiento de modelos de aprendizaje de máquina en el presente trabajo “Comparación de Métodos de Aprendizaje de Máquina Aplicados a la Predicción de Movimientos a Corto Plazo de Acciones del DJIA Durante Periodos de Reportes Financieros Trimestrales”.

5.4.9 Rationality of analysts’ earnings forecasts: evidence from dow 30 companies

En este trabajo, se prueba la “racionalidad” de las predicciones de analistas de las ganancias publicadas en informes financieros trimestrales de compañías del DJIA. Se usa un enfoque estadístico mejorado que da cuenta de la no estacionalidad en los datos de series de tiempo, la no normalidad en la regresión de integración y la correlación serial de los errores de pronóstico. Se recopilan datos (en I/B/E/S Inc.) de estimaciones de analistas de beneficios por acción y de valores reales desde 1984-Q4 hasta 2000-Q1. Se encuentra, además, que en 2/3 de los casos, las predicciones de ganancias de analistas son consistentes con los resultados reales [40]. Este trabajo se tomó en cuenta para el trabajo “Comparación de Métodos de Aprendizaje de Máquina Aplicados a la Predicción de Movimientos a Corto Plazo de Acciones del DJIA Durante Periodos de Reportes Financieros Trimestrales” en diferentes aspectos. Principalmente, se pudo observar la relación que se puede generar entre expectativas de inversores y resultados reales que podrían causar movimientos en el mercado, además, se usó como referencia para la recopilación de datos usando la fuente usada en este trabajo, aunque sin éxito debido a que I/B/E/S Inc. es una empresa que suele vender datos por volumen, por lo cual no fue posible que respondieran la solicitud de datos que se realizó.

6 RECOPIACIÓN DE DATOS Y CONSTRUCCIÓN DE SETS DE DATOS

Este trabajo tiene como objetivo aplicar y comparar métodos de aprendizaje de máquina para la predicción de decisiones de comprar/vender/mantener de acciones pertenecientes al índice industrial Dow Jones (DJIA), el cual consiste en las 30 empresas más representativas que cotizan en la bolsa de estados unidos. En este capítulo se seleccionaron 5 empresas de este índice y se definieron 8 sets de datos para realizar los experimentos de aprendizaje de máquina. La variable que se usó en los métodos de aprendizaje de máquina se denomina sorpresa de beneficios por acción, y la variable a predecir fue, en el caso de algoritmos de regresión, la variación porcentual esperada efecto de la sorpresa de beneficios por acción, y, en los algoritmos de clasificación, se identificaron con las etiquetas BUY/HOLD/SELL.

Por otra parte, la variable de sorpresa de beneficios por acción corresponde a la diferencia porcentual entre los beneficios por acción esperados por analistas y los beneficios por acción presentados en reportes financieros trimestrales por parte de las empresas que cotizan en bolsa. Estos informes ocurren 4 veces al año, de esta manera, la variable a predecir (variación porcentual), corresponde al incremento/decremento porcentual del precio de la acción de una empresa desde el día oficial del reporte financiero trimestral a un día posterior (desde 1 hasta 30 días después) para el cual corresponderá esta predicción. Por lo anterior, esta variable puede tener un valor diferente dependiendo del número de días contados después de la publicación de resultados financieros trimestrales, por lo que en este trabajo se realizaron pruebas de los métodos de aprendizaje de máquina, variando el número de días desde 1 hasta 30 días para así encontrar los modelos que obtengan mejores resultados[41].

6.1 ELECCIÓN DE EMPRESAS PARA LA IMPLEMENTACIÓN DE APRENDIZAJE DE MÁQUINA

Las empresas pertenecientes al índice DJIA son empresas grandes, con ingresos estables y consistentes, que además son consideradas representativas de un sector de estados unidos[41]. Por su misma definición e importancia, son empresas con gran relevancia en los mercados nacionales e internacionales, de las cuales además se puede encontrar información con mayor facilidad que otro tipo de empresas. Este es el razonamiento por el cual, para este trabajo, se optó por elegir empresas de este tipo para la correspondiente aplicación de aprendizaje de máquina, pues la cantidad, y facilidad de obtención de los datos es de gran importancia.

Inicialmente, este conjunto de empresas corresponde a 30 compañías de diversos sectores, las cuales representan el mercado estadounidense. Sin embargo, se decide reducir este número a un número igual o inferior a 10 debido a la dificultad de recopilar algunos datos, pues algunos de estos se recopilaron de manera manual. Para esto, se usó como criterio principal el tiempo que estas empresas hubiesen permanecido en el índice DJIA, seleccionando inicialmente a las empresas con mayor antigüedad en el DJIA. En un artículo

del 2011 de Forbes fueron mencionadas las 10 compañías más antiguas que hacían parte del DJIA[42]. Sin embargo, a la fecha actual de julio de 2022, 5 de estas compañías ya no hacen parte de este índice. Se muestra en la tabla 1 dichas empresas, estando en rojo las que actualmente no pertenecen al índice.

Empresas	Símbolo	Año de inclusión
General Electric	GE	1896
Exxon Mobil	XOM	1928
Procter & Gamble	PG	1932
DuPont	DD	1935
United Technologies	RTX	1934
Alcoa	AA	1959
3M	MMM	1976
IBM	IBM	1979
Merck	MRK	1979
American Express	AXP	1982

Tabla 1. 10 empresas más antiguas del DJIA a mayo del 2011

Por lo tanto, en la tabla 2 se listan las empresas elegidas para la posterior recopilación de datos.

Empresas	Símbolo	Año de inclusión
Procter & Gamble	PG	1932
3M	MMM	1976
IBM	IBM	1979
Merck	MRK	1979
American Express	AXP	1982

Tabla 2. 5 empresas más antiguas del DJIA a julio del 2022

6.2 DESCRIPCIÓN DE DATOS REQUERIDOS PARA LA IMPLEMENTACIÓN

A continuación, se muestran los datos recopilados para la implementación de aprendizaje de máquina planteada en este trabajo:

1. **Precios y cambios diarios de las acciones:** se refiere al cambio diario en el precio de las acciones de una determinada compañía. Es especialmente relevante recopilar los datos del precio de apertura y salida de cada día de una determinada acción. Con estos datos se calculó el porcentaje de variación del precio en un periodo. En este caso, dado que los reportes financieros suelen ser antes de la apertura del mercado, para calcular la variable objetivo se tomó como porcentaje de variación el cambio porcentual entre el precio de apertura del día de un reporte financiero, hasta el precio de cierre del día a predecir el modelo.
2. **Datos históricos de beneficios por acción de cada empresa:** este valor corresponde al número en dólares de beneficio neto que corresponde por cada acción de una determinada compañía. A modo de ejemplo, si una compañía X tiene beneficios netos de \$1,000,000 de dólares en el trimestre 1, y tiene además 1,000,000 de acciones emitidas que corresponden al 100% de la empresa, eso significa que el beneficio por acción es de \$1 dólar. Este valor es dado por periodos determinados. Para efectos de este trabajo se usará este valor en periodos trimestrales[43].
3. **Datos históricos de estimaciones de analistas de ganancias por acción de cada empresa (previo a las publicaciones oficiales):** cada trimestre, diversas compañías relacionadas con las recomendaciones de inversión, como Zacks, publican artículos y predicciones en cuanto a las expectativas financieras de las compañías más importantes del mercado bursátil. Una de estas predicciones suele ser la de las ganancias por acción esperadas que tendrán dichas compañías. Este dato es de suma importancia al invertir, ya que puede servir de punto de referencia respecto al crecimiento o no de una determinada empresa en caso de esta presente resultados significativamente mejores o peores de lo que se esperaba en un inicio desde este tipo de predicciones de especialistas.

6.3 RECOPIACIÓN DE DATOS

Los datos más sencillos de recopilar fueron los correspondientes al precio y cambios diarios de las acciones. Estos fueron encontrados en Yahoo Finance, dónde se pueden conseguir datos inclusive desde 1962 en el caso de IBM, o desde 1972 en el caso de American Express. Así, se puede observar que el año de inclusión de una compañía al DJIA (ver tabla 1 y 2) no tiene que ser el mismo que el año de inclusión al mercado de acciones, puesto que una compañía tiene que ser juzgada de forma cualitativa y cuantitativa por un panel selecto de personas de Wall Street y del S & P Global para ingresar al DJIA[44]. Además, en algunos

casos como el de Procter en Gamble, es muy complicado encontrar información histórica más allá de 1961, pues la fuente confiable más antigua que logra recopilar este tipo de información se encuentra en el libro “Daily Stock Price Record. New York Stock Exchange by Standard & Poor’s Corp”, el cual contiene este tipo de información desde 1961 hasta 2010[45]. Estos datos se descargan en formato csv, además, la información que está contenido se puede ver en la figura 5.

Date	Open	High	Low	Close*	Adj Close**	Volume
Jul 29, 2022	90.08	91.08	88.28	89.34	89.34	14,937,200
Jul 28, 2022	91.04	91.13	87.42	89.94	89.94	13,909,800

Figura 5. Ejemplo de formato de datos históricos del precio de acciones

En segundo lugar, la información de beneficios por acción de cada compañía y la información de estimación de beneficios por acción fue buscada en páginas especializadas en investigación de acciones y de inversión. Para la información de beneficios por acción reportados, se logró encontrar una mayor cantidad de datos de la fuente Ychart. Por otra parte, la información de estimaciones de beneficios por acción fue encontrada en Yahoo Finance, dónde es posible encontrar información desde el año 2000. En la tabla 3 se puede visualizar información de la fecha desde que se pudieron recopilar datos, no siendo estos necesariamente los datos usados debido a que se acotaron al periodo más cercano y se logró encontrar la variable sorpresa de beneficios por acción en Zacks Investment Research, la cual conjuga los datos de beneficios por acción y estimaciones de beneficios por acción en una sola, lo cual facilita la implementación e, inclusive, se obtiene una mayor cantidad de datos.

Empresas	Símbolo	Cambios diarios del precio de acciones	Beneficios por acción	Estimaciones de beneficios por acción
Procter & Gamble	PG	desde enero de 1962	desde trimestre 1 - 1984	desde trimestre 1-2000
3M	MMM	desde enero de 1970	desde trimestre 1 - 1984	desde trimestre 1-2000
IBM	IBM	desde enero de 1962	desde trimestre 1 - 1984	desde trimestre 1-2000
Merck	MRK	desde enero de 1962	desde trimestre 1 - 1984	desde trimestre 1-2000
American Express	AXP	desde enero de 1972	desde trimestre 1 - 1984	desde trimestre 1-2000

Tabla 3. Comparación fecha mínima de recopilación para cada tipo de dato

Por otra parte, fue posible encontrar la información conjugada de beneficios por acción y de estimaciones de beneficios por acción en un valor llamado sorpresa de beneficios por acción en la página de Zacks Investment Research, la cual es la firma líder en investigación de acciones, análisis y recomendaciones. Este valor conjugado de estimaciones y el valor real presenta inclusive un mayor número de datos que la variable de estimaciones de beneficios por acción, puesto que la página de Zacks, que se especializa en brindar este tipo de datos, solamente publica esta variable ya conjugada. La sorpresa de beneficios por acción corresponde a la diferencia porcentual entre las estimaciones de beneficios por acción y los beneficios por acción oficiales reportados[46]. Esta variable es de suma importancia para el análisis financiero por parte de inversionistas, debido a que la variable de beneficios por acción solo da una visualización acerca de las ganancias absolutas que corresponden a cada inversionista. Sin embargo, para poder visualizar una empresa de alto crecimiento o empresas que se están “estancando”, es necesario poder compararse con algún valor conservador que de una idea del crecimiento “normal” de una empresa. Este valor son las estimaciones de beneficios por acción realizadas por analistas de Zacks, Wall Street, y otras grandes firmas financieras.

Para visualizar de mejor manera este dato, pensaremos en él cómo la representación de los datos de beneficios por acción, y de estimaciones de beneficios por acción a través de la fórmula 6.

$$\frac{\text{Beneficios por acción}_{\text{publicación oficial}} - \text{Beneficios por acción}_{\text{estimación de Zacks}}}{\text{Beneficios por acción}_{\text{publicación oficial}}} * 100\%$$

Fórmula 6. Fórmula de sorpresa de beneficios por acción

Esta nueva variable conjugada, la cual será usada para las implementaciones de modelos de aprendizaje de máquina, fue encontrada en Zacks con datos desde el primer trimestre de 1992 hasta el último trimestre de 2021, al día que se recopiló esta información (marzo de 2022). Además, al ser información reportada trimestralmente, tienen 4 datos por año, lo que significa que solo hay 120 registros de información por cada compañía de la cual se recopiló información.

6.4 DEFINICIÓN DE SETS DE DATOS

Los datos de las empresas mencionadas se usaron para entrenar diferentes modelos de aprendizaje de máquina. Para esto, se experimentó con 8 sets de datos diferentes, lo cuales corresponden a 5 sets de datos correspondientes respectivamente a los datos de cada empresa, 2 sets de datos que combinan datos de 2 y 3 empresas respectivamente (buscando una agrupación por sectores), y 1 set de datos que corresponde a combinar los 5 sets de datos de las 5 empresas.

Primero, se experimentó con un set de datos el cual tiene los datos de las 5 empresas combinadas, por lo que se tienen 600 datos en este caso. Este posible modelo se prevé que podría funcionar debido a que los datos son porcentajes de un valor relativo a cada empresa (beneficios por acción), por lo que se podría encontrar una relación más directa. Adicionalmente, se buscaron características en común de las empresas como el sector en el que operan y el tipo de consumidor al que se dirigen y se experimentó con estos 2 sets de datos formados para entrenar los modelos de aprendizaje de máquina. Igualmente, en la sección 7.1, se puede observar un análisis por diagrama de cajas de la distribución de la variable de sorpresa de beneficios por acción, dónde se pueden observar similitudes entre los sets de datos que se usaron para formar estos 2 sets de datos por similitud de las empresas. En la tabla 4 se muestra información relevante de las empresas escogidas.

Empresa	Símbolo	Categoría	Capitalización (usd)	Peso en el DJIA	Dividendo	PER
International Business Machines Corporation	IBM	Tecnologías de la información	\$121,04 billones	2,96%	4,93%	21,79
American Express	AXP	Financiera	\$124.34 billones	2,96%	1,25%	17,01
3M	MMM	Industrial	\$86,72 billones	2,73%	3,91%	21,29
Merck Sharp and Dophne	MRK	Cuidado a la salud	\$230,58 billones	1,95%	3,03%	13,93
Procter & Gamble	PG	Productos básicos de consumo	\$350,48 billones	3,06%	2,49%	25,26

Tabla 4. Información general de las empresas seleccionadas

En la tabla 4 se muestra información general de cada empresa, como la categoría o sector al que pertenece, la capitalización (cuánto vale en el mercado actualmente), el peso en el DJIA (porcentaje en el que afecta el DJIA), dividendo (porcentaje de dinero respecto a la capitalización que da una empresa anualmente a sus inversionistas), y el índice PER (razón entre las ganancias netas absolutas de una empresa y su capitalización). Además, en la tabla 4 se ve una división por colores, la cual divide entre las empresas que se agruparan para formar así 2 sets de datos adicionales para entrenar modelos de aprendizaje de máquina.

La división se hizo dado la naturaleza del sector de cada empresa y su negocio, además de que en términos absolutos ambas tienen una capitalización muy similar. En el caso de IBM y American Express se definieron ambas como compañías que desarrollan productos tecnológicos, o financieros, con el uso de la tecnología (American Express), de alto valor agregado, pero que no son productos básicos ni esenciales. En el caso de IBM, esta compañía se dedica a la investigación, desarrollo, y comercialización de diferentes productos de hardware y de software[47]. Por otra parte, American Express fue concebida en esencia como una Fintech, que significa que es una empresa que pretende dar servicios financieros y para ello aplica nuevas tecnologías. Por otra parte, en el caso de 3M, Merck, y P&G, éstas se definieron como consecuencia cómo compañías de productos esenciales y básicos (aunque tenga algunas excepciones debido a que tienen cierto grado de diversificación). P&G tiene una cartera amplia de productos de cuidado personal, Merck productos del cuidado a la salud, y 3M de seguridad industrial.

Recopilando lo dicho en esta sección, tendremos 8 sets de datos para el entrenamiento de aprendizaje de máquina. Dichos sets de datos son los mencionados a continuación:

1. IBM
2. AMEX
3. 3M
4. MERCK
5. P&G
6. IBM – AMEX
7. 3M – MERCK – P&G
8. IBM – AMEX – 3M – MERCK – P&G (**TODOS**)

7 ADECUACIÓN DE DATOS, Y DEFINICIÓN DE PARÁMETROS Y MÉTRICAS

Se tienen 8 sets de datos, los cuales fueron usados para entrenar modelos de aprendizaje de máquina. Además, estos sets de datos consisten en 2 variables: sorpresa de beneficios por acción, y variación porcentual diaria del precio de las acciones. La sorpresa de beneficios por acción se usará en el set de datos, como es definido en Zacks. Sin embargo, la variación porcentual diaria se definió teniendo en cuenta la fecha en que se reportó oficialmente cada reporte financiero trimestral y el tiempo (BFO o AMC) del día en el cual se presentó este reporte. A continuación, se puede ver una descripción de la información que se recopila inicialmente de Zacks y de yahoo finance.

- Los datos de sorpresa de ganancia por acción de cada una de las 5 empresas elegidas. Estos contienen la fecha exacta en que se publicaron estos datos al mercado. Además, estos datos son desde el primer trimestre(Q1) de 1992, hasta el último trimestre(Q4) del 2021. La fuente de los datos es Zacks Investment Research.
- Precio de apertura, precio más alto y precio más bajo intradía, y precio de cierre de las acciones de las 5 compañías elegidas de todos los días en los que el mercado estuvo abierto (sin contar fines de semana, festivos y otros días especiales) desde el primero de enero de 1962. La fuente de los datos es Yahoo Finance.

Por la diferencia de fechas mínimas de obtención de cada variable (1992 para sorpresa de beneficios por acción y 1962 para variación porcentual diaria), solo se toman en cuenta los datos a partir del trimestre 1(Q1) de 1992. Además, tendrá que encontrarse un valor único de la variable variación porcentual diaria que se obtuvo, el cual esté relacionado(temporalmente) a cada dato de sorpresa de beneficios por acción. En este sentido, se busca que, por ejemplo, la variable de sorpresa de beneficios por acción del Q1 de 1992 del set de datos de Merck, que es de 1.24%, esté asociada a un solo valor de porcentaje de variación porcentual, el cual podrá estar para un número desde 1 día hasta 30 días después de la fecha exacta de publicación de los beneficios por acción del Q1 de 1992, por ejemplo, la variación porcentual a los 10 días de la publicación oficial de los beneficios por acción de este periodo es de -7.4%.

7.1 SORPRESA DE BENEFICIOS POR ACCIÓN

La primera variable es la sorpresa de beneficios por acción. Cómo se explica en la fórmula 6, este valor resulta de la diferencia porcentual entre el valor esperado, por parte de analistas de Zacks en este caso, y el valor de beneficios por acción de una compañía publicados en un trimestre. Este tipo de dato no es necesario preprocesarlo de ninguna manera en los sets de datos. A modo de ejemplificación, en otro tipo de trabajos es necesario filtrar aquellos datos

atípicos, sin embargo, este trabajo busca relacionar aquellos parámetros financieros con un tipo de reacción en el precio de las acciones subyacentes, por lo que en este caso al haber un resultado atípico (resultados financieros muy por encima o muy por debajo de lo esperado) tendría sentido relacionarlo con un movimiento positivo o negativo en el precio de la acción de la compañía relacionada.

En las figuras 6 a 13 se muestran gráficos de cajas e histogramas de los datos de beneficio por acción de cada uno de los sets de datos.

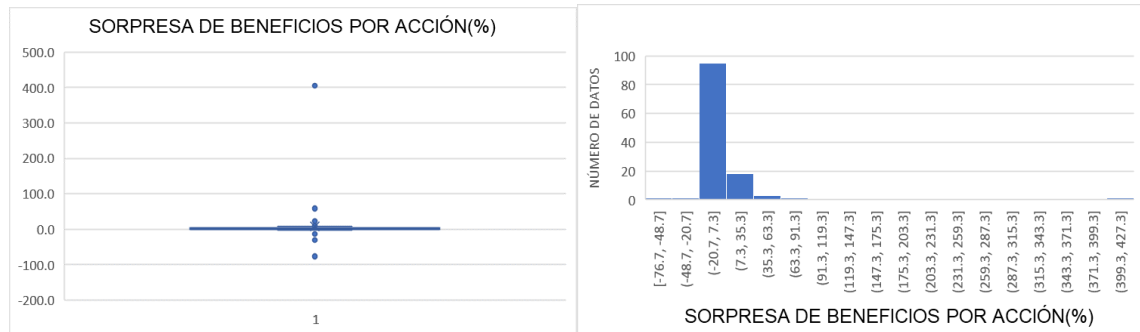


Figura 6. Gráficos de sorpresa de beneficios por acción de IBM

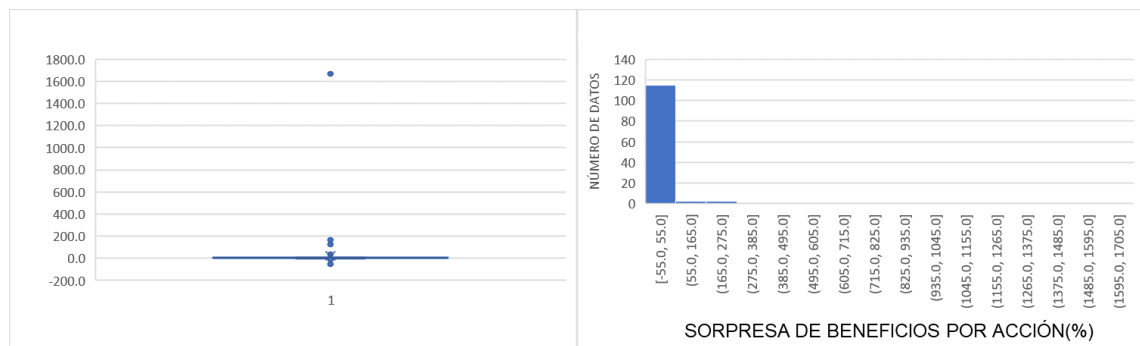


Figura 7. Gráficos de sorpresa de beneficios por acción de American Express

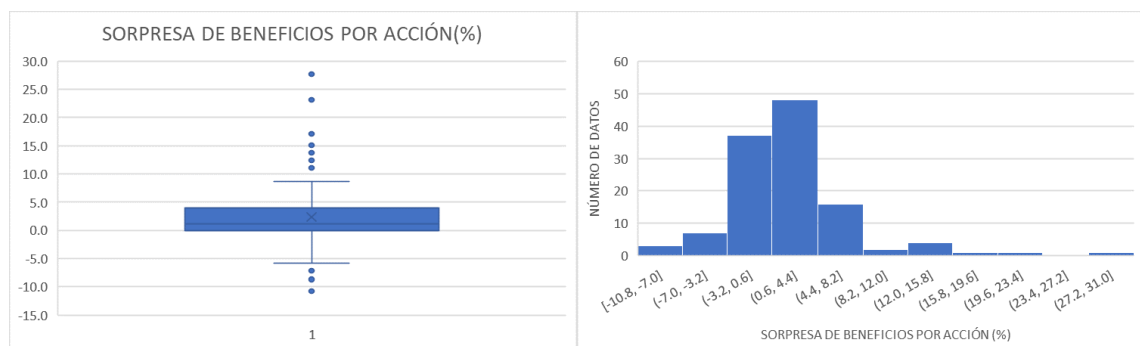


Figura 8. Gráficos de sorpresa de beneficios por acción de 3M

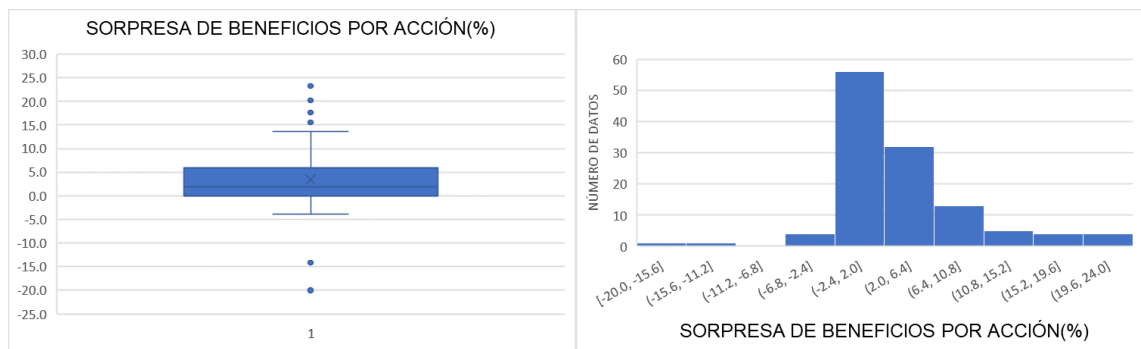


Figura 9. Gráficos de sorpresa de beneficios por acción de Merck

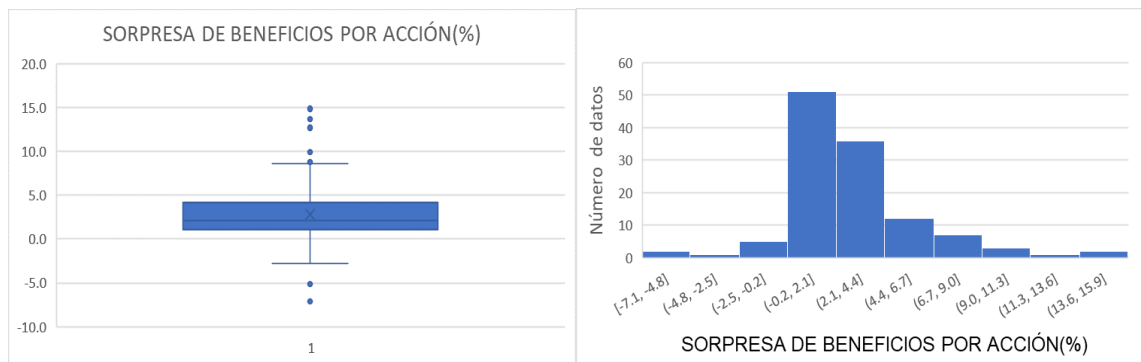


Figura 10. Gráficos de sorpresa de beneficios por acción de P&G

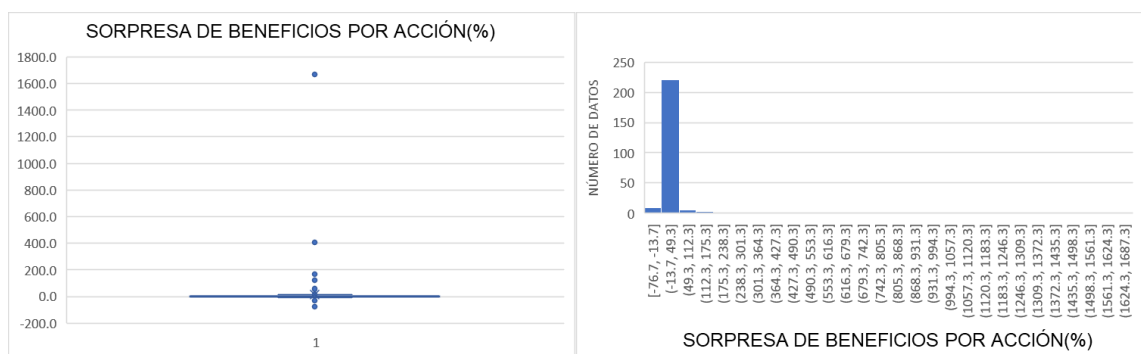


Figura 11. Gráficos de sorpresa de beneficios por acción del grupo IBM + American Express

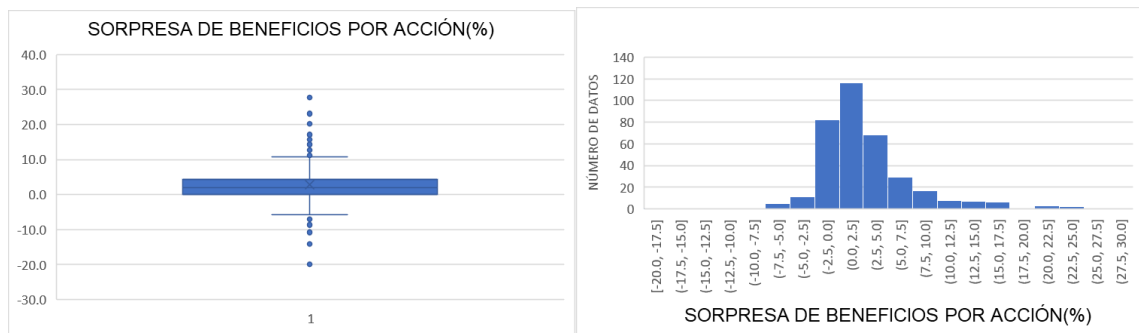


Figura 12. Gráficos de sorpresa de beneficios por acción del grupo Merck + P&G + 3M

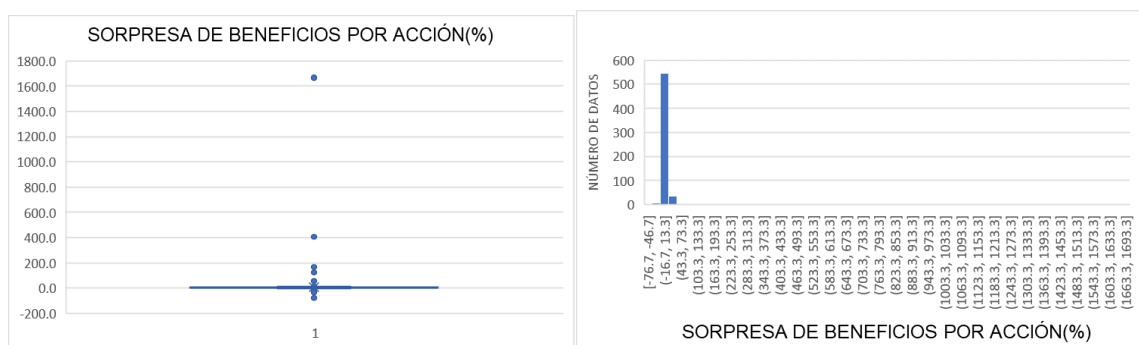


Figura 13. Gráficos de sorpresa de beneficios por acción del grupo de todas las acciones

Los métodos de visualización como el diagrama de cajas y el histograma mejoran nuestro entendimiento de la información, así como nos ayudan a hacer comparaciones entre la información de distintos conjuntos de datos, como en este caso[48]. El diagrama de cajas (gráfico a la izquierda) nos permite inclusive ver en detalle la distribución de cada conjunto de datos. Este diagrama es usado para el análisis de datos en el sector financiero. Esto se debe a que se puede evidenciar fácilmente información como la mediana, el rango intercuartílico (datos entre Q1 y Q3), y los datos atípicos[49]. Además de esto, se usan histogramas debido a su utilidad al ver la distribución de datos, aun cuando estos no son fácilmente reconocibles por haber valores atípicos extremos, como en el caso del conjunto de datos de IBM y American Express.

Se puede observar en los gráficos una distribución mayormente normal, habiendo pocas excepciones, como en el caso de Merck, que tiene un sesgo a la derecha, o en el caso de American Express, que tiene un sesgo a la izquierda. Adicionalmente, se nota que los conjuntos de acciones que incluye IBM y American Express, así como el que incluye todas las compañías, presenta un sesgo a la izquierda. Por otra parte, se puede notar a simple vista que los datos de sorpresa de beneficio por acción de ambas compañías que se incluyen en la

categoría de tecnológicas y Fintech tienen valores atípicos “extremos”, ya que inclusive se distorsiona el gráfico de cajas, esto pudiéndose explicar cómo el crecimiento explosivo prematuro que normalmente es intrínseco a este tipo de compañías tecnológicas y Fintech cuando alcanzan el punto de equilibrio y empiezan a generar beneficios[50]. Además, este tipo de valores atípicos causan que la escala del histograma se haga tan amplia, que pareciese que muchos valores se concentran en un solo rango, sin embargo, estos rangos donde pareciese se concentran la mayoría de los datos, son realmente amplios en comparación a otros sets de datos. Por ejemplo, en el caso de IBM este rango está entre -20.7% y 7.3%, sin embargo, un rango donde se encuentra la mayoría de los datos en Merck se encuentra entre -2.4% y 2.0%, por lo que los rangos comprenden un menor número de valores porcentuales.

7.2 CAMBIO DE PRECIO DE LAS ACCIONES

Para realizar el algoritmo, es necesario encontrar un valor único que podamos asociar con el cambio porcentual del precio de las acciones desde el momento de publicación de cada resultado financiero en el tiempo. Para esto, tenemos que tomar en cuenta que existen dos posibles tiempos de publicación de resultados financieros: BFO, y AMO. BFO significa que los resultados son publicados por parte de la compañía en la página de la comisión de bolsa y valores (ente regulador para las empresas que cotizan en el mercado estadounidense) antes de la apertura de la bolsa de Nueva York. De la misma manera, AMO significa que los resultados financieros son publicados después del cierre de la bolsa de Nueva York. Es de resaltar que la bolsa de Nueva York tiene horario de apertura a las 9:30 am hora ET, y horario de cierre a las 4:00 pm ET.

En la tabla 5 se pueden ver los horarios de publicación de resultados usuales de las empresas seleccionadas.

EMPRESA	HORA USUAL DE PUBLICACIÓN DE RESULTADOS TRIMESTRALES
IBM	AMO
AMEX	BFO
3M	BFO
MERCK	BFO
P&G	BFO

Tabla 5. Horario de publicación de resultados financieros de cada empresa seleccionada

Cada empresa que cotiza en la bolsa estadounidense está obligada a tener una página dedicada a inversionistas llamada “investor relations”. Dicha página es abierta al público general y se puede encontrar información importante como pasados reportes financieros e

inclusive la hora en que cada uno fue publicado. De dicha página se extrajo esta información de la hora usual de publicación de resultados después de revisar los últimos 4 reportes financieros de cada empresa.

De esta manera, la fórmula 7 se usó para calcular el porcentaje de variación porcentual del precio de una acción después de la publicación de resultados financieros. Esta fórmula se usó para el caso de empresas que publican resultados antes de la apertura de la bolsa (BFO). Además, la fórmula 8 se usó para empresas que publican resultados después del cierre de la bolsa (AMC). El resultado de variación porcentual (fórmula 7 o 8 dependiendo de la hora de normal de publicación de resultados financieros de cada compañía) es la variable que se usó para entrenar los modelos de aprendizaje de máquina junto a la sorpresa de beneficios por acción.

$$\frac{\text{Precio de apertura}_{\text{día de publicación}} - \text{Precio de cierre}_{\text{día a predecir movimiento}}}{\text{Precio de apertura}_{\text{día de publicación}}} * 100\%$$

Fórmula 7. Fórmula de variación porcentual del precio de las acciones para reportes
BFO

$$\frac{\text{Precio de cierre}_{\text{día de publicación}} - \text{Precio de cierre}_{\text{día a predecir movimiento}}}{\text{Precio de cierre}_{\text{día de publicación}}} * 100\%$$

Fórmula 8. Fórmula de variación porcentual del precio de las acciones para reportes
AMC

El uso de una fórmula que considera el precio de cierre del día de publicación de resultados se debe a que no se pueden operar acciones después del cierre de la bolsa de Nueva York, por lo que, al poderse ver los resultados financieros después de esto, esta información pesará en el mercado en la siguiente apertura de la bolsa.

De esta manera, se recopilan estos nuevos datos transformados junto a los datos de sorpresa de beneficios por acción para su uso en los algoritmos de aprendizaje de máquina, tanto de regresión, sección 7.4, cómo de clasificación, capítulo 8. Además, se encuentra el coeficiente de correlación de los datos en cada set de datos. Esto, debido a que la correlación de dos variables, es muy útil al invertir en los mercados financieros y suele ser usada para diferentes aplicaciones como al evaluar el rendimiento de un fondo de inversión respecto a un índice del mercado bursátil[51].

Para elegir el periodo de días después de la publicación de resultados, para el cual se tiene una mejor predicción, se generaron modelos de aprendizaje de máquina para los 30 días siguientes. Se usaron las fórmulas definidas anteriormente para predicción de movimientos

desde 1 día hasta 30 días, teniendo en cuenta que son días en los que el mercado estén abiertos y no días calendario. Para cada uno de estos periodos, se usó como variable objetivo (para métodos de regresión) el porcentaje de variación del precio de la acción después de cada publicación financiera de resultados, definida dependiendo del periodo, entre 1 y 30 días.

En la figura 14 se muestran las gráficas con los coeficientes de correlación de los datos de porcentaje de variación del precio de la acción respecto a los datos de sorpresa de beneficios por acción:

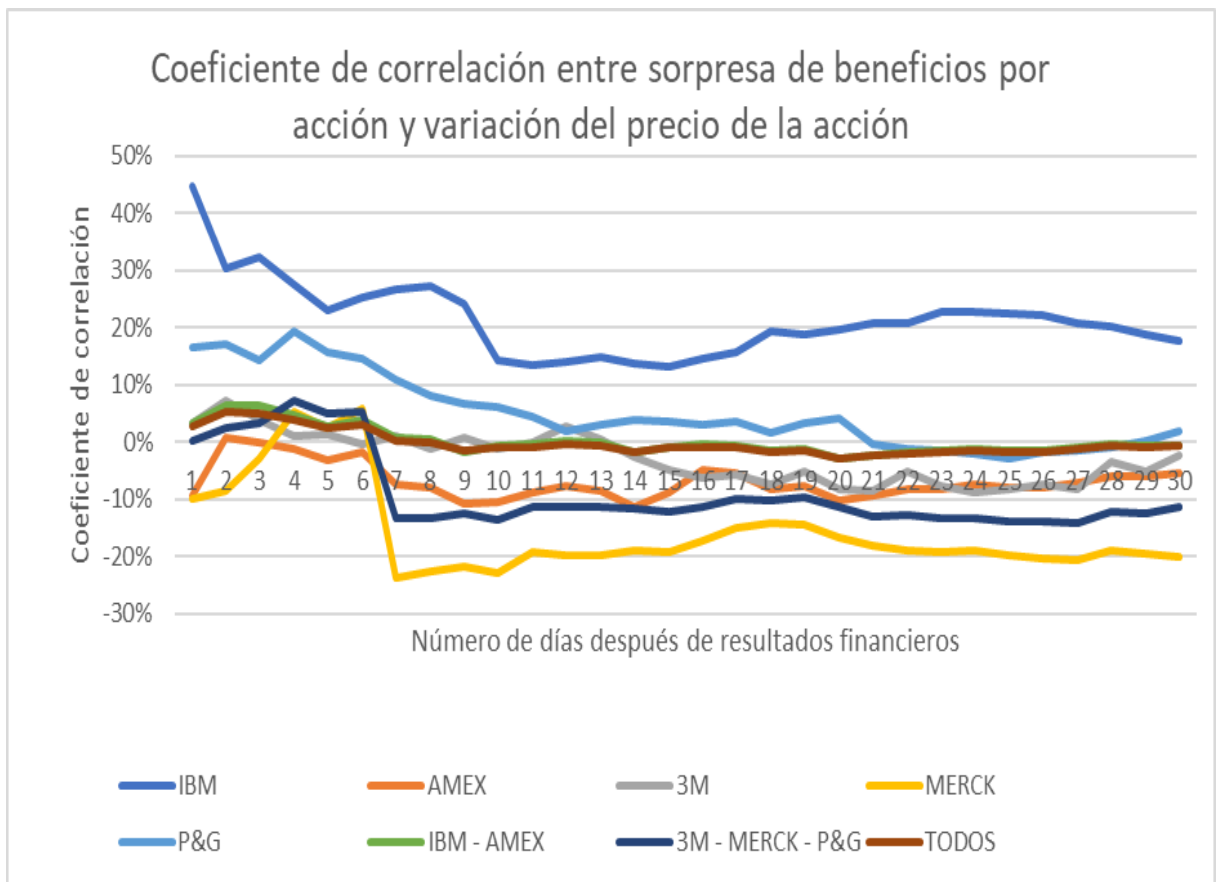


Figura 14. Coeficiente de correlación entre sorpresa de beneficios por acción y variación porcentual del precio de la acción para cada set de datos

Tal como se puede observar en la gráfica, existe una baja correlación entre estos parámetros en su mayoría. Sin embargo, este parámetro no es definitivo ni define si este problema tiene solución. Existen dos tipos de métodos de aprendizaje de máquina que suelen usarse en problemas financieros: regresión, y clasificación. En el caso de la correlación, este tipo de parámetro relaciona principalmente la parte lineal y directamente relacionada de dos variables, por esta razón podría dar indicios de que utilizar métodos de aprendizaje de

máquina de regresión no es la mejor opción. Por esta razón y por la tendencia a usar el enfoque BUY/HOLD/SELL o BUY/SELL al hacer predicciones para el mercado bursátil, se compararon diferentes métodos de clasificación en este trabajo. Algunos ejemplos de este enfoque son los trabajos “Algorithmic Financial Trading with Deep Convolutional Neural Networks: Time Series to Image Conversion Approach”, “A Deep Learning Model for Predicting BUY and SELL Recommendations in Stock Exchange of Thailand using Long Short-Term Memory”, y “On Stock Market Movement Prediction Via Stacking Ensemble Learning Model”[52] [37][38].

7.3 REDEFINICIÓN DE LOS NOMBRES DE LOS SETS DE DATOS

Para simplicidad de los gráficos, ejecuciones, y menciones posteriores; se definieron nombres más cortos para cada uno de los sets de datos usados en este trabajo. Estos nombres serán usados principalmente en las gráficas mostradas. Estas nuevas definiciones se pueden ver en la tabla 6.

Set de datos	Nombre a usar en el trabajo y gráficas
IBM	IBM
American Express	AMEX
3M	3M
Merck	MERCK
P & G	PG
IBM + AMEX	TECH
3M + MERCK + PG	NONTECH
IBM + AMEX + 3M + MERCK + PG	ALL

Tabla 6. Redefinición de los nombres de los sets de datos definidos

7.4 IMPLEMENTACIÓN CON ALGORITMOS DE REGRESIÓN

Se realizó una prueba preliminar sin optimización de hiperparámetros para evaluar la factibilidad de usar métodos de regresión para la solución de este problema. Se usa el método de regresión polinómica de grado 1(modelo lineal) hasta grado 4. De esta forma, se pudo observar que este tipo de modelos no es suficientemente satisfactorio para este tipo de implementación. Estos métodos fueron usados en el curso “Machine learning with Python: a practical introduction by IBM” y se implementaron con los sets de datos usados en este trabajo[53]. Para esto se usó las librerías pandas, sklearn.preprocessing.PolynomialFeatures, y sklearn.linear_model. En las figuras 15 a 22 se pueden ver los gráficos obtenidos con la métrica R2 square para cada set de datos. Igualmente, se puede observar en series de tiempo la evolución del dato de beneficios por acción en el tiempo.

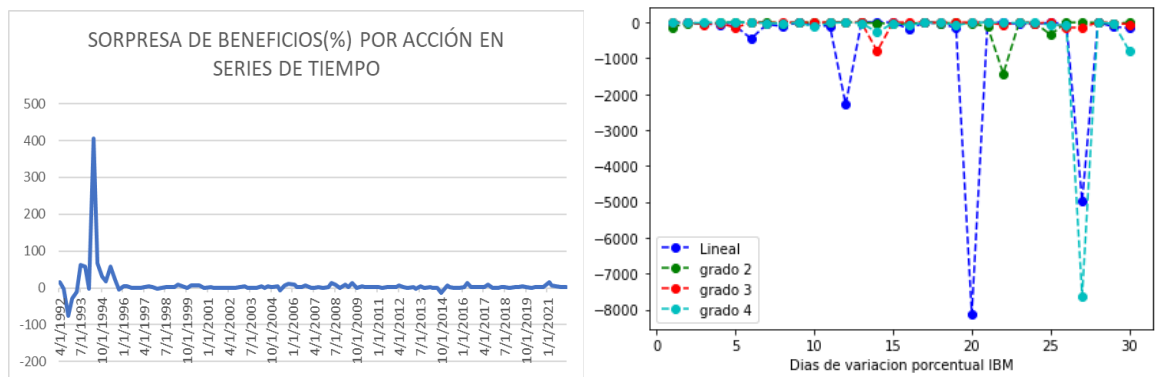


Figura 15. Sorpresa en series de tiempo (izquierda) y resultados de R2 square (derecha) para los diferentes periodos del set de IBM en métodos de regresión polinómica

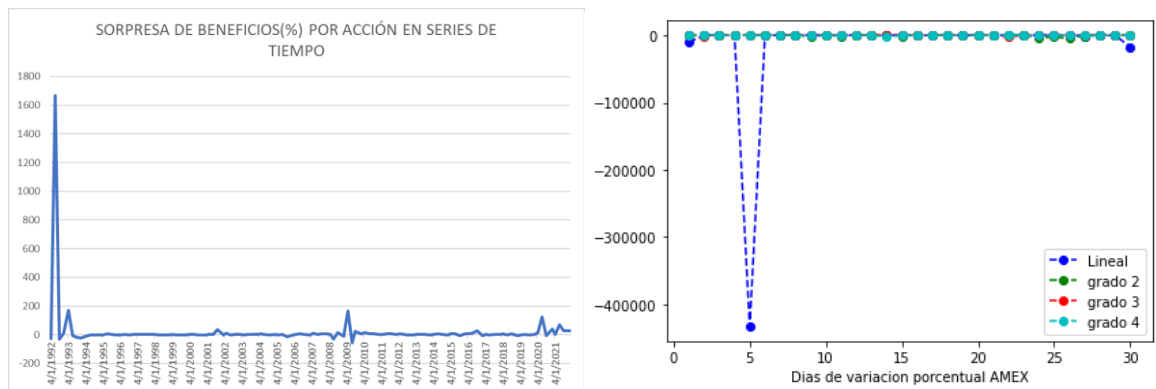


Figura 16. Sorpresa en series de tiempo (izquierda) y resultados de R2 square (derecha) para los diferentes periodos del set de AMEX en métodos de regresión polinómica

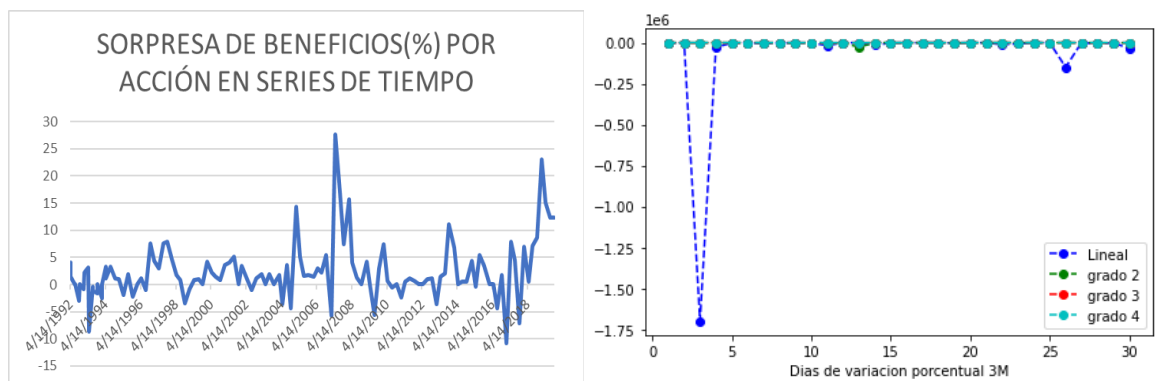


Figura 17. Sorpresa en series de tiempo (izquierda) y resultados de R2 square (derecha) para los diferentes periodos del set de 3M en métodos de regresión polinómica

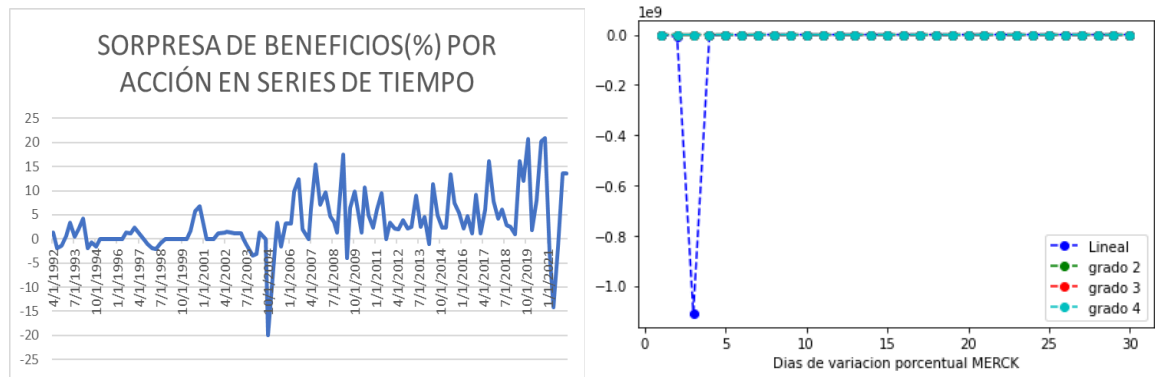


Figura 18. Sorpresa en series de tiempo (izquierda) y resultados de R2 square (derecha) para los diferentes periodos del set de MERCK en métodos de regresión polinómica

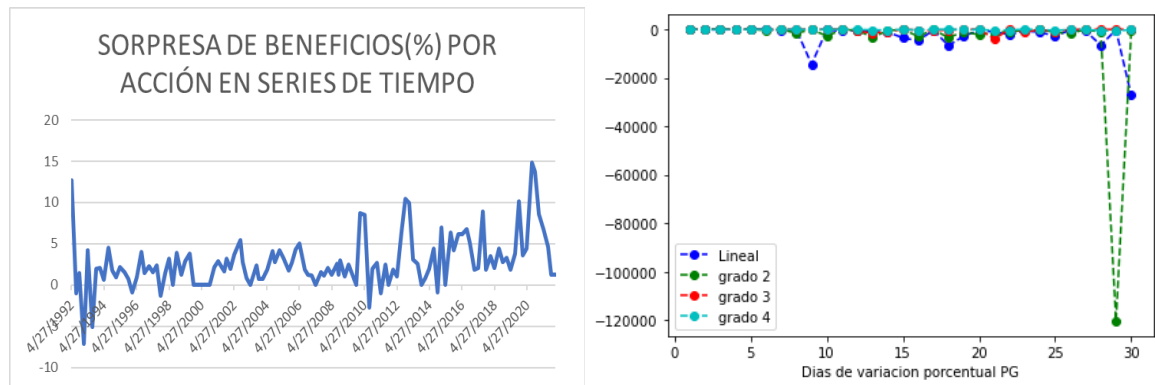


Figura 19. Sorpresa en series de tiempo (izquierda) y resultados de R2 square (derecha) para los diferentes periodos del set de PG en métodos de regresión polinómica

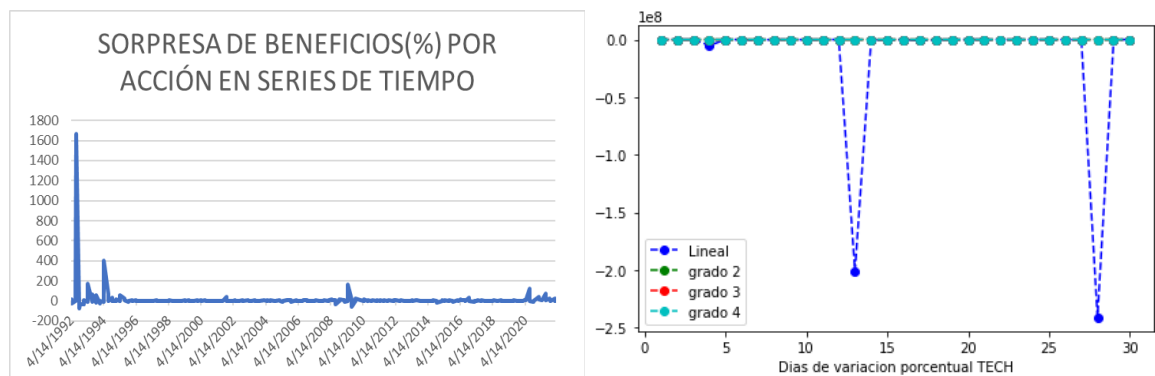


Figura 20. Sorpresa en series de tiempo (izquierda) y resultados de R2 square (derecha) para los diferentes periodos del set de TECH en métodos de regresión polinómica

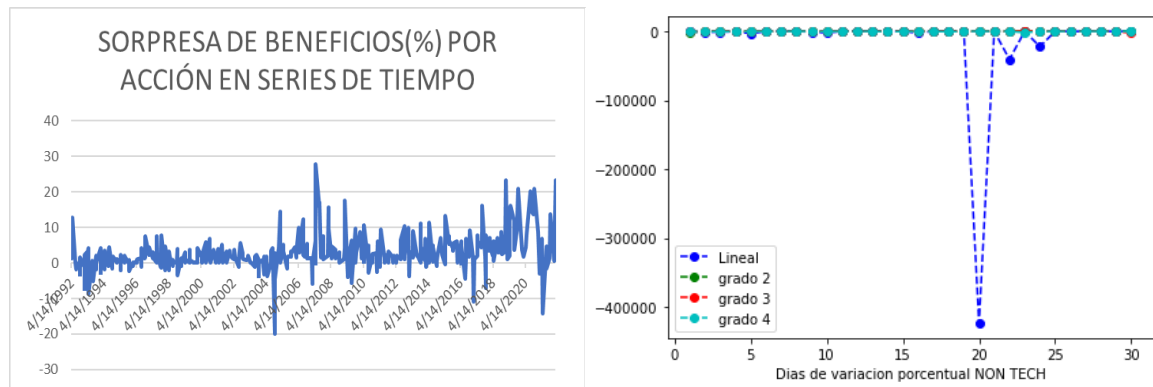


Figura 21. Sorpresa en series de tiempo (izquierda) y resultados de R2 square (derecha) para los diferentes periodos del set de NONTECH en métodos de regresión polinómica

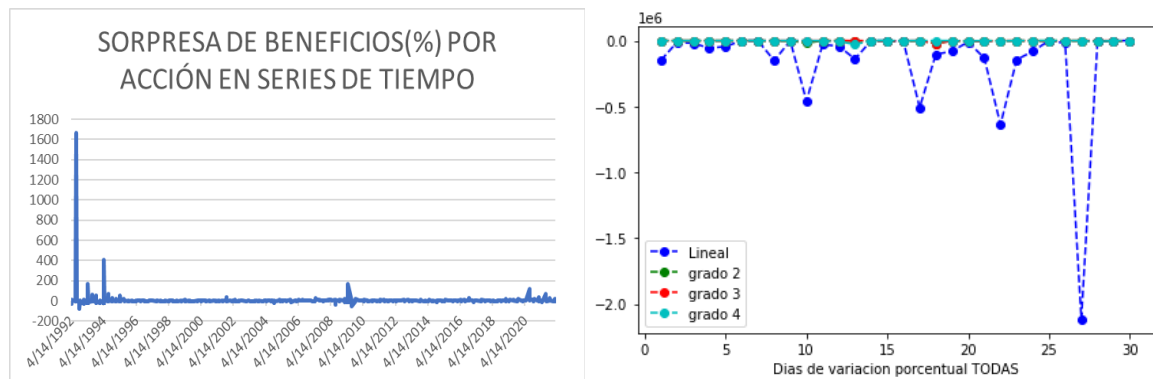


Figura 22. Sorpresa en series de tiempo (izquierda) y resultados de R2 square (derecha) para los diferentes periodos del set de ALL en métodos de regresión polinómica

La métrica R2 square es usada para medir la dispersión de los datos de la línea de regresión obtenida después de implementar el método de aprendizaje de máquina de regresión. Esta métrica es típicamente usada para medir el rendimiento de los modelos de aprendizaje de máquina de regresión y típicamente tiene un valor entre -1 y 1, dónde 1 significa una relación directa muy fuerte, y 0 una relación inversa muy fuerte[31]. Sin embargo, estos valores pueden ser superiores a 1 o inferiores a -1 cuando el modelo encontrado, después de aplicar el método de regresión, es muy malo para acoplarse a la dispersión de los datos. En las gráficas X y Y podemos observar cómo, en general, hay muchos casos en los que el modelo es muy inferior a -1, lo que implica que este tipo de modelos de regresión no es bueno para esta implementación. Además, en la figura 23 podemos observar los mayores valores de R2 square obtenidos en cada set de datos. Podemos observar que aún en el mejor de los casos el modelo solo alcanza un valor de r2 square de 0.5, siendo este un dato muy raro, debido a que el segundo mejor valor obtenido fue de 0.2. Además, al observar el promedio de este valor

en la figura 24, se puede observar que tiene un valor muy bajo, por lo que hace que este tipo de métodos probablemente no sea una buena opción para este tipo de implementaciones. Se planteó usar diferentes métodos de aprendizaje de máquina de clasificación para solucionar este problema.

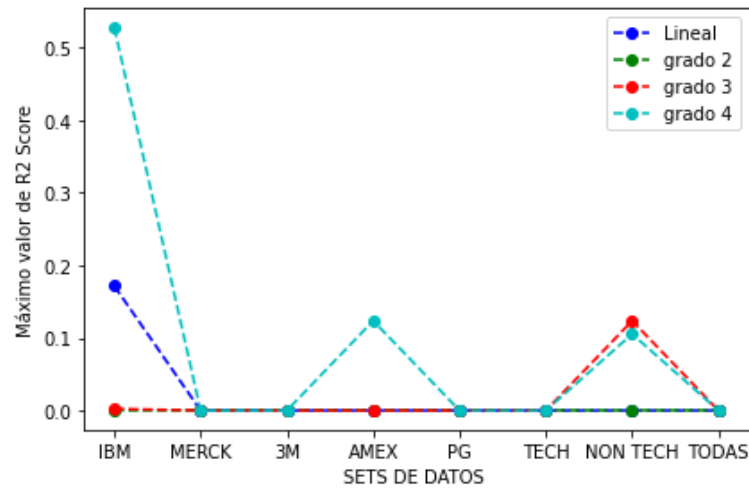


Figura 23. Mejores valores de R2 square para cada set de datos en diferentes métodos de regresión polinómica

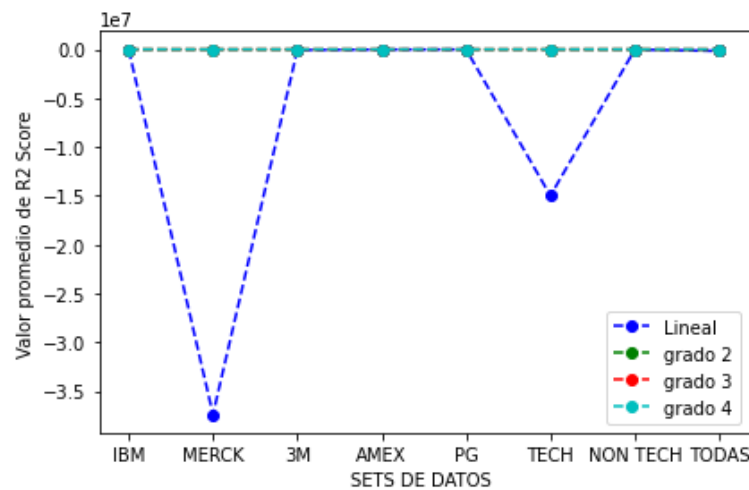


Figura 24. Valores promedio de R2 square para cada set de datos en diferentes métodos de regresión polinómica

7.5 ETIQUETAS PARA ALGORITMOS DE CLASIFICACIÓN

Cómo se muestra por los resultados de los modelos de regresión, y en consecuencia y para seguir el tipo de implementaciones que se hacen al realizar algoritmos para tomar decisiones en el mercado bursátil, se decide optar por métodos de clasificación para este problema. Existen dos aproximaciones en trabajos similares: BUY/SELL, y BUY/HOLD/SELL. En esencia, el método BUY/SELL es una extensión simplificada del que considera tres categorías (BUY/HOLD/SELL), por esto se usaron las tres etiquetas para buscar generalización del problema.

No existe una forma específica ni acordada para nombrar estas etiquetas. En los enfoques BUY/SELL, se suele asignar la categoría dependiendo del signo de variación porcentual predicha. Sin embargo, en el caso BUY/SELL, aunque haya ganado un porcentaje pequeño (+1%) o perdido un porcentaje pequeño (-1%), esto se contará como un BUY o SELL respectivamente, lo cual lógicamente no es coherente porque este tipo de volatilidad es “normal” en el mercado.

En la tabla 7, se muestra la desviación estándar de la variación porcentual histórica de 5 días y de 30 días de las compañías elegidas.

EMPRESA	DESVIACIÓN ESTÁNDAR DE 5 DÍAS	DESVIACIÓN ESTÁNDAR DE 30 DÍAS
IBM	3.48%	8.60%
AMEX	4.75%	11.85%
3M	3.20%	7.10%
MERCK	3.53%	7.94%
P&G	2.95%	7.94%

Tabla 7. Desviación estándar periodo de 5 y 30 días para cada una de las empresas seleccionadas

Pequeños cambios porcentuales no deberían ser de gran preocupación para el inversionista, por ejemplo, los valores más pequeños del caso BUY/SELL. Esto es visible al ver la tabla 7, pues, por ejemplo, en un periodo de 5 días puede ser normal una variación de entre un 2,93% y 4,75%. De hecho, se puede perder dinero al vender una acción con una ganancia muy pequeña. Esto se debe a los costos transaccionales de comprar y vender acciones, los cuales íntegramente cuestan entre 1% y 2% dependiendo del bróker que se use para realizar estas transacciones[54].

Por esta razón, se eligieron estas etiquetas basándose en identificar si cada movimiento corresponde o no a un movimiento especialmente alto dentro del histórico de cada acción para el periodo de tiempo que se predice. Se hizo uso para este caso del rango intercuartílico, el cual representa los datos “comunes” de un conjunto. Se definieron las etiquetas, para cada valor de tiempo específico (ej. 5 días o 10 días), de la siguiente manera:

- BUY: un valor en el extremo superior del rango intercuartílico (mayor a Q3)
- HOLD: un valor perteneciente al rango intercuartílico (entre Q1 y Q3)
- SELL: un valor en el extremo inferior del rango intercuartílico (menor a Q1)

En la tabla 8 se muestran los límites Q1 y Q3 de cada conjunto de datos para cada periodo (de 1 a 30 días). Además, es importante mencionar que, ya definidas las etiquetas para las 5 empresas. Los sets de datos que son compuestos por 2 o más empresas tienen los datos de los conjuntos individuales de las empresas que conforman el set de datos, por esa razón, se usaron las mismas etiquetas asignadas a los conjuntos individuales, con la diferencia de que al tener, ejemplo, las etiquetas de 2 empresas sets de datos individuales, tendrá este set de datos el doble de registros, sin embargo, se conservaron las mismas etiquetas y datos de beneficios por acción de los sets de datos individuales de las empresas que lo conforman.

NÚMERO DE DÍAS DESPUÉS DE RESULTADOS FINANCIEROS										
	IBM		AMEX		3M		MERCK		P&G	
	Q1	Q3	Q1	Q3	Q1	Q3	Q1	Q3	Q1	Q3
1	-0.81%	0.85%	-1.00%	1.07%	-0.69%	0.77%	-0.82%	0.90%	-0.67%	0.71%
2	-1.14%	1.27%	-1.46%	1.60%	-1.03%	1.17%	-1.19%	1.38%	-0.95%	1.10%
3	-1.39%	1.57%	-1.73%	2.02%	-1.24%	1.49%	-1.49%	1.75%	-1.15%	1.38%
4	-1.60%	1.85%	-2.02%	2.38%	-1.44%	1.71%	-1.70%	2.10%	-1.30%	1.63%
5	-1.76%	2.08%	-2.22%	2.63%	-1.59%	1.91%	-1.89%	2.39%	-1.44%	1.86%
6	-1.94%	2.31%	-2.40%	2.93%	-1.75%	2.12%	-2.06%	2.66%	-1.58%	2.03%
7	-2.09%	2.52%	-2.59%	3.18%	-1.90%	2.30%	-2.18%	2.90%	-1.68%	2.21%
8	-2.25%	2.74%	-2.65%	3.43%	-1.96%	2.47%	-2.35%	3.14%	-1.75%	2.43%
9	-2.35%	2.93%	-2.79%	3.69%	-2.04%	2.64%	-2.45%	3.43%	-1.80%	2.57%
10	-2.50%	3.12%	-2.91%	3.87%	-2.15%	2.80%	-2.57%	3.63%	-1.90%	2.73%
11	-2.64%	3.33%	-3.02%	4.07%	-2.22%	2.95%	-2.68%	3.81%	-1.95%	2.86%
12	-2.73%	3.55%	-3.13%	4.23%	-2.32%	3.15%	-2.81%	4.02%	-2.04%	3.01%
13	-2.86%	3.71%	-3.18%	4.40%	-2.40%	3.31%	-2.92%	4.20%	-2.08%	3.15%
14	-2.92%	3.86%	-3.27%	4.60%	-2.50%	3.44%	-2.97%	4.38%	-2.17%	3.32%
15	-3.04%	4.05%	-3.39%	4.75%	-2.57%	3.59%	-3.05%	4.50%	-2.18%	3.48%
16	-3.13%	4.19%	-3.48%	4.89%	-2.62%	3.70%	-3.11%	4.72%	-2.21%	3.63%
17	-3.21%	4.34%	-3.54%	5.14%	-2.66%	3.84%	-3.14%	4.88%	-2.28%	3.73%
18	-3.33%	4.50%	-3.63%	5.32%	-2.78%	4.00%	-3.26%	4.99%	-2.31%	3.91%
19	-3.35%	4.67%	-3.64%	5.52%	-2.86%	4.08%	-3.33%	5.17%	-2.36%	4.03%
20	-3.44%	4.85%	-3.70%	5.78%	-2.94%	4.22%	-3.34%	5.29%	-2.39%	4.19%
21	-3.56%	5.04%	-3.79%	5.91%	-3.00%	4.31%	-3.45%	5.43%	-2.45%	4.31%
22	-3.59%	5.20%	-3.88%	6.09%	-3.03%	4.49%	-3.47%	5.60%	-2.47%	4.42%
23	-3.70%	5.31%	-3.99%	6.31%	-3.13%	4.59%	-3.45%	5.75%	-2.51%	4.55%
24	-3.80%	5.46%	-4.07%	6.47%	-3.16%	4.66%	-3.51%	5.89%	-2.55%	4.67%
25	-3.88%	5.57%	-4.16%	6.61%	-3.22%	4.76%	-3.48%	6.05%	-2.56%	4.76%
26	-3.92%	5.76%	-4.27%	6.81%	-3.26%	4.90%	-3.53%	6.20%	-2.64%	4.85%
27	-3.95%	5.84%	-4.27%	6.97%	-3.23%	5.05%	-3.59%	6.38%	-2.65%	4.93%
28	-3.99%	5.98%	-4.29%	7.11%	-3.28%	5.14%	-3.63%	6.52%	-2.69%	5.03%
29	-4.06%	6.11%	-4.35%	7.26%	-3.31%	5.31%	-3.67%	6.62%	-2.73%	5.18%
30	-4.09%	6.28%	-4.35%	7.48%	-3.38%	5.43%	-3.75%	6.78%	-2.76%	5.29%

Tabla 8. Límites Q1 y Q3 usados para definir las etiquetas BUY/HOLD/SELL en cada set de datos

Con los valores de Q1 y Q3 definidos ya es posible asignar etiquetas para su uso en las implementaciones con aprendizaje de máquina del capítulo 8. Sin embargo, al ser el rango intercuartílico el que contiene la mayoría de los datos por definición, esto implica que la mayor parte de etiquetas fueron HOLD. Esto puede causar problemas, ya que un modelo con alta certeza podría significar solamente que predice la mayoría de las etiquetas como HOLD, lo cual no es deseable. En las figuras 25 a 28 se visualiza la cantidad de etiquetas BUY, HOLD, y SELL que hay por cada set de datos para corroborar la posibilidad que se plantea.

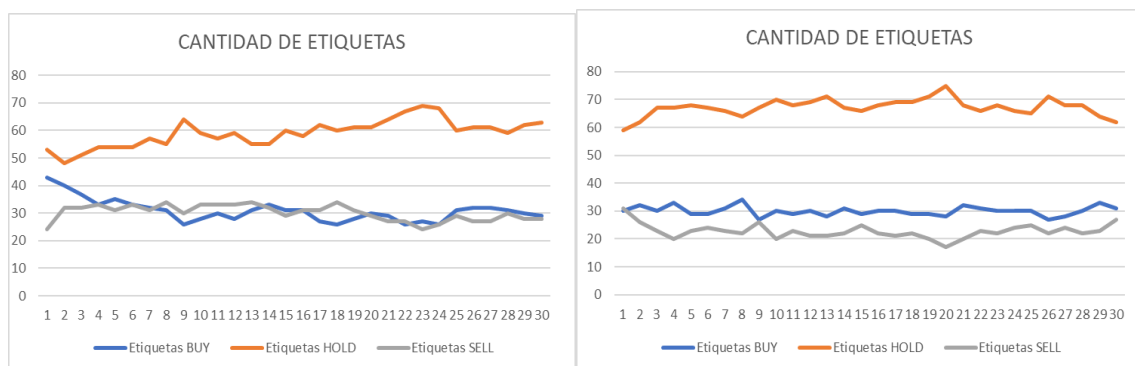


Figura 25. Cantidad de etiquetas BUY, HOLD, y SELL de IBM (izquierda) y AMEX (derecha)

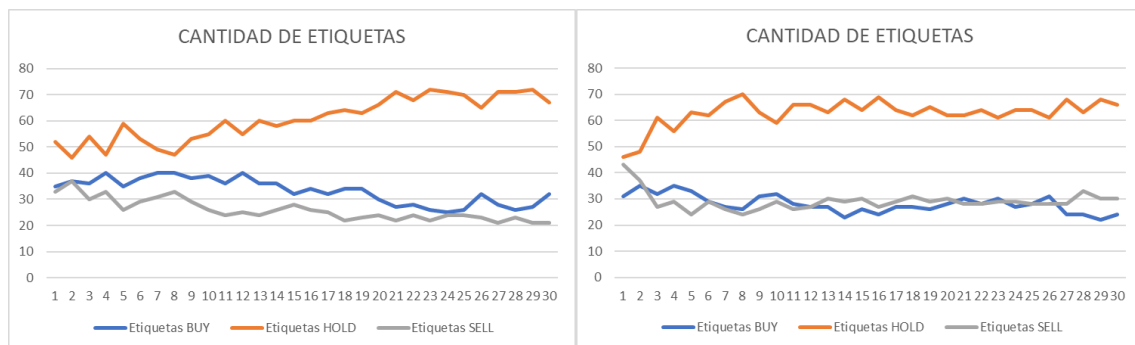


Figura 26. Cantidad de etiquetas BUY, HOLD, y SELL de 3M (izquierda) y MERCK (derecha)

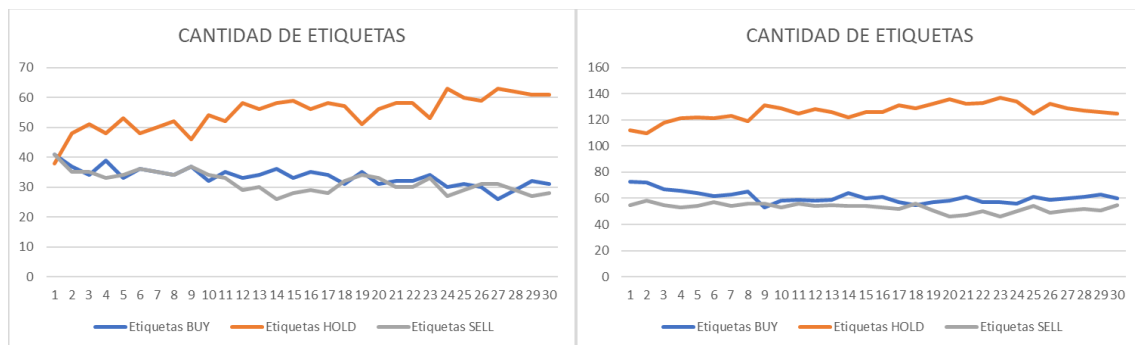


Figura 27. Cantidad de etiquetas BUY, HOLD, y SELL de P&G (izquierda) y del grupo de AMEX - IBM (derecha)

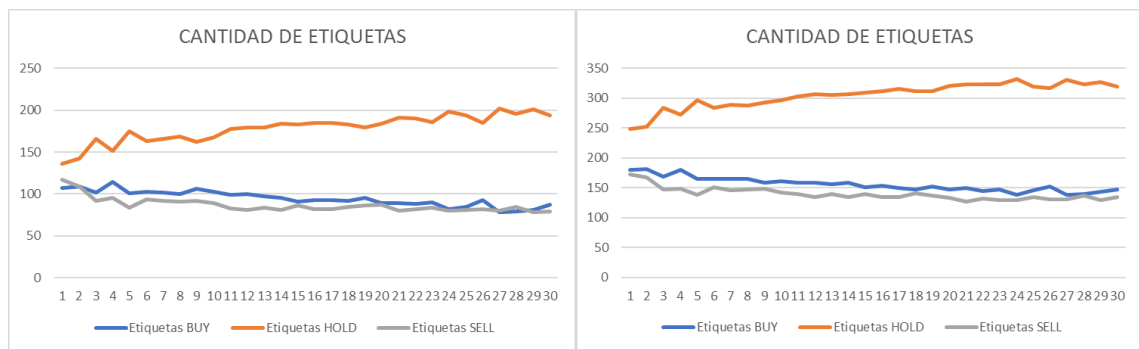


Figura 28. Cantidad de etiquetas BUY, HOLD, y SELL del grupo de 3M - MERCK - P&G (izquierda) y del grupo de todas las acciones (derecha)

Se puede observar cómo la distribución de las etiquetas es claramente desigual, siendo la parte HOLD aproximadamente el doble de cualesquiera BUY, o SELL. Esto puede generar problemas. Sin embargo, para la evaluación del modelo se usará no solo la certeza del modelo, igualmente se evaluará el comportamiento de la predicción de SELL, esto debido a que este proyecto está inicialmente dirigido a proteger personas con poca experiencia, por lo que es de suma importancia el evitar posibles pérdidas que inclusive produzcan un retiro definitivo de las inversiones en el mercado bursátil. De esta manera, y al evaluar al menos el comportamiento del modelo general y la predicción de las etiquetas SELL, se puede encontrar una forma más objetiva de evaluar y entrenar el modelo.

7.6 MÉTRICAS DE OPTIMIZACIÓN EN LOS MÉTODOS DE CLASIFICACIÓN

En este trabajo se hizo uso de diferentes métricas para evaluar el rendimiento y comportamiento de los modelos de aprendizaje de máquina entrenados. Las principales métricas usadas fueron la certeza del modelo y la puntuación F1 de la categoría “SELL”. La certeza es esencial debido a que mide la proporción de casos en los que el modelo entrenado predijo correctamente las categorías del conjunto de prueba. Por tanto, esta métrica nos dice que tan bueno es un modelo en términos generales. Por otra parte, se usó la puntuación F1, o F1-score por su traducción al inglés, para determinar la media armónica de la precisión y el “recall” cuando se calcula una categoría en especial. La puntuación F1 de “SELL” se usa debido a que está en nuestro interés predecir esta categoría sobre “BUY” y “HOLD”, debido a que el predecirla correctamente conllevará predecir una posible pérdida de capital por parte de las personas que posean acciones de las compañías usadas para las pruebas. En este caso, al ver que la métrica de certeza tiende a ser consistentemente más alta (ver sección 8.1) que la de F1-Score SELL, y teniendo en cuenta la importancia de la predicción de la clase SELL para pequeños inversionistas, se decidió optimizar ambas métricas de certeza y de F1-score SELL en los ciclos FOR. (ciclos para la optimización de hiperparámetros). Se realizó esto debido a que se considera importante la correcta predicción de todas las clases del modelo de

aprendizaje de máquina entrenado, así como la correcta predicción de la clase SELL, debido a su importancia especial para prever una posible pérdida de capital por parte de inversionistas.

Finalmente, se usó la matriz de confusión para presentar los valores predichos en función de los valores reales. Esta matriz es usada para los modelos después de una optimización completa de parámetros y después de una ardua selección de set de datos y parámetros a usar para la predicción[32].

7.7 PARÁMETROS ADICIONALES PARA CONSIDERAR PARA LA IMPLEMENTACIÓN DE LOS MÉTODOS DE APRENDIZAJE DE MÁQUINA

Los modelos de aprendizaje de máquina que se implementaron en este trabajo tuvieron 2 variables:

1. Sorpresa de beneficios por acción (único parámetro de entrada)
2. Etiqueta BUY/HOLD/SELL, que se determina dependiendo de la variación porcentual del precio de la acción (parte supervisada del modelo)

Estos ya fueron descritos anteriormente. Sin embargo, para las pruebas y optimización final de los modelos de aprendizaje de máquina, se usó la variable de sorpresa de beneficios por acción en 4 datasets diferentes definidos de la siguiente manera:

1. **Dataset normal:** tiene el parámetro de sorpresa de beneficios por acción únicamente
2. **Dataset histórico 1:** tiene el parámetro de sorpresa de beneficios por acción, así como el dato de sorpresa de beneficios por acción 1 periodo inmediatamente anterior. Este método tiene 2 columnas de entrada en el documento csv. Además, solo tiene 119 datos debido a que, al necesitar un dato previo, esto limita el primer dato que se tiene.
3. **Dataset histórico 2:** tiene el parámetro de sorpresa de beneficios por acción, así como el dato de sorpresa de beneficios por acción de los 2 periodos inmediatamente anteriores. Este método tiene 3 columnas de entrada en el documento csv. Además, solo tiene 118 datos debido a que, al necesitar dos datos previos, esto limita el primer y segundo dato que se tiene.
4. **Dataset histórico 3:** tiene el parámetro de sorpresa de beneficios por acción, así como el dato de sorpresa de beneficios por acción de los 3 periodos inmediatamente anteriores. Este método tiene 4 columnas de entrada en el documento csv. Además, solo tiene 117 datos debido a que, al necesitar tres datos previos, esto limita los primeros 3 datos que se tienen.

8 IMPLEMENTACIÓN DE MÉTODOS DE APRENDIZAJE DE MÁQUINA DE CLASIFICACIÓN

8.1 PRUEBAS INICIALES DE MÉTODOS DE CLASIFICACIÓN

Se hizo uso de los conocimientos adquiridos en el curso de “Introducción al aprendizaje de máquina en Python: un enfoque práctico” para realizar las pruebas iniciales respectivas. Se usó el método KNN, y el método árbol de decisión para realizar las pruebas iniciales de métodos de clasificación. Se sabe que cada método de aprendizaje de máquina tiene valores variables llamados hiperparámetros que pueden influir en el comportamiento del algoritmo de aprendizaje de máquina, y, por tanto, modificar los valores que este predice, afectando positiva o negativamente el comportamiento del modelo. Estos valores es recomendable modificarlos a conveniencia para lograr tener un modelo con mejores resultados en la métrica que se desee. Además, existen valores que se asignan por defecto a cada hiperparámetro en caso de que el usuario no lo especifique. Para la prueba inicial no se realizará ningún tipo de modificación de hiperparámetros. En la tabla 9 se pueden observar los hiperparámetros disponibles para optimizar el método KNN y el método árbol de decisión, y una pequeña descripción de cada uno, así como los valores que se toman por defecto[55][56]. Las técnicas usadas pertenecen a la librería sklearn, y la función específica de la librería que permite esto puede verse más adelante en el trabajo, en la tabla 11.

Hiperparámetros	KNN	Descripción	Default	Arbol de decisión	Descripción	Default
1	n_neighbors	Número de vecinos a usar para el entrenamiento del modelo	5	criterion	Define la función para medir la calidad del split	gini
2	weights	Función de peso utilizada en la predicción	uniform	splitter	Define la estrategia para elegir el split en cada nodo	best
3	algorithm	Algoritmo utilizado para calcular los vecinos más cercanos	auto	max_depth	La máxima profundidad del arbol	None
4	leaf_size	Controla el número mínimo de puntos en un nodo dado	30	min_samples	El número mínimo de muestras requeridas para dividir un nodo interno	2
5	p	Parámetro de potencia para la métrica de Minkowski	2	min_samples_leaf	El mínimo número de muestras requerido para estar en la hoja de un nodo	1
6	metric	Métrica a utilizar para el cálculo de la distancia	minkowski	random_state	Controla la aleatoriedad del estimador	None
7	metric_params	Argumentos de palabras clave adicionales para la función "metric"	None	max_features	El número de características a considerar al buscar el mejor split	None
8	n_jobs	El número de trabajos paralelos a ejecutar para la búsqueda de vecinos	None	max_leaf_nodes	El número máximo de muestras que puede tener la hoja de un nodo	None

Tabla 9. Hiperparámetros de los métodos KNN y árbol de decisión

El método KNN es un método de clasificación supervisado que consiste en la ubicación espacial de los puntos de información, para luego usar la proximidad de los datos a predecir con otros datos, pertenecientes al modelo entrenado, para definir la categoría con base en las categorías de los puntos más cercanos dentro de un perímetro o espacio geométrico definido [57]. Se implementó este modelo en cada uno de los sets de datos, cambiando la variable variación porcentual en el rango desde 1 a 30 días. En las figuras 29 a 36 se pueden ver la evolución de las métricas de certeza y F1score SELL en función de la variable variación porcentual. Es importante mencionar que para cada set de datos se usó el 80% de los elementos para el entrenamiento del modelo y un 20% para probar las métricas de certeza, y el F1-score SELL del modelo.

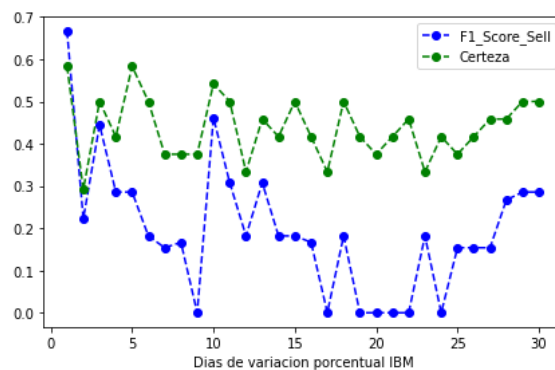


Figura 29. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos IBM al entrenar el modelo KNN sin optimización de parámetros

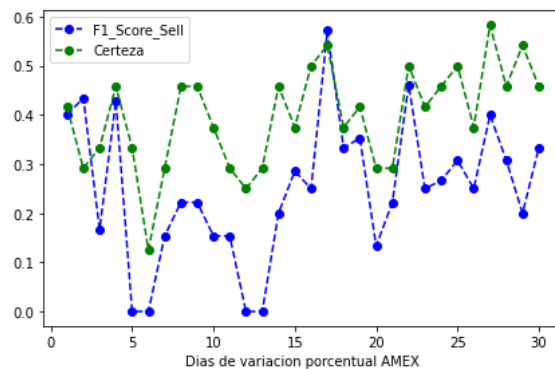


Figura 30. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos AMEX al entrenar el modelo KNN sin optimización de parámetros

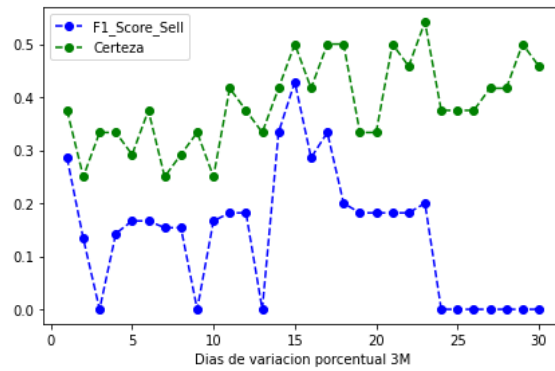


Figura 31. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos 3M al entrenar el modelo KNN sin optimización de parámetros

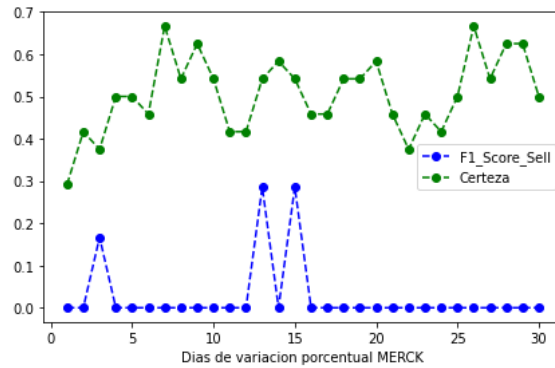


Figura 32. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos MERCK al entrenar el modelo KNN sin optimización de parámetros

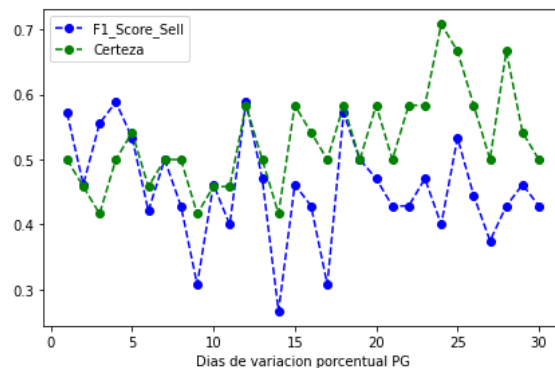


Figura 33. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos PG al entrenar el modelo KNN sin optimización de parámetros

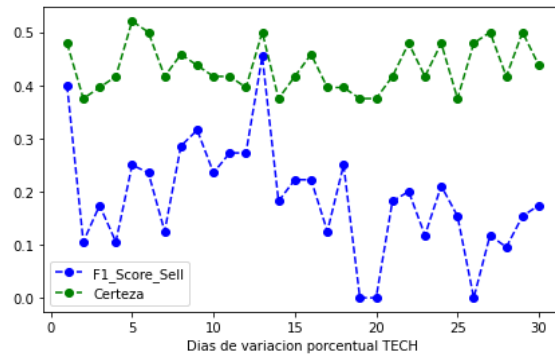


Figura 34. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos TECH al entrenar el modelo KNN sin optimización de parámetros

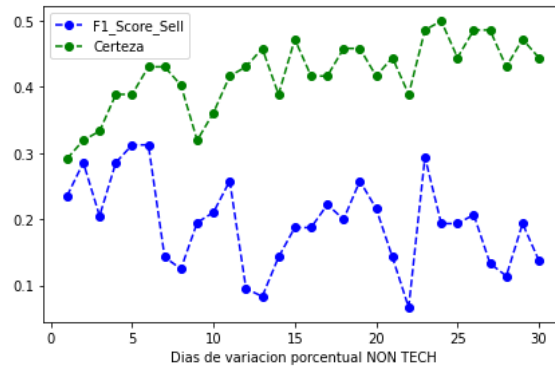


Figura 35. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos NONTECH al entrenar el modelo KNN sin optimización de parámetros

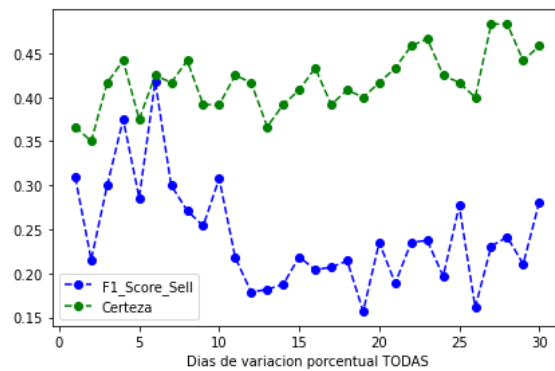


Figura 36. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos ALL al entrenar el modelo KNN sin optimización de parámetros

Por otra parte, el método árbol de decisión es un método supervisado que consta de nodos, ramas y hojas, el cual organizar los datos en una estructura jerárquica dentro de las distintas partes del árbol [58]. Se implementó este modelo en cada uno de los sets de datos, cambiando la variable variación porcentual en el rango desde 1 a 30 días. En las figuras 37 a 44 se pueden ver la evolución de las métricas de certeza y F1score SELL en función de la variable del periodo en días de variación porcentual. Se recuerda que se usó el 80% de los elementos para el entrenamiento del modelo y un 20% para la prueba de métricas.

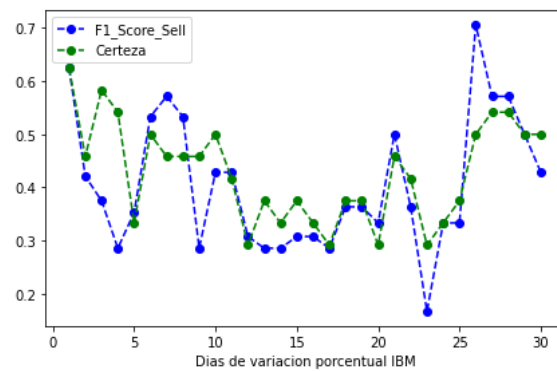


Figura 37. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos IBM al entrenar el modelo árbol de decisión sin optimización de parámetros

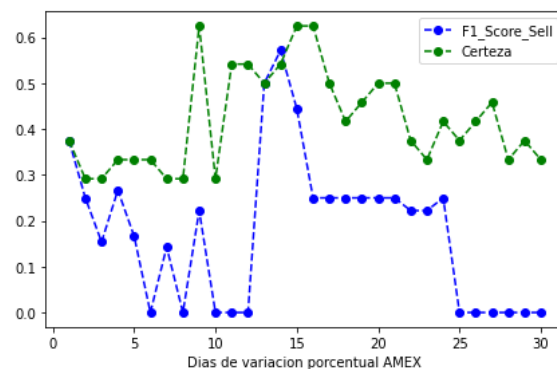


Figura 38. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos AMEX al entrenar el modelo árbol de decisión sin optimización de parámetros

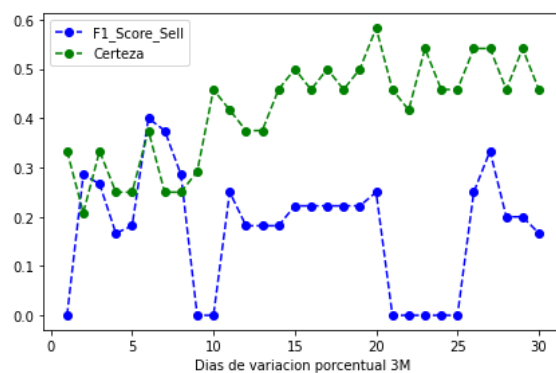


Figura 39. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos 3M al entrenar el modelo árbol de decisión sin optimización de parámetros

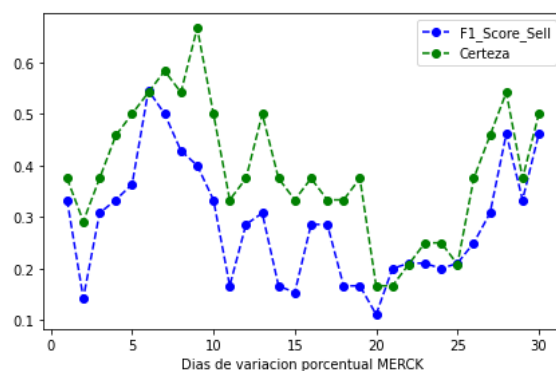


Figura 40. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos MERCK al entrenar el modelo árbol de decisión sin optimización de parámetros

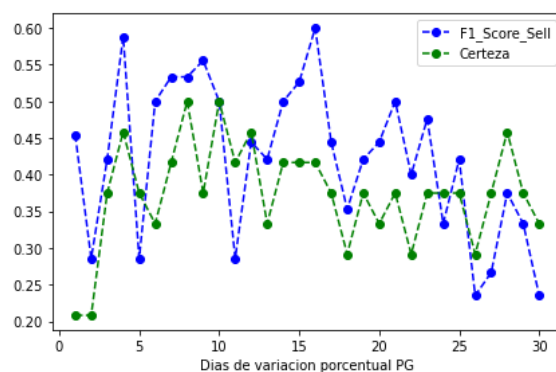


Figura 41. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos PG al entrenar el modelo árbol de decisión sin optimización de parámetros

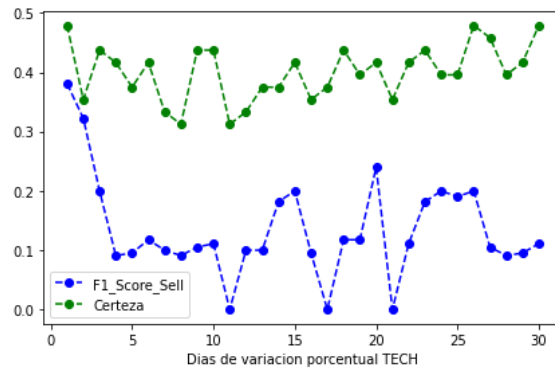


Figura 42. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos TECH al entrenar el modelo árbol de decisión sin optimización de parámetros

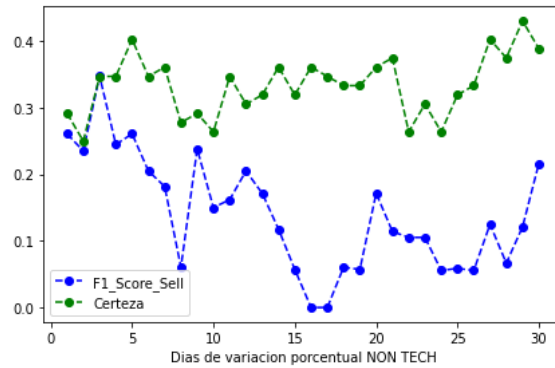


Figura 43. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos NONTECH al entrenar el modelo árbol de decisión sin optimización de parámetros

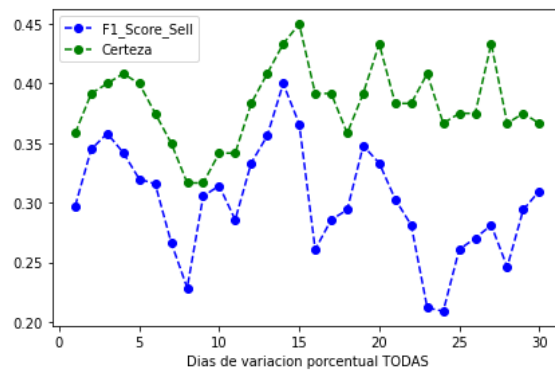


Figura 44. Evolución (en función de los días de variación porcentual) de F1-score SELL y certeza del set de datos ALL al entrenar el modelo árbol de decisión sin optimización de parámetros

En la figura 45 a 48 se muestran los mejores resultados de cada uno de los sets de datos para cada método de aprendizaje de máquina. Adicionalmente, se muestra el promedio de resultados obtenidos en cada uno de los sets de datos. Esto tanto para la métrica de certeza, como para la métrica de F1-score SELL.

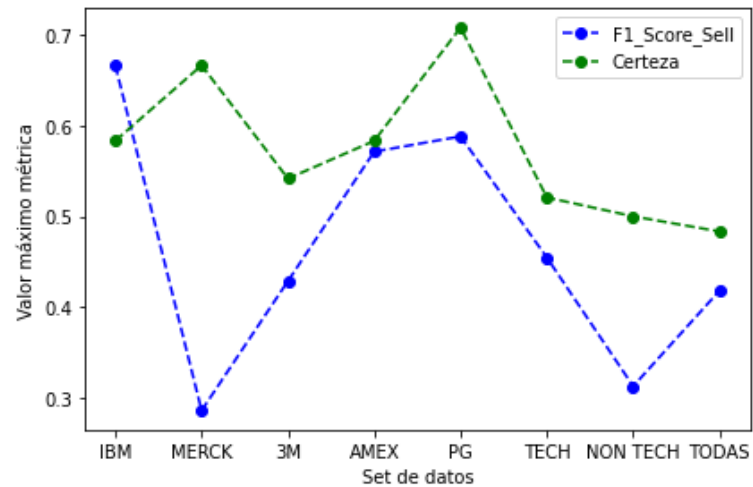


Figura 45. Mejores resultados de cada uno de los sets de datos para el método KNN

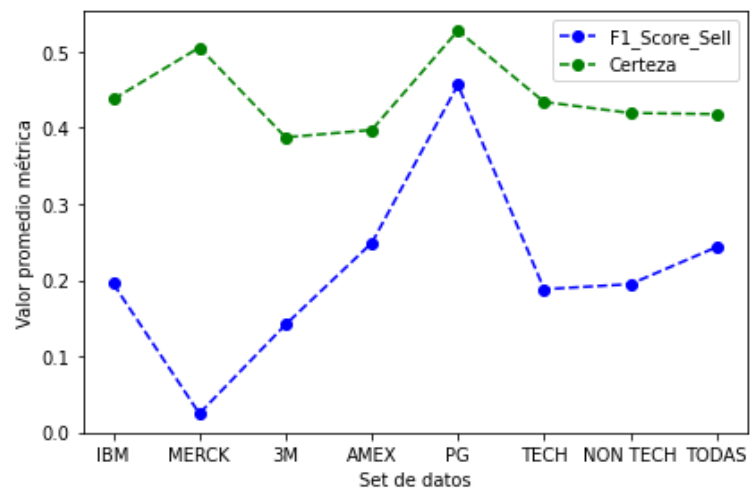


Figura 46. Resultado promedio de cada uno de los sets de datos para el método KNN

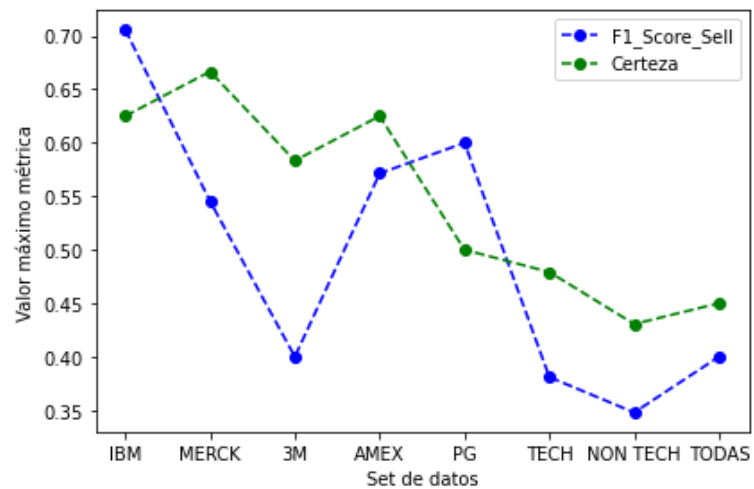


Figura 47. Mejores resultados de cada uno de los sets de datos para el método árbol de decisión

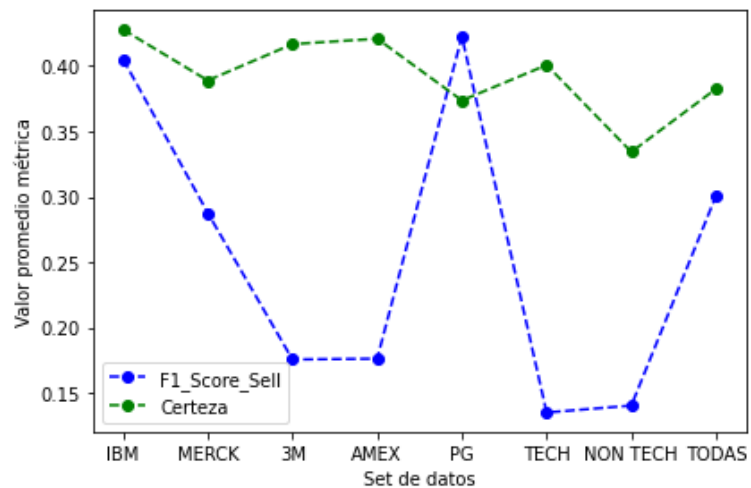


Figura 48. Resultado promedio de cada uno de los sets de datos para el método árbol de decisión

Después de esta prueba inicial se realizó una optimización de algunos métodos de aprendizaje de máquina que fue definida en la sección 8.3. Para esta optimización, se seleccionaron los 2 datasets que obtuvieron mejores resultados. En este caso, P and G obtuvo los mejores resultados debido a que tanto en F1-score SELL como en certeza logró los resultados más altos según la figura 45. Además, igualmente logró el mejor promedio de ambas métricas, tal como se puede apreciar en las figuras 46 y 48. Además, se seleccionó el set de datos IBM debido a que obtuvo resultados altos en ambos métodos de aprendizaje de máquina, aunque especialmente en KNN de forma comparativa a los otros sets de datos.

Además de esto, se seleccionaron únicamente 3 periodos de P and G, e IBM (diferentes entre sí), dependiendo del periodo de días de variación porcentual que obtuvo mejores resultados. Estos modelos fueron seleccionados dependiendo los resultados mostrados tanto en F1-score SELL, cómo en certeza del modelo. Estos modelos fueron sujetos a una optimización inicial de parámetros. Esta optimización se realizó con el fin de evaluar de forma general rangos de hiperparámetros óptimos que los que se tomaron en cuenta para la optimización. Esto debido a que en la sección 8.4 se realizará una optimización general teniendo en cuenta todos los sets de datos y todos los periodos de variación porcentual de N días, por lo que el tiempo de ejecución del programa será un problema para tener en cuenta. En la tabla 10 se muestran los tiempos de ejecución para el programa que realizó el modelo KNN y de árbol de decisión. Este programa no tiene ningún tipo de optimización.

SETS DE DATOS	Tiempo de ejecución /modelo	Tiempo de ejecución /set de datos	Tiempo de ejecución total
Sets de pruebas - KNN - Modelo No Optimizado	0,012 segundos	0,38 segundos	3,01 segundos
Sets de pruebas - Decision Tree - Modelo No Optimizado	0,011 segundos	0,33 segundos	2,67 segundos

Tabla 10. Tiempos de ejecución para cada tipo de modelo de aprendizaje de máquina sin optimización de hiperparámetros

8.2 ELECCIÓN DE MÉTODOS DE CLASIFICACIÓN

Resulta evidente que es necesario encontrar un método de aprendizaje de máquina apropiado para el tipo de datos y la cantidad de datos que se emplean en este trabajo. Para ello, se consultaron diferentes trabajos de aprendizaje de máquina enfocados al mercado bursátil. Sin embargo, no fue posible encontrar trabajos abiertos al público con la limitación de datos que se tiene en este caso (aproximadamente 120 por set de datos), esto principalmente debido a que trabajos relacionados con el mercado bursátil suelen estar ligados a métodos de análisis del precio histórico de las acciones, y no suelen usar parámetros financieros como el usado en este trabajo. Por esta razón, se usaron métodos de aprendizaje de máquina especialmente buenos con pocos datos. Para esto, se encontró un trabajo dónde se probaron 108 datasets con diferentes métodos de aprendizaje de máquina que se consideraron como los más usados, en este caso los más usados por la fundación Bill and Melinda Gates. En la figura 49 se presentan los resultados del rendimiento del modelo en función del número de datos[59].

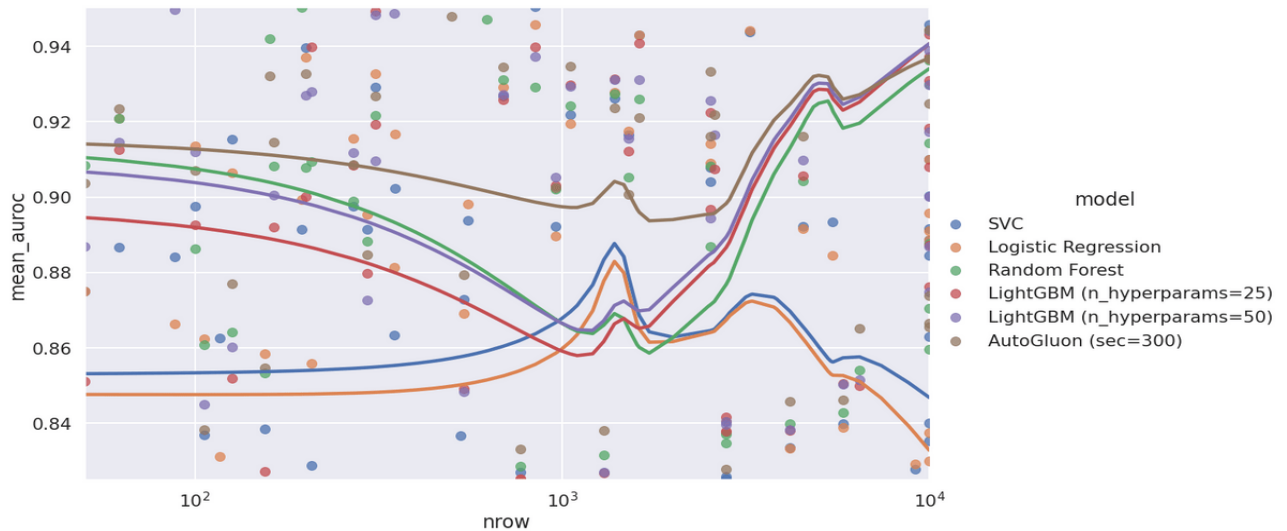


Figura 49. Mejores métodos de aprendizaje de máquina de clasificación para sets de datos con pocos elementos[59]

Sin embargo, al probar estos métodos de aprendizaje de máquina se encontraron problemas de compatibilidad con el método Auto Gluon, el cual requiere una gran cantidad de librerías para su funcionamiento, las cuales en muchos casos tienen que ser de versiones que no permiten el funcionamiento de las librerías de los otros métodos de aprendizaje de máquina. Además, se presentaron problemas de la herramienta que permite escanear información tabular, problema, que según la documentación de AutoGluon, puede ser debido inclusive a la falta de memoria del equipo dónde es ejecutado el método de aprendizaje de máquina[60]. Por dichos motivos, este método no se tendrá en cuenta en este trabajo. En la tabla 11 se muestran los métodos de aprendizaje de máquina usados, así como una pequeña descripción

de cada uno. Es de notar que también fueron usados y optimizados los métodos probados preliminarmente (KNN y árbol de decisión).

Método de aprendizaje de máquina	DESCRIPCIÓN	LIBRERÍA
KNN	Es un método que utiliza la proximidad de los datos para generar la predicción o hacer la clasificación	<code>sklearn.neighbors.KNeighborsClassifier</code>
Arbol de Decisión	Es un método que tiene una estructura jerárquica y consiste de un nudo principal, ramas, nodos internos y hojas	<code>sklearn.tree.DecisionTreeClassifier</code>
Random Forest	Es un método que combina la salida de múltiples árboles de decisión para lograr un resultado único	<code>sklearn.ensemble.RandomForestClassifier</code>
Lightgbm	Es un "marco de mejora de gradiente" que usa algoritmos de árbol de decisión como base. Está diseñado para ser eficiente y distribuido	<code>lightgbm.LGBMClassifier</code>
Support vector classifier(SVC)	Es un método que maximiza la precisión de un modelo sin causar overfitting. Este método es especialmente útil para datasets de muchos elementos	<code>sklearn.svm.SVC</code>
Logistic regression	Este método estima las probabilidades de que ocurra un evento. Este método usa una transformación logit a las probabilidades.	<code>sklearn.linear_model.LogisticRegression</code>

Tabla 11. Descripción de métodos de aprendizaje de máquina usados[57], [58], [61]–[64]

8.3 PRUEBAS INICIALES DE OPTIMIZACIÓN DE HIPERPARÁMETROS

La optimización de hiperparámetros es una parte importante de la implementación de métodos de aprendizaje de máquina. Esta permite evaluar modelos de aprendizaje de máquina con diferentes hiperparámetros para lograr encontrar la hiperparámetros que generan mejores resultados[65]. Métodos comunes para hacer esto consiste en aplicar la función “gridsearch”, la cual busca la mejor combinación de parámetros. Sin embargo, gridsearch tiene las desventajas de tener la posibilidad de overfitting. Además, se quiere optimizar un modelo por dos parámetros: certeza, y F1-score SELL, por lo que se optó por realizar un programa con ciclos FOR anidados para lograr la optimización de hiperparámetros. En este caso, al ver que la métrica de certeza tiende a ser consistentemente más alta que la de F1-Score SELL, y teniendo en cuenta la importancia de la predicción de la clase SELL para pequeños inversionistas, se decidió optimizar ambas métricas de certeza y de F1-score SELL en los ciclos FOR. Esto, debido a que se considera importante la correcta predicción de todas las clases del modelo de aprendizaje de máquina entrenado, así como la correcta predicción de la clase SELL, debido a su importancia especial para prever una posible pérdida de capital por parte de inversionistas.

Para esta optimización inicial, se espera reducir los rangos de cada parámetro de optimización, ya que los tiempos de ejecución es un problema real que se muestra en la siguiente sección de este capítulo. Además, usando los resultados de los métodos usados en la primera sección de este capítulo, se hizo la optimización de parámetros para el set de datos tanto de P&G como de IBM, debido a que en términos medios y de mejores resultados, obtuvieron los mejores resultados comparativos de las métricas F1-score de “SELL”, así como de certeza. De hecho, se hizo la optimización únicamente para los 3 modelos que obtuvieron mejores resultados dependiendo de la variación porcentual en días usada para la predicción (rango entre 1 y 30 días), esto debido a los altos tiempo de ejecución en algunos de los métodos de aprendizaje de máquina. Estos periodos fueron seleccionados comparando los mayores picos de ambas métricas de cada uno de los sets de datos, así como con el análisis de resultados promedios y mejores resultados realizados en la sección 8.1. En la tabla 12 se muestran los días de variación porcentual seleccionados.

Set de datos	Días de variación porcentual seleccionados
IBM	1 días, 10 días, 30 días
P & G	12 días, 18 días, 24 días

Tabla 12. Periodos de días de variación porcentual seleccionados para la optimización de hiperparámetros en la prueba inicial

En las tablas 13 a 18 se pueden ver los hiperparámetros usados para cada método de aprendizaje de máquina, así como el rango usado y una pequeña descripción de cada método. Para la selección de los rangos se consultaron proyectos que usaban y optimizaban cada uno de estos métodos.

KNN		
HIPERPARÁMETROS	Descripción	Rango Usado
n_neighbors	Número de vecinos a usar para el entrenamiento del modelo	1 a 21
metric	Métrica a utilizar para el cálculo de la distancia	['euclidean', 'manhattan', 'minkowski']
leaf_size	Controla el número mínimo de puntos en un nodo dado	1 a 50

Tabla 13. Rangos de hiperparámetros usados para la optimización del método KNN en la etapa de prueba inicial[55]

Arbol De Decisión		
HIPERPARÁMETROS	Descripción	Rango Usado
max_depth	La máxima profundidad del arbol	1 a 7
max_leaf_nodes	El número máximo de muestras que puede tener la hoja de un nodo	2 a 10
max_features	El número de características a considerar al buscar el mejor split	['sqrt','log2','auto',None]

Tabla 14. Rangos de hiperparámetros usados para la optimización del método árbol de decisión en la etapa de prueba inicial[56]

Random Forest		
HIPERPARÁMETROS	Descripción	Rango Usado
max_depth	La máxima profundidad del arbol	1 a 7
max_leaf_nodes	El número máximo de muestras que puede tener la hoja de un nodo	2 a 10
n_estimators	el número de arboles en el modelo	[10,100,1000]
max_features	El número de características a considerar al buscar el mejor split	['sqrt', 'log2','auto', None]

Tabla 15. Rangos de hiperparámetros usados para la optimización del método Random Forest en la etapa de prueba inicial[66]

Logistic Regression		
HIPERPARÁMETROS	Descripción	Rango Usado
solver	El algoritmo a usar en la optimización del problema	['newton-cg', 'liblinear','lbfgs', 'sag', 'saga']
penalty	Se imponen penalizaciones por tener demasiadas variables. Reduce los coeficientes de las variables que menos contribuyen. Esta variable define que tipo de penalidad se impone	[['none', 'l2'], ['l1','l2'],['none', 'l2'],['none', 'l2'],['elasticnet', 'l1', 'l2', 'none']]
C	Es el inverso de la fuerza de regularización. Este dato corresponde a cuánto se quiere evitar clasificar erróneamente cada ejemplo de entrenamiento.	[100, 10, 1.0, 0.1, 0.01]

Tabla 16. Rangos de hiperparámetros usados para la optimización del método “Logistic Regression” en la etapa de prueba inicial[67]

Support Vector Classifier (SVC)		
HIPERPARÁMETROS	Descripción	Rango Usado
kernel	Especifica el tipo de filtro que se va a usar en el algoritmo	['linear', 'rbf','poly']
gamma	EL coeficiente del kernel	[0.1, 1, 10, 100]
C	Es el inverso de la fuerza de regularización. Este dato corresponde a cuánto se quiere evitar clasificar erróneamente cada ejemplo de entrenamiento.	[0.1, 1, 10, 100, 1000]
degree	Grado de la función polinómica del kernel	[1, 2, 3]

Tabla 17. Rangos de hiperparámetros usados para la optimización del método SVC en la etapa de prueba inicial[68]

Light GBM		
HIPERPARÁMETROS	Descripción	Rango Usado
num_leaves	Número máximo de hojas en un arbol	2 a 50
n_estimators	el número de arboles en el modelo	1 a 100
max_depth	La máxima profundidad del arbol	1 a 7

Tabla 18. Rangos de hiperparámetros usados para la optimización del método Light GBM en la etapa de prueba inicial[69]

Los rangos usados para la optimización fueron consultados de proyectos realizados en comparación y optimización de hiperparámetros de métodos de aprendizaje de máquina. El rango usado para el algoritmo LightGBM fue encontrado en la página oficial de su documentación, donde se nombran los hiperparámetros recomendados para su optimización, además, se recomienda el rango máximo a tener en cuenta dependiendo de la variable max_depth. Por otra parte, el rango usado para el método SVC se encontró en el artículo “In Depth: Parameter tuning for SVC”. Finalmente, los rangos para los métodos árbol de decisión, Random Forest, Logistic regression, y KNN; fueron encontrados en el artículo “Tune Hiperparameters for Classification Machine Learning Algorithms”[70][71][69].

En las figuras 50 a 55 se pueden ver, para cada método de aprendizaje de máquina, los mejores resultados obtenidos en los sets de datos seleccionados y con los periodos de días de variación porcentual seleccionados.

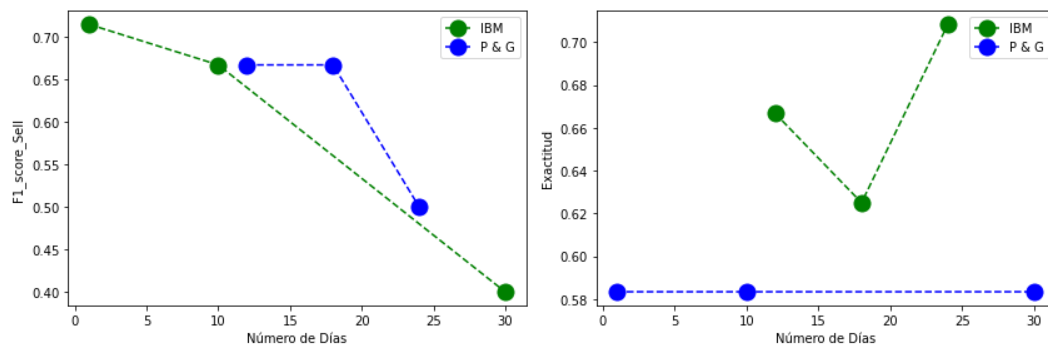


Figura 50. Mejores resultados de F1-score SELL (izquierda) y de certeza (derecha) para el método de KNN

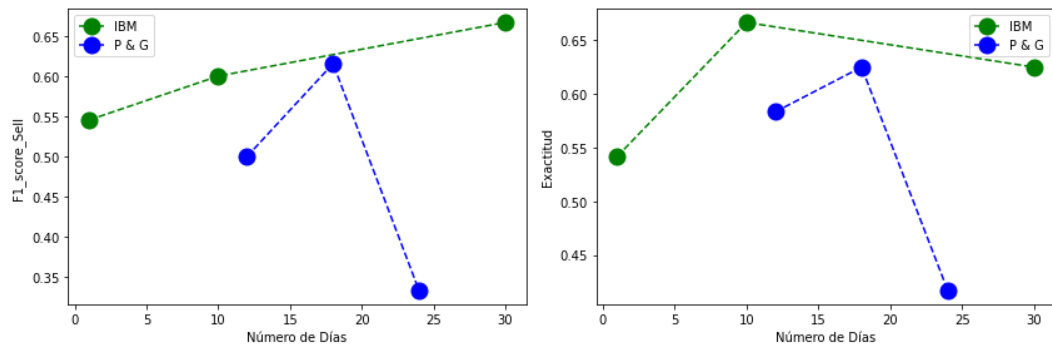


Figura 51. Mejores resultados de F1-score SELL (izquierda) y de certeza (derecha) para el método de árbol de decisión

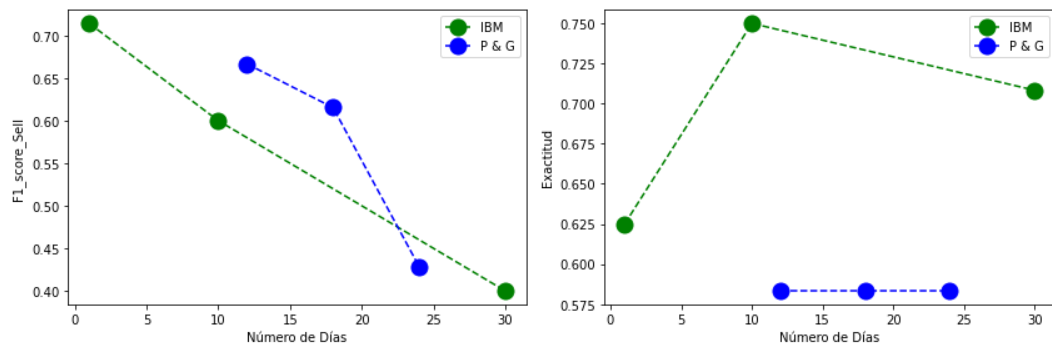


Figura 52. Mejores resultados de F1-score SELL (izquierda) y de certeza (derecha) para el método de Random Forest

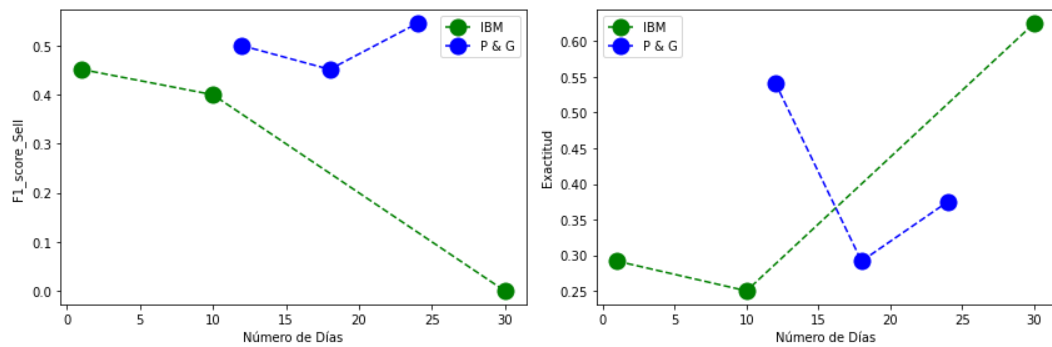


Figura 53. Mejores resultados de F1-score SELL (izquierda) y de certeza (derecha) para el método de Logistic Regression

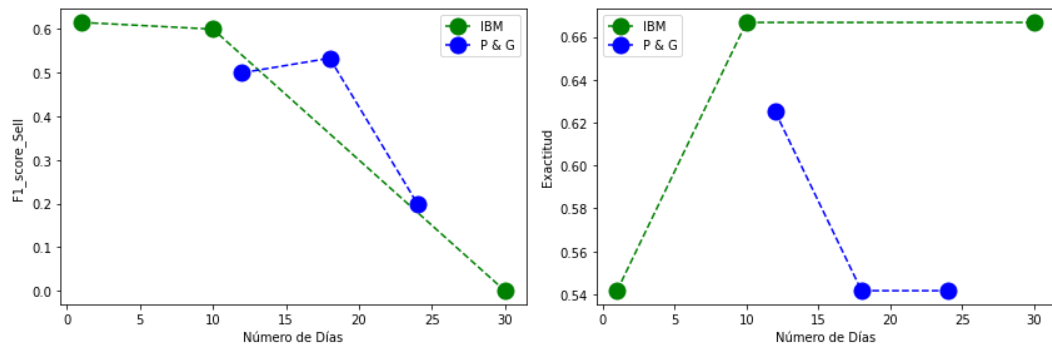


Figura 54. Mejores resultados de F1-score SELL (izquierda) y de certeza (derecha) para el método de SVC

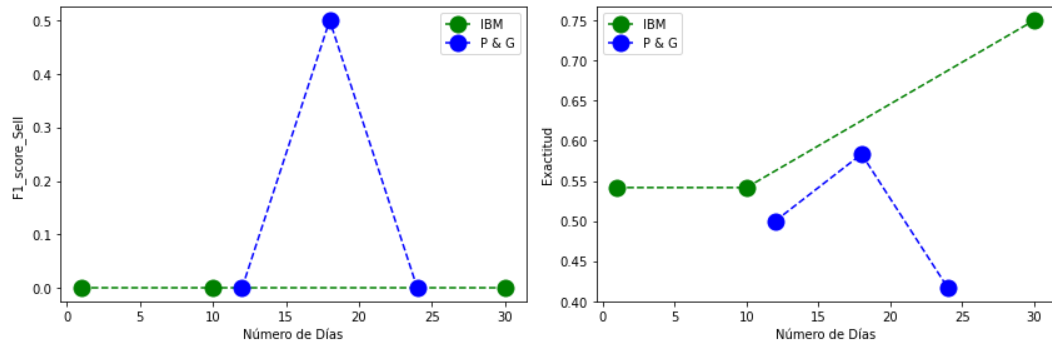


Figura 55. Mejores resultados de F1-score SELL (izquierda) y de certeza (derecha) para el método de Light GBM

Además, se logró recopilar las hiperparámetros que en cada caso generaron los mejores resultados. Estos resultados fueron usados para acotar los rangos de hiperparámetros a optimizar, para así conseguir tiempos de optimización manejables computacionalmente, ya que, como se verá en siguientes secciones, los tiempos de optimización pueden llegar a ser de inclusive días para algunos modelos.

En las tablas 19 a 24 se pueden observar los hiperparámetros que lograron los mejores resultados al entrenar los sets de datos con cada método de aprendizaje de máquina.

	PORCENTAJE DE VARIACIÓN(DÍAS)- KNN					
	SET DE DATOS DE IBM			SET DE DATOS DE P&G		
HIPERPARÁMETRO	1 DÍA	10 DÍAS	30 DÍAS	12 DÍAS	18 DÍAS	24 DÍAS
n_neighbors	5	10	4	8	6	5
metric	Euclidean	Euclidean	Euclidean	Euclidean	Euclidean	Euclidean
leaf_size	1	12	1	1	1	12

Tabla 19. Hiperparámetros que lograron los mejores resultados para el método KNN

	PORCENTAJE DE VARIACIÓN(DÍAS)- DECISION TREE					
	SET DE DATOS DE IBM			SET DE DATOS DE P&G		
HIPERPARÁMETRO	1 DÍA	10 DÍAS	30 DÍAS	12 DÍAS	18 DÍAS	24 DÍAS
max_depth	5	4	2	6	4	6
max_leaf_nodes	9	6	3	9	10	8
max_features	sqrt	sqrt	sqrt	sqrt	sqrt	sqrt

Tabla 20. Hiperparámetros que lograron los mejores resultados para el método Decision Tree

	PORCENTAJE DE VARIACIÓN(DÍAS)- RANDOM FOREST					
	SET DE DATOS DE IBM			SET DE DATOS DE P&G		
HIPERPARÁMETRO	1 DÍA	10 DÍAS	30 DÍAS	12 DÍAS	18 DÍAS	24 DÍAS
max_depth	6	5	5	3	2	6
max_leaf_nodes	9	9	6	10	3	4
n_estimators	10	10	10	10	10	10
max_features	log2	None	log2	sqrt	None	log2

Tabla 21. Hiperparámetros que lograron los mejores resultados para el método Random Forest

	PORCENTAJE DE VARIACIÓN(DÍAS)- LOGISTIC REGRESSION					
	SET DE DATOS DE IBM			SET DE DATOS DE P&G		
HIPERPARÁMETRO	1 DÍA	10 DÍAS	30 DÍAS	12 DÍAS	18 DÍAS	24 DÍAS
solver	newton-cg	liblinear	liblinear	liblinear	liblinear	newton-cg
penalty	none	l1	l1	l1	l1	l2
C	10	0.01	0.01	0.01	0.01	0.01

Tabla 22. Hiperparámetros que lograron los mejores resultados para el método Logistic Regression

	PORCENTAJE DE VARIACIÓN(DÍAS)- SVC					
	SET DE DATOS DE IBM			SET DE DATOS DE P&G		
HIPERPARÁMETRO	1 DÍA	10 DÍAS	30 DÍAS	12 DÍAS	18 DÍAS	24 DÍAS
kernel	linear	rbf	rbf	rbf	rbf	rbf
gamma	0.1	100	100	100	100	100
C	100	1000	1000	10	1	1
degree	0	0	0	0	0	0

Tabla 23. Hiperparámetros que lograron los mejores resultados para el método SVC

	PORCENTAJE DE VARIACIÓN(DÍAS)- LIGHT GBM					
	SET DE DATOS DE IBM			SET DE DATOS DE P&G		
HIPERPARÁMETRO	1 DÍA	10 DÍAS	30 DÍAS	12 DÍAS	18 DÍAS	24 DÍAS
num_leaves	2	3	2	3	2	2
n_estimators	1	58	1	32	1	60
max_depth	1	2	1	2	1	1

Tabla 24. Hiperparámetros que lograron los mejores resultados para el método Light GBM

Adicionalmente, en la tabla 25 se pueden ver los tiempos de ejecución para cada método de aprendizaje de máquina. Aquí podemos notar que, a diferencia de los métodos sin optimizar y métodos sencillos, algunos de los métodos ya pueden tardar más de una hora en ejecutarse.

MODELO DE APRENDIZAJE DE MÁQUINA	Tiempo de ejecución /modelo	Tiempo de ejecución /set de datos	Tiempo de ejecución total
KNN	8,73 segundos	26,19 segundos	52,38 segundos
DECISION TREE	0,48 segundos	1,22 segundos	2,44 segundos
RANDOM FOREST	6,52 minutos	19,57 minutos	39,13 minutos
LOGISTIC REGRESSION	0,36 segundos	1,08 segundos	2,17 segundos
SVC	1,26 segundos	3,78 segundos	7,57 segundos
LIGHT GBM	14,64 minutos	43,94 minutos	87,87 minutos

Tabla 25. Tiempos de ejecución para la optimización de cada método de aprendizaje de máquina en la etapa de prueba inicial

8.4 OPTIMIZACIÓN GENERAL DE HIPERPARÁMETROS

Se realizó un programa que busca encontrar la mejor combinación posible entre las siguientes variables, parámetros, y modelos:

1. Modelo de aprendizaje de máquina
2. Tipo de dataset: Normal, histórico 1, histórico 2, histórico 3

Este algoritmo también incluyó la optimización de cada método de aprendizaje de máquina según los rangos definidos anteriormente. Además, estos rangos fueron reducidos de acuerdo con los resultados encontrados en la prueba inicial. En la tabla 26 se muestran los rangos usados para cada uno de los métodos de aprendizaje de máquina considerados. En la tabla fueron ubicados los rangos en el mismo orden que fueron ordenados en las tablas 13 a 18.

HIPERPARÁMETROS	KNN	DECISION TREE	RANDOM FOREST	LOGISTIC REGRESSION	SVC	LIGHT GBM
Hiperparámetro1	1 a 21	1 a 4	1 a 5	['newton-cg', 'liblinear', 'lbfgs', 'sag', 'saga']	['linear', 'rbf']	2 a 8
Hiperparámetro2	1 a 50	2 a 7	1 a 8	[['none', 'l2'], ['l1', 'l2'], ['none', 'l2'], ['none', 'l2'], ['elasticnet', 'l1', 'l2', 'none']]	[0.1, 1, 10, 100]	1 a 8
Hiperparámetro3	['euclidean', 'manhattan', 'minkowski']	['sqrt', 'log2', 'auto', 'None']	[10, 100]	[100, 10, 1.0, 0.1]	[0.1, 1, 10, 100]	1 a 3
Hiperparámetro4			['sqrt', 'auto', 'log2', 'None']			

Tabla 26. Rangos reducidos de hiperparámetros para cada método de aprendizaje de máquina

Se implementó este programa con cada set de datos. Las figuras 56 a 63 muestran las gráficas correspondientes a cada set de datos.

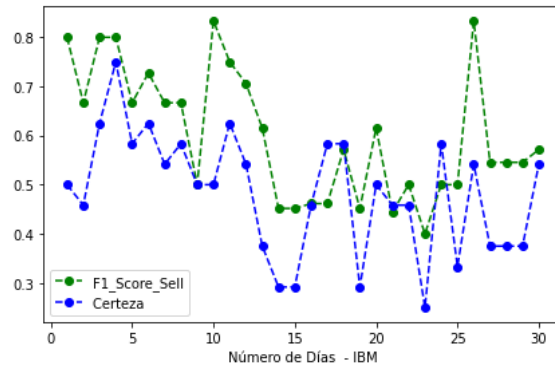


Figura 56. Mejores resultados de F1-score y certeza para el set de datos IBM con optimización reducida de hiperparámetros

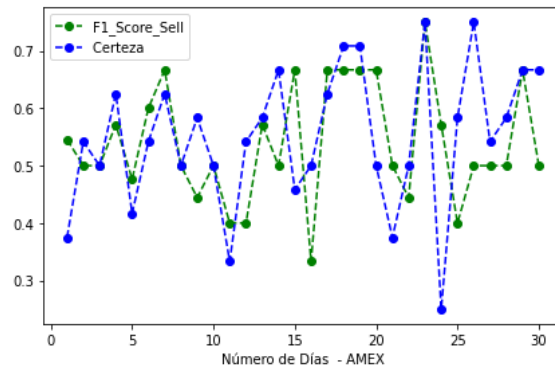


Figura 57. Mejores resultados de F1-score y certeza para el set de datos AMEX con optimización reducida de hiperparámetros

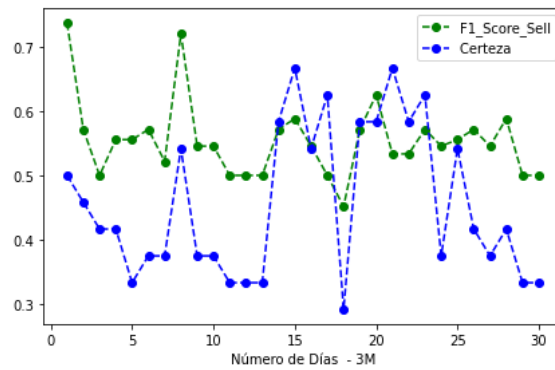


Figura 58. Mejores resultados de F1-score y certeza para el set de datos 3M con optimización reducida de hiperparámetros

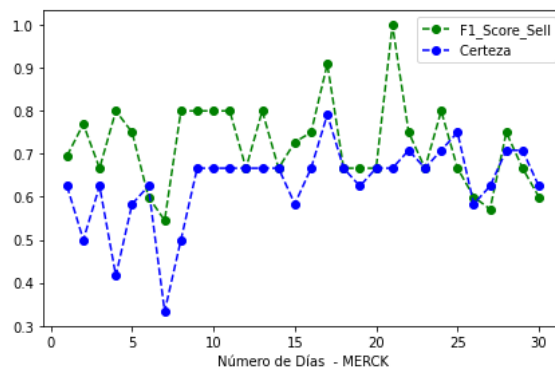


Figura 59. Mejores resultados de F1-score y certeza para el set de datos MERCK con optimización reducida de hiperparámetros

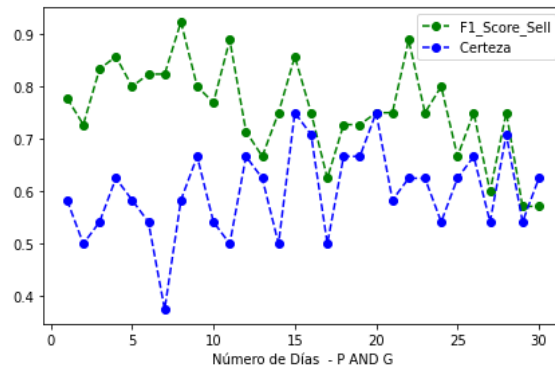


Figura 60. Mejores resultados de F1-score y certeza para el set de datos PG con optimización reducida de hiperparámetros

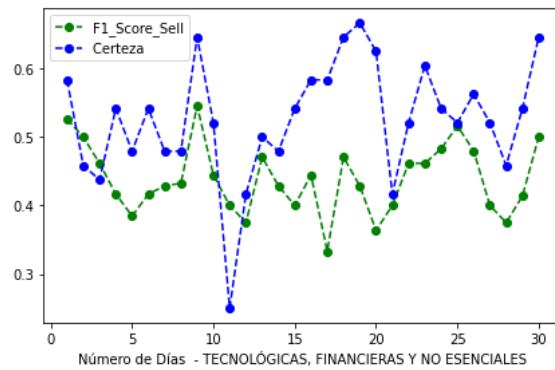


Figura 61. Mejores resultados de F1-score y certeza para el set de datos TECH con optimización reducida de hiperparámetros

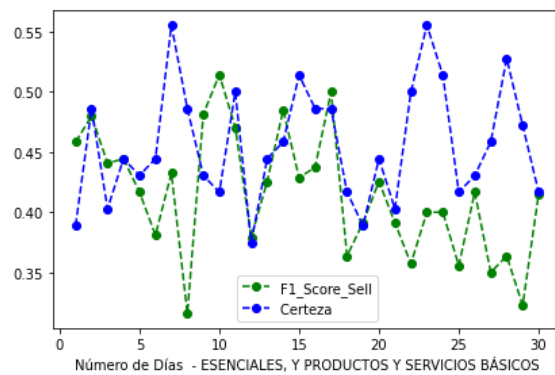


Figura 62. Mejores resultados de F1-score y certeza para el set de datos NONTECH con optimización reducida de hiperparámetros

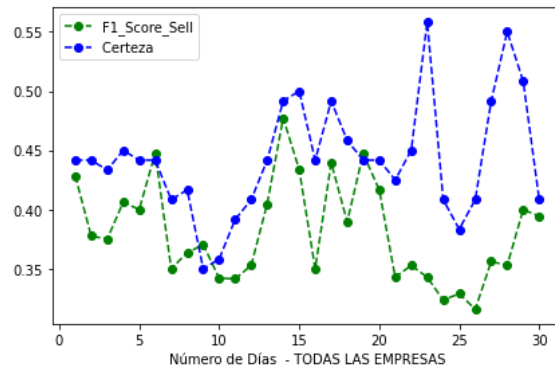


Figura 63. Mejores resultados de F1-score y certeza para el set de datos ALL con optimización reducida de hiperparámetros

Adicionalmente, se recopiló la información más importante de cada una de estas implementaciones, como la correspondencia a cada tipo de data set y modelo de aprendizaje de máquina que haya conseguido los mejores resultados. Las tablas 27 a 30 muestran la recopilación de esta información. En estas tablas se resaltan los periodos optimizados que generaron resultados promedios entre F1-score SELL y certeza mayores o iguales a 0.7, los cuales serán analizados a partir de la figura 66.

Día	Mejor modelo de ML	DATASET	F1_SCORE_ SELL	EXACTITUD	PROMEDIO	Día	Mejor modelo de ML	DATASET	F1_SCORE_ SELL	EXACTITUD	PROMEDIO
1	SVC	HISTORICO2	0.8	0.5	0.65		LOGISTIC				
2	RANDOM FOREST	NORMAL	0.67	0.46	0.565	1	REGRESSION DECISION	NORMAL	0.55	0.38	0.465
3	KNN	NORMAL	0.8	0.63	0.715	2	TREE	NORMAL	0.5	0.54	0.52
4	RANDOM FOREST	NORMAL	0.8	0.75	0.775	3	SVC	HISTORICO2	0.5	0.5	0.5
5	SVC	NORMAL	0.67	0.58	0.625	4	DECISION TREE	NORMAL	0.57	0.63	0.6
6	KNN	NORMAL	0.73	0.63	0.68	5	DECISION TREE	NORMAL	0.47	0.42	0.445
7	KNN	NORMAL	0.67	0.54	0.605	6	DECISION TREE	NORMAL	0.6	0.54	0.57
8	DECISION TREE	NORMAL	0.67	0.58	0.625	7	RANDOM FOREST	NORMAL	0.67	0.63	0.65
9	DECISION TREE	NORMAL	0.5	0.5	0.5	8	DECISION TREE	NORMAL	0.5	0.5	0.5
10	SVC	HISTORICO2	0.83	0.5	0.665	9	KNN	NORMAL	0.44	0.58	0.51
11	SVC	HISTORICO2	0.75	0.63	0.69	10	RANDOM FOREST	HISTORICO1	0.5	0.5	0.5
12	SVC	HISTORICO2	0.71	0.54	0.625	11	RANDOM FOREST	HISTORICO1	0.4	0.33	0.365
13	SVC	HISTORICO1	0.61	0.38	0.495	12	LIGHTGBM	HISTORICO2	0.4	0.54	0.47
14	LOGISTIC REGRESSION	NORMAL	0.45	0.29	0.37	13	RANDOM FOREST	HISTORICO2	0.57	0.58	0.575
15	LOGISTIC REGRESSION	NORMAL	0.45	0.29	0.37	14	RANDOM FOREST	HISTORICO1	0.5	0.67	0.585
16	RANDOM FOREST	HISTORICO3	0.46	0.46	0.46	15	SVC	HISTORICO3	0.67	0.46	0.565
17	SVC	HISTORICO3	0.46	0.58	0.52	16	DECISION TREE	NORMAL	0.33	0.5	0.415
18	KNN	NORMAL	0.57	0.58	0.575	17	SVC	HISTORICO3	0.67	0.63	0.65
19	LOGISTIC REGRESSION	NORMAL	0.45	0.29	0.37	18	KNN	NORMAL	0.67	0.71	0.69
20	DECISION TREE	NORMAL	0.62	0.5	0.56	19	KNN	NORMAL	0.67	0.71	0.69
21	DECISION TREE	HISTORICO2	0.44	0.46	0.45	20	KNN	NORMAL	0.67	0.5	0.585
22	DECISION TREE	HISTORICO2	0.5	0.46	0.48	21	KNN	NORMAL	0.5	0.38	0.44
23	LOGISTIC REGRESSION	NORMAL	0.4	0.25	0.325	22	SVC	HISTORICO3	0.44	0.5	0.47
24	SVC	HISTORICO2	0.5	0.58	0.54	23	DECISION TREE	HISTORICO3	0.75	0.75	0.75
25	LOGISTIC REGRESSION	NORMAL	0.5	0.33	0.415	24	KNN	NORMAL	0.57	0.25	0.41
26	DECISION TREE	NORMAL	0.83	0.54	0.685	25	RANDOM FOREST	HISTORICO3	0.4	0.58	0.49
27	LOGISTIC REGRESSION	NORMAL	0.55	0.38	0.465	26	RANDOM FOREST	HISTORICO3	0.5	0.75	0.625
28	LOGISTIC REGRESSION	NORMAL	0.55	0.38	0.465	27	SVC	HISTORICO3	0.5	0.54	0.52
29	LOGISTIC REGRESSION	NORMAL	0.55	0.38	0.465	28	DECISION TREE	NORMAL	0.5	0.58	0.54
30	DECISION TREE	NORMAL	0.57	0.54	0.555	29	SVC	HISTORICO3	0.67	0.67	0.67
						30	RANDOM FOREST	HISTORICO3	0.5	0.67	0.585

Tabla 27. Recopilación de mejores métodos de aprendizaje de máquina y tipos de datasets para el dataset de IBM (izquierda) y AMEX (derecha)

Día	Mejor modelo de ML	DATASET	F1_SCORE_SELL	EXACTITUD	PROMEDIO	Día	Mejor modelo de ML	DATASET	F1_SCORE_SELL	EXACTITUD	PROMEDIO
1	SVC	HISTORICO3	0.74	0.5	0.62	1	KNN	NORMAL	0.7	0.63	0.665
2	RANDOM FOREST	HISTORICO1	0.57	0.46	0.515	2	DECISION TREE	HISTORICO1	0.77	0.5	0.635
3	SVC	HISTORICO3	0.5	0.42	0.46	3	RANDOM FOREST	HISTORICO2	0.67	0.63	0.65
4	DECISION TREE	NORMAL	0.56	0.42	0.49	4	RANDOM FOREST	HISTORICO2	0.8	0.42	0.61
5	DECISION TREE	NORMAL	0.56	0.33	0.445	5	RANDOM FOREST	HISTORICO2	0.75	0.58	0.665
6	DECISION TREE	NORMAL	0.57	0.38	0.475	6	RANDOM FOREST	NORMAL	0.6	0.63	0.615
7	DECISION TREE	NORMAL	0.52	0.38	0.45	7	DECISION TREE	NORMAL	0.55	0.33	0.44
8	DECISION TREE	NORMAL	0.72	0.54	0.63	8	KNN	NORMAL	0.8	0.5	0.65
9	LOGISTIC REGRESSION	NORMAL	0.55	0.38	0.465	9	KNN	NORMAL	0.8	0.67	0.735
10	LOGISTIC REGRESSION	NORMAL	0.55	0.38	0.465	10	RANDOM FOREST	HISTORICO3	0.8	0.67	0.735
11	LOGISTIC REGRESSION	NORMAL	0.5	0.33	0.415	11	KNN	NORMAL	0.8	0.67	0.735
12	LOGISTIC REGRESSION	NORMAL	0.5	0.33	0.415	12	RANDOM FOREST	HISTORICO3	0.67	0.67	0.67
13	LOGISTIC REGRESSION	NORMAL	0.5	0.33	0.415	13	SVC	HISTORICO2	0.8	0.67	0.735
14	KNN	NORMAL	0.57	0.58	0.575	14	SVC	HISTORICO2	0.67	0.67	0.67
15	KNN	NORMAL	0.59	0.67	0.63	15	SVC	HISTORICO2	0.73	0.58	0.655
16	KNN	NORMAL	0.55	0.54	0.545	16	KNN	NORMAL	0.75	0.67	0.71
17	KNN	NORMAL	0.5	0.63	0.565	17	RANDOM FOREST	HISTORICO3	0.91	0.79	0.85
18	LOGISTIC REGRESSION	NORMAL	0.45	0.29	0.37	18	RANDOM FOREST	HISTORICO2	0.67	0.67	0.67
19	DECISION TREE	NORMAL	0.57	0.58	0.575	19	RANDOM FOREST	NORMAL	0.67	0.63	0.65
20	KNN	NORMAL	0.63	0.58	0.605	20	SVC	HISTORICO2	0.67	0.67	0.67
21	KNN	NORMAL	0.53	0.67	0.6	21	RANDOM FOREST	HISTORICO3	1	0.67	0.835
22	KNN	NORMAL	0.53	0.58	0.555	22	RANDOM FOREST	HISTORICO3	0.75	0.71	0.73
23	KNN	NORMAL	0.57	0.63	0.6	23	SVC	HISTORICO2	0.67	0.67	0.67
24	LOGISTIC REGRESSION	NORMAL	0.55	0.38	0.465	24	RANDOM FOREST	HISTORICO2	0.8	0.71	0.755
25	DECISION TREE	HISTORICO3	0.56	0.54	0.55	25	RANDOM FOREST	NORMAL	0.67	0.75	0.71
26	KNN	NORMAL	0.57	0.41	0.49	26	SVC	NORMAL	0.6	0.58	0.59
27	LOGISTIC REGRESSION	NORMAL	0.55	0.38	0.465	27	DECISION TREE	NORMAL	0.57	0.63	0.6
28	LOGISTIC REGRESSION	NORMAL	0.59	0.42	0.505	28	RANDOM FOREST	HISTORICO2	0.75	0.71	0.73
29	LOGISTIC REGRESSION	NORMAL	0.5	0.33	0.415	29	RANDOM FOREST	HISTORICO2	0.67	0.71	0.69
30	LOGISTIC REGRESSION	NORMAL	0.5	0.33	0.415	30	RANDOM FOREST	NORMAL	0.6	0.63	0.615

Tabla 28. Recopilación de mejores métodos de aprendizaje de máquina y tipos de datasets para el dataset de 3M (izquierda) y MERCK (derecha)

Día	Mejor modelo de ML	DATASET	F1_SCORE_SELL	EXACTITUD	PROMEDIO	Día	Mejor modelo de ML	DATASET	F1_SCORE_SELL	EXACTITUD	PROMEDIO
1	RANDOM FOREST	HISTORICO1	0.78	0.58	0.68	1	RANDOM FOREST	NORMAL	0.52	0.58	0.55
2	DECISION TREE	HISTORICO1	0.73	0.5	0.615	2	DECISION TREE	NORMAL	0.5	0.45	0.475
3	DECISION TREE	HISTORICO2	0.83	0.54	0.685	3	DECISION TREE	HISTORICO2	0.46	0.44	0.45
4	RANDOM FOREST	HISTORICO2	0.86	0.63	0.745	4	DECISION TREE	NORMAL	0.42	0.54	0.48
5	RANDOM FOREST	HISTORICO3	0.8	0.58	0.69	5	DECISION TREE	NORMAL	0.38	0.48	0.43
6	RANDOM FOREST	HISTORICO3	0.82	0.54	0.68	6	DECISION TREE	NORMAL	0.42	0.54	0.48
7	DECISION TREE	HISTORICO2	0.82	0.38	0.6	7	DECISION TREE	NORMAL	0.43	0.48	0.455
8	RANDOM FOREST	HISTORICO2	0.92	0.58	0.75	8	DECISION TREE	NORMAL	0.43	0.48	0.455
9	RANDOM FOREST	HISTORICO2	0.8	0.67	0.735	9	DECISION TREE	HISTORICO1	0.55	0.65	0.6
10	DECISION TREE	HISTORICO2	0.77	0.54	0.655	10	KNN	NORMAL	0.45	0.52	0.485
11	RANDOM FOREST	HISTORICO1	0.89	0.5	0.695	11	LOGISTIC REGRESSION	NORMAL	0.4	0.25	0.325
12	KNN	NORMAL	0.71	0.67	0.69	12	DECISION TREE	NORMAL	0.38	0.42	0.4
13	KNN	NORMAL	0.67	0.63	0.65	13	DECISION TREE	HISTORICO3	0.47	0.5	0.485
14	SVC	HISTORICO3	0.75	0.5	0.625	14	DECISION TREE	NORMAL	0.43	0.48	0.455
15	DECISION TREE	HISTORICO2	0.86	0.75	0.805	15	LIGHTGBM	HISTORICO1	0.4	0.54	0.47
16	RANDOM FOREST	HISTORICO2	0.75	0.71	0.73	16	DECISION TREE	HISTORICO1	0.45	0.58	0.515
17	DECISION TREE	NORMAL	0.63	0.5	0.565	17	RANDOM FOREST	HISTORICO2	0.33	0.58	0.455
18	RANDOM FOREST	NORMAL	0.73	0.67	0.7	18	RANDOM FOREST	HISTORICO2	0.47	0.65	0.56
19	SVC	NORMAL	0.73	0.67	0.7	19	SVC	HISTORICO2	0.43	0.67	0.55
20	DECISION TREE	HISTORICO2	0.75	0.75	0.75	20	SVC	HISTORICO3	0.36	0.63	0.495
21	DECISION TREE	HISTORICO1	0.75	0.58	0.665	21	DECISION TREE	NORMAL	0.4	0.42	0.41
22	RANDOM FOREST	HISTORICO1	0.89	0.63	0.76	22	DECISION TREE	NORMAL	0.46	0.52	0.49
23	RANDOM FOREST	NORMAL	0.75	0.63	0.69	23	DECISION TREE	NORMAL	0.46	0.61	0.535
24	DECISION TREE	HISTORICO1	0.8	0.54	0.67	24	DECISION TREE	NORMAL	0.48	0.54	0.51
25	RANDOM FOREST	HISTORICO1	0.67	0.63	0.65	25	DECISION TREE	NORMAL	0.52	0.52	0.52
26	DECISION TREE	NORMAL	0.75	0.67	0.71	26	DECISION TREE	NORMAL	0.48	0.56	0.52
27	DECISION TREE	NORMAL	0.6	0.54	0.57	27	DECISION TREE	NORMAL	0.4	0.52	0.46
28	DECISION TREE	NORMAL	0.75	0.71	0.73	28	DECISION TREE	NORMAL	0.38	0.46	0.42
29	SVC	HISTORICO1	0.57	0.54	0.555	29	DECISION TREE	NORMAL	0.41	0.54	0.475
30	SVC	HISTORICO2	0.57	0.63	0.6	30	RANDOM FOREST	NORMAL	0.5	0.65	0.575

Tabla 29. Recopilación de mejores métodos de aprendizaje de máquina y tipos de datasets para el dataset de PG (izquierda) y TECH (derecha)

Día	Mejor modelo de ML	DATASET	F1_SCORE_SELL	EXACTITUD	PROMEDIO	Día	Mejor modelo de ML	DATASET	F1_SCORE_SELL	EXACTITUD	PROMEDIO
1	DECISION TREE	NORMAL	0.46	0.39	0.425	1	DECISION TREE	NORMAL	0.43	0.44	0.435
2	RANDOM FOREST	HISTORICO2	0.48	0.49	0.485	2	DECISION TREE	NORMAL	0.38	0.44	0.41
3	DECISION TREE	NORMAL	0.44	0.4	0.42	3	DECISION TREE	NORMAL	0.38	0.43	0.405
4	KNN	NORMAL	0.45	0.45	0.45	4	KNN	NORMAL	0.41	0.45	0.43
5	DECISION TREE	NORMAL	0.42	0.43	0.425	5	DECISION TREE	NORMAL	0.4	0.44	0.42
6	DECISION TREE	NORMAL	0.38	0.45	0.415	6	KNN	NORMAL	0.45	0.44	0.445
7	DECISION TREE	HISTORICO2	0.43	0.56	0.495	7	DECISION TREE	NORMAL	0.35	0.41	0.38
8	DECISION TREE	NORMAL	0.32	0.49	0.405	8	DECISION TREE	NORMAL	0.36	0.42	0.39
9	DECISION TREE	NORMAL	0.48	0.43	0.455	9	DECISION TREE	NORMAL	0.37	0.35	0.36
10	SVC	HISTORICO1	0.51	0.42	0.465	10	KNN	NORMAL	0.34	0.36	0.35
11	KNN	NORMAL	0.47	0.5	0.485	11	DECISION TREE	NORMAL	0.34	0.39	0.365
12	DECISION TREE	NORMAL	0.38	0.38	0.38	12	DECISION TREE	NORMAL	0.35	0.41	0.38
13	DECISION TREE	NORMAL	0.43	0.45	0.44	13	DECISION TREE	NORMAL	0.4	0.44	0.42
14	SVC	HISTORICO1	0.49	0.46	0.475	14	DECISION TREE	NORMAL	0.47	0.49	0.48
15	SVC	HISTORICO1	0.43	0.51	0.47	15	DECISION TREE	NORMAL	0.44	0.5	0.47
16	KNN	NORMAL	0.44	0.49	0.465	16	DECISION TREE	NORMAL	0.35	0.44	0.395
17	SVC	HISTORICO1	0.5	0.49	0.495	17	DECISION TREE	NORMAL	0.44	0.49	0.465
18	KNN	NORMAL	0.36	0.42	0.39	18	DECISION TREE	NORMAL	0.39	0.46	0.425
19	KNN	NORMAL	0.39	0.39	0.39	19	DECISION TREE	NORMAL	0.45	0.44	0.445
20	KNN	NORMAL	0.43	0.45	0.44	20	DECISION TREE	NORMAL	0.42	0.44	0.43
21	KNN	NORMAL	0.39	0.4	0.395	21	DECISION TREE	NORMAL	0.34	0.43	0.385
22	SVC	HISTORICO1	0.36	0.5	0.43	22	DECISION TREE	NORMAL	0.35	0.45	0.4
23	SVC	HISTORICO1	0.4	0.56	0.48	23	KNN	NORMAL	0.34	0.56	0.45
24	DECISION TREE	NORMAL	0.4	0.51	0.455	24	KNN	NORMAL	0.32	0.41	0.365
25	KNN	NORMAL	0.36	0.42	0.39	25	KNN	NORMAL	0.33	0.38	0.355
26	KNN	NORMAL	0.42	0.43	0.425	26	DECISION TREE	NORMAL	0.32	0.41	0.365
27	KNN	NORMAL	0.35	0.46	0.405	27	KNN	NORMAL	0.36	0.49	0.425
28	SVC	HISTORICO1	0.36	0.53	0.445	28	DECISION TREE	NORMAL	0.35	0.55	0.45
29	DECISION TREE	NORMAL	0.32	0.47	0.395	29	DECISION TREE	NORMAL	0.4	0.51	0.455
30	DECISION TREE	NORMAL	0.42	0.42	0.42	30	KNN	NORMAL	0.39	0.41	0.4

Tabla 30. Recopilación de mejores métodos de aprendizaje de máquina y tipos de datasets para el dataset de NONTECH (izquierda) y ALL (derecha)

Se resaltaron especialmente aquellos resultados que entre las métricas F1-score SELL y certeza tuvieran un valor medio de 0.7 o superior, esto debido a que para el alcance de este trabajo se busca optimizar de forma completa los resultados de un set de datos, especialmente los 3 mejores periodos de variación porcentual. Se puede que los dos sets de datos con el mayor número modelos con esta característica corresponden a MERCK, y PG. Sin embargo, el set de datos MERCK contiene un mayor número de estos, además del único modelo que logra una métrica, en este caso F1-score SELL, en 1.0 al validar con los datos de prueba. Por esta razón, los modelos resaltados en verde fueron usados para una optimización completa

de datos. Las características de este modelo se pueden ver en la tabla 28, la imagen de la derecha. Por otra parte, es de notar que el set de datos histórico3, el cual contiene datos de sorpresa de beneficios por acción de 4 trimestres (1 año), es el que mejor resultados consigue. Además, los datos con mejores resultados corresponden a KNN y Random Forest, siendo este último el que produce los 3 modelos seleccionados. Sin embargo, la optimización completa se realizará con estos dos modelos para poder comparar resultados.

En la figura 64 se puede ver una gráfica con la distribución porcentual de la cantidad de veces que se encontró que cada método de aprendizaje de máquina fue el mejor. En la figura 65, de forma similar, se encuentra esta gráfica para el tipo de dataset que se encontraron que obtuvieron mejores resultados.

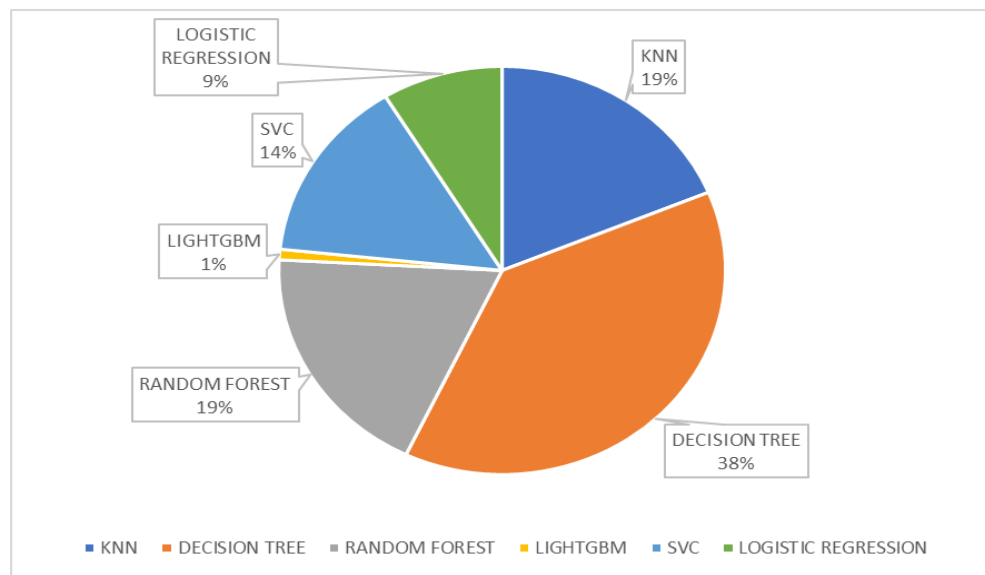


Figura 64. Cantidad porcentual de veces que cada modelo de aprendizaje de máquina generó mejores resultados

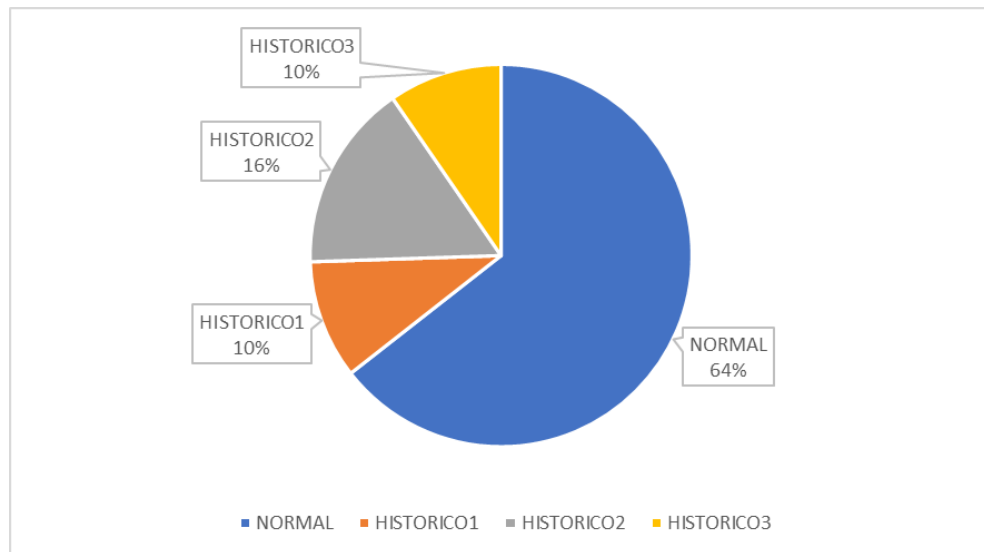


Figura 65. Cantidad porcentual de veces que cada tipo de dataset generó mejores resultados

Se puede observar que, aunque el modelo árbol de decisión no generó los mejores resultados máximos entre los distintos datasets, si fue recurrentemente el modelo que generaba mejores resultados. Además, se observa que igualmente, en proporción, el tomar datos históricos no generaba mejores resultados para la mayoría de los modelos. Sin embargo, al observar los datos resaltados en las tablas 27 a 30, es posible notar que una gran proporción de los modelos con mejores resultados aún entre los datasets, tenían algún tipo de consideración de parámetros históricos. En las gráficas 65 y 66 se puede ver la distribución en métodos de aprendizaje de máquina y tipos de sets de datos que generaron resultados promedios entre F1-score SELL y certezaes mayores a 0.7.

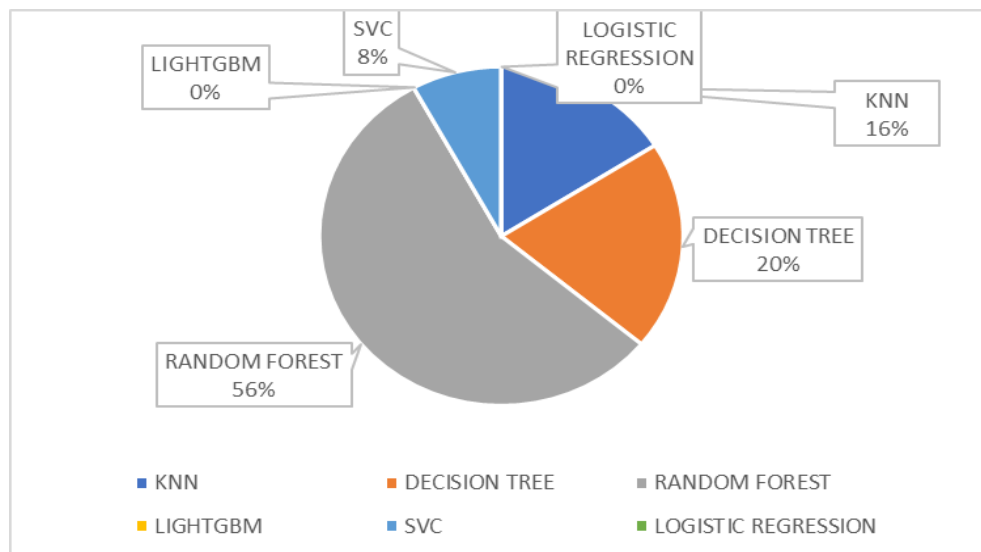


Figura 66. Cantidad porcentual de veces que cada modelo de aprendizaje de máquina generó resultados promedios entre F1-score SELL y certeza mayores o iguales a 0.7

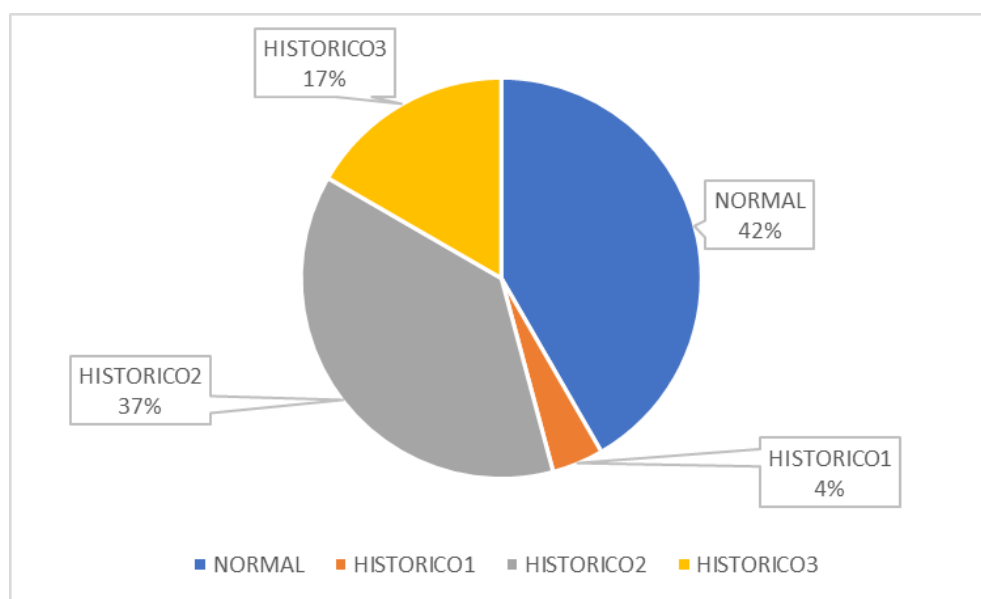


Figura 67. Cantidad porcentual de veces que cada tipo de dataset generó resultados promedios entre F1-score SELL y certeza mayores o iguales a 0.7

Además de esto, se compararon los resultados en términos de los datasets que los generaron. Se pudo ver, que para resultados conjuntos de F1-score SELL y certeza, los datasets Merck y P and G generaron un mayor número de resultados con 0.7 en promedio entre las dos métricas o más, esto se puede ver en la figura 68. Igualmente, se observa en la figura 69, que

Merck logra generar el doble de modelos, respecto a P and G, con resultados promedio conjuntos mayores a 0.8. Siendo esta la principal razón por la que únicamente se usó Merck para la optimización extensiva de hiperparámetros.

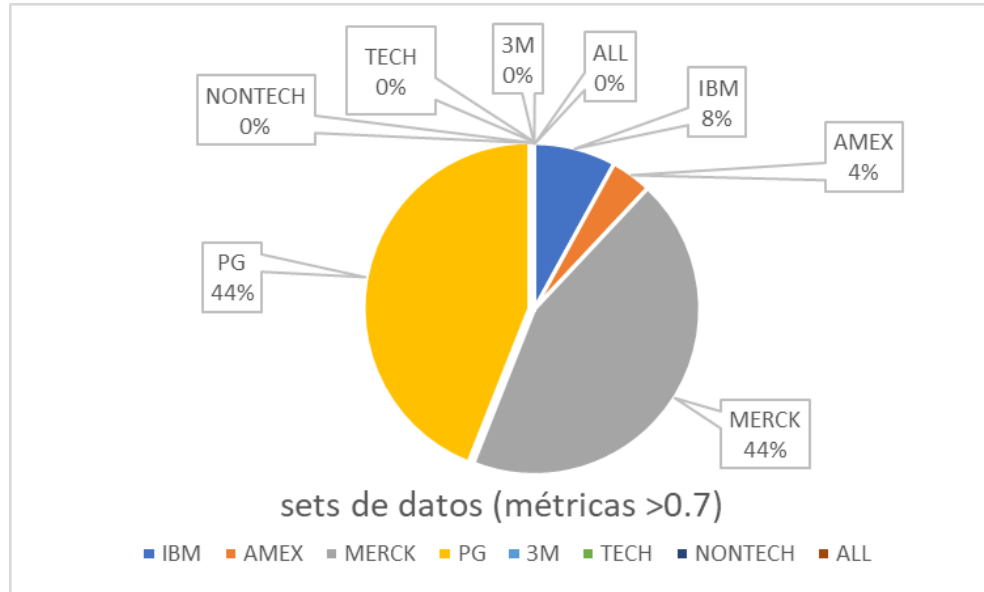


Figura 68. Cantidad porcentual de veces que cada dataset generó resultados promedios entre F1-score SELL y certeza mayores o iguales a 0.7

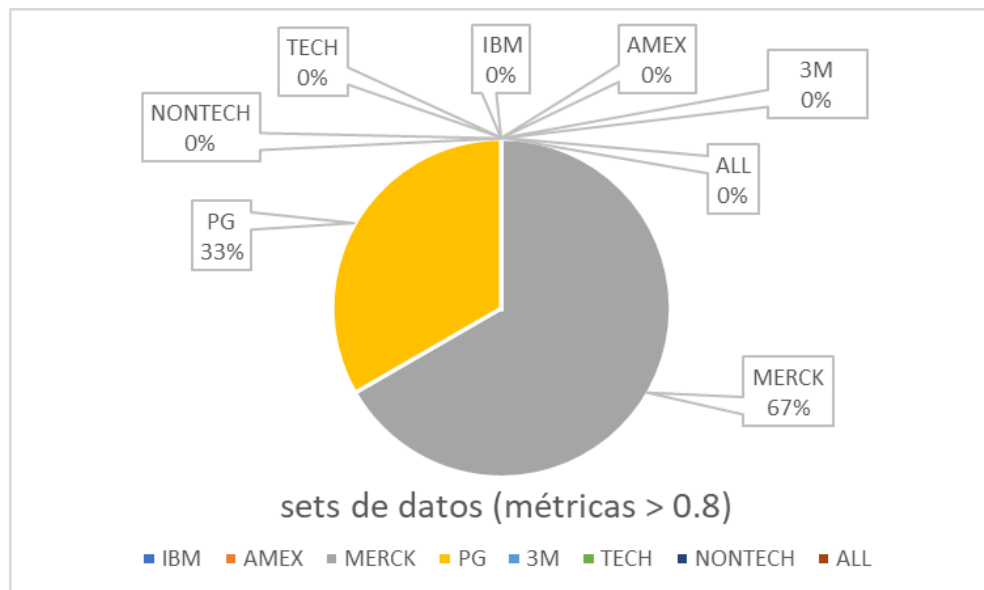


Figura 69. Cantidad porcentual de veces que cada dataset generó resultados promedios entre F1-score SELL y certeza mayores o iguales a 0.8

Adicionalmente, en la tabla 31 se muestran los tiempos de ejecución para la optimización de cada uno de los sets de datos.

SET DE DATOS	Tiempo de ejecución/ modelo	Tiempo de ejecución total
IBM	3,15 minutos	1,56 horas
AMEX	2,31 minutos	1,16 horas
3M	2,31 minutos	1,16 horas
MERCK	16,61 minutos	8,3 horas
PG	24,39 minutos	12,2 horas
TECH	7,37 minutos	3,69 horas
NONTECH	5,14 minutos	2,57 horas
ALL	4,99 minutos	2,50 horas

Tabla 31. Tiempos de ejecución de cada set de datos para una optimización general de hiperparámetros y considerando todos los tipos de sets de datos y métodos de aprendizaje de máquina

8.5 OPTIMIZACIÓN EXTENSIVA DE HIPERPARÁMETROS

Teniendo en cuenta los resultados obtenidos, se implementaron los métodos de aprendizaje de máquina Random Forest y KNN para una optimización extensiva de parámetros, esperando obtener mejores resultados en cada uno de los modelos. Sin embargo, este proceso tiene una duración en tiempo considerable, por lo que únicamente se seleccionaron tres periodos para generar el entrenamiento de datos y la predicción. Los periodos seleccionados se pueden ver en la tabla 32. Además, se puede ver los métodos de aprendizaje de máquina a implementar, y el tipo de set de datos a usar, teniendo en cuenta los resultados de la tabla 28.

Set de datos	Días de variación porcentual seleccionados	Tipo de set de datos	Métodos de aprendizaje de máquina
MERCK	10 días, 17 días, 21 días	Historico 3	KNN, Random Forest

Tabla 32. Periodos de variación porcentual seleccionados para la optimización extensiva de hiperparámetros

Además, se seleccionó para este proceso únicamente el set de datos de MERCK, debido a que produjo los resultados más altos tanto en F1-score SELL como en certeza en la optimización general de hiperparámetros. Al revisar la tabla 28, se nota que el tipo de set de datos histórico 3 fue el que produjo los 3 modelos seleccionados para esta optimización, además de haberse producido al usar el método Random Forest. Sin embargo, al analizar los mejores resultados del conjunto MERCK, se puede ver en la figura 70, que KNN es el segundo que produce mejores resultados en cantidad.

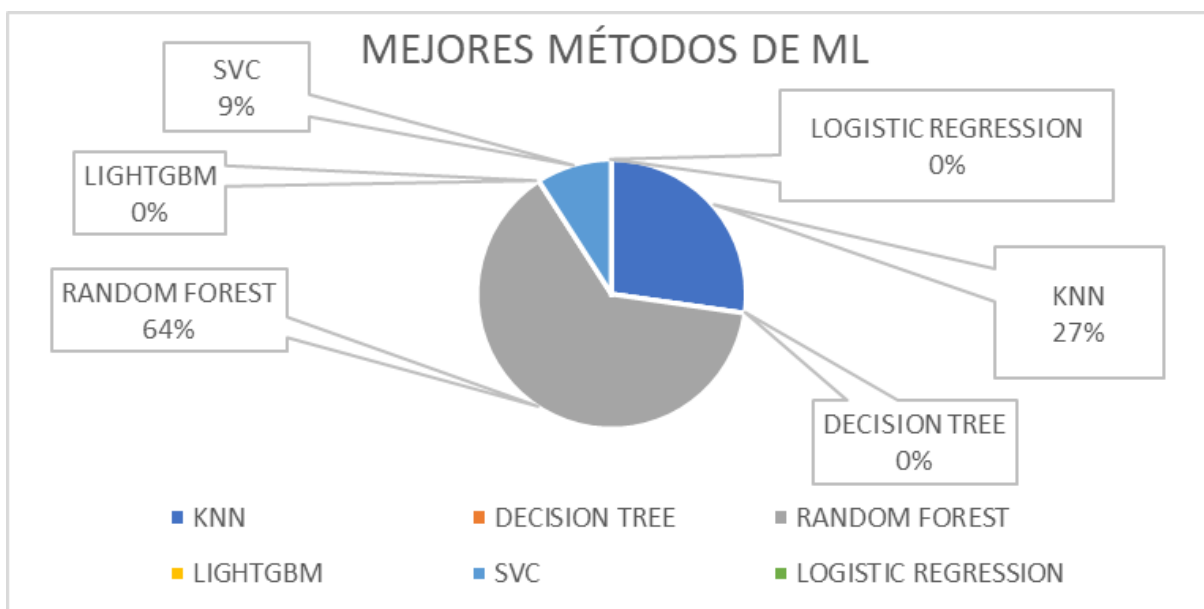


Figura 70. Cantidad porcentual de veces que cada modelo de aprendizaje de máquina generó resultados promedios entre F1-score SELL y certeza mayores o iguales a 0.7 en el set de datos MERCK

La optimización extensiva se realizó usando los rangos de hiperparámetros que se pueden ver en la tabla 33.

Hiperparámetros	KNN	Rango	Random Forest	Rango
1	n_neighbors	1 a 50	bootstrap	['True','False']
2	weights	['uniform','distance']	max_depth	1 a 10
3	algorithm	['auto','ball_tree','kd_tree','brute']	max_features	['sqrt','log2']
4	leaf_size	1 a 50	min_samples_leaf	1 a 10
5	p	1 a 2	min_samples_split	2 a 10
6	metric	['euclidean','manhattan','minkowski']	n_estimators	1 a 10
7			max_leaf_nodes	2 a 10

Tabla 33. Rangos de hiperparámetros usados para optimización extensiva

Usando estos rangos, se implementaron los métodos KNN y Random Forest. En la figura 71 se pueden observar los resultados de F1-score y de certeza para el modelo de KNN. Por otra parte, en la figura 72 se pueden observar los resultados de ambas métricas para el modelo Random Forest.

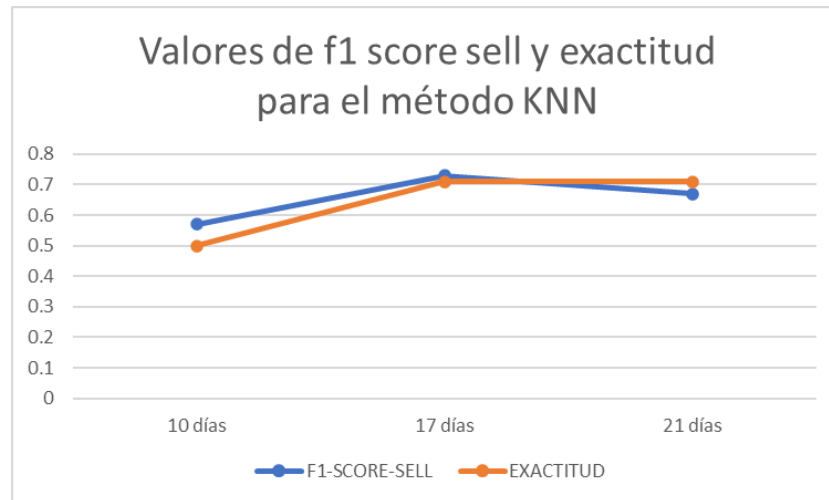


Figura 71. Resultados de F1-score SELL y Certeza después de optimización extensiva de hiperparámetros usando el modelo KNN

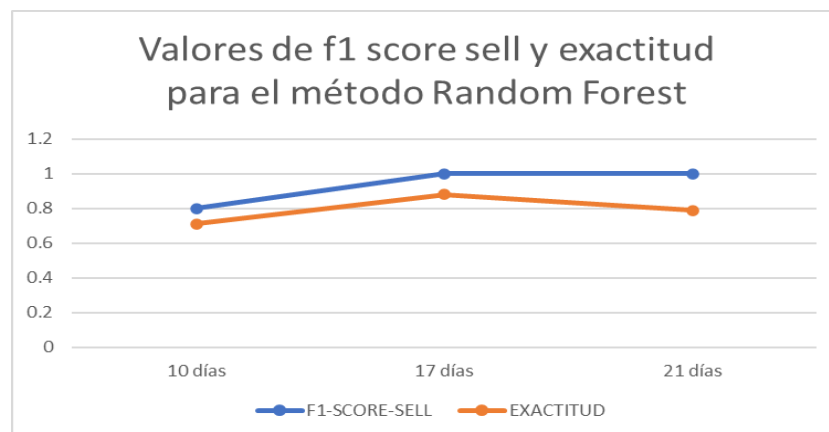


Figura 72. Resultados de F1-score SELL y Certeza después de optimización extensiva de hiperparámetros usando el modelo Random Forest

De estos resultados podemos notar cómo el método Random Forest presenta resultados significativamente mejores que los del método KNN. Principalmente, el método KNN tiene problemas con la métrica F1-score de la clase SELL que es de suma importancia para la protección de inversionistas “principiantes”. Se hizo uso de la matriz de confusión para poder

visualizar de mejor manera los valores predichos de cada modelo, y poder así analizar cuáles son las fallas o aciertos más significativos de cada modelo. En las figuras 73 a 75 se pueden ver las matrices de confusión de cada modelo trabajado del método KNN.

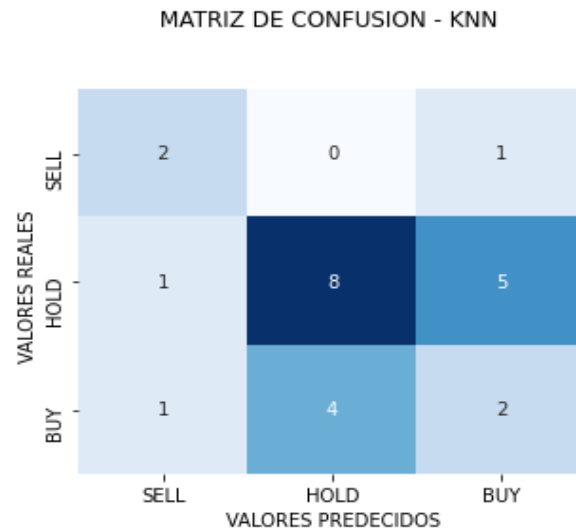


Figura 73. Matriz de confusión para el modelo KNN con optimización extensiva de hiperparámetros, para el periodo de 10 días de variación porcentual

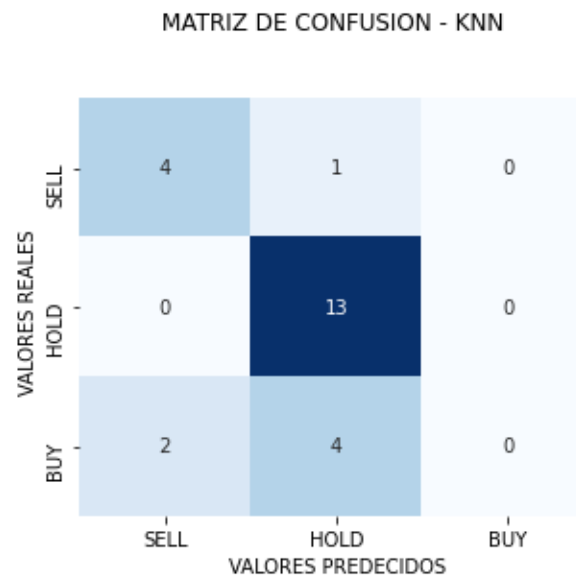


Figura 74. Matriz de confusión para el modelo KNN con optimización extensiva de hiperparámetros, para el periodo de 17 días de variación porcentual

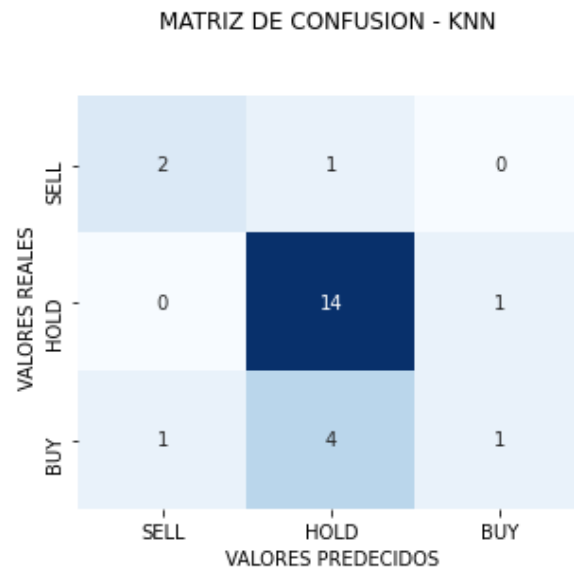


Figura 75. Matriz de confusión para el modelo KNN con optimización extensiva de hiperparámetros, para el periodo de 21 días de variación porcentual

En estas figuras, se puede observar cómo los modelos que usan KNN para este caso tienen valores no tan buenos. En el caso de F1-score de la clase SELL, que es una de las métricas más significativas, obtiene resultados de 0.57, 0.73, y 0.67 respectivamente, lo que es muy inferior a la cota de 0.7 definida anteriormente. Además, inclusive los resultados de F1-score de la clase HOLD son considerablemente superiores, siendo de 0.62, 0.84, y 0.82 respectivamente, lo cual a términos prácticos el modelo predice en una gran parte de los casos la etiqueta HOLD, lo cual significaría que el inversor o persona que se guíe por este tipo de consejo no debería hacer nada en la mayoría de los casos, siendo esto indeseable. En la tabla 34 se ven los valores de predicción, recall, y F1-score de la clase SELL de cada uno de los modelos.

10 DÍAS			
	PRECISIÓN	RECALL	F1-SCORE
SELL	0.5	0.67	0.57
HOLD	0.67	0.57	0.62
BUY	0.25	0.29	0.27
17 DÍAS			
SELL	0.67	0.8	0.73
HOLD	0.72	1	0.84
BUY	0	0	0
21 DÍAS			
SELL	0.67	0.67	0.67
HOLD	0.74	0.93	0.82
BUY	0.5	0.17	0.25

Tabla 34. Resultados de los modelos de KNN después de optimización extensiva de hiperparámetros.

En las figuras 76 a 78 se pueden ver las matrices de confusión de cada modelo trabajado del método KNN.

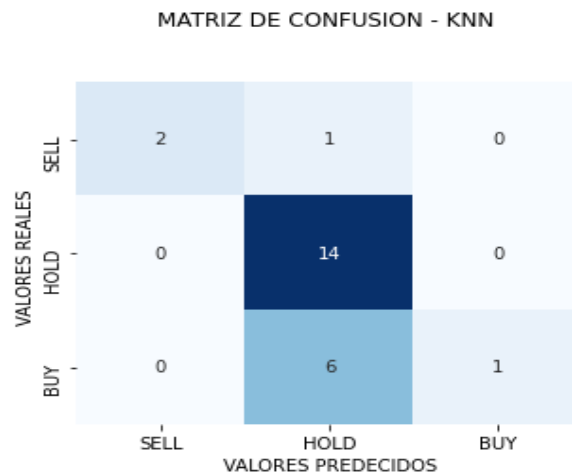


Figura 76. Matriz de confusión para el modelo Random Forest con optimización extensiva de hiperparámetros, para el periodo de 10 días de variación porcentual

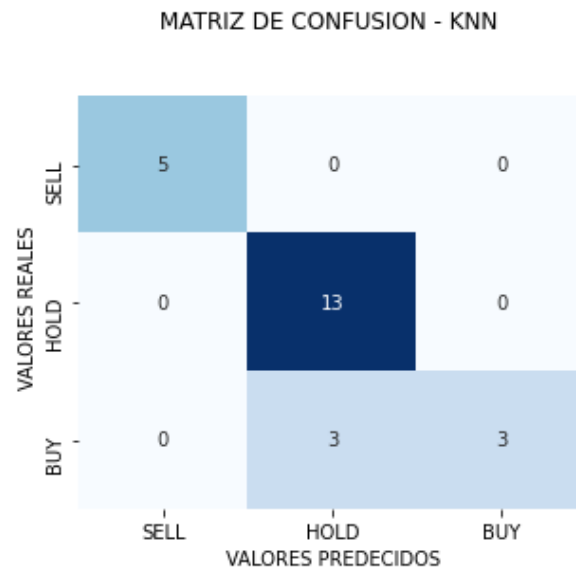


Figura 77. Matriz de confusión para el modelo Random Forest con optimización extensiva de hiperparámetros, para el periodo de 17 días de variación porcentual

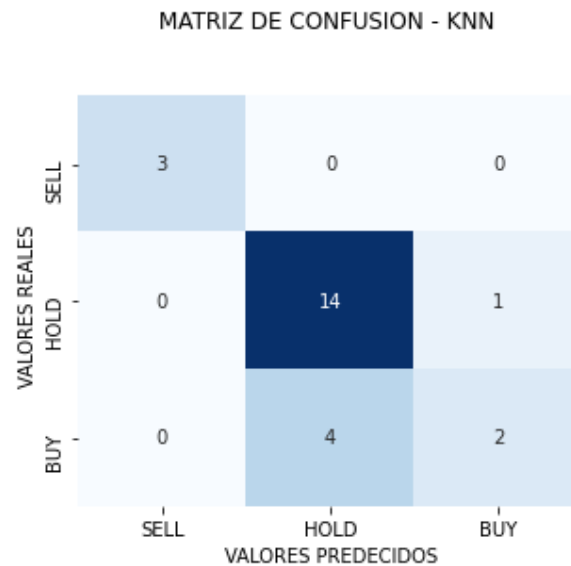


Figura 78. Matriz de confusión para el modelo Random Forest con optimización extensiva de hiperparámetros, para el periodo de 21 días de variación porcentual

En estas figuras, se puede observar cómo los modelos de Random Forest tienen una predicción casi perfecta de los valores, en dos casos generando un F1-score de la clase SELL de 1, y en los 3 casos generando una precisión de SELL de 1. Además, en los tres casos se

obtiene un F1-score de la clase HOLD mayor o igual a 0.8, lo cual es un resultado muy positivo. En este tipo de modelos, lamentablemente se consiguen resultados no tan positivos de BUY, ya que el F1-score de la clase BUY es de 0.25, 0.67, y 0.44 respectivamente. Sin embargo, es un modelo con resultados muy positivos en las métricas que definimos como importantes en este trabajo, F1-score de la clase SELL y la certeza del modelo. En la tabla 35 se ven los valores de predicción, recall, y F1-score de la clase SELL de cada uno de los modelos.

10 DÍAS			
	PRECISIÓN	RECALL	F1-SCORE
SELL	1	0.67	0.8
HOLD	0.67	1	0.8
BUY	1	0.14	0.25
17 DÍAS			
SELL	1	1	1
HOLD	0.81	1	0.9
BUY	1	0.5	0.67
21 DÍAS			
SELL	1	1	1
HOLD	0.78	0.93	0.85
BUY	0.67	0.33	0.44

Tabla 35. Resultados de los modelos de Random Forest después de optimización extensiva de hiperparámetros.

En la tabla 36 se pueden ver los tiempos de ejecución de estos modelos. Se resalta la alta duración de cada modelo Random Forest para realizar una optimización completa, pues dura aproximadamente 83 veces en realizar esta optimización en comparación al método KNN. Sin embargo, resulta muy positivo que los resultados obtenidos por los modelos de Random Forest fueron significativamente mejores que los del modelo KNN.

MÉTODO DE APRENDIZAJE DE MÁQUINA	Tiempo de ejecución/ modelo	Tiempo de ejecución total
KNN	4,98 minutos	14,94 minutos
RANDOM FOREST	6,93 horas	20,8 horas

Tabla 36. Tiempos de ejecución de optimización extensiva de hiperparámetros para cada método de aprendizaje de máquina

Finalmente, en la tabla 37 se pueden ver los hiperparámetros que generaron los mejores resultados tanto para los modelos de KNN, así como para los de Random Forest.

Hiperparámetros	KNN	Periodo(10 - 17 - 21 días)	Random Forest	Periodo(10 - 17 - 21 días)
1	n_neighbors	[8,28,9]	bootstrap	['False', 'False','False']
2	weights	['distance','uniform','distance']	max_depth	[4,4,5]
3	algorithm	['auto','auto','auto']	max_features	['log2','log2','log2']
4	leaf_size	[24,1,24]	min_samples_leaf	[10,6,5]
5	p	[1,1,1]	min_samples_split	[8,4,9]
6	metric	['manhattan','euclidean','manhattan']	n_estimators	[10,10,10]
7			max_leaf_nodes	[9,7,7]

Tabla 37. Hiperparámetros de cada modelo de KNN y Random Forest que generaron los mejores resultados de F1-Score SELL y de certeza

9 CONCLUSIONES

Actualmente, se está dando en los diferentes mercados financieros, y bolsas de valores, un crecimiento sustancial en el número de inversionistas con poca experiencia y conocimiento. En este trabajo se implementaron y compararon modelos de aprendizaje de máquina en 8 sets de datos diferentes para la predicción, en un periodo de hasta 30 días, de los movimientos del precio de algunas acciones. Esto se realizó por medio de recomendaciones dadas en clases/etiquetas BUY/HOLD/SELL, que pueden servir como un instrumento de apoyo para decisiones de inversión durante periodos volátiles como lo son los periodos de reportes de beneficios y otras métricas financieras de las empresas seleccionadas. Así, para la implementación de este trabajo, se lograron una serie de objetivos bien definidos. Los siguientes son las conclusiones de cada uno de los objetivos planteados para la realización de este trabajo.

- a. Fueron elegidas 5 empresas del DJIA y fueron definidos 8 set de datos para la implementación de los métodos de aprendizaje de máquina, dichos sets se pueden ver en la sección 6.4. Se concluyó que, para este trabajo, que se puede considerar “inicial” (para una implementación futura más completa de software), era preferible decidirse por sets de datos que fueran fácilmente recopilables, obtenibles de fuentes primarias de entidades confiables como Zacks, y yahoo finance, y que, por tanto, sean acciones reconocibles, en este caso compañías pertenecientes al DJIA, que está compuesto de las 30 compañías más representativas del mercado estadounidense. De igual manera, se definieron dos variables para la construcción de los modelos de aprendizaje de máquina: sorpresa de beneficios por acción, y variación porcentual del precio de la acción, siendo la segunda redefinida por medio de límites del rango intercuartílico, en la tabla 8, a las clases BUY/HOLD/SELL. Se encontró que el set de datos correspondiente a la empresa MERCK, generó los mejores resultados comparativos, esto se puede ver en la figura 68 y 69. Además, se usó la variable sorpresa de beneficios por acción de 4 periodos consecutivos (en 4 columnas separadas) para generar los mejores resultados obtenidos en este trabajo.
- b. Los datos de las variables financieras son, en muchos casos, información protegida y comercializada para fines de lucro. Esta información, sin embargo, es publicada de forma pública desde 2001 de forma digital al sistema “EDGAR” por parte de la comisión de bolsa y valores de estados unidos. Esta información fue recopilada, desde el periodo de 1992, del portal de Zacks, el cual es un líder en proveer servicios financieros relacionados con la bolsa de valores. Igualmente, se recopiló la información diaria de precios de acciones por medio de yahoo finance. Además, se encuentra que no hay periodos de predicción claros ni trabajos similares dónde se aborda este problema de forma pública, por lo que se recopiló la información necesaria para construir modelos de predicción desde 1 día hasta 30 días, para en la

sección 8.5 poder elegir los modelos que generen mejores resultados. Se encontró, que los modelos de predicción de 10, 17, y 21 días generaron los mejores resultados en el set de datos de la compañía MERCK, esto se puede ver en la figura 29. Sin embargo, estos periodos son diferentes para cada set de datos.

- c. Algunos de los métodos de aprendizaje de máquina más usados para implementaciones con pocos datos, consiste en los métodos LightGBM, Random Forest, SVC, y Logistic regression[59]. Estos métodos fueron usados junto a los métodos KNN y Decision tree para las comparaciones de resultados de cada uno de los modelos, dependiendo de las métricas usadas para la optimización de estos métodos. Algunos de estos métodos, sin embargo, generaron problemas especialmente relacionados con el tiempo de ejecución, ya que, en los casos más extremos, como en el caso de la optimización extensiva de hiperparámetros de Random Forest, se obtuvieron tiempos de ejecución de hasta 20.8 horas, lo que dificultó grandemente la posible inclusión de otros sets de datos para la implementación de los métodos de aprendizaje de máquina.
- d. El preprocesamiento consistió principalmente en la redefinición de la variable porcentaje de variación del precio de acciones a las clases BUY/HOLD/SELL, esto se logró usando límites intercuartílicos relativos a cada set de datos, estos se pueden ver en la tabla 8. Además, no fue necesario para esta aplicación eliminar datos atípicos, ya que, al estar ligados a una clase BUY/SELL, no importaría la magnitud de la variación porcentual siempre y cuando se supere la cota definida por los rangos intercuartílicos de cada set de datos. Además de esto, los datos atípicos corresponden a datos verificados, por lo tanto, estos datos corresponden a comportamientos reales que se dieron.
- e. Se validaron los modelos de aprendizaje de máquina con la métrica F1-score de la clase SELL, puesto que esto permitiría a los inversores protegerse de pérdidas financieras en muchos casos irremediables, así como con la certeza del modelo, debido a que se quiere lograr un óptimo comportamiento del modelo a entrenar en general. Por tanto, se determinó que para esta causa es preferible priorizar la optimización de los modelos de aprendizajes de máquina por las métricas F1-score de la clase SELL y certeza. Así, se compararon los resultados y se encontró que los métodos más recurrentes que producen los mejores resultados son árbol de decisión, y Random Forest (figura 66). Sin embargo, para el set de datos MERCK, el cual obtuvo mejores resultados de forma comparativa respecto a otros sets de datos, fueron más recurrentes los métodos KNN y Random Forest (figura 70). Se encontró, igualmente, que el parámetro F1-score de la clase SELL puede llegar a ser 1 (tabla 35), lo cual, en una aplicación real, podría permitir que inversores con poca experiencia tengan cierta tranquilidad de posibles pérdidas durante periodos de

reportes financieros. Igualmente, se identificó que los peores resultados obtenidos son de la predicción de la clase BUY, siendo su mayor valor 0.5 aún después de la optimización extensiva de hiperparámetros, cuyos resultados se muestran nuevamente en la tabla 35.

10 DIFICULTADES

En el presente trabajo se implementaron y compararon diferentes métodos de aprendizaje de máquina para predecir las variaciones en el precio de las acciones de ciertas empresas del DJIA posterior a la publicación de sus reportes financieros trimestrales. Él trabajo incluyó implementaciones tanto con modelos de regresión como de clasificación y siguió un proceso sistemático que buscó generar modelos que mejoraron la certeza, así como la métrica F1-score de la clase SELL. Sin embargo, el trabajo presentó dificultades que se esperan superar en trabajos futuros. Los siguientes son puntos en los cuales se presentaron las mayores dificultades para la realización de este trabajo, y, que se esperan superar.

- a. Cada compañía incluida en los sets de datos pertenece y fue incluida en el DJIA al menos en el año 1982. Sin embargo, por las limitaciones de formas de guardar la información en esa época, esa información es, en muchos casos, difícil o costosa de conseguir. Se usó en este trabajo información de libre acceso, por lo que la información usada corresponde al periodo 1992-2022, lo cual podría ser mejorado aumentando el tamaño el tamaño de los datasets. En la sección 11A se encuentra una recomendación del periodo para el cual sería factible recopilar datos adicionales, así como las fuentes de estos datos.
- b. Se presentaron problemas de compatibilidad para la implementación del método Autogluon, por lo que, al no lograr solucionar estos problemas en un periodo de una semana, se decidió descartar este método para la comparación posterior de resultados.
- c. La optimización de algunos métodos de aprendizaje consumió mucho tiempo, por lo que fue necesario recortar algunas opciones de hiperparámetros para hacer este tiempo más corto. Especialmente esto sucedió con el método SVC, ya que el hiperparámetros kernel en la configuración ‘poly’ emplea un tiempo excesivo para su ejecución, lo que causó su descarte después de casi 30 horas de ejecutado el algoritmo y con un avance de aproximadamente la mitad de los valores requeridos para construir la gráfica con los resultados.
- d. Al evaluar los resultados, se encontró que los algoritmos obtienen bajos resultados en la métrica F1-score de la clase BUY, lo que significa que es poco fiable en caso de que se recomiende esta clase.

11 TRABAJOS FUTUROS Y RECOMENDACIONES

El trabajo realizado aún tiene limitaciones claras para una implementación comercial, por esta razón hay algunos puntos a mejorar. Los siguientes son puntos en los que se podría darle seguimiento al presente trabajo.

- a. Ampliar la base de datos limitada que se tiene actualmente (desde 1992). Para esto, se recomienda consultar la base de datos “Refinitiv's I/B/E/S Estimates”, la cual tiene este tipo de datos desde 1984. Igualmente, se recomienda buscar la guía “Daily Stock Price Record. New York Stock Exchange by Standard & Poor's Corp”. Sin embargo, obtener datos de ambas fuentes requiere inversión en capital debido a que, en el primer caso, se requiere solicitar esta información antes de tener el derecho a comprarla, además pudiendo pasar que no respondan debido a que el volumen requerido no es suficiente. En el segundo caso, la guía no se encontró digitalizada para los años anteriores a 1992 (que sí se tienen datos), por lo que sería un poco tedioso recopilar los datos por este medio. Además, se requiere recopilar la información de las compañías del DJIA que no se presentan en este trabajo.
- b. Implementar y comparar los resultados del método Autogluon con los resultados obtenidos en este trabajo o resultados propios con sets de datos actualizados. Esta implementación no fue posible debido a problemas de compatibilidad del método. Además, se puede investigar otros métodos de aprendizaje de máquina adecuados para este tipo de aplicación.
- c. Incluir algunos de los hiperparámetros que generan un alto tiempo de ejecución para la comparación de resultados de optimización de hiperparámetros, tales como ‘poly’ del hiperparámetro kernel de SVC, o aumentar el número de `n_estimators` de los métodos relacionados con árbol de decisión y derivados como Random Forest y LightGBM. Se recomienda el uso de un computador de alto rendimiento para esta causa. El computador usado para este trabajo cuenta con 12 GB de RAM, Intel Core i5-7200, CPU @2.5 GHz, y no cuenta con una memoria gráfica o de otro tipo destacable.
- d. Incluir nuevos parámetros que permitan una mejor predicción de las variaciones futuras de los precios de las acciones. Se recomienda usar parámetros que permitan determinar el “sentimiento” del mercado e inversores en cada compañía, por lo que se podría implementar “análisis de sentimiento”. Por otra parte, es posible incluir una variable que compare la volatilidad de una empresa con la volatilidad promedio del mercado. Adicionalmente, es posible que la variación porcentual diaria o en cierto periodo de tiempo del DJIA sea una variable útil para este tipo de modelo.

- e. Incluir y comparar modelos que obtengan una predicción de los movimientos del precio de las acciones en periodos superiores a 1 mes. Se recomienda realizar modelos que generen una predicción de hasta 1 año (o incluso más), ya que se prevé la posibilidad de, al usar variables financieras e intrínsecas al buen funcionamiento de la empresa, este tipo de modelos sea, de hecho, aún más fiable para predicciones a largo plazo, lo cual requeriría un mayor grado de experimentación con los modelos para comprobar, o no, esta suposición.
- f. Se requiere optimizar los modelos de aprendizaje de máquina de forma que se obtengan resultados más altos, sin sacrificar los ya encontrados, de la métrica F1-score de la clase BUY, cuyo bajo valor hace que no sea confiable el modelo en caso de predecir la etiqueta BUY.

12 BIBLIOGRAFÍA

- [1] L. Saad, “What Percentage of Americans Owns Stock?,” *Gallup*, 2022. <https://news.gallup.com/poll/266807/percentage-americans-owns-stock.aspx> (accessed Sep. 17, 2022).
- [2] R. Furhmann, “How The Internet Has Changed Investing,” 2022. <https://www.investopedia.com/financial-edge/0212/how-the-internet-has-changed-investing.aspx> (accessed Sep. 17, 2022).
- [3] S. Swanson, “The Impact of Zero Commissions on Retail Trading and Execution | Coalition Greenwich,” 2020. <https://www.greenwich.com/equities/impact-zero-commissions-retail-trading-and-execution> (accessed Sep. 18, 2022).
- [4] yahoo finance, “GameStop Corp. (GME) Stock Price, News, Quote & History - Yahoo Finance,” *Yahoo finance*. <https://finance.yahoo.com/quote/GME?p=GME> (accessed Sep. 03, 2022).
- [5] A. Van Buskirk, “Volatility Skew, Earnings Announcements, and the Predictability of Crashes,” *SSRN Electron. J.*, Jan. 2012, doi: 10.2139/SSRN.1740513.
- [6] Y. Li, “80% of the stock market is now on autopilot,” *cnn*, 2019. <https://www.cnn.com/2019/06/28/80percent-of-the-stock-market-is-now-on-autopilot.html> (accessed Sep. 16, 2022).
- [7] I. Ph Jansen Associate Professor of Accounting and A. L. Nikiforov Assistant Professor of Finance, “Fear and Greed: a Returns-Based Trading Strategy around Earnings Announcements,” 2016.
- [8] Portafolio, “Indicador Dow Jones cumplió 125 años: su historia | Internacional | Portafolio,” *Portafolio*, 2021. <https://www.portafolio.co/internacional/indicador-dow-jones-cumplio-125-anos-su-historia-552429> (accessed Sep. 03, 2022).
- [9] World Economic Forum, “6 issues that will define the future of capital markets | World Economic Forum,” 2021. <https://www.weforum.org/agenda/2021/09/six-issues-to-define-the-future-of-capital-markets/> (accessed Sep. 16, 2022).
- [10] Mirae Asset Mutual Fund, “What Is Greed Cycle | How Does It Influence Investors | Mirae Asset.” <https://www.miraeassetmf.co.in/knowledge-center/importance-of-investor-behaviour-in-market-correction-greed-and-fear-cycle> (accessed Sep. 15, 2022).
- [11] R. L. Peterson, “Trading on sentiment : the power of minds over markets”.
- [12] A. Hayes, “Financial Markets Definition,” *investopedia*, 2022. <https://www.investopedia.com/terms/f/financial-market.asp> (accessed Sep. 03, 2022).
- [13] A. Sevilla, “Índice bursátil - Qué es, definición y concepto | 2022 | Economipedia,” *Economipedia*. <https://economipedia.com/definiciones/indice-bursatil.html> (accessed Sep. 03, 2022).
- [14] A. Hayes, “Stocks: What They Are, Main Types, How They Differ From Bonds,” *Investopedia*, 2022. <https://www.investopedia.com/terms/s/stock.asp> (accessed Sep. 16,

2022).

- [15] J. Montes, “Dow Jones - Qué es, definición y concepto | 2022 | Economipedia,” *Economipedia*. <https://economipedia.com/definiciones/dow-jones.html> (accessed Sep. 03, 2022).
- [16] C. Murphy, “Financial Statements: List of Types and How to Read Them,” *Investopedia*, 2022. <https://www.investopedia.com/terms/f/financial-statements.asp> (accessed Sep. 16, 2022).
- [17] A. Sevilla, “Beneficio por acción (BPA) - Qué es, definición y concepto | 2022 | Economipedia,” *Economipedia*. <https://economipedia.com/definiciones/beneficio-por-accion.html> (accessed Sep. 03, 2022).
- [18] CFI team, “Earnings Estimate - Overview, Revisions, and Impact on Stocks,” *CFI*, 2020. <https://corporatefinanceinstitute.com/resources/knowledge/trading-investing/earnings-estimate/> (accessed Sep. 03, 2022).
- [19] J. López, “Análisis técnico - Qué es, definición y concepto | 2022 | Economipedia,” *Economipedia*. <https://economipedia.com/definiciones/analisis-tecnico.html> (accessed Sep. 03, 2022).
- [20] J. F. López, “Análisis fundamental - Qué es, definición y concepto | 2022 | Economipedia,” <https://economipedia.com/definiciones/analisis-fundamental.html> (accessed Sep. 16, 2022).
- [21] R. Schlotmann and M. Czubatinski, “Trading : technical analysis masterclass : master the financial markets,” p. 171, 2019.
- [22] M. Grant, “Stock Market Opening Times Around the World,” *Investopedia*, 2022. <https://www.investopedia.com/ask/answers/040115/when-do-stock-market-exchanges-close.asp> (accessed Sep. 05, 2022).
- [23] J. Fernando, “Sell-Off Definition,” *Investopedia*, 2022. <https://www.investopedia.com/terms/s/sell-off.asp> (accessed Sep. 03, 2022).
- [24] J. Chen, “Rally Definition,” *Investopedia*, 2021. <https://www.investopedia.com/terms/r/rally.asp> (accessed Sep. 03, 2022).
- [25] W. Kenton, “Crash Definition,” *Investopedia*, 2021. <https://www.investopedia.com/terms/c/crash.asp> (accessed Sep. 03, 2022).
- [26] A. Sevilla, “Cobertura financiera - Qué es, definición y concepto | 2022 | Economipedia,” *Economipedia*. <https://economipedia.com/definiciones/cobertura-financiera.html> (accessed Sep. 03, 2022).
- [27] R. Balasubramanian, “Predicting the Volatility of Stock Data | by Ramji Balasubramanian | Analytics Vidhya | Medium,” *Analytics Vidhya*, 2020. <https://medium.com/analytics-vidhya/predicting-the-volatility-of-stock-data-56f8938ab99d> (accessed Sep. 03, 2022).
- [28] IBM México, “Acerca de la regresión lineal - México | IBM,” *IBM México*. <https://www.ibm.com/mx-es/analytics/learn/linear-regression> (accessed Sep. 03, 2022).
- [29] IBM España, “Clasificaciones - Documentación de IBM,” *IBM*, 2021. <https://www.ibm.com/docs/es/maximo-for-aviation/7.6.1?topic=module-classifications->

application (accessed Sep. 03, 2022).

- [30] G. cloud Team, “Hyperparameter tuning | AI platform training | Google cloud,” *Google*. <https://cloud.google.com/ai-platform/training/docs/hyperparameter-tuning-overview> (accessed Sep. 11, 2022).
- [31] J. Fernando, “R-Squared Formula, Regression, and Interpretations,” *Investopedia*, 2021. <https://www.investopedia.com/terms/r/r-squared.asp> (accessed Aug. 30, 2022).
- [32] N. Singh Chauhan, “Métricas De Evaluación De Modelos En El Aprendizaje Automático,” *DataSource.AI*, 2020. <https://www.datasource.ai/es/data-science-articles/metricas-de-evaluacion-de-modelos-en-el-aprendizaje-automatico> (accessed Aug. 30, 2022).
- [33] N. Naik and B. R. Mohan, “Stock price movements classification using machine and deep learning techniques-the case study of indian stock market,” *Commun. Comput. Inf. Sci.*, vol. 1000, pp. 445–452, 2019, doi: 10.1007/978-3-030-20257-6_38.
- [34] W. Kenton, “S&P 500 Index: What It’s for and Why It’s Important in Investing,” 2022. <https://www.investopedia.com/terms/s/sp500.asp> (accessed Sep. 22, 2022).
- [35] T. A. Smith and A. Rajan, “A Regression Model to Predict Stock Market Mega Movements and/or Volatility using both Macroeconomic indicators & Fed Bank Variables,” *Int. J. Math. Trends Technol.*, vol. 49, no. 3, pp. 165–167, Sep. 2017, doi: 10.14445/22315373/IJMTT-V49P522.
- [36] S. Jansen, “Machine Learning for Algorithmic Trading - Second Edition,” 2020.
- [37] T. Sanboon, K. Keatruangkamala, and S. Jaiyen, “A Deep Learning Model for Predicting Buy and Sell Recommendations in Stock Exchange of Thailand using Long Short-Term Memory,” *IEEE*, pp. 757–760, Sep. 2019, doi: 10.1109/CCOMS.2019.8821776.
- [38] S. A. Gyamerah, P. Ngare, and D. Ikpe, “On Stock Market Movement Prediction Via Stacking Ensemble Learning Method,” *CIFER 2019 - IEEE Conf. Comput. Intell. Financ. Eng. Econ.*, May 2019, doi: 10.1109/CIFER.2019.8759062.
- [39] M. L. Mitchell and J. H. Mulherin, “The Impact of Public Information on the Stock Market,” *J. Finance*, vol. 49, no. 3, p. 923, Jul. 1994, doi: 10.2307/2329211.
- [40] S. K. Mohanty and E. N. W. Aw, “Rationality of analysts’ earnings forecasts: evidence from dow 30 companies,” <http://dx.doi.org/10.1080/09603100500426564>, vol. 16, no. 12, pp. 915–929, Aug. 2006, doi: 10.1080/09603100500426564.
- [41] A. Ganti, “Dow Jones Industrial Average (DJIA) Definition,” *Investopedia*, 2022. <https://www.investopedia.com/terms/d/djia.asp> (accessed Aug. 30, 2022).
- [42] S. Schaefer, “The 10 Oldest Dow Components,” *Forbes*, 2011. https://www.forbes.com/2011/05/26/dow-at-115-longest-tenured-stocks_slide.html?sh=5a094f6c35e4 (accessed Aug. 30, 2022).
- [43] SEC, “SEC.gov | Beginners’ Guide to Financial Statements,” *SEC*, 2017. <https://www.sec.gov/oiea/reports-and-publications/investor-publications/beginners-guide-financial-statements> (accessed Aug. 30, 2022).
- [44] E. Wolff-Mann, “How a stock gets added to the Dow Jones Industrial Average,” *Yahoo*

- finance*, 2020. <https://finance.yahoo.com/news/how-stock-gets-added-to-dow-jones-industrial-average-181311632.html?guccounter=1> (accessed Aug. 30, 2022).
- [45] E. Terrell, “Research Guides: Doing Historical Company Research: A Resource Guide: Stock Price Sources,” *Libr. Congr.*, Accessed: Aug. 30, 2022. [Online]. Available: <https://guides.loc.gov/historical-company-research/stocks>
 - [46] N. D. Link, “ZES | Zacks Earnings Surprises Documentation | Nasdaq Data Link,” *Nasdaq Data Link*. <https://data.nasdaq.com/databases/ZES/documentation> (accessed Aug. 30, 2022).
 - [47] IBM, “Conozca IBM, Productos y Soluciones,” *IBM*, 2009, Accessed: Aug. 30, 2022. [Online]. Available: <http://www.ibm.com/mx>
 - [48] M. Krzywinski and N. Altman, “Visualizing samples with box plots,” *Nat. Methods*, vol. 11, no. 2, pp. 119–120, Feb. 2014, doi: 10.1038/NMETH.2813.
 - [49] Y. Altunbas, “Box plot distribution of the stock market returns of individual banks... | Download Scientific Diagram,” 2014. https://www.researchgate.net/figure/Box-plot-distribution-of-the-stock-market-returns-of-individual-banks-The-diagram-below_fig1_321224673 (accessed Aug. 30, 2022).
 - [50] P. Davidsson, P. Steffens, and J. Fitzsimmons, “Growing profitable or growing from profits: Putting the horse in front of the cart?,” *Acad. Manag. 2005 Annu. Meet. A New Vis. Manag. 21st Century, AOM 2005*, 2005, doi: 10.5465/AMBPP.2005.18778649.
 - [51] J. Fernando, “Correlation Coefficient Definition,” *Investopedia*, 2021. <https://www.investopedia.com/terms/c/correlationcoefficient.asp> (accessed Aug. 30, 2022).
 - [52] O. B. Sezer and A. M. Ozbayoglu, “Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach,” *Appl. Soft Comput.*, vol. 70, pp. 525–538, Sep. 2018, doi: 10.1016/J.ASOC.2018.04.024.
 - [53] IBM, “Course | Machine Learning (aprendizaje automático) con Python: una introducción práctica | edX,” *edx*. <https://learning.edx.org/course/course-v1:IBM+ML0101SP+1T2022/home> (accessed Sep. 05, 2022).
 - [54] A. Ganti, “Brokerage Fee Definition,” *Investopedia*, 2022. <https://www.investopedia.com/terms/b/brokerage-fee.asp> (accessed Aug. 30, 2022).
 - [55] Scikit Learn, “sklearn.neighbors.KNeighborsClassifier — scikit-learn 1.1.2 documentation.” <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html> (accessed Aug. 30, 2022).
 - [56] Scikit Learn, “sklearn.tree.DecisionTreeClassifier — scikit-learn 1.1.2 documentation.” <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html> (accessed Aug. 30, 2022).
 - [57] ibm, “¿Qué es el algoritmo de k vecinos más cercanos? | IBM.” <https://www.ibm.com/co-es/topics/knn> (accessed Aug. 30, 2022).
 - [58] IBM, “¿Qué es un árbol de decisión? | IBM,” *IBM*. <https://www.ibm.com/es-es/topics/decision-trees> (accessed Aug. 30, 2022).

- [59] “Which Machine Learning Classifiers are Best for Small D...,” 2021. <https://www.data-cowboys.com/blog/which-machine-learning-classifiers-are-best-for-small-datasets> (accessed Aug. 30, 2022).
- [60] AutoGluon, “FAQ — AutoGluon Documentation 0.5.2 documentation.” https://auto.gluon.ai/stable/tutorials/tabular_prediction/tabular-faq.html (accessed Sep. 22, 2022).
- [61] LightGBM, “Welcome to LightGBM’s documentation! — LightGBM 3.3.2 documentation,” *LightGBM*, 2022. <https://lightgbm.readthedocs.io/en/v3.3.2/> (accessed Aug. 30, 2022).
- [62] IBM Cloud Education, “What is Random Forest? | IBM,” *IBM Cloud Education*, 2020. <https://www.ibm.com/cloud/learn/random-forest> (accessed Aug. 30, 2022).
- [63] IBM, “What is Logistic regression? | IBM,” *IBM*. <https://www.ibm.com/topics/logistic-regression> (accessed Aug. 30, 2022).
- [64] IBM, “About SVM - IBM Documentation,” *IBM*, 2021. <https://www.ibm.com/docs/en/spss-modeler/saas?topic=models-about-svm> (accessed Aug. 30, 2022).
- [65] Analytics Vidhya, “Hyperparameter Tuning | Evaluate ML Models with Hyperparameter Tuning,” *Analytics Vidhya*, 2021. <https://www.analyticsvidhya.com/blog/2021/04/evaluating-machine-learning-models-hyperparameter-tuning/> (accessed Aug. 30, 2022).
- [66] Scikit Learn, “sklearn.ensemble.RandomForestClassifier — scikit-learn 1.1.2 documentation,” *Scikit Learn*. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html> (accessed Aug. 30, 2022).
- [67] Scikit Learn, “sklearn.linear_model.LogisticRegression — scikit-learn 1.1.2 documentation,” *Scikit Learn*. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (accessed Aug. 30, 2022).
- [68] Scikit Learn, “sklearn.svm.SVC — scikit-learn 1.1.2 documentation,” *Scikit Learn*. <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html> (accessed Aug. 30, 2022).
- [69] LightGBM, “Parameters — LightGBM 3.3.2.99 documentation,” *LightGBM*. <https://lightgbm.readthedocs.io/en/latest/Parameters.html> (accessed Aug. 30, 2022).
- [70] M. Ben Fraj, “In Depth: Parameter tuning for SVC | by Mohtadi Ben Fraj | All things AI | Medium,” 2018. <https://medium.com/all-things-ai/in-depth-parameter-tuning-for-svc-758215394769> (accessed Aug. 30, 2022).
- [71] J. Brownlee, “Tune Hyperparameters for Classification Machine Learning Algorithms,” 2019. <https://machinelearningmastery.com/hyperparameters-for-classification-machine-learning-algorithms/> (accessed Aug. 30, 2022).