PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

Desarrollo de un modelo para la predicción del precio del cobre empleando herramientas de Machine Learning

Tesis para obtener el título profesional de Ingeniera Industrial

AUTORA:

Almudena Fosca Gamarra

ASESOR:

Oscar Enrique Miranda Castillo

Lima, Diciembre, 2020

Resumen

A lo largo del presente trabajo de investigación se exploró el uso de herramientas de inteligencia artificial *(Machine Learning)* en la predicción del precio de cobre. Este proyecto de investigación se desarrolla dentro de las actividades del Grupo de investigación en finanzas aplicadas (GIFA) de la PUCP. En el cual, a partir de conocimientos interdisciplinarios se busca explotar metodologías de Inteligencia Artificial aplicando *Machine Learning* en el ámbito de inversión financiera.

En el capítulo 1, se exponen trabajos previos para el pronóstico de acciones, índices bursátiles y *commodities*, pudiendo comparar y contrastar los resultados obtenidos al aplicar diversos algoritmos. De esta forma, se emplean los estudios previos presentados como base para la ejecución y selección del modelo en esta tesis. Asimismo, de manera más detallada se presentan los factores más importantes en el comportamiento del precio de un *commodity.*

En el capítulo 2, se estructura la metodología a emplear en el desarrollo de la investigación. Se especifica el tipo de investigación y diseño, así como las métricas de evaluación a emplear.

El tercer capítulo corresponde al modelo de predicción basado en herramientas estadísticas tradicionales. Se presenta la metodología Box Jenkins como punto de partida para la ejecución del modelo ARIMA, posteriormente se evalúan los resultados obtenidos con este.

A partir del cuarto capítulo se introducen los conceptos de *Machine Learning.* Inicialmente se presenta un flujograma base para la elaboración de un algoritmo, y con este se estructuran dos modelos: regresión lineal y SVR. A lo largo de este capítulo se construyen ambos algoritmos de manera básica, desde la categorización del problema hasta la validación, según el flujo de procesos presentado.

El quinto capítulo tiene como objetivo evaluar la importancia de realizar un análisis de selección de atributos con el fin de mejorar el modelo. De esta forma, se utilizan dos algoritmos de selección y posteriormente se comparan los resultados obtenidos. El sexto y último capítulo del cuerpo de esta investigación se centra en optimizar el modelo SVR a través de la implementación de algoritmos de selección de hiperparámetros.

TABLA DE CONTENIDOS

Índice de Tablas

Índice de Figuras

Introducción

La motivación para la elección de este commodity es que el Perú es el segundo mayor productor de cobre del mundo. La producción nacional de este mineral ha crecido sustancialmente, pasando de 1.27 millones de toneladas en el 2008 a 2.46 millones en el 2019 significando un incremento del 97% en tal periodo[1].Sin lugar a duda, la explotación y manejo de este producto tiene un efecto directo en el desarrollo de la economía nacional. Por otro lado, el cobre es uno de metales con mayor aplicación en diversas industrias a nivel mundial, un pronóstico adecuado de su precio tiene gran valor tanto para las empresas, quienes se beneficiarían de la capacidad de evaluar proyectos y negocios estratégicos adecuadamente, así como para las entidades del estado que al conocer la fluctuación del precio de este mineral podrían anticiparse frente a situaciones de riesgo y disminuir las pérdidas causadas por la volatilidad de este.

En cuanto al enfoque del estudio, en la actualidad, es cada vez mayor el interés de la aplicación Inteligencia Artificial en diversas áreas de conocimiento, uno de ellos, el ámbito financiero. Se ha demostrado, que al emplear aprendizaje automático se pueden realizar operaciones a mayor escala y de forma más rápida, pues es posible procesar grandes volúmenes de datos para solucionar un problema específico y debido a la importancia que tienen los datos financieros históricos en el modelamiento de pronósticos, los algoritmos de *Machine Learning* resultan ser muy atractivos[2].

Conociendo el valor de la capacidad de pronóstico del precio del cobre, en este trabajo se plantea emplear herramientas de inteligencia artificial para obtener un modelo más preciso y eficiente.
Si bien, se han realizado modelos estadísticos, y también algoritmos complejos para el modelamiento de este problema, el objetivo primordial de esta tesis es construir desde un nivel básico un algoritmo simple que utilice como variables de entrada factores analizados y estudiados en estudios previos.

---

[1] Instituto de ingenieros de minas del Perú. http://www.iimp.org.pe/actualidad/produccion-de-cobre-en-peru-crecio-97-en-periodo-2008-2019

[2] How Big Data and AI are Accelerating Business Transformation. Big Data and AI Executive Survey 2019._https://newvantage.com/wp-content/uploads/2018/12/Big-Data-Executive-Survey-2019-Findings-Updated-010219-1.pdf

1 Capítulo I: Marco Teórico:

El objetivo de este capítulo es proveer de un fundamento académico al desarrollo de la presente tesis de investigación, Se brindará una visión acerca de estudios realizados hasta el momento en el campo de investigación propuesto. Asimismo, se brindarán las definiciones clave del marco de investigación.

1.1 Big Data, Machine Learning y Data Analytics en la actualidad

En primer lugar, es importante hacer una distinción entre las definiciones que se abordarán a lo largo de la investigación:

En el contexto de la era de la información, Shneiderman (2008) describe un conjunto de datos como *Big Data* cuando es demasiado grande para caber en una pantalla. En pocas palabras, cuando la información no puede ser procesada de una forma directa debido a la cantidad de elementos presentes, estamos hablando de "Big Data".

Por otro lado, "Machine Learning" es una de las ramas de los algoritmos computacionales diseñados para emular la inteligencia humana en el conocimiento y entendimiento del entorno.

Finalmente, el término *analytics* a menudo se usa ampliamente para señalar metodologías o procedimientos que permitan la toma de decisiones basada en datos. En el mundo corporativo, un equipo de análisis. usa su experiencia en estadística, *big data and data mining*, *machine learning*, y *visualization* para solucionar cuestiones que se plantean sus líderes.

La interrelación de estos tres componentes junto con otros elementos de análisis y con el uso de herramientas informáticas posibilitan la evolución de la ciencia de datos.

Hoy en día la información es tal vez el activo más preciado de las compañías, quienes la recolectan de forma "cruda" (es decir en forma de datos, la mayoría desestructurados) y procesan de acuerdo con los requerimientos de cada organización, con el fin de utilizarla para los procesos de toma de decisiones.

La inspección, limpieza, transformación y modelado de datos es el objetivo de los analistas de datos, quienes después de estudiar la data e interpretar los resultados tienen la posibilidad de brindar una recomendación o ejecutar una acción determinada.

1.1.1 Aplicaciones en el sector financiero

New Vantage publicó un informe titulado *"Big Data and AI Executive Survey 2019*" en el cual brinda una visión de cómo la ciencia de datos y la inteligencia artificial están acelerando la transformación de los negocios. En él exponen que el 97.2% de las empresas encuestadas están invirtiendo en Big Data e iniciativas de inteligencia artificial. De las 65 empresas participantes, la mayoría (el 74%) brinda servicios financieros.

Las empresas pertenecientes a este sector se han visto obligadas a implementar capacidades analíticas basadas en datos para aumentar el crecimiento, rentabilidad, optimizar procesos y reducir costos, minimizar riesgos, así como facilitar la regulación de cumplimientos.
Existen cuatro frentes en los que la inversión en *"Data Science*" puede beneficiar al sector financiero, estos son enfocados en: el cliente, las operaciones, el manejo y construcción de portafolios, y la regularización de cumplimientos.
Para los objetivos de esta investigación se desarrollarán los casos de uso relacionados al enfoque operativo, como la optimización de capital y el análisis de impacto o reacción de mercado.
Gracias a una gran variedad de desarrollos tecnológicos en finanzas el beneficio de la ciencia de datos en estas áreas está siendo cada vez mayor, por ejemplo, la proliferación de plataformas de *trading* electrónico ha sido acompañada por un aumento en la disponibilidad de datos de mercado en formatos estructurados, logrando un incremento en la participación del mercado. Asimismo, la disponibilidad de acceso a data del mercado y protocolos de trading ayuda a los participantes del mercado a descubrir o pronosticar el precio y obtener una rentabilidad de las transacciones. (IOSCO, 2017, p. 12).

El caso de uso clave para el desarrollo de esta investigación es básicamente la identificación de ciertos factores o indicadores del mercado que pueden ser distinguidos como *"Trading Signals".* El aprendizaje automático puede ayudar a las empresas a aumentar la productividad y a reducir los costos escaneando rápidamente y tomando decisiones basadas en más fuentes de información y en información oculta que puede perderse a los ojos de cualquier experto. (FSB, 2017, p.11)
La clave entonces es poder identificar señales en la data, de la cual se puede hacer predicciones relacionadas al precio o al nivel de volatilidad sobre distintos horizontes de inversión. Queda entonces como objetivo principal identificar los factores determinantes de un "estado positivo" para invertir rentablemente.

El estudio del comportamiento de activos en el mercado financiero tiene dos bases, el análisis fundamental, bajo el cual se estudian factores que pueden afectar el valor de un activo, desde variables

macroeconómicas hasta variables internas de la empresa, y el análisis técnico en el cual se usan datos de comportamiento pasado para predecir el movimiento del precio de los activos en el futuro.

Durante mucho tiempo, se creía que los cambios en los precios de las acciones no eran previsibles. La conocida hipótesis de "Random Walk" (Malkiel y Fama, 1970; Malkiel, 2003) y la Hipótesis del mercado eficiente (Jensen, 1978), establecen que un mercado es eficiente con respecto a un conjunto de información actual si es imposible hacer ganancias en este mercado. Sin embargo, a principios del siglo XXI, algunos economistas indicaron que los precios futuros de las acciones son al menos parcialmente predecibles. Por lo tanto, se han explorado muchos algoritmos de predicción y han demostrado que el comportamiento del precio de las acciones puede predecirse. (Malkiel, 2003).
No obstante, han habido varios estudios relacionados al modelado de algoritmos para el pronóstico de tendencias de comportamiento de acciones, bonos, índices bursátiles, entre otros.

1.1.1.1 Precios de las acciones

Seyed Enayatolah Alavi, Hasanali Sinaei, Elham Afsharirad (2015), desarrollaron un estudio para la predicción de la tendencia de precios de acciones usando técnicas de "Machine Learning". En este estudio se utilizan diez años de datos del índice total de precios de acciones del banco Tejarat de Irán desde 2002 hasta 2012, el objetivo del experimento fue comparar el rendimiento y la precisión entre *Support Vector Machines*, *Random Forest* y *K nearest neighbour* (máquinas de vectores, bosques aleatorios y vecinos más cercanos). Los resultados del estudio mostraron que la mejor precisión promedio y el estimador F se obtienen del uso de *Random Forest*, seguido de *SVN* y finalmente de *KNN.* Concluyeron, además, que esto se debió a que *Random Forest* es un método de *clasificación conjunta* mientras que SVM and KNN son clasificadores simples, logrando ser una gran alternativa para un conjunto de datos no estacionarios con la presencia de valor atípicos y ruido.

Por otro lado, Muhammad Waqar, Hassan Dawood,Muhammad Bilal Shahnawaz, Mustansar Ali Ghazanfar (2017), plantearon el desarrollo de un algoritmo para predecir el mercado de valores por análisis de componentes principales. La técnica de componentes principales (PCA), es muy efectiva para reducir la dimensionalidad de la data.
Esta técnica estadística se basa principalmente en el análisis de componentes principales, para enfrentar el problema presente en la recolección de información de una muestra de datos. Se sabe que para lograr encontrar representatividad en los estadísticos obtenidos se puede tender a tomar el mayor número posible de variables, sin embargo, al incrementar el número de variables también se incrementa el número de coeficientes de correlación, dificultando la visualización de relaciones entre estas El objetivo

del PCA es reducir la dimensionalidad existente en un *data set*, extrayendo toda la información perteneciente a este, pero con pocas variables no correlacionadas entre sí.
Las nuevas variables son combinaciones lineales de las anteriores y se van construyendo según el orden de importancia en cuanto a la variabilidad total que recogen de la muestra.

En su estudio investigaron el efecto de la aplicación de PCA en tres conjuntos de datos del mercado de valores (New York Stock Exchange, London Stock Exchange, Y Karachi Stock Exchange) y analizaron la precisión relativa del modelo de clasificación. Los resultados del estudio demostraron que el uso de PCA logra reducir la redundancia de los datos, lo que resulta en la reducción de datos altamente correlacionados manteniendo únicamente aquellos necesarios para explicar el 99% de la varianza. Logrando obtener un modelo con mayor precisión. No obstante, para uno de los índices se encontró que la precisión disminuyó y esto puede explicarse por el hecho de que en algunos casos la aplicación de PCA da como resultado la pérdida de información crítica que podría ser importante para la clasificación y causa una disminución en la precisión del modelo.
De los resultados se puede concluir entonces que, la selección adecuada de componentes principales apropiados es vital para mejorar la utilidad de PCA.

Suryoday Basaka, Saibal Karb,c, Snehanshu Sahaa, Luckyson Khaidema, y Sudeepa Roy Deya (2019), hicieron un modelo para pronosticar la dirección de los precios de las acciones usando clasificadores de árbol. Los árboles de decisión son una herramienta de clasificación utilizada en machine learning con el objetivo de crear un modelo que predice el valor de una variable de destino en función de diversas variables de entrada. En este tipo de estructuras, las “hojas” representan etiquetas de cada clase y las “ramas” las características que se atribuyen a cada clase.

El proceso de navegación inicia desde la raíz del árbol hasta cada hoja, de acuerdo con el resultado de las pruebas a lo largo del camino.
La Figura 1.1 describe un árbol de decisión que razones por las cuales un cliente potencial responderá o no a un envío directo. Los nodos internos se representan como círculos, mientras que las hojas se denotan como triángulos. Dado este clasificador, el analista puede predecir la respuesta de un potencial cliente y comprender las características de comportamiento de toda la población de clientes potenciales con respecto al correo directo. (Rokach, Maimon. 2007)

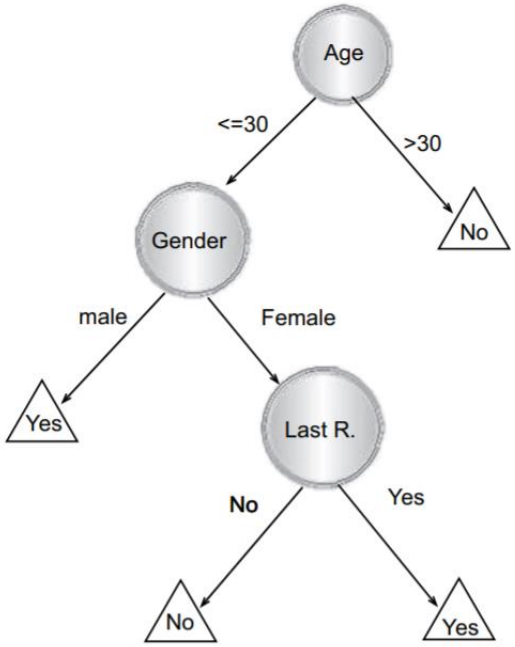

Figura 1.1: Árbol de decisión representativo de la respuesta de un cliente al mail directo
Elaboración: (Rokach, Maimon 2007: 2, figura 1)

Las técnicas más usadas basadas en la lógica del proceso de decisión del árbol de decisiones son métodos conjuntos híbridos entre árboles de clasificación y regresión, como *gradient boosted descision trees (GBDT)* y *Random Forest* (RF), que consiste en una gran cantidad de árboles de decisión individuales que operan como un conjunto. Cada árbol individual en el bosque aleatorio emite una predicción de clase y la clase con más "votos" se convierte en la predicción del modelo. El uso de esta metodología permite explotar los beneficios de la lógica de un árbol de decisión pues una gran cantidad de modelos (árboles) relativamente no correlacionados que operan en conjunto superará a cualquiera de los modelos constituyentes individuales (árbol de decisión evaluado individualmente).

La razón de este efecto es que los árboles se protegen entre sí de sus errores individuales (siempre que no se equivoquen constantemente en la misma dirección). Si bien algunos árboles pueden estar equivocados, muchos otros árboles estarán en lo correcto, por lo que, como grupo, los árboles pueden predecir en la dirección correcta. (Yiu, 2019)

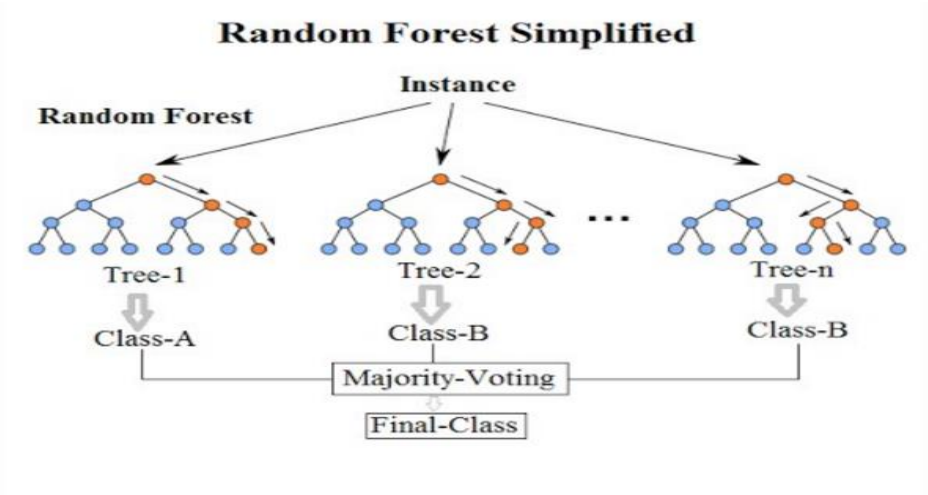

Figura 1.2: Estructura del algoritmo Random Forest

Elaboración: (Koheresen 2019: 1, figura 1)

Por otro lado, los modelos GBDT están enfocados en un proceso llamado *"Boosting",* este es un método de ensamble ponderado. Cada uno de los algoritmos base se agrega secuencialmente, uno por uno. Una serie de clasificadores N aprendidos iterativamente. Los pesos se actualizan para permitir que los clasificadores posteriores "presten más atención" a las tuplas de entrenamiento que fueron clasificadas erróneamente por el clasificador anterior. *Gradient Boosting*, es conocido por ser el algoritmo líder en métodos de ensamble, utiliza el método de gradiente descendiente (algoritmo de optimización iterativa para encontrar el mínimo de una función) para optimizar la función de pérdida (o error).

La metodología del ensayo de Suryoday Basak et al (2019) fue comparar ambos algoritmos *c*on el fin de encontrar y discutir las ventajas de estos dos sobre técnicas de análisis no ensambladas.

Las técnicas de análisis no ensambladas son llamadas también algoritmos de clasificación lineal pues utilizan técnicas de toma de decisión basadas en el valor de una combinación lineal de sus características, es decir se asume que existe un correlación lineal entre los atributos de las variables de la data.

En cuanto a los algoritmos de ensamble, son aquellos que utilizan diversos métodos base para generar un modelo óptimo, y no necesariamente asumen linealidad de datos. Cómo se mencionó, la aleatoriedad de los RF ayuda a hacer que el modelo sea más robusto que un solo árbol de decisión y que sea menos probable que se sobreajuste en los datos de entrenamiento. En comparación, los modelos GBDT se basan en construir un árbol a la vez, donde cada nuevo ayuda a corregir los errores cometidos por un

árbol previamente entrenado logrando aproximar los regresores a las muestras de entrenamiento y encontrando la mejor división para estos.
En el experimento se utilizaron anchos de ventana *(trading window)* variante para cada corrida, entre (3,5,10,15,30,60 y 90). Este es el período en el que una compañía permite a sus ejecutivos y empleados clave transar sus acciones.

Es importante enfatizar que la diversidad de los antecedentes de las compañías elegidas para el análisis de los precios de las acciones es crucial para garantizar la eficacia de los algoritmos.
Los resultados demuestran que para ambas técnicas la precisión y el valor del estadístico F aumentan a medida que se incrementa el ancho de la ventana de trading. Más aún, la capacidad de la clasificación observada para cada caso al usar GBDT es comparable con el de RF. Por otro lado, al usar algoritmos de clasificación lineal se consiguió como máximo una precisión de 55.65%. En Conclusión, las metodologías de análisis ensamblado tienen un mejor performance en comparación a clasificadores lineales en el pronóstico del comportamiento de acciones. Además de eso, parte importante de este trabajo es la selección de indicadores técnicos y la aplicación de estos como características determinantes en el precio de las acciones. Dado que el trasfondo de esta selección es de análisis financiero, los autores recomiendan tener flexibilidad para la elección de las características, pues la interpretación de cada una depende del contexto y del mercado que se está estudiando.

Las redes neuronales artificiales (ANN) (Kimoto, Asakawa, Yoda y Takeoka, 1990; Kohara, Ishikawa, Fukuhara y Nakamura, 1997) se han empleado para lograr buenas predicciones incluso en el caso de relaciones complejas de variables.

Las redes neuronales se basan en redes de múltiples capas (nodos azules y rojos en la figura 1.3) utilizadas para clasificar variables, hacer predicciones, entre otras cosas. Existen tres partes fundamentales en una red neuronal: una capa de entrada, con unidades que representan los campos de entrada; una o varias capas ocultas; y una capa de salida. Las unidades se conectan por ponderaciones o probabilidades.

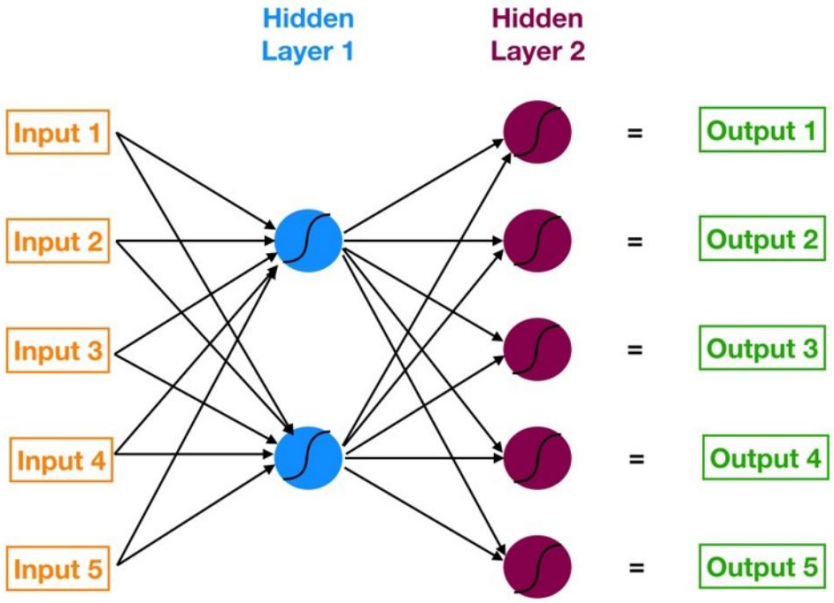

Figura 1.3: Estructura de una red neuronal

Elaboración: (Yiu 2019: 1, figura 2)

En un inicio, las ponderaciones son aleatorias, a medida que la red aprende examinando registros individuales, genera una predicción para cada registro y realiza ajustes a las ponderaciones iniciales dependiendo de la verdad o falsedad de cada predicción. El proceso se realiza indefinidamente hasta lograr alcanzar un criterio determinado.

Se les denomina caja negra, dado que, si bien puede abordar cualquier función, estudiar su estructura no brindará ninguna idea sobre la estructura de la función que se modela. Incluso la red neuronal más simple con una sola capa oculta es difícil de entender.

Existen redes neuronales de clases múltiples, una de ellas es la red neuronal del tipo *one-againstall (OAA-NN).* Proporciona una manera de aprovechar la clasificación binaria, en este tipo de red neuronal se clasifica bajo un clasificador binario para cada resultado posible. Se dice que este enfoque es eficiente cuando la cantidad total de clases es pequeña. Por último, las redes neuronales del tipo *one-against-one (OAO-NN)*

Boonpeng y Jeatrakul (2016) implementaron una red neuronal *one-againstall* (OAA-NN) y *one-against-one* (OAO-NN) para clasificar los datos de compra, retención o venta de acciones y compararon su desempeño con una red neuronal tradicional. Se encontró que las redes OAA-NN tenían mejor

performance que los modelos tradicionales neuronales, produciendo en promedio una exactitud de 72.5%.

1.1.1.2 Índices Bursátiles

Luciana S. Malagrino, Norton T. Roman, Ana M. Monteiro, (2018), llevaron a cabo el primer estudio de *machine learning* que tomaba en consideración la dependencia de índices bursátiles alrededor del mundo. El objetivo de su investigación fue pronosticar la dirección diaria del índice bursátil utilizando como metodología una Red Bayesiana (BN).

Las redes Bayesianas son una herramienta muy importante para comprender la dependencia entre los eventos y asignarles probabilidades, y con ello determinar cuán probable y cuál es el cambio de ocurrencia de un evento dado al otro. Bajo el fundamento teórico del teorema de Bayes, se utilizan probabilidades condicionales y la regla de la cadena para conseguir la distribución conjunta, es decir la probabilidad del evento final considerando todos los eventos dependientes.
En la figura 1.4 se puede visualizar un ejemplo de una red bayesiana para determinar la probabilidad de tardanza de una persona teniendo como inputs la probabilidad de situaciones como: el bus estuvo tarde y la alarma estuvo apagada.

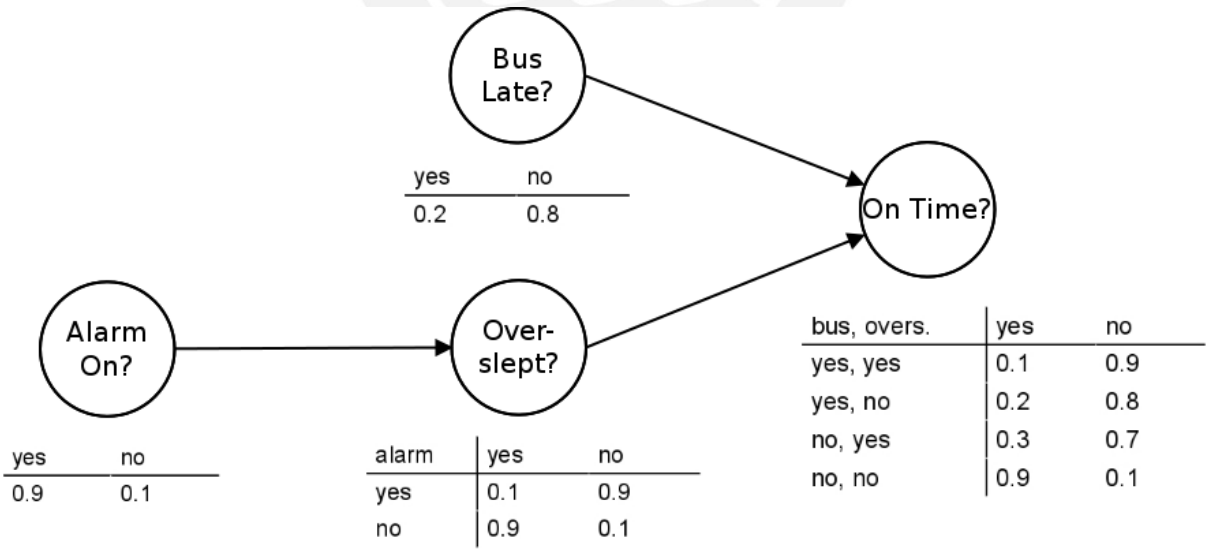

Figura 1.4: Estructura de una red Bayesiana
Elaboración:(University of Bergen 2019:1, figura 1)

En el estudio realizado por Luciana S. et al (2018), se centraron en la eficiencia de las BN para lidiar con data continua y discreta, lo que las hace atractivas en la estimación del valor del precio como en el pronóstico de dirección del precio de cierre. La hipótesis de estudio estuvo basada en que los índices se comportan bajo un enfoque "*follow the sun*"[1], bajo el supuesto de que, siguiendo la secuencia de los tiempos de cierre en los mercados de todo el mundo, cada mercado cerrado influye en el siguiente.
Frente a esta hipótesis, tales redes tienen la ventaja de dar una explicación legible de las dependencias entre los mercados de valores, por lo que uno puede comprender fácilmente cuán fuertes son las dependencias al verificar la probabilidad de un cambio en el mercado dado que otro ha cambiado.

Se usaron dos metodologías, la primera tomó en cuenta las direcciones de cierre dentro de un periodo de 24 horas y otra usando un periodo de 48 horas. Agruparon los mercados bursátiles por continente e hicieron pruebas con conjuntos de 1 a 3 índices de mercado por continente. Al hacerlo, su objetivo principal era probar la viabilidad de las BN para: Identificar los mercados y, más específicamente, los índices de mercado que más influyen en iBOVESPA (índice bursátil brasileño y el elegido como objeto de estudio), y también para distinguir qué ventana de tiempo otorga el mejor análisis. Hicieron la asunción de que todos los índices dentro de un mismo continente eran independientes a la hora de la predicción, y que todos los índices de un mercado son influenciados por el índice cerrará justo antes que ellos. Es por ello que se eligen tres nodos padres sobre iBOVESPA los cuales vendrían a ser los índices bursátiles de Asia, Europa y América.

Para el experimento con 48h se sigue la misma metodología con la diferencia de que en vez de usar 3 nodos padre, se usa uno más que vendría a ser la dirección de cierre de iBOVESPA el día anterior.
Como resultado, para el caso de las 24h, se observa que la precisión media obtenida al usar un sólo índice por continente es mayor (71.08%), que las obtenidas al usar dos índices (70.20%) o tres índices (56.45%). Lo mismo ocurre con la tipología de 48h, en las que con un sólo índice se obtiene 68.17%, con dos índices 31.04% y por último con tres índices 8.21%. Frente a estos resultados, se puede deducir que cuando se añaden más índices no solo se añade más ruido a la data, sino que el modelo falla en sus predicciones a tal punto que sería mejor negar el output obtenido para que la precisión aumente.
Por otro lado, el hecho de que las 48h tuvieron peor performance que las 24h puede ser un indicador de que la reducción en la fuerza de influencia de un mercado en iBOVESPA a lo largo del tiempo se reduce, haciendo que estos mercados se comporten más como ruido conforme retrocedemos el tiempo desde el día de predicción.

[1] Según su definición básica, "seguir al sol" significa que el efecto de un índice literalmente sigue al sol: es un tipo de flujo de efecto global en el que los problemas pueden ser manejados y pasados entre diferentes zonas horarias.

En conclusión, con este modelo no solo se puede determinar las influencias en el comportamiento de un mercado sino también cuantificarlas y verificar la tasa de decrecimiento a lo largo del tiempo. Asimismo, la importancia de este es la utilidad para distinguir a los mercados candidatos a contaminarse más a causa de una crisis, en este caso los mercados más influenciables serán rankeados como más probables a pasar el efecto de la crisis en el mercado objeto de estudio. Bajo el supuesto inicial, de que el capital se mueve a los lugares donde el retorno a la inversión será mayor, se puede utilizar este modelo como herramienta de decisión para la inversión en un mercado específico.
Por último, las Redes Bayesianas son una metodología mucho más amigable y entendible para los usuarios, comparado con otros modelos tipo *black box*, como las ANNs o SVMs.

Parag C. Pendharkat, Patrick Cusatis (2018), realizaron una investigación para modelar índices financieros con *reinforcement learning agents,* un tipo de herramienta de machine learning.
El aprendizaje por refuerzo es un área de *Machine Learning.* Que trata de tomar medidas adecuadas para maximizar la recompensa en una situación particular, a diferencia del *Supervised Learning* este no se procesa con la respuesta correcta dentro del training data, sino más bien, un agente decide que acción tomar de acuerdo con los datos presentados, y en consecuencia a falta de un *training data*, aprende paulatinamente de acuerdo con el *feedback* que brinde el output del modelo.

Cusatis et al, se basaron en la dificultad de predecir el comportamiento financiero utilizando el modelo tradicional de toma de decisiones según la teoría de Markowitz (MDP). Dado que en la realidad ningún mercado financiero se comporta como un modelo Markovitziano, plantean abordar el problema al dividir el modelado en dos partes, en primer lugar el diseño del algoritmo basado para aprender, el modelo de la función de ganancia y las transiciones y comportamiento de los indicadores financieros, el segundo paso del MDP se encargaría de aprender el mapeo del proceso de decisiones basado en la prueba y error del modelo diseñado en la primera parte (aprendizaje reforzado).
Se aborda el diseño del algoritmo para modelar índices financieros como si fuera un problema del tipo *multi-armed bandit problem*[2] en el cual se desconoce la distribución de ganancia atada a cada decisión del inversor.
En este estudio se utiliza dos portafolios distintos compuestos por 2 activos financieros. El primer porfolio consistió en retornos anuales de S&P y AGG, y el segundo portafolio consistió en 3 escenarios distintos de retornos trimestrales, semi anuales, y anuales para S&P y 10 years TN. Los investigadores distinguieron a los agentes en 3: *static knowledge agent*, *continuous learning agent y adaptive*

[2] En la teoría de la probabilidad, los problemas llamados "multi-armed bandit problem" son aquellos en los que cada persona debe elegir entre múltiples acciones para maximizar su ganancia u obtener el resultado más rentable, teniendo en cuenta que cada acción tiene un resultado desconocido. Es un problema muy común en Casinos, al comienzo del experimento, cuando se desconocen las probabilidades y los pagos, el jugador debe determinar qué máquina tirar, en qué orden y cuántas veces. (Optipedia, 2019).

continuous knowledge agent. Cada uno con una función de aprendizaje distinta que se diferenciaba por la información utilizada en el proceso de estudio del training *data set*. Al analizar el impacto de la frecuencia de aprendizaje y como los agentes reaccionaron a la información brindada se obtuvo que los agentes de aprendizaje discreto o estático eran en general incapaces de utilizar todos los beneficios de una mayor frecuencia de aprendizaje, y que los agentes de aprendizaje continuo y adaptativos siempre conseguían mejor performance.

1.1.1.3 Commodities

Un *commodity* es un bien o producto, especialmente agrícola o minero, que puede ser procesado y revendido. Se comercializan en grandes cantidades en todo el mundo. Dependemos de ellos para las necesidades básicas de una vida cotidiana: la electricidad, alimentos, ropa, y transporte.
Antes de ahondar en el entendimiento del comportamiento de un *commodity* en el mercado financiero, es importante destacar su naturaleza física. Fundamentalmente, estos son productos creados por fuerzas naturales. Eso tiene ciertas implicaciones. En primer lugar, es que cada producto es único: su forma química depende exactamente de cuándo y dónde se originó, es decir no se pueden estandarizar en forma perfecta, como en el caso de productos manufacturados.
En general, podemos definirlos como "todo bien que es producido en masa por el hombre o del cual existen enormes cantidades disponibles en la naturaleza, que tiene valor o utilidad y un muy bajo nivel de diferenciación o especialización" (Castelo, 2003).
Para ser comercializables, los *commodities* deben ser puestos en una forma utilizable y trasladados a donde puedan ser usados, en el momento en que se necesiten. Esta relación, entre espacio, tiempo y forma, es un factor clave para entender el negocio.

Existen dos formas de clasificar los *commodities*, en primer lugar, los primarios son aquellos que se extraen o capturan de forma directa de la naturaleza, en minas, granjas y pozos, sus cualidades y características varían en gran medida. Por otro lado, los *commodities* secundarios son obtenidos a partir de los primarios para satisfacer una necesidad específica del mercado. Por ejemplo, el petróleo crudo se refina para producir gasolina y otros combustibles; Los concentrados se funden para producir metales. En este caso, puede haber variaciones menores en la calidad dependiendo de cómo se producen.
Otra forma de clasificar estos productos es por su naturaleza. Los *commodities* energéticos, son los más importantes a nivel mundial, gracias a la preponderancia del petróleo. En cuanto a los productos agrícolas, las principales categorías incluyen granos y oleaginosas (maíz, soja, avena, arroz, trigo), ganado (ganado, cerdos, aves de corral), lácteos (leche, mantequilla, suero), madera, textiles (algodón, lana) y softs (cacao, café, azúcar).

Por último, la categoría de metales y minerales, incluyen metales no ferrosos, metales preciosos y minerales.

En cuanto a la comercialización de estos, se realiza en dos tipos de mercado. En primer lugar, se encuentran los mercados al contado, el cual es un término que refiere a muchos sitios descentralizados en los que el producto puede ser vendido o comprado a un precio spot acordado.
Por otro lado, el segundo tipo de mercado son los mercados listados, también llamados Bolsas de materias primas tienen una modalidad operativa que comercializa los productos mediante instrumentos derivados cuyo activo subyacente es el *commodity*. La mayoría de ellos son comercializados bajo contratos futuros, en los cuales se establece un pacto en un tiempo t0. Por la venta de una cantidad determinada de productos (*x*) a un precio *y* para el tiempo t1. Las transacciones futuras representan una buena estrategia para ambas partes pues reducen el riesgo, los primeros contratos de este tipo fueron desarrollados por agricultores como una forma de reducir el riesgo de sus transacciones. Una razón para que el comercio futuro sea elegido por individuos, es la gran volatilidad que existe en el comportamiento de los precios de un *commodity*. (Yagüe, 2014).

Se han realizado investigaciones acerca de modelos para la predicción de estos precios, como el realizado por Manel Hamdi Chaker Aloui, en el cual utilizan redes neuronales para predecir el precio del petróleo crudo. Se basan en la ineficiencia de estudios anteriores que utilizaron técnicas lineales para el pronóstico y obtuvieron errores significativos, dado que a pesar de haber empleado varias variables exógenas para predecir el precio del petróleo (inventario, oferta y demanda, entre otros), la oferta y demanda son relativamente inelásticas a los cambios de precios, por lo tanto, un ajuste de inventario puede ser lento, lo que explica la mayor parte de la diferencia entre los precios reales y los pronosticados, especialmente para el corto plazo (Hamilton, 2008). Sabiendo entonces que, el mercado del petróleo crudo es el mercado de materias primas más volátil, plantean que pronosticar el precio del petróleo a través de modelos no lineales es la opción adecuada.

Por otro lado, Massimo Panella, Francesco Barcellona and Rita L. D'Ecclesia (2012), especificaron en su estudio que la proyección de los precios de los productos básicos a diario no se puede obtener fácilmente utilizando modelos estructurales estándar, dada la falta de datos diarios sobre la oferta y la demanda, normalmente disponibles mensualmente. Por ello, propusieron un aprendizaje computacional de tipo ANN utilizando un enfoque de estimación de máxima precisión para calibrar los parámetros. Las redes neuronales fueron aplicadas con éxito en el modelo, logrando describir la dinámica del mercado de valores y sus volatilidades.
Como se mencionó líneas arriba, siguiendo la analogía del agricultor, al transar su cosecha a un precio determinado incluso antes de plantarla, el agricultor no debía preocuparse por los futuros cambios de

precios, sino más bien su único riesgo era la producción real de su cosecha. Por lo tanto, la compensación de un agricultor dependería de su capacidad para producir un cultivo en lugar de las posibles fluctuaciones en el precio del producto que cosecha. Al agricultor, como a cualquier otra persona que busca reducir su riesgo, se le llama *hedger*.

Dónde, cómo y cuándo: Los fundamentos del *pricing* de commodities

Para entender el comportamiento de un c*ommodity* específico, es importante primero entender de manera general los factores macroeconómicos que intervienen en el *pricing* de estos bienes.

Como se ha enfatizado, los precios de los productos básicos tienen una gran importancia, por su potencial impacto en la producción agregada, en el equilibrio transaccional entre mercados y en la transmisión de perturbaciones del ciclo económico entre países, al conectar a los exportadores e importadores de *commodities* de países desarrollados con aquellos de los países en desarrollo. Asimismo, cambios en los niveles de precios de los productos básicos pueden crear presiones internacionales sobre la economía de un país que podrían dificultar la ejecución de las políticas monetarias en este. Si la producción de estos *commodities* constituye un porcentaje representativo del producto agregado, entonces la volatilidad de sus precios debe tenerse en cuenta en el diseño de las políticas económicas dentro de los planes de un país. (Borenzstein y Reinhart, 1994).

Identificar el comportamiento de los *commodities* permite diversificar el riesgo internacional no solo para las autoridades monetarias. Por ejemplo, si los agentes económicos en un país exportador de productos básicos supieran qué productos probablemente experimentarán aumentos de precios y el grado en que el comportamiento de estos se correlaciona pueden diversificar parte del riesgo al expandir el portafolio de productos exportadores en los que invierten actualmente. Cashin et al (1999), destaca la importancia de este análisis, pues menciona que la diversificación mediante el comercio de productos que tienen vínculos débiles y no comparten ciclos de comportamiento comunes puede disminuir sustancialmente el riesgo y el impacto de la fluctuación del precio de un único producto en la economía nacional.

Existen dos puntos de vista, según Kurgman (2008), para explicar las razones que subyacen en los movimientos de los precios de un *commodity.* En primer lugar, el crecimiento acelerado de los niveles de vida de China y países de Asia emergente que poseen una alta elasticidad ingreso de la demanda de *commodities.* Dentro de este punto de vista, el valor del dólar ocupa una responsabilidad central. Además, las condiciones monetarias laxas, y el posible exceso de liquidez internacional por la

volatilidad del dólar han ocasionado que se sumen presiones inflacionarias en mercados de precios flexibles, como en el caso de los *commodities.*

El segundo punto de vista, se enfoca en la especulación como causal de las fluctuaciones de las cotizaciones de estos productos, específicamente enfatiza la relevancia del fenómeno conocido como "financiarización de los *commodities*", el cual se profundizará en el siguiente capítulo, esto debido a que, en la última década, se ha observado un incremento consistente del precio de los *commodities* que coincide con el volumen incrementado de transacciones e individuos en el mercado de instrumentos financieros relacionados a estos.

Si bien es cierto ambos puntos de vista están interrelacionados en gran medida tienen un impacto en el comportamiento a corto plazo de los *commodities.* Es importante entonces, profundizar en un estudio detallado de las variables y factores que intervienen también en el equilibrio de largo plazo de los *commodities.*

1.2 El Cobre

El cobre ha desempeñado desde siempre un papel crítico en la civilización del hombre desde tiempo prehistóricos hasta la fecha. La importancia de este metal no ferroso se basa en su alta conductividad eléctrica y térmica. Aproximadamente la mitad del cobre producido en el mundo se utiliza en aplicaciones eléctricas en distintas industrias, así como para la construcción de edificios, equipos de transporte, productos de consumo, y maquinarias y equipos industriales.

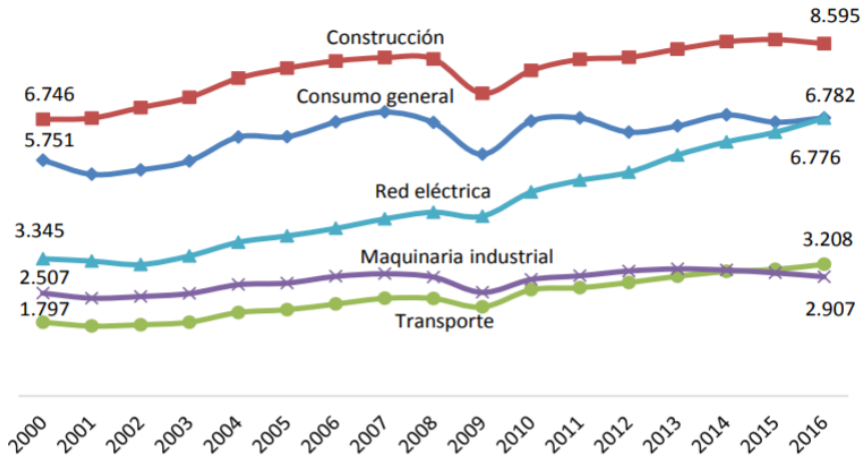

Figura 1.5: Distribución del consumo de cobre según el tipo de mercado

Elaboración: (Wood Mackenzie 2019: 1, figura 1)

En la Figura 1.5, se puede evidenciar la distribución del consumo de cobre según el mercado. Aproximadamente el 31% de la producción mundial de cobre es utilizada en construcción, siendo China el principal consumidor de cobre refinado del mundo, país en el que este metal es crucial para las metas que se ha trazado el gobierno en la última década y que actualmente están en su etapa de desarrollo. Dentro de estos planes se encuentra la modernización entera del país, electrificación del sistema nacional, disminución del carbón y petróleo incentivando energías sostenibles, implementación de proyectos de trenes rápidos, entre otros.

El cobre en la historia

El precio del cobre ha sufrido diversos picos en la historia que vale la pena especificar, para poder encontrar los factores que han influido, y probablemente seguirán influyendo en el comportamiento de este.

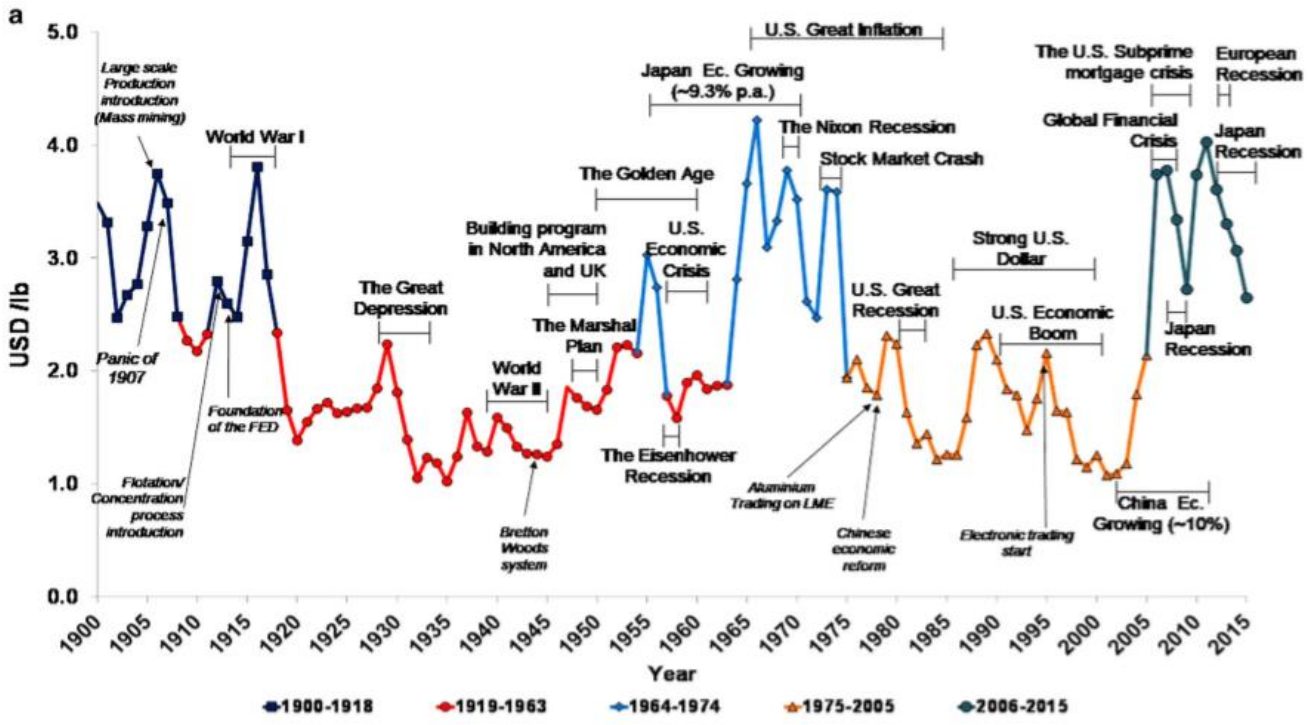

Figura 1.6: Evolución del precio del cobre entre 1900 y 2015. Los datos incluyen eventos económicos y financieros notables en el período.

Elaboración: (Cortez et al 2018: 4, figura 1)

Vale la pena ahondar en los hechos presentados en la figura 1.6 con el fin de comprender los picos del precio en función a los hechos contextuales contemporáneos.

En primer lugar, la industria del cobre reaccionó positivamente (alza de precios) a las épocas de guerra, esto debido a que las máquinas de ataque requerían grandes cantidades de este metal para los cartuchos, balas y piezas de armas. Cómo se puede observar entre los años 1914-1919, el precio del cobre alcanzó el que sería su pico máximo en los siguientes 50 años.

A diferencia de la Primera Guerra Mundial, los controles de precios se impusieron desde el comienzo de la Segunda Guerra Mundial y las transacciones de cobre no se reanudaron hasta agosto de 1953.

Por otro lado, entre los años 1929 y 1935 se registró una caída del precio del cobre, coincidente con la gran depresión americana. Asimismo, si se analiza la figura detenidamente, se puede distinguir un comportamiento similar a un *momentum* entre la época dorada estadounidense y la recesión del estado americano de 1960 durante el gobierno de Eisenhower. Esta tendencia es clara también en la recesión

del gobierno de Nixon, la gran recesión americana de 1980, y las crisis financieras en el mundo a partir del 2005.

En cuanto a la valorización del cobre en la historia, se identifican el periodo de crecimiento de Japón (aproximadamente 9% anual) durante el excluyendo[3] los años de recesión americana se registra una tasa de crecimiento anual compuesta (*CAGR*) de 9% para el precio del cobre. Lo mismo ocurre analizando el periodo de crecimiento de China en los inicios del 2000, obteniendo como resultado un *CAGR* 23% para el precio del cobre[4].

Del mismo gráfico se puede observar los rastros de bajo precio están conectados por períodos de transición de alto precio que describen un momento de bajo precio "estable" que se mueve a un estado de alto precio más "volátil" durante un período corto para luego regresar al estado de bajo precio (*momentum*). De esta figura entonces, se concluye que las disminuciones en el precio global del cobre coinciden con la inflación y las recesiones en la economía más grande del mundo, Estados Unidos, y con periodos de expansión de otros países.

Si bien, se ha realizado una breve comparación entre los hechos contextuales contemporáneos y el comportamiento del precio del cobre a lo largo de los años, se debe comprender que esta es una vaga y simple forma de identificar los factores influyentes en la variabilidad de este *commodity*. Dado que los hechos no son circunstancias apartadas, sino que en su mayoría afectan al mismo tiempo y, en direcciones y magnitudes distintas al precio del cobre.

En un intento por resumir los estudios sobre la dinámica de los precios de los productos básicos, Frankel y Rose (2009) enumeran tres teorías que explican el auge de estos productos en los últimos años. Las primeras dos teorías coinciden con los supuestos presentados por Kurgman.

- El primero es el "crecimiento de la demanda global", que se aceleró con la inclusión de países de alta demanda como China e India, causando los altos precios observados. Como se mencionó líneas arriba, se puede observar en la figura 1.6, que entre los años 1999 y 2010, la economía china presenció un crecimiento exponencial de aproximadamente 10% anual, el cual coincide con una tasa de crecimiento del precio del cobre. La participación de China en las importaciones de metales aumentó de menos del 10 por ciento en 2002 al 46 por ciento en 2014.
- La segunda teoría se centra en los mercados financieros y argumenta que la "especulación" fue la causa principal del auge de los productos básicos. Dada la existencia de mercados futuros, los participantes del mercado tienen la posibilidad de mantener sus bienes a largo plazo cuando

[3] Esto es, calculando el CAGR para el precio del cobre de 1965 a 1970.
[4] Calculado en base al 2003 como fecha inicial hasta el 2010.

se espera que el precio del *commodity* aumente, ejerciendo una presión aún más fuerte en el precio de este.

- La última teoría se centra en las bajas tasas de interés causantes de liquidez en el mercado, desencadenando que más individuos transfieran sus inversiones de fondos a contratos de *commodities,* incrementando la demanda de estos y por lo tanto elevando el precio.

Cortez et al, (2018), investigaron el comportamiento "*caótico"* del precio promedio anual del cobre entre los años 1900 y 2015, con el fin de examinar la dependencia del tiempo y de atractor extraño[5] mediante un análisis visual de series de tiempo. Utilizaron un *dataset* de precios anual con 116 observaciones, y confirmaron que este tamaño de muestra fue adecuado para distinguir el comportamiento caótico.

Concluyeron que las series de tiempo no expresaban ningún comportamiento periódico ni eran generadas por un proceso estocástico. Además, se observó la presencia de un agente extraño que describe un momento de bajo precio "estable", interrumpido después por varios años de aumentos de precios. Llaman a estas fluctuaciones "períodos de transición de precios" que deben abordarse como períodos de ajuste excepcionales en lugar de ciclos. Revelaron finalmente que las variables que impulsan los precios están relacionadas con el tiempo, evolucionan de una manera de causa y efecto, y los efectos se propagan con el tiempo. En su caso de estudio, se observó que las fluctuaciones de los precios de las materias primas minerales tienen efectos acumulativos a lo largo del tiempo, donde los cambios afectan no solo a los estados actuales sino también a los futuros que luego se convierten en el punto de partida para el próximo cambio de precios que describe una relación temporal continua. Este estudio es de gran valor para el entendimiento del comportamiento a largo plazo de los precios de *commodities.*

1.3 Factores determinantes en el comportamiento del precio de un commodity

En base a las teorías presentadas, y al gráfico presentado, se procede a distinguir los principales factores en la determinación del comportamiento del precio del cobre.

[5] Llámese a un factor exógeno

1.3.1 Tipo de cambio

Todos los trabajos de investigación utilizados en esta tesis concuerdan en el rol predominante del dólar, el cual es imposible de ignorar al estimar o modelar el comportamiento de todos los *commodities* en el mercado.

El dólar es el mecanismo de referencia universal para la fijación de precios de la mayoría de *commodities*, pues es la moneda de reserva de otros países alrededor del mundo, lo cual significa que lo mantienen como activo de reserva.

Si bien cada *commodity* tiene características idiosincráticas, históricamente, los precios de todos los productos básicos han tendido a caer cuando el dólar se fortalece frente a otras monedas, mientras que cuando el valor del dólar se debilita frente a otras monedas principales los precios generalmente tienden a aumentar.

Figura 1.7: Comparación del índice del dólar y el precio del cobre

Elaboración: (LME 2018: 1, figura 1)

1.3.2 Tasa de interés

La tasa de interés es un factor que impacta directamente en el proceso de toma de decisiones de los inversionistas, lo que puede derivar en el comportamiento del precio de *commodities* a corto plazo. En un entorno de bajas tasas de interés, el costo de financiar las existencias es menor que cuando las tasas de interés son altas. Es más barato almacenar los bienes a largo plazo; el costo de inventarios es un

término que los consumidores de productos básicos (y productores) usan para describir los costos asociados con la tenencia de inventarios durante un período.

Las tasas de interés están asociadas a las políticas monetarias de países desarrollados, una política monetaria *"fácil",* se traduce normalmente en tasas de interés reales bajas. Barsky y Kilian (2002, 2004) han argumentado que los altos precios del petróleo y otros productos básicos en la década de 1970[6] no fueron exógenos, sino más bien un resultado de la facilidad de la política monetaria.

A mayor costo de mantenimiento de inventarios, el mercado responde con una baja en la demanda y por lo tanto una disminución en el precio de los *commodities.*

Asimismo, se puede decir también que en situaciones donde las tasas de interés internacionales se ubican en niveles muy bajos, los inversores buscan en otros activos financieros, como los c*ommodities,* alternativas más rentables, elevando la demanda de estos. (Doporto, I, Michelena,G. 2011)

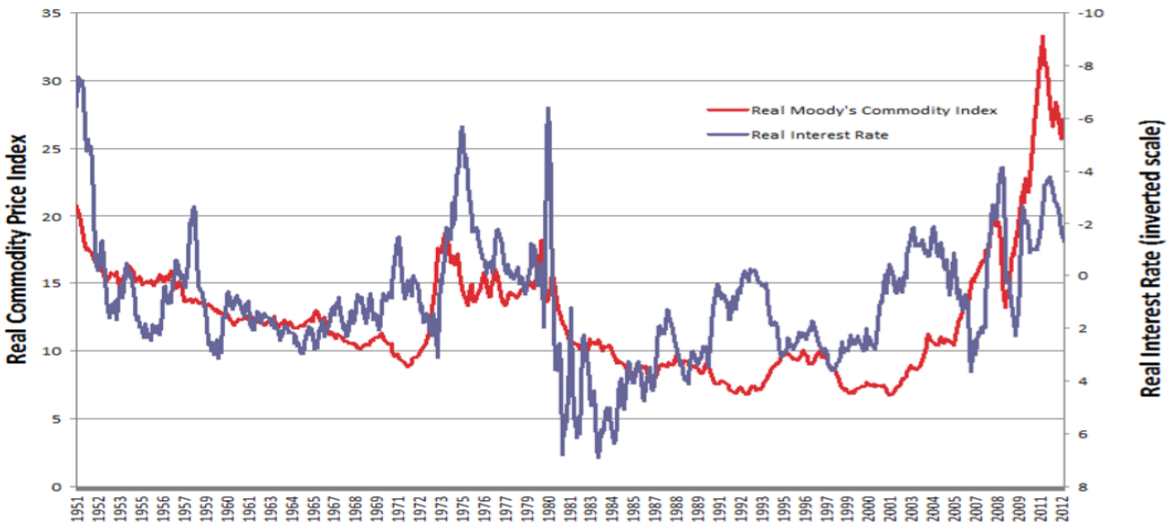

Figura 1.8: Comparación índice real precio de los commodities (Moody´s) y la tasa de interés (escala invertida)
Elaboración: (Frankel, J 2014: 4, figura 1)

1.3.3 Especulación

Podemos definir la especulación como la compra de los productos, ya sea en forma física o vía contratos negociados en una bolsa, en previsión de ganancias financieras al momento de la reventa. Esta incluye no solo la posibilidad de especulación desestabilizadora, que es la más común de considerar, sino también la posibilidad de la especulación estabilizadora (regresar el precio del *commodity* a su estado

[6] Ver figura 5

fundamental). El último caso es el fenómeno por el cual un aumento en el precio spot en relación con su equilibrio a largo plazo genera expectativas de una disminución de precios en el futuro, lo que lleva a los participantes del mercado a vender o vender en corto mercancía hoy y por lo tanto amortiguar el aumento de precios hoy. (Frankel, J 2014)

La explicación del fenómeno anterior se basa en la teoría financiera, según la cual existe especulación informada y especulación desinformada. En un mercado financiero, la primera deberá tener efectos en los precios, ya que esta es la forma en que la información privada se traduce en las transacciones del mercado. Mientras que, la especulación desinformada no debería tener tales efectos, o en casos de mercados poco líquidos no debería tener ningún efecto. Si un *trader* desinformado mueve los precios del mercado lejos de su equilibrio fundamental, *traders* informados *(arbitrageurs*), conocedores del verdadero valor, aprovecharán esta ventaja lo cual resultará en la estabilización del precio.

Si dividimos los efectos de la especulación en la transacción de *commodities* en ventajas y desventajas, obtenemos que: En los casos en los que los productores no necesariamente desean vender a un determinado precio, o en un determinado momento no coincidente con los consumidores, la presencia del *arbitrageur* hace que esta brecha sea eliminada, otorgando liquidez al mercado. Por otro lado, si no se tienen políticas reguladoras sobre el efecto de los especuladores en el precio de los *commodities*, estos pueden causar que las materias primas sean más caras para el comprador final, al estar presionando constantemente los precios.

1.3.4 Inflación

Los *commodities* representan en la actualidad activos financieros en la cartera de los inversores, por lo que los incentivos para adquirirlos como reserva de valor aumentan con el nivel de precios, es decir, con la inflación (Roache, 2010). Sin embargo, el efecto puede ser bidireccional, ya que el incremento en el precio de los *commodities* produce un incremento en el índice de precios, generando una mayor tasa de inflación.

El efecto de la inflación en el precio de los *commodities* tiene que ser estudiado detenidamente, pues si se evalúa como una relación directa con la capacidad de adquisición de los individuos participantes del mercado en cada país el efecto será independiente.

1.3.5 Productos substitutos

La demanda del cobre dependerá en gran medida de la existencia de posibles sustitutos. En la industria de las telecomunicaciones, el cobre es una materia prima importante, pero la promoción y aplicación de la tecnología de fibra óptica ha desafiado el estado del cobre.

El grafeno por su parte es un producto con cualidades "increíbles" según expertos, y es considerado el sustituto perfecto del cobre como conductor eléctrico y de calor. Según un estudio de Corfo, en 2006 sólo se habían hecho 138 publicaciones respecto al material en todo el mundo. En 2014, la cifra aumentó a casi 11 mil, observándose un crecimiento constante y explosivo.

1.3.6 Tendencias globales

Los planes masivos de urbanización de China e India tendrán un fuerte impacto en la demanda del cobre. Por otro lado, existe una tendencia de "Nacionalismo de recursos" por el lado de la oferta, lo que lleva a retrasos en los proyectos y a interrupciones en el suministro. Esto se ha demostrado en varios países en desarrollo donde existen deseos del gobierno de declarar un mayor control sobre los recursos naturales ubicados en sus territorios. Cabe resaltar también la suspensión de proyectos mineros en Chile, el mayor productor mundial de cobre. Estos proyectos tenían como objetivo contribuir a una mayor capacidad productiva de cobre en 2012-2025. (Yiming W, Kabwe E, 2015).

2 Capítulo II: Metodología

En este capítulo se presenta la metodología que permitió desarrollar el presente trabajo de tesis. Se muestran aspectos como el tipo de investigación, las técnicas y procedimientos que fueron utilizados para llevar a cabo esta investigación.

2.1 Estudios previos de predicción de los precios de *commodities*

En el ámbito financiero, la teoría del mercado eficiente ha sido por muchos años utilizada para explicar el comportamiento de los activos. En esta teoría planteada por Fama, se explica que el precio de un activo refleja toda la información disponible hasta el momento. Por lo tanto, desde el punto de vista de los mercados de *commodities*, esto conlleva a que no haya ningún sentido en tratar de predecir el comportamiento mercado, no sería útil investigar los fundamentos subyacentes de un producto que podría permitir obtener rentabilidad, no tiene sentido siquiera mirar las noticias o la variación histórica del precio de cada *commodity*. Subsecuentemente esta teoría conlleva a intuir que el movimiento de los precios es en su mayoría aleatorio y están impulsados por eventos imprevistos.

La suposición común del comportamiento aleatorio de los mercados de *commodities* minerales ha fomentado el uso de modelos estocásticos-gaussianos que trabajan dentro de límites preestablecidos y bien conocidos para pronosticar precios. Sin embargo, cada *commodity* tiene características propias con respecto al procesamiento, comercio, transporte y aplicación, lo que resulta en configuraciones particulares. Estas diferencias tienen implicaciones significativas para la fijación de precios no solo para un producto en particular, sino también para los productos complementarios y sustitutos.
Por ello existe hasta el día de hoy un debate sobre si los precios de los *commodities* exhiben un comportamiento aleatorio o no.

2.2 Tipo de investigación y diseño

El método de investigación que se utilizó es de carácter cuantitativo el cual está basado en la recolección de información a través de diferentes bases de datos. Se recopilaron los precios históricos del cobre, y demás variables detalladas más adelante para la generación del algoritmo.
Asimismo, al no tener total certeza de la naturaleza del comportamiento de estos mercados, es complicado determinar el modelo o algoritmo que será más adecuado para realizar la predicción. Por

ello el carácter que adopta la presente investigación es exploratorio, ya que otorgará una visión general de diversos métodos utilizados por otros autores antes de evaluar el performance de algoritmos de inteligencia artificial en el pronóstico de la variación del precio del cobre.

Se plantea iniciar el desarrollo de la investigación utilizando métodos estadísticos tradicionales, posteriormente se emplearán métodos de *Machine Learning*, a los cuales se aplicará un orden de procesos preestablecido para la obtención de óptimos resultados.

2.3 Métricas de Evaluación

La selección de una medida de rendimiento es crucial para poder evaluar y comparar los resultados obtenidos de los distintos algoritmos a emplear.

- MSE: Una medida de rendimiento típica para problemas de regresión es el error cuadrático medio (MSE). Se calcula como el promedio de la diferencia al cuadrado entre el valor objetivo y el valor predicho por el modelo de regresión. Al elevar al cuadrado las diferencias, penaliza incluso pequeños errores que conducen a una sobreestimación de cuán errado es el modelo.
- MAE: El error absoluto medio es la diferencia absoluta entre el valor objetivo y el valor predicho por el modelo. El MAE es más robusto para los valores atípicos y no penaliza los errores de forma tan extrema como el MSE.

Tanto el MSE como el MAE son formas de medir la distancia entre dos vectores: el vector de predicciones y el vector de valores objetivo.

- R^2: El coeficiente de correlación es una medida utilizada para evaluar el rendimiento de un modelo de regresión. La métrica compara el modelo de predicción con una línea base constante y expone cuán preciso es el este. La línea de base constante se elige tomando la media general de los datos. Es importante destacar que el R^2 es una puntuación libre de escala lo cual implica que no importa si los valores son demasiado grandes o pequeños, este siempre será menor o igual a 1.

  Para la evaluación de los algoritmos de *Machine Learning* el R^2 será denominado puntaje.

3 Capítulo III: Método Estadístico

Los métodos de aprendizaje automático y aprendizaje profundo pueden lograr resultados impresionantes en problemas de predicción de series temporales desafiantes. Sin embargo, hay muchos problemas de pronóstico en los que los métodos clásicos como ARIMA y el suavizado exponencial logran superar fácilmente a los métodos más sofisticados.

Por lo tanto, es importante comprender cómo funcionan los métodos clásicos de pronóstico de series de tiempo y evaluarlos antes de explorar métodos más avanzados.

El fin de empezar esta investigación utilizando los métodos clásicos de pronósticos, es tener un escenario base o, mejor dicho, una línea base en el rendimiento de los resultados obtenidos para poder compararlos posteriormente con otros modelos de *Machine Learning* y saber que tan bien se desempeñarán realmente cada uno.

El uso de métodos estadísticos y modelos de series de tiempo para analizar tendencias futuras cobraron relevancia frente a las inconsistencias e inconvenientes de la naturaleza estática que presentaron los modelos econométricos. Los modelos de series temporales se usan comúnmente para pronosticar precios de metales.

3.1 Metodología Box-Jenkins:

G.P.E Box y G.M Jenkins propusieron en la década de los 70, un nuevo conjunto de herramientas de predicción en su publicación *"Times Series Análisis: Forecasting and Control".* El análisis de Jenkins se refiere a un método sistemático de identificación, ajuste, verificación y uso de series temporales integradas, autorregresivas y de promedio móvil (ARIMA).

El objetivo de la metodología Box – Jenkins es identificar y estimar un modelo estadístico que puede ser interpretado como generador de la información de la muestra, para el desarrollo adecuado del modelo, la metodología se clasifica en tres etapas:

- Etapa de identificación y selección del modelo
- Etapa de estimación de parámetros
- Etapa de verificación

3.2.1 Etapa de identificación y selección del modelo:

El primer paso de la metodología propuesta es evaluar si la serie de datos a trabajar es estacionaria. Un proceso estadístico es estacionario si la distribución de probabilidad es la misma para todos los valores de t. Esto implica que la media y la varianza sean constantes para todo t. El supuesto de estacionariedad permite establecer afirmaciones sobre la correlación entre dos valores sucesivos del modelo predictivo. Los rendimientos financieros observados con alta frecuencia (por horas, días…), como es el caso del presente estudio, no son estacionales y suelen tener: (a) una media estable y periodos de alta y baja volatilidad.

Modelo autoregresivo (AR):

Los modelos de autorregresión, también conocidos como modelos AR, se utilizan para realizar pronósticos sobre variables ex-post (observaciones que conocemos completamente su valor) en determinados momentos del tiempo normalmente ordenados cronológicamente.

El valor en el momento t de la serie temporal se expresa como combinación lineal de las p observaciones anteriores de la serie más la innovación. En este tipo de modelo entonces la variable explicativa es la misma variable dependiente retardada y se debe poder probar que existe una correlación entre la variable de salida y los valores en los pasos de tiempo anteriores. Una fuerte correlación entre la variable de salida y la variable rezagada específica permite al modelo AR otorgarle más peso dentro del modelo a esa variable.

Modelo AR (p) $X_t = c + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t$

Donde c es una constante, φ son los parámetros del modelo, y se pronostica Xt en función de los valores pasados de x y se incluye un término de error que se comporta como ruido blanco.

Proceso integrado (I):

El componente integrado de los modelos ARIMA, se refieren al estado de la variable a tratar. Cada diferencia se entiende como una variación en el estado de la variable. Por ejemplo, en el caso del precio de un commodity, si se trabaja con la serie de precios de cierre la integración sería 0, pues la variable utilizada es "pura". No obstante, no es recomendable trabajar con variables puras pues suelen tener tendencia y no se pueden modelar en esas condiciones. Un grado de integración 2 significa que el

modelo se construirá sobre la variación de la serie en estudio, es decir, no se modela el precio, sino la variación del precio. (Venegas.P, Viveros C, 2018)

Proceso de Media Móvil (MA):

Los modelos de media móvil son aquellos en que el valor de la variable Xt para un instante t está en función de un término independiente εt y de una sucesión ponderada de errores correspondientes a los instantes precedentes.

Modelo MA(q) $X_t = \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} \cdots + \theta_2 \varepsilon_{t-k} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$

Modelo Autorregresivo de media móvil (ARIMA):

Si combinamos la diferenciación con autorregresión y un modelo de promedio móvil, obtenemos un modelo ARIMA con parámetros p, d y q. Donde:

p=orden de la parte autorregresiva

d=grado de primera diferencia involucrada

q=orden de la parte media móvil

Modelo ARIMA(q) $X'_t = c + \phi_1 x'_{t-1} + \cdots + \phi_p x'_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_q \varepsilon_{t-q} + \varepsilon_t$

Donde X'_t es la serie diferenciada, los valore sde la derecha incluyen los valores rezagados de Xt y los errores rezagados.

Procedimiento:

Siguiendo las etapas propuestas por Box-Jenkins, se debe iniciar por la fase de Identificación.
El objetivo de esta fase es determinar si el conjunto de datos cunple los requerimientos del modelo ARIMA, o si es necesaria una transformación de los datos. Para ellos se evalúo la estacionariedad de la serie original.

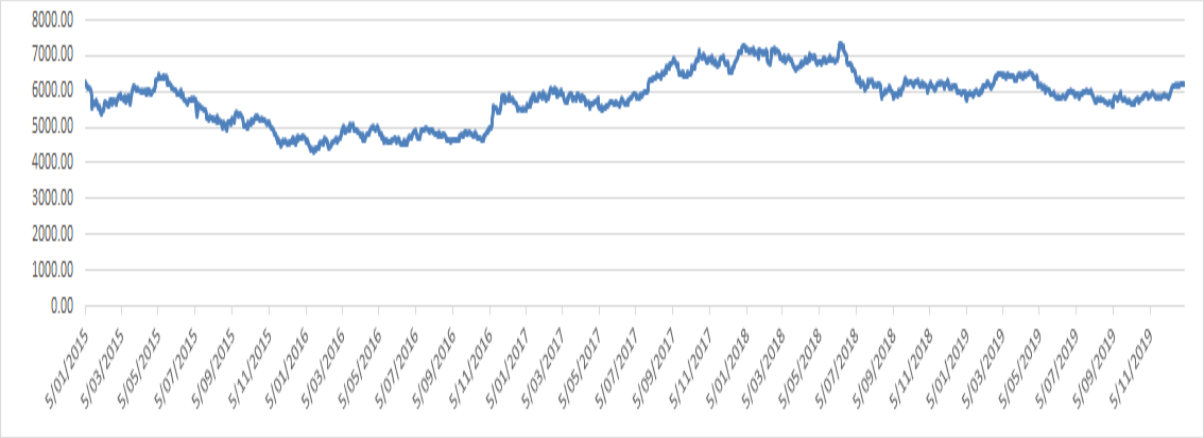

Figura 3.1: Precio diario del cobre $/ton.

Graficando de manera simple el comportamiento del precio a lo largo de los cuatro años mencionados, se puede observar que la variación del precio de cierre del cobre en el lapso de tiempo estudiado sigue una serie no estacionaria, como se observa en el gráfico de la figura 3.1.

Por lo general la no estacionariedad es causada por una tendencia, un cambio en la media local o variación estacional. Dado que para usar la metodología propuesta se debe conseguir una serie estacionaria, se debe ajustar la data antes de modelar la serie. Por ellos se procedió a trabajar con la variación del precio, diferenciando la serie y se obtuvo el resultado que se muestra en el gráfico 2.

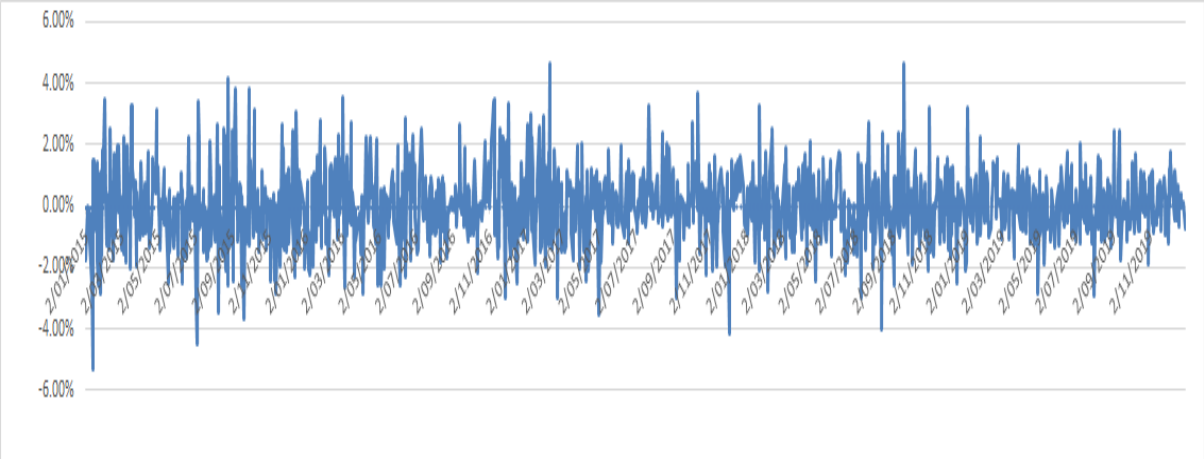

Figura 3.2: Variación porcentual del precio diario del cobre $/ton.

Una vez obtenida la serie de datos transformada, se inició por evaluar la capacidad predictiva del modelo Autoregresivo (AR).

El modelo ARIMA asume que la variable a predecir debe tener correlación con la variable rezagada. Una forma de verificar la presencia de autocorrelación es con el diagrama de retraso del conjunto de datos. (Figura 3.2). Al trazar la observación en el tiempo t-1, versus la observación en el siguiente lapso de tiempo (t+1) como un diagrama de dispersión, se advierte que no existe alguna correlación ni tendencia entre ambas variables.

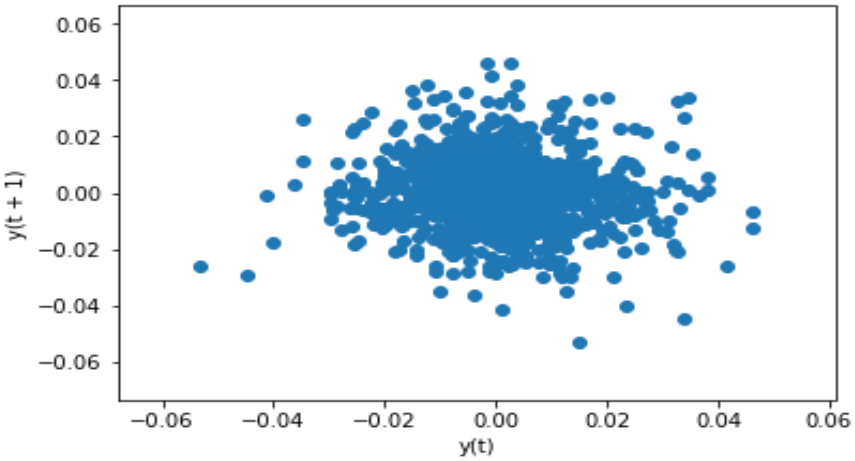

Figura 3.3: Diagrama de retraso del dataset de la variación de precio del cobre.

Se obtuvo, asimismo, el coeficiente de Pearson con un valor de -0,0579, y como se puede observar en el gráfico 4, en el que se traza el coeficiente de autocorrelación para cada variable de retraso del modelo, este se mantuvo muy por debajo del 0.5 (o muy por encima del -0.5), por lo tanto, no hay indicio de que alguna de las variables de retraso pueda usarse en un modelo predictivo.

La relación entre la observación y los valores históricos de esta no presenta ningún cambio significativo a lo largo del tiempo.

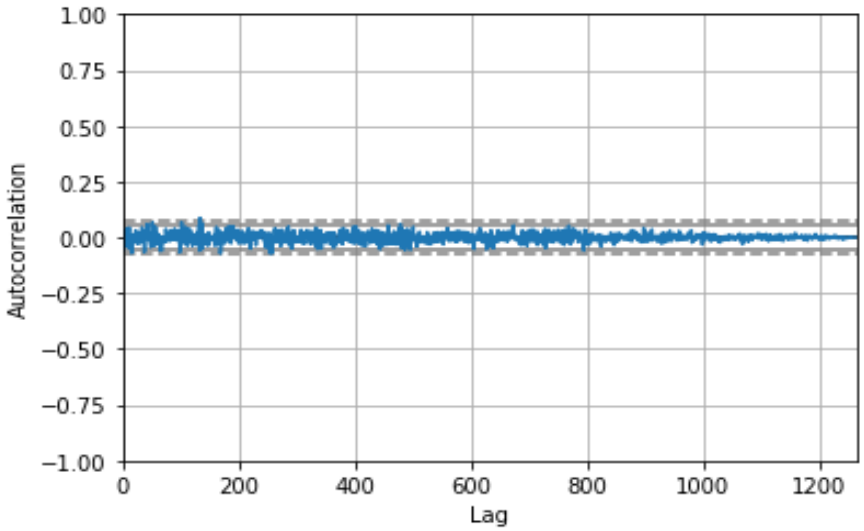

Figura 3.4: Pandas Autoccorrelation Plot

3.2.2 Etapa de estimación de parámetros:

Modelo de Persistencia:

Como siguiente paso, después de evaluar la estacionalidad y autocorrelación de la variable, como prueba para evaluar la capacidad predictiva del modelo, se empleó el modelo de persistencia. Esta clase de modelo es una forma simple de realizar predicciones persistiendo en la última observación.

Se dividen las observaciones en un grupo de datos para prueba llamado "test set" y el grupo usado para evaluar el performance del pronóstico, llamado *"unseen data".* Las predicciones se realizan utilizando un modelo de validación de avance para que se puedan persistir las observaciones más recientes para el día siguiente. Con lo cual se realizan X pronósticos por día, donde X es el número de días (u observaciones pasadas) utilizadas para hacer el pronóstico del día siguiente también llamado *lag value* o valor de retraso.

Se realizaron pruebas variando el valor de X, para comparar el performance del modelo. El indicador utilizado fue el MSE y el R2. Los resultados obtenidos se muestran en la tabla 3.1:

Tabla 3.1: Resultados obtenidos según el *Lag Value*

Lag value (X)	R^2	MSE
1	0	0.000006
5	-0.01	0.000070
10	-0.93	0.000601
12	-0.75	0.000553
15	-0.73	0.000525
20	-1.00	0.000718
30	-0.98	0.000636
35	-1.00	0.000601

El fin de emplear este modelo es tener una idea de cuál es el valor de X con el que se obtiene un menor error. Como se puede observar, en cuanto al R^2 obtenido, entre los 20 y 35 días se obtiene una correlación negativa casi perfecta.

Modelo Autoregresivo:

Con la biblioteca *stats model* disponible en Python se diseñó un modelo de autorregresión que de manera automática selecciona un valor de retraso apropiado mediante pruebas estadísticas y entrena un modelo de regresión lineal.
Una vez obtenido el valor de retraso óptimo se utiliza el modelo para realizar las predicciones de un determinado número de días "Y"

Se eligió un Y= 7 días y al correr el modelo, se obtuvo el *Lag* óptimo para X = 23 días, no obstante, la capacidad predictiva es ínfima, tal como se esperaba. En la tabla 3.2 se observan los valores obtenido en base a la predicción y los valores esperados.

Tabla 3.2: Valores obtenidos en la predicción del modelo AR

Predicción	Esperado
-0.034%	-1.198%
-0.222%	-0.229%
0.153%	-0.179%
0.244%	-0.488%
-0.003%	0.000%
0.273%	-0.018%
-0.101%	-0.200%

El MSE de este modelo es 0.000140.

Modelo ARIMA:

Como ya se mencionó, los modelos ARIMA integran el concepto de media móvil, auto regresión e integración. Cada uno de estos componentes se especifica explícitamente en el modelo como un parámetro. Se utiliza una notación estándar de ARIMA (p, d, q) donde los parámetros se sustituyen con valores enteros para indicar rápidamente el modelo ARIMA específico que se está utilizando.

Al usar la biblioteca de *Python statsmodel* se adaptó un modelo ARIMA con los siguientes parámetros:

p= 5

d=1

q=0

Donde p establece el valor de retraso *(Lag value)*, y en este caso son 5 días.

Tal como se mencionó anteriormente, el conjunto de datos inicial no cumplía los requisitos de serie estacionaria, por lo tanto, el orden de diferencia a usar en este caso es 1, para así obtener la variación en los precios como variable a pronosticar. Por último, el modelo de promedio móvil a usarse es de 0.

Se obtuvo un modelo con los siguientes coeficientes presentes en la tabla 3.3.

Tabla 3.3: Modelo ARIMA

	coef	std err	z	P>|z|	[0.025	0.975]
const	-7.411e-06	9.82e-05	-0.075	0.940	-0.000	0.000
ar.L1.D.VPrecio	-0.8902	0.028	-32.280	0.000	-0.944	-0.836
ar.L2.D.VPrecio	-0.7037	0.036	-19.684	0.000	-0.774	-0.634
ar.L3.D.VPrecio	-0.5133	0.038	-13.428	0.000	-0.588	-0.438
ar.L4.D.VPrecio	-0.3767	0.036	-10.548	0.000	-0.447	-0.307
ar.L5.D.VPrecio	-0.2001	0.028	-7.266	0.000	-0.254	-0.146

A continuación, se muestra el diagrama de los valores de error residual, los cuales presentan aparentemente una distribución gaussiana.

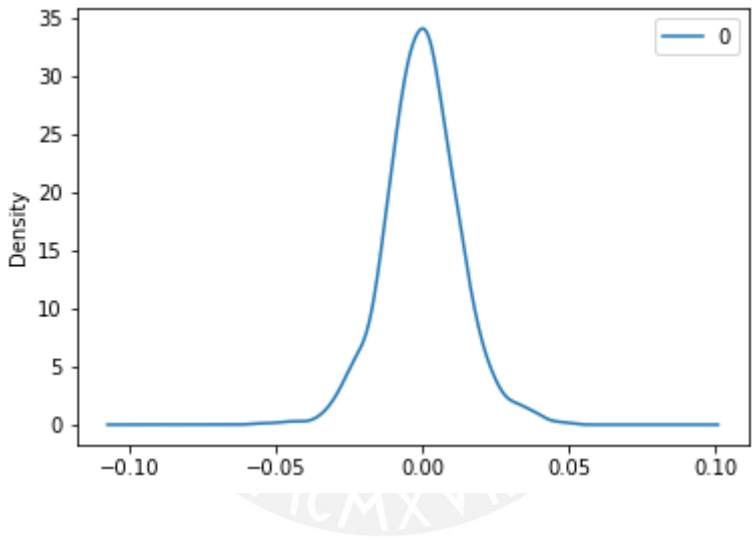

Figura 3.5: Diagrama de densidad de error residual de ajuste ARMA

La media de los errores es 0.000010, por lo tanto, se puede decir que el sesgo de la predicción es bastante reducido. Hay que tener en cuenta, sin embargo, que en este primer modelo se ha usado todo el conjunto de datos para realizar el análisis.

3.2.3 Etapa de verificación:

Modelo ARIMA de pronóstico continuo:

El objetivo de este modelo es realizar un pronóstico continúo dada la dependencia de las observaciones en los pasos de tiempo anteriores. La forma de realizar este pronóstico es recrear el modelo ARIMA anterior después de cada nueva observación recibida.

Se utilizó en primera instancia, la siguiente división de la data set (1264 datos): 66% para el conjunto de datos de entrenamiento inicial y el resto para el conjunto de datos de prueba. Asimismo, se mantuvo el modelo con los parámetros anteriores, ARIMA (5,1,0)

Los resultados se modelan en el siguiente grafico (figura 3.6), diagrama de líneas que muestra los valores esperados (color azul) en comparación con los valores obtenidos del pronóstico continuo (rojo). El MSE obtenido es 0.0000210, y el r2 -0.14776

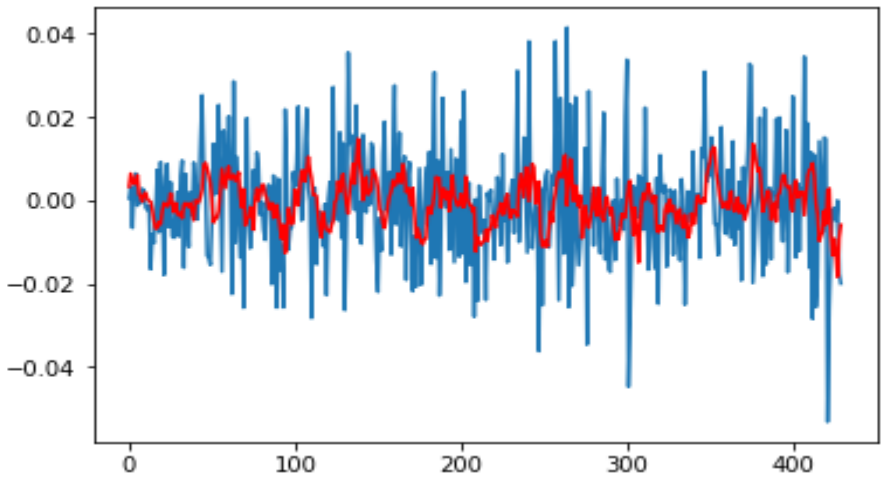

Figura 3.6: Predicción móvil ARIMA de la variación de precio del cobre diario

Resulta complejo evaluar la precisión del modelo en un pronóstico de data frecuente, en este caso de valores diarios. Por ellos, se ejecutó la prueba del modelo sobre la serie de datos con frecuencia mensual, disminuyendo el tamaño del *data set* de 1264 datos a 60 datos y manteniendo los parámetros en p=5, d=1 y q=0.

Los resultados se modelan en el gráfico 7, el MSE obtenido fue 0.004917 y el r2 -0.50240.

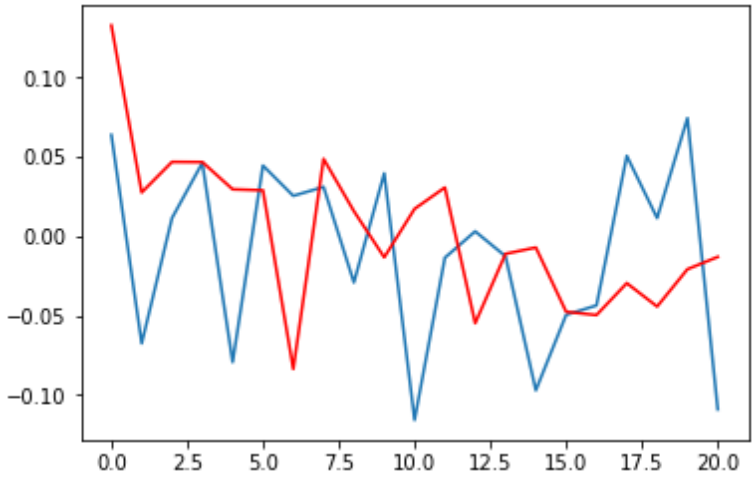

Figura 3.7: Predicción móvil ARIMA de la variación de precio del cobre mensual

Como se puede observar, la exactitud del modelo no mejora al cambiar la frecuencia de la variación de días a meses, y en general tal como lo demuestran ambos gráficos y los valores de r2 obtenidos, la capacidad de predicción de este método es deficiente.

Los modelos de series temporales son menos complejos que los modelos econométricos y pueden replicar el comportamiento de los precios de commodities en diferentes horizontes. El modelo ARIMA posee ventajas, como la flexibilidad para representar diferentes series de tiempo, sus propiedades estadísticas y el uso de la metodología Box-Jenkins, no obstante, posee desventajas significativas al estar basada en el supuesto de que las variables aleatorias dependen del tiempo y en la existencia de una correlación lineal preestablecida entre variables, además de la suposición de cambios uniformes a largo plazo y la inexistencia de una relación causal entre variables, lo cual resulta en una aproximación insatisfactoria de problemas del mundo real.

Kriechbaum et al (2014). afirmó en su investigación, que los modelos ARIMA tradicionales no son adecuados para predecir los precios de los metales básicos, no obstante, estos se pueden optimizar a través de diferentes métodos para mejorar la predicción, como por ejemplo la integración de inteligencia artificial a estos modelos. Por otro lado, Parisi, Améstica y Chileno (2016) realizaron una investigación centrada en evaluar la eficacia del modelo ARIMA al optimizarlo con fuerza bruta computacional para predecir el precio del petróleo. Gracias a esta implementación, se pudo obtener una capacidad predictiva superior al 60% al obtener el precio semanal de este *commodity*.

En la siguiente sección de este capítulo se profundizará en los distintos métodos que involucran *machine learning* y su eficiencia en la predicción del precio del cobre.

4 Capítulo IV: Algoritmos de *Machine Learning:*

El pronóstico de series de tiempo es complicado. A diferencia de los problemas más simples de clasificación y regresión, las series de tiempo agrega la complejidad del orden temporal de los datos o la dependencia temporal entre las observaciones. Esto puede elevar la complejidad del modelo ya que se requiere un manejo especializado de los datos al ajustar y evaluar modelos.

La estructura temporal debe estar presente en el modelado, añadiendo elementos de importancia como tendencias y estacionalidad, que se pueden igualmente aprovechar para mejorar la habilidad del modelo. Tradicionalmente, la predicción de series temporales ha estado dominada por métodos lineales como ARIMA porque se comprenden bien y son efectivos en muchos problemas. Pero estos métodos clásicos se enfocan en relaciones lineales, utilizan data univariable, y asumen una dependencia temporal fija, donde el número de observaciones de retraso proporcionadas deben diagnosticarse y especificarse.

4.1 Flujograma de procesos a aplicar en *Machine Learning:*

Muchos de los algoritmos de Machine Learning son capaces de aprender automáticamente asignaciones complejas arbitrarias de datos de entrada y salida, al mismo tiempo de poder admitir múltiples entradas y salidas. Estas son características poderosas para el pronóstico de series de tiempo, particularmente en problemas con dependencias no lineales complejas, entradas multivalentes y pronósticos de varios pasos.

Dentro de esta rama de la inteligencia artificial, existen diversos algoritmos que pueden ser utilizados para un sin fin de objetivos, como los mencionados en el marco teórico. Por ello, es clave reconocer el flujo de trabajo que se debe llevar a cabo con el fin de elegir el que se ajuste de manera óptima. Se distinguieron las siguientes seis etapas:

1. Categorización del problema: Consta en entender cómo representar el problema propuesto, identificar qué tipo de problema es. Determinar qué variables serán utilizadas, recolectar los datos y posteriormente representar los datos y adaptarlos para el modelado del problema.
2. Análisis de datos: Evaluar la estructura de los datos recolectados, realizar el pre-procesamiento adecuado y reorganizar la data en caso sea necesario. En esta etapa se hace la división correspondiente del Data Set en entrenamiento y validación.

3. Construcción del algoritmo: Abarca desde la transformación de los datos (en caso sea necesario estandarizar la data, o convertirla a algún formato específico), hasta la optimización del modelo en base a prueba y error con el set de entrenamiento.
4. Validación: Consta en la verificación de la capacidad predictiva del modelo, en caso sea necesario se retorna al paso 3 para la optimización de los hiper parámetros.
5. Evaluación: Se analiza la performance del modelo en base a los resultados predictivos sobre el *data set* de verificación.
6. Predicción: En esta última etapa se emplea el modelo sobre datos nuevos para pronosticar la variable *target*.

En el siguiente esquema se muestra de manera más detallada los procesos que involucran estas seis etapas, es importante mencionar, que este flujo puede ser usado independientemente del algoritmo que se elija.

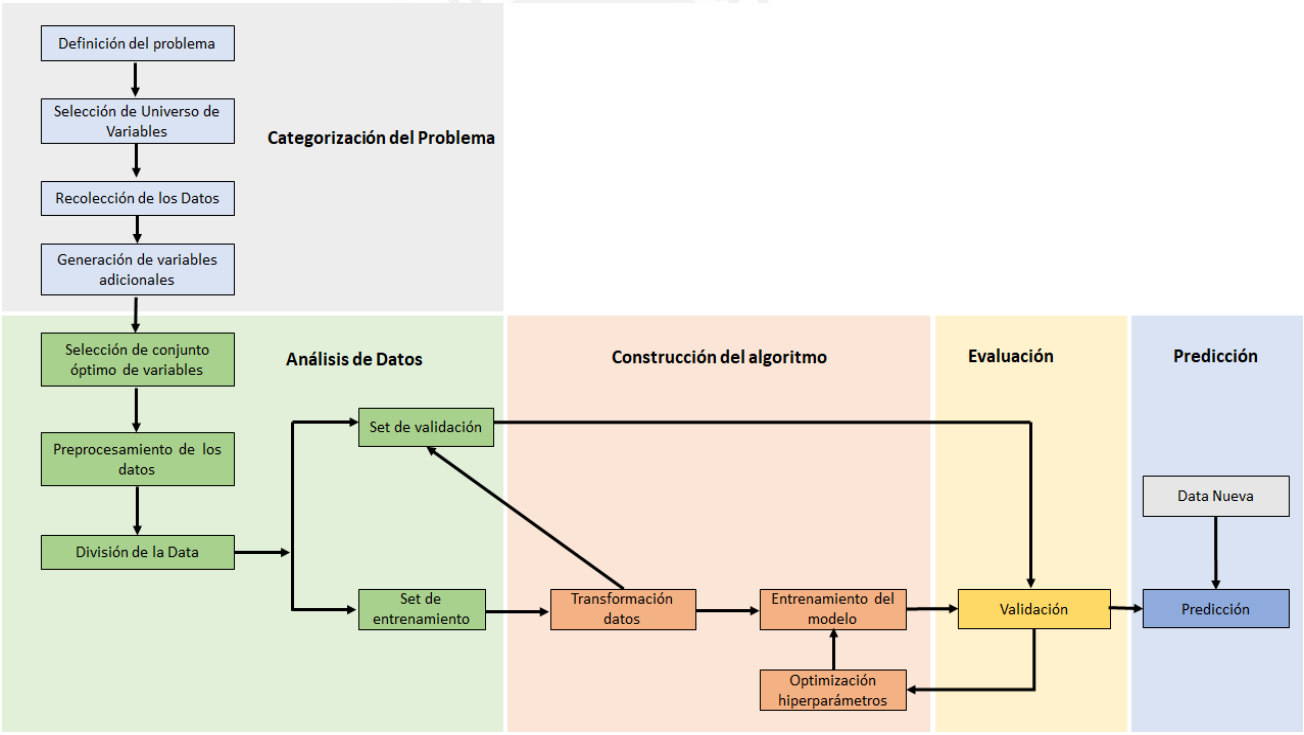

Figura 4.1: Flujograma de elaboración de un modelo predictivo con algoritmos de ML.

En este capítulo se desarrollarán las 6 etapas y se irá detallando cada una en la construcción del modelo. Es importante tener claro que tanto la categorización del problema como el análisis de los datos se desarrollarán una única vez, dado que son etapas independientes de la elección del algoritmo. Cómo se

verá más adelante, al final de esta tesis se habrá evaluado la capacidad predictiva de dos algoritmos de *Machine Learning* frente al problema de esta investigación.

4.2 Categorización del problema

4.2.1 Definición del problema

La primera etapa del proceso es la categorización, esta involucra la formulación y entendimiento del problema, determinación de las variables a utilizar y la recolección de datos para cada una.

Se debe poder clasificar el problema en una de las tres grandes clases de algoritmos existentes: *"Supervised Learning", "Unsupervised Learning"* y *"Reinforcement Learning".* En la siguiente tabla se estructuran las diferencias básicas entre estos.

Tabla 4.1: Métodos de Machine Learning

	Tipo de Data	Análisis	Objetivo	Enfoque
Supervised Learning	Datos de entrada y salida son conocidos y pre-categorizados	*Offline*	El resultado se predice utilizando la data de entrada categorizada	➔ Regresión ➔ Clasificación
Unsupervised Learning	Solo los datos de entrada son conocidos	En tiempo real	El resultado se predice en base a los patrones presentes en los datos de entrada.	➔ *Clustering* ➔ *Representation Learning*
Reinforcement Learning	Data no predefinida	En tiempo real y con interacción de un agente	El resultado se predice en base a patrones y al enfoque de prueba y error	➔ Optimización de estrategias ➔ *Behavior learning*

La elección del algoritmo dependerá también de los siguientes elementos característicos: la precisión, la interpretabilidad, la complejidad y la escalabilidad.

En primer lugar, dado que el fin es predecir la variación del precio del cobre, el *"output"* final del modelo será un valor numérico, el enfoque será entonces de regresión, además la data con la que se está trabajando es del tipo continua y está previamente clasificada es decir tanto el input como el output son conocidos, asociaremos entonces modelos pertenecientes al método de aprendizaje supervisado.

Una vez definido el problema, el siguiente paso es seleccionar el universo de variables que se tomará en consideración.

4.2.2 Selección del universo de variables

En el marco teórico se identificaron ciertas variables o factores que han influenciado en el comportamiento del precio de *commodities* a lo largo de la historia. Sin embargo, específicamente para la construcción de este modelo inicialmente se eligieron nueve variables explicativas para que conformen el universo de posibles factores a emplear en los algoritmos.

Precio de cobre:

La variable *target* de a estimar será el precio del cobre. Se utilizará la información histórica del precio del futuro de cobre a 3 meses, correspondiente al nemónico LMCADS03. Los futuros sobre *commodities* permiten a los *traders* ganar exposición a los precios de estos sin tomar posesión física del activo. Con estos contratos, los *traders* acuerdan comprar una cierta cantidad de un producto básico en una fecha futura (la fecha de vencimiento).

Inventarios de cobre:

La disponibilidad del bien, en este caso el cobre, se cuantifica en términos de inventario. En este caso, se presentan los valores correspondientes a los inventarios rastreados por la bolsa de Londres (LME).
La lógica del mercado explica que, la inestabilidad de los precios presentes en el corto y largo plazo se reflejan a través de los inventarios. Estos son, una caída en la demanda, dada la baja elasticidad de corto plazo y la gran elasticidad de largo plazo, lleva a una gran acumulación de inventarios. Esto en el tiempo deprimirá al precio, el cual se mantendrá así mientras se siga acumulando inventario. Un giro en las condiciones económicas lleva a un cambio en la demanda y desacumulación de inventarios provocando un incremento en el precio (Fisher et al, 1972). En general, la relación del precio del mineral y los inventarios es inversamente proporcional, tal como se puede observar en la figura 4.2.

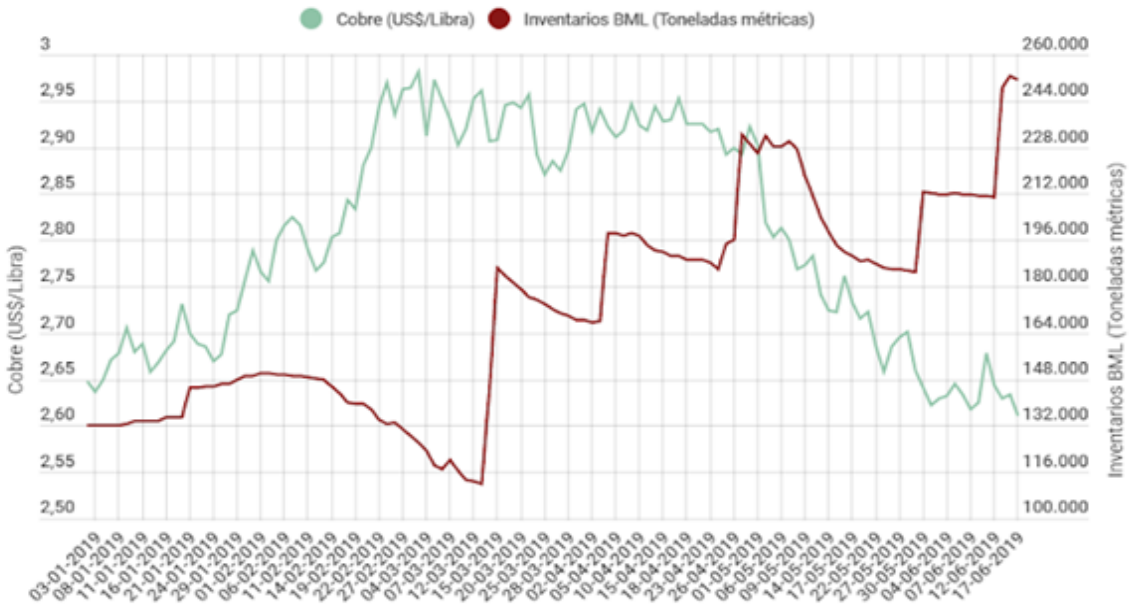

Figura 4.2: Evolución del precio e inventarios de cobre

Elaboración: (Cochilco 2019: 1, figura 1)

Purchasing Managers Index:

Este indicador económico brinda una visión del futuro en base a información de las condiciones comerciales actuales y esperadas. Se podría decir que es una medida de la dirección predominante de las tendencias económicas en los sectores de fabricación y servicios. En este trabajo se utilizarán los datos del PMI de Estados Unidos y China, siendo ambas potencias mundiales y principales consumidores de cobre en el mundo, se espera, tal como se observa en la figura 4.3, que ante un incremento en el PMI el precio del cobre suba dado el aumento en la demanda.

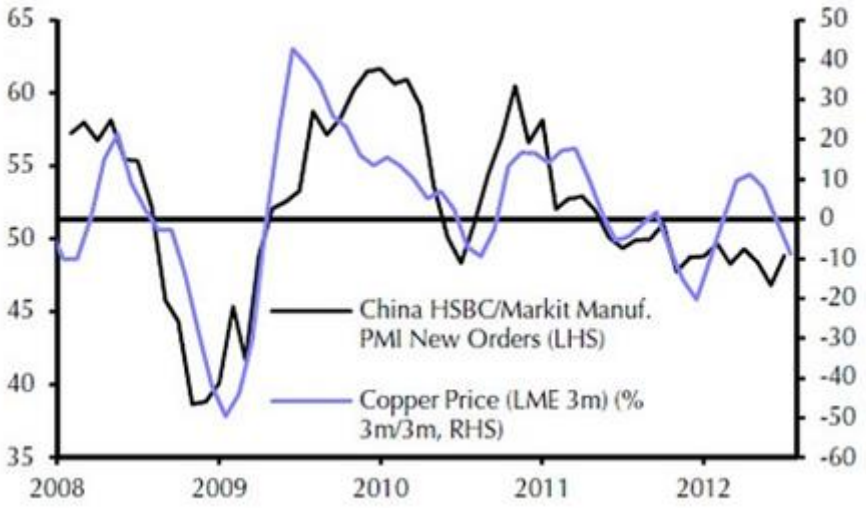

Figura 4.3: China PMI & Precios del cobre

Elaboración: (Capital Economics 2019)

PBI China:

Dentro de las fuerzas de mercado que impulsan el precio de cobre, China es responsable de más de la mitad del consumo mundial de cobre y de la mayor parte de la producción mundial. El cobre se importa a China en todas sus formas diferentes: semi y medios productos (por ejemplo, para el sector de la construcción) y componentes (por ejemplo, para el sector electrónico).

Existen diversos trabajos de investigación realizados a lo largo de los años, que demuestran la fuerte correlación entre indicadores económicos y productivos de este país y el precio del cobre. La contribución a la literatura empírica sobre la relación entre la actividad económica y el precio de las materias primas, en particular del cobre, es que se cuantifica el impacto del crecimiento del PIB de China usando datos en frecuencia mensual.

Ercio Muñoz, realizó un trabajo de investigación titulado “El efecto de sorpresas en el crecimiento de China sobre el precio del cobre”. Los principales resultados son: que el precio real del cobre responde de forma positiva y estadísticamente significativa a las sorpresas de crecimiento de China, obteniéndose una respuesta de 1,1% frente a una revisión en el crecimiento anual de 0,1%, más aún al comparar con otras economías avanzadas, sólo Estados Unidos presenta un impacto significativo de 0,9% frente a una revisión de 0,1%. (Muñoz E, 2013)

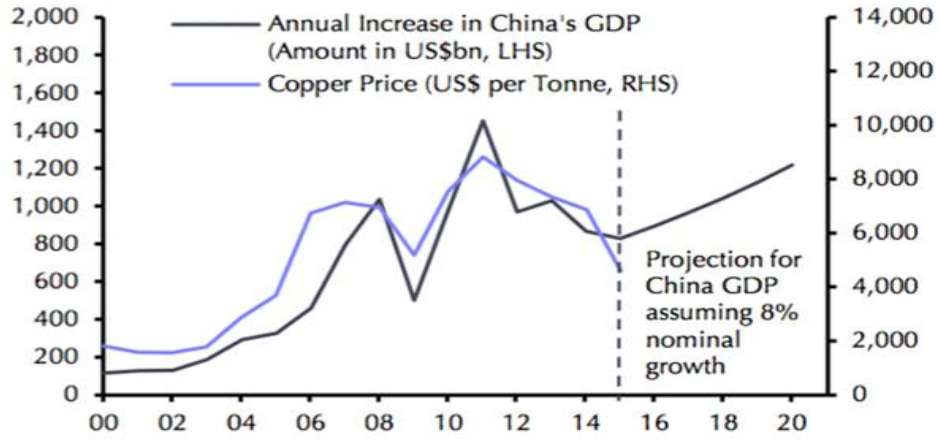

Figura 4.4: PBI China & Precio del cobre

Elaboración: (Capital Economics 2109)

Volumen transado:

El volumen de comercio o volumen transado es la cantidad total de un producto o *commodity* negociados por un valor específico durante un periodo de tiempo específico. Existen diversas interpretaciones entre la relación del volumen negociado y el precio, o la tendencia de este, de un *commodity*. Se puede decir que el volumen representa una medida de intensidad o presión detrás de una tendencia de precios. Cuanto mayor sea el volumen, se puede esperar con mayor certeza que la tendencia existente continúe en lugar de revertirse.

"Volume Always precedes Price", es la base de una de las principales creencias de la mayoría de los técnicos en la bolsa de valores, y abarca el mercado de *commodities*, acciones, y bonos. Este análisis se realiza calculando el *OBV (On Balance Volume),* indicador desarrollado por Joe Granville (1963), el cual mide la presión de compra y venta como un indicador acumulativo, agregando volumen en los días al alza y restando en los días a la baja. El volumen de un periodo será positivo cuando el cierre está por encima del cierre del día anterior y es negativo cuando el cierre está por debajo del cierre anterior.
Granville planteó que el volumen precedía al precio por dos situaciones: el OBV aumenta cuando el volumen en días hábiles supera el volumen en días inactivos. El OBV cae cuando el volumen en los días de baja es más fuerte.

Un OBV creciente refleja una presión de volumen positiva que puede conducir a precios más altos. Por el contrario, la caída del OBV refleja una presión de volumen negativa que puede presagiar precios más

bajos. Granville señaló en su investigación que OBV a menudo se movería antes del precio. Espere que los precios suban si OBV está subiendo mientras que los precios están planos o bajan. Espere que los precios bajen si OBV está cayendo mientras los precios están planos o subiendo.

Tipo de cambio y China Spot Exchange Rate Yuan/US$:

En cuanto al tipo de cambio, se utilizaron los datos del índice del dólar estadounidense (USDX), el cual es una medida del valor del dólar en relación con el valor de un grupo de monedas de los países que son los socios comerciales más importantes para Estados Unidos (euro, franco suizo, yen japonés, dólar canadiense, libra esterlina y corona sueca.). Se recopilaron los datos para el período de tiempo estudiado en una frecuencia diaria. Se espera que la correlación entre el índice USDX y el precio del cobre sea negativa, hipótesis derivada de la ley del precio único para los bienes comerciables, la cual explica que una disminución en el valor del dólar debe ser compensada por un aumento en el precio del c*ommodity* en dólares y / o una caída en sus precios en moneda extranjera para garantizar el mismo precio cuando se mide en dólares.

Además, como muchos productos tienen un precio en dólares en los mercados internacionales, una caída en el valor del dólar puede aumentar el poder adquisitivo y la demanda de productos básicos de los consumidores extranjeros, al tiempo que reduce los rendimientos de los proveedores de productos extranjeros y potencialmente sus suministros. (Farooq A, 2008) Se incluyó también la tasa de cambio entre el Yuan y el dólar, teniendo en cuenta que China es el principal consumidor y como una medida más del efecto del dólar en el cobre.

Tasa de interés FED USA:

La tasa de interés utilizada es la de la reserva federal estadounidense, en este caso la frecuencia de datos obtenidos es mensual. La relación de la tasa de interés con el precio de los *commodities* es más compleja, pues existen diversas implicancias del interés en el proceso de adquisición, transporte, mantenimiento y financiamiento de futuros de *commodities.* Se sabe, por rendimientos históricos, que la tasa de interés y el índice global de *commodities* tienen una relación inversa (Ver figura 7). No obstante, la naturaleza del *commodity* influye en el efecto esperado.

Como menciona Hull en su libro “Introducción a los mercados de futuros y opciones” (2009) existen dos tipos de activos: los de inversión y los de consumo. El cobre pertenece a esta última clase, pues se mantiene sobre todo para el consumo y no generalmente con propósitos de inversión, es por ello por lo

que los argumentos de arbitraje generalmente utilizados para determinar los precios a plazo y de futuros de activos de inversión como el oro o la plata, no serán válidos al analizar el precio de *un commodity* como el cobre. Los individuos y las empresas que mantienen en inventario un *commodity* de este tipo lo hacen por su valor de consumo, no por su valor como una inversión, no tiene sentido alguno para ellos vender el *commodity* y comprar contratos a plazo porque éstos no pueden consumirse. Entran en juego otro tipo de rendimiento llamado el rendimiento de conveniencia (Los beneficios de mantener el activo físico se conocen en ocasiones como el rendimiento de conveniencia que proporciona el *commodity.)* que al igual que el manejo y costo de inventario resulta complejo de cuantificar e incorporar en el precio futuro del cobre.

Tasa de inflación China:

Por último, se utiliza la tasa de inflación China, debido a que históricamente en tiempos de inflación el cobre ha tendido a tener un rendimiento superior, siendo utilizado por inversores como una buena cobertura a sus carteras. En un análisis de *Bloomberg Intelligence* , el cobre superó a todas las principales clases de activos, como cobertura frente a la inflación, y durante los períodos de aumento de los precios al consumidor. El cobre a diferencia de otros metales industriales que son especializados y de metales preciosos que son impulsados básicamente por el sentimiento de inversores, se emplea en una amplia gama de industrias, convirtiéndolo en un representante de la economía en su conjunto.

"El cobre es más sensible a la inflación y al dólar debido a sus usos y su crecimiento con la economía". - Jodie Gunzberg, S&P Dow Jones Indices

4.2.3 Recolección de Datos:

Los datos de las variables mencionadas anteriormente se recolectaron para los periodos del 01/01/2015 al 31/12/2019, siendo 1264 observaciones por variable.

Cada una cuenta con una frecuencia de publicación individual, como se puede observar en la siguiente tabla:

Tabla 4.2: Data obtenida para las variables según frecuencia

	Inventarios	PMI	PBI	Volumen transado	Precio del Cobre	Índice del dólar	China exchange rate	Tasa FED USA	Tasa inflación China
Frecuencia	Semanal	Mensual	Trimestral	Diario	Diario	Diario	Diario	Mensual	Mensual

Para poder utilizar la data obtenida se tuvo que homogeneizar la frecuencia de las nueve variables, manteniendo constante el valor de aquellas variables de menor frecuencia (por ejemplo, trimestral), a lo largo de los periodos en los que no se contaba con datos.

En este caso para las variables PBI, PMI, tasa FED, tasa de inflación China e Inventarios se tuvieron que estandarizar a la frecuencia diaria determinada por el volumen transado y la variación del precio del commodity, definiendo así finalmente una serie temporal de alta frecuencia, es decir con datos diarios.

4.2.4 Generación de variables adicionales:

En base al conjunto de datos recolectados se generan variables secundarias calculadas como la variación de aquellas variables primarias con frecuencia diaria. Al sobre generar atributos, el algoritmo tendrá un universo más amplio de donde seleccionar aquellos relevantes, y en base a ello se podrá concluir acerca de la importancia de cada factor en el modelo, se verá en detalle este análisis en el capítulo 5.

4.3 Análisis de Datos

Como parte de la segunda etapa, se analizan las variables y la configuración de los datos recolectados. En este caso, se desea pronosticar una variable que sigue el orden de una serie temporal, como se explicó en el capítulo 3. Como se puede observar en el flujograma presentado en la figura 11, en esta etapa se realiza principalmente el preprocesamiento de los datos, y la división y preparación de los dos data sets.

4.3.1 Selección de conjunto óptimo de variables:

Como línea base se seleccionó aleatoriamente 5 variables para trabajar inicialmente, estas serán el nivel de inventario de cobre, el PMI de China, el PMI de Estados Unidos, el PBI de China y el volumen transado de cobre. Más adelante se evaluará incrementar el número de factores a considerar en el modelo, e incluso se analizará cuál es el conjunto óptimo de variables.

4.3.2 Preprocesamiento de los datos:

Pronóstico de series temporales

Ante un problema de pronóstico de series de tiempo hay muchos elementos que se deben considerar para la construcción correcta del modelo. Parte fundamental del análisis de los datos es el entendimiento de la estructura de estos datos, para poder definir correctamente el diseño del algoritmo.
Para ello, en la presente investigación se utilizaron siete pasos para comprender la taxonomía de la serie temporal y facilitar el análisis:

1. Datos de entrada y salida: Identificar detalladamente cuáles serán las variables proporcionadas al modelo para realizar el pronóstico, y qué se desea pronosticar. Para el propósito de esta investigación se detallaron las variables a utilizar en el inciso 2.3.
2. Regresión o Clasificación: Las dos clases de modelos para pronósticos se dividen en problemas de regresión y clasificación. Cómo se detalló anteriormente, un problema de pronóstico de series temporales en el que desea predecir uno o más valores numéricos futuros (como en este caso, el precio del cobre) es un problema de modelado predictivo de tipo regresión.
3. Estructurados o No estructurado: Es importante poder inspeccionar los datos con el fin de encontrar posibles patrones. Una serie sin patrón vendría a ser no estructurada, a menudo se puede simplificar el proceso de modelado identificando y eliminando las estructuras obvias de los datos, como una tendencia creciente o un ciclo repetitivo.
4. Univariable o Multivariable: Determinar si se trabajará con una variable o múltiples en el pronóstico, tanto para los *inputs* como *outputs.* En este caso, el modelo será multivariable.
5. *Single- Step* o *Multi-step:* Un problema de pronóstico que requiere una predicción del tiempo *t+1* se denomina modelo de pronóstico de un paso. Mientras que un problema de pronóstico que requiere una predicción de más de un paso de tiempo se denomina modelo de pronóstico

de varios pasos. Cuantos más pasos de tiempo se proyecten hacia el futuro, más complejo será el modelo. Se ha definido que el pronóstico a realizar para el precio del cobre sea de un paso.

6. Estático o Dinámico: Es posible desarrollar un modelo una vez y usarlo repetidamente para hacer predicciones. Cuando el modelo no se actualiza o cambia entre pronósticos, se hace alusión a que es estático. Por otro lado, cuando el modelo se ajusta a nueva data disponible para hacer predicciones futuras entonces se estaría hablando de un modelo dinámico. Lo ideal en el pronóstico del precio de un commodity es lograr un modelo dinámico, ya que de esta forma se podría realmente emplear en operaciones financieras.
7. Continuo o Discontinuo: Una serie de tiempo donde las observaciones son uniformes a lo largo del tiempo puede describirse como continua, a diferencia de las discontinuas donde las observaciones no son uniformes. También puede ser una característica del problema que las observaciones solo están disponibles esporádicamente o en intervalos de tiempo variables, como es el caso de las variables utilizadas en este trabajo, en este caso de data no continua, es necesario ajustar todas las variables a un formato de datos específico para que las observaciones sean uniformes a lo largo del tiempo, tal como se hizo al uniformizar la frecuencia de las variables en el inciso 4.2.3.

Una vez que se tiene una estructura clara de los datos de entrada y el tipo de pronóstico que se desea realizar, se procede a preparar la data para entrenar el modelo.

Los datos de series de tiempo deben transformarse antes de que puedan usarse para un modelo de aprendizaje supervisado, en el tercer paso del flujograma. el "pre-procesamiento", se convierte la serie de tiempo en muestras con datos de entrada y salidas.

Para series univariables como el modelo ARIMA realizado, las observaciones de tiempos *t-1* (*lag observations*) son utilizadas como input, no obstante, a partir de ahora, al implementar los algoritmos de Machine Learning, se busca tener en paralelo múltiples variables de entrada.

Las variables o series de tiempo de entrada son paralelas dado que cada una tiene una observación en el mismo paso de tiempo *t*. Antes de dividir los datos en d*atase*t de entrenamiento y validación lo correcto es organizar los datos de entrada en muestras manteniendo el orden de las observaciones para cada variable, de tal forma que se respete el orden temporal. Para las cinco variables de entrada y la variable de salida correspondiente se tiene los siguientes datos de los primeros 5 pasos de tiempo:

Tabla 4.3: Observaciones de los 5 primeros pasos de tiempo para cada variable

PBI China	PMI China	PMI USA	Volumen	Inventarios	Precio
6.0	47.2	50.2	19956	322817	6190
6.0	47.2	50.2	24877	322817	6215
6.0	47.2	50.2	27234	322817	6214
6.0	47.2	50.2	20195	309321	6219
6.0	47.2	50.2	18756	309321	6174

Se estructuran los datos utilizando los valores de las cinco primeras columnas como inputs, y la última como output, en este proceso se debe definir el número de pasos de tiempo *n* que se utilizarán para realizar el pronóstico. Por ejemplo, si se define un *n* = 2, el orden sería el siguiente:

```
[6  47.2  50.2  19956  322817]
[6  47.2  50.2  24877  322817] [6215]
[6  47.2  50.2  27234  322817]
[6  47.2  50.2  27234  322817] [6219]
```

Como se observa, al transformar las series de tiempo en muestras de entrada y salida, se descartan algunos valores de salida, por lo tanto, la elección del tamaño del número de pasos de tiempo de entrada tendrá un efecto importante en la cantidad de datos de entrenamiento que se utilizan. Como resultado de este proceso se obtiene un componente de entrada de tres dimensiones:

```
X.shape, y.shape (1263, 2, 5) (1263,)
```

La primera dimensión es el número de muestras, en este caso son 1263, la segunda dimensión es el número de pasos de tiempos *n* definidos, y la última dimensión es el número de variables ingresadas. Como los algoritmos trabajan en su mayoría con datos presentados en dos dimensiones se convierte o "aplana" el *dataset* de la siguiente forma:

```
(1263,10)
```

Donde el segundo componente es la multiplicación del número de pasos de tiempos por el número de variables de entrada. Posteriormente se hace la división del *Dataset* en entrenamiento y validación, se utilizará dos tercios de la base de datos para el entrenamiento y el tercio restante para la validación.

Este proceso es crucial cuando se trabaja con datos con orden temporal y se debe realizar independientemente del algoritmo a utilizar. Por ello, hasta este punto se han ejecutado de forma genérica los tres primeros pasos a realizar del flujo de procesos de elaboración de un problema de ML. A partir de este paso, se procede a implementar cada algoritmo bajo sus propios requerimientos.

4.3.3 División del dataset:

Se trabajará con un número de paso de tiempos *n* igual a 2, en este caso la división se realizó de la siguiente forma:

```
Train shape     : (842, 10)
Validation shape: (421, 10)
```

El set de entrenamiento es un tercio del de validación. Esta división también es un proceso genérico, es decir se realiza sin importar el algoritmo seleccionado.

Regresión Lineal

4.4 Construcción del Algoritmo: Regresión lineal

El algoritmo de regresión lineal es quizás uno de los más conocidos y mejor entendidos en estadística y aprendizaje automático, este modela un valor de predicción objetivo basado en variables independientes. Lo atractivo de este modelo es su simple representación, una ecuación lineal que combina un conjunto específico de valores de entrada (x) cuya solución es la salida pronosticada para ese conjunto de valores.

En un problema de regresión simple (con una sola variable explicativa) la ecuación sería:

$$y = B_0 + B_1 \times X$$

En caso se tenga más de una variable independiente estaríamos hablando de una Regresión lineal múltiple. Con la siguiente ecuación:

$$y = B_0 + B_1 \times x_1 + B_2 \times x_2 + \ldots + B_n x$$

Donde:
Y es la variable a predecir o variable dependiente
B_0 es el término de sesgo
$B_{i..n}$ son los parámetros del modelo
$x_{i..n}$son las variables independientes

El objetivo en un algoritmo de regresión lineal es entrenar al modelo para encontrar los parámetros óptimos ($B_{i..n}$) que se acoplen mejor a la data ingresada. Hay diversas formas de determinar el mejor *fit.* Se denomina *Regression Line* a la línea en la cual el error entre los valores obtenidos y los observados es mínima, los valores de error son llamados también residuos.

Para definir y medir el error del modelo se define la siguiente función de costo, calculada como la suma de los cuadrados de los residuos, por ello el modelo toma el nombre de regresión de mínimos cuadrados ordinarios.

$$Z(B) = \frac{1}{m} \sum_{i=1}^{m} (h(x^i) - y^i)^2$$

Donde:
$h(x) = B_0 + B_1 \times x_1 + B_2 \times x_2 + \ldots + B_n \times x_n$ es el valor obtenido valor objetivo previsto utilizando el modelo
y^ies el valor objetivo del set de entrenamiento
m es el número total de datos presentes en el conjunto de datos.

El objetivo entonces es encontrar los parámetros para que la función de costo Z sea mínima. Los parámetros B_0 y B_i se estiman utilizando el set de entrenamiento, existen muchos métodos para estimar estos valores, la elección dependerá de los criterios que se desean utilizar para definir un buen ajuste de los datos de entrenamiento y cómo se desea controlar la complejidad del modelo.
Para los modelos lineales, la complejidad del modelo se basa en la naturaleza de los valores B_i en los datos de entrada. Los modelos lineales más simples tienen un vector B_i que está más cerca de cero, es

decir, donde no se utilizan más variables con coeficiente cero ya que tienen menos influencia en el resultado.

Hasta este punto, no se han utilizado parámetros para controlar la complejidad del modelo, sin importar el valor de B_0 y B_i, el resultado obtenido será siempre una línea recta. Esto puede significar tanto una debilidad como fortaleza del algoritmo como se verá más adelante.

4.4.1 Transformación de los datos:

Preparación de los datos para regresión lineal:

Para que los resultados de la regresión sean insesgados es necesario transformar la data para así lograr tener un mejor ajuste en los datos de entrenamiento. Como parte de esta preparación es necesario que:

- La relación entre las variables sea lineal lo cual puede lograrse con alguna transformación de los datos de entrada.
- Se elimine la colinealidad de las variables
- Se evalúe la distribución de las variables, en caso estas tengan distribución gaussiana la predicción será más confiable.
- Se re escala la data de entrada utilizando estandarización o normalización

En primer lugar, se observó detenidamente la data en base a la descripción estadística de cada atributo ingresado, como se observa en la tabla 4.4:

Tabla 4.4: Descripción de las variables

	PBIChina	PMIChina	...	Inventario	Price
count	1264.000000	1264.000000	...	1264.000000	1264.000000
mean	6.738608	53.999842	...	536346.155854	5825.517801
std	0.331479	3.920101	...	118777.769816	720.922313
min	6.000000	47.200000	...	309351.000000	4331.000000
25%	6.500000	51.175000	...	458500.000000	5296.875000
50%	6.900000	53.000000	...	519976.500000	5881.000000
75%	6.900000	57.700000	...	577749.000000	6282.250000
max	7.100000	60.800000	...	923213.000000	7332.000000

Como se observa, la escala para las variables es muy dispersa, en especial comparando los valores de PBI y PMI con los de Inventarios. Asimismo, esta última variable posee una desviación estándar muy elevada. Se puede observar con mayor énfasis la distribución de cada variable en los siguientes histogramas:

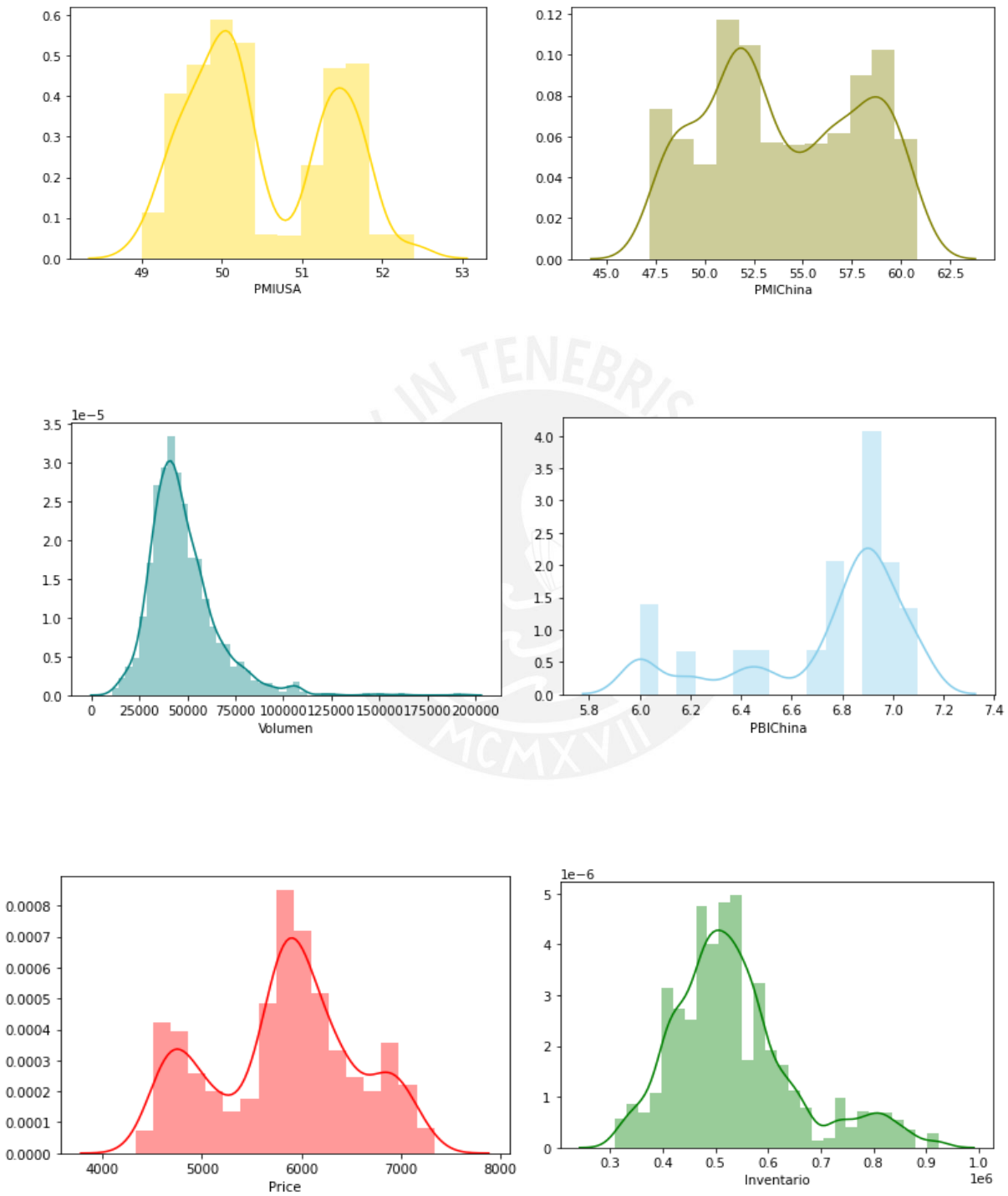

Figura 4.5: Histogramas de las variables de entrada: PMI USA, PMI China, PBI China, Volumen, Inventario y Precio

Aparentemente, ninguna de las variables posee una distribución normal o gaussiana, salvo el volumen transado, y en menor medida el Inventario y el Precio. Asimismo, como se mencionó, la escala de las variables es muy distorsionada. Frente a esto se sometió a los valores a un proceso de normalización.
La normalización es una técnica que a menudo se aplica como parte de la preparación de datos para el uso de algoritmos de *Machine Learning*. El objetivo de la normalización es cambiar los valores de las columnas numéricas en el conjunto de datos para usar una escala común, sin distorsionar las diferencias en los rangos de valores o perder información.

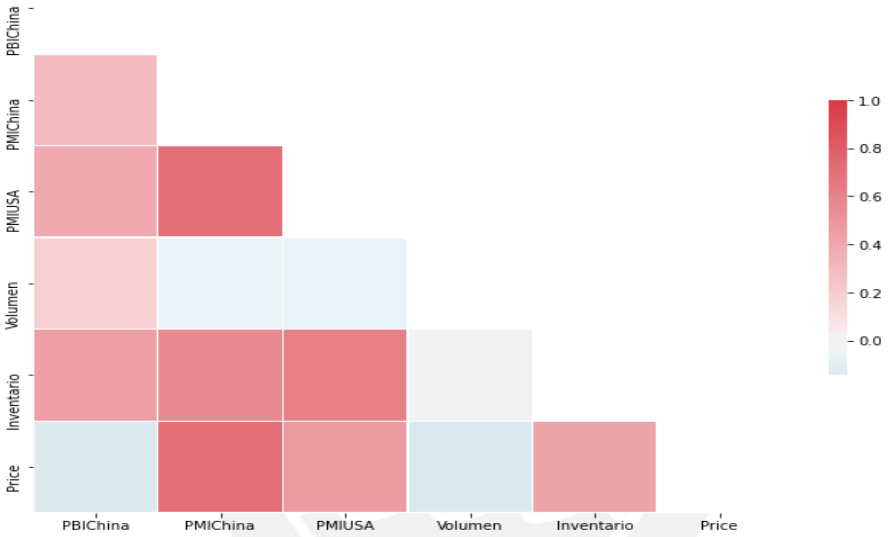

Figura 4.6: Matriz de correlación

Es primordial evaluar la multicolinealidad entre las variables, en otras palabras, la correlación entre las variables independientes.

Según la matriz de correlación, las únicas variables con una correlación considerable son el PMI de Estados Unidos y el de China. Asimismo, el PBI de China y el Volumen tienen una correlación baja con el Precio del cobre. No obstante, a pesar de que aparentemente no exista una fuerte correlación entre las variables y la variable a pronosticar, se debe tomar en cuenta que al hacer un análisis individual se omiten posibles combinaciones entre las variables que internamente pueden tener un efecto explicativo interesante en el modelo; se profundizará esta idea en el siguiente capítulo.
En este sentido, puede ser que una variable por sí sola no explique aparentemente el comportamiento del precio del cobre, no obstante, cuando está en conjunto con otras puede ser muy relevante.

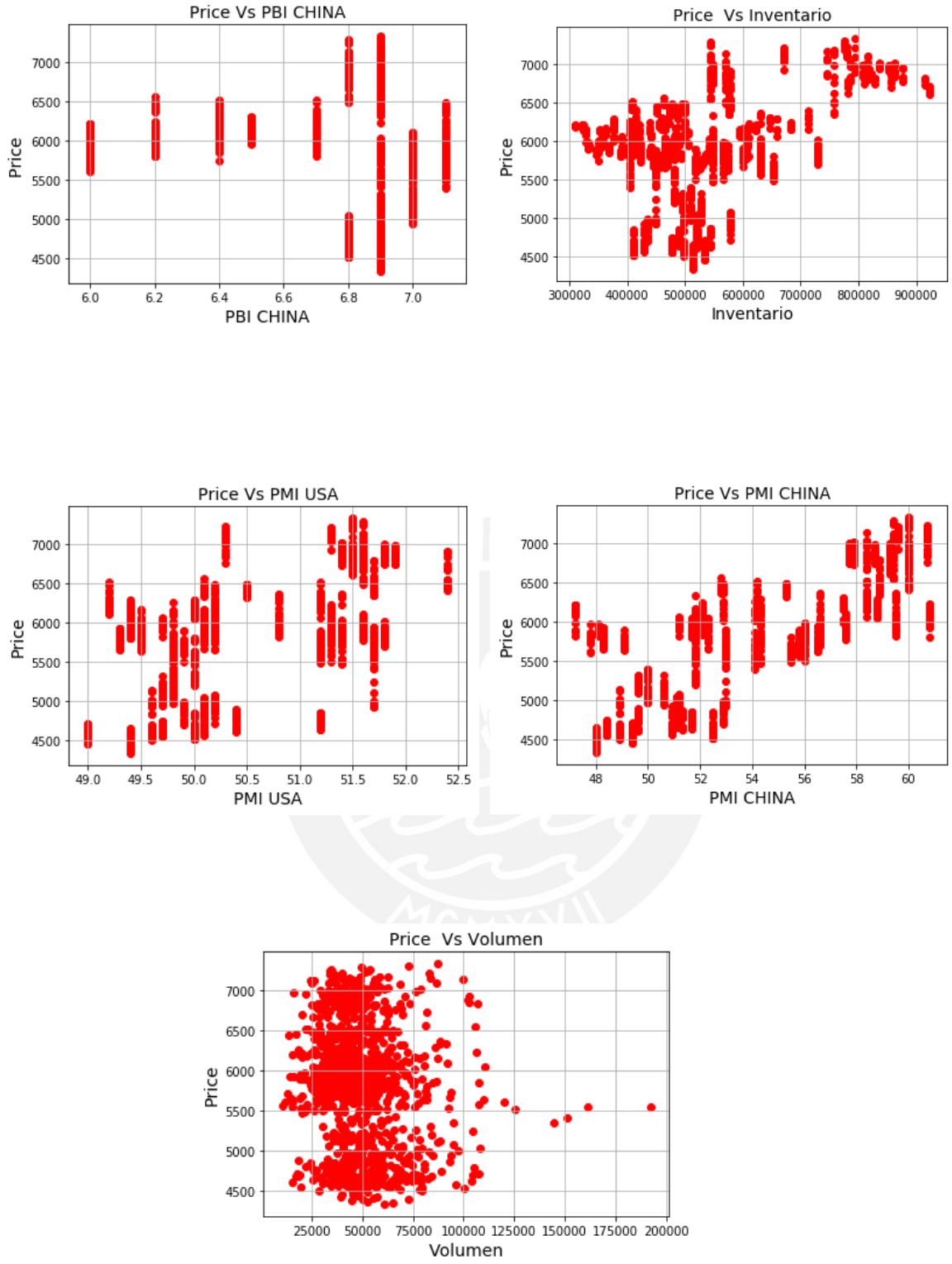

Figura 4.7: Gráficos de Linealidad: PMI USA, PMI China, PBI China, Volumen, Inventario y Precio

Antes de realizar la normalización, se debe recordar que ya se realizó la división del conjunto de datos en los *train set y test set* en el inciso 4.3.3. Estos serán los conjuntos de datos destinados a entrenar el modelo y posteriormente probarlo. Cada set está compuesto por pares de valores ($x_{i..n}$; y^i).
La librería de Python *Sklearn* permite transformar los datos con ayuda de la función Standard Scaler, la cual ajusta los parámetros transformadores a *test* set de las variables independientes ($x_{i..n}$.) y devuelve una versión transformada de los valores de $x_{i..n}$. Es por ello que en el flujograma (figura 4.1), existe un bucle entre la transformación de los datos y el set de validación, al cual se le aplica la transformada en base al ajuste realizado al set de entrenamiento.

4.4.2 Entrenamiento del modelo:

El algoritmo crea y ajusta la función objetivo de regresión lineal utilizando los datos de entrenamiento X_train y los valores correspondientes de Y_train, obteniendo los siguientes parámetros:

```
Linear model intercept (B_0):[5845.044536]
Linear model coeff (B_i): [[-331.07203699  379.6328856   50.85129861   -
6.12996995  -50.00893552  34.61277742  137.1377024   -59.3186314   -
10.33122549  212.491952]]
```

Con los valores obtenidos la ecuación de la regresión sería:

$$y = 5{,}811.9 - 320.9 \times x_1 + 508.3 \times x_2 + 14.3 \times x_3 - 33.9 \times x_4 + 156.2 \times x_5$$

Donde:
$x_1$toma los valores de la variable independiente PBI China
$x_2$toma los valores de la variable independiente PMI China
$x_3$toma los valores de la variable independiente PMI USA
$x_4$toma los valores de la variable independiente Volumen
$x_5$toma los valores de la variable independiente Inventario

4.5 Evaluación: Regresión Lineal

Validación

En primera instancia se obtuvieron los siguientes puntajes para la corrida del modelo:

```
Scores:
Train      = 0.6725
Validation = 0.6157
Mean Absolute Error: 336.0
Mean Squared Error: 450.0
```

Líneas arriba, cuando se mencionó acerca de los criterios para entrenar un algoritmo se hacía referencia a la forma en cómo se va a controlar la complejidad del modelo y la exactitud de la predicción obtenida de este, es importante poder establecer un balance entre ambas características.

Puede creerse, que un modelo con el más bajo valor de Suma de cuadrados residual es el mejor, no obstante los modelos de regresión lineal pueden tender a forzar el ajuste de datos de entrenamiento para cumplir con la minimización de la función objetivo en este caso, ocasionando que el modelo no funcione bien sobre los datos de validación porque está construido para los datos de entrenamiento de manera tan específica que puede no ajustarse a los datos nuevos, a esto se le llama *"overfitting"*

En este caso, analizando los puntajes obtenidos, se puede observar que en el caso del set de entrenamiento es de 0.6725 y para el de validación de 0.6157 estos puntajes son obtenidos calculando el R cuadrado de ambos sets. Se puede decir que los puntajes son cercanos, lo cual es un buen indicador de que no se ha cometido *"overfitting"* en el modelo.

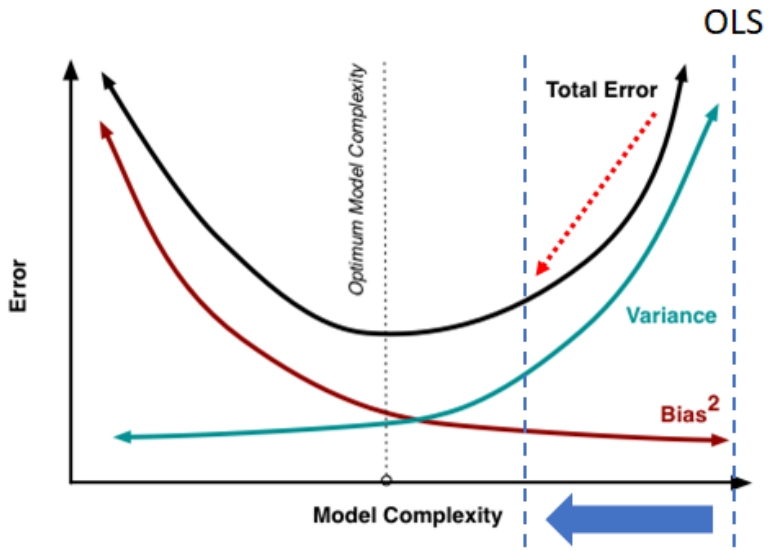

Figura 4.8: Gráfico de complejidad vs exactitud.

Elaboración: (Towards Science, 2019)

En la figura 4.8, se grafica en el eje Y el error también entendido como una medida de exactitud del modelo y el eje X representa la complejidad del modelo. La curva de sesgo está relacionada con un modelo que no se ajusta al conjunto de entrenamiento y la de varianza se relaciona con un modelo que no se ajusta al conjunto de prueba. El sesgo y la varianza están en una relación de compensación sobre la complejidad del modelo, lo que significa que un modelo simple tendría un alto sesgo y una baja varianza, y viceversa.

Como se puede observar, el error total, calculado como la suma de varianza y sesgo tiene un mínimo absoluto en un valor medio de complejidad. En el modelo revisado, de regresión de mínimos cuadrados o OLS por sus siglas en inglés, se trata a todas las variables por igual, por lo tanto, un modelo OLS se vuelve más complejo a medida que se agregan nuevas variables y puede ocasionar un *"overfitting"*, por ellos este tipo de modelos está ubicado en la parte derecha del gráfico, con el sesgo más bajo y la varianza más alta.

4.6 Optimización del modelo de Regresión Lineal

En el caso de la regresión lineal, no se aplica la optimización de hiper parámetros tal y como se presenta en la figura 11, dado que dentro de la estructura de la regresión lineal no existen hiper parámetros, por ello se ha procedido a optimizar el modelo empleando variaciones de la regresión lineal.

4.6.1 Ridge Regression:

Ridge Regression es una variación del algoritmo de regresión de mínimos cuadrado y se utiliza también para calcular los parámetros B_0 y B_i, la diferencia principal con el algoritmo previamente descrito es la siguiente: Durante la fase de entrenamiento, agrega una penalización por los valores de los parámetros B_i que son demasiado grandes como se muestra en la siguiente ecuación:

$$Z(B) = \frac{1}{m}\sum_{i=1}^{m}(h(x^i) - y^i)^2 + \alpha\sum_{j=1}^{p}B_j^2$$

La función objetivo es minimizar la función Z, los valores de B_i que son muy grandes ocasionan que la suma de los cuadrados sea mayor, por lo tanto, el segundo término de la ecuación actúa penalizando a los parámetros con valores muy elevados. Una vez que la *Ridge Regression* estima los parámetros del modelo, la predicción de los valores y^i sigue el mismo proceso que en el caso de la regresión lineal convencional.

La adición del término de penalización a la función objetivo de un algoritmo de aprendizaje se llama "Regularización", concepto clave en los modelos de Machine Learning. La regularización es una forma eficiente de evitar el "*overfitting"* y por lo tanto mejorar el rendimiento de generalización probable de un modelo, restringiendo las configuraciones de parámetros posibles de los modelos, lo cual resulta por lo general en la reducción de la complejidad del modelo estimado final.

El efecto de la regularización se visualiza con mayor claridad en modelos de regresión lineal con un número muy grande de variables independientes, el parámetro que controla la regularización es α, un valor de alfa mayor significa mayor regularización y modelos lineales más simples con parámetros cercanos a ceros. La configuración predeterminada para alfa es 1.

Se realizó la prueba con distintos valores de alfa obteniendo los siguientes resultados:

Tabla 4.5: R^2 **de entrenamiento y validación para cada alfa**

Alfa (α)	R^2 set entrenamiento	R^2 set validación
0.5	0.672396	0.615871
1	0.672378	0.615800
2	0.672348	0.615626
5	0.672196	0.614980
15	0.670156	0.612201
20	0.670201	0.610602

En este caso, el mejor resultado de R^2 para el set de validación se obtiene con un alfa de 1, sin embargo, la mejora obtenida en realidad es de poca significancia. Esto se debe a que el modelo que estamos utilizando es de complejidad baja, teniendo solo 5 parámetros independientes, y poca tendencia a que se genere el sobre ajuste o *"overfitting"* al set de entrenamiento. De todas formas, si se evalúa añadir más variables al modelo, este algoritmo sería de gran utilidad para evitar incrementar la complejidad del modelo.

4.6.2 Inclusión de nuevas variables explicativas

Cómo se mencionó al inicio del capítulo, esta investigación es de tipo exploratoria, en este punto del proceso vale la pena reflexionar acerca de los pasos previos a la elección y corrida del modelo. Por ello, se optó por incrementar el número de variables en el modelo. En esta oportunidad se utilizarán todos los factores del universo considerado en el acápite 4.2.2, sin considerar aquellas variables generadas posteriormente.

La etapa de Análisis de Datos se realiza de igual forma a la expuesta anteriormente, se pre-procesa y divide la data obteniendo:

```
Train shape     : (842, 18)
Validation shape: (421, 18)
```

Transformación de los datos:

Como parte de la etapa "Construcción del modelo", nuevamente se aplican procedimientos de transformación de datos para el correcto ajuste y entrenamiento de este.

En la figura 4.9, se puede observar los histogramas correspondientes a cada nueva variable. Se distingue que el índice del dólar es la única variable con distribución aparentemente Gaussiana. Como se explicó, es conveniente que las variables sigan este tipo de distribución, pues de esta forma el modelo se ajustará mejor. No obstante, no es un requisito limitante para aplicar el algoritmo de regresión lineal.

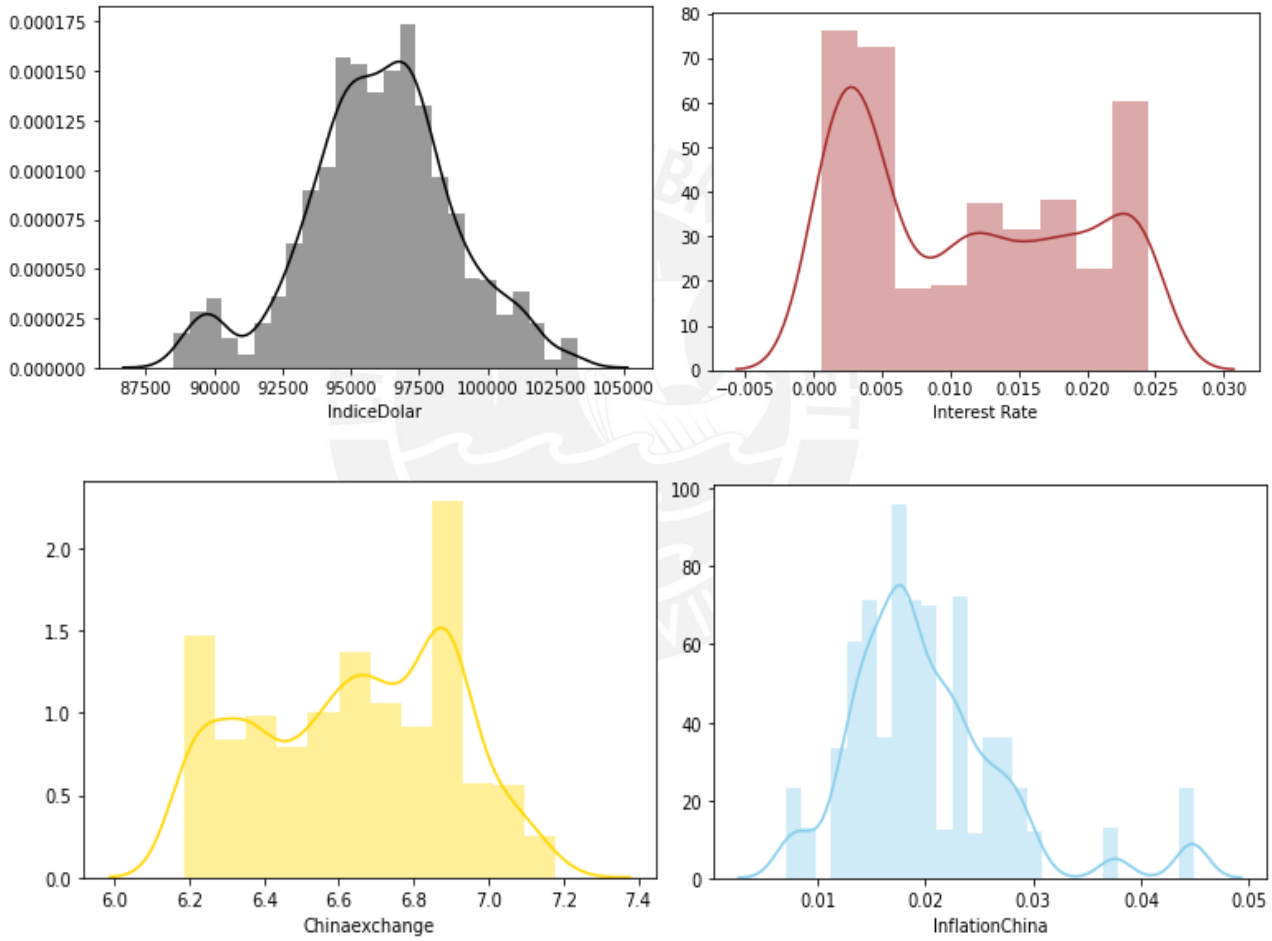

Figura 4.9: Histogramas de las variables de entrada: Índice del Dólar, Tasa Interés FED, Tipo de Cambio YUAN/S$, Tasa de inflación China

Al igual que con las variables iniciales, se efectúa una normalización en la base de datos utilizando la función *"StandardScaler"* de la biblioteca *Sklearn.*

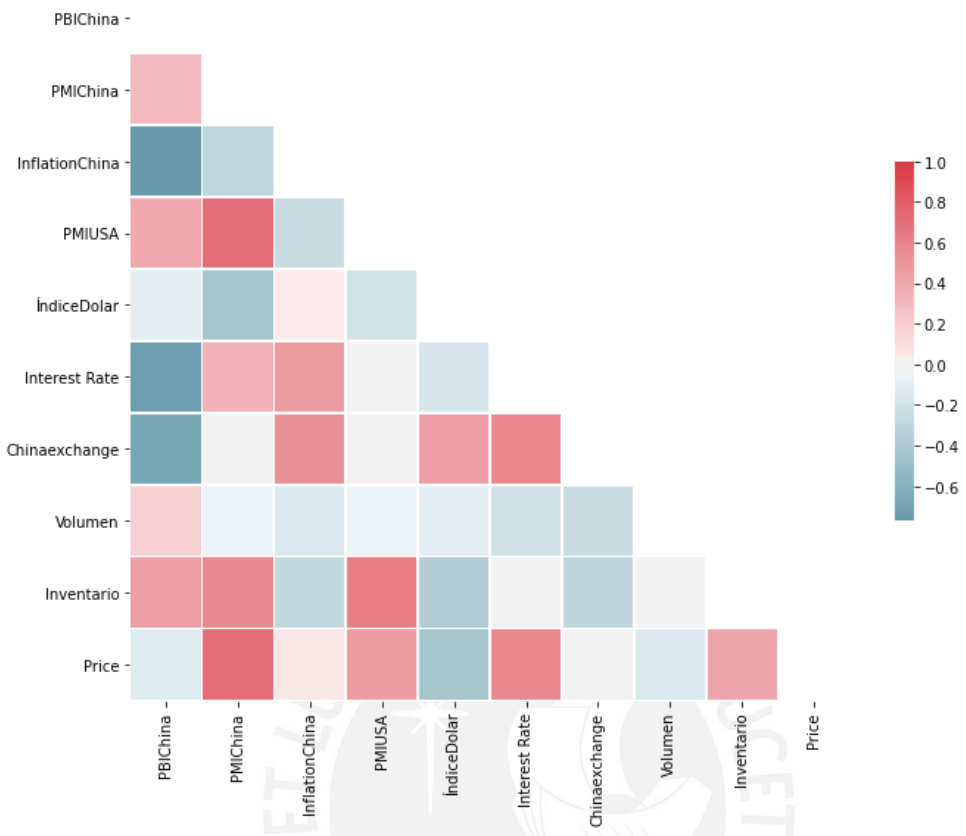

Figura 4.10: Matriz de correlación

En la figura 4.10, se puede apreciar, que el índice del dólar tiene una correlación negativa con el precio como era de esperarse; además la correlación entre la tasa de interés de la FED y el precio futuro del cobre es positiva. Las otras dos variables incluidas, tasa de inflación China y tasa de cambio entre el Yuan y el dólar presentan una baja correlación con la variable *target*.

En este caso la división de la data para entrenar y validar el modelo se realizó de la siguiente manera:

```
Train shape     : (842, 18)
Validation shape: (421, 18)
```

La regresión estimó los valores de parámetros:

```
Linear model intercept (B_0):[5806.8563]
Linear model coeff (B_i): [[-203.82205207  166.08058116    1.55601009
222.52415389   70.36894157  94.7996329   -388.25150246   -47.39331482
```

```
108.67058262  -291.79575956  252.84172301     51.56250617     92.41175509
120.35152249  148.70864409 -194.08903762  -11.61295305 -101.75636624]]
```

Con los valores obtenidos la ecuación sería:

$$y = 5{,}811.9 - 503.4 + 528.5.3 \times x_2 + 53.7 \times x_3 + 312.1 \times x_4 + 187.6 \times x_5 + 235.0 \times x_6 - 580.3 \times x_7 - 45.0 \times x_8 + 7.14 \times x_9$$

Donde:

$x_1$toma los valores de la variable independiente PBI China

$x_2$toma los valores de la variable independiente PMI China

$x_3$toma los valores de la variable independiente índice de inflación China

$x_4$toma los valores de la variable independiente PMI USA

$x_5$toma los valores de la variable independiente Índice Dólar

$x_6$toma los valores de la variable independiente Tasa de Interés

$x_7$toma los valores de la variable independiente China Exchange Rate

$x_8$toma los valores de la variable independiente Volumen

$x_9$toma los valores de la variable independiente Inventario

Evaluación:

```
Scores:
Train      = 0.8164
Validation = 0.7810
Mean Absolute Error: 281.0
Mean Squared Error: 345.4
```

En este caso, analizando los puntajes obtenidos, se puede observar que en el caso del set de entrenamiento es de 0.82 y para el de validación de 0.78. Se logra visiblemente una mejora en el ajuste del modelo tanto al set de entrenamiento como al de validación, comparándolos con los puntajes obtenido previamente al trabajar con 5 variables (0.67 y 0.62 para cada set respectivamente). Además, MAE y MSE disminuyen en 20% aproximadamente. En consecuencia, se logró optimizar el modelo al incluir cuatro variables que incluso individualmente presentaban correlaciones negativas con el precio como se presentó en la figura 20.

Como línea base se tiene el modelo de regresión lineal con el cual se obtuvieron resultados considerablemente buenos a pesar de ser un algoritmo simple y sin hiper parámetros. Como parte de la naturaleza exploratoria de esta tesis, se utilizará como segundo algoritmo el *Support Vector Regression*.

Support Vector Regression

Support Vector Machine (SVM) es un modelo de aprendizaje automático muy potente y versátil, capaz de realizar clasificaciones lineales o no lineales, de regresión e incluso puede realizar la detección de valores atípicos. Es uno de los modelos más populares en Machine Learning. Los SVM son particularmente adecuados para la clasificación de conjuntos de datos complejos, pero de volumen limitado.

El objetivo principal de este algoritmo es encontrar un hiperplano en un espacio p-dimensional, donde "p" es el número de características consideradas, que clasifica de manera eficiente los datos observados. Se podría decir que un hiperplano es un límite de decisión, y el óptimo es aquel que tiene el margen máximo *(Maximal Margin Classifier)*, o distancia máxima entre los puntos de datos de diferentes clases. Al maximizar la distancia, el modelo proporciona mayor precisión en la clasificación de futuros datos.

La definición matemática de un hiperplano es la siguiente:

$$\beta 0 + \beta 1x1 + \beta 2x2 = 0$$

Todos los puntos definidos por el vector $(x = x1, x2, \ldots, xp)$ que cumplen con la ecuación pertenecen al hiperplano. En caso un punto no satisfaga la ecuación se evalúan las siguientes dos opciones:

$$\beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta pxp < 0$$
$$\beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta pxp > 0$$

El punto x caerá a un lado u otro del hiperplano. Quedando claro que este divide un espacio p dimensional en dos partes.

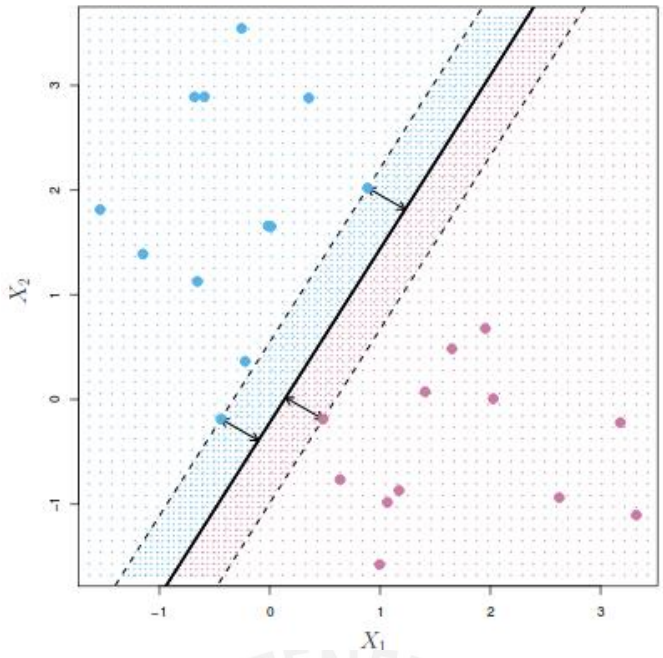

Figura 4.11: Maximal Margin Hyperplane

Fuente: (ISLR 2019)

En la figura 4.11, se muestra el *maximal margin hyperplane* para un conjunto de datos de entrenamiento, los tres valores distinguidos en el plano se encuentran equidistantes respecto al margen máximo y definen la anchura de este. A cada una de estas observaciones se les llama vector de soporte. Hay diversos casos para la división de hiperplanos, ya sean perfectamente separables linealmente, cuasi-separables linealmente o casos en los que una separación lineal no es viable.

Si bien el *Maximal Margin Classifier* es una herramienta útil y fácil de entender, en la mayoría de las situaciones reales, los datos no se pueden separar linealmente de forma perfecta, por lo que no existe un hiperplano de separación y no puede obtenerse un *maximal margin hyperplane*. En estos casos se emplea el clasificador denominado *Soft Margin Classifier* que basándose en el mismo concepto logra obtener una separación de las clases que no llega a ser perfecta, pero permite obtener mayor robustez en el modelo, y una mejor clasificación de la mayoría de las observaciones de entrenamiento y validación, pues no tiende a causar *overfitting*. Por ello se puede decir que el SVM tiene capacidad de ignorar valores atípicos y encontrar el hiperplano con margen máximo de igual forma.

A diferencia de la regresión lineal, primer algoritmo utilizado, en la construcción de un SVM se incluyen algunos hiper parámetros de ajuste, en primer lugar, el parámetro de regularización C, encargado de controlar la severidad permitida de las violaciones o clasificaciones erróneas de las n observaciones sobre el margen y el hiperplano. Cuando el valor de $C > 0$, se tiene que no más de C

muestras pueden encontrarse en el lado incorrecto del hiperplano, cuando C es grande, los márgenes son estrechos pues serán pocas las observaciones que podrán encontrarse mal ubicadas, pudiendo ocasionar sobreajuste. A menores valores de C mayor es la tolerancia del margen, siendo este más ancho y por lo tanto existirán mayores vectores de soporte. La selección de C por lo tanto, incide directamente en la varianza y *bias* del modelo.

Por otro lado, en los casos de división de hiperplanos no lineales, se utilizan *Kernels,* hiper parámetros de SVM que funcionan transformando un espacio de pocas dimensiones en uno con mayores y más complejas divisiones. La selección del Kernel va a influir directamente en la complejidad del algoritmo y en la flexibilidad del modelo.

Otro hiper parámetro a tomar en cuenta en los SVM es Gamma, usado en hiperplanos no lineales, a mayor valor de gamma el modelo intentará ajustarse de manera exacta al conjunto de datos de entrenamiento. La definición del valor de gamma incide en la influencia o efecto de cada vector de soporte sobre el área o región delimitada (dependiendo del Kernel utilizado), a un valor de gamma mayor, el radio de influencia de cada vector de soporte disminuye, obteniendo áreas muy definidas. Por el contrario, cuando gamma es muy pequeño, el modelo es limitado y no puede capturar la complejidad o la "forma" de los datos.

La elección de estos tres hiper parámetros se realiza en la fase de construcción del algoritmo, una vez validados los resultados obtenidos los hiper parámetros se optimizan. Se verá en detalle el flujo de este proceso en el capítulo 6.

Se ha explicado de manera general los principios de funcionamiento de los *Support Vector Machine*, y si bien es cierto, en la mayoría de los casos se emplea este algoritmo para problemas de clasificación, se aplica también en problemas de regresión manteniendo básicamente los mismos principios, con alguna diferencia menor.
El propósito de este subcapítulo es detallar el proceso de construcción del modelo utilizando un segundo algoritmo, en este caso SVR, para posteriormente comparar los resultados con el modelo base.

En los SVR se busca ajustar la mayor cantidad de muestras entre un margen limitado por un valor ε(epsilon), un hiper parámetro denominado margen de tolerancia, usado para poder trabajar con datos continuos. El valor de epsilon define por lo regiones donde se deben encontrar las observaciones.

En general el flujo de procesos es el mismo que el realizado para la regresión lineal, la categorización del problema es en principio el mismo, las variables a utilizar serán las 9 presentadas en el subcapítulo 4.6.2, asimismo la etapa de Análisis de datos se mantiene igual, pues el pre procesamiento y división del conjunto de datos se trabaja de igual forma independientemente del algoritmo a utilizar.

4.7 Construcción del Algoritmo: Support Vector Regression

Cómo ya se detalló, esta fase abarca desde la transformación de los datos hasta la optimización de los hiper parámetros. A diferencia de la regresión lineal, en este caso sí se tomará en cuenta el proceso de optimización pues tanto el valor de gamma, C y Kernel han de ser determinados.

4.7.1 Transformación de los datos:

Preparación de los datos para SVR:

Al igual que con la Regresión Lineal, se trabajará con un número de paso de tiempos *n* igual a 2, en este caso la división se realizó de la siguiente forma:

```
Train shape     : (842, 10)
Validation shape: (421, 10)
```

Con el fin de tener mejores resultados, los datos se deben estandarizar utilizando la función "StandardScaler" de la biblioteca sklearn.

4.7.2 Entrenamiento del modelo:

Primero se ajustan los datos de entrenamiento a la función objetivo del SVR teniendo como parámetros iniciales los siguientes:

```
regressor = SVR(kernel='linear', C=1,gamma='scale')
```

Se utiliza como instancia base un kernel lineal, con el cual no se transforman los datos a dimensiones más complejas pues se asumen planos lineales. Asimismo, el valor de C se establece como 1,epsilon como 0.1, y gamma como *"scale*", es decir se calcula como 1/(número de variables*varianza(X)).

4.8 Evaluación: Support Vector Regression

Validación

Al correr el modelo se obtienen los siguientes resultados

```
Scores:
Train      = 0.7907
Validation = 0.807
Mean Absolute Error: 257.6
Mean Squared Error: 299.1
```

Utilizando los hiper parámetros expuestos previamente se obtiene un puntaje de 0.79 para el set de entrenamiento y 0.8 para el de validación. Asimismo, el MAE es de 257.6 y el MSE de 299.1. Realizando una comparación con los resultados de la Regresión Lineal, el puntaje del set de entrenamiento es mayor para el primer algoritmo (0.82), no obstante, el SVR parece haber logrado un mejor ajuste en la validación frente a la RL (0.78). Considerando los valores del MAE y MSE, en ambos casos el SVR logra resultados más precisos.

5 Capítulo V: Análisis y selección óptima de atributos

Hasta el momento se ha comprobado la capacidad predictiva del algoritmo de regresión lineal y SVR, realizando además la comparación del desempeño del modelo de 5 y el modelo de 9 variables para el caso de regresión lineal. Si bien, como se mencionó esta selección fue aleatoria, fue un punto de partida para verificar cómo cambian los resultados al variar el conjunto de datos de ingreso, no obstante, para realizar una selección óptima de atributos se debe recurrir a herramientas que permitan analizar los factores o características a emplear, en este capítulo se explorarán algoritmos con dichas capacidades.

5.1 Análisis de componentes principales

A medida que aumenta el número de variables consideradas en el modelo, el número de muestras también aumenta proporcionalmente. Esto ocasiona que el entrenamiento sea cada vez más lento y al mismo tiempo complejo. Frente a este problema llamado *"the curse of dimensionality"* se recomienda recurrir a la reducción de la dimensionalidad.

Existen algoritmos que reducen la dimensionalidad de un modelo, llamados algoritmos de selección de características, estos tienen como función analizar los datos de entrada, clasificarlos en subconjuntos y definir una métrica con la cual se valorará la relevancia de la información proporcionada por cada componente. Aquellos componentes que brindan menos información son entonces descartados, permitiendo un ahorro en el almacenamiento de datos y tiempo de ejecución.

Principal Components Analysis

PCA es el algoritmo de reducción de dimensionalidad más utilizado, este procesa los datos y los analiza para encontrar la estructura de la información contenida en estos. La estructura está definida por los componentes con mayor varianza. La clave está en reducir la dimensionalidad manteniendo las variables que contienen la mayor proporción de varianza en los datos para así conservar la información relevante de los datos.

A partir del conjunto de datos del set de entrenamiento de *n* dimensiones se identifican como autovectores (o componentes principales), a la dirección de mayor varianza en los datos. Los autovectores elegidos definen un nuevo sistema de ejes sobre el cual se proyectan los datos, con la

finalidad de verlos con más claridad dado que los nuevos ejes son las direcciones en las que el aporte de información es el más significativo. El valor de la varianza representada por los autovectores se define como autovalor.

El proceso de aplicación de este algoritmo requiere que la data este normalizada o estandarizada. Para encontrar los componentes principales se calcula primero la matriz de covarianza:

Tabla 5.1: Matriz de covarianza

1.00	0.30	-0.77	0.40	-0.10	-0.75	-0.68	0.20	0.45
0.30	1.00	-0.29	0.70	-0.43	0.34	-0.03	-0.05	0.56
-0.77	-0.29	1.00	-0.25	0.05	0.47	0.52	-0.14	-0.28
0.40	0.70	-0.25	1.00	-0.21	-0.01	0.04	-0.06	0.62
-0.10	-0.43	0.05	-0.21	1.00	-0.17	0.45	-0.09	-0.38
-0.75	0.34	0.47	-0.01	-0.17	1.00	0.57	-0.21	-0.02
-0.68	-0.03	0.52	0.04	0.45	0.57	1.00	-0.25	-0.30
0.20	-0.05	-0.14	-0.06	-0.09	-0.21	-0.25	1.00	0.00
0.45	0.56	-0.28	0.62	-0.38	-0.02	-0.30	0.00	1.00

El siguiente paso es el cálculo de los autovectores y autovalores

Autovectores:

[-0.49	-0.20	-0.19	-0.01	0.69	-0.33	0.29	-0.09	0.00]
[-0.28	0.50	-0.04	0.13	-0.40	-0.19	0.55	-0.12	-0.37]
[0.41	0.17	0.20	0.00	0.11	-0.06	0.53	-0.22	0.65]
[-0.31	0.39	-0.33	0.20	0.09	0.49	-0.26	-0.46	0.28]
[0.21	-0.26	-0.68	0.28	-0.07	0.31	0.36	0.35	0.00]
[0.28	0.50	0.19	0.06	0.56	0.30	0.04	0.33	-0.35]
[0.38	0.27	-0.39	0.31	0.07	-0.64	-0.34	-0.09	-0.01]
[-0.12	-0.20	0.41	0.88	0.00	0.01	-0.02	0.04	0.03]
[-0.36	0.32	-0.01	-0.02	-0.11	-0.15	-0.12	0.70	0.48]

Autovalores:

```
[3.45        2.26        1.19        0.84        0.02        0.11        0.19        0.37        0.57]
```

Como se desea reducir la dimensionalidad del *dataset*, se procede a descartar aquellos autovectores que poseen autovalores más bajos, siendo ellos los que proporcionan menos información al modelo. Para ello se ordenan los autovalores en orden descendente y el siguiente paso es decidir cuál es el número óptimo de componentes principales con los que se puede describir mejor la varianza en el conjunto de datos. Se utiliza como métrica la "varianza explicada", la cual mostrará la variabilidad atribuida a cada componente.

En la figura 5.1, se aprecia que más del 60% de la varianza es atribuida a los dos primeros componentes, la tercera acumula aproximadamente 15%, y la cuarta un 10%. A simple vista, con las cinco primeras componentes la varianza acumulada alcanza el 85%.

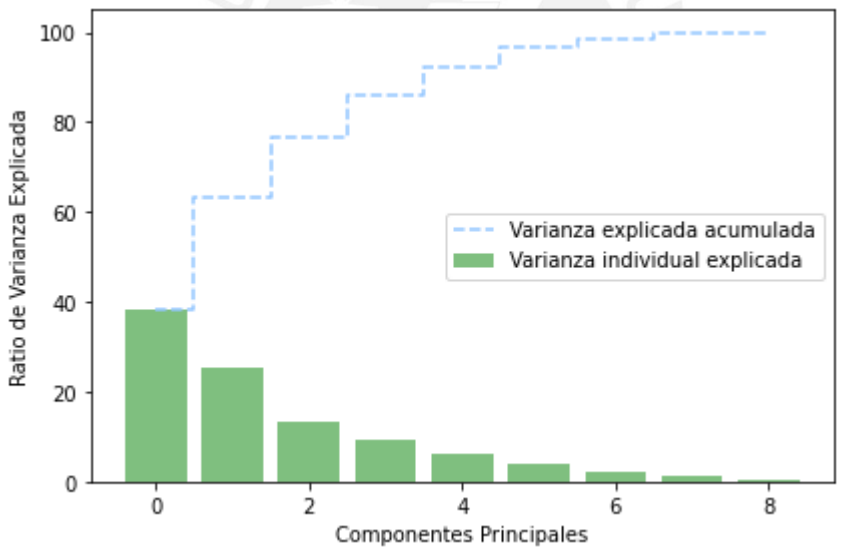

Figura 5.1: Gráfica de varianza explicada.

Si bien es cierto, el análisis de componentes principales puede servir como herramienta para examinar la variabilidad del modelo y seleccionar los componentes que describen mejor la varianza en un conjunto de datos, no considera el orden temporal de las muestras, simplemente descompone los datos de entrada en direcciones donde se observa la mayor variabilidad, no obstante no hay seguridad que los componentes principales sean realmente informativos al momento de predecir la variable dependiente.

Se debe tener cuidado y especial precaución en la selección de componentes en un problema de series temporales, en el que la correlación de las variables entre cada periodo muestral provee información relevante al modelo y no necesariamente está expresada en relaciones o dependencias lineales entre los atributos. PCA se puede utilizar como una base en el análisis, no obstante, es recomendable incluir análisis de selección de atributos, que sí tome en consideración el valor objetivo del modelo predictivo de modo que clasifique las variables de entrada en términos de utilidad en la predicción de la variable *target.*

Como se planteó en la introducción, el fin de esta investigación es poder utilizar los resultados en pronósticos reales y del día a día en el mercado del cobre, además de servir de base para próximas investigaciones que profundicen en el impacto de este commodity en los diversos sectores del país. Teniendo esto en consideración, es indispensable realizar un análisis de las características que son realmente relevantes en el desarrollo de los algoritmos para así poder explorar óptimas formas de obtención de datos en un contexto real para el pronóstico, o en todo caso profundizar acerca del comportamiento independiente de cada variable. El análisis que se debe realizar en este caso se llama "Selección de Características" *(Feature Selection).*

5.2 Feature Selection Analysis

Este análisis, consiste en la selección automática de variables que son más relevantes en el modelo predictivo para un problema específico. Tal como especifican Guyon y Elisseeff (2003) "El objetivo de la selección de variables es triple: mejorar el rendimiento de predicción de los predictores, proporcionar predictores más rápidos y rentables, y proporcionar una mejor comprensión del proceso subyacente que generó los datos." Teniendo los atributos correctamente seleccionados, se puede incrementar considerablemente la precisión de los resultados al mismo tiempo que se disminuye la complejidad del modelo volviéndolo más fácil de entender y explicar.

Si se recuerda, en la matriz inicial de correlación de la figura 16, el PBI de China, y el Volumen transado presentan un coeficiente de correlación bajo, igualmente en el caso de la matriz de correlación de las 10 variables (Figura 20) la tasa de cambio entre el Yuan y el dólar, así como la tasa de inflación China tienen de igual manera un coeficiente de correlación menor a 0.2. No obstante, a pesar de tener predictores con aparente correlación débil, el resultado finalmente obtenido es significativamente bueno. Por ello, se reitera la idea mencionada en la sección 4.3.1: Existen variables que por sí solas no explican nada aparentemente pero cuando se utilizan en conjunto con otras el resultado es positivo.

Para encontrar la mejor solución global, es decir, el subconjunto de predictores con el mejor rendimiento, se requeriría evaluar todas las posibles combinaciones de estos subconjuntos, lo cual puede llegar a ser computacionalmente inviable. Asimismo, en muchos modelos la forma en que se relaciona cada atributo con el resultado puede ser muy compleja, llegando a ser realmente engorroso encontrar la relación entre el predictor individual y el *target.*
Existen tres métodos de selección de atributos clasificados en: intrínsecos o integrados, métodos de filtro y métodos de envoltura.

Los métodos intrínsecos incorporan la selección de características en el proceso de modelado, es decir la estructura per sé del algoritmo conlleva a la búsqueda del mejor predictor, dos ejemplos claros son los modelos basado en árboles de decisión y los modelos de regularización como el Ridge Regression presentado en el capítulo previo, que penaliza los coeficientes de aquellos predictores que no tienen relevancia. Los métodos intrínsecos poseen ventajas como la rapidez de corrida al tener integrado el proceso de selección al proceso de entrenamiento y ajuste del modelo, evitando así el reflujo a la selección del conjunto óptimo de variables. El vínculo directo entre la selección y el objetivo del modelo (la función objetivo) hace que sea más fácil tomar decisiones informadas entre la posible escasez de características y el rendimiento predictivo. El inconveniente de este tipo de herramientas es que la selección depende de la estructura del modelo clasificador y puede que no trabajen de manera óptima con cualquier otro modelo.

Por otro lado, cuando un modelo no tiene un proceso de selección de características intrínsecas requiere de un procedimiento previo que mejore el rendimiento predictivo, en este caso se utilizan los otros dos métodos: filtro y envoltura
.

Los métodos de filtro realizan un análisis supervisado inicial de los predictores, generalmente como parte del pre-procesamiento, para determinar cuáles son importantes y luego solo los proporcionan al modelo. La selección se realiza de manera independiente de cualquier algoritmo de aprendizaje automático, las características son seleccionadas en función a puntajes obtenidos de pruebas estadísticas que evalúan su correlación con la variable a predecir. Se puede visualizar este procedimiento en la figura 5.3.

En cuanto a los métodos de envoltura, estos se ejecutan al seleccionar un subconjunto de características y entrenar un modelo con este. Posteriormente, según las inferencias que se extraen del modelo anterior, se decide agregar o eliminar características del subconjunto. Este tipo de selector tiende a tener un mejor desempeño en la selección de características ya que prueba y ajusta de manera iterativa recibiendo una

evaluación comparativa constante entre cada iteración. Sin embargo, ante modelo complejos, la utilización de métodos de envoltura involucra altos costos de procesamiento

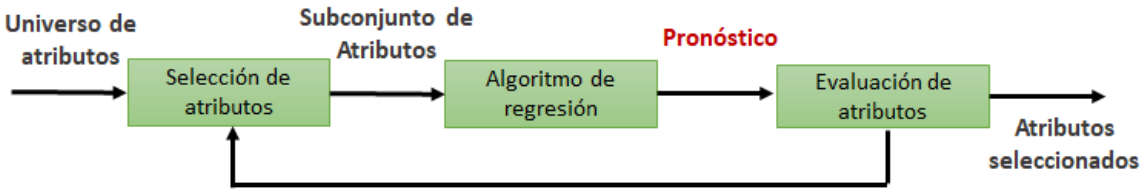

Figura 5.2: Método de envoltura para la selección de características

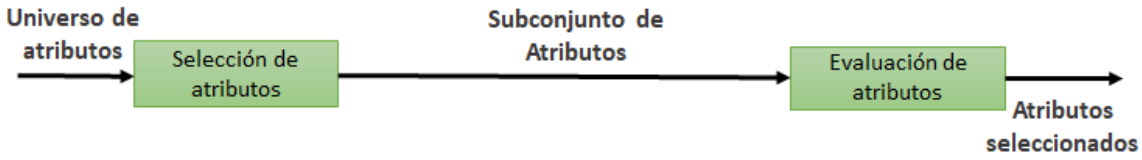

Figura 5.3: Método de filtro para la selección de características

Sintetizando, al realizar una comparación de los tres métodos considerando el flujo de procesos que se ha utilizado, se tendría que en el caso de los métodos de filtro la selección se realiza en la transformación o pretratamiento de los datos, paso previo al entrenamiento del algoritmo; en el caso de los métodos de envoltura esta selección es parte del entrenamiento y ajuste del algoritmo, se genera un bucle entre la generación del subconjunto y el entrenamiento hasta que se obtenga el grupo de predictores óptimos y se verifique el modelo. Por último, a diferencia de los de envoltura, los métodos intrínsecos realizan la validación del *performance* del subconjunto elegido dentro del mismo diseño del algoritmo.

5.2.1 Selección de atributos: Método Filtro

Como su nombre indica, en este método, se filtra y se toma solo el subconjunto de las funciones relevantes. El filtrado aquí se realiza utilizando la matriz de correlación y se realiza con mayor frecuencia mediante la correlación de Pearson**.**

Empleando este estadístico para seleccionar a las variables relevantes y estableciendo como límite mínimo un coeficiente de 0.4 entre las variables independientes con la variable de salida, se determina que solo las siguientes variables tienen una correlación significativa y deben usarse en el modelo.

```
PMIChina          0.709152
PMIUSA            0.473815
IndiceDolar       0.437114
Interest Rate     0.576025
Inventario        0.417694
```

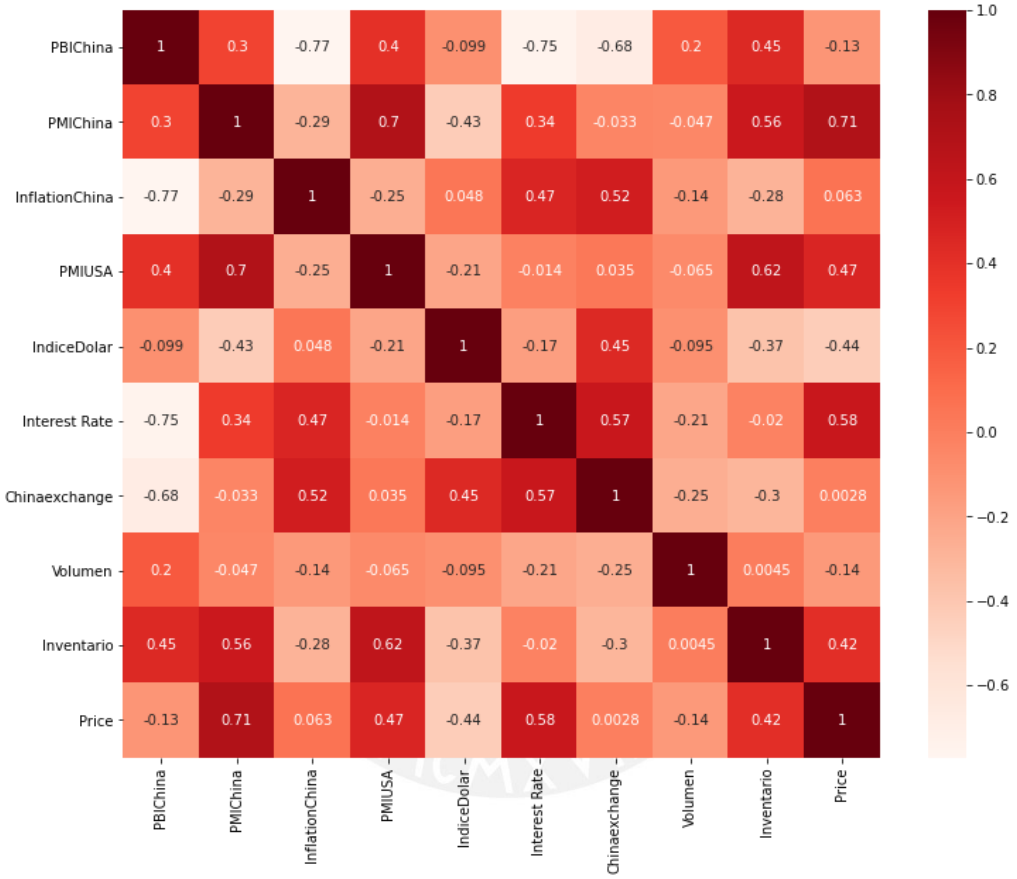

Figura 5.4: Matriz de correlación de Pearson

En base al resultado obtenido, se podría intuir entonces que las cinco variables referidas serán más relevantes en la predicción del precio del cobre. Sin embargo, como se detalló antes, este tipo de análisis excluye toda aquella información que proviene de la interrelación de las variables explicativas, y más aún al no utilizar un modelo para el ajuste de la selección el resultado es usualmente impreciso. Por ello, como siguiente método se explorará un método de envoltura denominado eliminación de características recursivas o RFE.

5.2.2 Eliminación de características recursivas

Este es un método de selección de características que se basa en eliminar las variables que presentan un peor desempeño en el modelo particular que se está generando. Como su nombre lo indica, al ser recursivo se realiza tal eliminación de forma reiterada hasta conseguir el resultado esperado, en otras palabras, es básicamente una selección hacia atrás *(backward selection)* de las variables.
Esta técnica comienza construyendo un modelo con todo el conjunto de predictores y calculando un puntaje de importancia para cada predictor. Luego se eliminan los predictores menos importantes, se reconstruye el modelo y se vuelven a calcular los puntajes de importancia. En el proceso, se debe especificar el número de subconjuntos de predictores a evaluar, así como el tamaño de cada subconjunto. Por lo tanto, el tamaño del subconjunto es un parámetro de ajuste para RFE. El tamaño del subconjunto que optimiza los criterios de rendimiento se utiliza para seleccionar los predictores en función de las clasificaciones de importancia. Finalmente, el subconjunto óptimo se usa para entrenar el modelo final.

Para poder utilizar RFE se debe proporcionar un algoritmo específico al cual se ajustan los datos, por ello los resultados obtenidos variarán en base al modelo usado como estimador. En primer lugar, se aplica RFE teniendo como modelo base la regresión lineal.

Regresión Lineal:

Se emplea la función RFECV obtenida de la librería sklearn, la cual tiene como parámetros, el estimador a utilizar, el número de características a eliminar en cada iteración denominado como step, por último, el puntaje obtenido en cada iteración se calcula como el MAE negativo, scikit-learn utiliza el valor negativo para que este se maximice, con lo cual a valores más negativos mejores resultados.

```
rfecv = RFECV (estimator = LinearRegression (), step=1,
scoring='neg_mean_squared_error')
```

El número óptimo de regresores para el modelo de regresión lineal es ocho, el siguiente procedimiento implica determinar cuáles son los regresores que se deben mantener, para ello se ejecuta la siguiente línea de código:

```
rfe = RFE (LinearRegression, 8)
X_rfe = rfe.fit_transform(X,y)
```

```
model.fit(X_rfe,y)
print(rfe.support_)
```

Obteniendo

```
[ True  True  True  True  True  True  True  True False]
```

Donde “False” indica que la variable número 9 debe eliminarse, es decir, el valor del Inventario no debe considerarse en la predicción del precio del cobre. Si se desea obtener un resultado más detallado del ranking de las 9 variables se ejecuta el código agregando

```
print(rfe.ranking_)
```

Obteniendo el siguiente ranking:

```
[5 6 2 4 7 1 3 8 9]
```

Para verificar el resultado se procede a construir el modelo utilizando las 8 variables, teniendo una nueva estructura del Data Set:

```
Train shape     : (842, 16)
Validation shape: (421, 16)
```

El puntaje logrado es

```
Scores:
Train      = 0.8157
Validation = 0.7849
Mean Absolute Error: 280.2
Mean Squared Error: 336.67
```

Como parte del análisis, se incluirá en esta evaluación de atributos a las variables secundarias generadas en el subcapítulo 4.2.4, estas son las siguientes: Variación Índice del Dólar, Variación China exchange rate, Variación Volumen transado. Al realizar el mismo procedimiento de selección de atributos ajustando al algoritmo de regresión lineal, se obtiene el siguiente resultado:

```
Scores:
```

```
Train      = 0.8092
Validation = 0.8200
Mean Absolute Error: 262.42
Mean Squared Error: 298.72
```

Asimismo, el número óptimo de variables en este caso es 9, excluyendo al inventario, volumen transado e índice del dólar.

Tabla 5.2: Ranking de variables considerando el conjunto de variables primarias y el conjunto total para Regresión Lineal

Ranking	LR con variables primarias	LR con conjunto total de variables
1	Tasa de interés FED	Tasa de interés FED
2	Tasa Inflación China	Tasa Inflación China
3	China exchange rate Yuan/Dólar	Variación China exchange rate Yuan/Dólar
4	PMI USA	Variación Índice Dólar
5	PBI China	China exchange rate Yuan/Dólar
6	PMI China	PMI USA
7	Índice Dólar	PBI China
8	Volumen transado	PMI China
9	Inventario	Variación Volumen
10	-	Índice Dólar
11	-	Volumen transado
12	-	Inventario

Incluyendo las tres variables secundarias generadas, se puede evidenciar que el ranking de relevancia cambia, de tal forma que la variación del índice del dólar y la variación de la tasa de intercambio entre el Yuan y el dólar desplazan a las variables primarias que otorgaban inicialmente tal información. Asimismo, la variación de volumen ocupa también una posición de relevancia menor que la variable primaria de Volumen transado. Por otro lado, comparando los resultados, es evidente que al utilizar las 9 variables óptimas del conjunto total se logra aumentar en 0.04 el puntaje en la validación.

Support Vector Regression:

Se utiliza la misma librería de scikit learn para aplicar el algoritmo RFE empleando en este caso como estimador al SVR.

```
rfecv = RFECV (estimator = SVR (kernel="linear", step=1, scoring='neg_mean_squared_error')
```

El número óptimo de regresores para el modelo de regresión lineal es seis, el siguiente procedimiento implica determinar cuáles son los regresores que se deben mantener, para ello se ejecuta la siguiente línea de código:

```
rfe = RFE(SVR, 6)
X_rfe = rfe.fit_transform(X,y)
model.fit(X_rfe,y)
print(rfe.support_)
```

Obteniendo

```
[ True  True  True  True  True  False  False  True False]
```

El puntaje logrado al utilizar 6 variables, (excluyendo al Inventario, la tasa de interés FED y la tasa de intercambio Yuan/Dólar) es:

```
Scores:
Train      = 0.9696
Validation = 0.9693
Mean Absolute Error: 0.048
Mean Squared Error: 1.878
```

Al incluir las variables secundarias generadas, se logra el siguiente puntaje:

```
Scores:
Train      = 0.9713
Validation = 0.9730
Mean Absolute Error: 0.043
Mean Squared Error: 0.98
```

En este caso, el número óptimo de variables es 9, excluyendo la tasa de cambio entre el Yuan y el Dólar, el inventario y la tasa de interés de la FED.

Tabla 5.3: Ranking de variables considerando el conjunto de variables primarias y el conjunto total para SVR

Ranking	SVR con variables primarias	SVR con conjunto total de variables
1	PMI USA	PMI USA
2	Índice Dólar	Variación Índice Dólar
3	PMI China	Índice Dólar
4	PBI China	PMI China
5	Tasa Inflación China	Variación China exchange rate Yuan/Dólar
6	Volumen transado	PBI China
7	China exchange rate Yuan/Dólar	Tasa Inflación China
8	Inventario	Variación Volumen
9	Tasa de interés FED	Volumen transado
10	-	China exchange rate Yuan/Dólar
11	-	Inventario
12	-	Tasa de interés FED

De la tabla 5.3, se aprecia que, al incorporar las variaciones en el conjunto de variables, el ranking cambia priorizando a las variables secundarias incorporadas y desplazando a las primarias. Asimismo, se puede observar que el algoritmo elige tanto a la variación del índice del dólar como al índice del dólar en valor nominal, lo mismo ocurre con el volumen. En este caso, se excluyen en ambos análisis a las 3 mismas variables. Por último, el puntaje al considerar las 9 variables óptimas incrementa ligeramente (0.01), al igual que el MAE y el MSE.

5.2.3 Observaciones generales de los atributos seleccionados

Como síntesis de los resultados obtenidos, se pueden ir detectando ciertas conclusiones. En primer lugar, según el primer método aplicado, *"Filter Method"*, las cinco variables con mayor correlación

con la variable *target* son, PMI China, PMI USA, el Índice del Dólar, la tasa de interés FED y el inventario. Al contrastar este primer resultado con los obtenidos al usar RFE para la regresión lineal y SVR es evidente que, si bien la correlación es un buen punto de partida para tener un panorama de los atributos a considerar, el algoritmo de eliminación recursiva permite conseguir resultados más exactos y un modelo mucho más preciso para el pronóstico. Además, cómo se evidencia en la tabla 5.4, el RFE al ajustar los datos a un algoritmo regresor específico, selecciona variables que otorguen más información a ese algoritmo en particular, por ello el ranking de atributos para la regresión lineal difiere al del SVR.

Tabla 5.4: Ranking de variables según cada método empleado

Ranking	Método Filtro	RFE Regresión Lineal	RFE SVR
1	PMI China	Tasa de interés FED	PMI USA
2	Tasa interés FED	Tasa Inflación China	Variación Índice Dólar
3	PMI USA	Variación China exchange rate Yuan/Dólar	Índice Dólar
4	Índice Dólar	Variación Índice Dólar	PMI China
5	Inventario	China exchange rate Yuan/Dólar	Variación China exchange rate Yuan/Dólar

Por otro lado, del análisis ejecutado con las variables secundarias para ambos algoritmos se puede apreciar que las variaciones incluidas como variables ocupan en ambos casos una relevancia mayor en el ranking, por lo tanto, se concluye que se obtienen mejores resultados al incorporarlas en el modelo. Por último, como cierre de este capítulo, es relevante indicar que incorporar un análisis de selección de atributos es de suma importancia al construir un modelo. Si bien, en este caso el universo de variables inicial no es considerablemente extenso, de igual forma se pudo aumentar el puntaje obtenido eliminando ciertos factores que no valía la pena incluir; de este modo se simplifica el algoritmo aligerando la carga de datos consiguiendo resultados de forma más rápida y precisa. Otro beneficio de este análisis es la interpretación y comprensión de los resultados, que sirven como base para próximas investigaciones.

Con el objetivo de seguir optimizando el algoritmo construido en el capítulo 4, a continuación, se explorarán métodos de selección de hiper parámetros.

6 Capítulo VI: Optimización de Hiper Parámetros

La optimización de hiper parámetros forma parte de la etapa de construcción del algoritmo, este proceso tiene como objetivo encontrar el valor para cada hiper parámetro que otorgue el mejor rendimiento medido en el conjunto de validación. Para ello se debe definir una función de pérdida que sea minimizada en el proceso y que finalmente logre aumentar la precisión de los datos.

Hasta el momento, se han elegido aleatoriamente los hiper parámetros para el SVR, en este capítulo se presentará un método que efectúe esta selección de manera automática.
Las estrategias más comunes para esta optimización son: en primer lugar, el ajuste manual realizado en base a prueba y error, sin duda no es eficiente cuando se cuenta con algoritmos complejos y que requieren de muchas iteraciones. Por otro lado, otro método muy utilizado es el *Grid Search*, el cual consiste en construir modelos para cada combinación posible de todos los valores de hiper parámetros proporcionados, evaluando cada modelo y seleccionando el que brinde mejores resultados. Finalmente, la búsqueda aleatoria, es una estrategia de optimización que se centra en seleccionar combinaciones al azar de un conjunto de valores de hiper parámetros para entrenar al modelo. Se detallarán estas dos últimas estrategias usando como ejemplo el modelo SVR construido en el subcapítulo 4.7

Antes de evaluar los algoritmos descritos, se define el proceso de evaluación de los parámetros elegidos en cada iteración. Las tres funciones de optimización a presentar en este capítulo tienen como parte de su estructura un proceso de validación llamado validación cruzada. La validación cruzada es un procedimiento de remuestreo utilizado para evaluar modelos de aprendizaje automático en una muestra de datos limitada.
El procedimiento tiene un único parámetro llamado k que se refiere al número de grupos en los que se dividirá una muestra de datos determinada. La muestra de datos determinada viene dada por el *dataset* de entrenamiento obtenido en el proceso de división del conjunto de datos, este conjunto de datos es dividido en los k grupos. En cada iteración se utilizan k-1 grupos para correr el algoritmo, y el grupo restante para evaluarlo. De esta forma, se estima la capacidad o habilidad de un modelo en datos desconocidos, optimizando en cada iteración el ajuste de los hiper parámetros en base a los resultados obtenidos.

La principal ventaja de utilizar validación cruzada es que disminuye sustancialmente la probabilidad de que ocurra un sobre ajuste a los datos de entrenamiento con los que se entrena el modelo y la optimización de hiper parámetros.

En la figura 6.1 se esquematiza el procedimiento de la validación cruzada, la división del set de entrenamiento se realiza en 5 grupos. Los grupos o "pliegues" de color verde son aquellos utilizados para entrenar el modelo y encontrar los parámetros, los pliegues azules son los destinados a la validación.

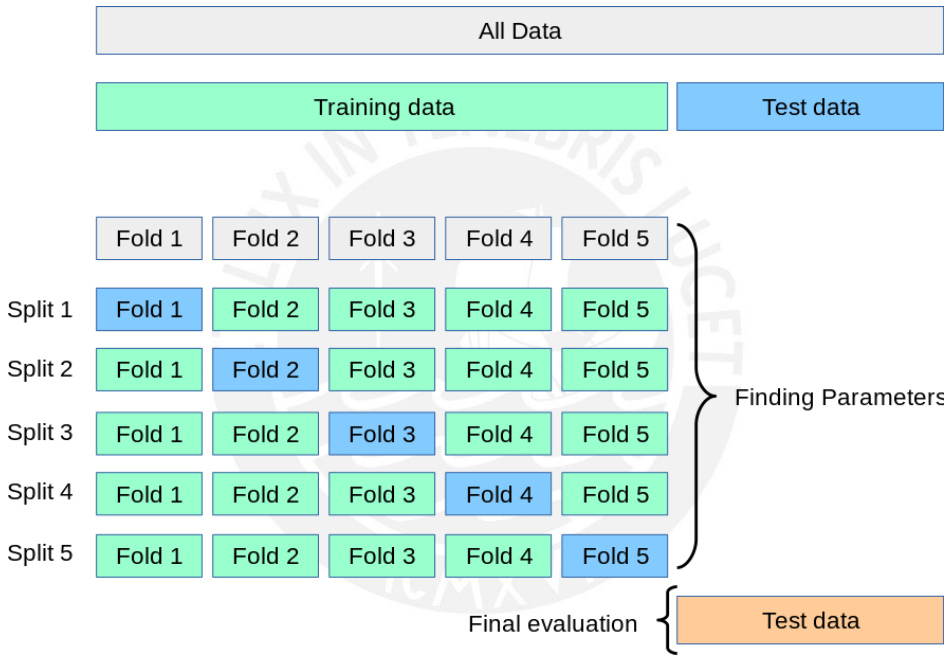

Figura 6.1: Esquema de Validación cruzada
Elaboración: (Sicki-learn.org 2019)

6.1 Grid Search

La función que realiza este tipo de búsqueda se llama GridsearchCV de la biblioteca de scikit-learn, esta genera candidatos a partir de una cuadrícula de valores especificados.

```
# Parameters for tuning
```

```
parameters = [{'kernel': ['rbf'], 'gamma': [1e-3, 1e-4], 'C': [1, 10,
100, 1000]},{'kernel': ['linear'],'gamma': [1e-3, 1e-4], 'C': [1, 10,
100, 1000]}].
```

Como se puede observar, los valores proporcionados para cada hiper parámetros una selección de un conjunto finito de valores, en el caso de Kernel se utilizará el *Radial basis function (rbf),* y el Kernel lineal, asimismo los valores de gamma podrán ser 0.001 o 0.001, y para C 1,10, 100 o 1000.
El *Grid Search* entrena un SVR con cada combinación de los valores otorgado, y evalúa el rendimiento en el conjunto de validación proporcionado.

Es importante mencionar, que esta función utiliza un tipo de validación cruzada en cada iteración, el número de grupos en los que se divide la muestra de datos, definido como k se asume igual a 5. Además, el valor de epsilon se toma como 0.001.

```
K = 5
svr = GridSearchCV(SVR(epsilon = 0.01), parameters, cv = K)
svr.fit(X_train, y_train)
```

El resultado al correr este código es el siguiente:

```
Tuning hyper-parameters
Grid scores on test set:
0.003 (+/-0.023) for {'C': 1, 'gamma': 0.001, 'kernel': 'rbf'}
0.013 (+/-0.023) for {'C': 1, 'gamma': 0.0001, 'kernel': 'rbf'}
0.138 (+/-0.025) for {'C': 10, 'gamma': 0.001, 'kernel': 'rbf'}
0.003 (+/-0.023) for {'C': 10, 'gamma': 0.0001, 'kernel': 'rbf'}
0.606 (+/-0.084) for {'C': 100, 'gamma': 0.001, 'kernel': 'rbf'}
0.144 (+/-0.025) for {'C': 100, 'gamma': 0.0001, 'kernel': 'rbf'}
0.831 (+/-0.054) for {'C': 1000, 'gamma': 0.001, 'kernel': 'rbf'}
0.594 (+/-0.090) for {'C': 1000, 'gamma': 0.0001, 'kernel': 'rbf'}
0.727 (+/-0.079) for {'C': 1, 'gamma': 0.001, 'kernel': 'linear'}
0.727 (+/-0.079) for {'C': 1, 'gamma': 0.0001, 'kernel': 'linear'}
0.791 (+/-0.058) for {'C': 10, 'gamma': 0.001, 'kernel': 'linear'}
0.791 (+/-0.058) for {'C': 10, 'gamma': 0.0001, 'kernel': 'linear'}
0.797 (+/-0.060) for {'C': 100, 'gamma': 0.001, 'kernel': 'linear'}
0.797 (+/-0.060) for {'C': 100, 'gamma': 0.0001, 'kernel': 'linear'}
0.793 (+/-0.064) for {'C': 1000, 'gamma': 0.001, 'kernel': 'linear'}
0.793 (+/-0.064) for {'C': 1000, 'gamma': 0.0001, 'kernel': 'linear'}
```

Por lo tanto, de los resultados obtenidos se concluye la selección óptima de hiper parámetros para el algoritmo SVR es un Kernel “rbf”, con gamma de 0.001 y C de 1000. Con esta combinación se obtiene un puntaje de 0.831.

Uno de los limitantes de este algoritmo está relacionado al problema de la dimensionalidad, cuando la evaluación de la cantidad de hiper parámetros crece, teniendo en cuenta que se debe proporcionar un conjunto de valores para cada uno, el número de iteraciones aumentará exponencialmente.

6.2 Randomized Search

Las posibilidades de encontrar el parámetro óptimo son comparativamente más altas en la búsqueda de tipo aleatoria debido al patrón de búsqueda donde el modelo puede terminar siendo entrenado en los parámetros optimizados sin ningún patrón o combinación predefinida. La función que realiza este tipo de búsqueda se llama *RandomizedSearchCV* de la biblioteca de *scikit-learn*.

A diferencia de *GridSearchCV*, no se prueban todos los valores de parámetros ingresados, sino que se muestrea un número fijo de configuraciones de parámetros de las distribuciones especificadas. El número de configuraciones de parámetros viene dado por n_iter.

```
# Parameters for tuning
parameters = [{'kernel': ['rbf'], 'gamma': expon(scale=.1),'C': expon(scale=100)},{'kernel': ['linear'],'gamma': [1e-3, 1e-4], 'C': [1, 10, 100, 1000]}, {'kernel': ['poly'],'gamma': expon(scale=.1), 'C': expon(scale=100)}]
```

Los valores que puede tomar gamma, y C quedan definidos por valores exponenciales de escala 0.1 y 100 respectivamente, a diferencia de la búsqueda *Grid*, el universo de valores es ahora infinito.

En cuanto a la validación, esta función utiliza de igual forma la validación cruzada para realizar la optimización en cada iteración.

```
K = 5
```

```
svr = RandomizedSearchCV(SVR(epsilon = 0.01), parameters, cv = K,
n_iter=10)
svr.fit(X_train, y_train)
```

Al especificar el número de iteraciones como n_iter = 10, se sabe de antemano que el algoritmo generará 10 combinaciones aleatorias, los resultados obtenidos son los siguientes

```
Tuning hyper-parameters
Random scores on test set:
0.765 (+/-0.040) for {'C': 11.60775957604531, 'gamma':
0.08662355642348757, 'kernel': 'poly'}
0.797 (+/-0.060) for {'C': 100, 'gamma': 0.0001, 'kernel': 'linear'}
0.811 (+/-0.028) for {'C': 33.66492987521718, 'gamma':
0.0781322265161743, 'kernel': 'poly'}
0.801 (+/-0.056) for {'C': 39.99708523808872, 'gamma':
0.008728795448371147, 'kernel': 'rbf'}
0.924 (+/-0.023) for {'C': 62.24600841059299, 'gamma':
0.03910948821302398, 'kernel': 'rbf'}
0.752 (+/-0.045) for {'C': 14.7757819801029, 'gamma':
0.04422347626333524, 'kernel': 'rbf'}
0.926 (+/-0.022) for {'C': 157.23474414917152, 'gamma':
0.013311855599634343, 'kernel': 'rbf'}
0.793 (+/-0.064) for {'C': 1000, 'gamma': 0.0001, 'kernel': 'linear'}
0.834 (+/-0.052) for {'C': 86.88074217143918, 'gamma':
0.15117648689720783, 'kernel': 'rbf'}
0.776 (+/-0.364) for {'C': 115.2264361599415, 'gamma':
0.14085556737923885, 'kernel': 'poly'}
```

Por lo tanto, de los resultados obtenidos se concluye la selección óptima de hiper parámetros para el algoritmo SVR es un kernel rbf, con gamma de `0.013` y C de 157.235, con esta combinación se obtiene un puntaje de 0.926. Comparando en general todos los resultados de las 10 iteraciones, el puntaje de la mayoría supera a las soluciones del *Grid Search*. La búsqueda aleatoria es un método útil para inicializar el proceso de búsqueda, ya que explora todo el espacio de configuración y, por lo tanto, a menudo encuentra combinaciones con un rendimiento razonable. Asimismo, es de gran valor pues no realiza ningún supuesto acerca del algoritmo que se está optimizando.

6.3 Optimización Bayesiana

Los dos algoritmos de optimización presentados si bien permiten obtener buenos resultados tienen una carencia común; Ambos son incapaces de tomar en cuenta los resultados obtenidos en iteraciones

pasadas para focalizar o restringir la búsqueda en regiones de mayor interés. Una buena alternativa frente a este inconveniente es la optimización bayesiana.

La optimización bayesiana es un algoritmo iterativo con dos compuestos clave, un modelo probabilístico y una función de adquisición para decidir qué punto se evaluará a continuación. En cada iteración, el modelo se ajusta a todas las observaciones de la función objetivo, función que es la métrica elegida para validar el modelo, de esta forma se consigue focalizar la búsqueda hacia los valores que devuelven un mejor puntaje, reduciendo el universo de combinaciones posibles a aquellas que conforman realmente las mejores candidatas. Lo cual conlleva a una mejora frente a las otras dos estrategias mencionadas. (Hutter F, Kotthoff L, Vanschoren J, 2019)

Se implementa la función *BayesSearchCV* de la biblioteca de *scikit-learn* de la siguiente forma:

```
parameters =  {
        'C': (1e-1, 1e+4, 'log-uniform'),
        'gamma': (1e-3, 1e-2, 'log-uniform'),
        'kernel': ['linear', 'poly', 'rbf']}
K = 5
svr = BayesSearchCV(SVR(epsilon = 0.01), parameters, cv = K, n_iter=10)
svr.fit(X_train, y_train)
```

De igual forma, al especificar el número de iteraciones como n_iter = 10, se sabe de antemano que el algoritmo generará 10 combinaciones aleatorias, los resultados obtenidos son los siguientes:

```
Bayesian Scores on test set:

0.622 (+/-0.074) for OrderedDict([('C', 2056.6872941698366), ('gamma',
0.008052280677715075), ('kernel', 'poly')])
-0.015 (+/-0.023) for OrderedDict([('C', 3.0519643336396807), ('gamma',
0.0019929418913664848), ('kernel', 'poly')])
0.797 (+/-0.059) for OrderedDict([('C', 78.21363020256017), ('gamma',
0.005277641896255444), ('kernel', 'linear')])
0.671 (+/-0.088) for OrderedDict([('C', 0.4440282434941538), ('gamma',
0.0011749410770069597), ('kernel', 'linear')])
0.777 (+/-0.061) for OrderedDict([('C', 2.9984467335922838), ('gamma',
0.001891805288551128), ('kernel', 'linear')])
0.797 (+/-0.056) for OrderedDict([('C', 39.592434972026176), ('gamma',
0.00481341899963295), ('kernel', 'linear')])
0.790 (+/-0.065) for OrderedDict([('C', 10000.0), ('gamma',
0.0011743501509672372), ('kernel', 'linear')])
0.793 (+/-0.064) for OrderedDict([('C', 528.7528745902974), ('gamma',
0.0024488916774940846), ('kernel', 'linear')])
```

```
0.793 (+/-0.064) for OrderedDict([('C', 778.1421417105589), ('gamma',
0.0032588867322390185), ('kernel', 'linear')])
0.965 (+/-0.009) for OrderedDict([('C', 10000.0), ('gamma', 0.01),
('kernel', 'rbf')])
```

De los resultados obtenidos se obtiene como puntaje máximo 0.97 para el set de validación, con la siguiente combinación: C igual a `10000`, Gamma de `0.01`, y Kernel del tipo `'rbf'`.

Se puede concluir que la búsqueda de cuadrícula y la búsqueda aleatoria son herramientas prácticas, sin embargo requieren tiempos de ejecución largos porque pierde tiempo evaluando áreas poco prometedoras del espacio de búsqueda, por ello es necesario contar con recursos suficientes de lo contrario, su implementación resulta ineficiente. Frente a esto la optimización Bayesiana logra mejores resultados, al realizar un seguimiento de resultados de evaluaciones anteriores. Finalmente, como resultado, se evidencia un incremento considerable en el puntaje obtenido al incluir en el flujo de procesos la optimización de hiper parámetros dentro de la etapa de construcción del algoritmo.

7 Capítulo VII: Conclusiones y recomendaciones

En este último capítulo se expondrán las conclusiones realizadas en base a los resultados presentados. Asimismo, se plantearán recomendaciones para futuros trabajos.

7.1 Conclusiones

Si bien, el objetivo principal de esta investigación no era construir un algoritmo complejo de predicción, sino más bien desarrollar una metodología para el proceso de selección, construcción y evaluación de este, los resultados obtenidos para los dos modelos presentados son bastante prometedores.

Como primera conclusión, queda evidenciado que la herramienta estadística explorada (ARIMA) no presenta una capacidad predictiva estadísticamente significativa. En parte por la complejidad que se le atribuye a la predicción del precio de este commodity. Como ya se ha revisado, la predictibilidad de este tipo de productos está ligada a la capacidad de pronosticar otras variables macroeconómicas más complejas, y es esta la limitante de los modelos estadísticos tradicionales simples como el ARIMA. Frente a esto, las herramientas de *Machine Learning* son capaces de generar modelos multivariables y con estructuras internas preparadas para no solo correlacionar linealmente las variables sino para encontrar patrones poco evidentes.

En cuanto a la metodología empleada, como segunda conclusión, del capítulo 4 se puede extraer un esquema base para el flujo de procesos que se debe seguir en la construcción de un modelo. Si bien se ha demostrado, que existen variantes para cada algoritmo, así como para cada objetivo específico de investigación, es crucial realizar en primer lugar una correcta categorización del problema, ya que la selección del tipo de herramienta de aprendizaje y variables dependerá directamente de este proceso. En la categorización del problema se debe entender el objetivo de la construcción del modelo, el tipo de enfoque a emplear y el tipo de datos con los que se cuenta, de esta forma se podrá definir el tipo de aprendizaje requerido. Sin embargo, como se mencionó al inicio de la tesis no es posible identificar de manera a priori el algoritmo más adecuado para modelar un problema. La selección de los datos y análisis de estos pueden ayudar a suponer ciertas tendencias como linealidad, por ejemplo, y bajo estas suposiciones se podría pensar que un modelo lineal sería el más adecuado. No obstante, como demostró David Wolpert en un famoso artículo de 1996 si no se realiza ninguna suposición sobre los datos entonces no habría razón para preferir un

modelo sobre cualquier otro. Este teorema llamado *No Free Lunch*, expone que no hay ningún modelo que esté garantizado a priori para funcionar mejor. Para ciertos conjuntos de datos, el mejor modelo puede ser una regresión lineal, mientras que si se cambia el conjunto de variables el mejor podría ser una red neuronal. La única forma de saber con certeza qué modelo es mejor es evaluarlos todos. Dado que esto no es posible, en la práctica se hacen algunas suposiciones razonables sobre los datos y se evalúa solo aquellos modelos que parecen ser más acordes. Por ello, la naturaleza de esta investigación, que al ser exploratoria permitió comparar dos algoritmos y explorar la optimización de estos mismos.

Siguiendo lo expuesto en el párrafo anterior, los resultados del capítulo 5 corroboran que la selección del conjunto de variables es distinta para cada algoritmo empleado. Si se observan la columna 3 y 4 de la tabla 10, queda demostrado que hay cierta diferencia en el conjunto óptimo de variables para cada algoritmo. Más aún, si se comparan estas dos columnas con la primera (método filtro), se confirma lo explicado por David Wolpert, y es que a pesar de realizar suposiciones acerca de la data, como en este caso la correlación de Pearson con la variable *target*, finalmente cada algoritmo estructura y emplea de forma distinta las observaciones.
De este capítulo entonces, se resalta el valor que agrega realizar un análisis y selección de atributos como parte del flujo de procesos, pues tanto para la regresión lineal como para el SVR se lograron mejores resultados. Por otro lado, se comprueba que la optimización de hiperparámetros realizada en el capítulo 6 es también un proceso clave pues el puntaje del SVR se incrementa en un 20% de 0.807 a 0.97.

Por último, acerca de la precisión de ambos métodos como predictores del precio del cobre, se constata que el SVR tiene mejores resultados. En un inicio ambos obtuvieron puntajes similares, (0.78 para la regresión lineal y 0.80 para SVR) y a medida que se seleccionaron de manera óptima el conjunto de variables para ambos casos los puntajes llegaron a ser 0.82 y 0.97 respectivamente. Asimismo, el MSE y MAE para el SVR fueron significativamente menores que los de la regresión lineal. Por lo tanto, se concluye que para el modelo de predicción del precio del cobre con el universo de datos presentado el algoritmo Support Vector Regression es más adecuado.

7.2 Recomendaciones

Por la naturaleza y objetivo de este trabajo, una de las limitaciones para la construcción del modelo fue el conjunto de variables seleccionadas para realizar el pronóstico. Si bien es verdad, que el resultado

final fue bastante prometedor, sería ideal en caso se desee ejecutar un modelo más complejo emplear más variables y con mayor rango histórico de datos.

Por otro lado, como se observó en el capítulo 5, las variaciones resultaron ser más importantes en el modelo que las variables primarias en sí. Por ello se sugiere profundizar la investigación en base al análisis de las variables seleccionadas y estudiar la implicancia que tiene emplear la derivada de los datos en el modelamiento. Al ahondar en el comportamiento de las variables más importantes según el ranking de la tabla 5.4, se pueden proponer mejoras en la toma de decisión de empresas y entidades comprometidas en el mercado del cobre.

Por último, la construcción de algoritmos que tengan en su estructura hiperparámetros permitirán que estos puedan optimizarse consiguiendo mejores resultados, como fue el caso del SVR. Por ello, se recomienda continuar la línea de esta investigación desarrollando modelos con tales características.

Bibliografía:

1. Libros:

CASTELO M.

2003 Diccionario Comentado de Términos Financieros Ingleses de Uso Frecuente Español. A Coruña: Netbiblo.

HUTTER Frank, KOTTHOFF Lars, VANSCHOREN Joaquin

2019 Automated Machine Learning. Suiza: Springer

GÉRON Aurélien

2019 Hands-on Machine Learning with Scikit-Learn, Keras and TensorFlow. Segunda Edición. Estados Unidos: O´reilly

KUHN Max, JOHNSON Kjell

2019 Feature Engineering and Selection: A practical Approach for Predictive Models. Bookdown.org.

2. Página Web:

AMAT Joaquín

2017 Máquinas de Vector Soporte. Recuperado de https://rpubs.com/Joaquin_AR/267926

BLOOMBERG

2019 Base de datos en Bloomberg.

INSTITUTO DE INGENIEROS DE MINAS DEL PERÚ

2020 Producción de cobre en Perú creció 97% en periodo 2008-2019. Recuperado de http://www.iimp.org.pe/actualidad/produccion-de-cobre-en-peru-crecio-97-en-periodo-2008-2019

KOHERSEN William
2017 Understanding Random Forest. Towards Data Science. Recuperado de https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

OPTIPEDIA
2019 Multi-Armed Bandit. Optimizely. Recuperado de https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

SINGH Nagesh
2019 Hyperparameter Optimization for Machine Learning Models. Recuperado de https://medium.com/swlh/hyperparameter-optimization-for-machine-learning-models-12582f00ae52

SICKIT-LEARN
2019 Sickit-learn machine learning in Python. Recuerado de https://scikit-learn.org/stable/

TRAFIGURA
2018 Commodities Demystified. A guide to trading and the global supply chain. Recuperado de https://www.bauer.uh.edu/spirrong/commoditiesdemystified-guide-en.pdf

YIU Tony
2019 Understanding Random Forest: How the Algorithm Works and Why it Is So Effective. Towards Data Science. Recuperado de https://towardsdatascience.com/understanding-random-forest-58381e0602d2

YIU Tony
2019 Understanding Neural Networks. Towards Data Science. Recuperado de https://towardsdatascience.com/understanding-neural-networks-19020b758230

3. Informes y estudios:

BASTOURRE D, CARRERA J, IBARLUCIA J.

2010 Precios de los commodities: Factores estructurales, mercados financieros y dinámica no lineal. Argentina: Banco Central de la República Argentina.

FINANCIAL STABILITY BOARD

2017 Artificial intelligence and machine learning in financial services. Market developments and financial stability implications. Recuperado de https://www.fsb.org/wp-content/uploads/P011117.pdf

INTERNATIONAL ORG OF SECURITIES COMMISSIONS

2017 IOSCO Research report on Fintech.*OICU-IOSCO*.Recuperado de https://www.iosco.org/library/pubdocs/pdf/IOSCOPD554.pdf

NEW VANTAGE PARTNERS

2019 How Big Data and AI are Accelerating Business Transformation. *Big Data and AI Executive Survey 2019*. Recuperado de https://newvantage.com/wp-content/uploads/2018/12/Big-Data-Executive-Survey-2019-Findings-Updated-010219-1.pdf

4. Tesis, tesinas y otros similares

AFSHARIRAD Elham, SINAEI Hasanali, ENAYATOLAH Seyed

2015 Predict the trend of stock prices using machine learning techinques. International Academic Journal of Economics. Vol2, No 12. pp1-11.

ALOUI Chaker, HAMDI Manel

2014 Forecasting Crude Oil Price Using artificial Neural Networks: A literature survey. Economics Bulletin. 35, 1339-1359.

BARSKY Robert, LUZT Kilian

2002 Do We Really Know That Oil Caused the Great Stagflation? A Monetary Alternative. NBER Macroeconomics Annual, B. Bernanke and K. Rogoff (eds), MIT Press, Cambridge, 137-183.

BARSKY Robert, LUZT Kilian

2002 Oil and the Macroeconomy Since the 1970s, Journal of Economic Perspectives, Vol 18(4), 115-134.

BORENSZTEIN Eduardo, REINHART C.M

1994 The Macroeconomic Determinants of Commodity Prices. IMF Sta§ Papers, Vol. 41 No. 2, 236-258.

BURTON Malkiel

2003 The efficient market hypothesis and its critics. The Journal of Economic Perspectives, Vol. 17. No. 1. 59–82.

CASIN P, MC DERMOTT C.J, SCOTT A

1999 The Myth of Co-Moving Commodity Prices. Bank of New Zealand Discussion Paper. No. G99/9.

CUSATIS Patric, PENDHARKAR Parag

2018 Trading financial indices with reinforcement learning agents. Expert Systems with Applications. Paper No. 103 1-13.

BASAK Suryoday, KAR Saibal, SAHA Sneahanshu, KHAIDEM Luckyson

2019 Predicting the direction of stock market prices using tree-based classifiers. North American Journal of Economics and Finance, Vol 47, pp. 552-567.

DOPORTO Miguez, MICHELENA Gabriel

2011 La volatilidad de los precios de los commodities: El caso de los productos agrícolas. Comercio Exterior e Integración, Vol 19, pp 35-53.

ELISSEFF André, GUYON Isabelle

2004 An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, Vol3. pp 1-26.

FAROOQ Akram
2008 Commodity prices, interest rates and the dollar. Norges Bank. Oslo

FRANKEL J.,ROSE A
2009 Determinants of Agricultural and Mineral Commodity Prices. Working Paper, Kennedy School of Government, Harvard University.

FRANKEL J.,ROSE A
2014 Effects of speculation and Interest Reates in a "Carry Trade" Model of Commodity prices. Journal of international Money and Finance, Cambridge.

FRANKLIN M, FISHER Paul Cootner, NEIL Martin
1972 An econometril model of the world copper industry. The Bell Journal of Economics and Management Science, Vol. 2, pp 568–609.

KABWE Eugine, YIMMING Wang
2015 Analysis of copper´s Market and price-focus on the last decade´s change and its future trend. International Journal of scientific & technology research, Vol. 4, pp 1-7.

KRIECHBAUMER T, ANGUS A, PARSONS D, CASADO MR.
2014 An improved wavelet-ARIMA approach for forecasting metal prices. Resource Policy, Vol. 39, pp 2–41.

HAMILTON J.D
2008 Understanding crude oil prices. The Energy Journal, International Association for Energy Economics, Vol. 30(2), 179-206.

MALAGRINO Luciana, ROMAN Norton, MONTEIRO Ana
2018 Forecasting stock market index daily direction: A Bayesian Network approach. Expert Systems with Applications, Vol. 105, pp.11-12.

MUÑOZ Ercio
2014 El efecto de sorpresas en el crecimiendo de China sobre el Precio del Cobre. Notas de investigación Journal Economía Chilena, Central Bank of Chile. 17, pp 110-123.

BARCELLONA Francesco, PANELLA Massimo, D´ECCLESIA Laura
2012 Forecasting Energy Commodity Prices Using Neural Networks. Advances in Decision Sciences, pp. 1-26.

PENDHARKAR P, CUSATIS P
2018 Trading financial indices with reinforcement learning agents. Expert Systems with Applications, Vol. 103, pp.1-13.

ROKACH Lior, MAIMON Oded
2007 Data Mining with Decision Trees. Tel-Aviv, Israel. World Scientific.

ROACHE Shaun
2010 What explains the rise in food price volatility?, IMF Working Paper WP/10/129.

SAVASCN Özge
2012 The Dynamics of Commodity Prices: A clustering Approach. North Carolina: University of North Carolina.

SHNEIDERMAN Ben
2008 Extreme visualization: Squeezing a billion data points into a million pixels. *ACM SIGMOD International Conference on Management of Data.* ACM, New York.Recuperado de http: https://www.cs.umd.edu/~ben/papers/Shneiderman2008Extreme.pdf

WAQAR Muhammad, DAWOOD Hassan, BILAL Muhammad, GHAZANFAR Ali
2017 Prediction of Stock Market by Principal Component Analysis. International Conference on Computational Intelligence and Security. Volume 1, pp 599-602.

WOLPERT David
1996 No Free Lunch Theorems for Optimization. Transaction on evolutionary compatation. Volume 1, pp 67-80.

YAGÜE Pablo
2014 Estudio de los commodities: El caso de los cereales. Madrid, España: ICADE.