# Análisis de proyección de ingresos económicos de buques tanqueros mediante el uso de análisis de datos en una compañía naviera

Pulla Mendieta, Andrés Santiago

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

Maestría en Gestión de Sistemas de Información e Inteligencia de Negocios

Trabajo de titulación, previo a la obtención del título de Magíster en Gestión de Sistemas Información e Inteligencia de Negocios

Msc. Mazón Quinde, Karina Inabel

14 de agosto del 2020

URKUND

## Urkund Analysis Result

Analysed Document: Tesis-Andres Pulla Mendieta_v_8_Rev Ago 26 2020 URKUND.docx (D78750718)

Submitted: 9/8/2020 4:28:00 AM

Submitted By: Gualotuña Alvarez Tatiana Marisol, tmgualotunia@espe.edu.ec

Significance: 0%

**Mazón Quinde, Karina Inabel**

**DIRECTOR**

**VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA**

**CENTRO DE POSGRADOS**

# CERTIFICACIÓN

Certifico que el trabajo de titulación, "**Análisis de proyección de ingresos económicos de buques tanqueros mediante el uso de análisis de datos en una compañía naviera**" fue realizado por el **Pulla Mendieta, Andrés Santiago** el mismo que ha sido revisado y analizado en su totalidad, por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 18 de agosto 2020

Mazón Quinde, Karina Inabel

Director

**C.C.: 0604597112**

**VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA**

**CENTRO DE POSGRADOS**

## RESPONSABILIDAD DE AUTORÍA

Yo **Pulla Mendieta, Andrés Santiago**, con cédula de ciudadanía n° 01092711090, declaro que el contenido, ideas y criterios del trabajo de titulación: **Análisis de proyección de ingresos económicos de buques tanqueros mediante el uso de análisis de datos en una compañía naviera** es de mí autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

**Sangolquí, 30 de abril 2020**

.....

Pulla Mendieta, Andrés Santiago

**C.C.:** ...01902711090

**VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA**

**CENTRO DE POSGRADOS**

**AUTORIZACIÓN DE PUBLICACIÓN**

Yo **Pulla Mendieta, Andrés Santiago** autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **Análisis de proyección de ingresos económicos de buques tanqueros mediante el uso de análisis de datos en una compañía naviera** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

**Sangolquí, 30 de abril 2020**

.....

Pulla Mendieta, Andrés Santiago

C.C.:...0102711090.

**Dedicatoria**

A mi esposa Evelin Gavilanes, por su apoyo incondicional, ánimo y comprensión para este proyecto.

A mis hijos Emilio y Gabriel que supieron entender el compromiso y aceptar permitirme este tiempo para estudiar, siempre les compensare cada día de mi vida.

A mis amigos de la universidad, por haber puesto la alegría y el entusiasmo en mí para poder terminar este largo camino de preparación profesional.

A mis padres Carmen y Arturo que siempre supieron darme los valores y fortaleza necesarios para terminar todos los proyectos.

**Agradecimiento**

Agradezco a la Universidad de las Fuerzas Armadas ESPE por abrirme las puertas y permitirme estudiar la esta maestría, así como a todos los docentes que participaron activamente en el desarrollo de este proyecto, y pudieron transmitir todo su conocimiento y poder guiar a este estudiante.

Agradezco a mi tutor de Tesis la Msc. Karina Mazón por su valioso tiempo, paciencia, conocimiento y brindarme la oportunidad de ejecutar un proyecto a su nombre.

De igual manera agradezco a todos mis compañeros y amigos, en especial a mi compañera de vida, porque su felicidad me motivo a culminar este proyecto.

**Índice de contenido**

## Índice de tablas

**Índice de figuras**

**Resumen**

Las compañías navieras alrededor del mundo están compuestas con una flota de buques tanqueros para la transportación de diferentes productos derivados del petróleo y realizar el transporte desde un puerto a otro, estudios realizados por la compañía Lloyd's Register compañía que cuenta con el aval de la OMI (Organización Marítima internacional) define que en el año 2018 la cantidad de buques tanqueros es aproximadamente de 8300 unidades en el mundo, definió también que, el crecimiento de buques entre 2017 a 2018 fue de un 4.1%, el cual representa un crecimiento en un numero de 340 unidades/año, lo que quiere decir, que la inversión de cada una de estas unidades tomando en consideración el costo promedio de un buque nuevo es de USD 55'000.000, representara un valor de 18.700 millones de dólares por año; el análisis de ingresos por renta de buques para inversión es un factor importante para las compañías navieras. Por lo que, este proyecto analiza la proyección de ingresos que un buque generaría en periodo de tiempo, haciendo uso de técnicas de minera de datos para establecer una proyección del mercado de precios de buques tanqueros. Así se definiría una herramienta que brinde soporte a la toma de decisiones, haciendo un análisis de los precios del mercado de buques de años anteriores y sometido a técnicas de análisis de datos, se podrá proyectar el impacto de la inversión de una compañía naviera, como apoyo a la toma de decisiones, la metodología para el desarrollo del proyecto será CRISP-DM.

PALABRAS CLAVE

- **PREDICCIÓN**
- **BUQUE TANQUERO**
- **RENTA DE BUQUES**
- **ANALISIS DE DATOS**

**Abstract**

Shipping companies around the world are made up of a fleet of tanker ships for the transportation of different oil products and to carry them from one port to another, studies carried out by the Lloyd's Register company, which has the endorsement of the IMO (International Maritime Organization) defines that in 2018 the number of tanker ships is approximately 8300 units in the world, also defined that, the growth of ships between 2017 and 2018 was 4.1%, which represents a growth in a number 340 units / year, which means that the investment of each of these units taking into account the 55,000,000, representing a value of 18.7 billion dollars per year; The analysis of income from rental of vessels for investment is an important factor for shipping companies. Therefore, this project analyzes the projection of income that a ship would generate in a period of time, using data mining techniques to establish a projection of the market for tanker ship prices around the world. This would define a tool that supports decision-making, making an analysis of the market prices of ships from previous years and subjected to advance data analysis techniques, to make a forecast of the impact of a shipping company's investment, as support to the decision-making, the methodology for the development of the project will be CRISP-DM.

KEY WORDS:

- **PREDICTION**
- **TANKER VESSEL**
- **CHARTERING**
- **DATA ANALYSIS**

## Capitulo I. Introducción

Para iniciar la creación del modelo de predicción es necesario establecer un punto de referencia del negocio marítimo y como se manejan los datos económicos dentro de una empresa naviera como en este caso EP FLOPEC, y la importancia que tienen los datos económicos sobre la renta de buques tanqueros.

Antecedentes

El éxito en el negocio naviero de hoy necesita establecer una visión precisa de la operación de la flota de buques petroleros en el mercado de transporte de crudo para poner en marcha los planes de distribución y crecimiento de la flota dentro de las compañías navieras, la proyección del crecimiento de la flota es una estrategia que conforme al establecimiento del mercado se convierte en una necesidad y es parte del modelo de negocio debido a que las unidades de transporte conforme pasan los años de vida útil del navío estos reducen su capacidad de ser comercializados.

Debido al alto nivel de inversión, el análisis de la información se ha tornado un proceso importante y ante esta necesidad, se han desarrollado algunos estudios alrededor de este tema; esto debido a que es un concepto que ha evolucionado con intensidad y que aporta datos muy importantes en la toma decisiones de una inversión que supera las decenas de millones de dólares. Y que tiene una tasa de crecimiento importante año a año.

Grandes empresas navieras como son Teekay Corp. (TK), Frontline Ltd. (FRO), Tsakos Energy Navigation, Nordic American Tanker (NAT), Ship Finance International Limited, y DHT Holdings Inc. están implementando proyectos de analítica avanzada de datos, relacionados al conocimiento de los precios de renta del buque en el mercado mundial. Los datos del mercado y sus precios son el activo principal de estas compañías, por tal motivo han optado por formar un

equipo centralizado de ciencia de datos organizado para compartir análisis, mejores prácticas y establecer proyecciones de inversión, basada en datos para la toma de decisiones en toda la organización.

En la actualidad las grandes compañías navieras como Maersk Tanker contratan servicios de compañías que brindan soluciones de BI para el análisis de sus datos en el año 2012 esta compañía contrato a la compañía KAPACITY como proveedor de una solución de BI. En la cual se establecieron el análisis de los datos focalizados en las siguientes áreas de la compañía como: costos operativos de las naves, operaciones comerciales, recursos humanos. Los elementos de la aplicación de BI estuvieron basados en Microsoft SQL Server 2014 Enterprise Data Warehouse, desarrollados de acuerdo a las mejores prácticas de Kapacity. El Data Warehouse cuenta con cubos OLAP multidimensionales, BIML para la generación de paquetes, y toda una aplicación que se alimenta de datos de forma automática, como herramienta de reportera se usó Microsoft SharePoint 2013 con Performance Point, Reporting Services, y Excel Services. El alcance de Maersk Tankers' es la integración de reportes móviles y el uso de la herramienta Microsoft Power BI como parte de la solución.

Los datos dentro de la solución son actualizados diariamente haciendo el proceso del negocio mucho más rápido y confiable. En Sudamérica la compañía naviera que marca la pauta en el transporte de crudo y que se encuentra con base en Ecuador es la FLOTA PETROLERA ECUATORIANA, FLOPEC EP compañía que pertenece al gobierno del Ecuador, y que cuenta con una flota de 7 buques petroleros como armados y 30 buques rentados para la transportación del crudo del Ecuador a todas partes del mundo.

Al ser esta compañía una de las más representativas en la zona cuenta con un departamento de inteligencia de mercado que se beneficiara de proyectos de analítica avanzada;

el análisis de los datos del mercado para establecer los ingresos económicos brutos de buques tanqueros es muy importante para sus operaciones futuras.

En el Ecuador, específicamente en el sector naviero, no existen proyectos que estén en desarrollo referentes a analítica avanzada cuya arquitectura, herramientas y modelos analíticos hayan sido publicados y expuestos, esto se debe a que en el país existe tan solo una compañía de transporte de crudo de tráfico internacional, por lo que, las técnicas de analítica avanzada no han sido aplicadas para el análisis de ingresos por renta de buques tanqueros en una flota naviera y los datos necesarios para el análisis son de carácter privado.

El presente proyecto, aportará con una propuesta de mejora para la proyección de los ingresos económicos para la adquisición de nuevos buques petroleros enfocado al modelo de negocio de armador, esta compañía no ha profundizado con estudios de este tipo, además el estudio puede ser considerado como un prototipo modelo considerando y ajustándolo a factores propios de otras compañías navieras.

**El problema de investigación.**

***Contexto del Problema***

En 1972 la Comandancia de Marina del Ecuador convocó a un concurso internacional para seleccionar una compañía que, junto con Transnave, conformase una compañía de economía mixta para el transporte marítimo del petróleo ecuatoriano.

Para esa época, el transporte internacional de hidrocarburos estaba totalmente centralizado en las empresas transnacionales del petróleo que habían acaparado el círculo total del negocio. El 14 de septiembre de 1972 se firmó un contrato por diez años con Kawasaki Kisen Kaisha, empresa ganadora de la licitación.

El 26 de marzo se establece legalmente la compañía de economía mixta FLOTA PETROLERA ECUATORIANA con la participación del 55% de sus acciones a nombre de Transnave y el 45% restante a la empresa japonesa. Ecuador fue el último país en ingresar a la Organización de Países Exportadores de Petróleo (OPEP), pero el primero en conformar una flota nacional para la exportación de su crudo.

FLOPEC inició sus operaciones con dos buques propios comprados a la compañía Gulf en 1973, estos buques fueron llamados Napo y Pastaza. Posteriormente, se unieron los buques tanques Ecuador de Transnave y el Zamora, construido en astilleros japoneses.

En 1978 se vendieron los buques Napo y Pastaza por haber concluido su vida útil, esto obligó a la empresa a fletar buques extranjeros para cubrir la demanda de CEPE. La asociación con la empresa naviera Kawasaki Kisen Kaisha otorgó experiencia al Ecuador para el manejo del transporte de hidrocarburos; en menos de seis años se dio paso a la nacionalización de la empresa a cargo de FLOPEC y de la Armada Nacional.

Durante este tiempo, FLOPEC llegó a transportar el 52 % del crudo nacional con buques propios y arrendados, entró al negocio naviero y puso a sus buques a transportar el crudo ecuatoriano. Después de 38 años la FLOPEC se ha convertido en la flota más importante del Pacífico oriental con 7 buques tanques propios de alta tecnología: Zaruma, Pichincha, Cotopaxi Chimborazo, Zamora Santiago y Aztec y 23 buques fletados según sus necesidades.

Doscientos sesenta hombres, todos ecuatorianos (as), forman parte de la tripulación de los 7 buques convirtiéndonos en el orgullo del Ecuador en los mares del mundo. Y en tierra, la empresa cuenta con alrededor de ciento diez funcionarios que contribuyen día a día para engrandecer a la primera empresa naviera del país.

El patrimonio de FLOPEC ha evolucionado en forma altamente positiva en los últimos diez años. Actualmente es una empresa líder en el campo del transporte marítimo internacional en América Latina y ha sido pionera en la consecución de Certificación Internacional de Protección Marítima, Certificación del Código Internacional de Gestión de Seguridad (ISM), Código Internacional para la Protección de los Buques y las Instalaciones Portuarias (ISPS), Normas de Calidad ISO 9001 y Norma Ambiental ISO 14001.

Actualmente FLOPEC cuenta con un proyecto en ejecución para la renovación de la flota y requiere recurrir a un acertado y preciso análisis del mercado, para definir la mejor alternativa de adquisición de tipo de buque tanquero tomando en consideración las rutas más convenientes y segmentos del mercado más oportunos.

Por lo que en la gráfica No. 1 se muestra la organización de la gestión comercial.

**Figura 1**

*Gestión de la Gerencia Comercial*


*Nota.* Sistema de Gestion Documental de EP FLOPEC, 2015.

En la Figura No. 1 se presenta el esquema que es utilizado por la Gerencia Comercial de EP FLOPEC para establecer un viaje de un buque de la flota, donde inicia con una planificacion anual de los viajes de los buques de la flota, a continuación el departamento de Fletamento se concretan los contratos de viaje planificados, en estos contratos de viaje se definen las condiciones y precios de renta de los buques tanqueros, una vez firmado el contrato la Gerencia de operaciones en conjunto con los buques de la flota ejecuta las ordenes de viaje y condiciones de operación del buque, una vez finalizada la operación del buque el departamento de reclamos y demorajes se encarga de conciliar el contrato con las operaciones realizadas por el buque, para entregar esta informacion a la Gerencia Financiera.

**Planteamiento del Problema**

El principal inconveniente que actualmente presenta la empresa FLOPEC EP es la predicción de tasa de los precios por renta de buques tanqueros y los ingresos económicos de nuevos buques tanqueros basándose en los precios del mercado.

Las predicciones de ingresos económicos para un buque petrolero no han sido establecidas mediante técnicas de análisis de datos, y actualmente este análisis es realizado de acuerdo al promedio de precio del mercado de años anteriores; este proceso es realizado calculando y registrando en hojas de calculo, por lo que la curva de ingresos proyectada será una recta constante, posteriormente esto genera un desajuste de acuerdo a la variación de precios en el mercado. El uso de herramientas tecnológicas de inteligencia de negocios adquiridas; a pesar que estas proyecciones se basan en el histórico de precios del mercado de buques, son débiles ya que no se ajustan al comportamiento del mercado y no han sido validados en un sistema o herramienta tecnológica. Esta situación podría causar un desacierto en la adquisición de buques tanqueros, además genera un factor adicional para de no lograr las metas de ingresos proyectados para los próximos años.

Las compañías navieras no cuentan con un sistema implementado con una arquitectura probada que esté orientado a la proyección de los ingresos por renta de buques tanqueros mediante el uso de herramientas y modelos analíticos, en la compañía EP FLOPEC en la actualidad no se utilizan este tipo de herramientas; por lo que incide a la planificación de ingresos anuales por renta de buques.

Debido a la deficiencia identificada que presenta la empresa naviera EP FLOPEC nace la viabilidad del presente trabajo de investigación, y permitirá conocer la proyección de precios del mercado de renta de buques para poder definir la inversión en un buque tanquero, consiguiendo así optimizar los precios de renta de su flota y obteniendo una ventaja sobre el mercado internacional de buques tanqueros, pero por sobre todo contribuirá que las empresa naviera EP FLOPEC continúe liderando el mercado de transporte de crudo.

**Objetivos.**

**Objetivo General.**

Analizar y diseñar una propuesta de mejora para el proceso de proyección de ingresos para la adquisición de nuevos tipos de buques tanqueros.

**Objetivos Específicos.**

OE1: Entender las arquitecturas, herramientas y modelos analíticos más utilizados para la proyección de ingresos por renta de un buque petrolero, para este objetivo específico se usará la metodología de revisión sistemática de literatura.

OE2: Mejorar la proyección de los ingresos en la empresa naviera, que permita tener una visión a corto plazo de los ingresos por renta de buques, a través del uso de una arquitectura definida, herramientas y modelos analíticos-predictivos, para este objetivo específico se utilizará la metodología de proyectos CRISP-DM.

OE3: Validar la propuesta de mejora para la proyección de ingresos por renta de buques, a través del uso de técnicas de validación implementadas en minería de datos, para determinar el nivel de confianza en un resultado predictivo, para el desarrollo de este sistema se usará la metodología de investigación.

**Hipótesis**

La tecnología de minería de datos mediante la analítica avanzada de datos puede mejorar la predicción de ingresos por renta de buques tanqueros una vez que se propone la adquisición de buques tanqueros para la planificación de ingresos.

Señalamiento de variables

**Variable dependiente**

Predicción de precios del mercado de buques tanqueros para la planificación de ingresos por adquisición de buques.

**Variable independiente**

Modelo de Análisis predictivo para predicción de los precios de renta del mercado de transporte de buques tanqueros.

Para la demostración de la hipótesis planteada se considera el método deductivo aplicado a los datos generados por compañías privadas dedicadas a establecer los precios del mercado de renta de buques que transportan crudo alrededor del mundo y en rutas definidas para diferentes tipos de buques, los precios de renta de buques son adquiridos por las compañías navieras con el propósito de establecer los precios de los buques de su flota, y definir el precio de renta en los contratos de viaje.

Los datos que fueron adquiridos por EP FLOPEC servirán para evaluar los resultados de la predicción a través del cálculo de la precisión del pronóstico de renta de buques utilizando la evaluación estadística. Para la construcción del modelo, se usarán los datos disponibles del histórico de la compañía naviera, del conjunto de datos se usarán sus tres cuartas partes y la tercera parte sobrante se usará para validar el modelo.

**Justificación, Importancia y Alcance.**

La planificación de una inversión tomando en cuenta los ingresos que genera un buque tanquero en la industria marítima es una necesidad que debe ser satisfecha. Se puede decir que el objetivo de la planeación es tratar de prever lo que sucederá en el futuro en base a una recopilación de hechos y eventos ocurridos con anterioridad.

La justificación del tema de estudio está basada en la necesidad que tiene la empresa naviera de conocer lo más preciso posible a cuánto ascenderán las tasas de ingreso por renta de buques y calcular cuánto será su ingreso por la adquisición de nuevos buques para la transportación de crudo. De esta forma, lograr una mejor proyección de gastos operativos, minimizar riesgos, y otras ventajas que conlleva una buena planificación. Si no se realiza este estudio la empresa continuará realizando la proyección de los ingresos mediante cálculos en Excel, es decir manejando un análisis plano no adecuado de los datos y además un uso deficiente de las herramientas tecnológicas de inteligencia de negocios, por consecuencia, podría causar una desestimación de los ingresos que cause pérdidas económicas en los próximos años.

El alcance de la presente investigación es realizar un estudio comparativo de las diferentes arquitecturas, herramientas y modelos analíticos implementados, con el fin de proponer una arquitectura fácil, entendible y realizable en una empresa naviera la que ayudará a mejorar la planificación de ingresos e impactará en la utilidad de la empresa.

Es importante considerar que en empresas navieras en donde no se ha explorado a detalle un análisis de datos e implementación de soluciones de inteligencia de negocio podría poner en riesgo la inversión necesaria para la renovación de su flota. La presente propuesta podrá ser considerada como una arquitectura o prototipo modelo de análisis para la proyección de la precios del mercado de renta de buques tanqueros considerando factores propios de la realidad a nivel mundial en el mercado de buques tanqueros, que pueden servir de base para el análisis en otras compañías del ámbito mercante, contribuyendo al marco referencial en investigaciones de proyectos de minería de datos y big data.

**Preguntas de Investigación**

Para el presente estudio, se ha propuesto la definición de una arquitectura, herramienta y modelo analítico para la proyección de precios de renta de buques en el mercado de buques tanqueros y requiere de un análisis, definiendo un conjunto de pasos por realizar enfocados a conseguir un objetivo específico, con este propósito se motivó las siguientes preguntas de estudio para cada objetivo:

OE1: Entender las arquitecturas, herramientas y modelos analíticos más utilizados para la proyección de ingresos por renta de un buque petrolero, para este objetivo específico se usará la metodología de revisión sistemática de literatura.

RQ1.1: ¿Qué arquitecturas y herramientas analíticas son las más utilizadas para proyectos predictivos?

RQ1.2: ¿Cuáles son las arquitecturas analíticas implementadas como casos de éxito a nivel mundial?

RQ1.3: ¿Cuáles son los modelos o algoritmos más utilizados en las herramientas analíticas implementadas en el sector precios del mercado de petróleo?

OE2: Mejorar la proyección de los ingresos en la empresa naviera, que permita tener una visión a corto plazo de los ingresos por renta de buques, a través del uso de una arquitectura definida, herramientas y modelos analíticos-predictivos, para este objetivo específico se utilizará la metodología de proyectos CRISP-DM.

RQ2.1: ¿Cuál es la arquitectura, herramientas y modelos analíticos más adecuados para la empresa naviera?

RQ2.2: ¿Es posible generar un modelo predictivo rápido, fácil y realizable con las herramientas analíticas seleccionadas para la empresa naviera?

OE3: Validar la propuesta de mejora para la proyección de ingresos por renta de buque, a través del uso de técnicas de validación implementadas en minería de datos, para determinar el nivel de confianza en un resultado predictivo.

RQ3.1: ¿Es viable y funcional validar el modelo propuesto con técnicas de evaluación de minería de datos?

RQ3.2: ¿Se puede determinar un nivel de confianza en los resultados que permita demostrar una mejora en los mismos?

RQ3.3: ¿Cuál es el margen de error que se debe considerar al implementar modelos de analítica predictiva?

**Metodología de la Investigación**

Para la ejecución de este proyecto se utilizará un caso de estudio basado en la compañía Naviera Empresa Publica Flota petrolera ecuatoriana, y se usará la metodología CRISP-DM. A continuación, se establece las actividades o acciones que se realizaran para el cumplimiento de los objetivos.

**Figura 2**

*Resumen de Objetivos*

| OBJETIVOS | | ACCIONES |
|---|---|---|
| 1 | Entender las arquitecturas, herramientas y modelos analíticos más utilizados para la proyección de ingresos por renta de un buque petrolero, para este objetivo específico se usará la metodología de revisión sistemática de literatura | Revisión de literatura sobre precios de buques tanquero usando SMS / SLR |
| | | Revisión de literatura sobre técnicas de minería de datos para la gestión de precios de buques tanqueros usando SMS / SLR |
| 2 | Mejorar la proyección de los ingresos en la empresa naviera, que permita tener una visión a corto plazo de los ingresos por renta de buques, a través del uso de una arquitectura definida, herramientas y modelos analíticos-predictivos, para este objetivo específico se utilizará la metodología de proyectos CRISP-DM | Evaluar algoritmos y herramientas de minería de datos usando estudios económicos |
| | | Evaluar los algoritmos con métricas de anélisis de datos y realizar comparativas |
| | | Se revisará cual algoritmo se adaptara mejor a nuestro requerimiento |
| 3 | Validar la propuesta de mejora para la proyección de ingresos por renta de buque, a través del uso de técnicas de validación implementadas en minería de datos, para determinar el nivel de confianza en un resultado predictivo | Analizar las fuentes de datos con que cuenta la organización para la predicción de precios del mercado de crudo |
| | | Realizar análisis con graficos de dispersion para identificar valores atípicos. |

**Metodología de investigación para prueba de hipótesis**

Para la prueba de la hipótesis se realizará la metodología de encuesta mediante un análisis de cuadros comparativos entre las variables dependientes comparadas por tipo de ruta para verificar que la hipótesis de cuál es el tipo de ruta que genere mayor ingreso.

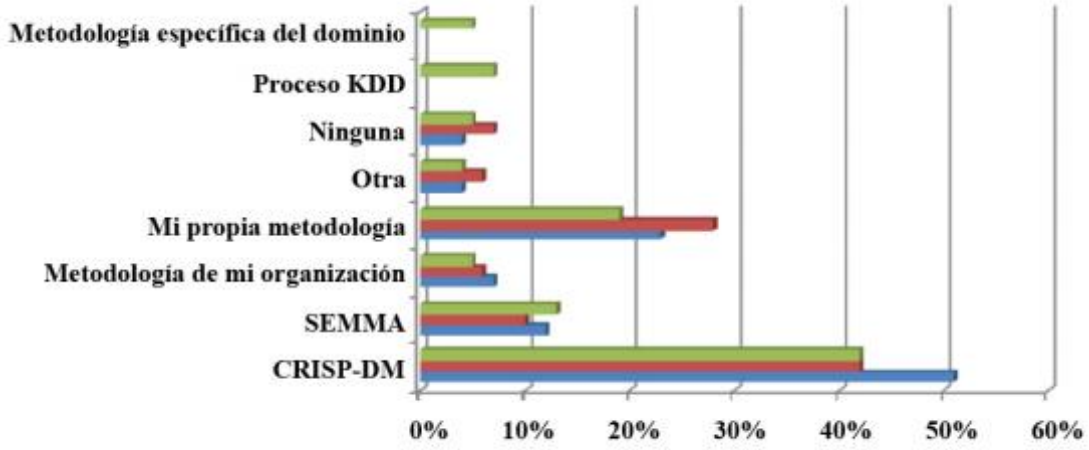
**Metodología de investigación para desarrollo del proyecto**

Para implementar una tecnología en un negocio es necesario establecer una metodología para el desarrollo de un proyecto tecnológico, la aplicación del método suele venir de las experiencias propias y también de los procedimientos estándar más conocidos. En el caso de los proyectos de implementación de minería de datos una de las metodologías que ha tenido más apoyo de las empresas privadas y organismos públicos es CRISP-DM, como se puede observar en la siguiente gráfica (figura 3), publicada en kdnuggets.com, y que representa el grado de utilización de las principales guías de desarrollo de proyectos de minería de datos según las encuestas realizadas.

Como se puede observar CRISP-DM ha experimentado un ligero descenso en los últimos años, pero sigue siendo la más empleada de las distintas metodologías, además esta metodología CRISP-DM se ajusta de mejor manera al tipo de investigación cuantitativa que se plantea en este proyecto de titulación.

**Figura 3**

*Metodologías aplicadas*




*Nota.*: Consumer Risk Analytics, Febrero 2014

CRISP-DM incluye un modelo y una guía, estructurados en seis fases, algunas de las cuales son bidireccionales, es decir que de una fase en concreto se puede volver a una fase anterior para poder revisarla, por lo que la sucesión de fases no tiene porqué ser ordenada desde la primera hasta la última. En la figura 4 se puede observar las fases en las que se divide CRISP-DM (Chapman, 1999) y las posibles secuencias a seguir entre ellas.

## Capitulo II. Revisión de literatura

En este capítulo se presenta la revisión de literatura para determinar las arquitecturas, herramientas y modelos analíticos más utilizados para la proyección de ingresos por renta de buques con base en los precios de renta de un buque petrolero, así también se presenta tres casos ejemplo donde se aplican arquitecturas ejemplo para el uso de analítica avanzada.

### Técnicas de analítica avanzada

Existen tres técnicas de analítica avanzada; descriptiva, predictiva y prescriptiva, la técnica de predicción que como su nombre lo indica predice el comportamiento de los datos, la técnica descriptiva detalla el comportamiento de la información en el presente y la exploratoria analiza el comportamiento histórico de los datos. (Davenport and Kim, 2013; Evans and Lindner, 2012 ).

**Figura 4**

*Clasificación de algoritmos*


*Nota.* "Top Algorithms and Methods Used by Data Scientists", Enero, 2014

Por lo que debido a lo expuesto en el planteamiento del problema y lo analizado en el Capítulo I, y de acuerdo a los objetivos del proyecto se selecciona la técnica predictiva de datos para establecer el comportamiento futuro de los precios del mercado para renta de buques tanqueros en una compañía naviera. Mediante el uso de una técnica de analítica predictiva.

Analítica predictiva, utiliza técnicas estadísticas para estimar qué comportamiento o resultado es probable. Es decir, intenta proyectar lo que puede ocurrir prediciendo situaciones futuras. Por tanto, su naturaleza es probabilística, porque nos dicen cuál es la probabilidad de que algo suceda. Y lo hace tratando de encontrar relaciones y patrones entre variables utilizando información actual e histórica para extraer conclusiones y predecir el futuro.(Evans and Lindner, 2012 )

Los modelos predictivos tienen como base un conjunto de datos que describen el comportamiento histórico de una variable, para nuestro caso de estudio el conjunto de datos será el precio de la renta de buques petroleros, donde existe una variable que es el valor en dólares en contraste con el tiempo.

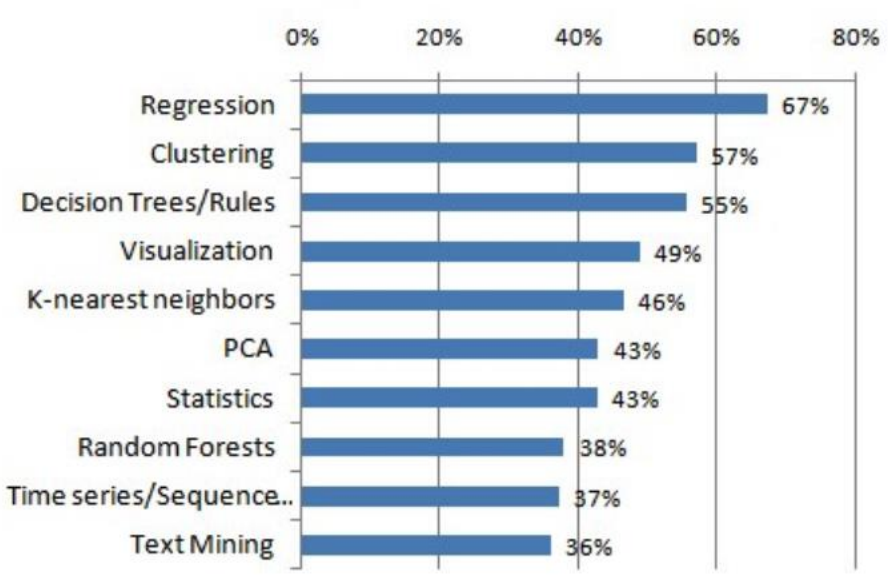
Entonces, en conclusión, la técnica analítica que se define para ser usada en este proyecto será la técnica predictiva, técnica que por los argumentos expuestos anteriormente se ajusta a los objetivos presentes en este proyecto de análisis avanzado de datos.

**Análisis de algoritmos predictivos.**

En la actualidad existen varios algoritmos predictivos que soportan la aplicación de la analítica predictiva, esto con la finalidad de obtener un algoritmo matemático de comportamiento de los datos, los algoritmos predictivos más usados conforme a la gaceta electrónica Kdnuggets.com donde se establecen una lista de los algoritmos por los analistas de datos haciendo uso de la técnica de encuestas, se enumeran a continuación: (Gregory Piatetsky, 2016)

Figura 5

*Cuadro de los 10 algoritmos más usados por los científicos de datos*




*Nota.*: ("Top Algorithms and Methods Used by Data Scientists", Enero, 2014)

A continuación, se describen los principales algoritmos analíticos, los modelos de regresión están basados en el establecimiento de una ecuación matemática, como modelo con el objeto de representar las interacciones entre las diferentes variables del evento físico que se pretende predecir, a continuación, se describen varios modelos que son utilizados en este grupo.

Existen varios algoritmos que se derivan de este como son:

**Regresión Lineal**

El modelo de regresión lineal relaciona la variable dependiente o de respuesta con el conjunto de variables independientes o predictores. Esta relación es expresada como una ecuación que predice la variable de respuesta como una función lineal de los parámetros. (Carlos Espino Timon, enero 2017)

El objetivo de la regresión es seleccionar los parámetros del modelo que minimizan la suma de los errores al cuadrado. Esto se conoce como estimación de mínimos cuadrados ordinarios y los resultados en las mejores estimaciones lineales no sesgadas de los parámetros si y sólo si se satisfacen las suposiciones de Gauss-Markov.

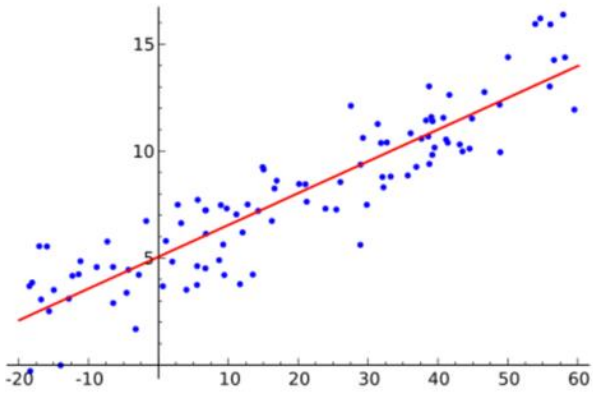
Una vez que se ha estimado el modelo, es necesario saber si las variables predictores pertenecen al mismo. Para ello podemos comprobar la significancia estadística de los coeficientes del modelo que pueden medirse utilizando el estadístico “t”. Esto equivale a probar si el coeficiente es significativamente diferente de cero.

El objetivo de la regresión es seleccionar los parámetros del modelo que minimizan la suma de los errores al cuadrado. Esto se conoce como estimación de mínimos cuadrados ordinarios y los resultados en las mejores estimaciones lineales no sesgadas de los parámetros si y sólo si se satisfacen las suposiciones. (Carlos Espino Timon, enero 2017).

En la siguiente imagen se puede ver un ejemplo de regresión linear simple.

Figura 6

*Ejemplo de regresión simple*



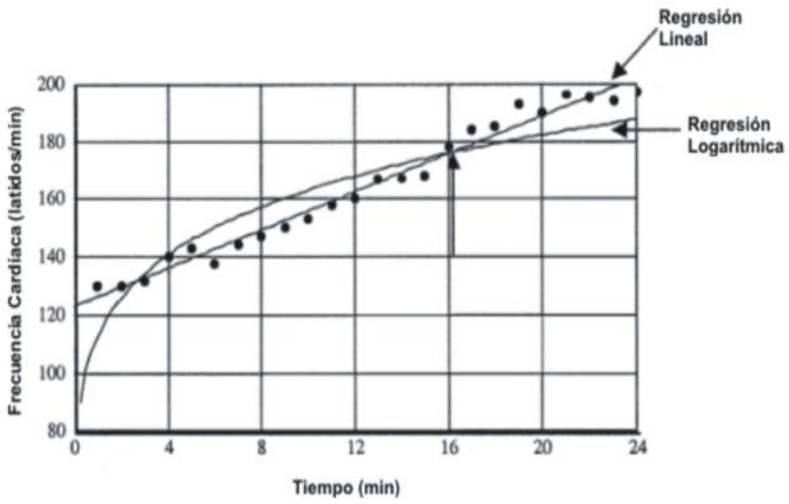
*Nota.* “Linear Regression”, 2020

**Regresión logarítmica**

Existen otros modelos de regresión como una alternativa cuando el modelo lineal no logra un coeficiente de determinación apropiado, o cuando el fenómeno en estudio tiene un comportamiento que puede considerarse potencial o logarítmico. La forma más simple de tratar de establecer la tendencia es a través de un diagrama de dispersión o nube de puntos. ( Bowerman Bruce, enero 2007), (José Manuel Molina López, 2006)

**Figura 5**

*Ejemplo de regresión logarítmica*




*Nota.* Wyatt et al., 2005

**Arboles de clasificación ( clustering )**

Conforme a lo citado por (George Seif, Febrero 2018) en su artículo para Clustering; es la tarea de dividir la población de datos in un números de grupos, estos grupos tienen las mismas características, por lo que se podría decir que se segregan los datos por sus características; es una técnica de machine learning donde se involucran los datos en un mismo grupo, se usa este

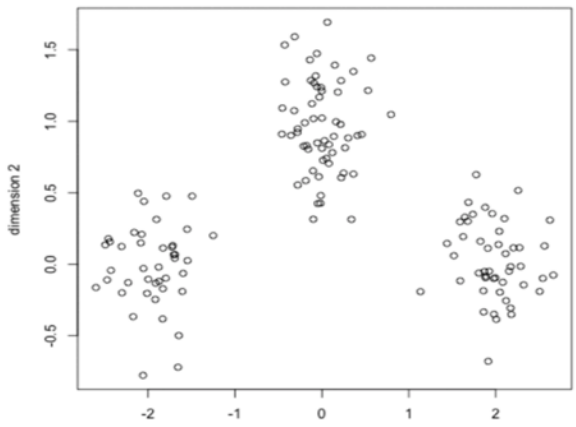
algoritmo para clasificar cada punto en un grupo específico, en teoría este es un algoritmo del tipo no supervisado.

Clustering tiene un largo número de aplicaciones en un número indefinido de dominios, algunos de las populares aplicaciones de clustering se enumera a continuación:

- Segmentación de mercados
- Análisis de redes sociales
- Agrupación de resultados de búsquedas
- Imágenes medicas
- Segmentación de imágenes
- Detección de anormalidades.

**Figura 6**

*Ejemplo de Clustering*




*Nota.* George Seif, Febrero 2018

**Decision trees/rules ( Gradient Boosted Trees )**

Un árbol de decisión es, para quien va a tomar la decisión, un modelo esquemático de las alternativas disponibles y de las posibles consecuencias de cada una, su nombre proviene de la forma que adopta el modelo, parecido a la de un árbol. El modelo está conformado por múltiples de nodos cuadrados que representan puntos de decisión y de los cuales surgen ramas (que deben leerse de izquierda a derecha), que representan las distintas alternativas, las ramas que salen de los nodos circulares, o causales, representan los eventos. La probabilidad de cada evento, P (E), se indica encima de cada rama, las posibilidades de todas las ramas deben sumar 1.0. (Krajewski & Ritzman, 2000, pág. 76).

Las ventajas del uso de árboles de decisión son los siguientes:

1) Facilita la interpretación de la decisión adoptada.
2) Proporciona un alto grado de comprensión del conocimiento utilizado en la toma de decisiones.
3) Explica el comportamiento respecto a una determinada tarea de decisión.
4) Reduce el número de variables independientes.
5) Los arboles de decisión son una magnifica herramienta para el control de la gestión empresarial

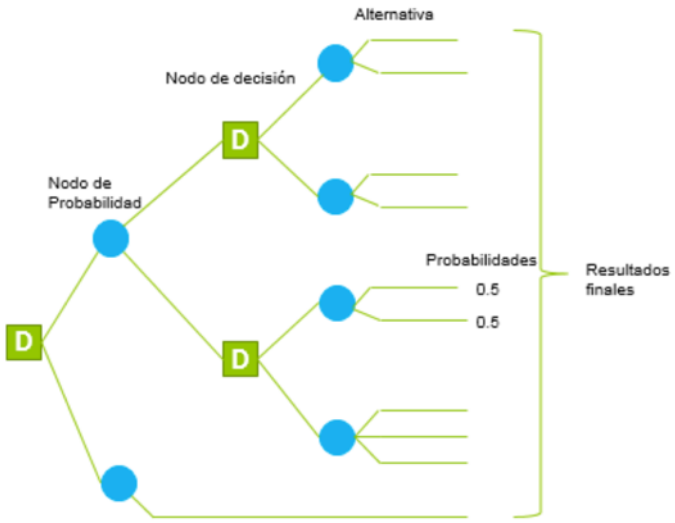
Los árboles de decisión se utilizan en cualquier proceso que implique toma de decisiones, ejemplos de estos procesos son:

- Búsqueda binaria.
- Sistemas expertos.

- Árboles de juego

**Figura 7**

*Ejemplo de árboles de decisión*




*Nota.* María Maya Lopera, 2018

**Visualización**

Además de los análisis matemáticos y empíricos de los algoritmos, existe todavía una tercera forma de estudiar el comportamiento de datos, y se denomina algoritmos de visualización, aunque este no es un algoritmo la gaceta electrónica Kdnuggets lo toma como análisis debido a que en la encuesta ocupa un lugar importante en su uso; se define como el uso de imágenes para transmitir alguna información útil acerca de los algoritmos y el comportamiento de los datos. Esta información puede ser una Figura visual de un algoritmo, o su comportamiento a diferentes tipos de entradas de datos o de la velocidad y ejecución en comparación con otros algoritmos para un mismo problema, para lograr este objetivo, un algoritmo de visualización utiliza elementos

gráficos como puntos, líneas, barras en dos o tres dimensiones con el fin de representar eventos interesantes durante la ejecución de un algoritmo.

Existen dos principales variaciones de un algoritmo de visualización:

- Visualización estática del algoritmo
- Visualización dinámica del algoritmo

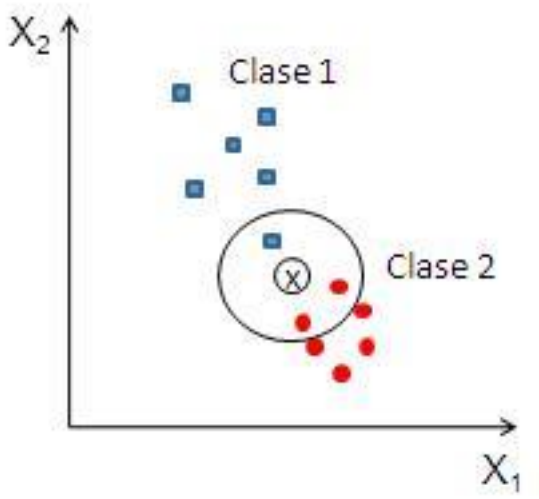
**K-nearest neighbors**

El algoritmo vecino más próximo k-NN (Nearest Neighbor) pertenece a la clase de a estadísticos de reconocimiento de patrones. El método no impone a priori ninguna suposición sobre la distribución de la que se extrae la muestra de modelado. Se trata de un conjunto de entrenamiento con valores positivos y negativos; una nueva muestra se clasifica calculando la distancia al vecino más cercano del conjunto de entrenamiento, el signo de ese punto determinará la clasificación de la muestra; en el clasificador k-vecino más cercano, se consideran los k puntos más cercanos y se utiliza el signo de la mayoría para clasificar la muestra. El rendimiento del algoritmo k-NN está influenciado por tres factores principales:

- La medida de distancia utilizada para localizar a los vecinos más cercanos
- La regla de decisión usada para derivar una clasificación de los k-vecinos más cercanos
- El número de vecinos utilizados para clasificar la nueva muestra. En la siguiente imagen se puede ver un ejemplo de modelo K-non. ( Carlos Espino Timon, Enero, 2017)

**Figura 8**

*Ejemplo de K-nn Vecinos*




*Nota.* Ruslan Unhich, 2019

En conclusión después de haber analizado cada uno de los algoritmos más usados en la actualidad para análisis de datos, se establece que debido al comportamiento de los datos y tomando en consideración que la variable independiente es el tiempo y que el proyecto propone una predicción del comportamiento de los datos desde un punto de vista económico, el algoritmo recomendado y más usado para predecir es el de tipo de regresión en todas sus variaciones, todo esto debido a que se requiere que el pronóstico sea un número cuantitativo ( valor económico ) con un nivel confianza.

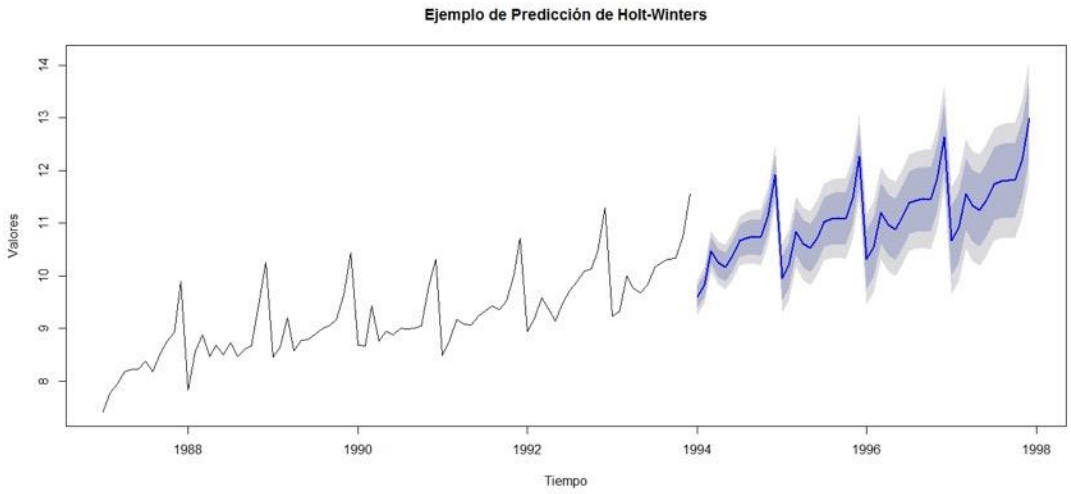
**Series de tiempo ( Holt Winters )**

EL método Holt-Winters es un método de pronóstico de triple exponente suavizante y tiene la ventaja de ser fácil de adaptarse a medida que nueva información real está disponible. El método Holt- Winters es una extensión del método Holt que considera solo dos exponentes suavizantes. Holt-Winters considera nivel, tendencia y estacional de una determinada serie de

tiempos. Este método tiene dos principales modelos, dependiendo del tipo de estacionalidad; el modelo multiplicativo estacional y el modelo aditivo estacional. El referente trabajo se concentra en el modelo multiplicativo.

**Figura 9**

*Ejemplo Holt Winters*




*Nota.* Rivero, 2016

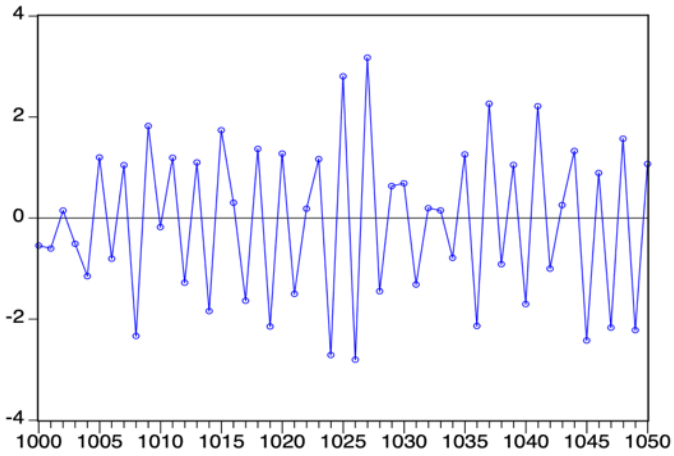
**Arima ( Autor regresivo integrado de promedio móvil )**

Son modelos paramétricos que tratan de obtener la representación de la serie en términos de la interrelación temporal de sus elementos. Este tipo de modelos que caracterizan las series como sumas o diferencias, ponderadas o no, de variables aleatorias o de las series resultantes, fue propuesto por Yule y Slutzky en la década de los 20. Fueron la base de los procesos de medias móviles y autor regresivos que han tenido un desarrollo espectacular tras la publicación en 1970 del libro de Box-Jenkins sobre modelos ARIMA.

El instrumento fundamental a la hora de analizar las propiedades de una serie temporal en términos de la interrelación temporal de sus observaciones es el denominado coeficiente de auto correlación que mide la correlación, es decir, el grado de asociación lineal que existe entre observaciones separadas k periodos. Estos coeficientes de auto correlación proporcionan mucha información sobre cómo están relacionadas entre sí las distintas observaciones de una serie temporal, lo que ayudaŕ a construir el modelo apropiado para los datos.

**Figura 10**

*Ejemplo de ARIMA*




*Nota.* Casimiro, 2009

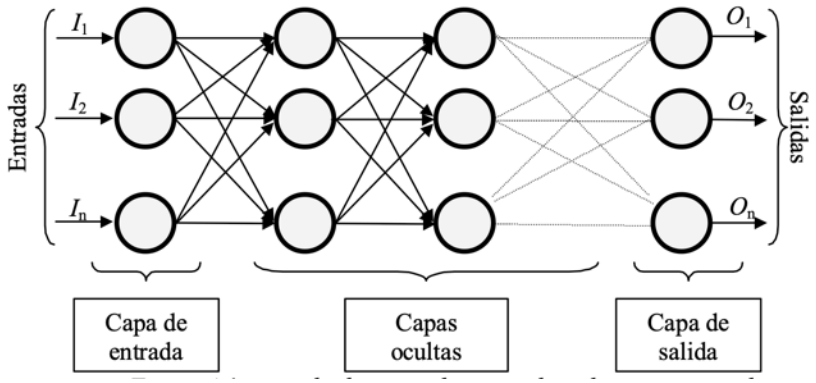
**Neural network**

Una Red Neuronal Artificial (RNA) es un modelo matemático inspirado en el comportamiento biológico de las neuronas y en cómo se organizan formando la estructura del cerebro. El cerebro puede considerarse un sistema altamente complejo, donde se calcula que hay aproximadamente 100 mil millones (10111011) neuronas en la corteza cerebral (humana) y que

forman un entramado de más de 500 billones de conexiones neuronales (una neurona puede llegar a tener 100 mil conexiones, aunque la media se sitúa entre 5000 y 10000 conexiones).

**Figura 11**

*Ejemplo de aplicación Redes Neuronales*




*Nota.* Ruiz et al., 2019

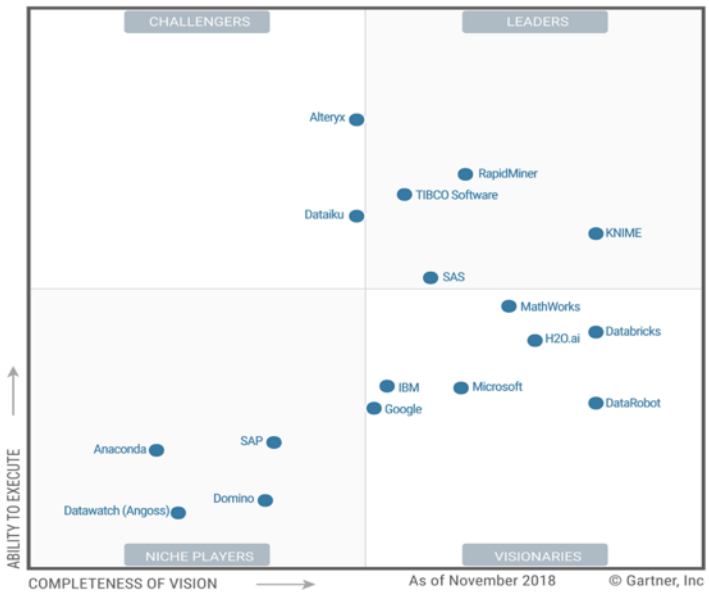
**Revisión de literatura para Herramientas de analítica avanzada**

Para la selección de la herramienta dentro del desarrollo del proyecto se tomó como referencia los aplicativos de software que actualmente en el mercado se han destacan y que son nombrados en el cuadrante de Gartner para BI y Analítica, de igual manera se contrasto esta evaluación de herramientas con el reporte emitido por Forrester para herramientas computacionales que realizan BI y análisis. (Carlie Idoine, Peter Krensky, Erick Brethenoux, Alexander Linden, 2019)

Conforme a lo indicado por el cuadrante de Gartner se define las siguientes herramientas:

**Figura 12**

*Cuadrante de Gartner (2019)*




*Nota*: Carlie Idoine, Peter Krensky, Alexander Linden, Erick Brethenoux, 2019

Los cuadrantes expuestos por el análisis de la compañía Gartner definen:

- Lideres (4): KNIME, RapidMiner, TIBCO Software, SAS
- Challengeres (2): Alteryx, Dataiku
- Visionarios (7): Mathworks, Databricks, H2O.ai, IBM, Microsoft, Google (new), DataRobot (new)
- Niche Players (4): SAP, Anaconda, Domino, Datawatch (Angoss)
- Dos nuevas compañías figuran en el cuadrante del 2019: Google and DataRobot.

Una vez más al analizar el cuadrante se define que la compañías KNIME y RAPIDMINER se mantienen fuertes y líderes en su posición , SAS disminuyo en ser el líder pero todavía se mantiene como uno de los lideres, y TIBCO la cual adquirió algunos compañías de analítica recientemente,

por primera vez alcanzo el cuadrante de los lideres, ASTERYX mientras tanto se encuentra en el cuadrante de los Challengers que en conjunto con DATAIKU mantiene la hegemonía en el cuadrante, Mathworks mejoro significantemente en su visión y se encuentra cercano al cuadrante de los lideres para el reporte del 2019. (Carlie Idoine, Peter Krensky, Alexander Linden, Erick Brethenoux, 2019)

**Knime**

Knime (Konstanz Information Miner) con Base en Surich, Suiza, provee la plataforma KNIME ANALYTICS de tipo open-source de libre distribución, mientras que la extensión comercial KNIME Server ofrece mayor cantidad de funciones como son trabajo en varios equipos, automatizaciones y desarrollo de capacidades propias. (Carlie Idoine, Peter Krensky, Erick Brethenoux, Alexander Linden, 2019)

Así también el análisis realizado por Gartner en el 2019 menciona: “Esta plataforma tiene un buen balance entre la ejecución y visión, además de un sofisticado y limpio producto que permite el fácil uso del mismo para desarrolladores de nivel intermedio.” (Carlie Idoine, Peter Krensky, Erick Brethenoux, Alexander Linden, 2019)

**Rapidminer**

Rapidminer es una herramienta perfecta para crear modelos y a posterior la realización de análisis predictivos de grandes volúmenes de datos. Es una plataforma que facilita el autoservicio de análisis predictivo permitiendo una avanzada analítica empleando solamente drag and drop y opcionalmente la generación de código, se utiliza para realizar análisis de minería de datos (Data Mining) en aplicaciones empresariales, gobierno y academias.

Por lo que Gartner nos menciona que “Una de las ventajas principales de la herramienta es que es una plataforma unificada donde tiene un entorno de programación visual fácil de usar,

permite arrastrar y soltar, de esta forma acelera el enfoque de análisis predictivo para mejorar la productividad". Carlie Idoine, Peter Krensky, Erick Brethenoux, Alexander Linden, 2019)

**Tibco Software**

Con base en Palo Alto, California, U.S. A través de la adquisición de informes empresariales y proveedores de plataformas de BI modernas como (Jaspersoft y Spotfire), proveedores de plataformas de análisis descriptivos y predictivos (Statistica y Alpine Data) y un proveedor de análisis de transmisión (StreamBase Systems), TIBCO ha creado una plataforma de análisis completo y potente.

TIBCO ha pasado del cuadrante Challengers al cuadrante Leader, gracias a una estrategia de integración bien orquestada que contribuye a su capacidad de ejecución y a sus esfuerzos por mantener el ritmo de la innovación en este mercado que cambia rápidamente.

TIBCO tiene una habilidad distintiva para servir a industrias centradas en activos, además de tener capacidades de desarrollo e implementación de extremo a extremo, TIBCO aborda con éxito el dominio analítico de IoT de ciencia de datos desatendida, en parte como resultado de sus raíces centradas en el proceso. ". (Carlie Idoine, Peter Krensky, Erick Brethenoux, Alexander Linden, 2019)

**Sas**

SAS tiene su sede en Cary, Carolina del Norte, EE. UU. Ofrece muchos productos de software para análisis y ciencia de datos. Para este Cuadrante Mágico, evaluamos SAS Enterprise Miner (EM) y SAS Visual Data Mining and Machine Learning (VDMML).

SAS conserva su estado de larga data como líder. Aunque la compañía enfrenta amenazas en múltiples frentes de otros grandes proveedores, descriptores maduros y soluciones de código abierto, conserva una fuerte presencia en el mercado.

Por lo que Gartner menciona: “La visión integral de SAS, está en la misma clase que muchos competidores altamente innovadores, pero la compañía se está quedando atrás en áreas clave como el aprendizaje profundo y las contribuciones a la comunidad de código abierto. Su capacidad de ejecución se ve obstaculizada por costos altos y, a veces, impredecibles, que hacen que los clientes existentes y potenciales exploren otras opciones. Al igual que otros veteranos del mercado de la ciencia de datos, además de enfocarse en nuevos clientes, SAS está asumiendo el desafío de brindar soporte a clientes y usuarios heredados mientras se adapta a un panorama que cambia rápidamente”. (Carlie Idoine, Peter Krensky, Erick Brethenoux, Alexander Linden, 2019)

**Tableu Software**

Tableau es una herramienta de visualización de datos potente utilizada en el área de la Inteligencia de negocios (más conocida como Business Intelligence). Simplifica los datos en bruto en un formato muy fácil de entender.

La esencia de Tableau es simple y a la vez muy relevante: ayudar a las personas y empresas a ver y comprender todos sus datos. Y esto lo consigue ofreciendo a los usuarios toda una selección de herramientas útiles e intuitivas de inteligencia de negocios.

A través de funciones simples como la de arrastrar y soltar, cualquier persona puede acceder y analizar de forma sencilla datos, e incluso, crear informes y compartir esta información con otros usuarios.

Para el proyecto actual al necesidad de contar con una herramienta que permita la visualización online es fundamental, por lo que recurrir a Tableau Online que es una versión de Tableau Server alojada en la nube. De esta forma, podremos acceder a nuestros datos sin necesidad de tener que pasar por un tedioso proceso de instalación.

Conforme al cuadrante de Gartner Tableu Software es por octavo año consecutivo el líder del mercado en analítica e inteligencia de negocios (James Richardson, Rita Sallam, Kurt Schlegel, Austin Kronz, Julian Sun. 2020).Adicional a esto la compañía naviera EP FLOPEC cuenta con licencias para manejar la herramienta Tableu Software.

Revisión de literatura para Arquitecturas

Existen algunas arquitecturas analíticas aplicadas a otros sectores de la industria, pero que su fin es realizar predicciones con el objeto de predecir ventas de productos, por ejemplo:

**Caso 1: Forecasting Nike**

Nike diseña, desarrolla, comercializa y vende calzado, ropa, equipo y accesorios deportivos en todo el mundo. Ofrece productos diseñados para niños, bolsas, calcetines, balones, gafas, relojes, dispositivos digitales, bates, guantes, equipos de protección, palos de golf, calzado, etc. La compañía vende sus productos a tiendas de calzado, tiendas especializadas, grandes almacenes, y minoristas, así como distribuidores independientes. NIKE, Inc. fue fundada en 1964 y tiene su sede en Beaverton, Oregon. Nike como todas las empresas innovadoras a nivel mundial recurren a estrategias de análisis de datos para proyectar y elevar sus ventas.

Por lo que, una de las estrategias usadas por NIKE es recurrir a las redes sociales como FACEBOOK para que una vez que recolecte los datos de esta red social haciendo uso de la herramienta Google Trends, luego procesarlas en la herramienta SAS con el objeto de establecer la reacción de sus clientes cuando se presenta un evento especifico de la marca. Para analizar esta estrategia de ventas se implementó la arquitectura descrita en la Figura No. 14.

**Figura 13**

*Arquitectura Forecasting ventas de NIKE Usando Facebook*


*Nota.* Boldt et al., 2016

Forecasting Nike´s sales using Facebook Data, en este paper se toma la información de Facebook para probar si las previsiones de ventas son precisas para Nike y cuan posibles los datos de Facebook y cómo los eventos relacionados con Nike afectan la actividad en las páginas de Facebook de Nike.

El documento se basa en el marco de ventas de AIDA (Conciencia, Interés, deseo y acción) del dominio de marketing y emplea el método de análisis de conjuntos sociales del dominio de Ciencias sociales computacionales para modelar las ventas de Big Social. (Boldt et al., 2016)

Datos. El conjunto de datos consta de (a) la selección de Facebook de Nike páginas con el número de me gusta, comentarios, mensajes, etc. que se han registrado para cada página por día y (b) de negocios. Datos en términos de cifras de ventas globales trimestrales publicadas en los informes financieros de Nike. También se lleva a cabo un estudio de eventos utilizando el visualizador de conjuntos sociales (SoSeVi).

Los hallazgos sugieren que los datos de Facebook tienen valor informativo, algunos de los modelos de regresión simple tienen una alta precisión de pronóstico. Las regresiones múltiples tienen una menor precisión de pronóstico y causar barreras de análisis debido a las características del conjunto de datos tales como perfecta multicolinealidad. El evento de estudio encontró anormal actividad en torno a varios eventos específicos de Nike, pero inferencias sobre esos picos de actividad, ya sean puramente relacionados con el evento o las coincidencias, solo se pueden determinar después de un detallado caso por caso análisis del texto. (Boldt et al., 2016)

**Caso 2: Caso predictivo cliente Bancario**

En este caso el dataset de una entidad bancaria fue obtenida por la Universidad de Irvine para realizar una investigación haciendo uso de la herramienta Amazon web Services y para luego aplicar análisis avanzado con la herramienta Amazon s3 con el objeto de predecir las reacciones de sus clientes a las promociones de marketing de la entidad bancaria, para realizar este proyecto de investigación se realizó la arquitectura descrita en la Figura No. 15.

**Figura 14**

*Arquitectura Análisis predictivo usando AWS*

*Nota.* Ramesh, 2017

Otro caso de estudio donde se establece una predicción usando datos es el paper "Predictive Analytics for Banking User Data using AWS machine learning web Services" desarrollado por Ranjith Ramesh del departamento de Computer Science de Johns Hopkins University Baltimore , Maryland, USA donde el objetivo del proyecto es desarrollar una máquina, con un modelo de aprendizaje para realizar análisis predictivos en la banca usando un conjunto de datos bancarios donde constan de detalles sobre los clientes y se define el gusto del cliente y si el cliente comprará un producto proporcionado por el banco o no, el conjunto de datos se obtiene del repositorio de la Universidad de California Irvine. Este conjunto de datos es utilizado para crear un modelo de clasificación binario utilizando Amazon Web, Plataforma de Aprendizaje de Máquina de Servicio (AWS), y se establece que el 70% de los datos son utilizados para entrenar el modelo de clasificación binaria y el 30% del conjunto de datos se utiliza para probar el modelo. Dependiendo del resultado de la prueba se analizan los parámetros esenciales como precisión, recuerdo, y tasas de falsos positivos. Estos parámetros evalúan la eficiencia del modelo.

Una vez que se ha diseñado el modelo se realiza las pruebas del modelo utilizando dos características en AWS Machine Learning. Primero, usando predicción en tiempo real donde se ingresan datos de entrada en tiempo real y se prueba el modelo, realizando predicción por lotes, donde se obtiene un conjunto de datos del cliente y se suben los datos para evaluar la predicción. (Ramesh, 2017)

**Caso 3: Predicción Financiera**

En este caso la el departamento de electrónica de la Universidad de Mumbai, India realizo una aplicación de predicción tomando como referencia los valores del mercado bursátil del índice Nifty 50, con el cual se realizó la investigación con un base de datos de 9 años para poder predecir

el valor de este índice para realizar esta investigación se estableció la arquitectura de la Figura No. 16 que se presenta a continuación:

**Figura 15**

*Arquitectura de predicción de precios usando análisis de datos*


*Nota*: Tiwari et al., 2017

Otro caso de uso de estudio es presentado en el paper de "Stock Price Prediction Using Data Analytics", desarrollado por Sashank Tiwari del Depto. Of Electronics del College of Engineering, Mumbai, India donde se relaciona la predicción financiera, que es de gran interés para inversores, el artículo propone el uso de análisis de datos para ser utilizado en ayudar a los inversores a hacer una predicción financiera correcta para que los inversores pueden tomar la decisión más adecuada sobre la inversión, dos plataformas se utilizan para la operación: Python y R; se aplican varias técnicas como Arima, Holt Winters, Redes neuronales (Feed Perceptron y multicapa), regresión lineal y series de tiempo. Se implementan series para pronosticar el índice de apertura del precio y el rendimiento en R. Mientras que en Python Multi-layer perceptron y Regresión de vectores de apoyo se implementan para el pronóstico de Nifty en este artículo se realizaron 50 cotizaciones y también análisis de sentimiento de la acción, haciendo uso de tweets recientes en la plataforma Twitter.

Los índices bursátiles de Nifty 50 (^ NSEI) son considerados como una entrada de datos para los métodos que se implementan, se utiliza un dataset compuesto de 9 años de datos. La precisión se calculó utilizando 2-3 años de resultados de previsión de R y 2 meses de resultados de previsión de Python después de comparar con el precio real de las acciones. Se utilizan la Error cuadrático medio y otros parámetros de error para cada predicción, para el sistema se encontró que la red de alimentación hacia adelante solo produce 1.81598342% de error cuando el precio de apertura de las acciones es pronosticado. (Tiwari et al., 2017)

## Capitulo III. Propuesta de un modelo para proyección de precios de buques tanqueros

En el siguiente capítulo se va realizar una propuesta de mejora para la proyección de ingresos por renta de buques en la empresa naviera, que permita tener una visión a corto plazo de los ingresos por renta de buques, a través del uso de una arquitectura definida, herramientas y modelos analíticos-predictivos. Para esto se definirá la arquitectura analítica propicia para la predicción de precios en el mercado de renta de buques tanqueros, así también se definirán las herramientas y modelos que se ajustan de mejor forma para la compañía naviera EP FLOPEC.

### Técnica Analítica Seleccionada

La técnica analítica seleccionada es la predictiva, tomando en consideración que los objetivos del proyecto apuntan a una predicción de los precios del costo de renta de un buque tanquero en una determinada ruta, y tomando en consideración las necesidades de la compañía naviera para establecer un valor del precio referencial en el futuro para los buques de su flota, entonces se requiere una predicción.

### Herramienta Analítica Seleccionada

En consecuencia, para la aplicación del modelo se usó la referencia de los líderes nombrados en el cuadrante de Gartner (Cuadrante, Gartner 2019) mencionado en el capítulo 2 punto 2.3, por lo que en base a la referencia citada y recopilada de acuerdo al análisis de la información establecida para líderes de Inteligencia de Negocios se tomaron en cuenta los siguientes puntos de valoración:

Para valorarlos en un cuadro comparativo tomando en consideración los siguientes aspectos funcionales:

- Licencia libre (Piatetsky, 2018)

- Multiplataforma(Gartner, 2019)
- Interfaz amigable (Gartner, 2019)
- Fácil configuración(Gartner, 2019)
- Fácil instalación(Gartner, 2019)
- Visualización de datos(Gartner, 2019)
- Conversión de datos(Gartner, 2019)
- Módulos de predicción(Gartner, 2019)
- Documentación en la web

Entonces se realiza el cuadro explicativo de cada uno de los parámetros calificados y que fueron recopilados de las *Nota.*s como Gartner, Kdnuggets que son compañías que se dedican a realizar encuestas a los usuarios para definir el comportamiento y usabilidad de las diferentes herramientas analíticas en el mercado, esto se describe en la Tabla 1:

Tabla 1

*Comparativo y descripción de cada una de las características de las herramientas analíticas.*

| PARAMETRO | FUENTE | RAPIDMINER | SAS | KNIME | TIBCO |
|---|---|---|---|---|---|
| Licencia Libre | (Piatetsky, 2018) | Freeware, diferentes versiones de pago | Freeware limitado a instituciones públicas, el precio se establece tras solicitud, diferentes modelos disponibles | Software libre (GPL) (a partir de la versión 2.1) | Freeware, diferentes versiones de pago |
| Multiplataforma y manejo de proyectos | (Gartner, 2019) | El software maneja de buena forma la seguridad y es flexible con el manejo de proyectos | No maneja linemaientos de uditoria y productividad | Es deficiente en el manejo de la seguridad | El manejo de proyectos es bastante complejo |
| Interfaz amigable con el usuario / Lenguaje de programacion | (Gartner, 2019) | El sistema maneja un concepto visual y intuitivo que facilita el desarrollo de aplicaciónes por parte del usuario | El sistema se mira y se siente igual | Maneja un sistema de forma intituitiva que facilita el manejod el usuario | Maneja un composicion visual intiutiva |
| Rendimiento y escabilidad | (Gartner, 2019) | Utiliza un servidos y escritorio para las implementaciones | No utiliza multinodos para implementacion | Utilizar la nube, un servidos y escritorio para las implementaciones | Utiliza la nube y escritorio pero no una version server para implementaciones |
| Fácil instalación / sistema operativo | (Piatetsky, 2018) | Windows, macOS, Linux | Windows, macOS, Linux | Windows, macOS, Linux | Windows, Mobile IOS, Android |
| Visualización de datos | (Gartner, 2019) | El software permite un alto rango de pasos de exploración | Incluye visualización interactiva | El software permite pasos limitados de exploracion | No permite visualizcion interactiva |
| Conversión de datos | (Gartner, 2019) | Maneja cualuqier fuente de datos de manera segura y eficiente, las conexiones con bases de datos son sencillas y se ejecutan intuitivamente | Maneja con facilidad fuentes de datos como imágenes, excel, csv | No maneja con facilidad las fuentes de datos como imágenes, excel, csv, audio, etc | Maneja deficientemente fuentes de imágenes y audio, fuentes como excel, csv y database |
| Módulos de predicción | (Gartner, 2019) | Modulos de simulacion no se encuentran incluidos en la herramienta, se requieren pago adicional, modulos de predicción básicos para la herrmienta gratis. | Modulos de prediccion limitados, y basicos se requiere otras caracteristicas adicionales para su uso | Modulos de prediccion inlcuidos en el software, modulos de simulacion integrados en la aplicación | Modulos de simulacio y analisis se encuentran integrados en la aplicación |
| Documentación en la web | | Acceso a informacion y manuales de uso via web, pagina dedicada a ayuda y desarrollo de proyectos, comunidad de desarrolladores | Acceso a informacion y manuales de uso via web, pagina dedicada a ayuda y desarrollo de proyectos, comunidad de desarrolladores | Acceso a informacion y manuales de uso via web, pagina dedicada a ayuda y desarrollo de proyectos, comunidad de desarrolladores | Acceso a informacion y manuales de uso via web, pagina dedicada a ayuda y desarrollo de proyectos, comunidad de desarrolladores |

*Nota.*: Gartner, 2019, Piatetsky, 2018

Para la calificación de cada uno de los fundamentos se define la aplicación de la siguiente escala calificativa de la Tabla 2:

**Tabla 2**

*Valores para calificación de la herramienta*

| Valor | Significado |
|---|---|
| 0 | Totalmente no satisfecho |
| 5 | Completamente satisfecho |

Los resultados de la calificación de la herramienta analítica (tabla 1), se muestran en la Tabla 3:

**Tabla 3**

*Calificación de la herramienta de analítica avanzada*

| **HERRAMIENTA BI** | | | | |
|---|---|---|---|---|
| PARAMETRO | RAPIDMINER | SAS | KNIME | TIBCO |
| Licencia Libre | 4.3 | 4.0 | 4.5 | 5.0 |
| Multiplataforma y manejo de proyectos | 4.5 | 4.2 | 4.0 | 4.0 |
| Interfaz amigable con el usuario | 4.4 | 3.8 | 4.3 | 3.8 |
| Rendimiento y escabilidad | 4.4 | 4.4 | 4.2 | 4.3 |
| Fácil instalación | 4.5 | 3.8 | 3.9 | 4.1 |
| Visualización de datos | 4.2 | 4.2 | 4.1 | 4.8 |
| Conversión de datos | 4.6 | 4.6 | 4.5 | 4.7 |
| Módulos de predicción | 4.6 | 4.4 | 4.5 | 4.8 |
| Documentación en la web | 4.3 | 4.2 | 4.2 | 3.6 |
| **PROMEDIO** | **39.7** | **37.6** | **38.1** | **39.0** |

*Nota.*: Gartner, 2019, Piatetsky, 2018, TIBCO Data Science vs RapidMiner, s/f

De acuerdo a los resultados obtenidos dentro en la Tabla No. 3, la herramienta seleccionada para crear el modelo de predicción es RapidMiner con una calificación de 39,7 puntos.

Con este resultado se asegura a EP FLOPEC que el entregable no tendrá mayor costo para su mantenimiento y una vez que entre en funcionamiento, puede ser modificado por el personal del área de TIC's y será de fácil comprensión mediante los manuales y tutoriales que se encuentran en internet, el departamento de TICs podrá alterar el modelo o crear uno nuevo a partir de los datos futuros.

**Selección del Algoritmo**

Para la selección del algoritmo se ha seguido lo estipulado por (Simeone, 2018) al respecto del aprendizaje Supervisado, y no supervisado por lo que a continuación se detalla lo siguiente:

- Aprendizaje Supervisado

En el aprendizaje supervisado, los algoritmos trabajan con datos "etiquetados" (labeled data), intentado encontrar una función que, dadas las variables de entrada (input data), les asigne la etiqueta de salida adecuada. El algoritmo se entrena con un "histórico" de datos y así "aprende" a asignar la etiqueta de salida adecuada a un nuevo valor, es decir, predice el valor de salida.

Conforme a la revisión de literatura realizada en el capítulo II del presente proyecto se definieron los algoritmos más utilizados para la predicción de datos supervisados y que requieren se pronosticados, y se enlistan a continuación:

- Holt Winters
- Regresión Lineal
- Arima
- Red Neuronal
- K – nn
- Gradient Boosted Trees

Para el caso de este proyecto donde se requiere hacer un entrenamiento supervisado de una variable dependiente en función del tiempo, por lo que la variable independiente será el tiempo y se usarán los algoritmos descritos anteriormente como algoritmos el modelo de predicción y estos algoritmos serán validados de acuerdo a los siguientes parámetros:

- Error medio promedio al cuadrado
- Error Absoluto
- Error Relativo
- Coeficiente de correlación

Estos parámetros de calificación son los que representativamente nos demuestran la mejor valides del modelo.

Para el entrenamiento y validación del modelo se definió el uso de un dataset referencial que será la ruta Arzew – Trieste. El dataset está constituido por un total de 29 rutas con 261 registros lo que hace un total de 7569 registros, este dataset cuenta con una columna de registros de tiempo con un historial de 4 años de datos a partir de enero 2015 y hasta llegar a noviembre del 2019, estos de registros de tiempo están ordenados con un salto de tiempo de 7 días entre ellos, los registros de cada una de las rutas es el índice World Scale para renta de buques petroleros tipo aframax.

**Tabla 4**

*Dataset Nota. de datos de los precios de renta de buques.*

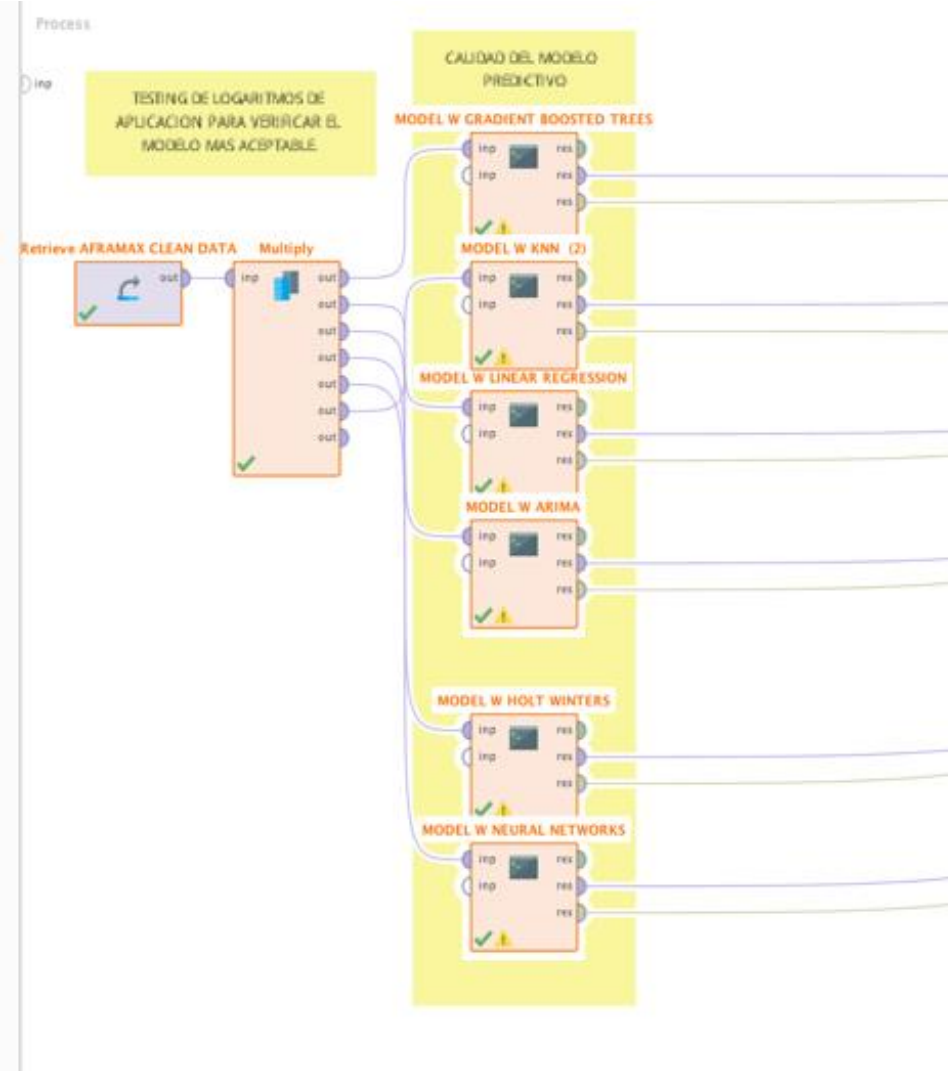
| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | DATE | Curacao Hamburg Aframax 80K Worldscale Rates | Hound Point Bayway Aframax 80K Worldscale Rates | Arzew Philadelphia Aframax 80K Worldscale Rates | Jakarta - Chiba Aframax Worldscale Rates | Ras Tanura Chiba Clean 75K Worldscale Rates | Ras Tanura Singapore Aframax 80K Worldscale Rates | Sidi Kerir Trieste Aframax 80K Worldscale Rates |
| 2 | Friday, January 2, 2015 | 95.00 | 127.50 | 90.00 | 110.00 | 90.00 | 110.00 | 95.00 |
| 3 | Friday, January 9, 2015 | 95.00 | 115.00 | 90.00 | 105.00 | 90.00 | 102.50 | 90.00 |
| 4 | Friday, January 16, 2015 | 105.00 | 120.00 | 100.00 | 112.50 | 87.50 | 105.00 | 110.00 |
| 5 | Friday, January 23, 2015 | 105.00 | 120.00 | 120.00 | 120.00 | 85.00 | 112.50 | 140.00 |
| 6 | Friday, January 30, 2015 | 95.00 | 110.00 | 115.00 | 120.00 | 105.00 | 115.00 | 125.00 |
| 7 | Friday, February 6, 2015 | 105.00 | 105.00 | 175.00 | 107.50 | 105.00 | 112.50 | 200.00 |
| 8 | Friday, February 13, 2015 | 110.00 | 100.00 | 105.00 | 107.50 | 102.50 | 110.00 | 117.50 |
| 9 | Friday, February 20, 2015 | 140.00 | 95.00 | 85.00 | 107.50 | 94.50 | 110.00 | 87.50 |
| 10 | Friday, February 27, 2015 | 150.00 | 95.00 | 95.00 | 107.50 | 100.00 | 110.00 | 105.00 |
| 11 | Friday, March 6, 2015 | 150.00 | 95.00 | 95.00 | 105.00 | 110.00 | 107.50 | 105.00 |
| 12 | Friday, March 13, 2015 | 140.00 | 97.50 | 95.00 | 107.50 | 102.50 | 110.00 | 100.00 |
| 13 | Friday, March 20, 2015 | 110.00 | 97.50 | 115.00 | 107.50 | 105.00 | 110.00 | 122.50 |
| 14 | Friday, March 27, 2015 | 150.00 | 97.50 | 120.00 | 110.00 | 102.50 | 112.50 | 115.00 |
| 15 | Friday, April 3, 2015 | 130.00 | 105.00 | 97.50 | 105.00 | 102.00 | 110.00 | 105.00 |
| 16 | Friday, April 10, 2015 | 102.50 | 95.00 | 95.00 | 105.00 | 96.00 | 110.00 | 100.00 |
| 17 | Friday, April 17, 2015 | 110.00 | 120.00 | 95.00 | 102.50 | 96.00 | 102.50 | 105.00 |
| 18 | Friday, April 24, 2015 | 100.00 | 110.00 | 100.00 | 100.00 | 93.00 | 102.50 | 110.00 |
| 19 | Friday, May 1, 2015 | 85.00 | 105.00 | 112.50 | 97.50 | 102.50 | 100.00 | 112.50 |
| 20 | Friday, May 8, 2015 | 80.00 | 105.00 | 92.50 | 107.50 | 103.00 | 102.50 | 102.50 |
| 21 | Friday, May 15, 2015 | 90.00 | 95.00 | 87.50 | 115.00 | 103.00 | 117.50 | 92.50 |
| 22 | Friday, May 22, 2015 | 95.00 | 120.00 | 120.00 | 125.00 | 105.00 | 125.00 | 140.00 |
| 23 | Friday, May 29, 2015 | 80.00 | 110.00 | 120.00 | 130.00 | 120.00 | 127.50 | 140.00 |
| 24 | Friday, June 5, 2015 | 80.00 | 110.00 | 110.00 | 135.00 | 126.00 | 130.00 | 120.00 |
| 25 | Friday, June 12, 2015 | 100.00 | 125.00 | 120.00 | 155.00 | 125.00 | 142.50 | 145.00 |
| 26 | Friday, June 19, 2015 | 110.00 | 135.00 | 120.00 | 190.00 | 125.00 | 175.00 | 135.00 |
| 27 | Friday, June 26, 2015 | 110.00 | 135.00 | 125.00 | 190.00 | 125.00 | 160.00 | 150.00 |
| 28 | Friday, July 3, 2015 | 105.00 | 95.00 | 105.00 | 180.00 | 140.00 | 150.00 | 115.00 |
| 29 | Friday, July 10, 2015 | 90.00 | 85.00 | 80.00 | 150.00 | 145.00 | 140.00 | 80.00 |
| 30 | Friday, July 17, 2015 | 90.00 | 90.00 | 85.00 | 125.00 | 170.00 | 130.00 | 80.00 |
| 31 | Friday, July 24, 2015 | 95.00 | 87.50 | 97.50 | 115.00 | 155.00 | 117.50 | 105.00 |
| 32 | Friday, July 31, 2015 | 95.00 | 82.50 | 92.50 | 100.00 | 155.00 | 110.00 | 107.50 |
| 33 | Friday, August 7, 2015 | 95.00 | 82.50 | 92.50 | 100.00 | 170.00 | 110.00 | 105.00 |
| 34 | Friday, August 14, 2015 | 75.00 | 85.00 | 95.00 | 100.00 | 170.00 | 102.50 | 102.50 |
| 35 | Friday, August 21, 2015 | 75.00 | 90.00 | 80.00 | 97.50 | 160.00 | 97.50 | 80.00 |
| 36 | Friday, August 28, 2015 | 70.00 | 90.00 | 72.50 | 95.00 | 120.00 | 95.00 | 77.50 |
| 37 | Friday, September 4, 2015 | 70.00 | 87.50 | 70.00 | 92.50 | 110.00 | 92.50 | 75.00 |
| 38 | Friday, September 11, 2015 | 65.00 | 87.50 | 70.00 | 97.50 | 92.50 | 92.50 | 75.00 |
| 39 | Friday, September 18, 2015 | 85.00 | 82.50 | 70.00 | 95.00 | 82.50 | 90.00 | 75.00 |
| 40 | Friday, September 25, 2015 | 95.00 | 80.00 | 65.00 | 92.50 | 82.50 | 86.00 | 70.00 |
| 41 | Friday, October 2, 2015 | 90.00 | 87.50 | 65.00 | 92.50 | 80.00 | 86.00 | 67.50 |
| 42 | Friday, October 9, 2015 | 90.00 | 95.00 | 95.00 | 92.50 | 77.50 | 86.00 | 100.00 |

*Nota.* Gerencia Comercial, EP FLOPEC

El dataset, una vez aplicado el proceso extracción, transformación y carga “ETL” se usó para el entrenamiento y validación de cada uno de los algoritmos aplicados para construir uno a uno el modelo en la herramienta de analítica avanzada, cada uno de los modelos fueron construidos en un espacio de trabajo para poder ser evaluados en conjunto como se muestra en la Figura No. 17:

**Figura 16**

*Procesos de modelado para entrenamiento y evaluación*

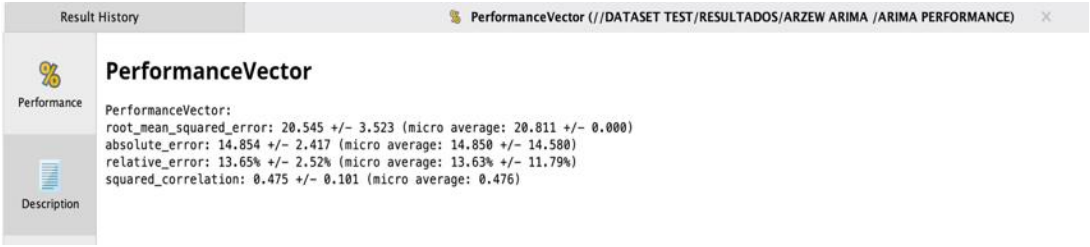



A continuación, se presenta cada una de las pruebas de aplicación, para la elección del mejor algoritmo se utilizó un dataset especifico el cual fue utilizado para realizar un benchmarking entre los siguientes algoritmos y con el uso de la herramienta analítica se obtienen resultados concluyentes los cuales se detallan a continuación para cada uno de los modelos desarrollados:

Para el modelo generado en base a algoritmo ARIMA se presentan los siguientes resultados:

**Figura 17**

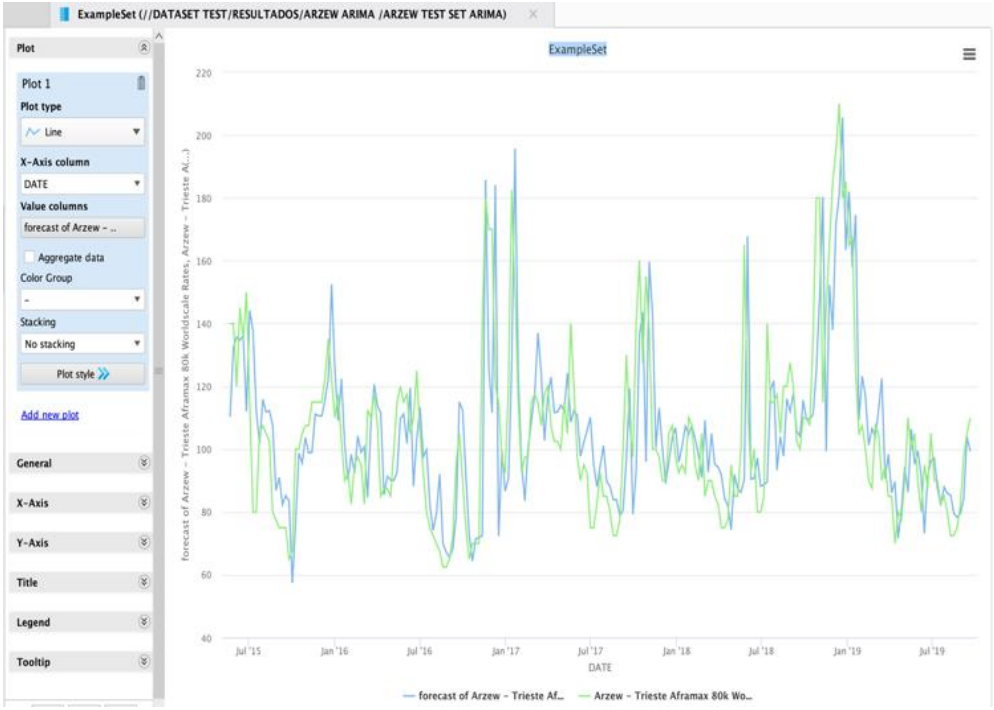
*Resultados validación algoritmo ARIMA*




La Figura 18, muestra la comparativa de la curva de precios de la ruta de referencia vs la ruta pronosticada con el modelo construido a partir del algoritmo ARIMA:

**Figura 18**

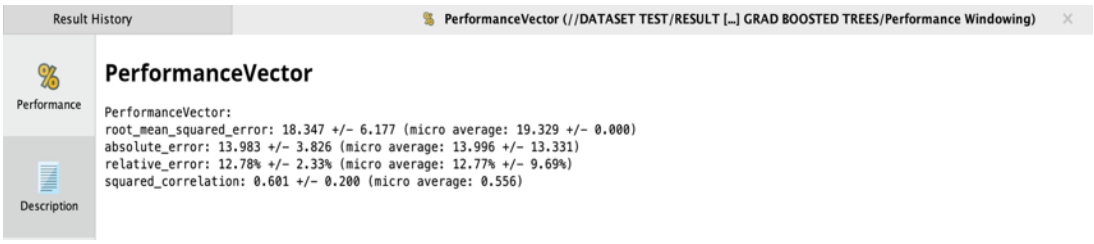
*Modelado real vs Pronostico con modelo ARIMA*




Para el modelo generado en base a algoritmo GRADIENT BOOSTED TREE:

**Figura 19**

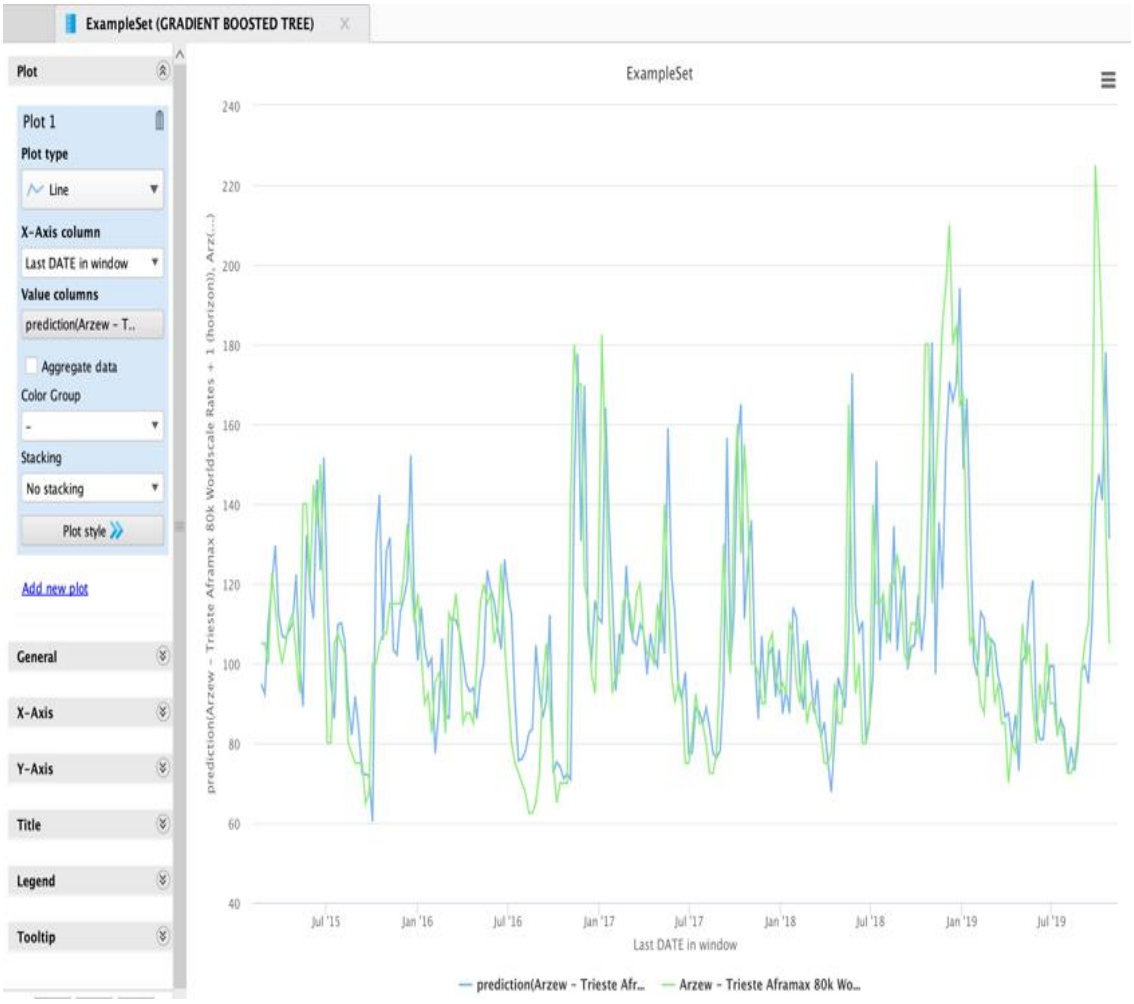
*Resultados validación algoritmo Gradient Boosted Tree*




La Figura 21 muestra la comparativa de la curva de precios de la ruta de referencia vs la ruta pronosticada con el modelo construido a partir del algoritmo GRADIENT BOOST TREE:

**Figura 20**

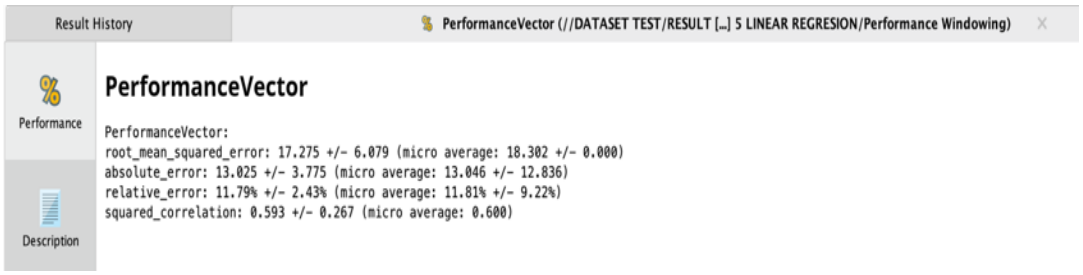
*Modelado real vs Pronostico con modelo Gradient Boosted Tree*




Para el modelo generado en base a algoritmo LINEAR REGRESSION:

**Figura 21**

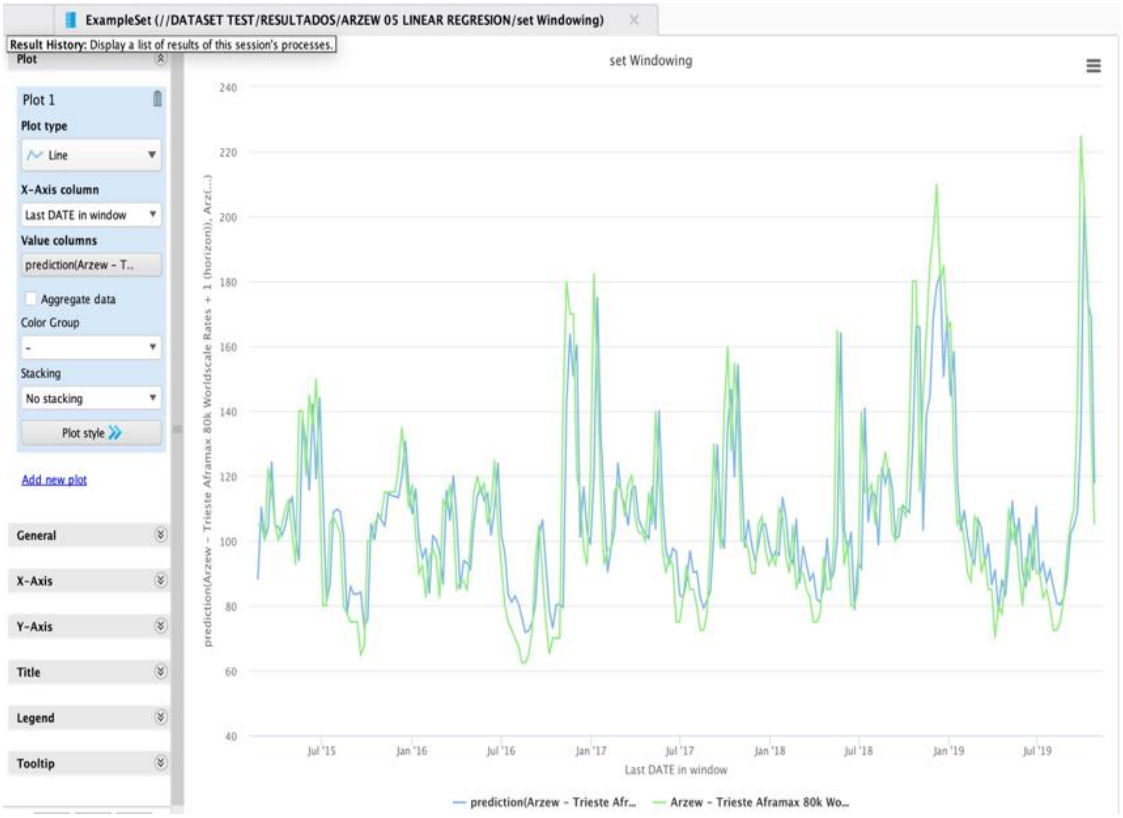
*Resultados validación algoritmo Linear Regression.*




La Figura 23 muestra la comparativa de la curva de precios de la ruta de referencia vs la ruta pronosticada con el modelo construido a partir del algoritmo LINEAR REGRESSSION:

**Figura 22**

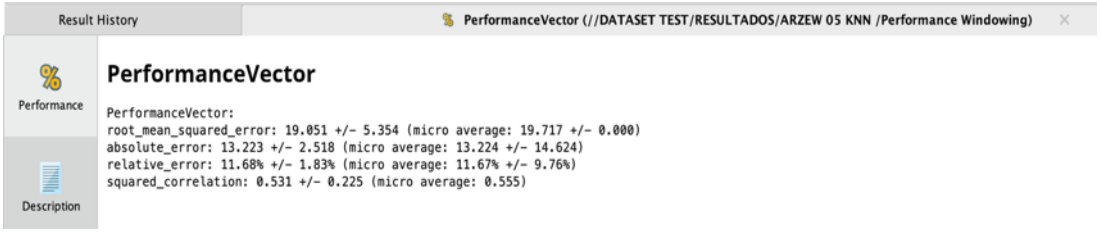
*Modelado real vs Pronostico con modelo Linear Regression*




Para el modelo generado en base a algoritmo K-NN:

**Figura 23**

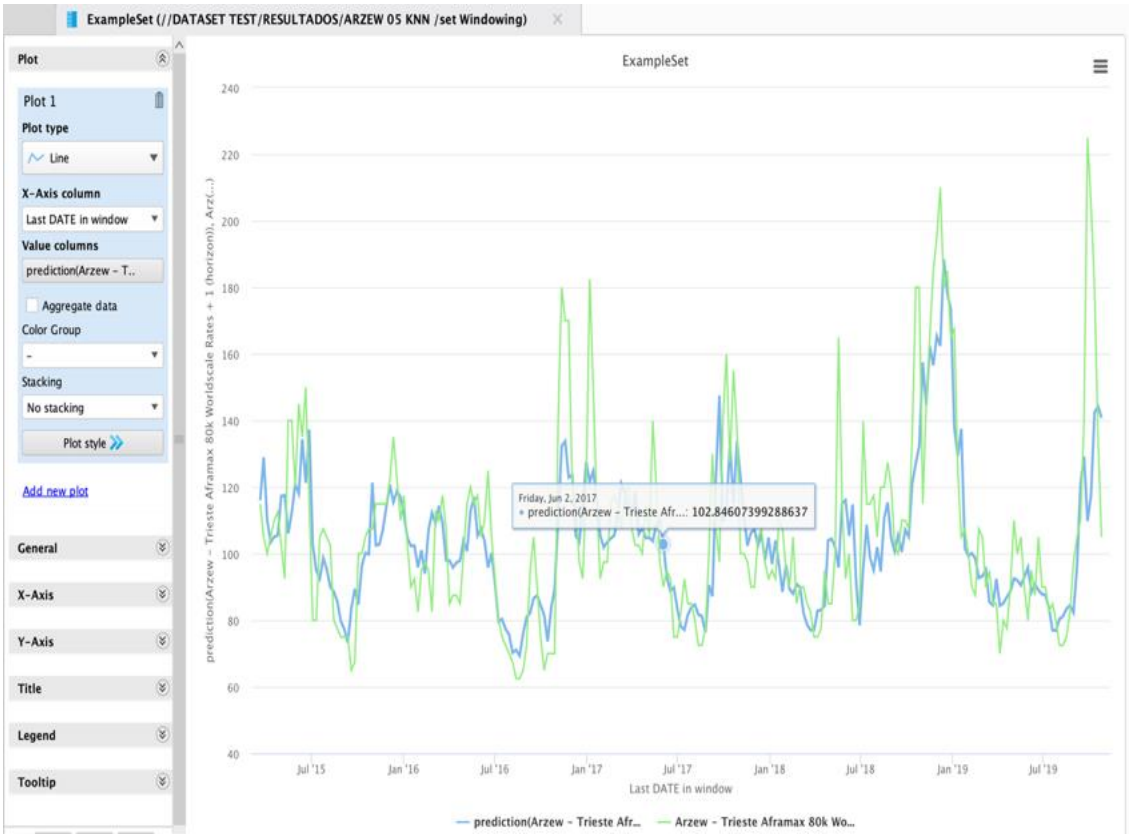
*Resultados validación algoritmo K-NN*




La Figura 25 muestra la comparativa de la curva de precios de la ruta de referencia vs la ruta pronosticada con el modelo construido a partir del algoritmo K-NN:

**Figura 24**

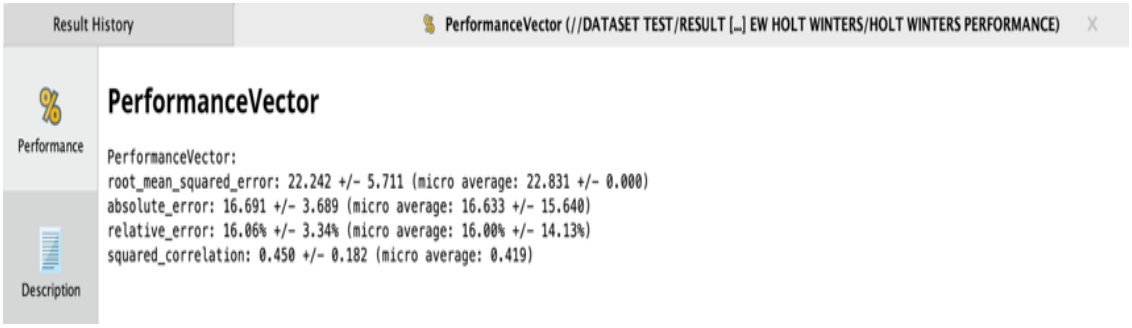
*Modelado real vs Pronostico con modelo K-NN*



Para el modelo generado en base a algoritmo HOLT WINTERS:

**Figura 25**

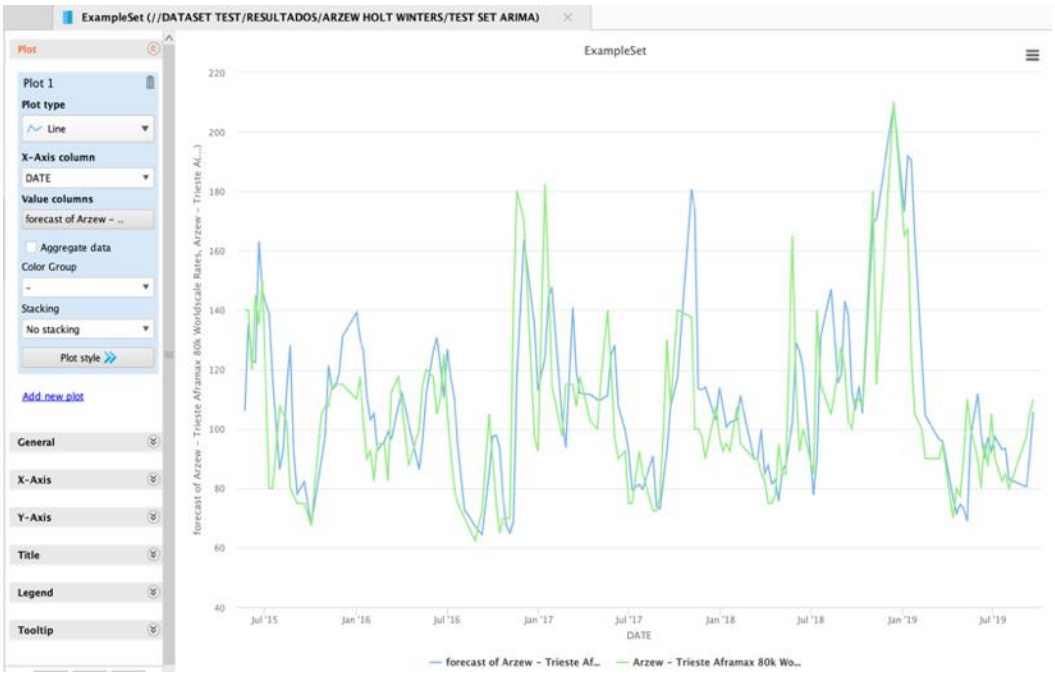
*Resultados validación algoritmo Linear Regression Holt-Winters*




La Figura 27 muestra la comparativa de la curva de precios de la ruta de referencia vs la ruta pronosticada con el modelo construido a partir del algoritmo HOLT - WINTERS:

**Figura 26**

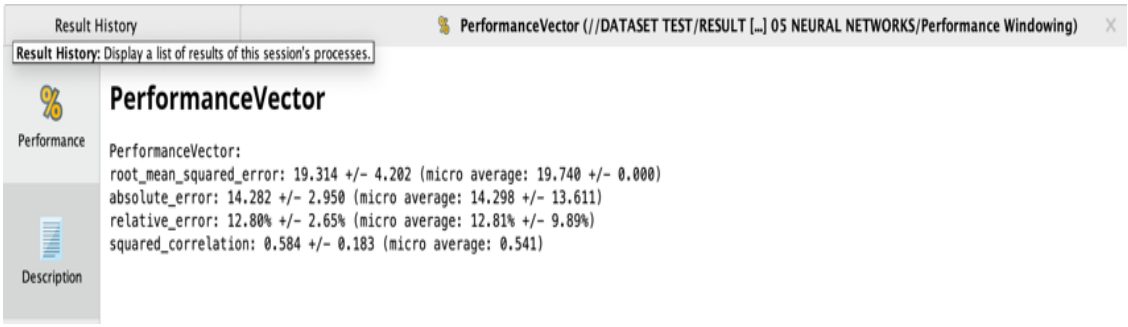
*Modelado real vs Pronostico con modelo Holt Winters*




Para el modelo generado en base a algoritmo NEURAL NETWORKS:

**Figura 27**

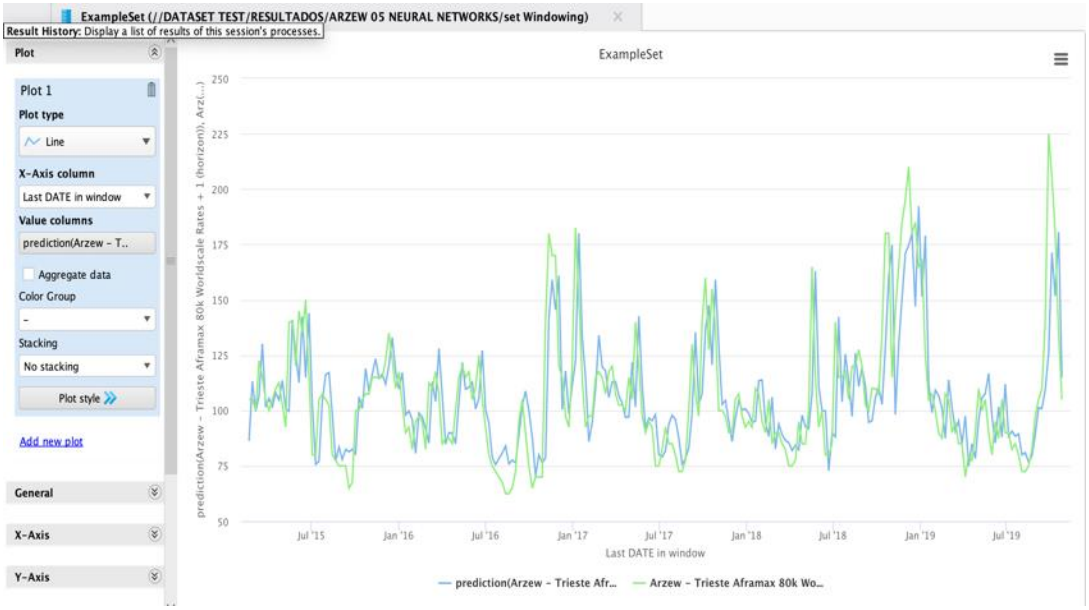
*Resultados validación algoritmo Neural Networks*




La Figura 29 muestra la comparativa de la curva de precios de la ruta de referencia vs la ruta pronosticada con el modelo construido a partir del algoritmo NEURAL NETWORKS:

**Figura 28**

*Modelado real vs Pronostico con modelo Neural Network*




Una vez tabulados todos los datos obtenidos, estos resultados se plantean en la siguiente Tabla comparativa:

**Tabla 5**

*Cuadro comparativo de resultados de los errores de los modelos*

| PARAMETRO DE CALIFICACION ( ERROR ) | K-NN | Neural Networks | Árboles de Clasificación y Regresión ( Gradient Boosted Tree ) | Holt Winters | Modelos lineales generalizados ( Linear Regresion ) | ARIMA |
|---|---|---|---|---|---|---|
| ROOT MEAN SQUARED ERROR | 19.051 +/- 5.354 % | 19.314 +/- 4.202 % | 18.347 +/- 6.177 % | 22.242 +/- 5.711 % | 17.275 +/- 6.079 % | 20.54 +/- 3.523 % |
| ABOSLUTE ERROR | 13.223 +/-2.518 % | 14.282 +/-2.95 % | 13.983 +/-3.826 % | 16.691 +/-3.689 % | 13.025 +/-3.775 % | 14.854 +/-2.417 % |
| RELATIVE ERROR | 11.68 +/- 1.83 % | 12.8 +/- 2.65 % | 12.78 +/- 2.33 % | 16.06 +/- 3.34 % | 11.79 +/- 2.43 % | 13.65 +/- 2.52 % |
| SQUARED CORRELATION | 0.531 +/- 0.225 | 0.584+/- 0.183 | 0.601 +/- 0.200 | 0.450 +/- 0.182 | 0.593 +/- 0.267 | 0.475 +/- 0.101 |

De acuerdo al análisis de los datos de la Tabla No. 5 se pueden obtener la siguiente información:

- El algoritmo que obtuvo el menor error una vez aplicado el Dataset de prueba es el de Modelos Lineales generalizados ( Linear Regresión )
- El algoritmo que presento el mayor error absoluto es el de Holt Winters.
- El error que presento menor coeficiente de correlación es el de Holt Winters
- El algoritmo que presento el mayor coeficiente de correlación es el de Gradient Boosted Trees

Es necesario establecer que el nivel de error obtenido por la aplicación de cada uno de los modelos estadísticos al dataset de referencia, se encuentra entre el 13% y 16% esto quiere decir que, la diferencia entre los precios de renta de buques real y el calculado para un punto especifico estará alrededor de 13 +/- 3.77 puntos porcentuales aproximadamente, para la proyección de precios de renta de buques tanqueros realizada de forma manual por la Gerencia Comercial donde el error absoluto de estimación mes a mes varia y que este puede llegar hasta un valor del 36.31% como sucedió en octubre del 2017, por lo que contar con una proyección del precio real que contenta un 13% de error mejora la precisión de la predicción y establece la aceptación de la

proyección lo que permitirá establecer una referencia para la planificación de la renta de un buque tanquero.

Por las razones antes expuestas y de acuerdo al análisis realizado, para el desarrollo de nuestro proyecto se utilizará el algoritmo que ha presentado el menor error una vez se validó el entrenamiento del modelo, este será Linear Regresión.

Arquitectura Propuesta

Una vez analizada las arquitecturas de cada uno los de los ejemplos de aplicación en el Capítulo II se define la composición de cada uno de los componentes de la arquitectura más adecuada para ser aplicada en este proyecto, siempre tomando en consideración los parámetros concluyentes citados con anteriormente, se propone el uso de la siguiente arquitectura para el desarrollo del proyecto:

**Figura 29**

*Arquitectura de solución propuesta*


Los componentes de la arquitectura propuesta se detallan a continuación:

1. Fuente de información

Las fuente de información o dataset serán obtenidos a partir de los datos históricos de los precios de renta de buques tanqueros, estos datos son adquiridos por la compañía naviera para obtener la información correspondiente a los precios de renta de los buques tanqueros por ruta y conforme a la variación del WS (World Scale) que es el índice transaccional diario para renta de buques tanqueros. Estas fuentes serán de la compañía Clarkson.

2. Calidad de datos Proceso ETL (Extracción, transformación y carga)

Para el tratamiento de los datos se trabajará en la herramienta de software definida (Rapidminer) para lo cual se obtendrá datos limpios, filtrados y definidos apropiadamente, estos datos limpios servirán como base para la construcción del modelo.

Adicional a esto, se realizará el análisis de los datos en descomposición y de proyección con el fin de observar tendencias.

3. Analítica avanzada

Las herramientas para la construcción del modelo en este proyecto se definieron la de Rapidminer por contar con la flexibilidad de adaptación a la aplicación desarrollada y que es la herramienta elegida en el capítulo anterior mediante una calificación objetiva.

4. Exploración de la información

Para la exploración de los datos obtenidos a partir de la aplicación del modelo de predicción elegido se usará el modulo para reportes de Tableau Online, por los motivos expuestos en el Capitulo II y enfáticamente debido a que la compañía EP FLOPEC cuenta con la licencia de este software y que facilitaría la observación de las predicciones de las diferentes rutas en comparación de los datos reales y demás observaciones que representan métricas para el usuario.

3.5 Desarrollo del modelo de análisis predictivo.

Para el desarrollo de la solución analítica se aplicara a la metodología de implementación planteada en este proyecto de titulación, la cual es CRISP-DM a continuación se describe cada una de las fases de esta metodología aplicada al modelo analítico para el proceso de predicción de precios de renta de buques tanqueros dentro de la compañía naviera FLOPEC EP, esta metodología nos permitirá establecer las relaciones que existen entre los datos y su utilidad para la compañía naviera de una manera normalizada y con actividades específicas que faciliten la implementación del modelo.

**Figura 30**

*Esquema de la metodología CRISP-DM*


*Nota*. Nicole, 2009

Fase 1: de Comprensión del negocio o problema

La primera fase de la guía de referencia CRISP-DM, denominada fase de comprensión del negocio o problema descrita en la Figura 31, es probablemente la más importante y aglutina las

tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. Sin lograr comprender dichos objetivos, ningún algoritmo por muy sofisticado que sea, permitirá obtener resultados fiables.

Fase 2: de Comprensión de los datos

La segunda fase descrita en la Figura 31, es la fase de comprensión de los datos, comprende la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, familiarizándose con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis.

Fase 3: de Preparación de los datos

En esta fase y una vez efectuada la recolección inicial de datos, se procede a su preparación para adaptarlos a las técnicas de Data Mining que se utilicen posteriormente, tales como técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para exploración de los datos.

Fase 4:  de Modelado

En esta fase de CRISP-DM, se seleccionan las técnicas de modelado más apropiadas para el proyecto de Data Mining específico. Las técnicas a utilizar en esta fase se eligen en función de los siguientes criterios:

- Ser apropiada al problema.
- Disponer de datos adecuados.

- Cumplir los requisitos del problema.
- Tiempo adecuado para obtener un modelo
- Conocimiento de la técnica.

Fase 5: de evaluación

En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerarse, además, que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. Es preciso revisar el proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso anterior, en el que se haya posiblemente cometido algún error.

Fase 6: de implementación

En esta fase y una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, ya sea que el analista recomiende acciones basadas en la observación del modelo y sus resultados, ya sea aplicando el modelo a diferentes conjuntos de datos o como parte del proceso, Generalmente un proyecto de Data Mining no concluye en la implantación del modelo, pues se deben documentar y presentar los resultados de manera comprensible para el usuario, con el objetivo de lograr un incremento del conocimiento.

Una vez que se ha realizado una descripción de cada una de las fases de la metodología que se va a implementar, en este proyecto de titulación, a continuación, se describe el desarrollo y aplicación de cada una de las fases del proyecto:

**Comprensión del negocio.**

La compañía naviera EP FLOPEC ( Flota Petrolera Ecuatoriana) cuenta con una flota de buques petroleros que son los encargados de transportar el crudo producido en el Ecuador a diferentes puertos del mundo, por lo que, el objetivo del negocio es recibir una comisión por la renta de buques tanqueros que transporten el crudo, esta comisión esta referenciada en base a los costos establecidos a nivel mundial de los diferentes tipos de buques y la ruta que tomara el buque para transportar este crudo.

Por lo que la proyección de precios de buques tanqueros es fundamental para la compañía naviera, estos valores proyectados sirven para definir los parámetros de renta de buques, establecer la ruta más rentable, y de igual manera establecer el buque más rentable.

Las herramientas usadas actualmente por la compañía naviera se limitan al uso de una hoja de calculo donde se establecen los históricos de los precios de buques tanqueros. Por lo que al generar una proyección de los costos de renta de un buque tanquero motiva la solución planteada en este proyecto de análisis avanzado de datos.

En esta fase se mantuvieron algunas reuniones de trabajo vista que, parte de las funciones desarrollas en la compañía son de habitual desarrollo para las funciones inherentes al trabajos desarrollado en EP FLOPEC, y como parte del equipo de trabajo se puede establecer las mejores condiciones para conocer el negocio de la compañía naviera.

**Determinación de los objetivos del negocio.**

El principal objetivo de la Compañía naviera EP FLOPEC es obtener la mayor rentabilidad a partir de la renta de un buques tanqueros para transportar productos derivados del petrolero, para realizar este transportes existen varios factores que se deben tomar en cuenta como son la ruta, tipo de buque, producto que transportara, por lo que, al tener la posibilidad de proyectar el

precio de un buque tanquero en determinado periodo de tiempo futuro, la compañía naviera tendrá la posibilidad de tomar la mejor decisión en cuanto a cerrar contratos de viaje y direccionar de mejor forma la optimización de rentabilidad de los buques de la flota.

**Estudio y comprensión de los datos**

En el negocio naviero se cuenta con el conocimiento de cuáles son los valores que se pagan por transitar en rutas definidas, de igual manera se cuenta con los tiempos de vida útil de un buque tanquero, los valores del mercado y estos datos demuestran cómo se comportan de acuerdo a las diferentes épocas del año.

Los datos con los que se cuenta, son los generados durante una cierta cantidad de años, para este proyecto de titulación se define un tiempo de colección de datos aproximadamente a 5 años, los cuales son del histórico diario del precio de la renta de un buque tanquero para las rutas alrededor del mundo, estos datos se generan semanalmente por lo que se tiene al menos 52 datos por año y esto por 5 años.

El beneficio de usar la minería de datos nos permitirá generar un análisis, con el objeto de definir un algoritmo que intérprete el comportamiento de los valores del precio de renta de buques tanqueros y pueda servir para establecer una proyección del precio del mercado para un tiempo futuro determinado, teniendo en consideración un margen de error aceptable.

El dataset al que se tiene acceso, fue provisto por el departamento comercial de EP FLOPEC (Empresa Publica Flota Petrolera Ecuatoriana), estos datos establecen los precios de renta de buques anualmente para referenciar los contratos de transporte de crudo dentro de la flota de buques, la información tiene un inicio en enero del 2015 y finaliza en noviembre del 2019, este dataset responde a la Tabla 6 de *Nota.*s de datos:

**Tabla 6**

*Fuentes de datos*

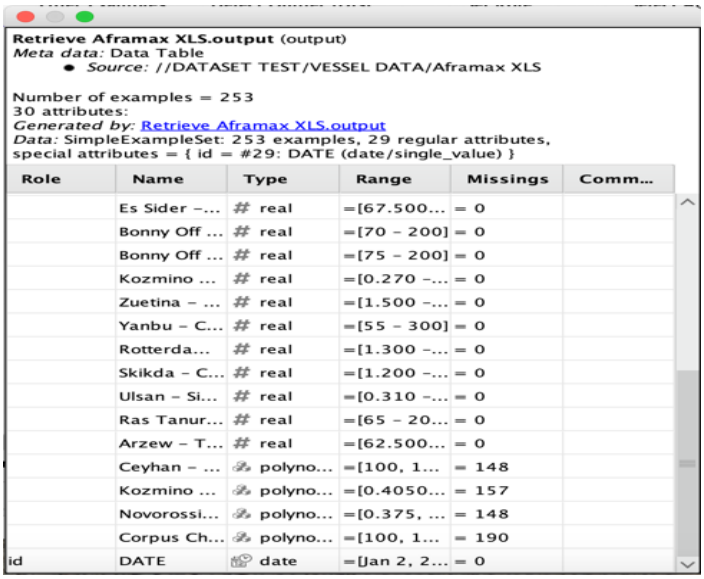
| Tema información | Canal | Archivo |
|---|---|---|
| Comercial (Valores comerciales, índice de precio de renta de buques.) | Página web https://www.clarksons.com/ | Excel, csv |
| Tiempo ( escala en semanas ) | Página web https://www.clarksons.com/ | Excel, csv |
| Rutas navieras | Página web https://www.clarksons.com/ | Excel, csv |

*Nota.* Gerencia Comercial, EP FLOPEC, 2019

Así también, se presenta la figura resumen de metadata generado en la herramienta analítica (Rapidminer):

**Figura 31**

*Metada del Dataset, precios de buques*



Y a continuación se realiza un resumen explicativo de cada uno de los campos del dataset en la siguiente Tabla:

**Tabla 7**

*Descripción de campos del Dataset*

| Campo | Descripción |
|---|---|
| DATE | Fecha de registro de valor de precio de renta un buque tanquero |
| TIPO_BUQUE | Tipo de buque, puede ser aframax, handymax, suezmax, panamax. |
| RUTA | |
| WORLSCALE | Indice de renta de un buque tanquero alrededor del mundo |

*Nota.* Dataset, EP FLOPEC 2019

Al realizar el análisis de los diferentes campos y su relación teniendo con consideración el planteamiento de los objetivos del proyecto se procede a realizar el histograma que relaciona el tiempo con el precio de renta de un buque (DATE vs PRICE_WORLDSCALE_RATES) para un determinado tipo de buque en una ruta determinada, estableciendo así la siguiente Figura:

**Figura 32**

*Índice de precio de renta de buques aframax en función del tiempo*

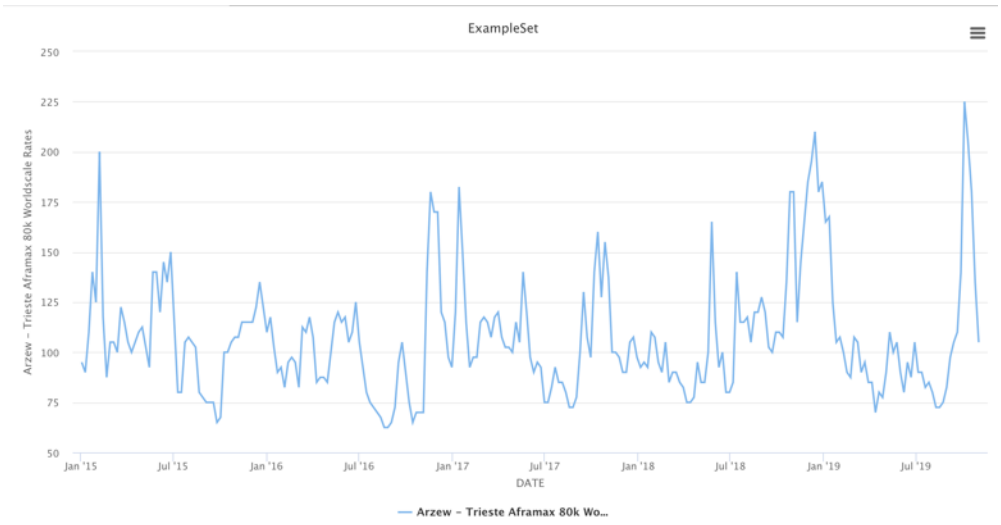



**Figura 33**

*Estadística básica de rutas en Rapidminer*


Conforme a lo verificado en la gráfica se tiene que existen diferentes variaciones dentro de los intervalos de tiempo y las variables del campo que serán útiles para el objetivo del proyecto será el precio de renta de un determinado tipo de buque. (DATE, RENT_PRICE).

Cada uno de los valores por renta de buque se encuentra definidos en el tiempo, por lo que se podría escoger un algoritmo de series temporales.

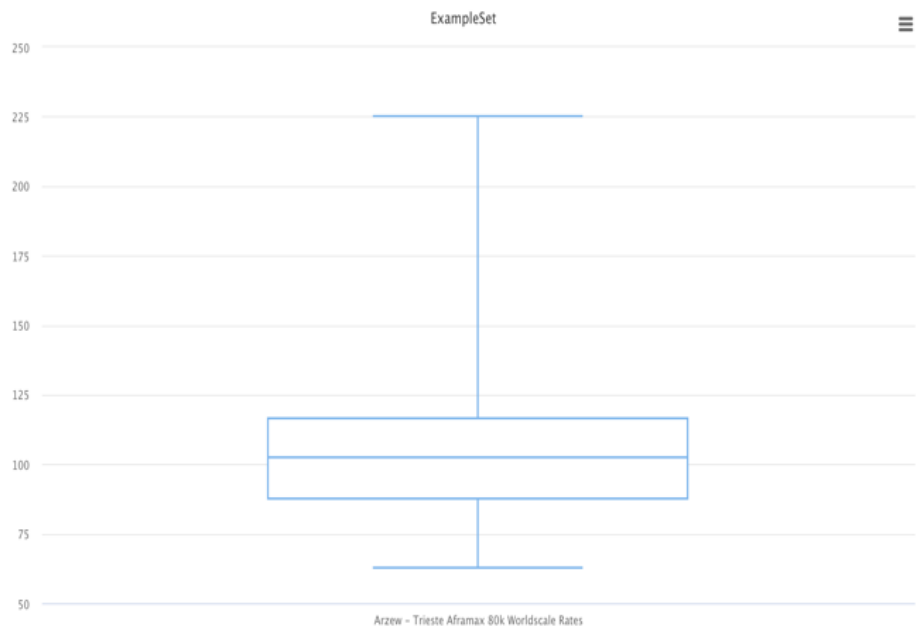
Para el análisis estadístico básico de las diferentes rutas se realizó un análisis de caja a la ruta que se escogió como ruta para el test o de referencia que es la ruta Arzew, Algeria – Trieste, Italia (Figura 35.) y se obtuvo los siguientes detalles:

**Figura 34**

*Ruta Arzew (Argelia)  hasta Trieste (Italia)*


**Figura 35**

*Diagrama de caja aplicado a la ruta Arzew a Trieste*



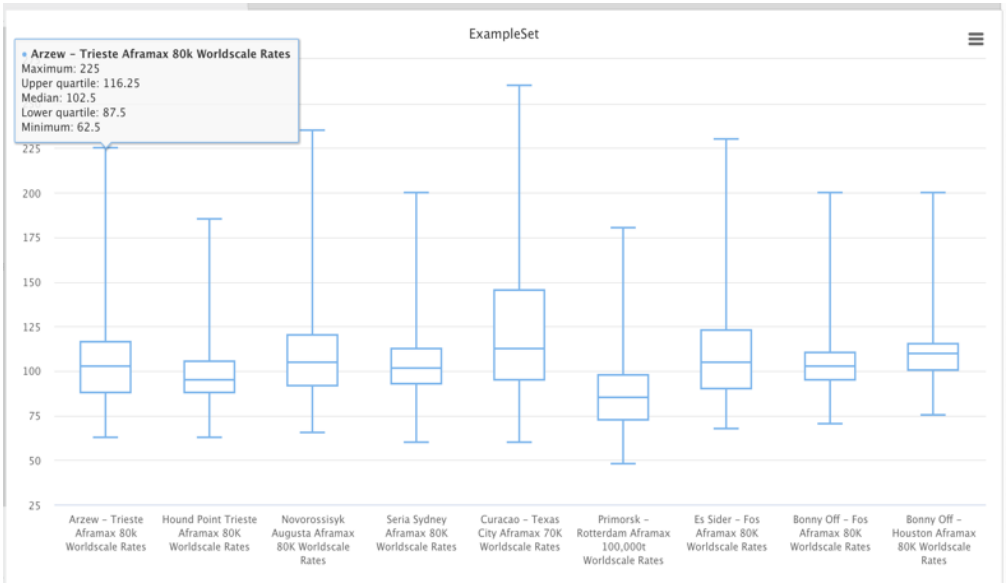

Del análisis de caja realizado a los datos de la ruta de Arzew-Trieste de la Figura 36. se tienen las siguientes conclusiones:

- Todos los datos de valor WS_Rates tienen un valor menor a 250
- Al menos el 75 % de los datos se encuentran en un rango menor del valor WS_Rates igual a 125
- El valor máximo de los datos es 225 y el valor mínimo de los datos se encuentra entre el rango de 75 y 50 WS_Rates
- Se tiene una mediana que se aproxima al valor de 100.

Este mismo comportamiento se repite en las siguientes rutas descritas en la Figura 37:.

**Figura 36**

*Diagrama de caja de otras rutas*

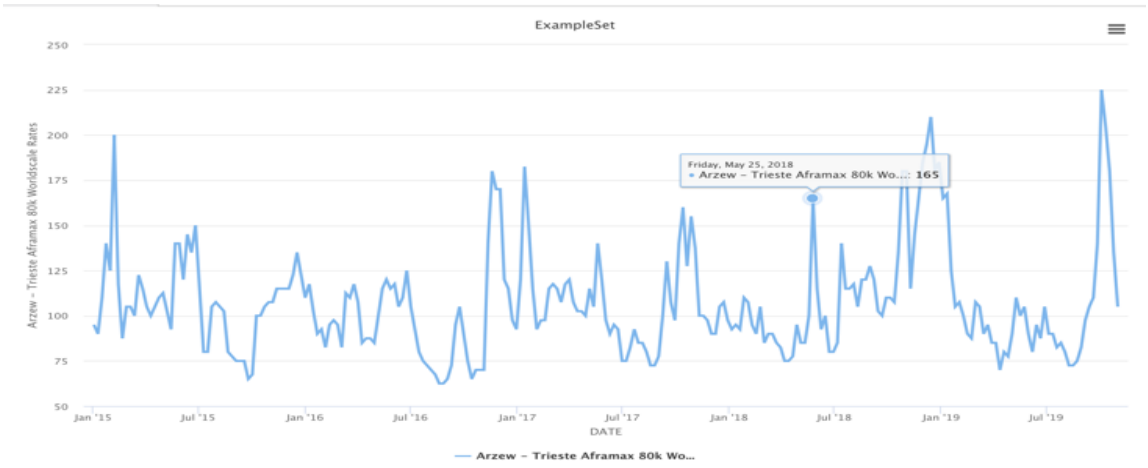



En conclusión, cualquiera de las rutas podrá ser tomada como referencia para la construcción del modelo, debido a que cada ruta cuenta con sus características propias de comportamiento, y que no es lo mismo realizar una ruta en el atlántico desde Curazao hasta Holanda, debido a factores como a millas náuticas recorridas, días de duración del viaje, volumen de combustible a ser consumido, etc.

Conforme a la gráfica obtenida para un tipo de buque, en este caso tipo aframax en la ruta Arzew -Trieste se establecen parámetros de World Scale Rates en ciertas épocas se elevan a un punto máximo y en septiembre del 2017 este alcanza un máximo de 225 por lo que se pensaría afecta a la gráfica, una vez conversado con los expertos en el mundo naviero conforme a la experiencia en el manejo de estos datos para establecer contratos, este salto importante dentro del comportamiento del dataset y en los precios es normal y es tomado en cuenta para el establecimiento de contratos por renta de buques, es decir, que si en esa semana se cerró un contrato de renta de buque con un valor alto, se tomara en cuenta este valor siendo un acierto y ventaja para el armador o compañía naviera.

**Figura 37**

*Grafica de los datos en función del tiempo para verificar su comportamiento.*




Como parte del análisis del dataset, es importante conocer los eventos que pudieron motivar esta elevación del precio de buques tanqueros como por ejemplo en los meses de Mayo - Junio del 2018, se estableció que el presidente de USA Donald Trump se reuniría con su homólogo el Presidente de Corea del Norte Kim Jong, lo que provoco un pico en la valor de renta de buques tanqueros. (Semana, 2018)

Por las razones antes expuestas se elige la ruta Arzew – Trieste por ser una ruta estándar y que cuenta con todos los factores más óptimos de contratación de rutas alrededor del mundo.

**Preparación de los datos**

Para la preparación de los datos en la herramienta de software Rapidminer se utilizaron los siguientes operadores propios de la aplicación:

- Aplicación del operador “Set Role”, para definir cuál es la variable dependiente y la variable independiente.

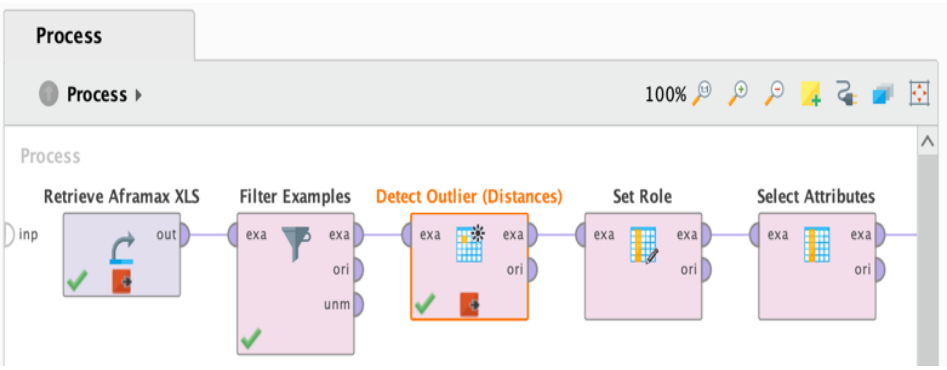
En nuestro caso la variable independiente está definida por la fecha y la variable dependiente será el valor WS de cada una de las rutas definidas en el dataset.

- Mediante la aplicación del operador “Filter Examples” se eliminación de registros fuera del rango, valores iguales a “0” y valores en blanco.
- Mediante la aplicación del operador “Select Attributes” establecemos la ruta referencial.

Estos operadores y su arreglo se encuentran visualmente descritos en la Figura 39.

**Figura 38**

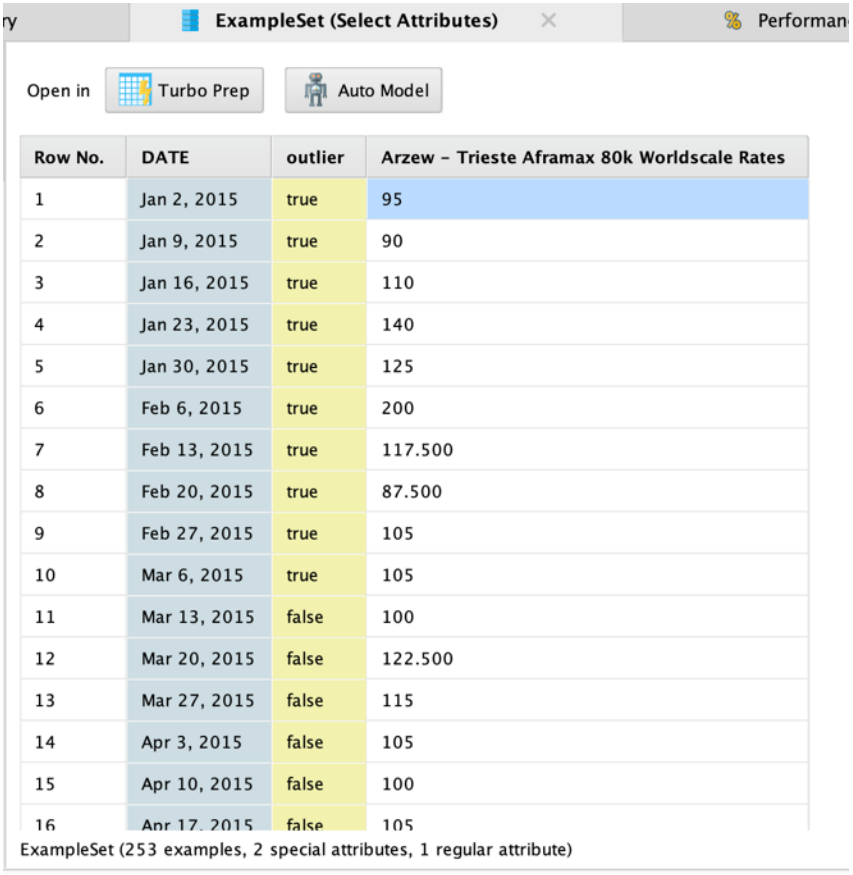
*Aplicación de operadores para limpieza de datos*



Una vez aplicados todos estos operadores de tratamiento de datos se obtiene como dataset la siguiente Figura 40. ejemplo de datos:

**Figura 39**

*Dataset después de la limpieza de datos*




| Row No. | DATE | outlier | Arzew – Trieste Aframax 80k Worldscale Rates |
|---|---|---|---|
| 1 | Jan 2, 2015 | true | 95 |
| 2 | Jan 9, 2015 | true | 90 |
| 3 | Jan 16, 2015 | true | 110 |
| 4 | Jan 23, 2015 | true | 140 |
| 5 | Jan 30, 2015 | true | 125 |
| 6 | Feb 6, 2015 | true | 200 |
| 7 | Feb 13, 2015 | true | 117.500 |
| 8 | Feb 20, 2015 | true | 87.500 |
| 9 | Feb 27, 2015 | true | 105 |
| 10 | Mar 6, 2015 | true | 105 |
| 11 | Mar 13, 2015 | false | 100 |
| 12 | Mar 20, 2015 | false | 122.500 |
| 13 | Mar 27, 2015 | false | 115 |
| 14 | Apr 3, 2015 | false | 105 |
| 15 | Apr 10, 2015 | false | 100 |
| 16 | Apr 17, 2015 | false | 105 |

ExampleSet (253 examples, 2 special attributes, 1 regular attribute)

**Modelado**

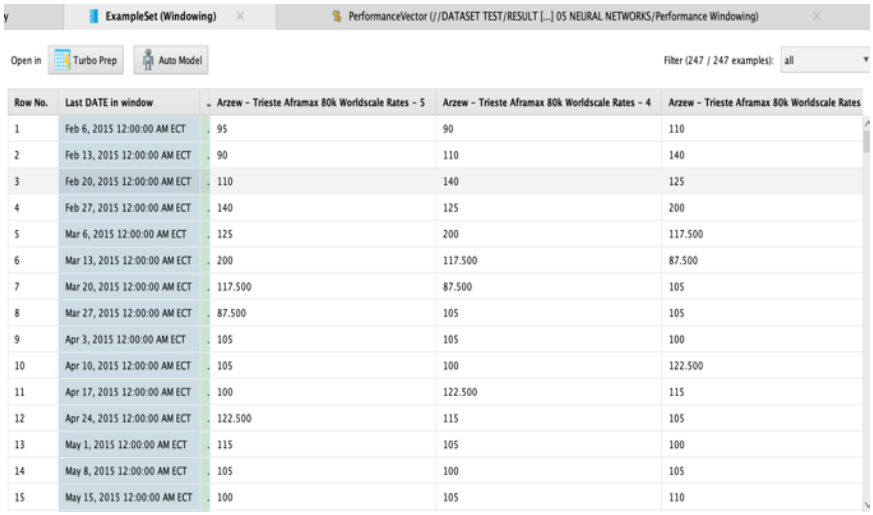
Toda vez que los datos han sido preparados en la herramienta analítica y que se ha definido el algoritmo modelo con el menor factor de error para nuestro dataset, y que es el de Linear Regresión, se procede a realizar el modelado para todos los datos disponibles del dataset:

Para lo cual se aplican los siguientes operadores lógicos de la herramienta analítica “Rapidminer”, esto se realiza de manera modular iniciando con el uso del operador

“WINDOWING” el cual se usa para construir la regresión polinomial en un dataset, y que pueda ser consumido en sus valores, para habilitar la posibilidad de que el atributo a ser pronosticado tenga un precedente, esto se demuestra en la Figura 41.

**Figura 40**

*Dataset después de aplicar "Windowing"*

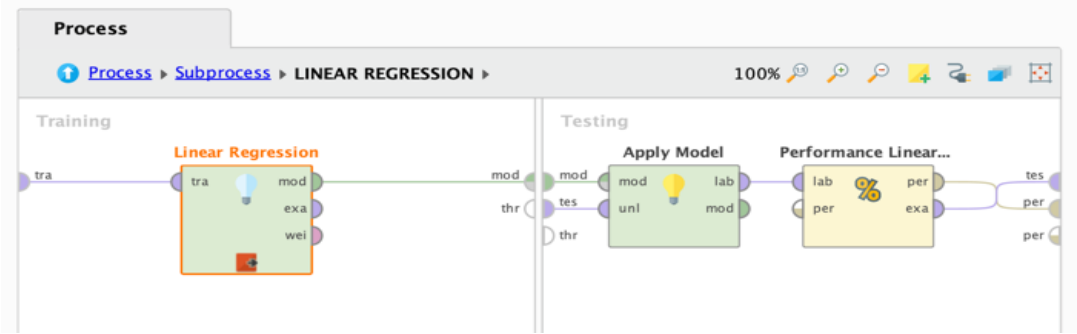



| Row No. | Last DATE in window | Arzew - Trieste Aframax 80k Worldscale Rates - 5 | Arzew - Trieste Aframax 80k Worldscale Rates - 4 | Arzew - Trieste Aframax 80k Worldscale Rates |
|---|---|---|---|---|
| 1 | Feb 6, 2015 12:00:00 AM ECT | 95 | 90 | 110 |
| 2 | Feb 13, 2015 12:00:00 AM ECT | 90 | 110 | 140 |
| 3 | Feb 20, 2015 12:00:00 AM ECT | 110 | 140 | 125 |
| 4 | Feb 27, 2015 12:00:00 AM ECT | 140 | 125 | 200 |
| 5 | Mar 6, 2015 12:00:00 AM ECT | 125 | 200 | 117.500 |
| 6 | Mar 13, 2015 12:00:00 AM ECT | 200 | 117.500 | 87.500 |
| 7 | Mar 20, 2015 12:00:00 AM ECT | 117.500 | 87.500 | 105 |
| 8 | Mar 27, 2015 12:00:00 AM ECT | 87.500 | 105 | 105 |
| 9 | Apr 3, 2015 12:00:00 AM ECT | 105 | 105 | 100 |
| 10 | Apr 10, 2015 12:00:00 AM ECT | 105 | 100 | 122.500 |
| 11 | Apr 17, 2015 12:00:00 AM ECT | 100 | 122.500 | 115 |
| 12 | Apr 24, 2015 12:00:00 AM ECT | 122.500 | 115 | 105 |
| 13 | May 1, 2015 12:00:00 AM ECT | 115 | 105 | 100 |
| 14 | May 8, 2015 12:00:00 AM ECT | 105 | 100 | 105 |
| 15 | May 15, 2015 12:00:00 AM ECT | 100 | 105 | 110 |

Una vez obtenido estos valores, se procede a aplicar un subproceso donde se aplica a este dataset con escenarios, el algoritmo “Linear Regresión”, el cual es aplicado para entrenamiento y validado por el operador de comportamiento y errores “Performance Linear Regresión”, donde se obtienen los errores como son error absoluto, error relativo, y coeficiente de correlación.

**Figura 41**

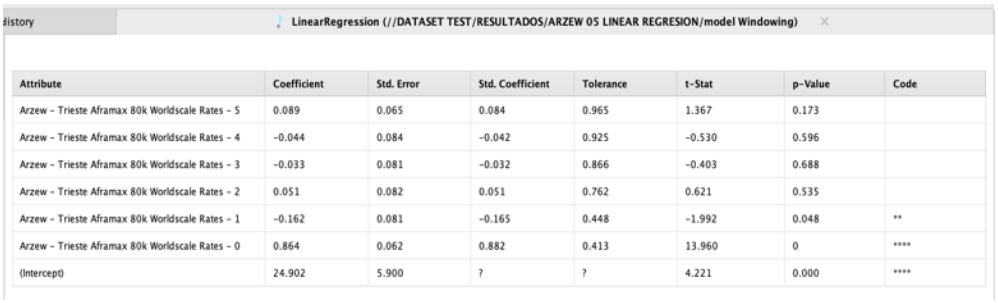
*Modelo Linear Regresión aplicado al dataset.*



Donde los resultados de la aplicación del modelo arrojan la siguiente ecuación polinomial:

**Figura 42**

*Modelo matemático polinomial*



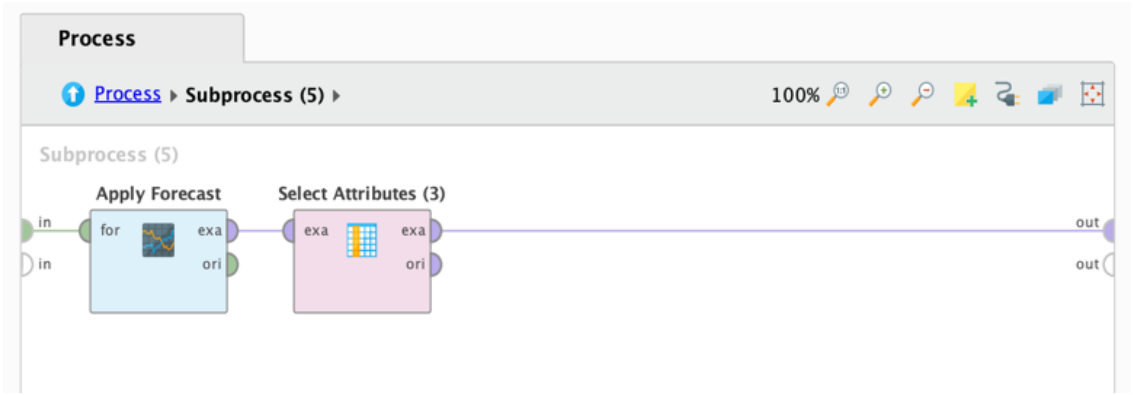
LinearRegression (//DATASET TEST/RESULTADOS/ARZEW 05 LINEAR REGRESION/model Windowing)

| Attribute | Coefficient | Std. Error | Std. Coefficient | Tolerance | t-Stat | p-Value | Code |
|---|---|---|---|---|---|---|---|
| Arzew – Trieste Aframax 80k Worldscale Rates – 5 | 0.089 | 0.065 | 0.084 | 0.965 | 1.367 | 0.173 | |
| Arzew – Trieste Aframax 80k Worldscale Rates – 4 | -0.044 | 0.084 | -0.042 | 0.925 | -0.530 | 0.596 | |
| Arzew – Trieste Aframax 80k Worldscale Rates – 3 | -0.033 | 0.081 | -0.032 | 0.866 | -0.403 | 0.688 | |
| Arzew – Trieste Aframax 80k Worldscale Rates – 2 | 0.051 | 0.082 | 0.051 | 0.762 | 0.621 | 0.535 | |
| Arzew – Trieste Aframax 80k Worldscale Rates – 1 | -0.162 | 0.081 | -0.165 | 0.448 | -1.992 | 0.048 | ** |
| Arzew – Trieste Aframax 80k Worldscale Rates – 0 | 0.864 | 0.062 | 0.882 | 0.413 | 13.960 | 0 | *** |
| (Intercept) | 24.902 | 5.900 | ? | ? | 4.221 | 0.000 | *** |

Una vez aplicado el modelo y con la ecuación matemática (Figura 43) que describe el comportamiento de los datos se procede a aplicar el operador “Forecast” para predecir los elementos del Atributo “Arzew – Trieste Aframax 80k Worldscale Rates”, para nuestro proyecto se ha definido que la variable tiempo se encuentre en semanas, y los valores a predecir serán de 8 semanas. Para esto se define el tiempo de estacionalidad en 52 semanas para la correcta aplicación de la predicción.

**Figura 43**

*Módulo de Pronostico*




Una vez aplicado el módulo de predicción (Figura 44.) se obtienen los valores:

**Figura 44**

*Dataset resultado de la aplicación del módulo de pronóstico*

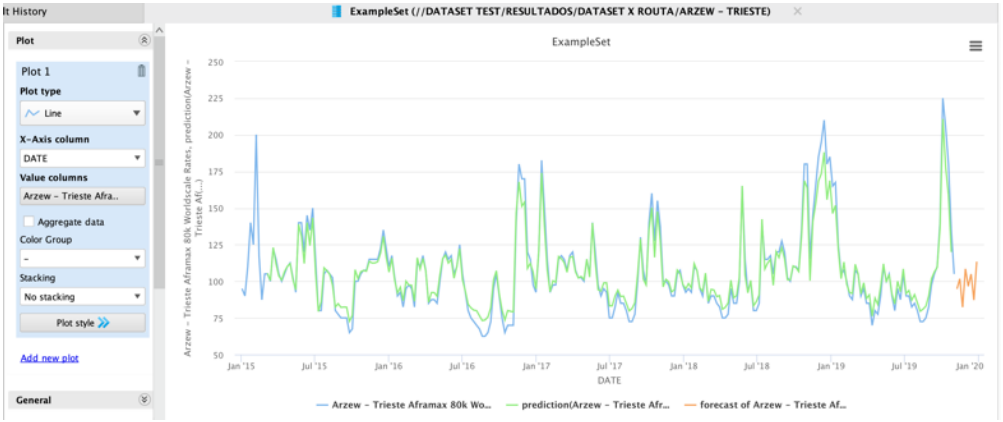
| Row No. | id | forecast of Arzew – Trieste Aframax 80k Worlds... | Arzew – Trieste Aframax 80k Worldscale Rates | DATE |
|---|---|---|---|---|
| 248 | id_248 | ? | 140 | Sep 27, 2019 12:00:00 AM ECT |
| 249 | id_249 | ? | 225 | Oct 4, 2019 12:00:00 AM ECT |
| 250 | id_250 | ? | 205 | Oct 11, 2019 12:00:00 AM ECT |
| 251 | id_251 | ? | 180 | Oct 18, 2019 12:00:00 AM ECT |
| 252 | id_252 | ? | 135 | Oct 25, 2019 12:00:00 AM ECT |
| 253 | id_253 | ? | 105 | Nov 1, 2019 12:00:00 AM ECT |
| 254 | id_254 | 94.778 | ? | Nov 8, 2019 12:00:00 AM ECT |
| 255 | id_255 | 101.702 | ? | Nov 15, 2019 12:00:00 AM ECT |
| 256 | id_256 | 82.395 | ? | Nov 22, 2019 12:00:00 AM ECT |
| 257 | id_257 | 108.487 | ? | Nov 29, 2019 12:00:00 AM ECT |
| 258 | id_258 | 96.651 | ? | Dec 6, 2019 12:00:00 AM ECT |
| 259 | id_259 | 104.898 | ? | Dec 13, 2019 12:00:00 AM ECT |
| 260 | id_260 | 87.263 | ? | Dec 20, 2019 12:00:00 AM ECT |
| 261 | id_261 | 113.409 | ? | Dec 27, 2019 12:00:00 AM ECT |

ExampleSet (261 examples, 2 special attributes, 3 regular attributes)

En la Figura No. 45 se puede observar la predicción obtenidas en el dataset para la ruta Arzew – Trieste, en la cual se predice las 8 semanas desde Noviembre 8, 2019 a Diciembre 27, 2019, obteniéndose:

**Figura 45**

*Resultado del pronóstico de la Ruta Arzew-Trieste*



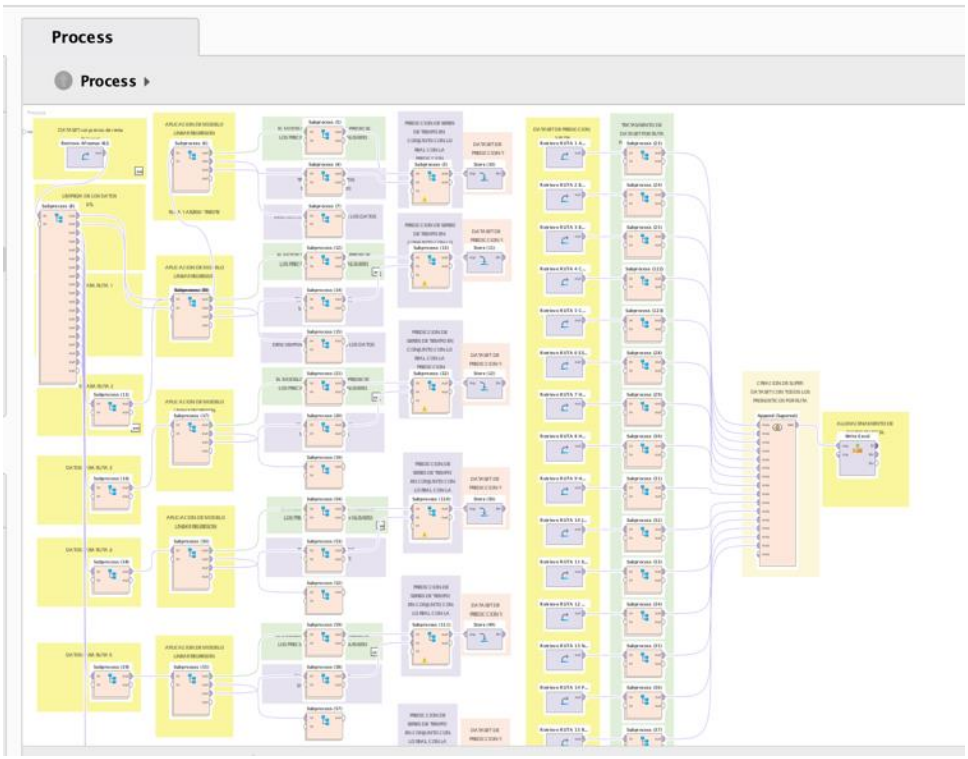

Como se puede observar en la Figura No. 46 la línea marcada con azul son los valores reales del dataset ruta Arzew – Trieste desde enero del 2015 a Noviembre 1 del 2019, la curva marcada con verde es la curva resultado de la predicción en base al modelo Linear Regresión, la

curva en naranja corresponde a los valores de la predicción de las 8 semanas desde Noviembre 8, 2019 a Diciembre 27 del 2019.

Este proceso se desarrolló para el 50% de las rutas establecidas en el dataset, a continuación, se muestra la aplicación (Figura 47 )para pronóstico de 15 rutas del dataset:

**Figura 46**

*Aplicación desarrollada para predicción de 15 rutas*




Una vez se aplica este desarrollo se obtiene un archivo en Excel, consumible en nuestra aplicación de exploración de datos, para nuestro proyecto se estableció como herramienta de exploración a Tableau, herramienta que fue escogida debido a que la compañía se encuentra en proceso de adquisición de esta plataforma para reportes.

**Evaluación o Validación del modelo aplicado a una ruta**

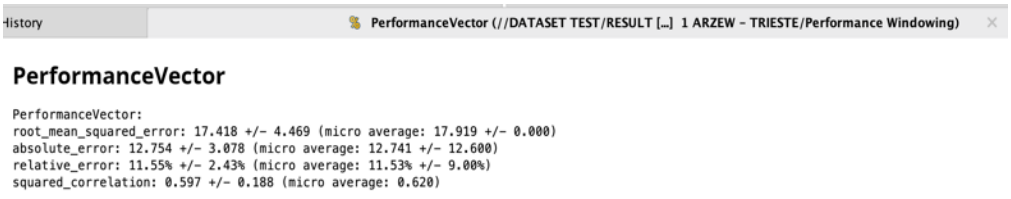
Una vez se aplicado del modelo y su predicción, las métricas obtenidas para la validación del modelo, se obtienen en base a la aplicación del operador "Performance" y sus respectivos resultados estadísticos que se definieron como se demuestran en la siguiente figura 48.

Para la validación del modelo se aplicó la técnica de "Validación Cruzada" de K iteraciones o "K-fold cross-validation", técnica donde los datos de muestra se dividen en K subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento. El proceso de validación cruzada es repetido durante k iteraciones, con cada uno de los posibles subconjuntos de datos de prueba.

Finalmente se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado, este método es muy preciso puesto que evaluamos a partir de K combinaciones de datos de entrenamiento y de prueba, para nuestra validación del modelo se eligió el número de iteraciones más común y es utilizar la validación cruzada de 10 iteraciones. (Pérez-Planells, Ll, 2015).

**Figura 47**

*Métricas del Modelo Linear regresión aplicado a la ruta Arzew – Trieste*




Así también se estableció, el valor de validación de la predicción en base al error entre los datos reales y los datos pronosticados con la siguiente formula:

$$PA = 1 - Error\ Abosluto/100$$

$$PA = 1 - \frac{12.74}{100} = 0.872$$

Entonces se puede concluir que la precisión del modelo a los datos pronosticados tiene un PA = 87.2 % ( PA: Precisión de la Predicción)

Este valor obtenido significa que, el margen de precisión de la predicción del precio de renta de un buque tanquero tiene un precisión del 87.2% y que el margen de error del valor calculado es de 12% aproximadamente, valores que son aceptables para la proyección a corto plazo en el mercado de buques tanqueros, ya que en comparación al valor calculado manualmente, aplicando el promedio mensual como por ejemplo para el mes de octubre del 2017 se tiene un PA=0.55 ( PA: Precisión de la Predicción), lo que quiere decir que se tiene un 45% de la precisión de la predicción, por lo que se establece que los valores obtenidos por el modelo propuesto ha mejorado la precisión de la predicción.

**Despliegue de la información.**

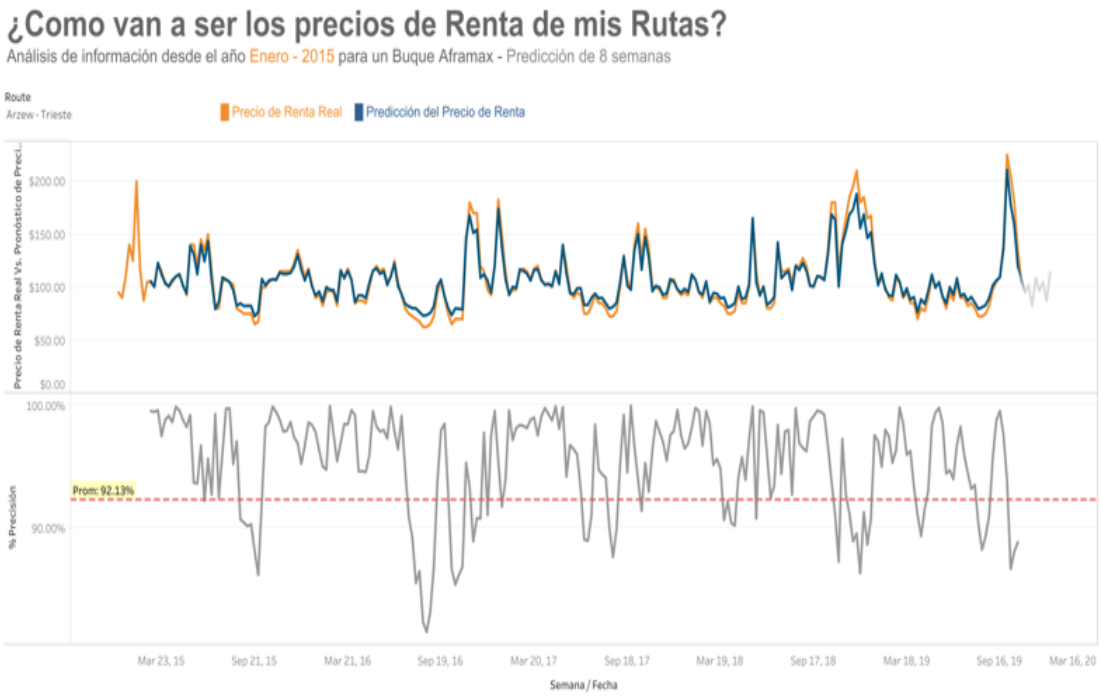
Con el fin que el usuario pueda comprender el análisis de datos y la predicción de datos, se utiliza la herramienta Tableau Software para realizar hacer una análisis simple, que permita hacer un despliegue y exploración de los datos pronosticados.

Esta exploración permite realizar un arreglo donde se muestren cada una de las rutas pronosticadas y que el usuario tenga la posibilidad de navegar en cada una de las rutas, una vez establecida una ruta deseada, el usuario puede visualizar los datos reales históricos de la ruta seleccionada, y el pronóstico de 2 meses con un escalamiento de datos semana a semana.

Adicionalmente se visualiza la curva donde con el promedio de la precisión del pronóstico, como muestra la figura a continuación:

**Figura 48**

*Grafica en la herramienta de exploración Tableau Software*



## Capitulo IV. Validación de resultados del modelo de análisis predictivo

Una vez terminada la aplicación de la metodología propuesta CRISP-DM en el proyecto, así como el procesamiento y aplicación del modelo de analítica avanzada a la empresa naviera y sus datos, se procede a realizar un análisis para validar la exactitud del modelo de predicción; esta validación se realizará mediante el análisis de cada uno de los resultados de la aplicación del modelo a un grupo de rutas, lo que permitirá la determinación del nivel de confianza del modelo, para determinar el nivel de confianza se realizarán aplicaran dos tipos validaciones:

- Validación mediante el uso de la herramienta de analítica avanzada (Rapidminer).
- Validación mediante uso de la herramienta de exploración de datos ( Tableau Online )

### Validación de la herramienta de analítica avanzada

Para la validación en la herramienta de analítica avanzada “RAPIDMINER” se utilizó el operador “Performance” el cual permite la calificación de los errores comunes en estadística, para definir el performance del modelo se establecieron los siguientes errores: raíz del error cuadrático medio, error absoluto, error relativo y coeficiente de correlación.

A continuación, se presenta el resumen los errores obtenidos para cada 5 rutas, los cuales se establecen en la siguiente Tabla 8:

**Tabla 8**

*Resumen de las validaciones del modelo en cada ruta.*

| RUTAS | DESCRIPCIÓN | ROOT MEAN SQUARED ERROR | ABOSLUTE ERROR | RELATIVE ERROR | SQUARED CORRELATION | Probability accuracy PA =( 1-EA%) |
|---|---|---|---|---|---|---|
| Ruta 1 | Arzew -Trieste | 17.418 | 12.754 | 11.55 | 0.597 | 0.87246 |
| Ruta 2 | Bonny Off - FOS | 9.688 | 6.675 | 6.675 | 0.729 | 0.93325 |
| Ruta 3 | Bonny Off - Houston | 10.232 | 6.937 | 6.05 | 0.679 | 0.93063 |
| Ruta 4 | Curazao - Texas City | 19.065 | 14.096 | 11.6 | 0.723 | 0.85904 |
| Ruta 5 | Curazao - Hamburg | 14.003 | 10.014 | 10.58 | 0.681 | 0.89986 |
| | **Promedio** | **14.0812** | **10.0952** | **9.291** | **0.6818** | **0.899048** |

Donde la precisión de la predicción “PA” tiene un promedio de 89.9 %, esto quiere decir que el modelo tiene un error de pronostico no mayor al 10%.

**Validación desde la herramienta de exploración de datos**

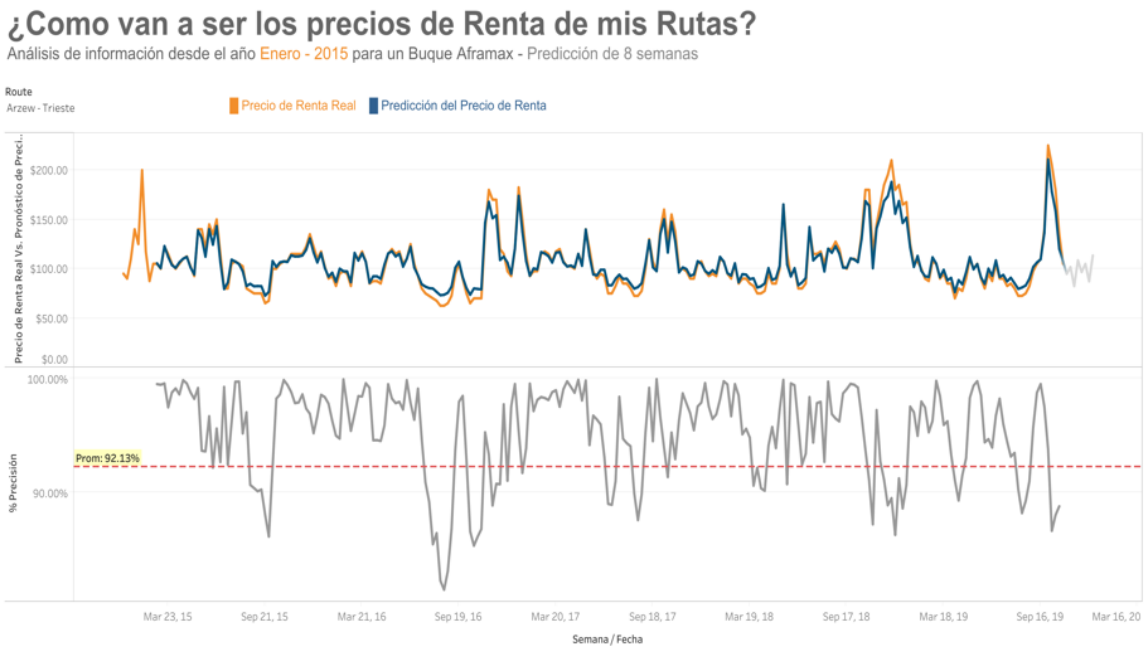
Para la validación mediante la herramienta de exploración de datos Tableau Software, se aplicó la fórmula de la precisión del pronóstico que se describe a continuación:

$$Precisión = \sum \frac{Valor\ pronosticado - Valor\ Real}{Valor\ Real}$$

Una vez aplicado esta fórmula al grupo de rutas de prueba se obtiene un cuadro como la Figura 49:

**Figura 49**

*Gráfica informativa de pronósticos de la ruta Arzew - Trieste*




El resultado de la aplicación de la fórmula a 5 rutas permite establecer la tabla 9, que se presenta a continuación:

**Tabla 9**

*Validación en herramienta de exploración*

| RUTAS | DESCRIPCIÓN | PRECISIÓN PROMEDIO |
|---|---|---|
| Ruta 1 | Arzew -Trieste | 92.13 |
| Ruta 2 | Bonny Off - FOS | 94.38 |
| Ruta 3 | Bonny Off - Houston | 94.31 |
| Ruta 4 | Curazao - Texas City | 92.84 |
| Ruta 5 | Curazao - Hamburg | 91.76 |
| | **Promedio** | **93.084** |

De la Tabla 9, se puede determinar que la precisión promedio del pronóstico tiene un promedio de 93.08 %, esto quiere decir que el modelo tiene un error de pronostico no mayor al 10%.

## Capitulo V. Conclusiones y recomendaciones

### Conclusiones

- El proceso de revisión de literatura planteado como parte de los objetivos fue necesaria, debido a que mediante este proceso de investigación se definieron los conceptos de analítica avanzada; la clasificación y comportamiento de los tipos de algoritmos a ser utilizados, nos permitió definir como realizar el análisis del comportamiento de los datos, y el uso de herramientas de análisis y exploración de datos, las cuales ayudaron a poder definir las mejores alternativas de aplicación.

- Se desarrolló el modelo de predicción de precios de renta de buques tanqueros tipo aframax utilizando la herramienta analítica Rapidminer, haciendo uso del algoritmo de regresión lineal, todos los elementos constituyentes del modelo fueron integrados de forma lógica y sucesivamente tomando en consideración la metodología CRISP-DM, al final se validaron los datos obtenidos del pronóstico de dos métodos en base a los errores estadísticos y en base a la precisión de la predicción resultando una exitosa implementación.

- Se construyó el modelo de pronóstico, haciendo uso de la analítica avanzada con base en modelos y se mejoró el proceso de proyección de ingresos por renta de buques tanqueros en una compañía naviera, logrando un porcentaje de precisión de la predicción en la herramienta de analítica avanzada de 89.9% y en la herramienta de inteligencia de

negocio igual al 93.3%, por lo que se valida la hipótesis planteada y se asume como válida para proyectar los ingresos por renta de buques y mejorar la predicción de ingresos, y establecer la mejor referencia para una la adquisición de buques tanqueros y para la planificación de ingresos en la flota de una compañía naviera.

- El proyecto se complementó con la generación de un aplicativo que facilita el análisis visual por parte del usuario final, este fue desarrollado sobre la herramienta de inteligencia de negocios llamado Tableau Software, con el objetivo una mayor interacción con el usuario final.

- Es importante aclarar que los datos de los precios de renta de buques tanqueros son de propiedad privada de la compañía naviera y que para el presente proyecto fueron afectados para evitar interpretaciones erróneas.

**Recomendaciones**

- Al momento de establecer las necesidad de cualquier proyecto analítico se recomienda levantar correctamente los requerimientos y sobre todo prestar el tiempo necesario para definir el alcance, y evitar la presencia de cualquier evento que retrase la ejecución del proyecto o que causen falsas expectativas en el usuario final.

- El presente modelo implementado, para la empresa naviera EP FLOPEC podrá servir como base para proyectos analíticos de referencia, en el sector naviero del transporte de hidrocarburos.

- Es necesario tener en cuenta que la predicción se realizó con base a un histórico de los costos de los buques tanqueros alrededor del mundo, y se debe considerar que el mercado de transporte de hidrocarburos mediante buques tanqueros, es afectado directamente por factores de índole social, económico y político a nivel mundial, y que pueden causar cambios drásticos en el pronóstico de los precios.

- Se recomienda que para que cualquier análisis predictivo analítico se tenga al menos de tres a cinco años de información histórica de tal manera que el entrenamiento del modelo sea el óptimo.

**Referencias bibliográficas**


Boldt, L. C., Vinayagamoorthy, V., Winder, F., Schnittger, M., Ekran, M., Mukkamala, R. R., Lassen, N. B., Flesch, B., Hussain, A., & Vatrapu, R. (2016). Forecasting Nike's Sales using Facebook Data. 10. https://doi.org/978-1-4673-9005-7

Carlie Idoine, Peter Krensky, Alexander Linden, Erick Brethenoux. (2019, enero). Magic Quadrant for Data Science and Machine Learning Platforms. https://www.gartner.com/en/documents/3899464

Casimiro, P. G. (2009). Análisis de Series Temporales: Modelos ARIMA. 169.

Edgar Martínez. (2014). Consumer Risk Analytics. Consumer Risk Analytics. https://edgarmartinezq.wordpress.com/

Gartner, G. (2019). Compare KNIME vs. RapidMiner vs. SAS vs. TIBCO Software in Data Science and Machine Learning Platforms | Gartner Peer Insights. Gartner. https://gartner.com/market/data-science-machine-learning-platforms/compare/knime-vs-rapidminer-vs-sas-vs-tibco

Gregory Piatetsky, KDnuggets. (2007, enero). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. KDnuggets. https://www.kdnuggets.com/crisp-dm-still-the-top-methodology-for-analytics-data-mining-or-data-science-projects.html/

Gregory Piatetsky, KDnuggets. (2016, septiembre). Top Algorithms and Methods Used by Data Scientists. KDnuggets. https://www.kdnuggets.com/top-algorithms-used-by-data-scientists.html/

Linear regression. (2020). En Wikipedia. https://en.wikipedia.org/w/index.php?title=Linear_regression&oldid=951272799

María Maya Lopera. (2018). ARBOLES DE DESICION. UNIVERSIDAD EAFIT.

Pérez-Planells, Ll. (2015). Análisis de Validación Cruzada Caso practico. https://core.ac.uk/download/pdf/71051261.pdf

Piatetsky. (2018, enero). Software de data mining: Las mejores herramientas. IONOS Digitalguide. https://www.ionos.es/digitalguide/online-marketing/analisis-web/software-de-data-mining-las-mejores-herramientas/

Ramesh, R. (2017). Predictive Analytics for Banking User Data using AWS Machine Learning Cloud Service. 6.

Rivero, O. M. (2016). El Método de Pronóstico Holt-Winters. 12.

Ruiz, C. A., Basualdo, M. S., & Matich, D. J. (2019). Redes Neuronales: Conceptos Básicos y Aplicaciones. 55.

Ruslan Unhich. (2019). Usuario: Ruslan unhicch—EcuRed. https://www.ecured.cu/Usuario:Ruslan_unhicch

Semana. (2018, mayo). Los hechos que marcaron el mundo en 2018. Los hechos que marcaron el mundo en 2018. https://www.semana.com/mundo/articulo/los-hechos-que-marcaron-el-mundo-en-2018/596458

Simeone, O. (2018). A Very Brief Introduction to Machine Learning With Applications to Communication Systems. ArXiv:1808.02342 [Cs, Math]. http://arxiv.org/abs/1808.02342

TIBCO Data Science vs RapidMiner. (s/f). G2. Recuperado el 26 de abril de 2020, de https://www.g2.com/compare/rapidminer-studio-vs-tibco-data-science

Tiwari, S., Bharadwaj, A., & Gupta, D. S. (2017). Stock Price Prediction Using Data Analytics. 5. https://doi.org/978-1-5386-3852-1

Wyatt, F. B., McCarthy, J. P., Heimdal, J., Godoy, S., & Autrey, L. (2005). Utilización de una Regresión Logarítmica para Identificar el Umbral de Frecuencia Cardíaca en Ciclistas—G-SE / Editorial Board / Dpto. Contenido. PubliCE. https://g-se.com/utilizacion-de-una-

regresion-logaritmica-para-identificar-el-umbral-de-frecuencia-cardiaca-en-ciclistas-708-sa-V57cfb271790f0