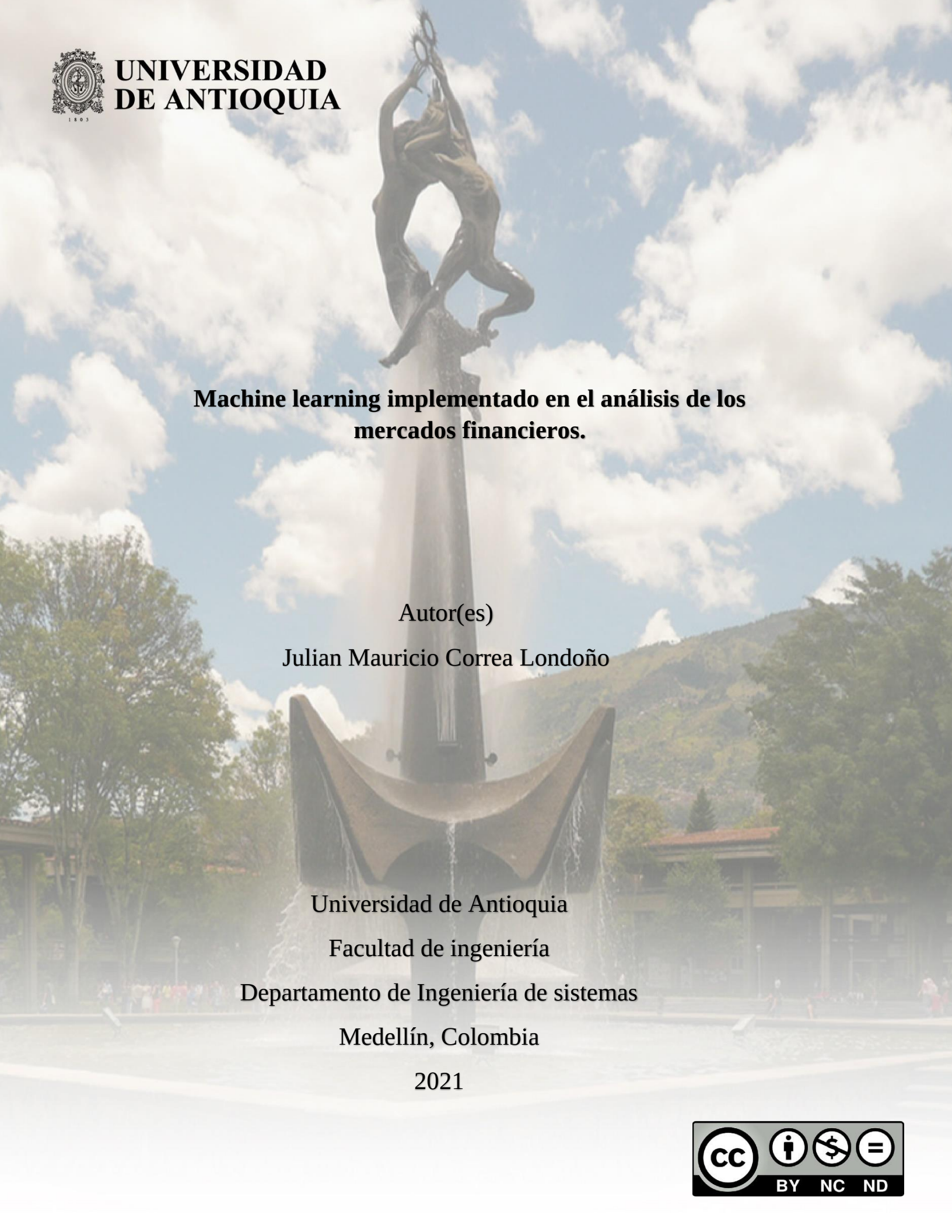


UNIVERSIDAD DE ANTIOQUIA

Machine learning implementado en el análisis de los mercados financieros.

Autor(es)

Julian Mauricio Correa Londoño

Universidad de Antioquia

Facultad de ingeniería

Departamento de Ingeniería de sistemas

Medellín, Colombia

2021

Machine learning implementado en el análisis de los mercados financieros.

Julian Mauricio Correa Londoño

Tesis o trabajo de investigación presentada(o) como requisito parcial para optar al título de:
Ingeniero de sistemas.

Asesores (a):
Sandra Patricia Zabala Orrego, Ing. Informática – Esp Gerencia

Julie Andrea Gonzalez Morales, Ing. Matemática – Esp en finanzas – Máster en economía

Universidad de Antioquia
Facultad de ingeniería
Departamento de Ingeniería de Sistemas
Medellín, Colombia
2021

Resumen

La gerencia de soluciones analíticas de tesorería, perteneciente a la vicepresidencia de tesorería del grupo Bancolombia, se enfocó en la búsqueda de información para la creación de nuevas herramientas basadas en machine learning, capaces de generar valores o señales de compra y/o venta de activos. Esto con el fin de aumentar el portafolio de servicios ya existente dentro de la gerencia. El desarrollo de esta herramienta es logrado después de una revisión bibliográfica con la cual se encontraron distintos campos de aplicación y enfoques, de los cuales el abarcado en este documento es el de predicción o clasificación de señales de compra y/o venta de activos.

Introducción

La gerencia de soluciones analíticas de tesorería (abreviado como GSAT por sus siglas), perteneciente a la vicepresidencia de tesorería del grupo Bancolombia, cuenta con más de 10 años de experiencia apoyando las funciones de las demás mesas pertenecientes a la vicepresidencia y desarrollando aplicativos/herramientas para el análisis de los diferentes mercados financieros (Renta Fija, Renta Variable, Divisas y Derivados).

Entre algunas de sus funciones se encuentra analizar y entender el comportamiento del mercado financiero, para generar estrategias que permitan aumentar las ganancias tanto de sus clientes como del grupo Bancolombia.

De esta función surge la idea de agregar dentro de su portafolio, herramientas basadas en el aprendizaje de máquinas (Machine learning), tales como: el análisis de sentimientos, pronósticos inteligentes, clasificadores de datos, identificación de patrones, entre otros.

Se muestra el resultado llegado al aplicar la adaptación del marco de desarrollo ágil SCRUM, el cual es el inicio de las nuevas funcionalidades basadas en machine learning para mejorar los pronósticos de la gerencia al tiempo que se expande el portafolio de herramientas existente.

Objetivo General

Investigar e Implementar un aplicativo/herramienta capaz de utilizar modelos de machine learning que puedan ser aplicados al mercado financiero.

Objetivos Específicos

- Investigar cuáles son los modelos/metodologías dentro de la tipología de machine learning, más exitosos para ser aplicados en los mercados financieros ya sea para pronóstico o para clasificación/análisis de sentimientos.
- Identificar diferentes formas de implementación del machine learning.
- Desarrollar prototipos de modelos que apliquen machine learning.
- Identificar de los modelos implementados cuáles son los más eficientes en términos de generación de valor en los mercados financieros.
- Migrar los modelos existentes de machine learning al lenguaje python e implementar nuevas funcionalidades.
- Apoyar en el desarrollo de nuevas funcionalidades dentro del área.
- Generar documentación para futuras implementaciones.

Marco Teórico

La inteligencia artificial es un campo de las ciencias de la computación con un gran foco de interés en la actualidad, el cual consiste en la generación de máquinas inteligentes capaces de operar por su propia cuenta sin la intervención humana. En el desarrollo de la inteligencia artificial se encuentran distintas técnicas y campos de aplicación tales como: el machine learning, la lógica difusa y el procesamiento del lenguaje natural.

Siendo una de las más mencionadas el machine learning, se define como la rama de la inteligencia artificial que le otorga a las máquinas la capacidad de aprender sin ser expresamente programadas. Esta rama está especialmente enfocada en crear programas que de forma automática sean capaces de generar modelos de clasificación, pronóstico y predicción como otros ejemplos posibles.

El machine learning, puede ser usado en muchas áreas, por ejemplo, la financiera, donde podemos buscar predecir valores, señales de compra, o encontrar patrones en los títulos contables, los cuales otorgan a su dueño ingresos futuros procedentes de quien los venda, o dichos en otras palabras en los activos financieros, los cuales son uno de los problemas más interesantes de afrontar desde el machine learning por la gran complejidad que pueden tener en algunos casos.

Metodología

El proyecto se realizó implementando una adaptación del marco de trabajo para desarrollo ágil de software SCRUM adaptada por GSAT, en la cual la función de "producto owner" fue desempeñada por la gerente de GSAT.

Dicha metodología fue ajusta para para tener cada semana una interacción con el nombre "Sprint", las tareas de cada sprint se realizaron usando la metodología Kanban, con la cual se definieron las tareas objetivo, para avanzar o terminar cada semana

Cada interacción concluyó con una reunión de aproximadamente una hora al final de la semana laboral, donde se analizó el Kanban y se organizaban los objetivos de la siguiente semana, basándose en la prioridad y avance de los ya existentes junto con lo nuevo que se agregaba.

Para una mejor distribución de las actividades, los objetivos relacionados con el proyecto tuvieron un plazo máximo de 2 Sprints, durante los cuales se evaluó el avance en cada Sprint, las actividades realizadas en cada sprint y su cronograma se pueden ver en las tablas 1 y 2 respectivamente.

Para el trabajo se usaron dos aplicaciones fundamentales para la creación de modelos: el framework de inteligencia artificial de keras, con el cual se manejó todo lo relacionado a redes neuronales, la librería de aprendizaje automático scikit – learn (sklearn), encargada de generar múltiples tipos de modelos y realizar sus respectivas pruebas con varios parámetros. Por último, también está la librería de Ta-Lib la cual es útil para el análisis técnico de los datos financieros.

Detalle de los sprint

Sprint	Actividades
Sprint 1 al 2	- Profundización en Machine learning. - Profundización en Deep Learning. - Investigación de aplicaciones del Machine learning en el mundo real. - Búsqueda de entornos de desarrollo para Machine learning.
Sprint 3 al 4	- Socialización y retroalimentación con GSAT. - Configuración del entorno necesario para la aplicación básica de machine learning. - Experimentación con librerías de machine learning. - Comienzan las reuniones para definir la información de los sets de datos. - Identificar modelos que no se podrán aplicar por limitantes de seguridad corporativa.
Sprint 5 al 6	- Búsqueda de sets de datos públicos para testeo. - Investigación de mejores estrategias de machine learning. - Socialización y retroalimentación de los avances. - Desarrollo inicial del set de datos.
Sprint 7 al 8	- Búsqueda de frameworks para la interfaz del sistema. - Inicio del diseño de la interfaz. - Implementación de modelos usando los sets de datos desarrollados. - Socialización del avance y retroalimentación de los temas y códigos desarrollados.
Sprint 9 al 10	- Configuración de los modelos para ser consumidos desde un servidor. - Socialización de avances.
Sprint 11 al 12	- Optimizar los detalles de los modelos. - Socialización de avances.
Sprint 13 al 14	- Ajustes finales de los modelos. - Socialización de avances y cambios menores finales.
Sprint 15 al 16	- Socialización final del trabajo con el equipo y product owner. - Pruebas finales de los modelos para depurar detalles faltantes.

Tabla 1. Detalles de los Sprints

Cronograma de Actividades

Sprints	Inicio	Final	Septiembre				Octubre				Noviembre				Diciembre			
Sprint 1 al 2	07/09/2020	21/09/2020																
Sprint 3 al 4	21/09/2020	05/10/2020																
Sprint 5 al 6	05/10/2020	19/10/2020																
Sprint 7 al 8	19/10/2020	02/11/2020																
Sprint 9 al 10	02/11/2020	16/11/2020																
Sprint 11 al 12	16/11/2020	30/11/2020																
Sprint 13 al 14	30/11/2020	14/12/2020																
Sprint 15 al 16	14/12/2020	28/12/2020																

Tabla 2. Cronograma de los Sprints

Resultados y análisis

A través de la metodología planteada, se comenzó el proyecto realizando una revisión bibliográfica con el objetivo de hallar metodologías, aplicaciones y modelos que fuesen útiles al momento de trabajar con activos financieros (sea acciones o divisas), de esta revisión se destacan 3 implementaciones importantes:

- **Predicción de tendencias**: Se refiere a la aplicación basada en predecir la dirección que tomará el activo financiero, también por su funcionamiento puede servir para predecir un valor futuro del mismo.
- **Detección de patrones**: Busca detectar patrones, en los datos básicos de un activo financiero (apertura, cierre, máximo y mínimo) o en las gráficas de velas que son la representación visual de estos datos básicos, todo esto con el objetivo de encontrar posibles movimientos futuros en los mercados financieros.
- **Procesamiento de lenguaje**: Es la habilidad que desarrolla una máquina para interpretar y darle contexto a todo tipo de texto, en este caso su dominio sería el mercado financiero, permitiendo dar una vista de cómo se mueve y moverá un mercado específico.

Siendo la predicción de tendencias el foco principal de este trabajo, se realizó otra revisión bibliográfica, con el objetivo de encontrar implementaciones e información referente, esto trajo consigo dos enfoques, el de redes neuronales para la predicción del valor futuro de un activo financiero y los modelos clásicos de machine learning para la predicción de una señal (compra o venta). En la tabla 3 podemos ver un resumen relacionado con estos enfoques.

Enfoque	Que se esperaba	Resultado
Redes neuronales	Una red neuronal capaz de dar resultados confiables sobre el valor de un activo financiero en un futuro cercano, para tomar decisiones relacionadas a la compra y/o venta de este.	Una red capaz de ajustarse de forma aceptable a los datos de entrenamiento y test, pero poco confiable en las predicciones por tener resultados con ruido que pueden indicar un movimiento contrario en muchas ocasiones.
Modelos clásicos	Encontrar un modelo capaz de predecir la tendencia en la cual apuntará un activo financiero en un futuro cercano, para usar la misma al momento de realizar operaciones de compra y/o venta	Un modelo capaz de generar señales de compra y/o venta con un nivel de confianza alto dependiendo de cómo se defina inicialmente cuando una tendencia sube o baja (señal).

Tabla 3. Resumen enfoques de modelos

Para los resultados vistos en la tabla 3, se realizaron pruebas con distintos activos financieros, principalmente tasas de cambio debido a la gran liquidez comercial que tienen, permitiendo tener datos muy variables en nuestros modelos, algunas de las tasas de cambio usadas son: USDJPY (dólar - yen), EURUSD (Euro – dólar) y GBPUSD (libra – dólar), de las cuales todo el análisis a profundidad se basó en el EURUSD por ser la tasa de cambio más comercializada dentro del grupo de trabajo.

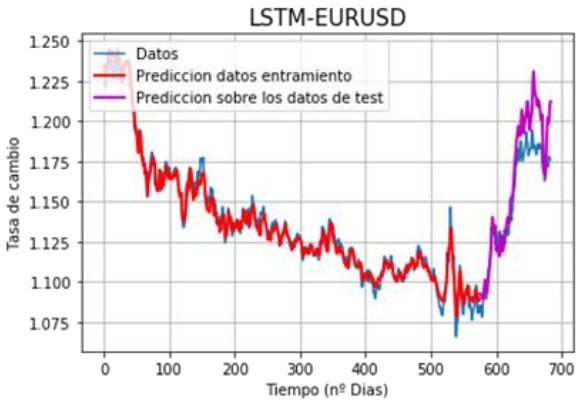
Enfoque de Redes neuronales

El objetivo de este enfoque es predecir el valor en un futuro cercano de un activo financiero, usando para ellos redes neuronales, específicamente usando redes LSTM (Long short-term memory). Estas redes en partículas son escogidas por su alta eficiencia con series de tiempo, por lo cual son ideales para ser usadas en datos históricos relacionados con el valor de cualquier activo financiero.

Para el caso a presentar, nuestro periodo de prueba es conformado por datos diarios que abarcan el periodo desde enero 1 del 2018, hasta octubre 9 del 2020 esto debido a la disponibilidad histórica que se tenía al momento de las pruebas.

parámetro	Error Cuadrático Medio (RMSE)
Puntuación del entrenamiento:	0.01342 ± 0.11586
Puntuación del test:	0.15398 ± 0.39241

Tabla 4. Puntajes LSTM - EURUSD



Gráfica 1. LSTM – EURUSD

La configuración de la red LSTM usada para este caso es: una ventana de tiempo de 40 días (una serie de 40 precios en el tiempo como características) con 3 capas donde las 2 primeras tienen 128 neuronas y la ultima 64, en medio de estas se descartó el 20% de la información para evitar un posible sobre ajuste de los datos, esto se repite durante 30 épocas, valor al que se llegó luego de múltiples pruebas y ver que la mejora no era significativa con un mayor número de épocas.

Nuestros resultados muestran exactamente lo esperado al momento de plantear una red LSTM, podemos observar cómo los datos de entrenamiento que representan al 85% de los datos totales, se ajustan de manera casi perfecta a los datos reales, caso similar se ve con los datos restantes correspondientes a la prueba, logrando comprobar lo fuerte que es una red LSTM al usar series de tiempo. No obstante, a pesar de visualizar resultados tan prometedores, este enfoque basado en predecir el valor futuro del activo financiero, se descartó debido a un análisis minucioso de los datos predichos por el modelo, en los cuales se observa que siempre el resultado es igual al último día de la ventana de tiempo , con un ruido , el cual no es muy preciso, causando en algunos casos que la predicción esté en una dirección contraria a la vista en los datos reales, esto se puede también colaborar al ver los resultados del error cuadrático medio del modelo , los cuales para el tipo de activo usado son considerablemente alto para ser viables.

Enfoque en modelos clásicos de Machine Learning.

Este enfoque busca predecir el movimiento de tendencia que seguirá un activo financiero (alcista, bajista o lateral) en un futuro cercano, con la intención de generar señales de compra y/o venta en el mismo.

Para este caso en particular se optó por hacer una implementación que ignora completamente la forma de una serie de tiempo, no teniendo en cuenta la estampa de tiempo y tomando los datos como muestras independiente, se tomó la decisión de trabajar de este modo después de una revisión minuciosa a los notebook , foros y comentarios de la competencia/problema de Jane Street Market Prediction de la plataforma kaggle , la cual al ser muy similar a nuestro enfoque , sirvió como un punto de apoyo muy grande para este.

El set de datos usado sigue cumpliendo la misma estructura básica de apertura, cierre, precio mayor y menor, de cada periodo, además se agregan indicadores técnicos y transformaciones de moneda, que sirven para dar un mayor entendimiento al modelo sobre el movimiento del activo financiero.

Para este enfoque los modelos fueron generados usando scikit – learn, librería de la cual también se implementaron varias de sus funciones al momento de programación, a su vez se usó Ta-Lib, para generar los indicadores técnicos y las transformaciones de moneda, todo gracias a que estos usualmente se manejan como datos para análisis técnico lo cual es el fuerte de esta librería.

Lamentablemente, nuestra variable dependiente que hace a su vez de señal de mercado no existe previamente en el modelo, por lo cual es necesario generar esta señal usando para ello los propios datos existentes.

Para que la señal se pueda considerar aceptable tiene que cumplir tres condiciones mínimas:

- Tener un retorno positivo al momento de hacer simulación de mercado
- Tener una métrica de Sharpe ratio (rendimiento de la inversión) mayor a 2
- Que el modelo generado con esta señal sea capaz de superar a la caminata aleatoria (el equivalente a lanzar una moneda y tomar las elecciones por la misma) al momento de pasarle los datos de prueba.

Para la construcción de nuestra señal, se pensaron varias estrategias, algunas de estas son:

- **Cambio porcentual simple:** Consiste en realizar el cambio porcentual entre cada muestra de la serie de tiempo, usualmente usando el valor de cierre dando como resultado una señal, que puede ser negativa o positiva dependiendo del valor del cambio, estos valores se desplazan un periodo en el tiempo hacia atrás haciendo que la muestra del día anterior tenga como señal el movimiento esperado para el día siguiente.

- **Cambio porcentual con margen:** Su funcionamiento inicial es el mismo que en el cambio porcentual simple, con la diferencia de que este agrega un margen porcentual que es necesario sobrepasar para decir que la señal es positiva o negativa, si dicho margen no es superado, la señal se clasifica como lateral, la cual en nuestra simulación de mercado se considera como un cierre de la operación actual.
- **Moda del cambio porcentual:** Puede ser una variante de alguno de los dos casos anteriores, donde al resultado obtenido de las señales, se le aplica una ventana móvil de tiempo, esta ventana cumple la función de tomar la moda de la señal en dicho periodo de tiempo, y convertir esta moda en la nueva señal.

Para este caso nuestro periodo de información aumentó considerablemente al tener acceso a datos históricos más antiguos, permitiendo hacer pruebas con datos que van desde Enero 01 del 2010 hasta Diciembre 28 del 2020, estos datos a su vez sufren un pequeño recorte en el extremo inicial dependiendo de los parámetros usados para los indicadores técnicos, también existe otro recorte que tiene que ver con el caso de usar una señal basada en la moda del cambio porcentual , donde dependiendo de la implementación puede disminuir la cantidad de datos en cualquiera de los extremos.

Ya una vez tenemos nuestro set de datos y la variable dependiente o señal que representa cada periodo, se procede a hacer uso de las herramientas de scikit – learn, más específicamente de los modelos y el módulo GrindSearchCV, que nos permite probar los modelos con diferentes parámetros preprogramados, los modelos probados durante este enfoque fueron:

Random Forest Classifier, Ada Boost Classifier y el Gradient Boosting Classifier, cada uno con su respectiva combinación de parámetros.

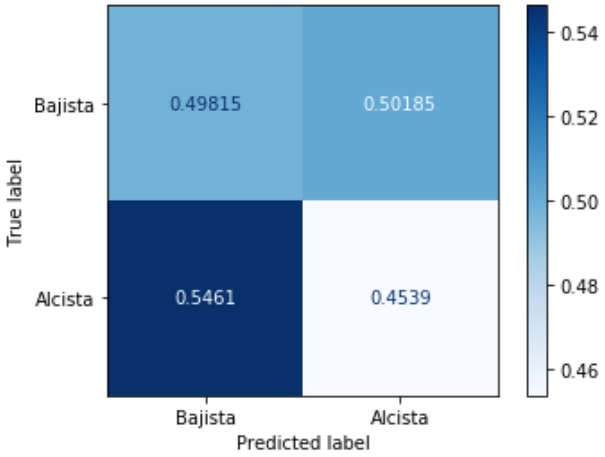
Todas las pruebas se realizaron entrenando el modelo con el 80% de los datos finales del set de datos y haciendo su respectiva prueba con los datos restantes, dichos datos fueron tomados de forma aleatoria asegurándonos de que no se repitiera ni excluyera datos en dichos porcentajes.

El modelo con mejor resultado fue el Random Forest Classifier, superando de forma considerable a los demás modelos, esto puede ser posiblemente causado por usar los datos en su forma más pura sin ningún tipo de normalización o escalado y también por la definición de la señal usada, siendo esta una versión personalizada basada en la moda del cambio porcentual.

También la señal juega un papel importante, como podemos ver en las gráficas 2 y 3 y sus tablas (tabla 5 y 6 respectivamente), en las cuales se compara una señal de cambio porcentual simple y la personalizada con la cual se encontró el mejor modelo expuesto.

Métricas Simulación	**Valor (Euro)**
Inicial	$100000.00
Portafolio Final	$2954750305.92
Sharpe	16.81
Métricas modelo	**Porcentaje**
F1	0.46892 ± 0.08375
Exactitud	0.48685 ± 0.09204
Aciertos	0.47542 ± 0.08567
Error	0.52458 ± 0.08567

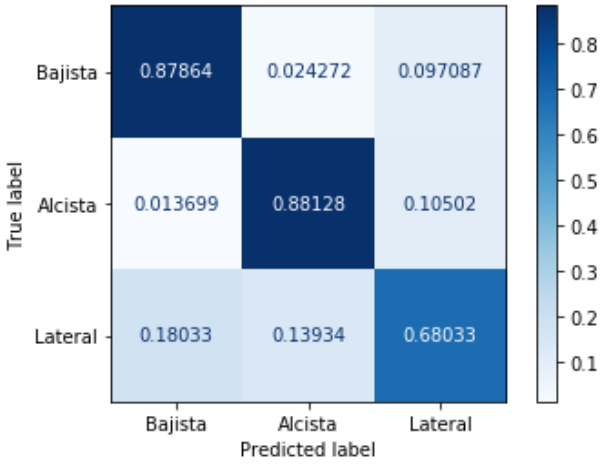
Tabla 5. Test Porcentual Simple



Gráfica 2. Test Porcentual Simple

Métricas Simulación	**Valor (Euro)**
Inicial	$100000.00
Portafolio Final	$514312.39
Sharpe	2.48
Métricas modelo	**Porcentaje**
F1	0.83568 ± 0.04734
Exactitud	0.84236 ± 0.04563
Aciertos	0.83539 ± 0.04879
Error	0.16461 ± 0.04879

Tabla 6. Test personalizado



Gráfica 3. Test personalizado

En esta comparación podemos ver las matrices de confusión acompañadas de sus respectivas métricas, se puede observar un buen ejemplo de una señal que no sirve para predicción de activos financiero como es el caso de la porcentual simple (Gráfica 2, tabla 5), la cual tiene un retorno y un sharpe muy buenos, pero al momento de pasar al modelo, sus resultados no superar la caminata aleatoria (para superarla tiene que existir un 60% de acierto mínimo).

Por otro lado, si observamos la señal personalizada (Gráfica 3, Tabla 6) vemos que su retorno y sharpe son inferiores, aun cumpliendo el criterio mínimo establecido, pero sus métricas al momento de usarse para entrenar el modelo son superiores, dando una victoria más que aceptable sobre la caminata aleatoria, además podemos colaborar que tan bueno es el modelo mirando sus demás métricas y la propia matriz de confusión (Gráfica 3) en la que se observa cómo se clasificó cada muestra de la señal.

Este enfoque resultó ser el más acertado, cumpliendo ciertamente con lo que se buscaba en el mismo, dando resultados mayores a los esperados con respecto a las señales de compra y venta de activos financieros, dejando un modelo de señales que se puede considerar óptimo para la toma de decisiones en un entorno real.

Conclusiones

- La generación de modelos de machine learning para la predicción del mercado financiero es muy compleja, todo gracias a la cantidad enorme de ruido que suelen tener los datos. El efecto causado por este ruido se puede apreciar en el enfoque de redes neuronales, donde este se descarta por el ruido generado en cada predicción.

- En ocasiones es mejor representar el problema de formas alternativas, tal como se ve con el enfoque clásico, en el cual, al ver el problema de una manera diferente a la normal que sería tomar los datos en orden por ser una serie de tiempo, se pudo generar resultados más que aceptables, como se aprecia en este enfoque, al usar la señal personalizada la cual sigue respetado la serie de tiempo por la forma de su construcción.

- El método perfecto para general señales no existe, muchas implementaciones pueden ser mejores que otras, pero esto puede variar dependiendo del activo financiero en muchos casos, haciendo que en ocasiones un tipo de señal solo sea funcional para un activo específico.

- El mejor modelo mostrado al final del enfoque clásico puede ser inferior a otros dependiendo de varios factores como: la estructura del set de datos, la forma como se genera la señal y el activo financiero usado en el mismo, incluso la cantidad de datos, esto es debido en muchos casos a las cualidades propias de cada modelo al momento de entrenarse.

- Predecir el valor futuro de un activo financiero, es muy complicado como pudimos ver en el enfoque de redes neuronales, por ello se recomienda reducir la complejidad del problema al buscar clasificar en señales tal como sucede en el enfoque clásico, cambiando de un problema de predicción a uno de clasificación, que es más fácil de solucionar en ciertos casos.

Referencias Bibliográficas

- BBVA [2020] Machine Learning que es y cómo funciona. Recuperado de: https://www.bbva.com/es/machine-learning-que-es-y-como-funciona/
- BBVA [2020] Activos financieros, ¿Qué son? Recuperado de: https://www.bbva.es/finanzas-vistazo/ef/fondos-inversion/activos-financieros.html
- Sinnaps [2020] Metodología SCRUM. Recuperado de: https://www.sinnaps.com/blog-gestion-proyectos/metodologia-scrum
- Kanbanize [2020] Qué es Kanban: Definición, Características y ventajas. Recuperado de: https://kanbanize.com/es/recursos-de-kanban/primeros-pasos/que-es-kanban
- Jubert de Almeida Bernardo, Ferreira Neves Rui, Horta Nuno [2016] Combining Support Vector Machine with Genetic Algorithms to optimize investments in Forex markets with high leverage. Recuperado de: https://www.sciencedirect.com/science/article/abs/pii/S1568494618300036
- scikit – learn [2020] scikit – learn (0.24) [Software – Librería] recuperado de: https://scikit-learn.org/stable/
- [2020] Talib (0.4) [Librería open source] recuperado de: http://mrjbq7.github.io/ta-lib/
- Kaggle [2020] Jane Street Market Prediction recuperado de: https://www.kaggle.com/c/jane-street-market-prediction