UNIVERSIDAD MAYOR DE SAN ANDRÉS

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA PETROLERA

PROYECTO DE GRADO:

´´ PREDICCIÓN DEL PRECIO DEL GAS NATURAL USANDO PYTHON MEDIANTE MODELOS DE APRENDIZAJE AUTOMATICO´´

POSTULANTE: Univ. Pamela Meliza Álvarez Barco

TUTOR: MSc. Ing. Marco Antonio Montesinos Montesinos

La Paz – Bolivia

2022

UNIVERSIDAD MAYOR DE SAN ANDRÉS
FACULTAD DE INGENIERIA

LA FACULTAD DE INGENIERIA DE LA UNIVERSIDAD MAYOR DE SAN ANDRÉS AUTORIZA EL USO DE LA INFORMACIÓN CONTENIDA EN ESTE DOCUMENTO SI LOS PROPÓSITOS SON ESTRICTAMENTE ACADÉMICOS.

LICENCIA DE USO

El usuario está autorizado a:

a) Visualizar el documento mediante el uso de un ordenador o dispositivo móvil.
b) Copiar, almacenar o imprimir si ha de ser de uso exclusivamente personal y privado.
c) Copiar textualmente parte(s) de su contenido mencionando la fuente y/o haciendo la cita o referencia correspondiente en apego a las normas de redacción e investigación.

El usuario no puede publicar, distribuir o realizar emisión o exhibición alguna de este material, sin la autorización correspondiente.

Dedicatoria

Este proyecto de grado se lo dedico primero a Dios por haberme otorgado una familia maravillosa, quienes han creído en mí siempre, dándome ejemplo de superación, humildad y sacrificio, enseñándome a valorar todo lo que tengo.

Dedico de todo corazón mi tesis a mi madre María pues sin ella no lo habría logrado, tu bendición a diario a lo largo de mi vida me protege y me lleva por el camino del bien, sé que desde el cielo seguirás guiándome para lograr mis objetivos.

Mi padre Edwin por su amor y sacrificio en todos estos años de la carrera y brindándome siempre consejos importantes para mi vida, gracias a ustedes hemos logrado llegar hasta acá, gracias por todo su sacrificio.

A mis hermanas por estar presentes, acompañándome y por el apoyo moral, que me brindaron por todos estos años.

A mi hijita Alison que es mi inspiración para nunca rendirme y poder culminar mis estudios.

En general a todas las personas que me han apoyado que han hecho que el hoy pueda terminar una etapa muy importante de mi vida.

Agradecimiento

Con este Proyecto de Grado, se cierra una etapa muy importante de mi vida por lo cual he pasado mucho trabajo, pero ese trabajo sin duda sirvió para formarme como futuro Ingeniero.

A la Universidad Mayor de San Andrés sin duda la mejor universidad del país, que me acogió en todo en todos estos años.

A la Facultad de Ingeniería por inculcarme las bases de lo que es la Ingeniería y hace querer más mi querida futura profesión.

A la Carrera de Ingeniería Petrolera, al personal Docente en especial al Ing. Marco Antonio Montesinos Montesinos gracias por ser mi tutor y gracias por la enseñanza que me brindo, al personal administrativo a la Sra. Marcela gracias por siempre atender con amabilidad que la caracteriza y aclarar las dudas sobre las documentaciones a lo largo de estos años.

Sin olvidar a Dios, nuestra fuente de amor y sabiduría.

A todos, eternamente agradecido.

Pamela Melisa Álvarez Barco

Índice General

Índice de Figuras

Índice de Tablas

Índice de Ecuaciones

GLOSARIO DE TERMÍNOS

Accuracy: Es una métrica para evaluar los modelos de clasificación. Informalmente, la precisión es la fracción de predicciones que nuestro modelo acertó.

Alcanos: Los alcanos son hidrocarburos, es decir, compuestos que contienen solo átomos de carbono e hidrógeno.

Aprendizaje automático (machine learning): Es una ciencia de la inteligencia artificial que crea sistemas de aprendizaje automático. Aprender en este contexto significa identificar patrones complejos en millones de datos.

Aprendizaje no Supervisado: Esta es una forma en que el aprendizaje automático (ML) "aprende" los datos. El aprendizaje no supervisado contiene datos sin etiquetar que el algoritmo debe intentar comprender.

Aprendizaje supervisado: Es un derivado del aprendizaje automático, una técnica de análisis de datos que utiliza algoritmos que examinan los datos de manera iterativa para permitir que las computadoras encuentren información oculta sin tener que programar explícitamente dónde buscar.

Commodities: Es una sustancia física que se puede intercambiar, comprar o vender. A menudo se utilizan como materiales de partida para fabricar otros productos más complejos.

DataFrame: Es una estructura de datos bidimensional en la que datos de varios tipos (como caracteres, números enteros, valores de punto flotante, coeficientes, etc.) Es como una hoja de cálculo, tabla o datos SQL, cada vez que se lo use se ara referencia con la abreviatura de ‘df’.

Deep learning: Este es un tipo de aprendizaje automático que enseña a las computadoras a realizar las mismas tareas que los humanos.

Deficit: Es la escasez es una situación que se produce cuando falta algo esencial. En finanzas, se entiende por déficit un superávit de los gastos sobre los ingresos (falta de dinero).

Google colaboratory: Es una extensión de Google que permite a cualquier usuario escribir y ejecutar cualquier código de Python en el navegador. Es especialmente adecuado para aprendizaje automático, análisis de datos y tareas educativas.

Inteligencia Artificial: Es un área científica de la informática que se enfoca en crear programas y mecanismos que puedan exhibir un comportamiento que se perciba como inteligente.

Kaggle: Es una plataforma web que reúne la comunidad Data Science más grande del mundo, con más de 536 mil miembros activos en 194 países, recibe más de 150 mil publicaciones por mes, que brindan todas las herramientas y recursos más importantes para progresar al máximo en data science.

Matplotlib: Es una biblioteca para crear gráficos a partir de datos contenidos en listas o matrices utilizando el lenguaje de programación Python y su extensión matemática NumPy.

MMBTU: Millones de unidades térmicas británicas.

NaN: Significa Not a Number que representa los valores faltantes en Pandas.

Overfitting: Es el efecto de sobreentrenar un algoritmo de aprendizaje con unos ciertos datos para los que se conoce el resultado deseado.

Pandas: Es una biblioteca de código abierto que es muy popular entre los desarrolladores de Python, especialmente en el campo de la ciencia de datos y el aprendizaje automático, ya que

proporciona marcos muy potentes y flexibles que facilitan la manipulación y el procesamiento de datos.

Python: Es un lenguaje de programación de alto nivel que se utiliza para desarrollar aplicaciones de todo tipo.

Train: Es la parte de un modelo preciso que responda las preguntas correctamente la mayor parte del tiempo. La máquina necesita ser "entrenada", alimentándola explícitamente con las respuestas correctas adjuntas.

Scikit-Learn: es una de estas librerías gratuitas para Python. Cuenta con algoritmos de clasificación, regresión, clustering y reducción de dimensionalidad.

Underfitting: Este fenómeno mostrará una incapacidad para determinar u obtener resultados precisos debido a muestras de entrenamiento insuficientes o a un entrenamiento muy deficiente.

RESUMEN

Hoy es común escuchar que los datos son representan el nuevo petróleo, ya que gracias a los datos recolectados en un pasado es posible predecir, clasificar, ordenar estos datos.

El precio del Gas Natural es medido por u/MMBTU (dólares por MILLON de BTU), las exportaciones de Gas Natural desde el 2006 hasta el 2020 representaron el 43 % del total de exportaciones del país.

Esto rebela que la economía boliviana es bastante dependiente del precio del Gas Natural, de aquí la importancia de ser capaz de predecir el precio del Gas Natural, ya que los contratos que tenemos con la Argentina y Brasil no son precios estáticos.

Estos precios de exportación son dinámicos, en el presente trabajo no trataremos esos datos, solo trabajaremos con los precios internacionales del Gas Natural.

Con lo anterior descrito nace la necesidad de ser capaz de predecir el precio internacional del Gas Natural mediante Aprendizaje Automático. Los modelos que se usaron fueron: Regresión Lineal, Árbol de decisión y bosque aleatorio.

Se planteo tener al menos una exactitud del 70%, ahora para cada modelo para el entrenamiento y la evaluación se trabajó con un rango de 70% para el entrenamiento y 30% para la evaluación, para no caer en problemas de sobre ajuste o sub ajuste.

El modelo más eficiente hablando en tiempos de meses sería el modelo de bosque aleatorio que obtuvo una exactitud de 95%.

1. CAPÍTULO I: GENERALIDADES

1.1.INTRODUCCIÓN

El gas natural es un hidrocarburo de mezcla de gases ligeros de origen natural de la familia del Alcanos donde principalmente componente es el metano.

Cada día se empieza a ver más cerca el principio del fin de la era del petróleo para ser sustituido por energías verdes, pero el Gas Natural será un actor clave en la transición por lo que varios medios científicos le estiman como mínimo 30 años más de materia energética al Gas Natural y en 10 años seria la materia energética principal en el mundo.

El Estado Plurinacional de Bolivia depende de gran magnitud de las exportaciones de Gas Natural, se tiene datos que desde el 2006 hasta el 2020 el Gas Natural represento el 43% del total de las exportaciones.

Estas divisas que entran al país le son de gran ayuda al Estado para mantener a las cuentas públicas.

El precio de venta del Gas Natural en un futuro podría estar en función al precio internacional del Gas Natural, cabe mencionar que el precio de venta de Gas Boliviano es mayor al del precio internacional.

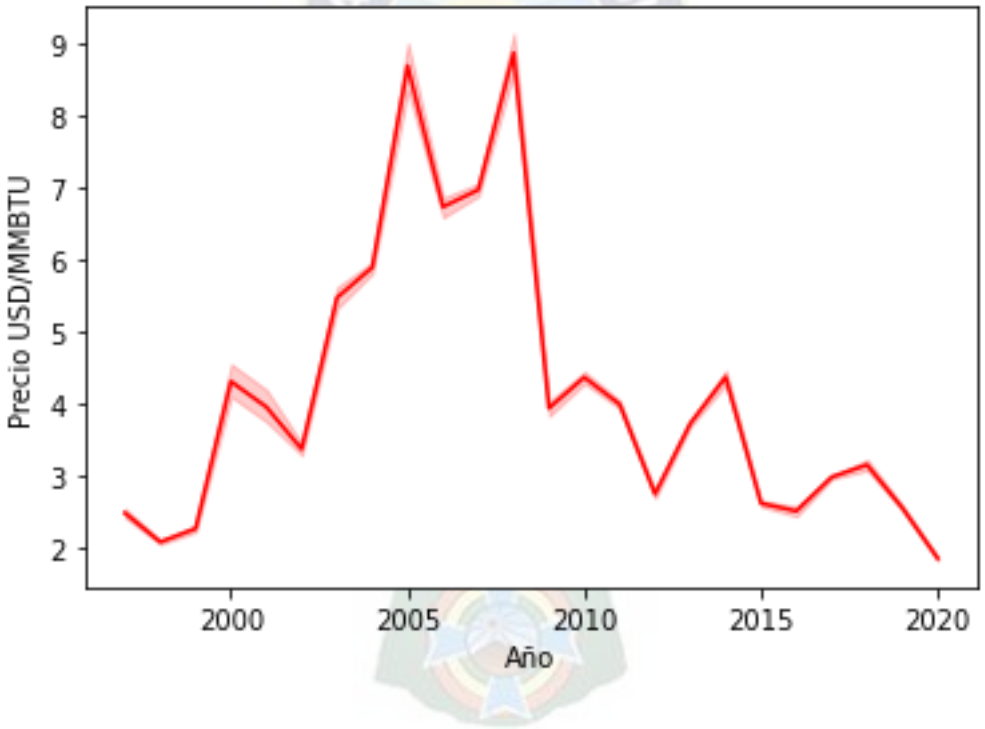
Se desea encontrar un modelo y ajustar para predecir el precio internacional del Gas Natural, para que luego esta pueda servir de referencia para futuros ingresos que tendrá en país por la venta de este Gas.

El impacto que causará la aplicación de los modelos que cumplirán con nuestras exigencias podría ser que el Estado sepa en que épocas el precio del Gas Natural es más bajo y en otros meses y/o años es más alto. Y con esto administrar mejor los recursos por la venta del Gas Natural.

1.2.ANTECEDENTES

Los datos con los que se cuenta son precios diarios de precio de gas natural en EUROPA del ESTE desde el año de 1997 hasta el 2020 con el cual fue extraída en la página de Kaggle en el siguiente link: https://www.kaggle.com/datasets/tunguz/natural-gas-prices. Donde se puede ver en la siguiente figura su evolución del precio. La figura 1.1 representa como vario el precio del gas natural por año.

Figura 1.1: Evolución del precio del Gas Vs el año.

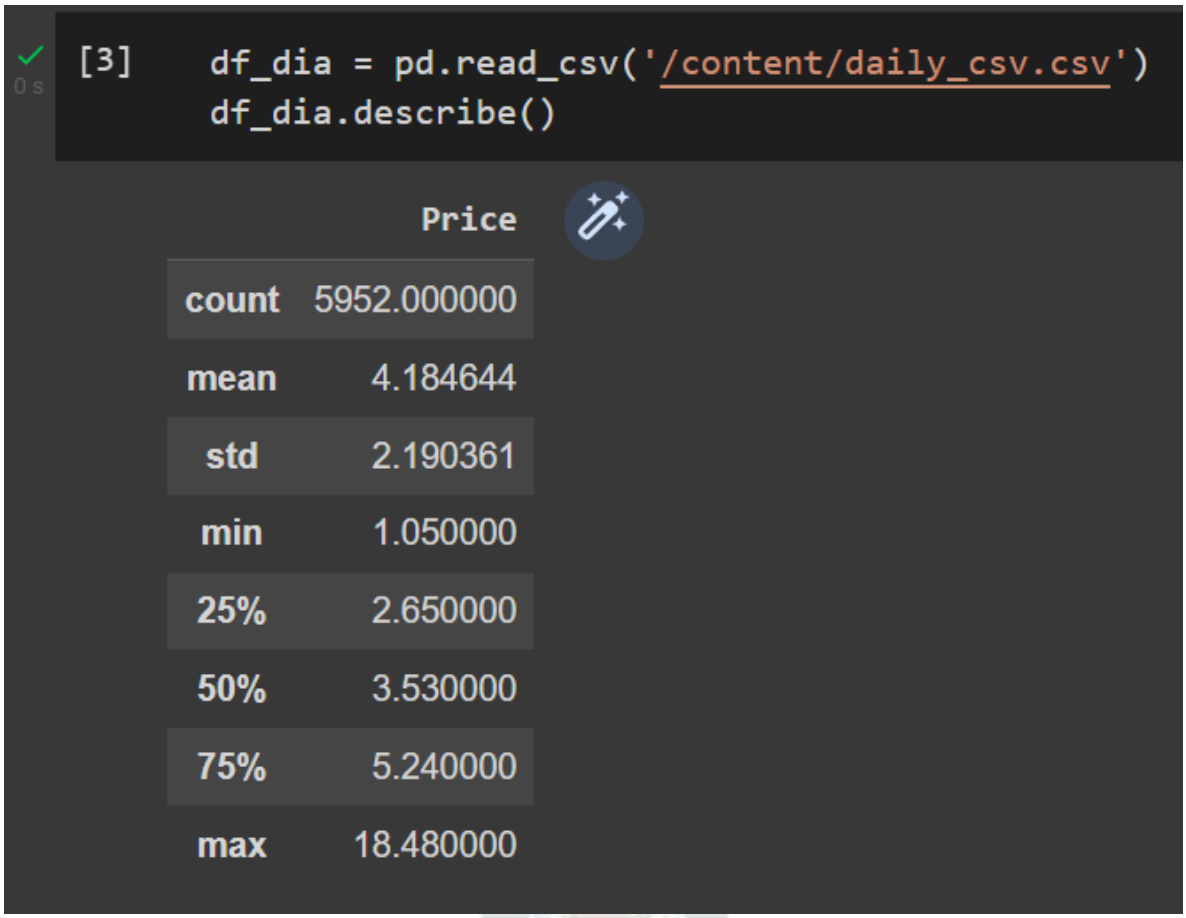


Fuente: Kaggle Natural Gas Prices, 2021.

Se puede observar que es bastante variante el precio a simple vista no es uniforme de año a año, con ayuda de pandas y Python se describirá en el siguiente grafico el máximo y mínimo y otros datos para el Gas Natural.

Gracias al lenguaje de programación python y su librería panda es posible tener toda la descripción de estos datos de manera rápida y de manera sencilla, en la figura 1.2 se presentará una estadística descriptiva del df_dia que representa los datos del precio del gas desde 1997 hasta el 2020.

Figura 1.2: Descripción del DataFrame(df_dia).



Fuente: Google Colaboratory con datos de kaggle, 2021.

En la figura 1.2 se presenta una descripción de los datos los cuales son en el mismo orden: Contar, promedio, desviación estándar, valor mínimo, quartil 25%, quartil 50% que es lo mismo que la mediana, quartil 75%, y el valor máximo.

La figura 1.2 represento los datos en formato de cada día es por ese motivo que se registró tal cantidad de datos de 5952 datos captados.

A continuación, la figura 1.3 representara a los datos del precio del gas natural por mes.

Figura 1.3: Descripción del DataFrame(df_mes)

Fuente: Google Colaboratory con datos de kaggle, 2021.

La figura 1.3 presenta una descripción de los datos los cuales son en el mismo orden: Contar, promedio, desviación estándar, valor mínimo, quartil 25%, quartil 50% que es lo mismo que la mediana, quartil 75%, y el valor máximo en referencia al precio del gas natural.

La información fue extraída del repositorio de Datos de Kaglee.

Otros intentos también fueron realizados para el precio del petróleo como por ejemplo (https://ichi.pro/es/predecir-los-precios-del-petroleo-con-aprendizaje-automatico-y-python-55256159639482). Donde utiliza el modelo de regresión lineal donde sus variables independientes son el promedio de 3 y 9 días anteriores.

Por lo tanto, partiremos de ese modelo, adicionalmente añadiremos el modelo de Arbol de decisión regresor y bosque aleatorio regresor cuyas variables independientes son el mes y el año, más adelante se explicará por qué se usó dichas variables. Como mínimo se desea una exactitud mínima de al menos un 70%.

1.3.PLANTEAMIENTO DEL PROBLEMA

1.3.1. Identificación del problema

Como se vio en la figura 1.2. los datos que se manejan son más de 5000 datos lo cual dificulta ser realizado con programas como Excel, es por ese motivo que está optando por el lenguaje de programación Python.

Como anteriormente se definió Bolivia depende de gran manera de la exportación del Gas Natural y como el precio de venta de este Gas USD/MMBTU en un futuro podría estar función al precio internacional del Gas Natural.

En un escenario donde el Estado no tenga en cuenta los precios predichos del Gas Natural Internacionalmente podría llevar podría llevar a tener déficit en las cuentas públicas y exportaciones. Lo cual traería consecuencias en la economía de la sociedad.

En caso contrario donde el Estado si tome en cuenta los precios predichos del Gas Natural podrá ser precavido para el gasto de las cuentas públicas que dependen de estos precios.

No solo el Estado estaría involucrado sino también instituciones y/o lugares que dependan de manera directa o indirecta del precio de venta del Gas Natural, como ser los Impuestos Directo a los Hidrocarburos (Universidades, Municipios, Gobernaciones, Departamentos).

El país en general depende de gran manera de un precio alto del Gas Natural, de no ser medido con moderación y son tomar los pronósticos del precio del Gas Natural, existiría un aumento de la deuda pública para cubrir cuando se avecinen bajos precios.

1.3.2. Formulación del problema

Debido a la dependencia del Estado por la venta del Gas, y como este precio de venta no es estático sino más bien dinámico, este precio de venta está en referencia al precio del Gas Natural internacional. Cabe recalcar que el presente proyecto de grado usara los datos internacionales ya la data de precios de venta en Bolivia es muy escasa.

Si se aplicaría nuestros modelos de predicción del precio del Gas Natural Internacional tanto el estado como instituciones que depende de manera directa de la venta de Gas Natural podría planificar sus cuentas para no caer en deudas públicas o falta de fondos para su financiamiento.

1.4.OBJETIVOS GENERALES Y ESPECÍFICOS

1.4.1. Objetivo general

- Predecir el precio del Gas Natural mediante el desarrollo de diferentes modelos de aprendizaje de maquina se debe desarrollar al menos un modelo de que intente predecir el precio internacional del Gas Natural.

1.4.2. Objetivos específicos

- Dar la limpieza necesaria a los datos originales.

- Desarrollar los modelos de regresión lineal, árbol de decisiones y bosque aleatorio.
- Proponer graficas para expresar los valores predichos y los valores reales.
- Obtener métricas que indiquen la exactitud de los modelos.
- Escoger el modelo que mejor se adapte a nuestro caso de estudio.

1.5.JUSTIFICACIÓN

Aporta en revelar nuevas herramientas y paradigmas de que tan importantes son los datos. Las nuevas tecnológicas como ser Ciencia de Datos y Aprendizaje Automático.

De ser aplicado el y/o los modelos para la predicción del precio del Gas Natural ayudaría al Estado a ser precavido ante futuros precios fluctuantes del Gas Natural Internacional.

1.5.1. Justificación técnica

Aplicando el y/o los modelos que cumplan nuestro objetivo de al menos 70% de exactitud en la predicción, ayudaría a tener unas cuentas públicas más claras para prevenir futuros precios fluctuantes del precio del Gas Natural Internacional.

1.5.2. Justificación social

No solo el Estado recibe los ingresos generados por la venta de Gas Natural, existen universidades, municipios, gobernaciones, programas sociales que reciben parte de la venta de este recurso el IDH (Impuesto Directo a los Hidrocarburos).

1.6.ALCANCE

1.6.1. Alcance temático

- Lo que se planea con el proyecto de grado es predecir el precio internacional del gas natural, mínimamente uno de nuestros modelos deberá ser capaz de lograr este objetivo, se planea que

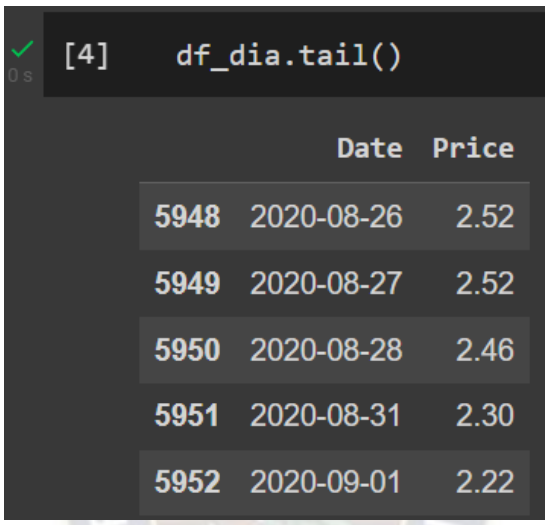
los 3 modelos sean capaces de realizar esta tarea. Cabe mencionar que los modelos tendrán un rango predecir como máximo hasta finales del 2020, esto debido a las desventajas que presentan los modelos de árbol de decisión regresor y bosque aleatorio regresor.

- En caso de que los datos presenten valores nulos se procederá con la limpieza de datos y se analizara ver si son convenientes borrarlos o reemplazarlos por alguna tendencia de los datos colineales a estos datos nulos.
- Desarrollar los modelos de Regresión Lineal, árbol de decisiones y bosque aleatorio, para esto serán necesario transformar los datos ya que se debe crear las variables ya que la base de datos solo cuenta con 2 columnas que son la fecha y el precio, probablemente se creen las variables a partir de la fecha.
- Se realizarán graficas para comparar valores reales con los valores predichos posiblemente con la parte reservada para el test del modelo.
- Se evaluarán los modelos de Regresión Lineal, árbol de decisiones y bosque aleatorio con la métrica del coeficiente de determinación en caso de que uno de los modelos tenga una exactitud menos al 70% será descartada.
- Se escogerá al modelo que presente mejores métricas y visualmente sea idéntica a los valores reales del precio del gas natural internacional, adicionalmente el que menor probabilidad tenga de caer en un sobre ajuste.

1.6.2. Alcance temporal

Los datos abarcan aproximadamente desde el año 1997 hasta el 2020. En la figura 1.4 se puede evidenciar cuales son los últimos 5 datos en orden cronológico.

Figura 1.4: Evolución del precio del Gas Vs Año últimas fechas



	Date	Price
5948	2020-08-26	2.52
5949	2020-08-27	2.52
5950	2020-08-28	2.46
5951	2020-08-31	2.30
5952	2020-09-01	2.22

Fuente: Google Colaboratory con datos de kaggle.

2. CAPÍTULO II: MARCO TEÓRICO

2.1.GAS NATURAL

El gas natural es una mezcla de gases, de los cuales el metano constituye una gran proporción. Se encontró que la proporción de este compuesto era del 75% al 95% del volumen total de la mezcla. Los componentes restantes son etano, propano, butano, nitrógeno, dióxido de carbono, sulfuro de hidrógeno, helio y argón. El desarrollo de la utilización del gas natural se lleva a cabo después de la utilización del petróleo. Casi todo el gas natural que se encuentra en los campos petroleros se quema como otro tipo de residuo. Aunque genera mucho calor, no se puede utilizar debido a los grandes problemas de almacenamiento y transporte.

La necesidad de encontrar nuevas fuentes de energía, el perfeccionamiento de la tecnología de licuefacción de gases y de los procedimientos de soldadura de tuberías para soportar altas presiones hace posible el aprovechamiento de todas estas fuentes de energía. Hoy, Europa es abastecida por una red de gasoductos que se extiende a lo largo de miles de kilómetros, y se están planificando nuevos gasoductos con Asia y Argelia.

2.2.GAS NATURAL BOLIVIANO

A partir de 1997 Bolivia emprendió reformas de gran envergadura transformando radicalmente el mapa gasífero de la región por sus potenciales de exportación a mercados externos y por el papel que deberá cumplir en la integración gasífera del Sur de América.

2.2.1. Desarrollo del 'UPSTREAM'

Las regiones con potencial hidrocarburífero se hallan divididas en ocho provincias geológicas que totalizan 450000 Km2. Las reservas probadas y probables de gas natural a enero de 2002 eran de

52,3 TCF (1.966,3 millones de m^3) mientras que en 1997 se estimaban en sólo 5,69 TCF (161,2 millones de metros cúbicos).

La exploración se rige por la Ley de Hidrocarburos N° 1689 del 30 de abril de 1996. La información es proporcionada por el Centro de Información de Hidrocarburos de Santa Cruz de la Sierra, perteneciente a la empresa estatal YPFB. YPFB debe suscribir contratos de riesgo compartido, exploración, desarrollo y comercialización en nombre del Estado. En la actualidad, existen 38 contratos de empresas conjuntas en el país y en el extranjero que participan en actividades exploratorias.

2.2.2. Consumo del gas natural en Bolivia.

Bolivia es conocido como un país exportador neto de gas natural, pero surgieron cambios entre el gas destinado al mercado interno y externo varió sustantivamente primero con las reformas de los años 90 y luego tras los cambios institucionales ocurridos entre 2005 y 2006. En la Tabla 2.1 se dará a conocer los porcentajes de que se destinan del gas natural para el mercado interno para el año 2009, el promedio de consumo del país fue 10.9 MMm3d en total para dicho año.

Tabla 2.1: Destino del Gas producido en el estado de Plurinacional de Bolivia 2020.

Destino	Consumo
Gas Natural Vehicular	19%
Comercial	2%
Domestico	2%
Industria	29%
Generadoras Eléctricas	48%

Fuente: Instituto Nacional de Estadísticas. Datos de 2009, preliminares

2.3.CÁLCULO DE PRECIOS DEL GAS NATURAL EN EUROPA Y ASIA

A mediados de la década de los cincuenta del siglo pasado se inicia un importante negocio del gas en Sobre todo en Estados Unidos y Europa. A su vez, se ha desarrollado una red que permite importar de Rusia, Argelia, Noruega y Holanda. Durante esos años,

Se realizó intercambio de gas natural licuado (GNL) en Asia, para brindar centrales térmicas de petróleo en Japón.

A fin de permitir estos intercambios que necesitaban grandes inversiones sobre todo en el transporte, fue necesario encontrar un mecanismo que garantice las transacciones tanto al vendedor como al comprador. Es así, que se implementaron los contratos a largo plazo.

Las principales características de estos contratos, fueron que tendrían una duración entre 20 y 25 años, con obligaciones mínimas de pago de la parte del comprador (cláusula take or pay) y el compromiso de entrega por parte del vendedor y, además, de un precio indexado a las energías concurrentes (Mecanismo aplicado en función del costo de oportunidad del sustituto).

El objetivo de tener un precio indexado a combustibles concurrentes como el fuel oil o directamente al petróleo, es el de simular un mercado de gas natural, a falta de la existencia de un mercado dedicado exclusivamente a este energético. Con este mecanismo, se asegura un precio más cercano al precio de los combustibles concurrentes y además, las inversiones ligadas al transporte, puedan ser amortizadas sin el riesgo de una disminución en los volúmenes entregados. Estas fórmulas y tipos de contratos, son vinculadas específicamente a los mercados europeos y asiáticos.

Los contratos en Europa, están indexados en porcentajes 50% con el fuel oil doméstico, 30% con el fuel oil pesado, 5% con el precio spot del gas y el 15% restante con los precios de la electricidad incluso con las tasas de inflación.

Las cláusulas de indexación que ligan el precio del gas con el del fuel pesado y con el fuel oil para uso doméstico, se deben a que durante el invierno en Europa se tiene un mayor consumo de fuel oil doméstico (heating oil) y en verano de fuel oil pesado, ya que este es el combustible que se usa para generar energía eléctrica (en verano para climatización).

En estos mercados (Europa y Asia), el principio de cálculo es definido sobre la base llamada "netback": es decir que los costos de transporte y de distribución son deducidos del precio medio de las energías concurrentes en el mercado final. El resultado corresponde al precio máximo de compra que el distribuidor gasífero está dispuesto a pagar al productor.

La indexación clásica permite mantener en el tiempo una relación entre el gas natural y los combustibles concurrentes, como se muestran en las siguientes fórmulas:

Ecuación 1 Determinación del Precio del Gas Natural en Europa

$$P = Po + A(G - Go) + B(F - Fo) \dots (1)$$

Ecuación 2 Determinación del Precio del Gas Natural en Asia

$$P = Po + A(B - Bo) \dots (2)$$

Donde:

Po= Precio mensual de compra del gas al productor.

o=fecha de inicio del contrato.

A/B=Coeficientes de equivalencia energética

G/F/B=Promedio del precio de 3,6,9 meses dependiendo del contrato

G=Fuel oíl domestico

F=Fuel oíl pesado

B=Petróleo, el promedio permite atenuar las alzas o las bajas del mercado petrolero.

El mercado europeo a día de hoy se encuentra en una situación dual, ya que los precios del gas en Reino Unido están sujetos a precios spot y gran parte de la producción de electricidad es en ciclos combinados y gas, y en el resto del continente el precio del continente es directamente afectados por los productos del petróleo y sus derivados. Sin embargo, estos dos mercados no son independientes ya que en realidad están vinculados al gasoducto que conecta Bélgica con el Reino Unido, que actúa como un equilibrio de precios.

La pregunta principal en el mercado europeo es si es probable que esta actividad dual continúe en el futuro. Los contratos a largo plazo representan actualmente el 90 % del suministro de gas en Europa continental, mientras que los precios al contado fuera del Reino Unido son limitados. La evolución del contrato al sistema spot en Europa dependerá de una mayor competencia y de la instalación de nueva infraestructura (gasoductos y terminales de GNL) en el mercado del Reino Unido.

En definitiva, los precios del gas en el mercado europeo estarán siempre ligados a los precios del petróleo, dada la importancia actual de los contratos a largo plazo, tradicionalmente clasificados sobre productos petrolíferos. Los indicadores de electricidad serán más frecuentes dada la importancia de este sector en la demanda de gas en Europa.

2.4.EL MERCADO DEL GAS Y CÁLCULO DE PRECIOS EN LATINOAMERICA

América Latina posee una amplia variedad de recursos naturales y entre ellos se destaca el gas natural. La dotación de este energético en la región, acompañada de una política de penetración del mismo en la matriz energética de los países de América Latina, ha permitido que en la actualidad el 26% de los recursos primarios utilizados en la región corresponda al gas natural.

Las reservas probadas de gas natural en Sudamérica ascienden a 7.528 miles de millones de m^3, representando el 4% de las reservas mundiales. Por su parte la producción de gas en la región se ubicó en el año 2016 en 178,8 miles de millones de m^3 (representando el 5% de la producción mundial). El cociente entre el volumen de reservas y la producción (R/P), para el conjunto de países sudamericanos, arroja un valor de 42,1 años (levemente por debajo de la media mundial, del orden de 52,4 años).

El consumo total de gas en la región alcanza los 170 miles de millones de m^3 (467,7 millones de m^3d), siendo Argentina, Brasil, Venezuela y Trinidad y Tobago, quienes explican el 83% de dicho volumen.

En cuanto al balance entre oferta y demanda de gas, se aprecia que la región presenta un saldo neto exportador, del orden de 7,5 miles de millones de m^3 (20,5 millones m^3d).

Si bien a nivel regional se observa un equilibrio entre la oferta y la demanda de gas natural y un elevado horizonte para la relación R/P, varios países de la región tienen faltantes de gas por lo que deben recurrir a importaciones. Tal es el caso de Argentina, Brasil, Chile, Uruguay y marginalmente Colombia y Venezuela. En total estos países importaron en 2016 unos 81,5 millones m^3d, dichas importaciones representan el 17,4% del consumo de gas de la región.

Por otra parte, países como Bolivia, Perú y Trinidad y Tabago son netos exportadores, presentando Ecuador una situación equilibrada. En total estos países exportaron en el año 2016 unos 102 millones m^3d.

Los seis países de Sudamérica que importan gas se abastecen utilizando gasoductos y en el caso de Argentina, Brasil, Chile y Colombia estos también poseen plantas de regasificación de gas natural licuado (GNL). Se observa en el siguiente cuadro que sobre el total de las importaciones registradas en el 2016 (ubicadas en el orden de los 81,5 millones de m^3d), el 57% correspondieron a importaciones de gas natural vía gasoducto y el 43% de gas natural como GNL. Países como Argentina y Brasil tienen diversificadas las fuentes de aprovisionamiento (gasoducto y GNL), mientras que otros tales como Chile y Colombia sólo lo hacen vía GNL y por su parte Venezuela y Uruguay solo vía gasoducto. SSEn la tabla 2.2 se expresará las importaciones de gas natural.

Tabla 2.2: Importaciones de gas natural por tipo de medio (Gasoducto o GNL).

Importaciones	**Gasoducto MMm3d**	**GNL MMm3d**	**Total, MMm3d**
Argentina	15,4	15,6	31,0
Brasil	27,7	8,2	35,9
Colombia	0	0,5	0,5
Venezuela	3,6	0,0	3,6
Uruguay	0,2	0,0	0,2
Chile	0,0	10,3	10,3
Total	46,8	34,6	81,5

Fuente: a BP Statistical Review of World Energy, 2017.

En cuanto a las fuentes de aprovisionamiento de gas importado, el proveedor regional más relevante de gas natural vía gasoducto es Bolivia (con 43,1 millones m^3d, representado el 53% de las importaciones de gas de la región), mientras que en el caso del GNL los proveedores son extra

regionales y regionales, destacándose entre estos últimos Trinidad y Tobago quien provee el 37% de las importaciones de GNL de la región (alrededor de 12,7 millones m^3d). Los restantes 22 millones m^3d de GNL (27% del total importado) corresponden a importaciones provenientes de países por fuera de la región (principalmente de Qatar, Nigeria y USA). Por su parte Perú, que exporta el equivalente a 16,7 millones de m^3d de gas en forma de GNL, lo hace a países fuera de la región sudamericana (tales como México, España, Francia, China, entre otros).

En cuanto a los gasoductos de integración, la región posee 16 gasoductos de integración (7 de ellos entre Argentina y Chile, 2 entre Argentina y Uruguay, 1 entre Argentina y Brasil, 3 entre Bolivia y Argentina, 2 entre Bolivia y Brasil y 1 entre Colombia y Venezuela). La capacidad de transporte instalada en dicha infraestructura alcanza los 121 millones m^3d.

2.5.REVISIÓN DE LOS CONTRATOS DE EXPORTACIÓN DE GAS NATURAL E IMPLICACIÓN EN LOS PRECIOS.

Históricamente, las exportaciones de gas natural se concentraron en los mercados de Argentina y Brasil. He aquí un breve resumen de los hitos históricos arraigados en estas décadas:

2.5.1. Brasil

La intención de Bolivia de exportar gas a Brasil se remonta a fines de la década de 1950, por lo que en el Acuerdo Robor (marzo de 1958) se incluyeron algunos temas de alineamiento energético. En la primera mitad de la década de 1970 se firmó un acuerdo para la construcción de un gasoducto, con el objetivo de exportar gas boliviano a Brasil.

A principios de la década de 1990, se firmó un acuerdo para la exportación de gas boliviano, que obligaba a las entonces autoridades bolivianas a participar en un proceso activo de exploración de hidrocarburos para cumplir con el acuerdo. Poco antes de la promulgación de la Ley de

Hidrocarburos 1689, el 30 de abril de 1996, se examinaron los aspectos técnicos inherentes a la construcción de un gasoducto para exportar gas a Brasil, operado por Gas Trans Boliviano.

El contrato de exportación de gas natural tiene una duración de veinte años, contados a partir del primero de julio de 1999, fecha de determinación de los criterios para la determinación de los precios del gas de exportación.

Undécima adenda al contrato de suministro de Gas Natural boliviano al Brasil

La petrolera estatal boliviana YPFB firmó en fecha 5 de agosto de 2022 una nueva adenda al contrato con Petróleo Brasileiro SA, Petrobras, para extender sus exportaciones de gas natural hasta el 2026, lo que representará a Bolivia un ingreso por hasta 6.000 millones de dólares.

El nuevo acuerdo brinda una mayor flexibilidad para el cumplimiento de los compromisos asumidos por YPFB y otorga certidumbre a los volúmenes comprometidos hasta la culminación total del contrato con Petrobras. Por otra parte, en compensación al costo de transporte asumido por YPFB, se consiguió una mejor valorización para el gas y un premio para cantidades adicionales al compromiso de YPFB.

La mejora en las condiciones contractuales permitirá a la estatal petrolera boliviana contar con una mayor disponibilidad del producto, escenario que permitirá incrementar los ingresos para el país.

La adenda prevé el mantenimiento del volumen máximo contratado de 20 millones de metros cúbicos por día (m^3d), con flexibilidad de entrega y recepción de acuerdo con la estacionalidad y disponibilidad de la oferta, asegurando así un suministro en equilibrio contractual para las empresas, la adenda le permitía adecuar los volúmenes comprometidos al mercado brasilero en la presente gestión, de manera de minimizar la posible exposición a multas por fallas de suministro.

2.5.2. Argentina

El contrato de exportación fue firmado el 23 de julio de 1968 entre YPFB/BOLIVIAN GULF OIL y GAS DEL ESTADO, especificando un volumen de 4 MMC/día para los primeros 7 años y 4,5 MMC/día del 8° al 20° año. En él se fijó un precio fijo de 0,2153 USD/MMBTU.

Las exportaciones comenzaron el 1 de mayo de 1972. El 22 de agosto de 1973 se firmó una ley según la cual se introdujo un nuevo precio de exportación (US$0,335 / MMBTU). Luego, el 11 de abril de 1975, se firmó un segundo contrato de prórroga según el cual los precios de exportación se revisaban semestralmente, pero no era posible establecer normas. El 29 de octubre de 1987 se fijaron y ajustaron los precios en base a una fórmula que incluía el precio de una canasta de fuel oil (Nueva York, Mediterráneo y Rotterdam) incluyendo la aplicación de un factor de ajuste adicional a partir de octubre de 1987.

El 1 de mayo de 1992, dos de los dos compradores terminaron el 20 de mayo del mismo año, el nuevo contrato de adquisición se firmó en 20 meses y US $ 1.00 / MMBTU. El 17 de marzo de 1994, el nuevo contrato de compra y venta se firmó a un precio válido durante 3 años, con el siguiente precio: 1994, en invierno, US $ 1.10 US / MMBTU y en verano.

US $ 1,05 / MMBTU; Para 1995 en el invierno 1.20 USD / MMBTU y en el verano.

1.15 USD / MMBTU; 1996 en invierno 1.25 USD / MMBTU y en verano.

1.20 USD / MMBTU, y finalmente, del 1 de julio de 1996, el precio del combustible de aceite de azufre se aplicó 1% de Nueva York. De esta manera, la exportación de gas natural comenzó en mayo de 1972 se implementó en agosto de 1999.

Luego de algunas interrupciones en los envíos y la fijación no programada de precios de exportación, entre los que se encontraba el "precio vinculado" de $0,98/MMBTU, se firmó un acuerdo marco, incluyendo del 15 de julio al 31 de diciembre de 2006, el precio se fijó en 5 u/MMBTU. En este convenio se acordó estudiar y diseñar una fórmula para el cálculo del precio del gas de exportación hasta el final de este proceso, y ya está vigente.

Quinta Adenda al Contrato de Exportación de Gas a la Argentina.

El último día de 2020, YPFB a nombre del gobierno de Bolivia firmó con la Estatal petrolera de la Argentina (IEASA) la Quinta Adenda al Contrato de Compra-Venta de Gas Natural existente entre ambos países. La citada adenda introduce nuevos elementos que son dignos de análisis y que responden a la progresiva declinación de la producción de gas en Bolivia, así como a cambios importantes en el mercado del gas natural en la región.

Esta progresiva reducción de volúmenes de entrega, desde el punto de vista del comprador, obedece a que Argentina alternativamente se abastece de LNG importado de ultramar, mientras que, por el lado de Bolivia, la reducción del volumen comprometido para venta, responde a la acelerada declinación en la producción de los principales campos productores de gas natural y la falta de nuevos descubrimientos de reservorios.

Los principales puntos de la quinta adenda son:

- La obligación de la Argentina, de comprar al menos una cantidad mínima que debe pagar a pesar de no requerir necesariamente ese volumen, se conoce como la cláusula “Tomar o pagar” (Take or pay). Esta obligación de compra, se ha reducido a 10MMm3d, respecto a los 14MMm3d señalados en la Cuarta Adenda para el período de baja demanda (verano) correspondiente a los meses de enero a abril y de octubre a diciembre. Mientras que para

el periodo de alta demanda (invierno) correspondiente a los meses de junio a agosto, el volumen se reduce de 18 MMm3d a 14MMm3d. Finalmente, para los meses mayo y septiembre, el volumen demandado se reduce de 16MMm3d a 13 MMm3d.

- La reducción de volúmenes se hace más notoria en las obligaciones de "Entregar o Pagar" (Delivery or Pay - CDG1) y de "Tomar o pagar" (Take or Pay - CDG2) de la adenda. Por el lado del Delivery or pay se reduce el compromiso de 11 a 10 MMm3d en el verano y de 14 a 13 MMm3d en invierno, este último es 4 MMm3d menor al existente en la Cuarta Adenda de 2019. Por el lado del demandante, la obligación del Take or Pay es de 10 MMm3d en verano y 14 MMm3d en invierno.
- En la Quinta Adenda se mantiene la fórmula de precios establecido en el contrato original para los volúmenes por debajo de la Cantidad Diaria Base de 9 MMm3d.

2.5.3. Precios actuales de exportación

EL cálculo de los precios de exportación del gas natural considera una fórmula determinística. Forman parte de ellas, algunos carburantes (fuel oil), que se cotizan en el mercado internacional bajo la denominación de commodities. Para el caso del contrato ENARSA con la Argentina de los cuatro fuels que forman la canasta, tres son también utilizados por el contrato GSA con el Brasil. Estos precios son publicados diariamente en el Platt's Oilgram Price Report Assessments.

A continuación, se detallen los aspectos relevantes de la determinación de estos precios:

2.5.3.1.Contrato con GSA con el Brasil

El precio de exportación al Brasil se fija cada tres meses y está vinculado a la cotización de tres fuel oils, de acuerdo a la siguiente fórmula:

Ecuación 3 Precio del Gas

$$PG = Pi(0.5 * \frac{FO1}{FO1_0} + 0.25 * \frac{FO2}{FO2_0} + 0.25 * \frac{FO3}{FO3_0}) \dots (3)$$

Donde

PG : Precio del Gas (US$/MMBTU) redondeado al cuarto decimal

Pi : Precio base (US$/MMBTU)

Para la Cantidad Diaria Contractual Base (QDCB) el Pi varía de 0,95 a 1,06 (ver el contrato) y para la Cantidad Diaria Contractual Adicional (QDCA) es 1,20 para todo el periodo de vigencia del Contrato.

FO1 : Fuel Oil de 3,5% de azufre, referido bajo el título de Cargoes FOB Med Basis Italy (US$/TM).

FO2 : Fuel Oil N°6 de 1% de azufre, referido bajo el título U.S. Gulf Coast Waterborne (US$/bbl).

FO3 : Fuel Oil de 1% de azufre, referido bajo el título Cargoes FOB NWE (US$/TM).

FO1, FO2 y FO3 son promedios aritméticos de cada día del trimestre inmediatamente anterior al trimestre correspondiente a la aplicación de PG.

Mientras que FO1o, FO2o y FO3oson promedios aritméticos para los mismos Fuel Oils definidos anteriormente para el periodo comprendido entre el 1ro de enero de 1990 hasta el 30 de junio de 1992, excluyendo el periodo comprendido entre el 1ro de agosto de 1990 al 31 de enero de 1991.

De acuerdo a lo establecido en el Contrato a partir del segundo trimestre de entrega y recepción del gas y para cada trimestre posterior, el precio del Gas (PG) será reajustado aplicándose la siguiente fórmula:

Ecuación 4 Precio del Gas para el Trimestre al Brasil

$$P_t = 0.5PG + 0.5P_{t-1} \dots (4)$$

Donde:

Pt : Precio del Gas para el trimestre pertinente (US$/MMBTU)

PG : Precio del Gas calculado de acuerdo a la fórmula (US$/MMBTU)

Pt-1 : Precio del Gas correspondiente al trimestre inmediatamente anterior (US$/MMBTU).

2.5.3.2.Exportaciones a la Argentina

La fórmula de fijación de precios en el Contrato de exportación de gas natural a Argentina es similar a la estipulada en el Contrato GSA, salvo que se añade a la canasta de tres fueles oils, el precio internacional de diesel oil, de acuerdo a la siguiente fórmula:

Ecuación 5 Precio del Gas de exportación a la Argentina

$$PG = P(0.2 * \frac{FO1}{FO1_0} + 0.4 * \frac{FO2}{FO2_0} + 0.2 * \frac{FO3}{FO3_0} + 0.2 * \frac{DO_i}{DO_0}) \dots (5)$$

Donde:

PG: Precio del Gas (US$/MMBTU)

P: Precio base igual a 4,0588 US$/MMBTU

FO1, FO2 y FO3 son los mismos del Contrato GSA.

FO1i, FO2i, FO3i y DOi son promedios aritméticos de cada día del semestre inmediatamente anterior al trimestre correspondiente a la aplicación de PG.

El precio “P” de 4,0588US$/MMBTU, fue determinado endógenamente para que en el inicio del contrato el precio PG tome un valor de 5US$/MMBTU.

Con la quinta adenda de venta de gas a la Argentina se tiene un parámetro de volúmenes adicionales a la Cantidad Diaria Base de 9 MMm3d se introduce una nueva fórmula de precios que elimina el componente LNG de la Cuarta Adenda y se introduce el Índice de Precios Henry Hub (HH), como precio de referencia para valorar los volúmenes adicionales que pueda solicitar Argentina durante los meses de invierno la cual se expresa en la ecuación 6.

Ecuación 6 Precios Henry Hub (HH)

$$\mathrm{PG}_{\text{volumenes Adicionales}} = \mathrm{HH}_{\mathrm{i}} + 2.25\,\frac{\mathrm{USD}}{\mathrm{MMBTU}} \ldots\ldots (6)$$

Donde:

$\mathrm{PG}_{\text{Volúmenes Adicionales}}$=Precio del gas de volúmenes adicionales.

HH_{i}=Precio Henry Hub.

La inclusión del precio de referencia HH, es claramente favorable para la Argentina, debido a que elimina el efecto de los altos precios del LNG que se observaron en invierno de 2020-2021. El precio Henry Hub, tiende a subir en invierno, (verano para el hemisferio sur) y a bajar en verano (invierno en el hemisferio sur), lo cual va en contra de lo que sería un buen negocio para Bolivia, sin embargo, es óptimo para los intereses de Argentina. Un aspecto importante a remarcar sobre la novedosa inclusión del precio HH como referente en el precio de venta de gas natural a la Argentina, es que no existe una razón técnica para ello. El HH es un indicador del mercado estadounidense que no tiene relación con el mercado regional de América del Sur, además que como se mencionó, estacionalmente es contra cíclico.

2.6.CIENCIA DE DATOS

La ciencia de datos combina varios campos, como estadísticas, métodos científicos, inteligencia artificial (IA) y análisis de datos para extraer valor de los datos. Un analista de datos se llama analista de datos y combina varias habilidades para analizar los datos recopilados de Internet, teléfonos inteligentes, clientes, sensores y otras fuentes para hacer realidad la información. La ciencia de datos implica la preparación de datos para el análisis, incluida la limpieza, la agregación y el procesamiento de datos para un análisis avanzado. Las aplicaciones de análisis y los analistas de datos pueden analizar los resultados para descubrir patrones y permitir que los líderes empresariales tengan información confiable.

2.6.1. Datos

Los datos recogen un conjunto de hechos (una base de datos, BD) y los patrones son expresiones que describen un subconjunto de los datos (un modelo aplicable a ese subconjunto). BD involucra un proceso iterativo e interactivo de búsqueda de modelos, patrones o parámetros. Los patrones descubiertos han de ser válidos, novedosos para el sistema (para el usuario siempre que sea posible) y potencialmente útiles.

Se han de definir medidas cuantitativas para los patrones obtenidos (precisión, utilidad, beneficio obtenido...). Se debe establecer alguna medida de interés que considere la validez, utilidad y simplicidad de los patrones obtenidos mediante alguna de las técnicas de minería de datos. El objetivo final de todo esto es incorporar el conocimiento obtenido en algún sistema real, tomar decisiones a partir de los resultados alcanzados o, simple- mente, registrar la información conseguida y suministrársela a quien esté interesado.

2.6.2. Minería de datos

La minería de Datos es un término genérico que engloba resultados de investigación, técnicas y herramientas usadas para extraer información útil de grandes bases de datos. Si bien minería de datos es una parte del proceso completo de BD, en buena parte de la literatura, los términos minería de datos y BD se identifican como si fueran lo mismo.

Algo muy importante de la minería de datos es la transformación de los datos, es posible que al principio antes de la transformación, sea más pequeño o grande que los datos finales.

Los procesos de BD involucran técnicas como:

- Bases de datos que permiten almacenar los datos de forma estructurada, tanto a nivel lógico (con la aparición en los últimos años de las bases de datos no relacionales) y a nivel de hardware (con la capacidad para el proceso en clúster).
- Técnicas de visualización que permiten realizar representaciones gráficas de los datos que facilitan la labor del usuario a la hora de entender los datos, para filtrarlos y procesarlos.
- Técnicas estadísticas, que permiten analizar analíticamente los datos almacenados en las bases de datos y desarrollar modelos estadísticos que los expliquen.
- Técnicas de aprendizaje automático, que permiten desarrollar modelos conceptuales que representan los datos almacenados en la base de datos.

2.6.2.1.Técnicas algebraicas y estadísticas

Este tipo de técnicas expresan modelos y patrones mediante el uso de fórmulas algebraicas, funciones lineales, funciones no lineales, distribuciones, varianzas, correlaciones y más. Suelen extraer el patrón con la ayuda de un modelo previamente predeterminado y generar a partir de él

unos parámetros o coeficientes. Esta es la razón por la que a menudo se les llama técnicas paramétricas.

2.6.2.2.Técnicas bayesianas

Utilizan el teorema de Bayes (como su nombre indica) para evaluar la probabilidad de pertenecer a una clase o grupo mediante la estimación de probabilidades condicionales inversas o previas. Los algoritmos más utilizados en dichas técnicas son los métodos basados en la máxima verosimilitud, los algoritmos EM y los clasificadores Naive Bayes. Gracias a las técnicas bayesianas se pueden representar gráficamente interacciones e interacciones probabilísticas entre variables.

2.6.2.3.Técnicas basadas en arboles de decisión y sistemas de aprendizaje de reglas

Este tipo de técnicas se representan en forma de reglas y se basan principalmente en dos tipos diferentes de algoritmos: algoritmos de tipo "divide y vencerás".

Arboles de Decisiones. Técnica que permite analizar decisiones secuenciales basadas en el uso de los resultados y probabilidades asociadas.

Los árboles de decisión se pueden usar para generar sistemas expertos, búsquedas binarios y árbol de juegos, los cuales serán explicados posteriormente.

2.7.INTELIGENCIA ARTIFICIAL

Las máquinas de IA pueden realizar varias operaciones similares al comportamiento humano, como devolver respuestas a cada entrada (similar a las reacciones de un organismo), o buscar un estado entre todas las operaciones posibles basadas en una acción o resolución de problemas a través de la lógica formal.

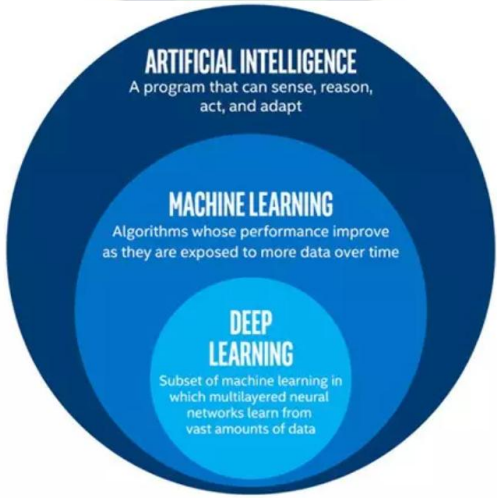
Cuando estos dispositivos tienen la capacidad de aprender y discriminar, se convierten en entidades cercanas a lo paranormal, alcanzando velocidades de procesamiento imposibles para los humanos y que no requieren descanso para funcionar, entre las ventajas que colocan por encima de los organismos vivos.

El machine learning es también condición necesaria para que un ente artificial pueda ser considerado inteligente. Si una entidad maquinaria no es capaz de aprender cosas nuevas, difícilmente será capaz de adaptarse al medio, condición exigible de investigación actuales busca hacer que las maquinas sean capaces de hacer generalizaciones a partes de los ejemplos sacados del entorno.

2.7.1. Aprendizaje automático

En la figura 2.1 se apreciará como se relacionan las diferentes áreas de Inteligencia Artificial, Aprendizaje Automático y Redes Neuronales.

Figura 2.1. Relación entre IA, ML, DL



Fuente: https://hardzone.es/tutoriales/rendimiento/diferencias-ia-deep-machine-learning/

Toda inteligencia para saber ha de aprender, por lo que necesita datos para sacar conclusiones a partir de los mismos. De la misma manera que el pensamiento lógico se basa en premisas para sacar la información, y es que los sistemas de nuestros ordenadores se basan en la lógica pura y dura. Un programa normal lo que hace es ejecutarse, pero un algoritmo de inteligencia artificial está pensado para aprender y esto significa corregir errores.

Es decir, esta innovación les da a los ordenadores el poder de resolver las cosas sin que las tengan programadas explícitamente. Como ejemplo de un aprendizaje automático, supongamos que quieres que un programa pueda identificar gatos en imágenes:

- Dale a tu IA un conjunto de características de cómo es un gato, para que sepa reconocerlo. Colores, formas, etc.
- Muéstrale imágenes (si alguna está etiquetada como «gato», la IA podrá identificarla más fácilmente).
- Una vez que el programa haya visto suficientes gatos, debería ser capaz de identificarlos en otras imágenes: «si la imagen contiene ciertas características, entonces hay un 95% de que sea un gato».

Puede sonar complicado, pero un ejemplo de la capacidad de aprendizaje la tenemos en una serie de utilidades, ya que gracias a ella cosas tan cotidianas como el filtro del correo basura o las recomendaciones de Netflix u otros servicios similares para cada uno de nosotros se hacen posibles, todo gracias a los datos que ha recopilado el sistema.

Es por ello que a día de hoy y con todo el mundo conectado a la red de redes y compartiendo sus datos el Machine Learning se ha vuelto tan importante, ya que ayuda a las diferentes empresas a

tener perfiles de consumo de los usuarios. Aunque no es la única aplicación, pero si la más sencilla de entender y la que está más al alcance de todos.

2.7.2. Aprendizaje supervisado

Aquí es cuando entrenamos un algoritmo de aprendizaje automático dándole preguntas (características) y respuestas (etiquetas). Entonces, en el futuro, el algoritmo puede hacer predicciones al conocer las características.

En este tipo de aprendizaje, existen dos tipos de algoritmos (de aprendizaje): clasificación y regresión.

2.7.2.1.Algoritmos de clasificación

Esperamos que el algoritmo nos diga a qué grupo pertenece el elemento en estudio. El algoritmo encuentra patrones en los datos que le damos y los clasifica en grupos. Luego compara los nuevos datos y los ubica en uno de los grupos y es así como puede predecir de que se trata.

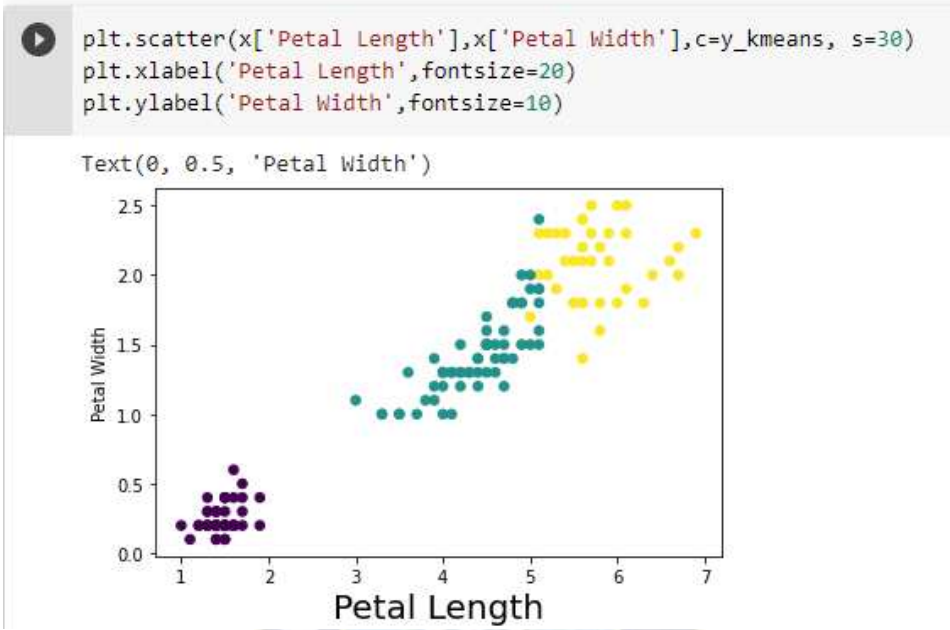
La variable por predecir es un conjunto de estados discretos o categóricos. Pueden ser:

- Binaria: {Sí, No}, {Azul, Rojo}, {Fuga, No Fuga}, etc.
- Múltiple: Comprará {Producto1, Producto 2…}, etc.

Ordenada: Riesgo {Bajo, Medio, Alto}, etc.

En la figura 2.2 se presentará un ejemplo muy conocido de clasificación que es el dataset de Iris, que es la clasificación de 3 distintos tipos de flores. Estas toman como parámetros como la longitud de sus pétalos, el ancho y otros parámetros.

Figura 2.2: Clasificador de Dataset Iris



Fuente: Ronald Fisher con los datos recolectados en 1936.

La figura 2.2 representa a la clasificación de 3 tipos de plantas por Ronald Fisher en 1936 hoy en día pese a su antigüedad es un dataset bastante usado para problemas de clasificación, ya que el objetivo suele ser predecir a que tipo de flor perene una flor de la especie iris.

2.7.2.2.Algoritmo de regresión

En este método lo que se espera es un número. No lo ubica en un grupo, sino que devuelve un valor específico.

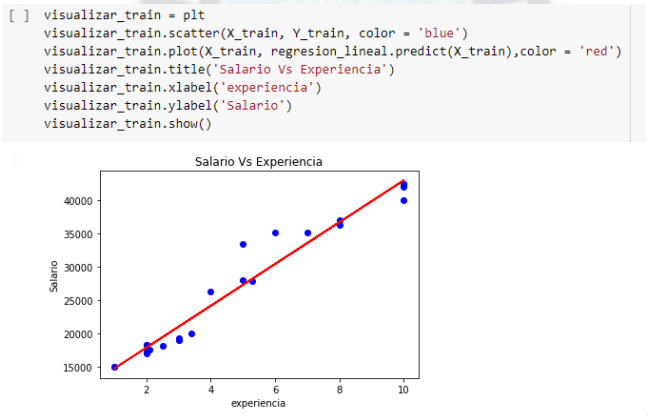
Los algoritmos de regresión más conocidos son:

- El algoritmo de regresión simple (en que solo se necesita una variable independiente y otra variable dependiente),

- Regresión lineal múltiple en que pueden a ver 2 o más variable independientes con un variable dependiente.
- Regresión logística donde el objetivo es clasificar la variable dependiente.

Por ejemplo, el precio de una casa. El algoritmo tiene el precio de diferentes casas, pequeñas, grandes, en el campo, en la ciudad, etc. y por medio de un gráfico de dispersión, puede predecir el precio correcto de una casa en consulta. En la figura 2.3 representa la relación que existe entre el salario de una persona y la experiencia.

Figura 2.3: Ejemplo de regresión lineal



SS

Fuente: Google Colaboratory con la librería Matplotlib

2.7.3. Aprendizaje no supervisado

Aquí solo le damos las características al algoritmo, nunca las etiquetas. Queremos que nos agrupe los datos que le dimos según sus características. El algoritmo solo sabe que como los datos comparten ciertas características, de esa forma asume que pueda que pertenezcan al mismo grupo.

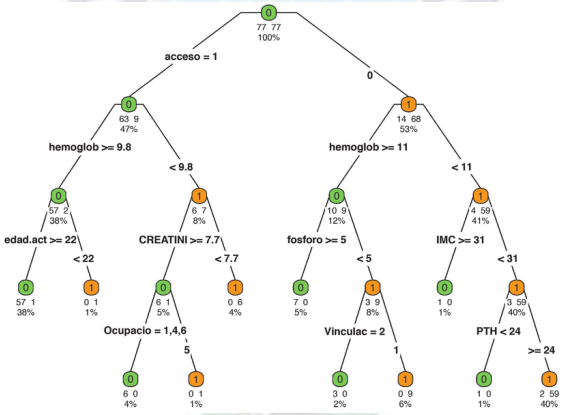
2.7.4. Modelos de aprendizaje automático

Los algoritmos de Machine Learning, se pueden agrupar en tres modelos:

2.7.4.1.Modelos de árbol

Son más precisos, estables y fáciles de interpretar como modelos porque construyen reglas de decisión que se pueden representar como árboles. A diferencia de los modelos lineales, pueden representar relaciones no lineales para resolver problemas. En estos modelos se distinguen árboles de decisión y bosques aleatorios (árboles de decisión promedio). Al ser más precisos y precisos, obviamente ganamos previsibilidad, pero perdemos rendimiento. La figura 2.4 representa un ejemplo de árbol de decisiones.

Figura 2.4: Ejemplo de Árbol de decisiones



Fuente: https://www.researchgate.net/figure/Figura-1-Arboles-de-clasificacion-y-regresion-Classification-and-Regression-Tree-CART_fig1_321073328

Las ventajas de un árbol de decisión son:

- Resumir los ejemplos iniciales y clasificar nuevos casos siempre que las condiciones que generan los ejemplos utilizados para la construcción no hayan cambiado sustancialmente.
- Facilitar la interpretación de las decisiones adoptadas.

- Proporcionar un alto nivel de comprensión del conocimiento utilizado en la toma de decisiones.
- Explicar el comportamiento sobre una tarea de decisión.
- Reducir el número de variables independientes.
- Es una excelente herramienta para el control de la gestión empresarial.

Los árboles de decisión se utilizan en cualquier proceso que implique toma de decisiones, ejemplos de estos procesos son:

- Búsqueda binaria.
- Sistemas expertos.
- Arboles de juego.

2.7.4.1.1. Árbol de decisión regresor.

Los árboles de decisión son una técnica de aprendizaje supervisado que predice valores de respuestas mediante el aprendizaje de reglas de decisión derivadas de características. Se pueden utilizar tanto en una regresión como en un contexto de clasificación.

Este algoritmo es muy bueno en el manejo de datos tabulares con características numéricas o características categóricas con menos de cientos de categorías.

Los Árboles de regresión es cuando el resultado predicho se puede considerar un número real (por ejemplo, el precio de una casa, o el número de días de estancia de un paciente en un hospital). En nuestro caso particular el precio del Gas Natural.

Las ventajas que presenta este modelo son:

- Fácil de entender la salida del árbol de decisión es muy fácil de entender.

- Útil en la exploración de datos el árbol de decisiones es una de las formas más rápidas para identificar las variables más significativas y la relación entre dos o más.

- Se requiere menos limpieza de datos. Requiere menos limpieza de datos en comparación con algunas otras técnicas de modelado. A su vez, no está influenciado por los valores atípicos y faltantes en la data.

- El tipo de datos no es una restricción. Puede manejar variables numéricas y categóricas.

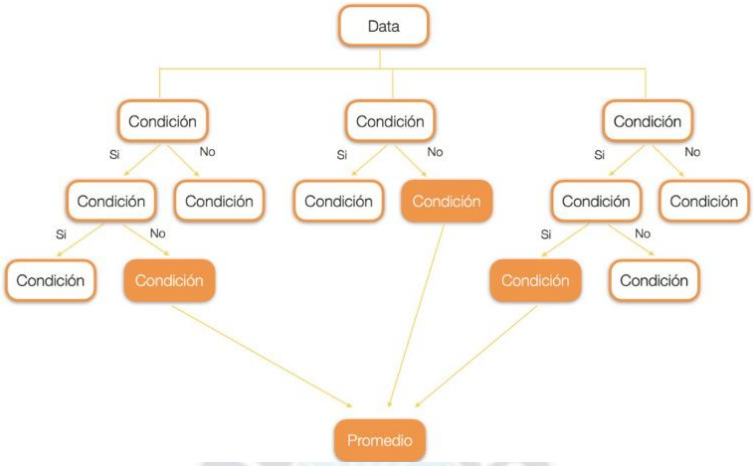
Las desventajas son las siguientes:

- Sobreajuste. Es una de las dificultades más comunes que tiene este algoritmo, este problema se resuelve colocando restricciones en los parámetros del modelo y eliminando ramas en el análisis.
- No apto para variables continuas. Al trabajar con variables numéricas continuas, el árbol de decisiones pierde información cuando categoriza variables en diferentes categorías.
- Los modelos basados en árboles no están diseñados para funcionar con características muy dispersas.

2.7.4.1.2. Bosque Aleatorio regresor.

Los Bosques Aleatorios es un algoritmo de aprendizaje supervisado que, como ya se puede ver en su nombre, crea un bosque y lo hace de alguna manera aleatorio. Para decirlo en palabras simples: el Bosque Aleatorio crea múltiples árboles de decisión y los combina para obtener una predicción más precisa y estable. En general, mientras más árboles en el bosque se vea, más robusto es el bosque.

Figura 2.5: Bosque aleatorio.



Fuente: Escuela AprendeIA

Las ventajas del modelo de bosque aleatorio son las siguientes:

- Puede resolver ambos tipos de problemas, es decir, clasificación y regresión, y realiza una estimación decente en ambos frentes.
- Unos de los beneficios que más llama la atención es el poder de manejar grandes cantidades de datos con mayor dimensionalidad. Puede manejar miles de variables de entrada e identificar las variables más significativas, por lo que se considera uno de los métodos de reducción de dimensionalidad. Además, el modelo muestra la importancia de la variable, que puede ser una característica muy útil.
- Tiene un método efectivo para estimar datos faltantes y mantiene la precisión cuando falta una gran proporción de los datos.

Las desventajas son las siguientes:

- Hace un buen trabajo en la clasificación, pero no es tanto bueno como para los problemas de regresión, ya que no proporciona predicciones precisas y continuas sobre la naturaleza.

- En ocasiones se puede parecer este algoritmo como una caja negra, ya que se tiene muy poco control sobre lo que hace el modelo. Puedes, en el mejor de los casos, probar diferentes parámetros y datos aleatorios.

En resumen, un bosque aleatorio es una colección de árboles de decisión, como se indicó con las figuras 3.11, 3.12 y 3.13 las ramas que se generarían de cada árbol son las respectivas, valores de día, mes y año que son valores enteros lo que se puede traducir en programación como una condición y se usara la versión regresor porque la variable dependiente el precio del gas natural es un dato continuo.

2.7.4.2.Modelos lineales

Estos tratan de encontrar una línea que se "ajuste" bien a la nube de puntos que se disponen. Aquí destacan desde modelos muy conocidos y usados como la regresión lineal (también conocida como la regresión de mínimos cuadrados), la logística (adaptación de la lineal a problemas de clasificación -cuando son variables discretas o categóricas-). Estos dos modelos tienen el problema del "overfit", esto significa que se ajustan "demasiado" a los datos disponibles, con el riesgo que esto tiene para nuevos datos que pudieran llegar. Al ser modelos relativamente simples, no ofrecen resultados muy buenos para comportamientos más complicados.

En este algoritmo se agrega aleatoriedad adicional al modelo, mientras crece los árboles, en lugar de buscar la característica más importante al dividir un nodo, busca la mejor característica entre un subconjunto aleatorio de características. Esto da como resultado una amplia diversidad que generalmente resulta en un mejor modelo. Por lo tanto, en Bosques Aleatorios, el algoritmo para dividir un nodo sólo tiene en cuenta un subconjunto aleatorio de las características. Incluso puede hacer que los árboles sean más aleatorios, mediante el uso adicional de umbrales aleatorios para

cada función en lugar de buscar los mejores umbrales posibles, como lo hace un árbol de decisión normal.

2.7.4.2.1. Regresión lineal simple

Los mínimos cuadrados implican calcular la suma de las distancias al cuadrado entre los puntos reales y los puntos definidos por las líneas estimadas por las variables introducidas en el modelo, de modo que la mejor estimación minimice estas distancias. Para determinar qué modelo se ajusta mejor a los datos disponibles en el modelo de regresión lineal, compare la F parcial obtenida en cada modelo de regresión construido.

Si tuviéramos que utilizar alguna de las técnicas de selección de variables descritas anteriormente, este coeficiente se calcularía cada vez que se eliminara o introdujera una variable, ya que al hacerlo, en realidad se estaría estimando un nuevo modelo de regresión. En todos los casos, el paquete de estadísticas hace esto automáticamente, a menos que usemos una técnica que fuerce la entrada de todas las variables, en cuyo caso estimamos manualmente todos los modelos posibles y luego elegimos. Otra forma de validar un modelo es evaluar los residuos de regresión, la diferencia entre el valor estimado por el modelo y el valor observado, y por lo tanto la parte del modelo de regresión que no se puede explicar.

Si el modelo de regresión es suficiente para explicar nuestros datos, los residuos deben distribuirse normalmente con una media de 0 y una varianza constante. Esta hipótesis se puede verificar gráficamente representando cómo se distribuyen los residuos de nuestro modelo de regresión como una nube de puntos. Con este enfoque, se puede diagnosticar la no linealidad o la heteroscedasticidad (cuando la varianza no es constante).

El peligro de introducir valores extremos en el modelo de regresión, aunque en realidad son valores registrados, pueden causar cambios significativos en los resultados de la regresión lineal debido a que se estima con base en el método de mínimos cuadrados, como se muestra, y en base a la diferencia entre el cálculo de la distancia de los puntos. Por tanto, es necesario tenerlos en cuenta a la hora de ajustar el modelo, estimar dos modelos: uno que incluya el valor declarado, otro que excluya el valor declarado y finalmente evaluar qué resultados son mejores para nuestros propósitos.

2.7.4.2.2. Regresión lineal múltiple

Este modelo lineal está relacionado a una variable dependiente Y con m variables regresoras Xj con j=1, 2,…..,m o cualquier transformación de estas que generan un hiperplano de parámetro βi desconocidos:

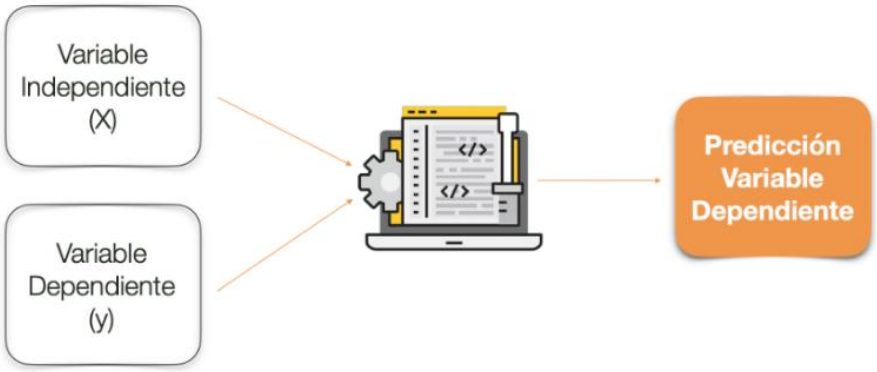
Ecuación 7 Regresión lineal múltiple

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m + e \ldots (6)$$

Donde e es una variable aleatoria que recoge todos aquellos factores de la realidad no controlables u observables y que por tanto se asocian con el azar.

En la figura 2.6 se representa como opera el modelo de regresión lineal.

Figura 2.6: Regresión lineal.



Fuente: Fuente: Escuela AprendeIA.

2.7.4.3.Redes neuronales

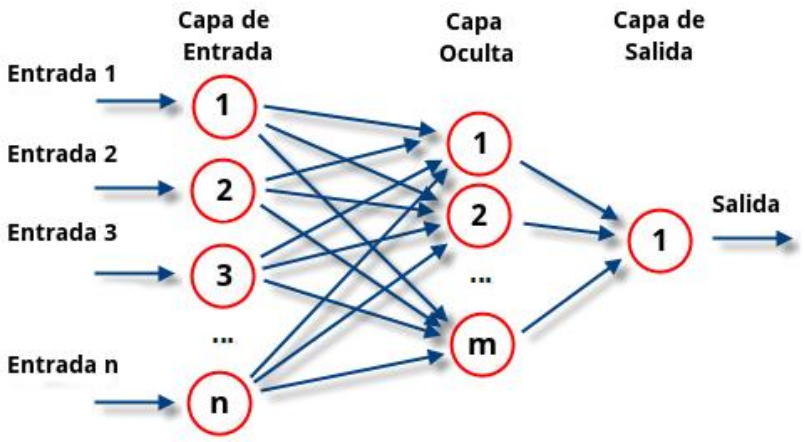
Las redes artificiales de neuronas tratan, en cierto modo, de replicar el comportamiento del cerebro, donde tenemos millones de neuronas que se interconectan en red para enviarse mensajes unas a otras. Esta réplica del funcionamiento del cerebro humano es uno de los "modelos de moda" por las habilidades cognitivas de razonamiento que adquieren.

El reconocimiento de imágenes o vídeos, por ejemplo, es un mecanismo complejo y una red neuronal es lo mejor para realizarlo.

El problema, como ocurre con el cerebro humano, es que son lentas de entrenar y necesitan mucha capacidad de cómputo. Quizás sea uno de los modelos que más ha ganado con la "revolución de los datos".

En la figura 2.7 se muestra cómo funciona a nivel básico una red neuronal y mostrando sus componentes básicos como ser las entradas, capas de entrada, capas ocultas, capas de salida y la salida de la red neuronal.

Figura 2.7: Redes neuronales

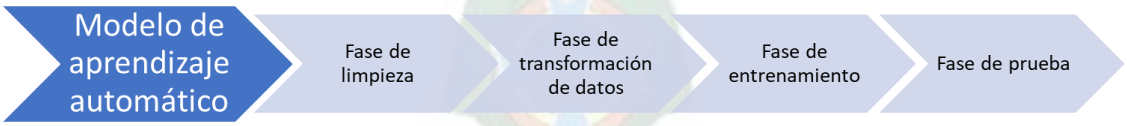


Fuente: https://www.atriainnovation.com/que-son-las-redes-neuronales-y-sus-funciones/

2.8.FASES DEL DESARROLLO PARA UN MODELO DE APRENDIZAJE AUTOMÁTICO

En la Figura 2.8 se representará las fases que se realizan para realizar un modelo de aprendizaje automático.

Figura 2.8: Fases para desarrollar ML



Fuente: Elaboración propia usando power point.

2.8.1. Fase de limpieza

La limpieza de datos es el proceso de corregir o eliminar datos incorrectos, corruptos, formateados incorrectamente, duplicados o incompletos dentro de un conjunto de datos.

Cuando se combinan varias fuentes de datos, se dan muchas circunstancias para que los datos se dupliquen o se etiqueten incorrectamente. Si los datos son incorrectos, los resultados y los algoritmos no son fiables, aunque parezcan correctos.

No existe una forma absoluta de prescribir los pasos exactos en el proceso de limpieza de datos porque los procesos variarán de un conjunto de datos a otro. Pero es fundamental establecer una plantilla para el proceso de limpieza de datos para que sepas que lo estás haciendo de la manera correcta en todo momento.

2.8.2. Fase de transformación

En esta fase básicamente se debe tener en cuenta que se pueden transformar los datos, imagínese el caso en el que se pregunta a los usuarios su sexo que podrían escribir masculino o femenino, donde el objetivo es predecir el preso de la persona estos valores descritos de manera literaria no aportarían a nuestro modelo. Pero si se transformarían como: masculino transformado en el valor de 1 y femenino transformado en el valor de 0. Sin duda los valores de 0 y 1 ayudarían a desarrollar y ser tomados en cuenta en nuestro modelo.

2.8.3. Fase de entrenamiento

En esta etapa hay una gran cantidad de datos, parte de los cuales se descomponen para entrenar al algoritmo y darle toda esta información hasta que encuentre los patrones necesarios y luego pueda hacer predicciones.

2.8.3.1.Overfitting

Este concepto es uno de los conceptos clave en aprendizaje automático. Se denomina sobreajuste al hecho de hacer un modelo tan ajustado a los datos de entrenamiento que haga que no generalice bien a los datos de test.

Hay que recordar que el objetivo de los modelos de aprendizaje automático es el de obtener patrones de los datos de entrenamiento disponibles de cara a predecir o inferir correctamente datos nuevos. Es decir, el concepto clave es el de entrenar y obtener patrones generales que sean extrapolables a nuevos datos. Algo similar ocurre en el aprendizaje de los seres humanos, el sobreajuste se produciría cuando aprendemos las cosas de memoria, sin entender el concepto.

El sobreajuste se produce cuando un sistema de aprendizaje automático se entrena demasiado o con datos anómalos, que hace que el algoritmo «aprenda» patrones que no son generales. Aprende características específicas, pero no los patrones generales, el concepto. Los modelos más complejos tienden a sobre ajustar más que lo modelos más simples. Además, ante un mismo modelo, a menor cantidad de datos es más posible que ese modelo se sobreajuste.

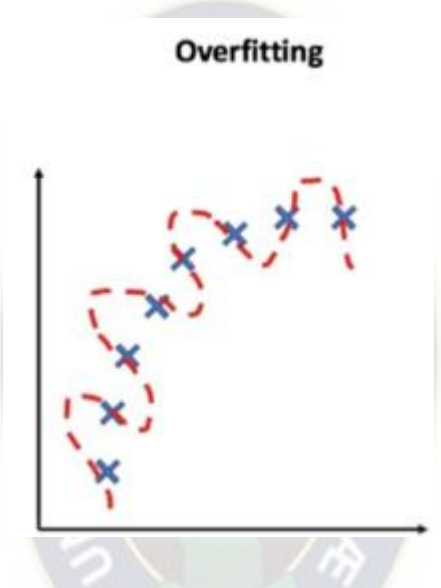
El sobreajuste se puede evitar de varias formas, las más claras son las siguientes:

- Incorporando mayor cantidad de datos: al tener más cantidad de datos es más probable que el algoritmo generalice mejor, al tener en cuenta más tipos de datos.
- Cambiando los parámetros de ciertos algoritmos, haciendo los algoritmos más simples: haciendo que el algoritmo sea más simple, se ajusta menos a los datos y es menos posible sobre ajustar a los datos de entrenamiento. Por ejemplo, reduciendo la profundidad de un árbol de decisión se ajusta menos al hacer el modelo más simple.

- Los parámetros que nos ayudarían a medir el modelo serian erróneos debido a que estamos sobre entrenado el modelo.
- Se obtiene una exactitud falsa. Próxima a la unidad o la unidad.

En la figura 2.9 representa como se puede dar un caso de Overfitting al entrenar el modelo.

Figura 2.9: Overfitting



Fuente: https://ichi.pro/es/investigacion-del-ajuste-insuficiente-y-excesivo-123386495099996

2.8.4. Underfitting

Underfitting se refiere al escenario en el que un modelo de aprendizaje automático no puede generalizarse o encajar bien en un conjunto de datos invisible. Una clara señal de sobreajuste del aprendizaje automático es si tu error en el conjunto de datos de prueba o validación es mucho mayor que el error en el conjunto de datos de entrenamiento.

El subajuste es un término utilizado en las estadísticas que se refiere a un error de modelado que se produce cuando una función se corresponde demasiado con un conjunto de datos. Como

resultado, el sobreajuste puede no ajustarse a datos adicionales y esto puede afectar la precisión de la predicción de observaciones futuras. Es lo contrario al overfitting.

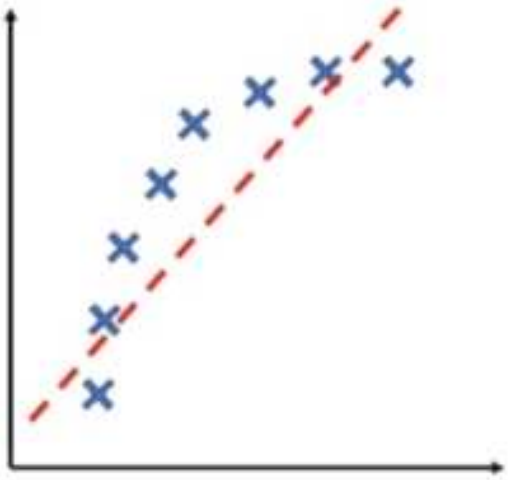
El underfitting ocurre cuando un modelo aprende los detalles y el ruido en el conjunto de datos de entrenamiento en la medida en que impacta negativamente el rendimiento del modelo en un nuevo conjunto de datos. Esto significa que el modelo capta y aprende el ruido o las fluctuaciones aleatorias en el conjunto de datos de entrenamiento como conceptos.

Las principales causas del underfitting o sobreajuste son:

- No hay suficientes parámetros o complejidad para modelar adecuadamente los datos.
- Los priores bayesianos son demasiado restrictivos o ciertos (baja entropía).
- No se le dio suficiente tiempo al algoritmo de aprendizaje automático para entrenar.

En la figura 2.10 representa un caso de cómo puede presentarse underfitting en el modelo.

Figura 2.10: Underfitting



Fuente: https://ichi.pro/es/investigacion-del-ajuste-insuficiente-y-excesivo-123386495099996

2.8.5. Fase de entrenamiento

En esta etapa hay una gran cantidad de datos, parte de los cuales se descomponen para entrenar al algoritmo y darle toda esta información hasta que encuentre los patrones necesarios y luego pueda hacer predicciones.

2.8.6. Fase de prueba

El resto de los datos restantes se utilizarán para las pruebas. De esta forma, podemos cuestionar el algoritmo y evaluar si las respuestas son verdaderas o falsas, y si aprenderá. Si encontramos que los datos no coinciden, necesitaremos agregar más datos o cambiar el método que usamos. Pero si notamos que hay un 70% a 90% de respuestas correctas, entonces podemos decir que hay un buen nivel de aprendizaje y se puede usar este algoritmo.

2.9.MÉTRICAS

La métrica que se usara para predecir el precio del Gas natural es la siguiente:

2.9.1. Coeficiente de determinación

En estadística, el coeficiente de determinación, denominado R^2 (se pronuncia R cuadrado), es un estadístico usado en el contexto de un modelo estadístico cuyo principal propósito es predecir futuros resultados o probar una hipótesis. El coeficiente determina la calidad del modelo para replicar los resultados, y la proporción de variación de los resultados que puede explicarse por el modelo.

Hay varias definiciones diferentes para R^2 que son algunas veces equivalentes. Las más comunes se refieren a la regresión lineal. En este caso, el R^2 es simplemente el cuadrado del coeficiente de correlación de Pearson, lo cual es sólo cierto para la regresión lineal simple. Si existen varios

resultados para una única variable, es decir, para una X existe una Y, Z... el «coeficiente de determinación» resulta del cuadrado del coeficiente de determinación múltiple. En ambos casos el R^2 adquiere valores entre 0 y 1. Existen casos dentro de la definición computacional de R^2 donde este valor puede tomar valores negativos.2

Ecuación 8: Coeficiente de determinación

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - y_{pred}\right)^2}{\sum_{i=1}^{N}\left(y_i - y_{promedio}\right)^2} \ldots\ldots (7)$$

Donde:

y_i= Representa al valor de la variable dependiente.

y_{pred}= Es el valor que predice el modelo.

$y_{promedio}$= Es el promedio de la variable dependiente.

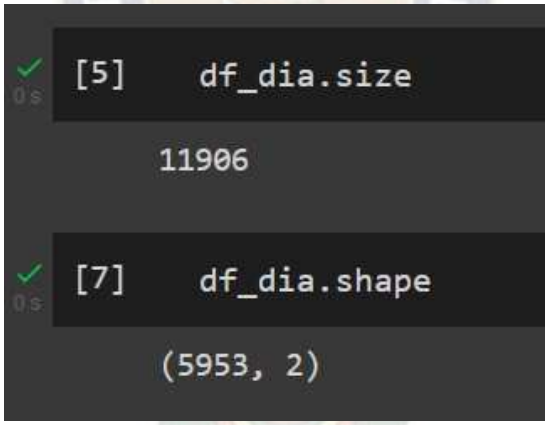
3. CAPÍTULO III: APLICACIÓN PRÁCTICA

Para el desarrollo de la aplicación se usará el lenguaje de programación Python 3.9 y en el entorno de desarrollo de Google Colaboratory además de sus librerías, sklearn, matplotlib, seaborn, numpy, pandas y stasmodels.

3.1.FASE DE EXPLORACIÓN

La data fue obtenida en kaggle en el siguiente link: https://www.kaggle.com/tunguz/natural-gas-prices. La cual representa los precios del gas natural recolectados desde 1997 hasta el 2020 que se compone de 2 columnas las cuales son la fecha y el precio. La información de los datos medidos para cada día se expresada en la figura 3.1.

Figura 3.1: Datos analizados por día.

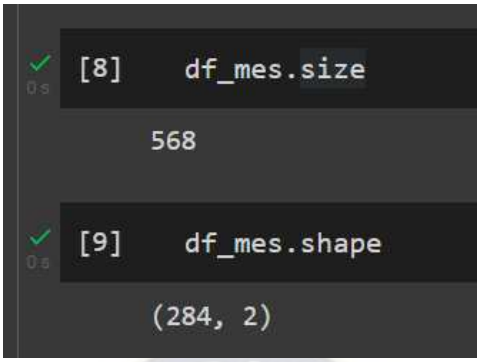


Fuente: Google Colaboratory usando datos de Kaggle.

De la figura 3.1 vemos que los datos en total representan 11906 datos, en la segunda parte se que son 5953 filas y 2 columnas.

La información de los datos medidos para cada mes se ve expresada en la figura 3.2.

Figura 3.2: Datos analizados por mes



Fuente: Google Colaboratory usando datos de Kaggle.

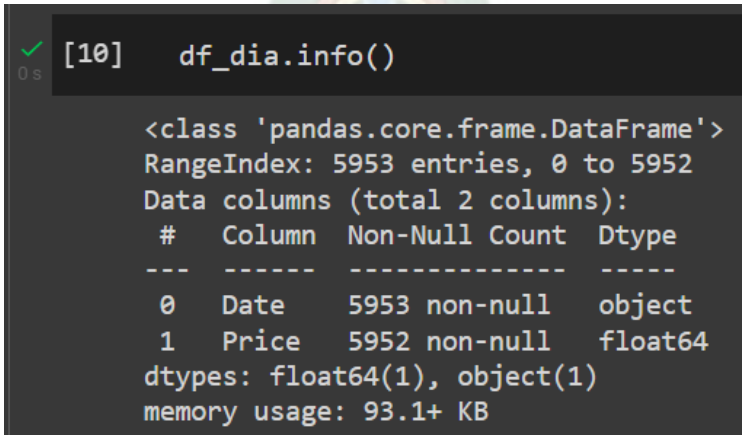
La figura 3.2 se ve que el tamaño total de datos es de 568 y para ser más precio se tiene 284 filas y 2 columnas.

3.2.FASE DE LIMPIEZA

En la parte de la limpieza de datos se debe identificar los datos que se expresen como NAN, en Python presenta facilidad para encontrar estos valores.

En la figura 3.3 se vera la cantidad de valor nulos para los datos examinados por día.

Figura 3.3: Valores nulos por datos examinados por día



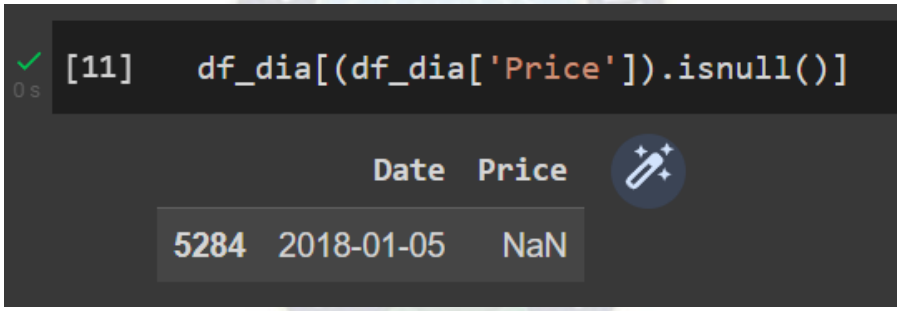
Fuente: Google Colaboratory usando datos de Kaggle.

La columna [‘Date’] representa la fecha, ahora bien, una observación anticipada que se puede hacer es que es de tipo object, posteriormente se tendrá que cambiar a tipo de dato datatime.

La columna [‘Price’] que representa el precio del u/MMBTU anotado en ese día, es de tipo float64 lo cual normal y no se presenta alguna observación. Pero al comparar el número de datos entre las columnas se nota que la columna [‘Date’] cuenta con 5953 datos mientras que la columna [‘Price’] cuenta con 5952, esto lleva a concluir que hay un valor nulo en la columna [‘Price’].

En la figura 3.4 se procederá a encontrar el valor nulo en el df_dia.

Figura 3.4: Encontrando el valor nulo en df_dia



Fuente: Google Colaboratory usando datos de Kaggle.

El valor NAN representa que se ha perdido el valor del precio en el índice 5284, perteneciente a la fecha del 5 de enero del 2018.

En la figura 3.5 se procederá a reemplazar el valor NAN por el promedio de los 4 datos anteriores y posteriores al NAN.

Figura 3.5: Promedio de valores superiores y posteriores al NAN

```
[14]  promedio = df_dia.loc[5280:5289,['Price']].median()
      df_dia = df_dia.fillna(value=promedio)
      df_dia.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5953 entries, 0 to 5952
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Date    5953 non-null   object
 1   Price   5953 non-null   float64
dtypes: float64(1), object(1)
memory usage: 93.1+ KB
```

Fuente: Google Colaboratory usando datos de Kaggle.

Como se puede observar en el output ambas columnas tienen el mismo número de datos validos 5953.

En la figura 3.6 se vera la cantidad de valor nulos para los datos examinados por mes.

Figura 3.6: Valores nulos por datos examinados por mes

```
[15]  df_mes.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 284 entries, 0 to 283
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Month   284 non-null    object
 1   Price   284 non-null    float64
dtypes: float64(1), object(1)
memory usage: 4.6+ KB
```

Fuente: Google Colaboratory usando datos de Kaggle.

De la figura 3.6 se evidenciar que no existen datos nulos en las 2 columnas.

3.3.FASE DE TRANSFORMACIÓN DE DATOS

3.3.1. Transformaciones para la regresión lineal múltiple

Ya que solo tenemos 2 columnas, tendremos que crear nuevas columnas para llevar la regresión lineal múltiple es por eso que tendremos que recurrir a los PROMEDIOS MÓVILES de los últimos 3 y 9 días. La figura 3.7 representa la creación de las nuevas columnas.

Figura 3.7: Creación de nuevas columnas de promedios móviles.

```
df_dia['m_3'] = df_dia['Price'].shift(1).rolling(window=3).mean()
df_dia['m_9'] = df_dia['Price'].shift(1).rolling(window=9).mean()
df_dia.head(11)
```

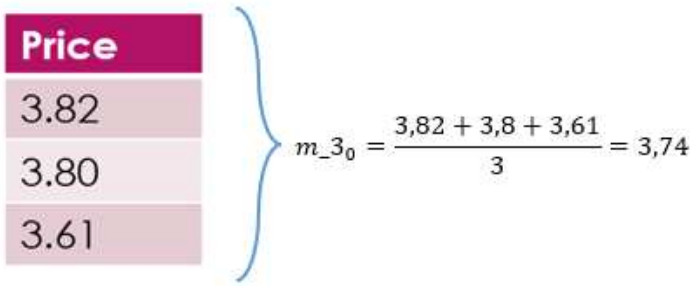
	Date	Price	m_3	m_9
0	1997-01-07	3.82	NaN	NaN
1	1997-01-08	3.80	NaN	NaN
2	1997-01-09	3.61	NaN	NaN
3	1997-01-10	3.92	3.743333	NaN
4	1997-01-13	4.00	3.776667	NaN
5	1997-01-14	4.01	3.843333	NaN
6	1997-01-15	4.34	3.976667	NaN
7	1997-01-16	4.71	4.116667	NaN
8	1997-01-17	3.91	4.353333	NaN
9	1997-01-20	3.26	4.320000	4.013333
10	1997-01-21	2.99	3.960000	3.951111

Fuente: Google Colaboratory usando datos de Kaggle.

Básicamente agarra los 3 datos anteriores y lo promedio, para la columna [m_3] agarrará los 3 datos anteriores y para la columna [m_9] agarra los 9 datos anteriores por ejemplo para verificar se pondrá un ejemplo para m_3 y m_9.

La figura 3.8 representa como se operó la columna Price para hallar m_3.

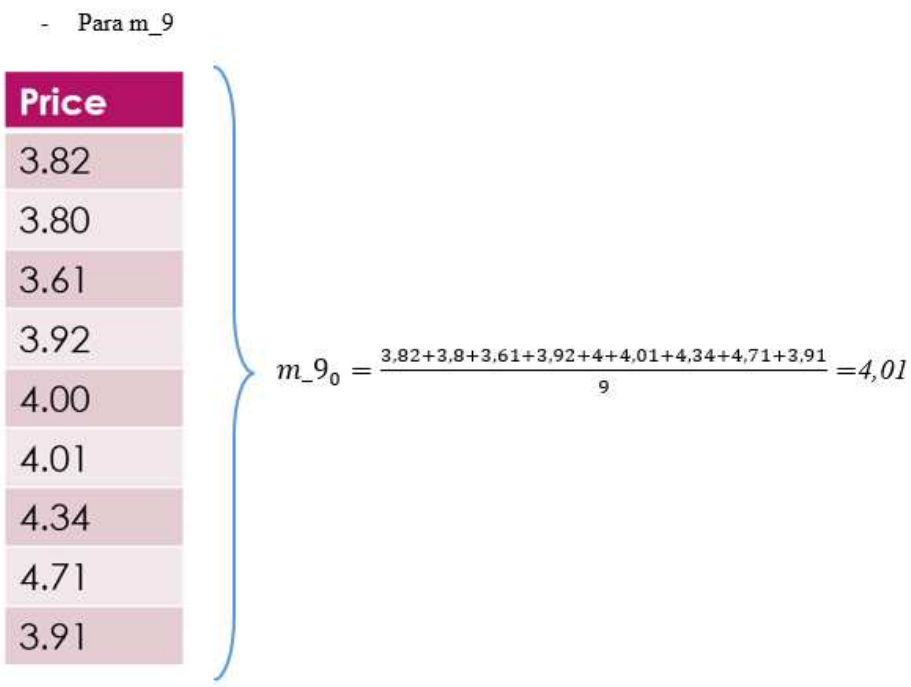
Figura 3.8: Obtención de m_3



Fuente: Elaboración propia en Power Point.

La figura 3.9 representa como se operó la columna Price para hallar m_9.

Figura 3.9: Obtención de m_9



Fuente: Elaboración propia en Power Point.

Como se puede verificar en la figura 3.7 se sacó los mismos resultados que el algoritmo, pero se ve claramente que se generaron valores nulos NAN. En la figura 3.10 se procederá a eliminar dichos valores formados.

Figura 3.10: Eliminando valores formados por m_3 y m_9.

```
df_dia = df_dia.dropna()
df_dia.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5944 entries, 9 to 5952
Data columns (total 4 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   Date    5944 non-null   object 
 1   Price   5944 non-null   float64
 2   m_3     5944 non-null   float64
 3   m_9     5944 non-null   float64
dtypes: float64(3), object(1)
memory usage: 232.2+ KB
```

Fuente: Google Colaboratory usando datos de Kaggle.

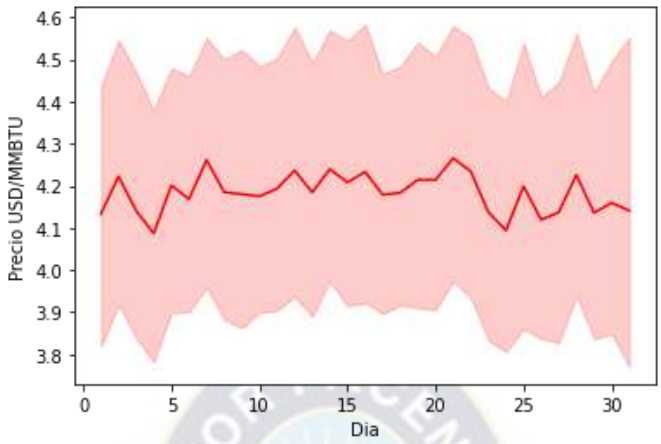
Se observa en la figura 3.10 claramente que directamente se borraron 9 filas por presentar valores NAN, y la creación de las nuevas columnas que son m_3 y m_9.

3.3.2. Transformaciones para los modelos de árboles de decisión

Para alimentar a nuestro árbol de decisiones debemos crear nuevas variables categóricas, para ello haremos un análisis de la incidencia de los días, meses y años que tanto intervienen en el modelo.

En la figura 3.11 presentaremos la incidencia de los días VS el precio del gas natural.

Figura 3.11: Análisis del precio del Gas Natural por día



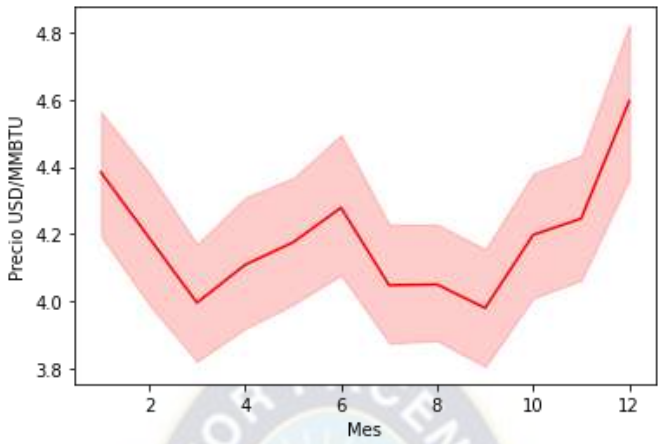
Fuente: Google Colaboratory usando datos de Kaggle.

Como se puede ver en la gráfica 3.11 se puede evidenciar que el día no toma mucha importancia en el precio del Gas Natural, razón por la cual posiblemente sea rechazado en el modelo de árbol de decisiones y bosque aleatorio.

A continuación, en la figura 3.12 se presentará la incidencia de los Meses VS el precio del gas natural.

En la figura 3.12 se puede ver claramente como esta varia de gran manera según el mes donde se puede evidenciar un incremento del precio desde los meses de octubre a diciembre que suele ser el periodo del invierno en Europa consideremos que Europa es uno de los principales consumidores de Gas Natural en el mundo.

Figura 3.12: Análisis del Precio del Gas Natural por mes.

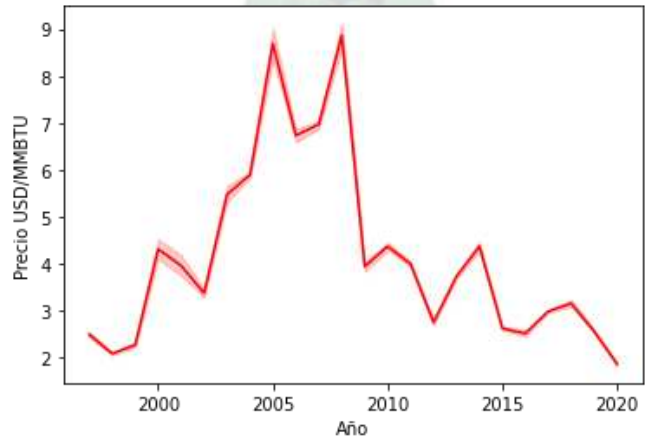


Fuente: Google Colaboratory usando datos de Kaggle.

En la figura 3.12 se puede evidenciar claramente que el precio se ve fuertemente alterado por el mes que se atraviesa, esto se puede ser un indicador de que posiblemente se convierta en un variable para los modelos.

A continuación, se presentará la figura 3.13 la cual es un análisis del año y el precio.

Figura 3.13: Análisis del Precio del Gas Natural por año.



Fuente: Google Colaboratory usando datos de Kaggle.

Como el Dataframe original solo tiene dos columnas se deberán crear columnas nuevas, que involucren el mes y el año, esto último explicado con las figuras anteriores.

En la figura 3.14 se observa que se añadieron las variables correspondientes de mes y año.

Figura 3.14: Añado Mes y Año

```
[31]  df_mess['Month'] = pd.to_datetime(df_mess['Month'])
      df_mess['Year'] = df_mess['Month'].apply(lambda x: x.year)
      df_mess['Mes'] = df_mess['Month'].apply(lambda x: x.month)
```

```
[32]  df_mess
```

	Month	Price	Year	Mes
0	1997-01-01	3.45	1997	1
1	1997-02-01	2.15	1997	2
2	1997-03-01	1.89	1997	3
3	1997-04-01	2.03	1997	4
4	1997-05-01	2.25	1997	5
...	...	...	...	...
271	2019-08-01	2.22	2019	8
272	2019-09-01	2.56	2019	9
273	2019-10-01	2.33	2019	10
274	2019-11-01	2.65	2019	11
275	2019-12-01	2.22	2019	12

276 rows × 4 columns

Fuente: Google Colaboratory usando datos de Kaggle.

3.4.FASE DE ENTRENAMIENTO

3.4.1. Selección de modelos y sus justificaciones de elección.

Los modelos seleccionados son:

- Regresión lineal múltiple.
- Árbol de decisión regresor.

- Bosque aleatorio.

Regresión lineal múltiple.

Se tomo como base el modelo de regresión lineal múltiple que se usaron para predecir el precio del petróleo el link de esta descrito en el punto 1.2 ANTECEDENTES. Cabe mencionar que además otra razón para usar el método de regresión lineal múltiple es que la variable dependiente es tipo de dato numérico continuo.

Árbol de decisión regresor.

La razón para usar este tipo de modelo es que si observamos las figuras 3.11, 3.12 y 3.13 las fechas representan números enteros como tal en el caso de los años desde los años 80 hasta el 2020 es decir, en el caso más puntual de los meses valores del 1 al 12 también enteros y de los días del 1 al 31 también números enteros y fijándose en la figura 2.4 es posiblemente que se produzca la regla de decisión derivada. Además, usamos su versión regresor ya que el PRECIO DEL GAS NATURAL retorna un continuo.

Bosque aleatorio regresor.

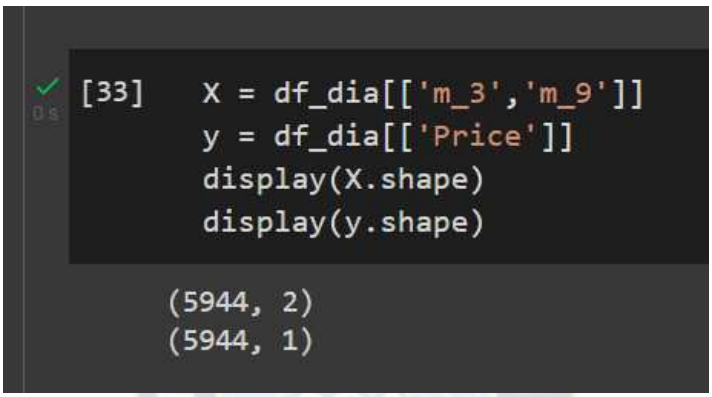
En resumen, un bosque aleatorio es una colección de árboles de decisión, como se indicó con las figuras 3.11, 3.12 y 3.13 las ramas que se generarían de cada árbol son las respectivas, valores de día, mes y año que son valores enteros lo que se puede traducir en programación como una condición y se usara la versión regresor porque la variable dependiente el precio del gas natural es un dato continuo.

Además, posee menor probabilidad de caer en un sobre ajuste.

3.4.2. Entrenamiento para la regresión lineal múltiple

Primeramente, se debe partir la data en la variable dependiente y variables independientes, lo cual será observado en la figura 3.15.

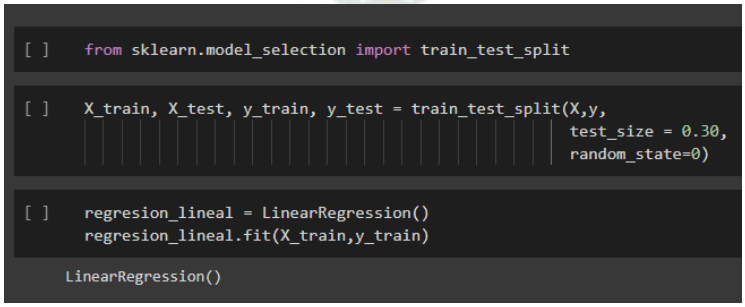
Figura 3.15: Designación de variable dependiente e independientes.



Fuente: Google Colaboratory usando datos de Kaggle.

Posterior a esto lo que se debe hacer es repartir los datos para en entrenamiento y prueba, en la figura 3.16 demarca como se repartieron los datos para el entrenamiento con el parámetro test_size que esta con un valor de 0.3 que denota que el 30% de los datos serán enviados a la prueba y el 70% restante al entrenamiento.

Figura 3.16: Entrenamiento de la regresión lineal múltiple

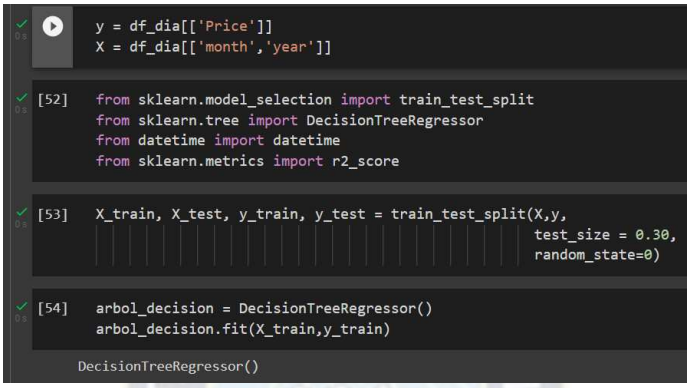


Fuente: Google Colaboratory usando datos de Kaggle.

3.4.3. Entrenamiento para el árbol de decisión

Primeramente, se debe partir la data en la variable dependiente y variables independientes, lo cual será observado en la figura 3.17.

Figura 3.17: Designación de variable dependiente e independientes en árbol de decisión

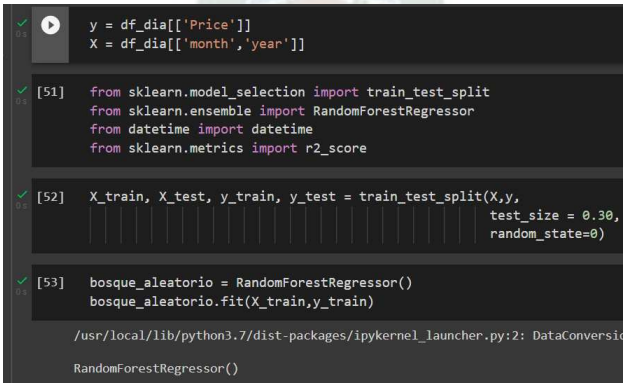


Fuente: Google Colaboratory usando datos de Kaggle.

3.4.4. Entrenamiento para el bosque aleatorio

Primeramente, se debe partir la data en la variable dependiente y variables independientes, lo cual será observado en la figura 3.18.

Figura 3.18: Designación de variable dependiente e independientes en el bosque aleatorio



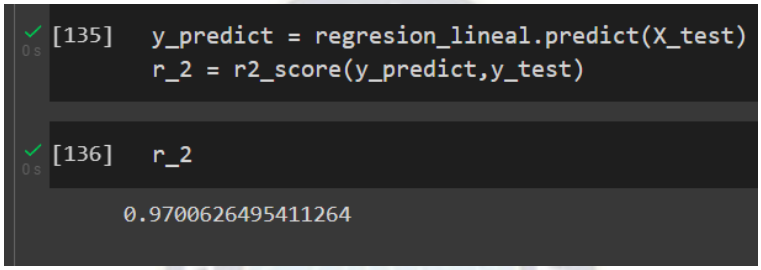
Fuente: Google Colaboratory usando datos de Kaggle.

3.5.FASE DE PRUEBA

3.5.1. Prueba para el modelo de regresión lineal múltiple

Se cargaron los datos con un 30% para la prueba, donde r_2 significa el coeficiente de determinación que se halló del modelo. Esto se verá reflejado en la figura 3.19.

Figura 3.19: Métricas de la regresión lineal múltiple



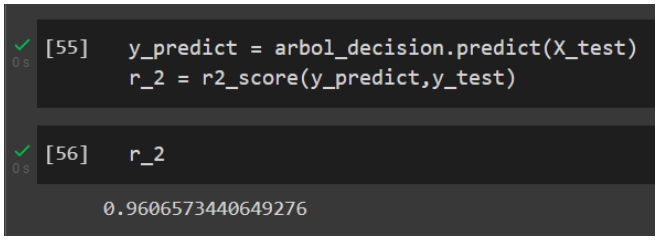
Fuente: Google Colaboratory usando datos de Kaggle.

En la parte de Anexos se colocarán los valores predichos por el modelo de regresión lineal múltiple como también la gráfica de comparación de los valores reales VS los predichos por el modelo.

3.5.2. Prueba para el modelo de árbol de decisión

En la gráfica 3.20 se puede evidenciar que el coeficiente de determinación es de 0.96 aproximadamente.

Figura 3.20: Métricas del árbol de decisión

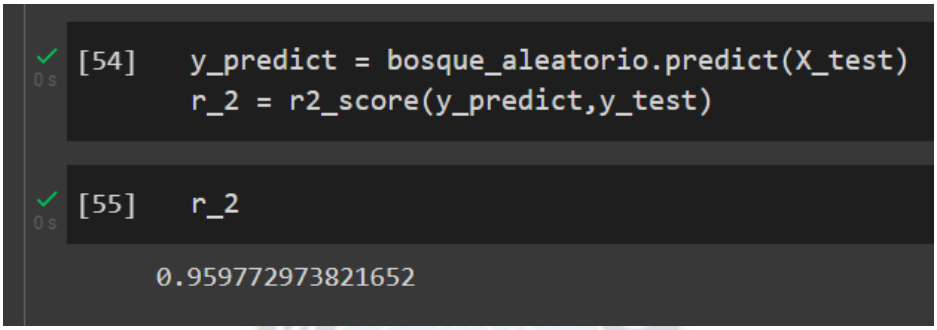


Fuente: Google Colaboratory usando datos de Kaggle.

3.5.3. Prueba para el modelo de bosque aleatorio

En la gráfica 3.21 se puede evidenciar que el coeficiente de determinación es de 0.959 aproximadamente.

Figura 3.21: Métricas de bosque aleatorio



Fuente: Google Colaboratory usando datos de Kaggle.

3.6.Memoria de cálculos

En la tabla 3.1 se tendrá una memoria de cálculos a lo largo del proceso.

Tabla 3.1: Memoria de cálculos.

ETAPA	RESULTADOS Y OBSERVACIONES
Exploración de los datos.	Se evidencia la existencia de los datos, el cual consta de 5953 filas y 2 columnas, una columna cuenta con la fecha y la otra con el precio del Gas Natural. En la figura 1.2 se puede evidenciar las características de la estadística descriptiva de los datos.
Fase de limpieza de los datos.	Se evidencia la existencia de un dato nulo en la figura 3.3, el cual se lo procede a reemplazar con el promedio de 4 días anteriores y 4 días posteriores.
Transformación	Se transformo para el modelo de regresión lineal múltiple con promedios móviles ver figura 3.7, es decir se crearon 2 nuevas columnas m_3 y m_9. También se transformó la fecha solo se agarró el mes y año de cada precio para asemejarlo a un árbol de decisión y un bosque aleatorio.
Entrenamiento	Los tres modelos se los alimento con un 70% de los datos para un entrenamiento.

Prueba	Modelo de regresión lineal. • Se crearon las variables independientes de m_3 y m_9 cuya explicación está en las figuras 3.11, 3.12 y 3.13 estas mismas son los promedios móviles. • Se obtuvo un R^2 de 0.97. • La ecuación del modelo es: $$Y = 0.038 + 0.9733X_{m\_3} + 0.01712X_{m\_9}$$ Modelo de árbol de decisión. • Se crearon las variables independientes de mes y año que son tipo de dato categóricos esto esta explicado en las figuras 3.12 y 3.13 debido a que estas variables no representan valores continuos. Debido a la variabilidad de estas variables con el precio son tomadas como variables independientes. • Se obtuvo un R^2 de 0.96. • El árbol de decisión no forma ecuación como su nombre indica forma un árbol, la imagen de este mismo está en anexo (Figura del Árbol de decisión)
	Modelo de bosque aleatorio. • Debido al comportamiento de las figuras 3.12 y 3.13 se puede evidenciar que los valores de mes y año son bastante dependientes del precio del Gas Natural como tal, es por eso que se toma al mes y año como variable independiente. • Se obtuvo un R^2 de 0.95. • En el caso de un bosque aleatorio resulta más complicada su visualización ya que como su mismo nombre indica se formará un conjunto de árboles aleatorios para luego sacar un promedio, su visualización. Revisar anexos para ver ejemplos de los árboles formados (Figura de Bosque Aleatorio)

Fuente: Elaboración propia con los datos de Kaggle.

4. CAPÍTULO IV: RESULTADOS

En la tabla 4.1 se tendrá un resumen de los modelos donde el accuracy se refiere al coeficiente de determinación este coeficiente mientras más cercano este a la unidad representa menor error en la predicción.

Tabla 4.1: Resumen de resultados

MODELO	TRAIN-TEST	VARIABLES	ACCURACY	ERROR PROMEDIO
Regresión lineal múltiple	70-30	Son m_3 y m_9 que son los promedios anteriores a 3 días y 9 días, estas variables fueron obtenidas y transformando la base de datos original, se tomó este modelo ya que otros autores usan este mismo modelo el link para dirigirse a la aplicación de otros autores para la predicción del precio del petróleo se encuentra en el punto 1.2 de justificación.	0.97	7.78%
Árbol de decisión	70-30	Observando las figuras 3.12 y 3.13 se puede evidenciar que el precio del gas natural depende de gran medida de la estación del año y al ser precios internacionales se ven fuertemente	0.96	5.33%

		afectados, por lo que se supone que esta subida de precios en los puede ser explicada por los meses de invierno en Europa. Razón por la cual se crearon nuevas variables de mes y año a partir de la base de datos original.		
Bosque Aleatorio	70-30	Viendo la dependencia del precio en las figuras 3.12 y 3.13 se transformó la columna fecha en la creación de 2 variables independientes que son el mes y año.	0.95	5.07%

Fuente: Elaboración propia en Excel usando datos de Kaggle, 2021.

En la tabla 4.1 se puede observar los errores obtenidos por cada modelo que es el promedio de los valores predichos y los reales. Como por ejemplo el error para cada fecha en las tablas 4.2, 4.3 y 4.4.

4.1.OBSERVACIONES DE LOS MODELOS

4.1.1. Observaciones del modelo de regresión lineal múltiple

Se creó puntos promedios con datos anteriores de 3 y 9 días (m_3 y m_9), se preferiría quedarse con el modelo que se entrenó con 70% y se lo evaluó con el 30%, ya que la diferencia entre sus exactitudes corresponde a un valor relativamente bajo, además así tenemos menos probabilidad de haber caído en un caso de overfitting.

Cabe recalcar que lo negativo del modelo es que solo nos daría rango de predecir el precio del Gas Natural de solo 1 día, lo cual no da mucho rango de predicción. Ya que se debe alimentar de datos anteriores para predecir un nuevo valor a un día.

4.1.2. Observaciones del modelo de árbol de decisiones

Se toma la decisión de eliminar los días del modelo esta decisión también parte de la figura 3.11 donde esta graficado los días Vs el precio, se puede observar que no varía mucho el precio con relación al día.

Pero en cambio la figura 3.12 claramente se puede ver que si inciden los meses con el precio del Gas Natural como también lo hacen los años.

Se prefiere que trabaje con un train de 70% y test de 30% para evitar over-fitting.

Cabe recalcar con lo anterior explicado el modelo es capaz de predecir precios del Gas Natural con un rango de meses.

4.1.3. Observaciones del modelo de bosque aleatorio

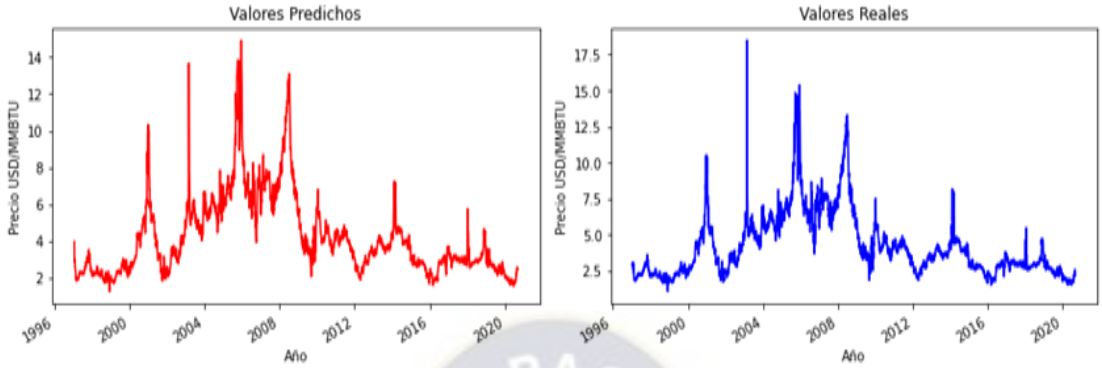
Si bien se obtiene un menor resultado que trabajando con el árbol de decisión, por conceptos técnicas un bosque aleatorio tiene menor probabilidad de caer en overfitting.

4.2.COMPARATIVA DE LOS MODELOS

4.2.1. Comparativa modelo de regresión lineal

En la figura 4.1 se presentarán los datos predichos por el modelo de regresión lineal múltiple.

Figura 4.1: Comparativa regresión lineal múltiple.



Fuente: Elaboración propia usando datos de Kaggle.

En la tabla 4.2 se presentará los valores predichos y los reales en una fecha específica, debido a el tamaño que datos (30%) que fueron puestos a prueba que son 1784, razón por la cual se procederá a tomar 100 datos de muestra para la correspondiente tabla 4.2.

Tabla 4.2: Comparativa regresión lineal múltiple con fechas.

Date	Predichos	Price	m_3	m_9	Error
24/1/1997	3,02174	2,62	3	3,692222	15,33%
13/2/1997	2,411006	2,22	2,393333	2,508889	8,60%
6/6/1997	2,179049	2,19	2,16	2,225556	0,50%
1/8/1997	2,233861	2,24	2,216667	2,205556	0,27%
13/8/1997	2,502867	2,45	2,49	2,378889	2,16%
21/1/1998	2,118158	2,09	2,1	2,08	1,35%
12/10/1998	1,97865	1,75	1,956667	2,08	13,07%
15/12/1998	1,666937	1,86	1,646667	1,496667	10,38%
15/3/1999	1,892671	1,75	1,873333	1,795556	8,15%
7/4/1999	2,004749	2,03	1,986667	1,898889	1,24%
13/4/1999	2,094325	2,14	2,076667	2,014444	2,13%
23/8/1999	2,873104	2,95	2,863333	2,78	2,61%
15/10/1999	2,729658	2,66	2,72	2,55	2,62%
19/1/2000	2,322011	2,4	2,306667	2,237778	3,25%
25/7/2000	3,832013	3,63	3,826667	4,023333	5,57%
11/8/2000	4,448262	4,44	4,456667	4,202222	0,19%
18/8/2000	4,276167	4,38	4,276667	4,383333	2,37%
6/9/2000	4,747922	4,89	4,756667	4,65	2,91%
13/12/2000	8,828639	7,8	8,89	8,02	13,19%

16/2/2001	5,5954	5,57	5,606667	5,827778	0,46%
9/8/2001	3,122994	3,1	3,113333	3,163333	0,74%
16/10/2001	2,330545	2,51	2,316667	2,167778	7,15%
10/1/2002	2,442707	2,32	2,426667	2,465556	5,29%
14/1/2002	2,335339	2,32	2,316667	2,447778	0,66%
9/5/2002	3,620389	3,72	3,616667	3,601111	2,68%
22/10/2002	4,144139	4,2	4,146667	4,062222	1,33%
4/12/2002	4,254741	4,24	4,256667	4,268889	0,35%
3/2/2003	5,636778	5,71	5,65	5,781111	1,28%
14/2/2003	6,044939	5,88	6,063333	6,123333	2,81%
11/4/2003	5,153067	5,28	5,166667	5,005556	2,40%
17/4/2003	5,461861	5,54	5,48	5,228889	1,41%
22/8/2003	5,04331	5,24	5,053333	5,037778	3,75%
15/10/2003	4,894508	4,93	4,906667	4,684444	0,72%
14/11/2003	4,621345	4,62	4,63	4,457778	0,03%
11/5/2004	6,156438	6,24	6,18	6,003333	1,34%
8/3/2005	6,599363	6,81	6,626667	6,481111	3,09%
24/5/2005	6,33949	6,45	6,36	6,462222	1,71%
13/6/2005	7,085542	7,08	7,12	6,832222	0,08%
14/9/2005	10,75036	10,8	10,80667	11,30333	0,46%
8/12/2005	13,81035	14,25	13,93	12,47222	3,09%
1/2/2006	8,383434	8,71	8,426667	8,356667	3,75%
24/2/2006	7,360159	7,39	7,393333	7,333333	0,40%
5/4/2006	7,006419	6,89	7,033333	7,137778	1,69%
19/4/2006	7,120515	7,72	7,153333	6,98	7,77%
20/6/2006	6,68898	6,62	6,723333	6,22	1,04%
7/9/2006	5,462721	5,64	5,463333	6,226667	3,14%
7/3/2007	7,345079	7,5	7,376667	7,4	2,07%
3/4/2007	7,449632	7,57	7,486667	7,253333	1,59%
25/5/2007	7,524916	7,47	7,556667	7,671111	0,74%
2/7/2007	6,62386	6,24	6,643333	6,964444	6,15%
9/8/2007	6,218527	6,45	6,24	6,218889	3,59%
3/10/2007	6,2353	6,96	6,256667	6,251111	10,41%
5/5/2008	10,55114	10,77	10,61333	10,65778	2,03%
6/6/2008	12,22356	12,71	12,31	11,88667	3,83%
9/6/2008	12,3685	12,71	12,45667	12,01444	2,69%
2/1/2009	5,697956	5,41	5,716667	5,564444	5,32%
13/4/2009	3,568746	3,46	3,563333	3,616667	3,14%
10/2/2010	5,608703	5,48	5,626667	5,467778	2,35%
31/3/2010	3,850833	3,93	3,846667	3,985556	2,01%
8/10/2010	3,569944	3,36	3,563333	3,686667	6,25%

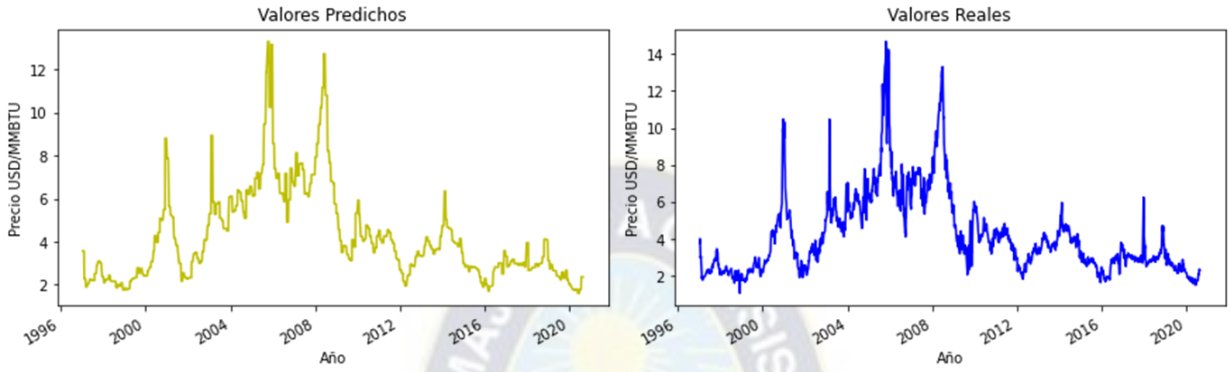
22/3/2011	3,939735	4,05	3,94	3,872222	2,72%
2/11/2011	3,59766	3,39	3,593333	3,6	6,13%
11/11/2011	3,488256	3,29	3,483333	3,463333	6,03%
22/5/2012	2,598342	2,55	2,586667	2,46	1,90%
11/7/2012	2,877128	2,72	2,866667	2,825556	5,78%
25/9/2012	2,792945	2,84	2,78	2,835556	1,66%
23/10/2012	3,404883	3,34	3,4	3,331111	1,94%
28/12/2012	3,326333	3,4	3,32	3,291111	2,17%
22/2/2013	3,293471	3,27	3,286667	3,266667	0,72%
17/6/2013	3,747765	3,78	3,743333	3,84	0,85%
5/11/2013	3,47386	3,37	3,466667	3,57	3,08%
6/3/2014	7,125079	4,89	7,176667	5,92	45,71%
6/5/2014	4,743601	4,8	4,75	4,776667	1,17%
5/6/2014	4,57069	4,66	4,576667	4,531111	1,92%
5/8/2014	3,809009	3,87	3,806667	3,816667	1,58%
18/2/2015	2,797608	2,92	2,786667	2,728889	4,19%
6/4/2015	2,648215	2,63	2,633333	2,72	0,69%
22/6/2015	2,900523	2,79	2,89	2,865556	3,96%
25/6/2015	2,829223	2,8	2,816667	2,87	1,04%
2/2/2016	2,253612	2,07	2,236667	2,222222	8,87%
10/2/2016	2,177831	2,16	2,16	2,154444	0,83%
28/4/2016	1,959252	1,88	1,94	1,894444	4,22%
12/9/2016	2,871857	3,04	2,86	2,896667	5,53%
19/10/2016	3,220154	3,18	3,213333	3,153333	1,26%
26/1/2017	3,202647	3,44	3,193333	3,267778	6,90%
11/4/2017	3,210193	3,11	3,203333	3,14	3,22%
27/7/2017	2,977978	2,95	2,966667	3,031111	0,95%
31/7/2017	2,955134	2,87	2,943333	3,023333	2,97%
11/1/2018	3,021395	3,16	2,993333	4,051111	4,39%
23/1/2018	3,426735	3,35	3,416667	3,66	2,29%
14/5/2018	2,772129	2,84	2,76	2,756667	2,39%
15/5/2018	2,801499	2,83	2,79	2,766667	1,01%
17/5/2018	2,850469	2,75	2,84	2,784444	3,65%
18/7/2018	2,822734	2,8	2,81	2,87	0,81%
18/1/2019	3,564305	3,43	3,566667	3,167778	3,92%
5/3/2019	3,441315	3,18	3,443333	2,995556	8,22%
24/5/2019	2,686329	2,6	2,673333	2,672222	3,32%
1/4/2020	1,709693	1,69	1,686667	1,72	1,17%
11/5/2020	1,847022	1,7	1,826667	1,782222	8,65%
9/6/2020	1,74206	1,68	1,72	1,715556	3,69%

Fuente: Elaboración propia con datos de Kaggle.

4.2.2. Comparativa modelo de árbol de decisión regresor.

En la figura 4.2 se presentarán los datos predichos por el modelo de árbol de decisión regresor.

Figura 4.2: Comparativa árbol de decisión regresor.



Fuente: Elaboración propia usando datos de Kaggle.

En la tabla 4.3 se presentará los valores predichos y los reales en una fecha específica, debido a el tamaño que datos (30%) que fueron puestos a prueba que son 1784, razón por la cual se procederá a tomar 100 datos de muestra para la correspondiente tabla 4.3 que también mostrará las variables de meses y años.

Tabla 4.3: Comparativa árbol de decisión regresor con fechas.

Date	Predichos	Price	year	month	day	Error
20/1/1997	3,5625	3,26	1997	1	20	9,28%
24/6/1997	2,197222	2,32	1997	6	24	5,29%
21/11/1997	3,004545	2,59	1997	11	21	16,01%
24/11/1997	3,004545	2,63	1997	11	24	14,24%
27/1/1998	2,092	2,06	1998	1	27	1,55%
24/2/1999	1,76125	1,73	1999	2	24	1,81%
26/3/1999	1,809	1,83	1999	3	26	1,15%
24/8/1999	2,771875	3,01	1999	8	24	7,91%
31/8/1999	2,771875	2,84	1999	8	31	2,40%
16/2/2000	2,639375	2,65	2000	2	16	0,40%
27/3/2000	2,793889	2,82	2000	3	27	0,93%

16/5/2000	3,563529	3,45	2000	5	16	3,29%
22/6/2000	4,26	4,44	2000	6	22	4,05%
3/8/2000	4,424667	4,23	2000	8	3	4,60%
5/9/2000	5,078235	4,81	2000	9	5	5,58%
22/1/2001	7,864615	7,7	2001	1	22	2,14%
22/2/2001	5,625	5,11	2001	2	22	10,08%
23/4/2001	5,14	5,09	2001	4	23	0,98%
25/6/2001	3,830769	3,56	2001	6	25	7,61%
21/6/2002	3,224545	3,18	2002	6	21	1,40%
31/7/2002	2,991333	3,02	2002	7	31	0,95%
3/2/2003	8,931111	5,71	2003	2	3	56,41%
4/3/2003	5,837143	7,71	2003	3	4	24,29%
3/7/2003	5,085625	4,96	2003	7	3	2,53%
28/10/2003	4,602	4,45	2003	10	28	3,42%
18/12/2003	6,07375	6,98	2003	12	18	12,98%
19/12/2003	6,07375	6,92	2003	12	19	12,23%
26/4/2004	5,668333	5,6	2004	4	26	1,22%
17/9/2004	5,085385	4,95	2004	9	17	2,74%
17/11/2004	6,198125	6,06	2004	11	17	2,28%
10/12/2004	6,561818	6,29	2004	12	10	4,32%
28/2/2005	6,136429	6,62	2005	2	28	7,30%
10/3/2005	6,947778	6,91	2005	3	10	0,55%
19/4/2005	7,205833	6,95	2005	4	19	3,68%
25/4/2005	7,205833	7,27	2005	4	25	0,88%
14/7/2005	7,582857	7,99	2005	7	14	5,10%
17/7/2006	5,861111	6,27	2006	7	17	6,52%
18/7/2006	5,861111	6,02	2006	7	18	2,64%
12/12/2006	6,760625	6,93	2006	12	12	2,44%
5/7/2007	6,219231	6,3	2007	7	5	1,28%
26/7/2007	6,219231	5,83	2007	7	26	6,68%
8/8/2007	6,259167	6,24	2007	8	8	0,31%
14/8/2007	6,259167	6,86	2007	8	14	8,76%
12/9/2007	6,095	6,13	2007	9	12	0,57%
24/3/2008	9,433077	8,99	2008	3	24	4,93%
12/11/2008	6,714286	6,65	2008	11	12	0,97%
14/1/2009	5,304615	5,47	2009	1	14	3,02%
12/5/2009	3,8075	4,41	2009	5	12	13,66%
23/7/2009	3,393125	3,66	2009	7	23	7,29%
11/9/2009	3,115455	2,94	2009	9	11	5,97%
16/12/2009	5,322	5,57	2009	12	16	4,45%
4/11/2010	3,703571	3,53	2010	11	4	4,92%

2/12/2010	4,243125	4,28	2010	12	2	0,86%
20/12/2010	4,243125	4,1	2010	12	20	3,49%
8/2/2011	4,083571	4,24	2011	2	8	3,69%
14/3/2011	3,941176	3,9	2011	3	14	1,06%
19/8/2011	4,07	3,99	2011	8	19	2,01%
10/1/2012	2,684615	2,97	2012	1	10	9,61%
16/5/2012	2,39375	2,5	2012	5	16	4,25%
16/8/2012	2,852353	2,78	2012	8	16	2,60%
14/12/2012	3,354615	3,15	2012	12	14	6,50%
18/1/2013	3,32	3,54	2013	1	18	6,21%
8/2/2013	3,344667	3,26	2013	2	8	2,60%
4/4/2013	4,201667	3,94	2013	4	4	6,64%
8/8/2013	3,429333	3,27	2013	8	8	4,87%
12/9/2013	3,617143	3,57	2013	9	12	1,32%
29/10/2013	3,664118	3,57	2013	10	29	2,64%
30/4/2014	4,664667	4,79	2014	4	30	2,62%
1/7/2014	4,014	4,47	2014	7	1	10,20%
29/9/2014	3,911875	3,89	2014	9	29	0,56%
14/10/2014	3,771429	3,91	2014	10	14	3,54%
10/12/2014	3,485	3,65	2014	12	10	4,52%
27/2/2015	2,819091	2,79	2015	2	27	1,04%
21/7/2015	2,819333	2,84	2015	7	21	0,73%
3/9/2015	2,68	2,67	2015	9	3	0,37%
11/3/2016	1,699286	1,74	2016	3	11	2,34%
17/3/2016	1,699286	1,84	2016	3	17	7,65%
27/4/2016	1,918	1,88	2016	4	27	2,02%
29/7/2016	2,811538	2,97	2016	7	29	5,34%
2/9/2016	2,995455	2,88	2016	9	2	4,01%
9/12/2016	3,57	3,75	2016	12	9	4,80%
26/1/2017	3,315333	3,44	2017	1	26	3,62%
6/2/2017	2,828125	2,92	2017	2	6	3,15%
17/4/2017	3,091111	3,09	2017	4	17	0,04%
21/11/2017	3,019375	3,05	2017	11	21	1,00%
12/1/2018	3,939231	4,06	2018	1	12	2,97%
7/2/2018	2,655385	2,74	2018	2	7	3,09%
20/6/2018	2,963571	2,96	2018	6	20	0,12%
29/6/2018	2,963571	2,96	2018	6	29	0,12%
19/7/2018	2,841765	2,75	2018	7	19	3,34%
13/9/2018	2,993333	2,94	2018	9	13	1,81%
5/10/2018	3,288125	3,25	2018	10	5	1,17%
4/12/2018	4,087	4,4	2018	12	4	7,11%

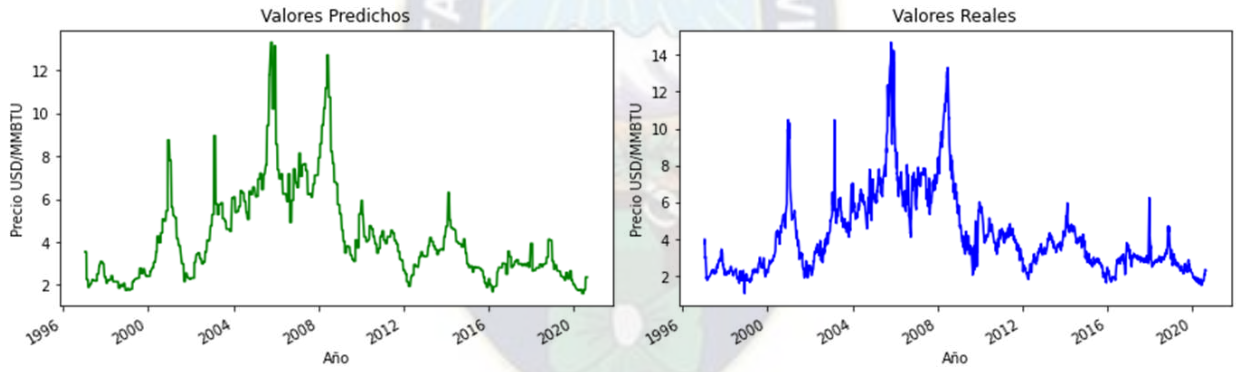
14/12/2018	4,087	3,99	2018	12	14	2,43%
9/1/2019	3,114118	2,92	2019	1	9	6,65%
17/4/2019	2,648333	2,55	2019	4	17	3,86%
11/7/2019	2,353529	2,54	2019	7	11	7,34%
6/12/2019	2,211176	2,31	2019	12	6	4,28%
3/6/2020	1,594167	1,84	2020	6	3	13,36%
4/8/2020	2,351875	2,07	2020	8	4	13,62%

Fuente: Elaboración propia con datos de Kaggle.

4.2.3. Comparativa modelo de bosque aleatorio regresor.

En la figura 4.3 se presentarán los datos predichos por el modelo de árbol de decisión regresor.

Figura 4.3: Comparativa bosque aleatorio regresor.



Fuente: Elaboración propia usando datos de Kaggle.

En la tabla 4.4 se presentará los valores predichos y los reales en una fecha específica, debido a el tamaño que datos (30%) que fueron puestos a prueba que son 1784, razón por la cual se procederá a tomar 100 datos de muestra para la correspondiente tabla 4.4 que también mostrará las variables de meses y años.

Tabla 4.4: Comparativa bosque aleatorio regresor con fechas.

Date	Predichos	Price	year	month	day	Error
28/1/1997	3,54615	3,05	1997	1	28	16,27%
9/4/1997	2,037815	1,96	1997	4	9	3,97%

18/6/1997	2,197454	2,22	1997	6	18	1,02%
17/7/1998	2,150436	2,15	1998	7	17	0,02%
18/9/1998	2,00929	2,27	1998	9	18	11,49%
17/11/1998	2,077492	2,12	1998	11	17	2,01%
23/2/1999	1,764269	1,75	1999	2	23	0,82%
5/4/1999	2,151922	2,03	1999	4	5	6,01%
19/4/1999	2,151922	2,1	1999	4	19	2,47%
19/8/1999	2,772735	2,87	1999	8	19	3,39%
29/8/2000	4,426359	4,6	2000	8	29	3,77%
31/8/2000	4,426359	4,76	2000	8	31	7,01%
8/1/2001	7,806218	10,31	2001	1	8	24,28%
14/2/2001	5,618641	5,89	2001	2	14	4,61%
10/4/2001	5,141477	5,55	2001	4	10	7,36%
27/7/2001	3,111273	3,06	2001	7	27	1,68%
29/8/2001	2,988301	2,46	2001	8	29	21,48%
12/12/2001	2,271011	2,57	2001	12	12	11,63%
28/12/2001	2,271011	2,4	2001	12	28	5,37%
18/1/2002	2,326179	2,28	2002	1	18	2,03%
26/7/2002	2,988392	2,94	2002	7	26	1,65%
29/7/2002	2,988392	3,07	2002	7	29	2,66%
6/9/2002	3,548513	3,39	2002	9	6	4,68%
24/9/2002	3,548513	4	2002	9	24	11,29%
24/1/2003	5,287015	5,91	2003	1	24	10,54%
6/2/2003	8,9592	6,08	2003	2	6	47,36%
30/5/2003	5,74963	5,99	2003	5	30	4,01%
30/6/2003	5,827363	5,31	2003	6	30	9,74%
10/11/2003	4,502517	4,42	2003	11	10	1,87%
24/2/2004	5,386594	5,08	2004	2	24	6,04%
15/4/2004	5,664788	5,68	2004	4	15	0,27%
16/6/2004	6,291901	6,38	2004	6	16	1,38%
29/4/2005	7,203186	6,64	2005	4	29	8,48%
8/8/2005	9,435399	8,93	2005	8	8	5,66%
16/8/2005	9,435399	9,66	2005	8	16	2,33%
19/10/2005	13,31633	13,52	2005	10	19	1,51%
26/10/2005	13,31633	14,68	2005	10	26	9,29%
5/1/2006	8,56461	9,24	2006	1	5	7,31%
18/1/2006	8,56461	8,86	2006	1	18	3,33%
13/6/2006	6,250115	5,95	2006	6	13	5,04%
6/3/2007	7,073052	7,55	2007	3	6	6,32%
9/3/2007	7,073052	7,05	2007	3	9	0,33%
27/8/2007	6,26299	5,34	2007	8	27	17,28%

2/11/2007	7,114685	6,63	2007	11	2	7,31%
12/11/2007	7,114685	6,83	2007	11	12	4,17%
27/11/2007	7,114685	7,42	2007	11	27	4,11%
14/12/2007	7,116005	7,09	2007	12	14	0,37%
11/1/2008	7,937281	8,13	2008	1	11	2,37%
24/1/2008	7,937281	7,85	2008	1	24	1,11%
9/5/2008	11,19749	11,29	2008	5	9	0,82%
21/7/2008	10,7594	10,58	2008	7	21	1,70%
12/8/2008	8,233855	8,23	2008	8	12	0,05%
1/4/2009	3,497594	3,56	2009	4	1	1,75%
17/4/2009	3,497594	3,47	2009	4	17	0,80%
18/6/2009	3,784193	4,19	2009	6	18	9,69%
8/12/2009	5,309161	5,1	2009	12	8	4,10%
27/4/2010	4,019928	4,18	2010	4	27	3,83%
10/6/2010	4,758486	4,68	2010	6	10	1,68%
1/2/2011	4,085059	4,42	2011	2	1	7,58%
12/5/2011	4,241144	4,1	2011	5	12	3,44%
5/10/2011	3,553265	3,63	2011	10	5	2,11%
10/11/2011	3,203076	3,48	2011	11	10	7,96%
20/12/2011	3,15415	3,06	2011	12	20	3,08%
9/2/2012	2,520819	2,5	2012	2	9	0,83%
16/5/2012	2,391814	2,5	2012	5	16	4,33%
3/7/2012	2,944311	2,78	2012	7	3	5,91%
3/10/2012	3,311591	3,21	2012	10	3	3,16%
2/5/2013	4,020417	4,28	2013	5	2	6,07%
28/8/2013	3,42737	3,54	2013	8	28	3,18%
15/5/2014	4,594106	4,42	2014	5	15	3,94%
10/6/2014	4,570169	4,67	2014	6	10	2,14%
31/10/2014	3,775527	3,82	2014	10	31	1,16%
25/2/2015	2,818231	3,21	2015	2	25	12,20%
2/3/2015	2,872783	2,79	2015	3	2	2,97%
15/5/2015	2,842952	2,96	2015	5	15	3,95%
3/6/2015	2,80009	2,65	2015	6	3	5,66%
17/7/2015	2,817818	2,88	2015	7	17	2,16%
10/8/2015	2,766422	2,85	2015	8	10	2,93%
29/9/2015	2,680227	2,57	2015	9	29	4,29%
21/12/2015	1,897571	1,76	2015	12	21	7,82%
8/1/2016	2,28081	2,47	2016	1	8	7,66%
25/3/2016	1,695706	1,78	2016	3	25	4,74%
27/4/2016	1,91848	1,88	2016	4	27	2,05%
26/5/2016	1,951264	1,77	2016	5	26	10,24%

15/7/2016	2,813576	2,7	2016	7	15	4,21%
14/9/2016	3,00475	3,07	2016	9	14	2,13%
27/6/2017	2,969312	2,99	2017	6	27	0,69%
8/2/2018	2,652525	2,74	2018	2	8	3,19%
11/4/2018	2,796304	2,74	2018	4	11	2,05%
25/6/2018	2,961207	2,96	2018	6	25	0,04%
23/7/2018	2,842205	2,79	2018	7	23	1,87%
14/8/2018	2,957153	3,02	2018	8	14	2,08%
6/2/2019	2,73045	2,58	2019	2	6	5,83%
25/6/2019	2,403812	2,31	2019	6	25	4,06%
30/10/2019	2,295071	2,71	2019	10	30	15,31%
3/3/2020	1,785642	1,78	2020	3	3	0,32%
24/3/2020	1,785642	1,73	2020	3	24	3,22%
28/5/2020	1,78889	1,79	2020	5	28	0,06%
11/6/2020	1,600581	1,77	2020	6	11	9,57%
24/6/2020	1,600581	1,64	2020	6	24	2,40%

Fuente: Elaboración propia con datos de Kaggle.

4.3.DECISIÓN

Si bien los 3 modelos obtuvieron la métrica del coeficiente de determinación mayor al 70%, se prefiere usar al bosque aleatorio regresor con un coeficiente de determinación mayor al 95% el cual tomo como variables independientes al mes y año si nos fijamos en la figura 3.12 donde está el mes contra el precio se puede evidenciar una subida de los precios en los meses de invierno en Europa que aproximadamente empieza en octubre hasta principios de marzo y mediante una transformación de la fecha que convertía los meses en números enteros del 1 al 12 nos brindaría una pista de usar este modelo de bosque aleatorio.

Debido al tamaño de la data es posible caer en un sobre ajuste y una de las bondades del bosque aleatorio es que tiene menor probabilidad de caer en un sobre ajuste que los otros modelos ya que su respuesta depende del promedio de las masas aleatorias.

4.3.1. Discusión de resultados.

En la tabla 4.5 se pondrá la discusión correspondiente de los resultados obtenidos por el modelo.

Tabla 4.5: Discusión de resultados.

MODELO	ACCURACY	ERROR PROMEDIO	OBSERVACIONES
Regresión lineal múltiple	0.97	7.78%	Se escogió este modelo debido a que se tomo como referencia el segundo link en la parte de 1.2. Antecedentes que es un trabajo que intentan predecir el precio internacional del petróleo.
Árbol de decisión regresor	0.96	5.33%	Se lo escogió debió a que a que la variable a determinar el precio del gas natural es un tipo de numero continuo. Además, que se puede ver en las figuras de 3.12 y 3.13 el precio cambia bastante según sea el mes y año.
Bosque Aleatorio regresor	0.95	5.07%	Este modelo es un tipo de árbol de decisiones pero que toma varios arboles aleatorios esto ultimo significa que se reduce la probabilidad de llegar a un sobre ajuste.

Fuente: Elaboración propia en Excel con los datos de Kaggle.

Regresión lineal múltiple.

Si bien presenta un accuracy de 0.97 que es bastante bueno, pero al ser alimentado por las variables m_3 y m_9 que son el promedio de los 3 días anteriores y 9 días anteriores resulta que este modelo no es muy aplicable, esto debió a que si se pasa de este rango de 3 y 9 días no podrá determinar un precio. Ejemplo si lo queremos determinar para 10 días en adelante no podrá ser posible debido a que se alimentan de la variable m_3 y m_9.

Árbol de decisión regresor.

El resultado obtenido es bastante bueno como un accuracy de 0.96 y un error de 5.33%, ya que este modelo se lo uso por la dependencia del precio según el precio del año y mes. Lo negativo de este modelo es que en base de datos grandes como es nuestro caso tiende al sobreajuste.

Bosque aleatorio regresor.

Sus medidas fueron 0.95 de el accuracy y 5.07% en el error, si el anterior modelo tiende al sobre ajuste la ventaja que muestra el bosque aleatorio regresor es que genera por defecto 100 arboles aleatorios y luego saca un promedio para cada precio, razón por la cual se escogió este ultimo modelo.

4.4.ANÁLISIS PARA BOLIVIA.

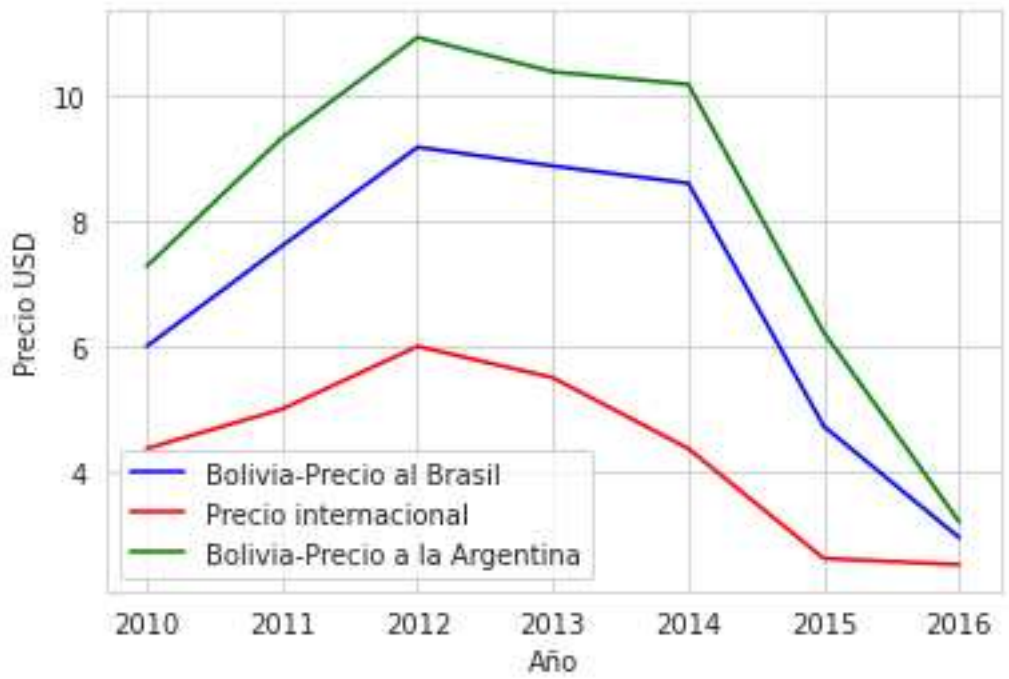
Si bien el desarrollo del presente proyecto de grado se usaron datos de precios internacionales del gas natural de Kaggle esto debido a que en Bolivia se observó que los datos de los precios no son tan cuantiosos en cantidad de datos y se los almacenan de manera trimestral y/o mensual causando una disminución significativa de los datos.

Se podría predecir el precio del gas natural en Bolivia mediante el bosque aleatorio regresor por las siguientes razones:

- En épocas como el inverno el país de la Argentina demanda más gas, tal como se observó en la figura 3.12 el precio de incrementa en esas épocas.
- El gas natural es un commodity y su precio se ve afectado por la oferta y la demanda internacional.
- La quinta adenda suscrita con la Argentina la cual toma el índice de Henry Hub el cual esta explicado en la ecuación 6, esta indica que el precio de venta gas boliviano ya se basa en un parámetro internacional ya que Henry Hub es usado en el mercado Nortea americano.

En la figura 4.4 se compara el precio del gas boliviano y el internacional para apreciar su semejanza, los datos que se pudieron recabar para Bolivia son a partir del año 2010 hasta el 2016 así que se compara con dichos años.

Figura 4.4: Análisis en Bolivia y el precio internacional 2010-2016.



Fuente: Elaboración propia con datos de Kaggle y HidrocarburosBolivia.com.

En la figura 4.4 se evidencia que existe cierto grado de correlación con los precios internacionales que con los precios del gas natural en este caso de exportación al Brasil y a la Argentina.

4.5.PREDICCIÓN DEL PRECIO FUTURO

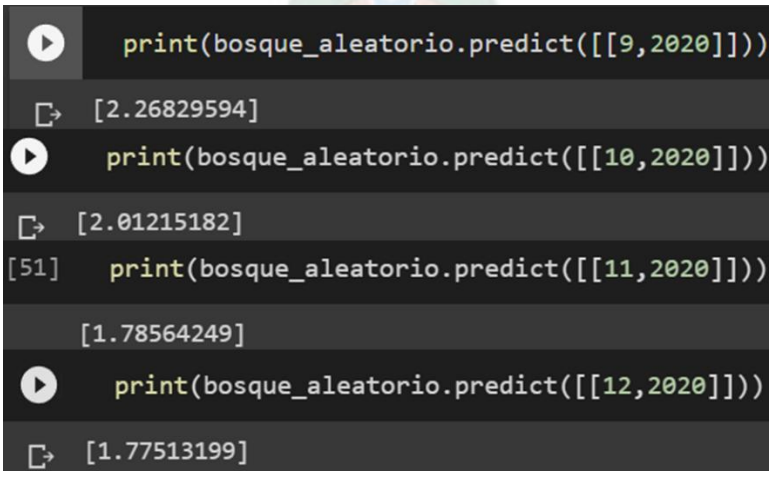
Cabe recordar que la predicción que realizo con anterioridad al momento de entrenar los modelos, como se evidencia en la figura 3.16 que el comando ‘test_size=0.3’ indica que se tomara un 70% de la base de datos para entrenar el modelo y el 30% restante para su correspondiente evaluación de esta misma se calcular parámetros como es el error y el coeficiente de determinación, en este último una parte del 30% de los datos de evaluación se pueden reflejar en las tablas 4.2, 4.3 y 4.4.

En puntos anteriores se escogió el modelo de bosque aleatorio regresor como el modelo escogido, cabe mencionar que tanto los modelos de árboles de decisión y bosque aleatorio presentan una

desventaja la cual es que sus variables independientes no pueden ir más allá del rango de entrenamiento, es quiere decir que no podrá predecir precios para años superior al 2021. Pero si será capaz de predecir para todos los meses del 2020 ya que la base de datos abarca como máximo agosto del 2020.

Si bien el 2021 representa una fecha pasada a la fecha de la elaboración de este proyecto de grado cabe recalcar que la base de datos abarca hasta agosto del 2020 como la ultima fecha donde se registró la última fecha de precio del gas lo cual se evidencia en la figura 1.4 con el comando 'tail' de Python que muestra los últimos 5 datos por defecto. En la figura 4.5 se alimentará muestro modelo con 2 variables las que son el mes y el año, donde deben estar en el mismo orden y respetando el rango de entrenamiento que se dio por ejemplo para el año con el rango de 1987 hasta el 2020. En caso de tener una nueva base de datos con precios de estos últimos años 2021 y 2020 sería capaz de predecir el precio hasta finales del 2022. Entonces se predecirá el precio del gas natural para los meses de septiembre, octubre, noviembre y diciembre para el año 2020. Dichas fechas de septiembre a diciembre del 2020 no forman parte de la base de datos.

Figura 4.5: Predicción a futuro para el 2020.



Fuente: Elaboración propia con datos de Kaggle.

En la figura 4.5 se puede ver que se pasa como parámetros al modelo llamado ‘bosque_aleatorio’ y gracias al comando ‘predict’ que es una función de python para predecir un valor luego de a ver entrenado un modelo.

Los datos de septiembre a diciembre para el año 2020, no están en la base de datos original la base de datos original solo tiene datos hasta agosto de 2020.

En la tabla 4.6 se presentará de forma entendible los resultados de la figura 4.5.

Tabla 4.6: Predicción de septiembre a diciembre del 2020.

Mes	Año	Precio USD/MMBTU
9 (septiembre)	2020	2.27
10 (octubre)	2020	2.01
11 (noviembre)	2020	1.79
12 (diciembre)	2020	178

Fuente: Elaboración propia en Excel con Datos de Kaggle, 2021.

Si bien el modelo no puede ser capaz de predecir valores con tiempos actuales, si puede predecir para valores que nunca vio en el momento de su entrenamiento en este caso como septiembre 2020, octubre 2020, noviembre 2020 y diciembre 2020.

La solución para predecir precios más actuales es entrenaron con una base de datos más actual que incluya fechas del 2021 y 2022.

5. CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES

5.1.CONCLUSIONES

A continuación, se detallan las conclusiones obtenidas por el análisis que se realizó en el presente trabajo:

- Se logro predecir el precio del gas natural internacional para dicha tarea se ejecutaron 3 modelos que fueron: regresión lineal, árbol de decisión y bosque aleatorio donde los 3 muestran ser capaces de predecir el precio del gas ya que muestran un coeficiente de determinación arriba de un 95%.

  Regresión lineal múltiple. Este modelo es alimentado por las variables de los promedios móviles de 3 días anteriores (m_3) y nueve días anteriores (m_9), su desventaja que presenta es que necesitaría dichos datos de tres y nueve días anteriores para funcionar, es decir que si usaríamos para predecir precios con 10 días o más de rango este modelo no se ejecutaría.

  Árbol de decisión regresor. Este modelo es alimentado por las variables de mes y año ya que estas demostraron mediante graficas que tiene incidencia en el precio es decir que en ciertos meses es casi habitual un incremento del precio. La desventaja que se noto es que al ser bastantes datos (5953) tiene una tendencia al sobre ajuste.

  Bosque aleatorio regresor. De la misma manera que el árbol de decisión este modelo tomo como variables el mes y el año. El aspecto favorable de este modelo en comparación a un árbol de decisión es que este bosque aleatorio genera varios árboles aleatorios y saca un promedio de la respuesta para presentar por este motivo tiene menor probabilidad a caer en un sobre ajuste. Este modelo solo puede predecir precios hasta diciembre del 2020 lo cual se lo demostró en la figura 4.5.

- La base de datos solo presento 1 dato nulo el que tenia de fecha 2018-01-05 el cual se lo reemplazo por la media de sus intervalos.

  Regresión lineal múltiple. En este modelo se presenta la existencia datos nulos provenientes de medias anteriores se generó más valores nulos, en dicho caso por no tener valores en su cabeza se procedió a eliminar todas las filas que tenían valores NAN aproximadamente fueron 9 filas.

  Árbol de decisión regresor. Al aplicar este modelo no se obtuvieron más valores nulos.

  Bosque aleatorio regresor. Al aplicar este modelo no se obtuvieron más valores nulos.
- Se desarrollaron los modelos planteados de aprendizaje automático de Regresión lineal, árbol de decisiones y bosque aleatorio cabe destacar que la manera de alimentar a cada modelo fueron distintas, salvo en el árbol de decisión y bosque aleatorio.
- Las gráficas que se propusieron denotaron la explicación del precio del gas natural unas más que las otras, sin duda las estaciones el año juegan un papel importante dichas graficas están en ANEXOS.
- Se obtuvo la métrica para cada modelo que fueron superiores al 95% que es el coeficiente de determinación.
- El modelo que se escogió fue el de bosque aleatorio regresor debido a que marcaba un coeficiente de determinación de 95 %, al formarse varios árboles de decisión posee menor probabilidad de caer en un sobre ajuste.

5.2.RECOMENDACIONES

- Las recomendaciones para aplicar el modelo serían las siguientes:

  Regresión lineal múltiple. Al ser alimentados de promedios móviles poseen un rango pequeño de predicción ya que los promedios móviles son alimentados por precios anteriores del gas natural.

  Árbol de decisión regresor. Para disminuir la probabilidad de caer en un sobre ajuste posiblemente se debería tomar solo los últimos 10 años.

  Bosque aleatorio regresor. Al ser el modelo escogido se tiene pocas recomendaciones una podría ser trabajar con un mayor número de árboles.
- Se recomendaría que tomen de base esta documentación se debería actualizar la data, la complejidad estaría en hacer crear un Bot para realizar Web Scripting en distintas paginas donde se cotizan los precios de las materias primas. Para ser almacenadas en la nueva data, es decir que la data se actualice cada día.
- Crear una página Web o un programa conectado a internet para que los interesados de la predicción del Precio del Gas Natural puedan poner una fecha, para que la pagina y/o programa sea capaz de predecirle el precio.
- La implementación de otros modelos de series temporales como ARIMA y/o AUTOARIMA.
- Cabe recalcar que el Precio del Gas Natural es muy susceptible a presentar varias y bruscos cambios de precio debido a geopolítica mundial. El modelo plateado en este trabajo es una ayuda en la predicción, ya que el Gas puede alcanzar picos altos como bajos debido a lo anterior descrito.

- Implementación de una cultura de datos y la importancia que estos tienen hoy en día, hubiera sido un mejor caso de estudio si tuviéramos los datos de por ejemplo el precio de venta de Gas al Brasil, venta a la Argentina para cada día, trimestral o anual.
- Implementación de aprendizaje automático en la industria energética del país, e incentivar a los profesionales y futuros profesionales de las habilidades que aporta las tecnologías inteligentes.

Bibliografía

ABUÍN, J. R. (2007). Regresión lineal múltiple. IdEyGdM-Ld Estadística, Editor, 32.

BAVIERA, T. (2017). Técnicas para el análisis del sentimiento en Twitter: Aprendizaje Automático Supervisado y SentiStrength. Dígitos, 1(3), 33-50.

BISHOP, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.

CAMPODÓNICO SÁNCHEZ, H. (1999). La industria del gas natural y su regulación en América Latina. Revista de la CEPAL.

CHÁVEZ, G. (2013). Ingresos fiscales por explotación de hidrocarburos en Bolivia. Banco Interamericano de Desarrollo Resumen de Políticas, 199, 1-54.

CHITARRONI, H. (2002). La regresión logística.

GARCÍA, A. (2012). Inteligencia Artificial. Fundamentos, práctica y aplicaciones. Rc Libros.

GARRÓN, M., & Cisneros, P. (2007). Metodologías para la determinación de precios de gas en la región. Artículos técnicos. Quito: OLADE.

GRANADOS, R. M. (2016). Modelos de regresión lineal múltiple. Granada, España: Departamento de Economía Aplicada, Universidad de Granada.

JEMIO, L. C. (2000). Impacto de las Exportaciones de Gas al Brasil sobre la Economía Boliviana. Draft, La Paz.

IGLESIAS, E. P. (2003). Petróleo y gas natural (Vol. 5). Ediciones Akal.

KOZULJ, R. (2004). La industria del gas natural en América del Sur: situación y posibilidades de la integración de mercados. CEPAL.

MEDINACELI MONRROY, S. M. (2021). Breve análisis y prospectiva de la industria del gas natural boliviano: 1980-2021. Revista Latinoamericana de Desarrollo Económico, (36), 169-226.

MILLÁN, D. B., Boticario, J. G., & Viñuela, P. I. (2006). Aprendizaje automático. Sanz y Torres.

Moreno, A. (1994). Aprendizaje automático.

PELÁEZ, I. M. (2016). Modelos de regresión: lineal simple y regresión logística. Revista Seden, 14, 195-214.

PÉRTEGA DÍAZ, S., & Fernández, P. (2000). Técnicas de regresión: Regresión lineal simple. Investigación.

RAUCH-HINDIN, W. B. (1989). Aplicaciones de la inteligencia artificial en la actividad empresarial, la ciencia y la industria. Ediciones Díaz de Santos.

SERÁN, S. D. U. H., & HAYAN, U. (2006). Aprendizaje automático.

URIEL, E. (2013). Regresión lineal múltiple: estimación y propiedades. Universidad de Valencia Versión, 09-2013

YPFB. (07 de mayo de 2021). Obtenido de https://www.ypfb.gob.bo/es/informacion-institucional/noticias/1317-gas-y-petr%C3%B3leo-representan-el-43-del-total-de-exportaciones-del-pa%C3%ADs-de-2006-a-2020.html

ANEXOS:

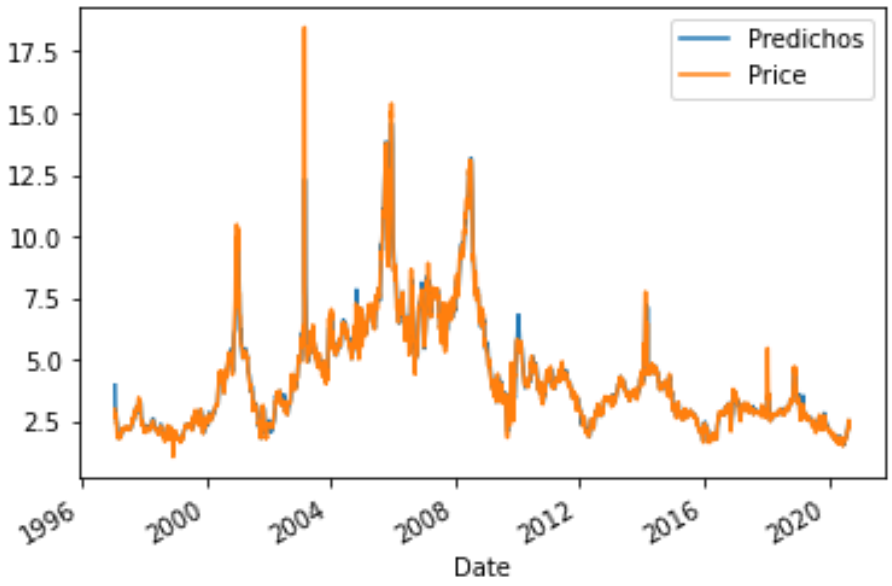
Modelo de regresión lineal múltiple

[150]
0 s

index	Predichos	Date	Price	m_3	m_9
10	3.960552	1997-01-21	2.99	3.960000	3.951111
13	3.021740	1997-01-24	2.62	3.000000	3.692222
18	2.950882	1997-01-31	2.77	2.940000	2.964444
21	2.633419	1997-02-05	2.65	2.616667	2.803333
23	2.600177	1997-02-07	2.39	2.583333	2.756667
...	...	...	...	...	...
5944	2.412948	2020-08-20	2.35	2.400000	2.243333
5947	2.450139	2020-08-25	2.54	2.436667	2.331111
5948	2.512714	2020-08-26	2.52	2.500000	2.385556
5951	2.514160	2020-08-31	2.30	2.500000	2.470000
5952	2.442498	2020-09-01	2.22	2.426667	2.453333

1784 rows × 5 columns

Grafica de comparación entre los valores reales y predichos del modelo de regresión línea múltiple

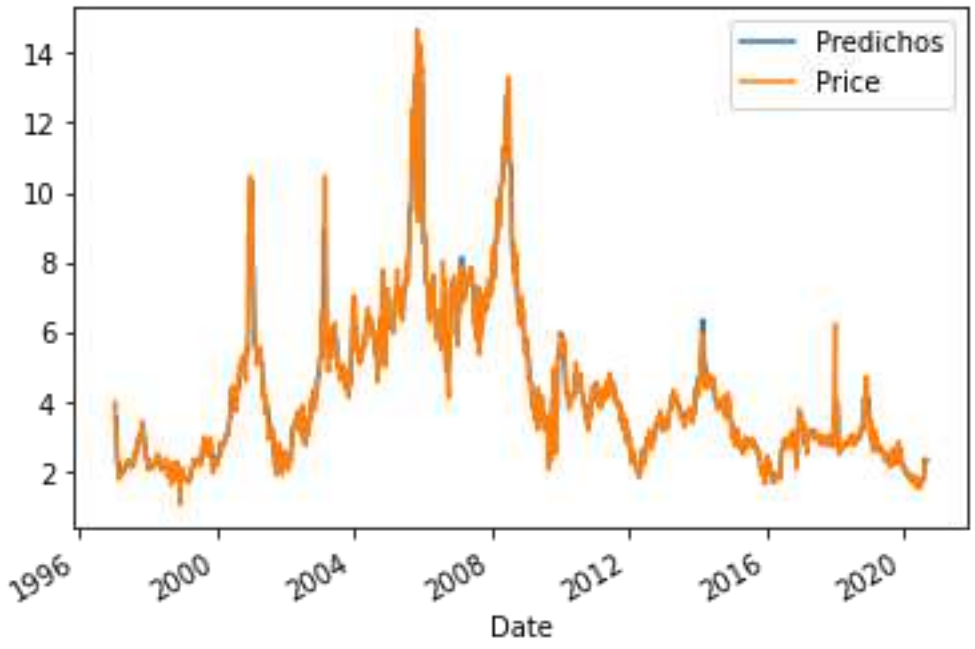


Modelo de árbol de decisión

index	Predichos	Date	Price	year	month	day
1	3.562500	1997-01-08	3.80	1997	1	8
4	3.562500	1997-01-13	4.00	1997	1	13
9	3.562500	1997-01-20	3.26	1997	1	20
12	3.562500	1997-01-23	2.96	1997	1	23
14	3.562500	1997-01-27	2.98	1997	1	27
...	...	...	...	...	...	...
5931	2.351875	2020-08-03	1.95	2020	8	3
5932	2.351875	2020-08-04	2.07	2020	8	4
5938	2.351875	2020-08-12	2.05	2020	8	12
5944	2.351875	2020-08-20	2.35	2020	8	20
5951	2.351875	2020-08-31	2.30	2020	8	31

1786 rows × 6 columns

Grafica de comparación entre los valores reales y predichos del modelo de árbol de decisión

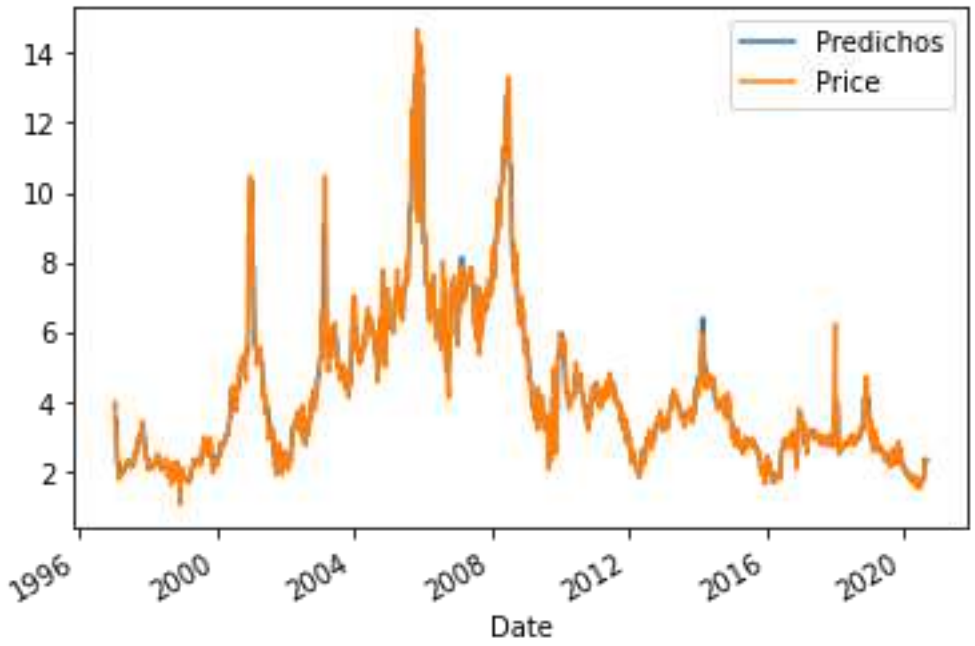


Modelo de bosque aleatorio

Date	Predichos	Price	year	month	day
1997-01-08	3.50576	3.80	1997	1	8
1997-01-13	3.50576	4.00	1997	1	13
1997-01-20	3.50576	3.26	1997	1	20
1997-01-23	3.50576	2.96	1997	1	23
1997-01-27	3.50576	2.98	1997	1	27
...	...	...	...	...	...
2020-08-03	2.35230	1.95	2020	8	3
2020-08-04	2.35230	2.07	2020	8	4
2020-08-12	2.35230	2.05	2020	8	12
2020-08-20	2.35230	2.35	2020	8	20
2020-08-31	2.35230	2.30	2020	8	31

1786 rows × 5 columns

Grafica de comparación entre los valores reales y predichos del modelo de bosque aleatorio

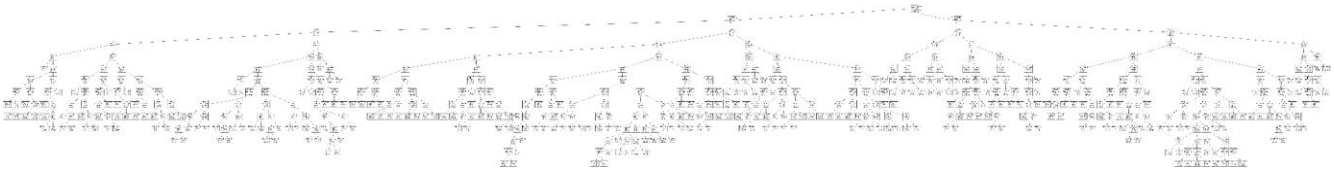


Ecuación regresión lineal simple.

```
print('El valor de la interseccion es (A): ', regresion_lineal.intercept_)
print('El valor de la pendiente es (B): ', regresion_lineal.coef_)
```

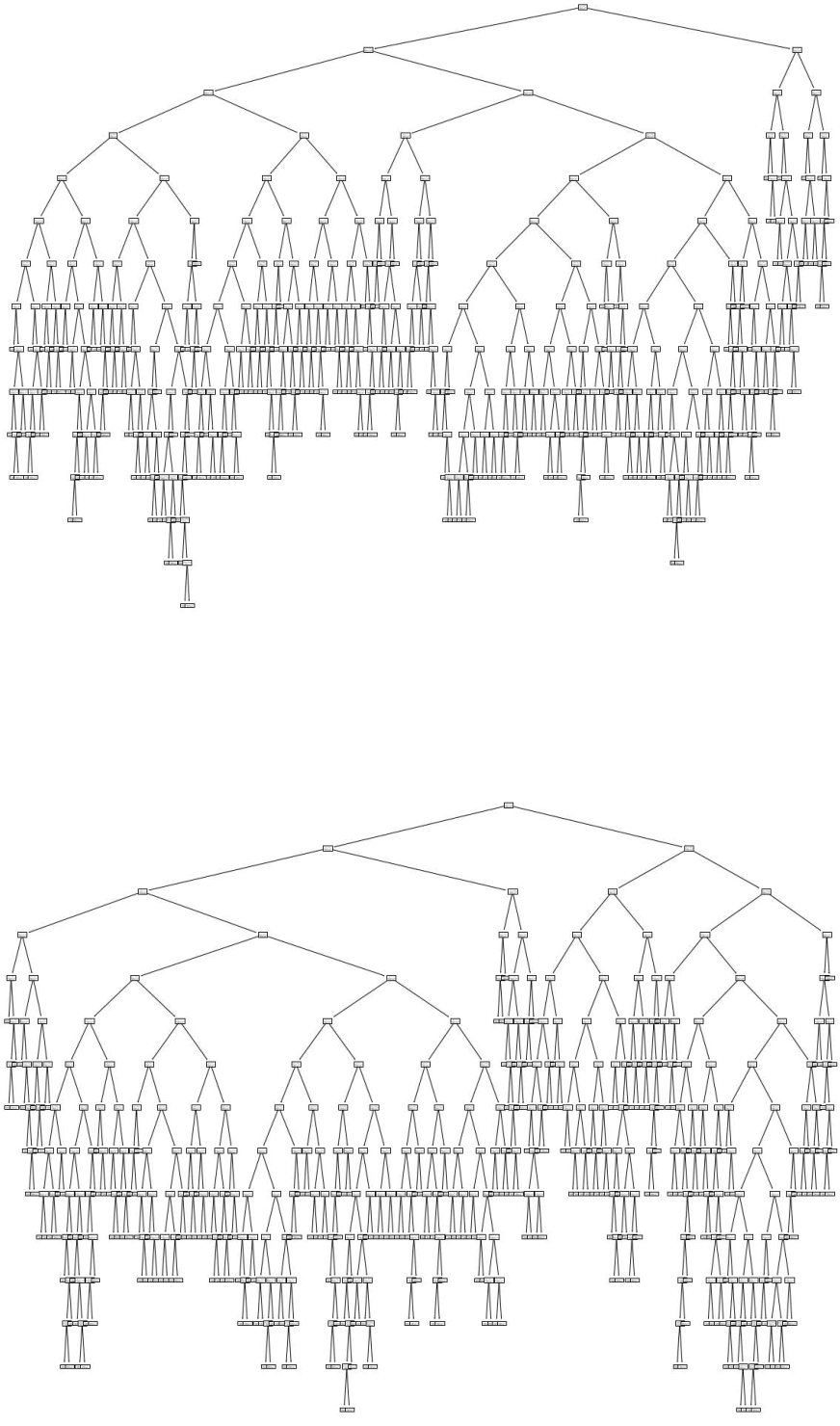
```
El valor de la interseccion es (A):  [0.03859385]
El valor de la pendiente es (B):  [[0.97331216 0.01711965]]
```

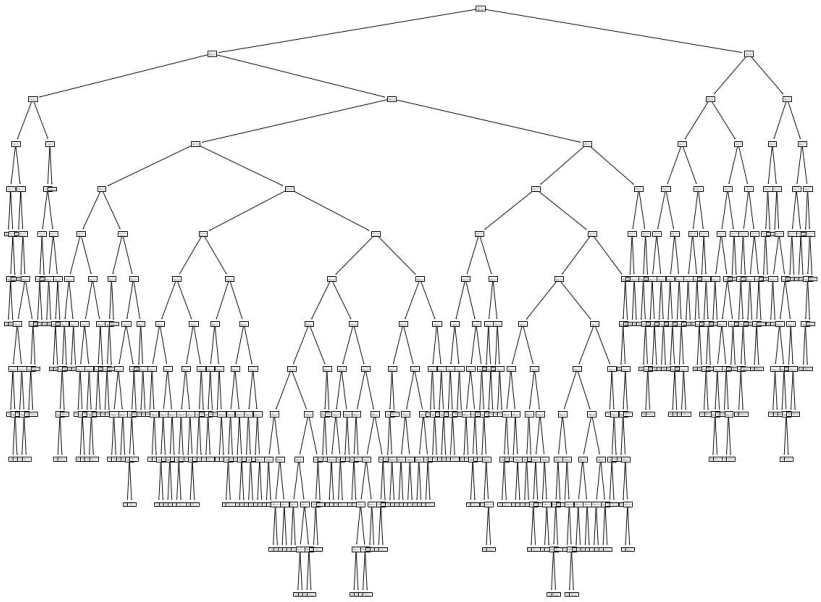
Figura del Árbol de decisión



Nota. Se puede evidenciar la existencia de el árbol de decisión, debido al tamaño de la Data más de 5000 datos no se puede colocar en una imagen que sea más visible en una sola hoja.

Figura de Bosque Aleatorio





NOTA. Como se indicó en el la memoria de desarrollo se formaron más de 100 árboles por defecto, solo se procedió a graficar unos ejemplos debido a la redundancia del mismo.

Código regression lineal multiple

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression
from datetime import datetime
df_dia = pd.read_csv('dia.csv')
display(df_dia.info())
promedio_8_dias = df_dia.loc[5280:5289,['Price']]
promedio = promedio_8_dias.median()
df_dia = df_dia.fillna(value=promedio)
df_dia['m_3'] = df_dia['Price'].shift(1).rolling(window=3).mean()
df_dia['m_9']= df_dia['Price'].shift(1).rolling(window=9).mean()
df_dia = df_dia.dropna()
X = df_dia[['m_3','m_9']]
y = df_dia[['Price']]
from sklearn.model_selection import train_test_split
#ENTRENAMIENTO CON EL 70%
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size = 0.30,
random_state=0)
regresion_lineal = LinearRegression()
regresion_lineal.fit(X_train,y_train)
regresion_lineal.score(X_test,y_test)
```

Código de árbol de decisiones

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
df_dia = pd.read_csv('dia.csv')
df_dia['year'] = pd.DatetimeIndex(df_dia['Date']).year
df_dia['month'] = pd.DatetimeIndex(df_dia['Date']).month
df_dia['day'] = pd.DatetimeIndex(df_dia['Date']).day
df_dia.drop('Date', axis=1, inplace=True)
df_dia['Price'].fillna(df_dia['Price'].mean(),inplace=True)
y = df_dia[['Price']]
X = df_dia[['day','month','year']]
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3,random_st
ate=0)
from sklearn.tree import DecisionTreeRegressor
dtr=DecisionTreeRegressor()
dtr.fit(X_train,y_train)
from sklearn.metrics import r2_score
y_pred=dtr.predict(X_test)
y_pred
accuracy=r2_score(y_test,y_pred)
accuracy
```

Código de bosque aleatorio

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
df_dia = pd.read_csv('dia.csv')
df_dia['year'] = pd.DatetimeIndex(df_dia['Date']).year
df_dia['month'] = pd.DatetimeIndex(df_dia['Date']).month
df_dia['day'] = pd.DatetimeIndex(df_dia['Date']).day
df_dia.drop('Date', axis=1, inplace=True)
df_dia['Price'].fillna(df_dia['Price'].mean(),inplace=True)
y = df_dia[['Price']]
X = df_dia[['day','month','year']]
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3,random_st
ate=0)
from sklearn.tree import RandomForestRegressor
dtr= RandomForestRegressor()
dtr.fit(X_train,y_train)
from sklearn.metrics import r2_score
y_pred=dtr.predict(X_test)
y_pred
accuracy=r2_score(y_test,y_pred)
accuracy
```