Pronóstico de la demanda del biodiesel mediante la aplicación de técnicas de machine learning

Forecasting the demand for biodiesel through the application of machine learning techniques

Leonardo Orjuela Camargo[1]

Cristina Díaz Ramos[2]

Resumen

Este artículo aborda la aplicación de un modelo de pronóstico de Machine Learning para establecer la demanda del principal material utilizado en la perforación de pozos petroleros. Se realizó un análisis de los datos disponibles de la compañía de estudio, permitiendo identificar los cinco materiales con mayor participación en las transacciones diarias de una operación, tomando como dato de entrada del modelo el material Biodiesel que representa el 42% de las transacciones registradas. Posteriormente, se entrenan dos modelos, Arboles de Regresión y Random Forest, se realizó una transformación y limpieza a los datos, se entrenaron dos algoritmos. Finalmente se evalúan estos modelos, usando medidas de error porcentual absoluto medio.

Palabras clave: Machine Learning, Perforación, Pozo, Biodiesel, Transacción.

Abstract

This article addresses the application of a Machine Learning forecasting model to establish the demand for the main material used in drilling oil wells. An analysis of the data available from the study company was carried out, allowing the identification of the five materials with the highest

[1] Fundación Universitaria Los Libertadores, Bogotá- Colombia. Contacto: lorjuelac@libertadores.edu.co
[2] Fundación Universitaria Los Libertadores, Bogotá- Colombia. Contacto: diazramos96@gmail.com

participation in the daily transactions of an operation, taking as input data of the model the Biodiesel material that represents 42% of the registered transactions. Subsequently, two models are trained, Regression Trees and Random Forest, a transformation and cleaning of the data was carried out, two algorithms were trained. Finally, these models are evaluated, using measures of mean absolute percentage error.

Keywords: Machine Learning, Drilling, Well, Biodiesel, Transaction.

Introducción

La industria petrolera ha mostrado un gran potencial en la economía nacional, no solo la producción ha incrementado notablemente, sino que también la inversión en exploración y desarrollo de nuevas tecnologías para aprovechar pozos no convencionales ha tenido un gran impacto en la explotación de recursos de hidrocarburos. (J. García 2014).

En la actualidad la industria petrolera colombiana, se convirtió en uno de los mayores generadores de recursos, anteriormente nuestra económica se caracterizó por ser exportadores de café, pasando a convertirse en exportador de commodities energéticos, como hidrocarburos, carbón, ferroníquel y recientemente en la explotación de reservas gasíferas con el desarrollo de nuevas tecnologías. (J. García 2014).

En el aprendizaje automático (ML: Machine Learning), se entrenan algoritmos para realizar tareas que son de alta complejidad y que podrían con otros métodos, llegar a requerir de una programación muy exhaustiva. En contraste, las técnicas de ML se limitan por la cantidad de datos o ejemplos de entrenamiento con los que se pueda alimentar un modelo en específico. Las aplicaciones del ML son tan variadas como campos del conocimiento existen, sin embargo, existen un grupo de usos comunes o de categorías donde se usan ciertas técnicas en específico y dentro de las cuales las 2 principales son clasificación y predicción (nombrada regresión por algunos autores). J.A Suarez Peña (2019) pág. 13. Modelo de Aprendizaje Automático para la Predicción de la Calidad del café. Tesis de Grado para obtener título de Magister en Ingeniería.

El presente artículo trata de explicar las dificultades que presenta la compañía de estudio, en el manejo de los inventarios de materiales, y cómo verificar el control para no incurrir en sobrecostos en la operación de perforación. Así mismo implementar una herramienta que permita a la empresa identificar el comportamiento de la demanda de los materiales más utilizados, y de esta manera contar con las cantidades requeridas al momento de la operación, sin que esto afecte los ingresos potenciales de forma significativa en los periodos subsiguientes.

Referentes Teóricos

Inventario

La palabra “inventario" se define de muchas maneras en la literatura. En general, los inventarios son las existencias de materias primas, materiales de empaque, trabajos en proceso y productos acabados que aparecen en numerosos puntos a lo largo de la producción y el canal logístico de la compañía (Ballou, 2005). (Pycraft, y otros, 2010) definieron el inventario como la acumulación almacenada de recursos materiales en un sistema de transformación, (Chase, Aquilano, & Jacobs, 2006) describieron el inventario como el stock de cualquier elemento o recursos utilizados en una organización.

Esencialmente, la administración de inventarios puede explicarse mejor como un conjunto de políticas, procedimientos y controles que monitorean y observan sistemáticamente los niveles de inventario e inteligentemente determinan el inventario que debe mantenerse, a qué hora debe reabastecerse el inventario, y la cantidad que debe ser pedida para stock. Es un proceso continuo de planificación, organización, y control cuyo objetivo es minimizar la inversión en inventario, al tiempo que se equilibra la oferta y la demanda (West, 2009).

La falla de la administración del inventario en cualquier compañía puede llevar al aumento en el monto de las pérdidas, lo que puede afectar el desempeño financiero de una empresa, según datos de Chase, Aquilano, & Jacobs, (2006) en su investigación revelaron que cerca del 80% de una muestra de empresas con anomalías en el inventario tienden a su desaparición en un periodo de 5 años.

El control interno tiene gran importancia en los inventarios existentes en las empresas, pues, de esto depende que no se presenten inconsistencias entre la información existente en las bases de datos y los productos que se tienen en el stock, tema que se ve reflejado en la productividad de la empresa, reducción de costos, mejor calidad de los productos, desempeño laboral, desempeño financiero, reducción en la presencia de delitos como fraudes, hurtos, etc. Pucuna Orozco, F. C. (2019),

Otro inconveniente que se puede presentar con la falta de control de inventarios es que debido a esa ausencia, se vean afectados los destinatarios finales del producto, es decir, los clientes; los cuales al observar esas fallas presentadas en los procesos adelantados por las empresas, decidan retirarse del acuerdo que tengan y acudir a otras industrias que le brinden la confianza y seguridad que se requiere en toda transacción comercial. C.J. Madariaga Fernández, Y. O. Lao León, D.A. Curra Sosa, R. L. (2020).

Este problema se puede evitar con el hecho, de tener una buena organización y comunicación entre las diferentes dependencias de la empresa en busca que todo funcione de manera engranada, efectiva y organizada, permitiendo mejores resultados en la productividad y satisfacción en el cliente, quien al ver este comportamiento empresarial, no dudara en conservar la relación

comercial, al punto que se puede convertir en una cadena comercial, en el entendido que puede referenciar esa competitividad y eficiencia, haciendo que más clientes hagan parte de las relaciones comerciales, toda vez que éstos van a ver en las empresas la seriedad en la fabricación, colocación, calidad de los productos entre otros ítems. Pucuna Orozco, F. C. (2019),

Concadenado a lo anteriormente enunciado, se puede establecer que con una excelente gestión de inventarios, automáticamente se verán reflejados la reducción de costos en materias primas, en procesos administrativos, de transporte, de producción, etc; tema que de manera inmediata permitirá reducir los valores del producto final, lo que generará beneficios económicos en cabeza de los clientes y de la empresa, pues, así las cosas, los compradores compraran los bienes a menor precio, con buena calidad y el productor tendrá mayor utilidad con menor esfuerzo, sin que se llegue a desmejorar la calidad del producto. Pucuna Orozco, F. C. (2019),

Así las cosas, la gestión de inventarios juega un papel muy importante para determinar con qué materia prima se cuenta, con que disponibilidad logística y de capital humano se dispone para llevar la producción a los resultados deseados y de esta manera lograr la satisfacción del cliente, generando ganancias para las partes.

La gestión del inventario se puede mejorar de varias maneras. Algunos de los métodos populares y técnicas que atraen muchas investigaciones y hallazgos empíricos, como sugiere (Williams & Tokar, 2008), incluir la implementación de un sistema de revisión periódica (Ballou, 2005), uso de tecnología como RFID ((Talavera, Banks, Smith, & Cárdenas-Barrón, 12 2015), centralización de las localidades de siembra (Evers & Beier, 1993), adecuada política de control de inventario adaptado (Fu, Ionescu, Aghezzaf, & De Keyser, 2015), e integración de almacenes con un sistema de control del inventario (Thomas & Tyworth, 2006).

Los costos de inventario al administrar el inventario, se deben considerar varios costos para determinar los niveles de inventario óptimos. (Owoeye, Adejuyigbe, Bolaji, & Adekoya, 2014) coincidió en que el costo de comprar y mantener el inventario puede representar entre el 60% y el 80% del costo total de un producto o servicio. Según (Gourdin, 2005), se deben considerar tres tipos de costos: costos de mantenimiento, orden y desabastecimiento.

Demanda se define: como la cantidad y calidad de bienes y servicios que serán adquiridos por un comprador en un lapso de tiempo, todo aquello guardando relación con la oferta que para el momento exista en el mercado, como para citar un ejemplo, con el tema de la pandemia por la que atraviesa el país, donde a inicios de la misma, la demanda en productos desinfectantes fue tan elevada que los productores no dieron abasto para cumplir con las necesidades creadas por los consumidores.

Definición de pronosticar

Es realizar un enunciado sobre el valor futuro de una variable de interés, fundamentado ya sea por el análisis de datos históricos disponibles, por el juicio de expertos en el tema o por una combinación de ambas cosas (Montemayor, 2013).

Predecir el futuro con la mayor precisión posible, dada toda la información disponible, incluidos los datos históricos y el conocimiento de cualquier evento futuro que pueda afectar los pronósticos. (Hyndman & Athnasopoulos, 2018)

El pronóstico, como su nombre lo dice es determinar lo que puede suceder a futuro y esto se logra efectuando un análisis de hechos pasados para así determinar el momento y cómo se realizarán los proyectos trazados por las empresas; claro es, que se debe ir realizando ajustes a tales pronósticos a medida que se va evolucionando en su desarrollo y por lo tanto es muy importante trazar por decirlo así, metas a corto, mediano y largo plazo, con el fin de llevar a buen término las metas establecidas. Chicaiza Ipiales, J. A. (2019).

Para la consecución de estos objetivos a través de los pronósticos, debemos remitirnos a un conjunto de números, metodologías, proyecciones, antecedentes, análisis de expertos, etc.

Las redes neuronales artificiales son métodos de pronóstico que basadas en modelos matemáticos simples del cerebro. Permiten relaciones complejas no lineales entre la variable de respuesta y sus predictores (Hyndman & Athnasopoulos, 2018).

El modelo de predicción de la red neuronal se establecerá en función de los factores de influencia, con las entradas siendo los factores que afectan la demanda, y la salida es la demanda (Hu, Sun, & Wen, 2014).

Pronóstico de la Demanda

El pronóstico de demanda es la predicción de las ventas futuras a través del uso de métodos ya sean éstos cualitativos o cuantitativos basados en recolección de datos históricos, experiencias o información de los vendedores. (Heizer & Render, Dirección de la producción y de operaciones: Decisiones Estratégicas, 2007).

Dentro del campo del pronóstico de la demanda y más específicamente el pronóstico de la demanda de productos terminados se han realizado varios estudios se puede citar el trabajo de (Aburto 2017), el cual desarrolla un enfoque hibrido para el pronóstico de la demanda de productos terminados partiendo del análisis de un modelo ARIMA e introduciéndolo en una red neuronal la cual intenta reproducir los patrones de una serie histórica obtenida sobre la demanda de productos de consumo diario en un supermercado (Aburto & Weber, 2007).

En las empresas que comercializan a nivel macro, es de gran importancia la clasificación de inventarios, pues, debido a su elevada oferta y demanda, se requiere la existencia de organización en sus procesos de producción, evitando problemas como disparidad entre las existencias de

materia prima y los compromisos adquiridos en lo que respecta al producto final; al igual, que las demás fases que se deben observar en el proceso de producción de bienes y servicios.

Dentro de las técnicas más utilizadas para adelantar estos procesos de manera eficiente y con los mejores resultados tanto para la empresa como para el cliente, se encuentra el denominado diagrama de Pareto, que no es otra cosa que permitir representar gráficamente la información desde la más importante hasta la de menor importancia para así enfocarse en la manera y el orden en que se debe dar solución a los mismos y así, llevar a un buen término los procesos de producción. C.J. Madariaga Fernández, Y. O. Lao León, D.A. Curra Sosa, R. L.

Para ejemplarizar, se puede tener en cuenta el análisis ABC, el cual es un método del principio de Pareto, método que permite identificar todos los productos que tienen mayor impacto frente a los demás, lo cual se establece de acuerdo a parámetros preestablecidos, yendo del más importante al menos; claro es, teniendo en cuenta que en los procesos de producción no existe temas sin relevancia, pues, estos se debe tener como un engranaje, donde una gestión es requisito esencial para el desarrollo de la otra. C.J. Madariaga Fernández, Y. O. Lao León, D.A. Curra Sosa, R. L.

Snyder, Koehler & Ord (2002) desarrollaron un modelo de inventarios que parte de pronósticos por suavización exponencial. Little & Coughlan (2008) estudiaron la optimización de los stocks de seguridad a partir de restricciones en instituciones hospitalarias. Ferbar (2010) propuso integrar el modelo de inventarios y el de pronósticos optimizando tanto los parámetros como los valores iniciales.

Teunter, Syntetos, Babai & Stephenson (2011) estudiaron el efecto de los modelos de pronósticos en los costos y en los niveles de servicio; ellos propusieron que se tengan en cuenta esos factores, además de la minimización del error para evaluar el modelo a aplicar. Mientras que la técnica de pronósticos Holt- Winters con el añadido de la diferenciación de nivel de servicio por clasificación ABC, fue empleada por Arango, Giraldo y Castrillón (2013) en empresas comercializadoras y de servicios. Descripción realizada por C.J. Madariaga Fernández, Y. O. Lao León, D.A. Curra Sosa, R. L. Martín. (2020 pag. 360. Retos de la Dirección 2020.

Los métodos de *machine learning* han demostrado mejor rendimiento que todas las técnicas estadísticas utilizadas para el análisis de serie de tiempo (Fry & Brundage, 2020; Makridakis, Spiliotis & Assimakopoulos, 2018). La capacidad de aproximación universal de las redes neuronales (NN, de las siglas de *Neuronal Networks*) para funciones continuas que tienen primera y segunda derivada en todo su dominio ha sido verificada matemáticamente. Adicionalmente, varios estudios demuestran que las NN pueden aproximar con exactitud diversos tipos de relaciones funcionales complejas. En tal sentido muchos son los estudios que versan sobre las distintas formas de conformar NN para realizar el pronóstico de la demanda. Medeiro, Tersvirta & Rech (2006) proponen un modelo híbrido entre un modelo autorregresivo (AR) y una red neuronal con una sola capa oculta. Tiene como principal ventaja el bajo costo computacional de su solución. Otros como

Li, Luo, Zhu, Liu & Le (2008) consideran la combinación de modelos AR y una NN con regresión generalizada (GRNN). El resultado indica que es un método efectivo para obtener lo mejor de los dos en un solo modelo.

Para producir pronósticos más precisos con datos incompletos proponen Khashei, Reza & Bijari (2008) un modelo híbrido basado en el concepto básico de una NN con un modelo de regresión difuso. Posteriormente en 2010, dos de estos mismos autores, Khashei y Bijari presentan un nuevo híbrido basado en un modelo de NN con el uso de la metodología ARIMA (modelo autorregresivo integrado de media móvil) para obtener un pronóstico más exacto (Khashei & Bijari, 2010). Esta misma combinación de ARIMA con NN fue la solución dada por Wang, Zou, Su, li & Chaudhry (2013), en la cual se usaron tres *datasets* (conjunto de datos) con similares resultados.

Wang, Fang & Niu (2016) proponen un modelo híbrido basado en redes neuronales recurrentes de Elman (ERNN) con series de tiempo estocásticas, en el cual demostraron que las redes neuronales presentan mejor rendimiento que la regresión lineal, la distancia invariante compleja (CID), multiescala (MCID). Las redes neuronales con percertrón multicapa fueron empleadas por Rivas (2017) con mejores resultados de error medio de estimación para el pronóstico. Descripción realizada por C.J. Madariaga Fernández, Y. O. Lao León, D.A. Curra Sosa, R. L. Martín. (2020 pag. 361. Retos de la Dirección 2020.

Esta misma tipología de NN fue empleada en Xu & Chan (2019) para el pronóstico de la demanda de materiales médicos. Por otra parte el aprendizaje de máquina para la demanda de energía, fue empleada con el algoritmo de clasificación KNN (*K-Nearest Neighbour*), por Grimaldo & Novak (2020).

Marco Conceptual

Herramientas Computacionales.

A continuación, se describe el lenguaje que se utiliza en este artículo.

Algoritmo Árbol de decisión: Según (Palma Méndez & Marín Morales, 2008) “los árboles de decisión son una representación gráfica de un procedimiento para clasificar o evaluar un concepto”. Dichos arboles están constituidos por nodos de decisión, los cuales se despliegan en ramas para cada una de las alternativas (Barber, 2012)

Árbol de Regresión Como primer algoritmo de aprendizaje automático, se utilizara un árbol de decisiones, los árboles de decisión son una clase de algoritmos de aprendizaje automático que crearán un mapa (un árbol en realidad) de preguntas para hacer una predicción. Llamamos a estos

árboles de regresión si queremos que predigan un número o árboles de clasificación si queremos que predigan una categoría o una etiqueta. Para hacer una predicción, el árbol comenzará en su base con una primera pregunta de sí / no; y según la respuesta, continuará haciendo nuevas preguntas de sí / no hasta llegar a una predicción final. D.W. Gareth James, R.T Trevor Hastie, Introduction to Statistical Learning.

Ventajas

- Fácil de entender
- Útil en exploración de datos
- Identificar importancia de variables a partir de cientos de variables.
- Menos limpieza de datos
- El tipo de datos no es una restricción
- Es un método no paramétrico

Desventajas

- Es recomendable para bases pequeñas.
- Para bases muy grandes tiene mucho gasto computacional.
- Sobreajuste: se ajusta mucho a los datos de entrenamiento por que va uno a uno.

Nodo raíz: Es el nodo de decisión más alto en un árbol de decisión.
Nodo de decisión: Un nodo de árbol o nodo principal que se divide en uno o más nodos secundarios se denomina nodo de decisión.
Nodo hoja o terminal: nodos inferiores que (en términos generales) no se dividen más.

4.2- Conceptos relacionados con la demanda de materiales para la perforación de pozos petroleros

Movimientos de inventario (transacciones)*:* Hace referencia a las transacciones de uno o varios materiales, que reflejan la cantidad de material que entra o sale de un almacén.
Transacciones de Entrada. Son traslados entre almacenes e ingresos de inventario.
Transacciones de Salida. Son movimientos de salida de inventario, consumos o utilización de materiales demandados para uso u operación.
Imputación: Es la asignación codificada del costo asociado a la transacción realizada
Biodiesel: Es el combustible utilizado para el funcionamiento de los taladros en su operación.
Centro Logístico: unidad organizativa que divide una empresa.
Almacén: Subdivisión de los centros logísticos donde se gestionan los materiales.
Taladro de Perforación: Es un equipo que se compone de varias partes, y cada una de estas partes se descompone en equipos o activos fijos, esto con el fin de controlar cada parte y equipo del taladro, ahora, cada uno de estos equipos se componen de materiales que aplican como repuestos,

Metodología

Para el desarrollo del problema planteado, en primer lugar, se procede a seleccionar una metodología para minería de datos se espera lograr la aplicación de la metodología CRISP-DM, la cual busca organizar de forma secuencial las acciones básicas y necesarias para el desarrollo de un proyecto basado en analítica de datos.

La metodología CRISP–DM fue creada por el grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000, es actualmente la guía de referencia más utilizada en el desarrollo de proyectos de Data Mining. Estructura el proceso en seis fases: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación e Implantación (Chapman, 2000)

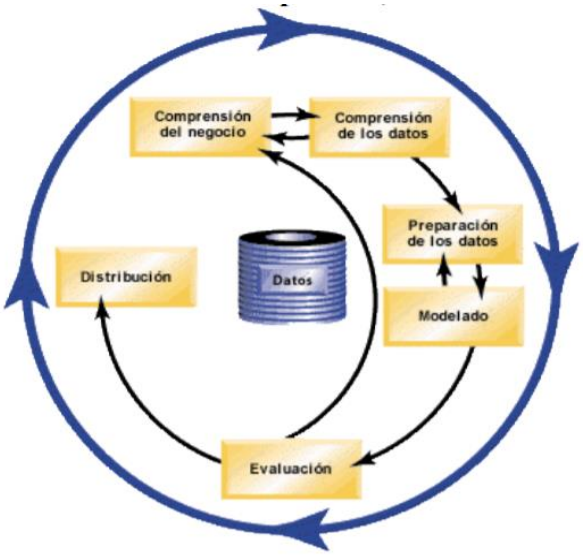

Imagen No. 1 Metodología CRISP-DM para proyectos de analítica de datos. Fuente: (IBM Corporation, 2012)

Este proyecto se desarrollará por etapas:

1) ***Conocimiento del negocio***: Antes de empezar a modelar datos, lo primero que se debe hacer es dedicar parte del tiempo a explorar las expectativas de la organización con respecto a la ciencia de datos. Conocer las razones del área encargada.

2) ***Compresión de los datos:*** en ella se analizarán a profundidad el conocimiento de los datos que tiene la Empresa, la preparación de los datos suele ser la fase más larga del proyecto; en ella se exploran los datos existentes, adquiridos y adicionales. la comprensión de estos requiere de herramientas tecnológicas que permita explorar y visualizar los datos de manera óptima. Para efectos del proyecto se utilizará Python

3) ***Preparación de los datos:*** La preparación de datos es una de las fases más importantes y con frecuencia que más tiempo exigen en la minería de datos. De hecho, se estima que la preparación de datos suele llevar el 50-70 % del tiempo y esfuerzo de un proyecto (IBM Corporation, 2012) dentro de las etapas tenemos:

    - Integración de datos (Etls, Integración en Nube)
    - Selección de muestras y variables (Variables sociodemográficas y de negocio de los usuarios)
    - Agregación de registros
    - Derivación de nuevos atributos
    - Clasificación de datos para el modelado
    - Tratamiento de valores ausentes
    - División en conjunto de datos de entrenamiento y prueba.

4) ***Modelado:*** Es aquí donde el proyecto empieza a tener sentido, el modelado implica la ejecución de múltiples iteraciones, los analistas de datos ejecutaran modelos paramétricos y no paramétricos utilizando parámetros o Hiperparámetros por defecto y ajustando los mismos. En ocasiones se devuelven a la fase de preparación para modificaciones necesarias del modelo. Existen mucha forma de abarcar un problema y tanto R como Python ofrecen paquetes estadísticos para la que la iteración de estos modelos sea optimas. Donde se escogerá el modelo con mayor índice de precisión.

Para fines del proyecto trataremos de modelar los siguientes modelos:

- Modelo de Machine Learning para series de tiempo.
- Modelo de series de tiempo.
- Evaluación: en esta parte se ha completado la mayor parte del modelo, se ha concluido que los modelos utilizados son pertinentes y los de mayor rendimiento. Sin embargo, se debe evaluar los resultados con los criterios de rendimiento de la compañía
- Los modelos finales seleccionados con la metodología CRISP-DM
- Las conclusiones y hallazgos nuevos obtenidos de los modelos y proceso de analítica de datos a esto se le llama descubrimientos de patrones de comportamiento.

5) Distribución o despliegue: En esta etapa se incluyen dos tipos de activadas para finalizar el proyecto:

- Planificación y control de la distribución de resultados
- Finalización de tares de presentación y la producción de un informe final técnico.

Resultados

La información aquí registrada, corresponde a los datos recolectados por la compañía de estudio entre julio de 2015 y febrero de 2020, a través del sistema SAP donde se encuentra la administración del inventario, datos que fueron obtenidos de experiencias y resultados empíricos de los procedimientos utilizados en la demanda del material biodiesel. A continuación, se presenta un resumen de las variables posibles a medir.

Análisis y Preparación de Datos

Con los datos disponibles, se realizó una depuración de la base donde se determinó no utilizar la información del año 2015 por contener datos de solo cinco meses, esto debido a la implementación del sistema ERP SAP para el manejo del inventario, y del año 2020 por ser un año atípico por la pandemia, lo que genero la paralización de las operaciones y la reducción del uso de materiales y combustible. A continuación, se describen los pasos para el análisis de los datos de la empresa de estudio.

Primer paso identificar los materiales con mayor participación en las transacciones diarias de la operación, y transacciones de entradas y salidas. Las transacciones de entradas no tienen imputación y se señalan como NO, y las de salida con imputación se indican con SI.

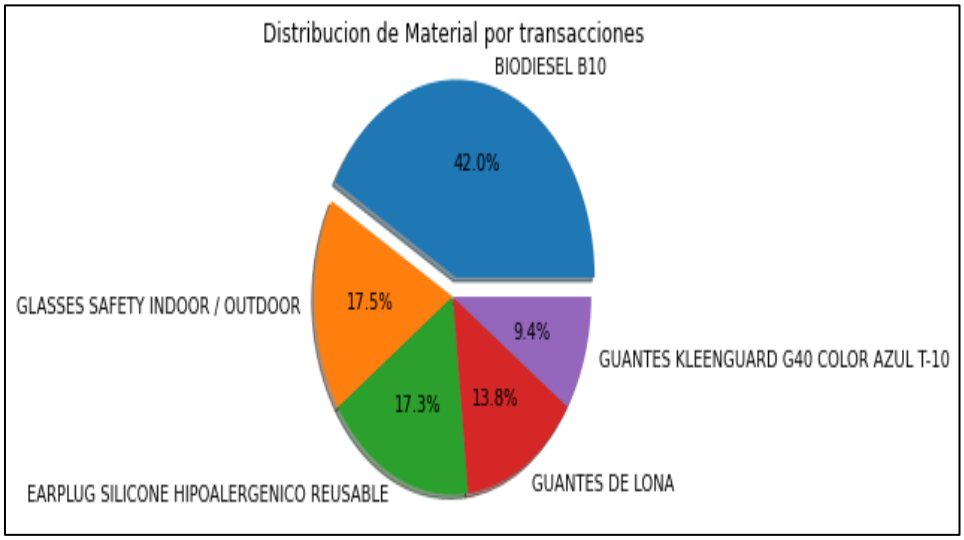

Figura 1. Gráfico de la distribución de los cinco materiales con mayores transacciones

Por lo anterior se evidencia en la figura 1, que el material con mayor participación de transacciones dentro del top cinco es el biodiesel, este material representa el 42% de las transacciones registradas en la operación.

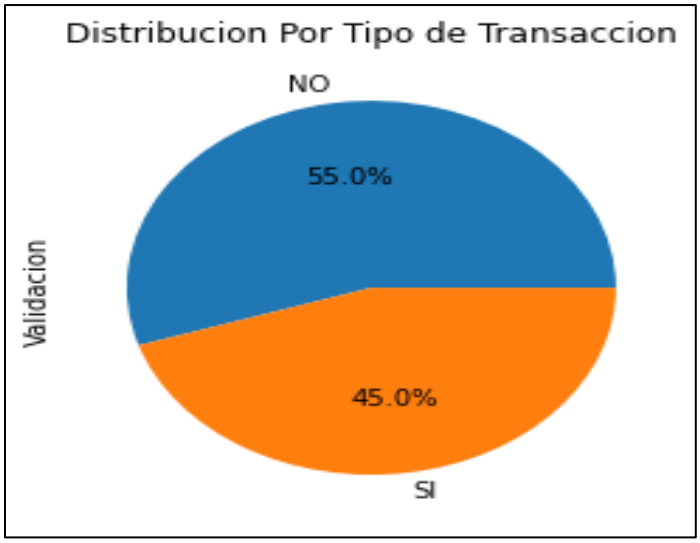

Figura 2. Gráfico por transacción SI - NO

En la figura 2, vemos que el 55% de las transacciones representan el NO tienen imputación, mientras que el 45% son las transacciones SI con imputación.

Como el objeto de estudio, consiste en el analizar los datos, y proyectar la demanda de materiales, se utilizará la variable transacciones de salida de consumo de materiales identificada con SI.

Segundo paso, Top cinco de los materiales con transacciones de salida (SI), siendo el biodiesel el material con mayor participación.

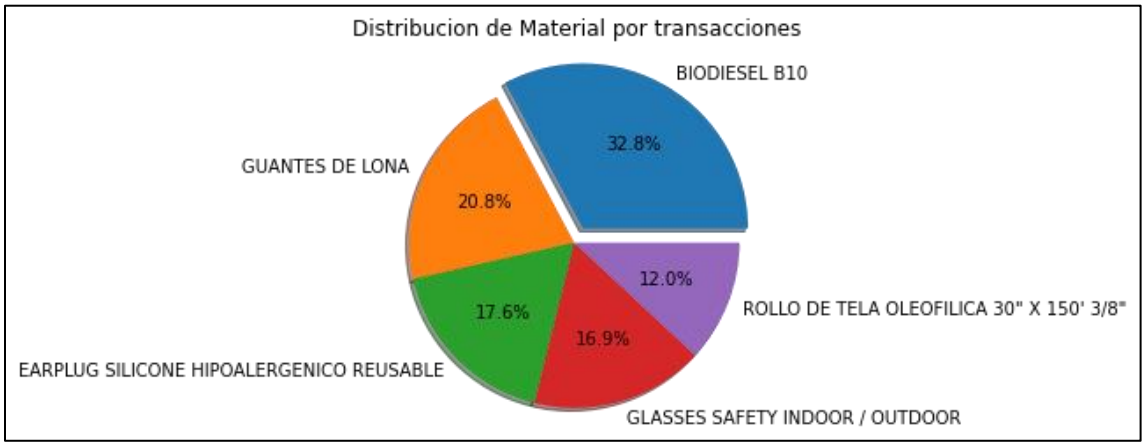

Figura 3. Gráfico de la distribución del top cinco de materiales con transacciones SI

Tercer paso, realizar la distribución de materiales por cantidades, evidenciando que el biodiesel se encuentra en el top 5 de los materiales que más consumo genera dentro de una operación, con una participación del 94.7%. Figura 4.

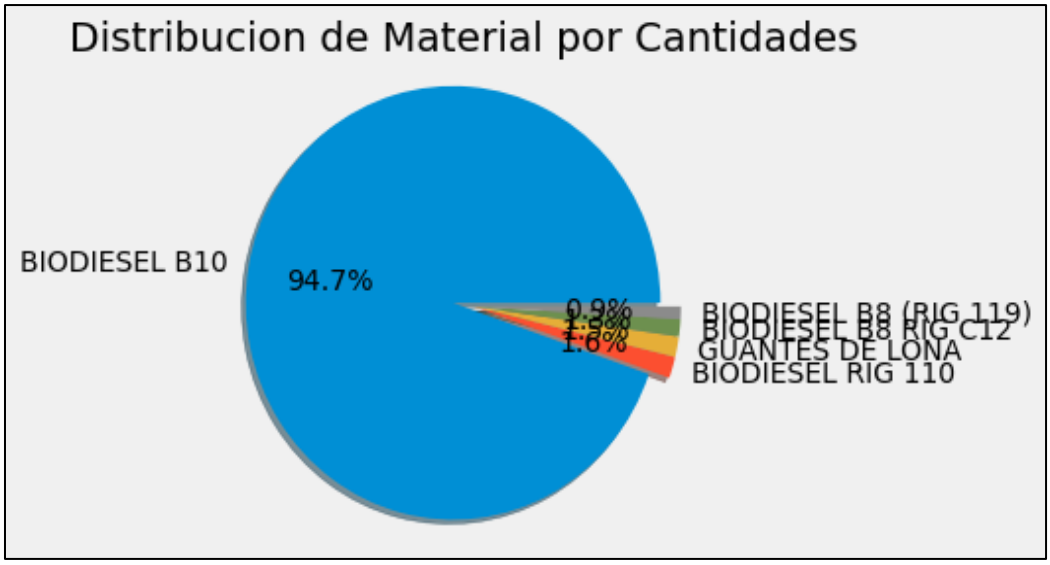

Figura 4 Gráfico de la distribución de materiales por cantidades

Cuarto paso, la gráfica 5 representa la correlación que existe entre entradas y salidas de inventario de materiales durante los últimos cuatro años, donde podemos detectar posibles anomalías, y así determinar si incrementar o disminuir la rotación para ajustar o modificar el margen de beneficios de la empresa.

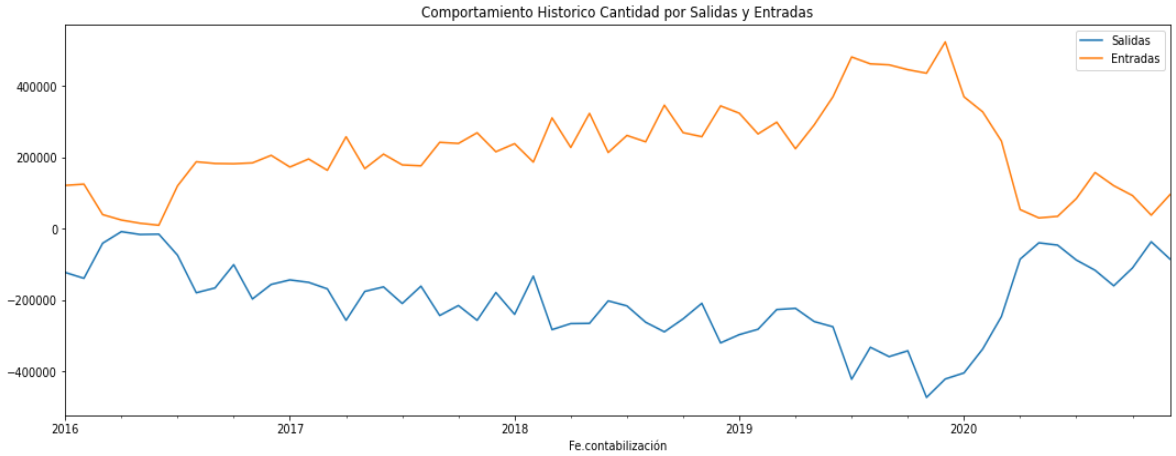

Figura 5. Gráfico de correlación de entradas vs salida de Biodiesel.

Quinto paso, identificar el top cinco de los taladros con mayor de consumo de biodiesel por cantidad de galones dentro de la operación.

Tabla 1. Distribución de Material por Cantidad

Almacén	Cantidad Galones	Participación %
R110	1.608.121	23,2%
R001	1.460.653	21,1%
R109	1.448.016	20,9%
R012	1.249.638	18,1%
R118	1.152.503	16,7%
Total	**6.918.931**	**100%**

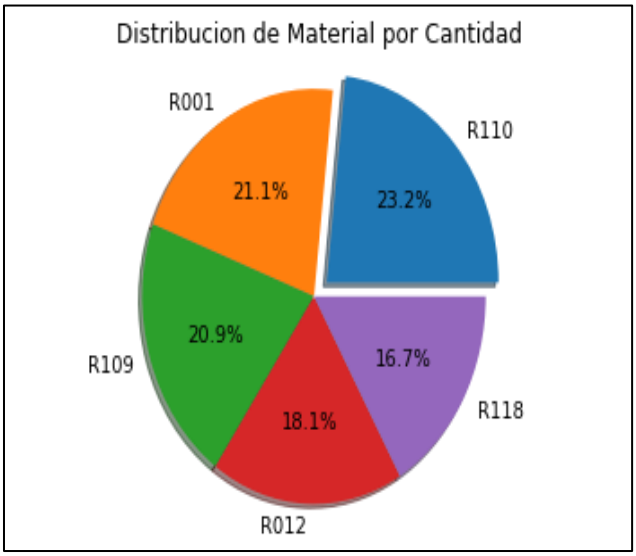

Figura 6. Gráfico del top cinco de taladros y su participación.

Realizada la limpieza del conjunto de datos históricos de la compañía de estudio, se estableció que el biodiesel es el material con mayor participación de demanda de los materiales utilizados en la operación, por lo anterior nuestra variable objeto para predecir las salidas de material en M.L es el biodiesel.

Tabla 2. Base de datos de las variables de entrada y salidas.

Make	2016-01	2016-02	2016-03	2016-04	2016-05	2016-06	2016-07	2016-08	2016-09	2016-10	2016-11	2016-12
R001	1520	2709	836	506	747	303	1532	883	983	2949	2523	3285
R012	0	0	0	0	0	2323	11331	58746	10550	22354	34243	19084
R109	0	0	0	0	0	0	0	608	2602	38158	40849	50574
R110	24739	26571	13191	9735	1059	293	85	22551,74	37034	673	4464	979
R118	0	0	0	0	0	0	0	0	0	0	0	0

Make	2017-01	2017-02	2017-03	2017-04	2017-05	2017-06	2017-07	2017-08	2017-09	2017-10	2017-11	2017-12
R001	1648	1041	0	7084	42093	26823	49224	23295	46379	46060	41825	43203
R012	250	2471	23527	28278	10493	18415	250	250	0	0	201	1453
R109	40168	30865	17402	43597	29491	29907	34703	17958	594	4313	27759	1444
R110	30	25582	53720	55285	32876	26239	50740	22855	108744	59949	25315	26758
R118	2158	25344	40049	42752	27914	40079	41978	25987	42871	21697	39612	21620

Make	2018-01	2018-02	2018-03	2018-04	2018-05	2018-06	2018-07	2018-08	2018-09	2018-10	2018-11	2018-12
R001	32428	31008	43385	40236	16390	18743	35118	28962	40167	44772	28329	38743
R012	0	0	0	776	110	32832	33686	42310	28075	30522	36303	34816
R109	23392	3216	19331	26358	41673	34476	27184	17537	44551	25500	18366	18728
R110	41653	24738	50372	28873	2330	1711	1071	897	37734	44132	35065	43546
R118	21073	29839	27901	32343	34626	31953	34089	34136	33110	34247	32988	24112

Make	2019-01	2019-02	2019-03	2019-04	2019-05	2019-06	2019-07	2019-08	2019-09	2019-10	2019-11	2019-12
R001	41109	31642	39234	39096	26313	32598	34251	35891	35916	38889	32786	36419
R012	39196	40589	35224	22075	39756	23463	35011	35734	36931	43341	46533	45740
R109	1147	22371	37540	7694	399	1792	979	8989	13289	39704	50336	33289
R110	30499	41886	39952	35354	41356	47212	17390	39250	23924	15598	37386	41321
R118	33229	23165	30015	36121	34674	35526	31736	30239	29841	33327	13580	48469

Sexto paso, identificado el top cinco de los taladros con mayor demanda de biodiesel, definimos el comportamiento del consumo de biodiesel por taladro durante el período comprendido entre los años 2016 a 2019.

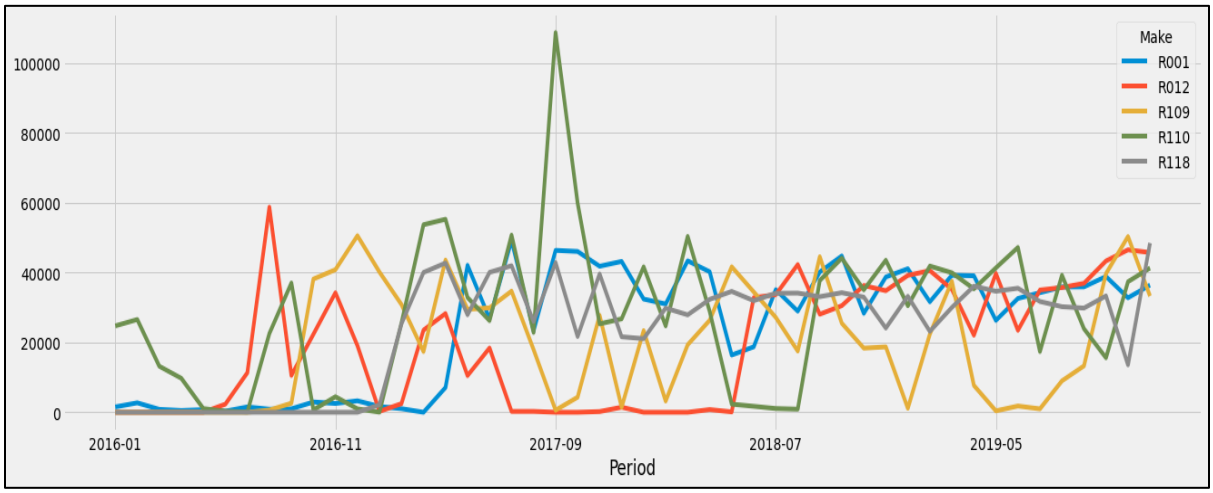

Figura 7. Gráfica del comportamiento del consumo de biodiesel por taladro durante cuatro años.

En la gráfica 7 se observa que los taladros tienen periodos de no consumo, esto debido a que un taladro opera por un periodo determinado, y después pasa a realizar un proceso de movilización hacia otro pozo, tiempo que no genera consumo de biodiesel.

Entrenamiento del algoritmo de M.L para la demanda de biodiesel

Analizadas las variables de la base de datos de la empresa de estudio, se procede a implementar el algoritmo de árbol de regresión, para nuestro problema de pronóstico básicamente mostraremos el algoritmo de aprendizaje automático de diferentes extractos de nuestro conjunto de datos de demanda histórica, como entradas y cada vez mostraremos cuál fue la siguiente observación de demanda. (N.Vandeput, 2019).

Product	Inputs				Output
	Q1	Q2	Q3	Q4	Q1 Y+1
#1	5	15	10	7	6
#2	7	2	3	1	4
#3	18	25	32	47	56
#4	4	1	5	3	3

Figura 8. Gráfica del comportamiento del consumo biodiesel por trimestre.

El algoritmo de aprendizaje automático toma diferentes extractos del conjunto de datos de la demanda histórica, como entrada la demanda mensual de cada taladro durante doce (12) meses, y para este conjunto de datos predictivos se estima la demanda del siguiente mes.

El algoritmo aprenderá la relación entre los últimos doce meses de demanda por año, y pronosticara la demanda del mes trece, como se evidencia en la figura 8, la demanda durante los cuatro periodos es 5, 15, 10 y 7 como las observaciones de demanda, la siguiente observación de demanda será 6, por lo que su predicción debería ser 6.

Evaluación del Modelo- Error de entrenamiento 21.3%

Para evaluar el error de entrenamiento se tomó tres años de consumo mensual de biodiesel del top cinco de los taladros.

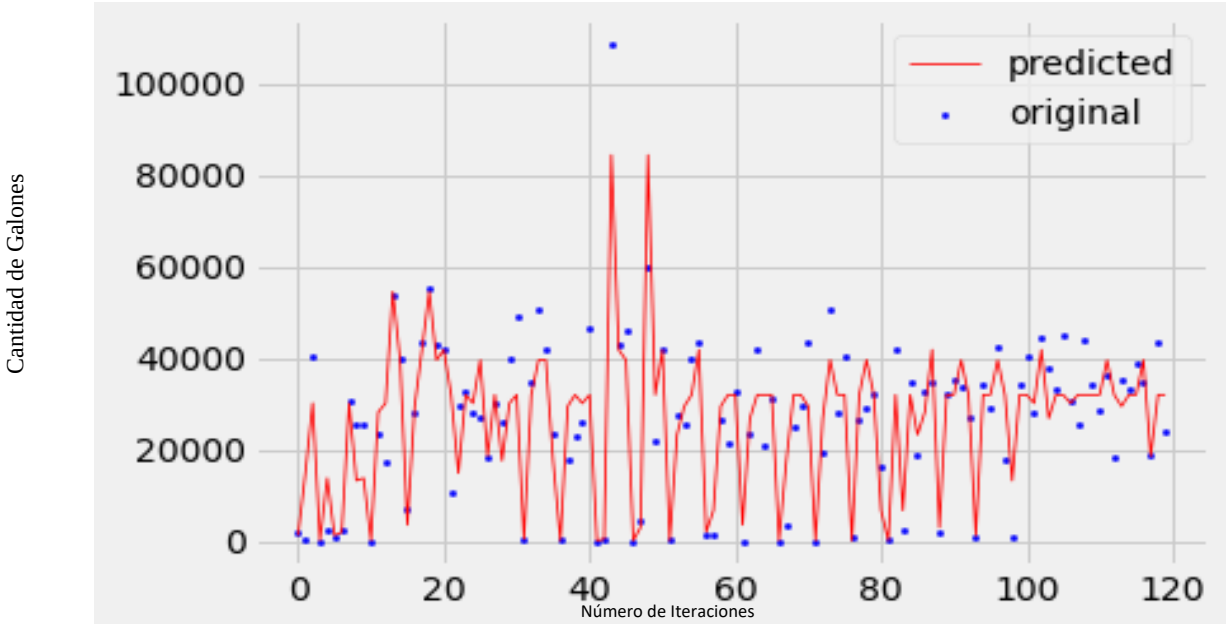

Figura 9. Gráfica del error de entrenamiento.

Evaluación del Modelo- Error de Prueba 25.0%

Para evaluar el error de prueba del modelo tomamos de la base de datos el año 2019, obteniendo los resultados de la demanda futura de periodos ya conocidos.

En la figura 9, podemos comparar las líneas como datos de salida del modelo y los puntos de color azul como los datos originales de la base de datos, existe un acercamiento de los datos como también se visualizan los datos que están sobre el eje que no alcanza a tomar.

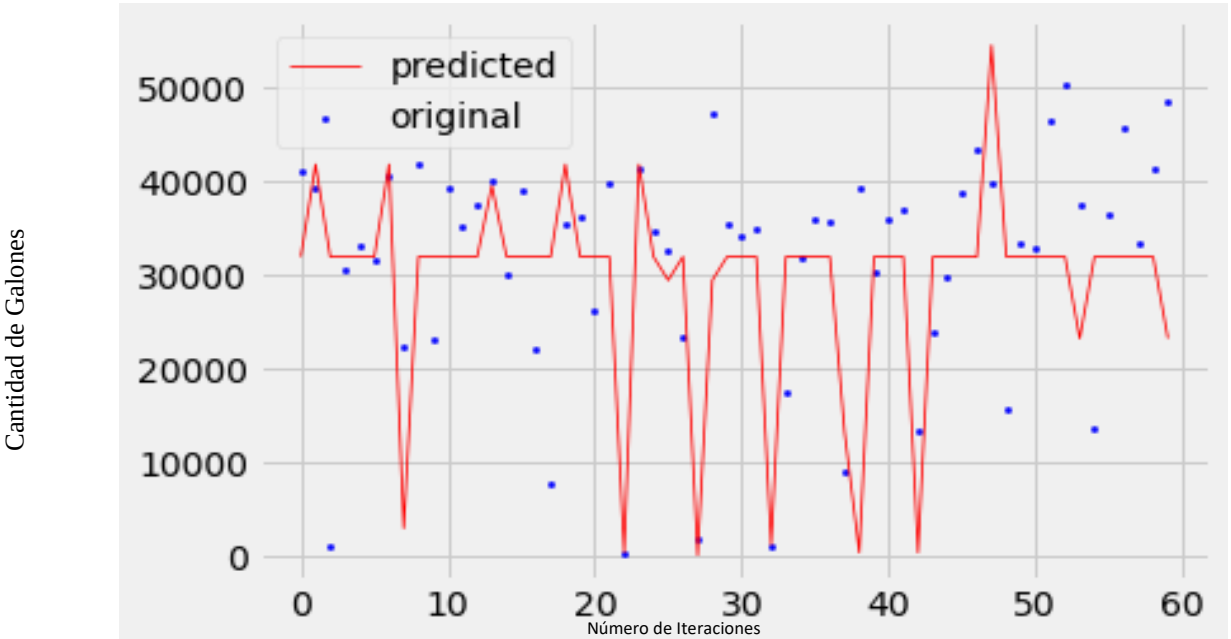

Figura 10. Gráfico del Error de prueba.

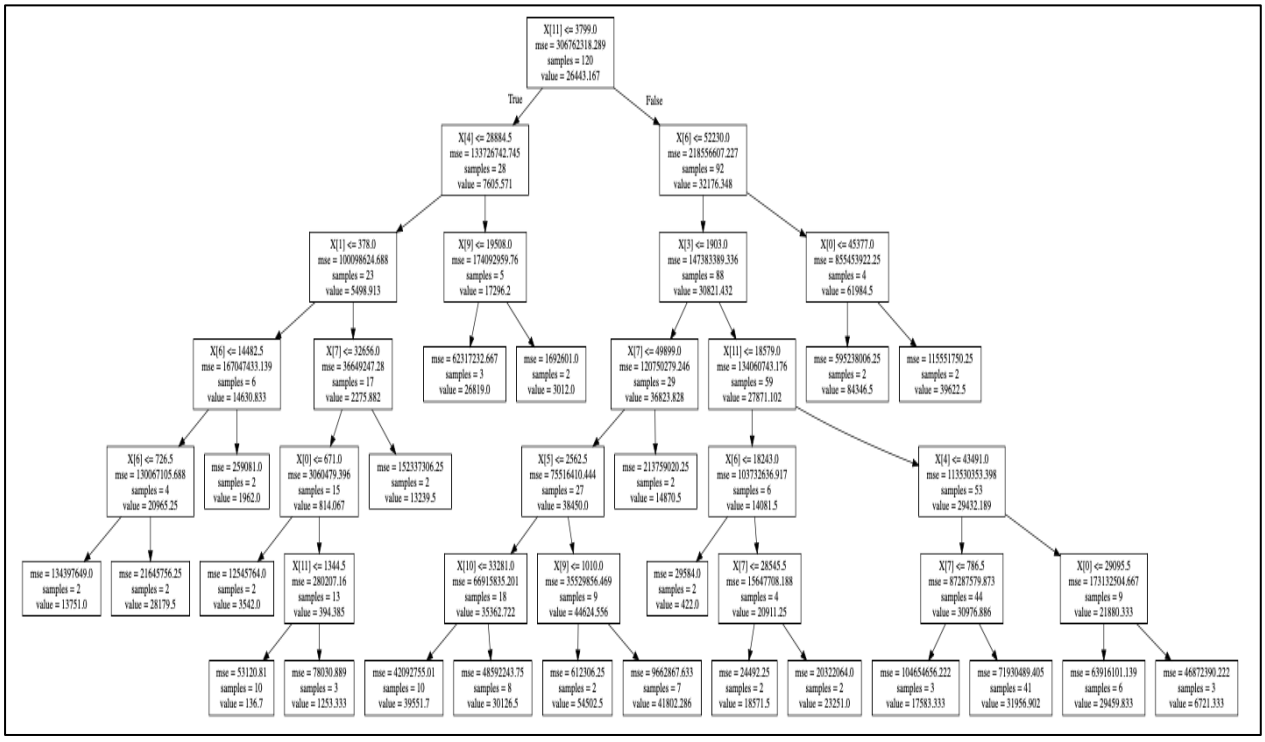

Figura 11. Diagrama de árbol de decisión para predecir la demanda de biodiesel

En la figura 11, se aprecia el modelo del árbol creado usando como variables de entrada el consumo de biodiesel mes por mes de cada taladro, tomando doce (12) meses, la variable de salida, es el mes siguiente, es decir el mes trece (13) de esta manera el modelo toma las doce (12) variables de entrada para predecir la variable de salida Y, como enero es el siguiente nivel, toma desde el mes de febrero, doce meses para predecir el mes de febrero del siguiente año, y así continua tomando desde marzo doce meses hasta marzo del siguiente año para predecir abril del siguiente año, y así sucesivamente.

Random Forest

El modelo Random Forest es nuestra segunda alternativa, ya que decidimos comparar el modelo inicial árbol de regresión con otro que nos permitiera obtener el mejor modelo, o el que más se ajusta al comportamiento de la demanda de consumo de material para los principales taladros.

El modelo Random Forest toma muestras aleatorias varias veces, comparándolas para calcular el menor margen de error, calculando el promedio de promedios.

El error generado durante la etapa de entrenamiento es del 27,9%, este es mayor que el modelo del árbol de regresión que es de 21,3%.

Error de entrenamiento % 27.9

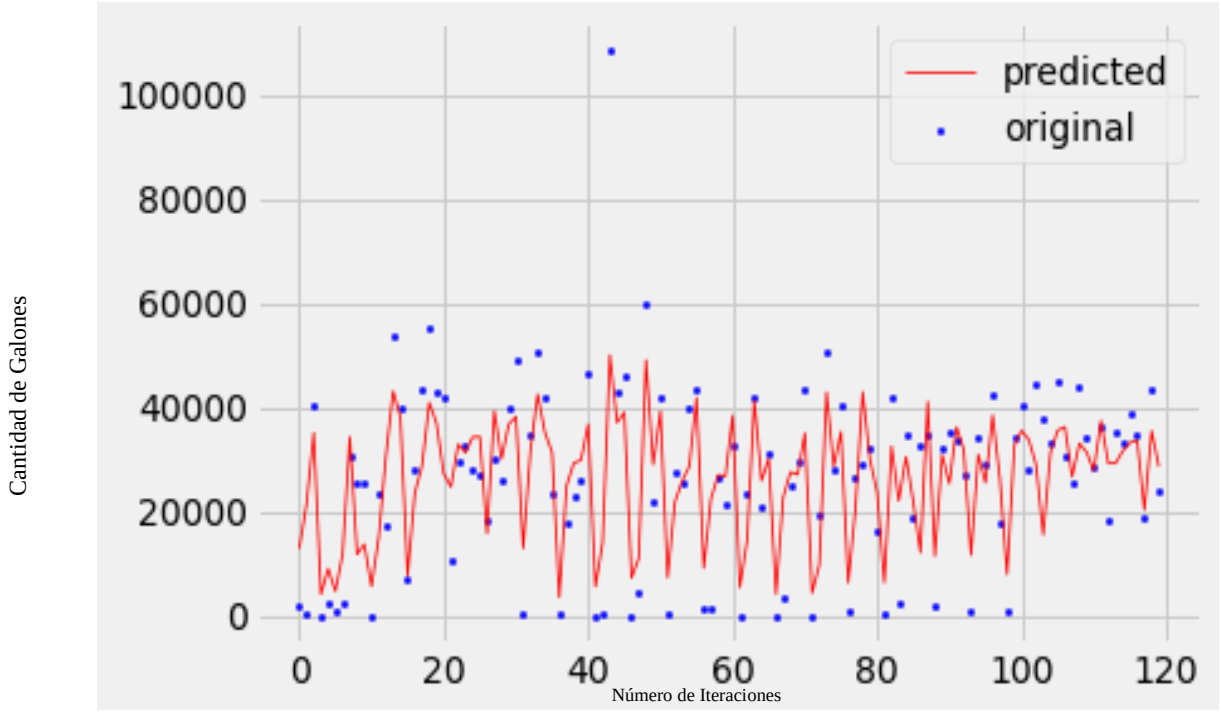

Figura 12. Gráfico Error de Entrenamiento- Random Forest

El error generado durante la etapa de prueba es del 25,3% para este modelo, siendo este muy similar al modelo de árbol de regresión, sin embargo, de acuerdo con el tipo de cálculo que realiza este modelo con la multiplicidad de árboles, genera una mejor relación entre los datos de entrada o entrenamiento, y los datos de salida de demanda futura de los taladros.

En la gráfica 13, se puede visualizar que se tienen menos datos sobre el eje X, y estos están siendo conectados con las líneas rojas, que permite identificar el acercamiento de los datos pronosticados con los datos de origen.

Error de Prueba % 25.3

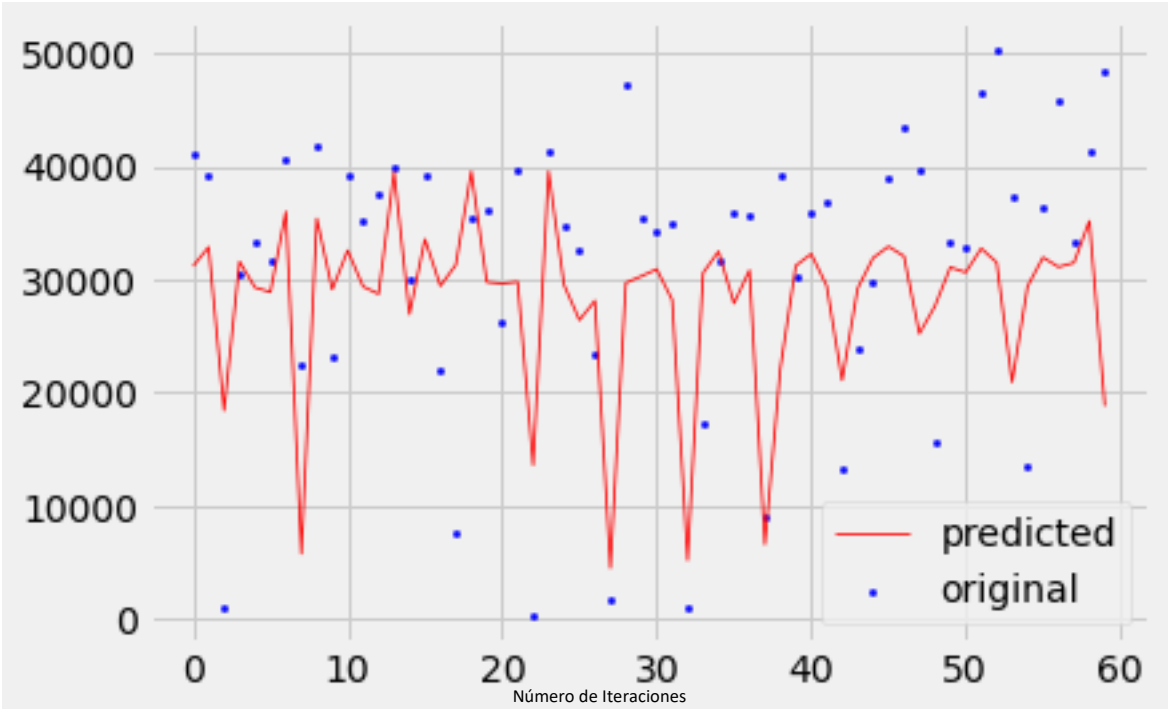

Figura 13. Grafico Erro de Prueba Random Forest

En las siguientes tablas se muestran los resultados de cada algoritmo empleado.

En las tablas 4 y 5 se aprecian los resultados de los algoritmos de clasificación empleados. Los datos disponibles nos permitieron realizar un pronóstico del consumo de biodiesel mes a mes por cada taladro, los mini gráficos nos muestran que el modelo de Random Forest es el que más se ajusta al comportamiento de los datos originales tomados para el Testeo. ver tabla 1.

Tablas 3. Datos para testeo del algoritmo

Datos para testeo													
Make	**2019-01**	**2019-02**	**2019-03**	**2019-04**	**2019-05**	**2019-06**	**2019-07**	**2019-08**	**2019-09**	**2019-10**	**2019-11**	**2019-12**	**Minigraficos**
R001	41109	31642	39234	39096	26313	32598	34251	35891	35916	38889	32786	36419	
R012	39196	40589	35224	22075	39756	23463	35011	35734	36931	43341	46533	45740	
R109	1147	22371	37540	7694	399	1792	979	8989	13289	39704	50336	33289	
R110	30499	41886	39952	35354	41356	47212	17390	39250	23924	15598	37386	41321	
R118	33229	23165	30015	36121	34674	35526	31736	30239	29841	33327	13580	48469	

Tablas 4. Resumen de resultados del algoritmo del árbol de decisión.

Prediccion arbol de Regresion													
Make	**2021-01**	**2021-02**	**2021-03**	**2021-04**	**2021-05**	**2021-06**	**2021-07**	**2021-08**	**2021-09**	**2021-10**	**2021-11**	**2021-12**	**Minigraficos**
R001	31.957	31.957	31.957	31.957	31.957	29.460	31.957	31.957	31.957	31.957	31.957	31.957	
R012	41.802	41.802	31.957	31.957	31.957	31.957	31.957	31.957	31.957	31.957	31.957	31.957	
R109	31.957	3.012	31.957	31.957	422	137	1.253	13.240	422	54.503	31.957	31.957	
R110	31.957	31.957	39.552	41.802	41.802	29.460	31.957	422	31.957	31.957	18.572	31.957	
R118	31.957	31.957	31.957	31.957	31.957	31.957	31.957	31.957	31.957	31.957	31.957	23.251	

Tablas 5. Resumen de resultados del algoritmo Random Forest

Prediccion ranfon forest													
Make	2021-01	2021-02	2021-03	2021-04	2021-05	2021-06	2021-07	2021-08	2021-09	2021-10	2021-11	2021-12	Minigrafico
R001	31.178	28.888	32.557	33.549	29.636	26.409	30.901	27.889	32.251	32.909	30.622	31.944	
R012	32.851	35.979	29.369	29.425	29.788	28.117	28.193	30.802	29.441	32.049	32.732	31.098	
R109	18.439	5.799	28.712	31.292	13.625	4.545	5.229	6.628	21.102	25.219	31.503	31.465	
R110	31.575	35.367	39.485	39.536	39.546	29.696	30.575	22.156	29.195	27.632	20.943	35.156	
R118	29.286	29.118	26.932	29.768	29.525	30.280	32.464	31.210	31.893	31.067	29.390	18.824	

Conclusiones

- ✓ Teniendo en cuenta los dos modelos de pronóstico utilizados, se concluye que el mejor es Random Forest con un error de prueba del 25,3%, utilizando muchas muestras varias veces aleatoriamente calculando promedio de promedios.

- ✓ La precisión en la predicción fue del 75% para el enfoque de clasificación, tanto con el algoritmo de árbol de regresión como con Random Forest

- ✓ Con estos resultados, la compañía tendrá mejor criterio de decisión en el momento de realizar las compras de combustible para los futuros periodos evitando sobre stock o faltantes de material, y reduciendo el costo del inventario al negociar tarifas de acuerdo con la demanda.

- ✓ Para el modelo de pronóstico, es importante no tener en cuenta todos los taladros ya que estos generan un sobre ajuste en los modelos.

Recomendaciones

- ✓ Optimizar los parámetros del árbol con una Grilla, función que selecciona los mejores parámetros del modelo, de forma automática.
- ✓ Teniendo en cuenta este modelo, la empresa podrá utilizarlo para pronosticar la demanda para diferentes materiales, es decir, el modelo es flexible para todos los materiales que utiliza la compañía.
- ✓ Se sugiere la implementación de un piloto en el taladro R110 para aplicar los datos e ir ajustándolo y mejorando los resultados.

- ✓ Teniendo en cuenta este modelo la empresa podrá utilizarlo para pronosticar la demanda de diferentes materiales, es decir, el modelo es flexible para todos los materiales que se utilizan en la compañía.

Referencias Bibliográficas

Aburto, L., & Weber, R. (2007). A Sequential Hybrid Forecasting System for Demand Prediction. *Machine Learning and Data Mining in Pattern Recognition. MLDM 2007. Lecture Notes in Computer Science, vol 4571.*, 518–532. https://doi.org/10.1007/978-3-540-73499-4_39

Arango, J., Giraldo, J. y Castrillón, O. (2013). Gestión de compras e inventarios a partir de pronósticos Holt-Winters y diferenciación de nivel de servicio por Clasificación ABC. *Scientia et Technica*, 18(4), 743-747. Recuperado de

Australia: Texts.

C.J. Madariaga Fernández, Y. O. Lao León, D.A. Curra Sosa, R. L. Martín Retos de la Dirección. Departamento de Desarrollo de Sistemas, Facultad de Ciencias Empresariales y Administración, Facultad de Ingeniería, Dirección de Ciencia Tecnología e Innovación, (2020) *Metodología para pronosticar demanda y clasificar inventarios en empresas comercializadoras de productos mayoristas.* Universidad de Holguín, Cuba.

Chicaiza Ipiales, J. A. (2019) *Diseño de un modelo para el pronóstico de la demanda para una empresa embotelladora de agua purificada mediante redes neuronales* [Tesis de grado publicada]. Universidad Técnica del Norte – Ibarra- Ecuador

D.W. Gareth James, R.T Trevor Hastie, Introduction to Statistical Learning, Springer Texts in Statistics, With Applications in R pag.82-150.

Ferbar, L. (2010). Joint optimisation of demand forecasting and stock control parameter. International Journal of Production Economics, 127(1), 173-179. Recuperado de http://doi.org/10.1016/j.ijpe.2010.05.009.

Fry, C. & Brundage, M. (2020). The M4 forecasting competition. A practitioner's view. International Journal of Forecasting, 36(1), 156-160. Recuperado de

Grimaldo, A. & Novak, J. (2020). Combining Machine Learning with Visual Analytics for Explainable Forecasting of Energy Demand in Prosumer Scenarios. *Procedia Computer Science*, 175, 525-532. Recuperado de http://doi.org/10.1016/j.procs.2020.07.074

Heizer, J., Render, B., & Munson, C. (2017). *Operations Management Sustainability and Supply Chain Management*. (PEARSON, Ed.) (12th ed.).

http://doi.org/10.1016/j.ijforecast.2019.02.013

http://repositorio.utn.edu.ec/handle/123456789/9372

http://www.revistas,utp.edu.co

Hu, Y., Sun, S., & Wen, J. (2014). Agricultural Machinery Spare Parts Demand Forecast Based on BP Neural Network. *Applied Mechanics & Materials*, (635–637), 1822–1825. Retrieved from http://10.0.15.188/www.scientific.net/AMM.635-637.1822

hybrid_model_for_time_series_forecasting/citation/download

Hyndman, R., & Athnasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd ed.).

J. A. Suárez Peña. (2019), Facultad de Ingeniería, Maestría en Ingeniería Industrial, *Modelo de aprendizaje automático para la predicción de la calidad del café.* Universidad Distrital Francisco José De Caldas

J. Garcia Caro, (2014), Administración de Empresas *Estructura y modelo de control de costos para una empresa dedicada a la perforación, mantenimiento y reacondicionamiento de pozos de hidrocarburo.* Colegio de Estudios Superiores de Administración –CESA.

J.P Ríos Ocampo, Y Olaya Morales, G. J. Rivera Leòn Revista Ingenierías Universidad de Medellín, (2017), scielo.org.co, Proyección de la demanda de materiales de construcción en Colombia por medio de análisis de flujos de materiales y dinámica de sistemas.

Khashei, M., Reza, S. & Bijari, M. (2008). A new hybrid artificial neural networks and fuzzy regression model for time series forecasting. *Fuzzy Set Systems*, 139(7), 769-786. doi: 101016/j.fiss.2007.10_11

Li, W., Luo, Y., Zhu, Q., Liu, J. & Le, J. (2008). Applications of AR-GRNN model for financial time series forecasting. *Neural Computing Applications*, 17(56), 441-448. doi: 10.1007/500521_17007_0131_9

Little, J. & Coughlan, B. (2008). Optimal inventory policy within hospital space constraints. *Health Care Management Science*, 11(2), 177-183. doi: 10.1007/510729_00890666_7

Makridakis, S., Spiliotis, E. & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLoS ONE*, 19(3), 1- 26. Recuperado de http://doi.org/10.1371/journal.pone.0194889

Medeiro, M., Tersvirta, T. & Rech, G. (2006). Building neural network models for time series: a statistical approach. *International Journal of Forecasting*, 25(1), 49-75.

Montemayor, E. (2013). *Métodos de pronósticos para negocios*. (Instituto Tecnológico y de Estudios Superiores & Monterrey, Eds.). México.

Pucuna Orozco, F. C. (2019), *Diseño de un plan estratégico de control de inventarios para la empresa Distribuidores de Industrias Nacionales Cía. Ltda.* [Tesis de grado publicada]. Universidad de Guayaquil. http://repositorio.ug.edu.ec/handle/redug/42586

Rivas, A. (2017). Procedimiento para el pronóstico de productos farmacéuticos mediante modelos de regresión (Tesis de maestría). Universidad de Holguín, Holguín, Cuba.

Saha, C., Lam, S. S., & Boldrin, W. (2015). Demand Forecasting for Server Manufacturing Using Neural Networks Demand Forecasting for Server Manufacturing Using Neural Networks State University of New York at Binghamton. *Proceedings of the 2014 Industrial and Systems Engineering Research Conference*, (June 2014).

Snyder, R., Koehler, A. & Ord, J. (2002). Forecasting for inventory control with exponential smoothing. *International Journal of Forecasting*, 18(1), 5-18. doi: 10.1016/50169_2070(01)109_1

Teunter, R., Syntetos, A., Babai, M. & Stephenson, D. (2011). Intermittent demand: Linking forecasting to inventory obsolescence. *European Journal of Operational Research*, 214, 606-615.

Wang, L., Zou, H., Su, J., li, L. & Chaudhry, S. (2013). An ARIMA-ANN hybrid model for time series forecasting. *Systems Research and Behavioral Science*. 30(3), 244- 259. DOI:

10.1002/sres.2179. Recuperado de https://www.researchgate.net/publication/263454208_An_ARIMAANN_

Xu, S. & Chan, H. K. (2019). Forecasting Medical Device Demand with Online SearchQueries: A Big Data and Machine Learning Approach. Procedia Computer Science, 39, 32-39. Recuperado de http://doi.org/10.1016/j.promfg.2020.01.225