SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

INSTITUTO TECNOLÓGICO DE CD. GUZMÁN

PROGRAMA DE MAESTRÍA EN CIENCIAS
DE LA COMPUTACIÓN

TESIS

TEMA:

ANALÍTICA DE DATOS PARA UN SISTEMA QUE AUTOMATIZA LA LOGÍSTICA EN LOS PROCESOS DE EXPORTACIÓN

QUE PARA OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

BLANCA EDITH GÓMEZ GÓMEZ

DIRECTORES:

M.C. CYNTHIA ALEJANDRA MARTÍNEZ PINTO

DRA. ROSA MARÍA MICHEL NAVA

M.C. RUBÉN ZEPEDA GARCÍA

CD. GUZMÁN JALISCO, MÉXICO, AGOSTO DE 2018

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Instituto Tecnológico de Cd. Guzmán

Cd. Guzmán, Jal. a 09/Agosto/2018

Oficio No. DEPI/59/18

ASUNTO : AUTORIZACIÓN DE IMPRESIÓN

C. BLANCA EDITH GÓMEZ GÓMEZ
N.C. M16290031

En cumplimiento con el documento normativo de las disposiciones para la operación de estudios de posgrado del Tecnológico Nacional de México y con base en la aprobación del Comité Tutorial comisionado para su revisión; la División de Estudios de Posgrado e Investigación le otorga la autorización de impresión de su trabajo de tesis intitulado:

"ANALÍTICA DE DATOS PARA UN SISTEMA QUE AUTOMATIZA LA LOGÍSTICA EN LOS PROCESOS DE EXPORTACIÓN"

dirigido por la **M.C. Cynthia Alejandra Martínez Pinto**, desarrollado como requisito parcial para la obtención del grado de Maestro en Ciencias de la Computación, de acuerdo al plan de estudios MCCOM-2011-05.

Sin otro asunto en particular, quedo de usted.

S.E.P. TecNM
INSTITUTO TECNOLOGICO
DE CD. GUZMAN
DIVISION DE ESTUDIOS
DE POSGRADO E
INVESTIGACION

A T E N T A M E N T E

DR. HUMBERTO BRACAMONTES DEL TORO
JEFE DE DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

C.p. Archivo

Agradecimientos

Agradezco a Dios las fuerzas para poder sacar adelante este proyecto personal.

A mi esposo y sobre todo a mis hijos Astrid, Braulio y Leonardo por la paciencia que me tuvieron para poder llevarlo a cabo, no fue fácil para ellos tener que cambiar su vida por completo por acompañarme en este nuevo reto que no fue únicamente para mí, se qué no fue fácil pero lo logramos.

A mis hermanas y demás familiares que me apoyaron con el cuidado de mis hijos para que fuera posible poder realizar este proyecto.

A mis compañeros en especial a Carlos Hernández por todos los momentos de risas, estrés, alegrías que compartimos durante este proyecto.

A mi directora de tesis la M.C. Cynthia Alejandra Martínez Pinto por la paciencia y la dedicación para concluir con el proyecto.

A la Doctora Rosa María Michel Nava por todo el apoyo brindado tanto como docente, como personal.

Al personal Administrativo y Docente que de alguna manera fueron piezas claves para terminar mi maestría.

A todos gracias pero muchas gracias, porque hoy concluyo una etapa académica.

ÍNDICE

Índice de figuras

Índice de tablas

Índice de cuadros

Capítulo I. Introducción

El Instituto Nacional de Estadística y Geografía (INEGI) y el Banco de México dieron a conocer el día 26 de enero de 2018 la información del comercio exterior mexicano y la balanza comercial de mercancías del país, donde informaron que las exportaciones de México hacia el resto del mundo cerraron el año 2017 con un aumento anual de 9.5%. Esto significa que el país alcanzó una cifra récord de exportaciones de 409,494 mdd en 2017. Tan sólo el valor de las exportaciones el mes de diciembre es de 2017 lo que sumó 35,825 mdd. (INEGI, 2018).

La empresa MCP Parnter se dedica a la exportación uno de sus objetivos es darse a conocer de manera internacional, por ello es necesario conocer que es lo que México exporta hacía el resto del mundo, para dar una mejor respuesta en tiempo y forma de los procesos de logística que requiere llevar una exportación.

El presente proyecto se divide en dos partes, primero el sistema que se entregó a la empresa que lleva el control de la logística para simplificar el envío de productos al resto del mundo, y segundo, la analítica de datos llevada a cabo para determinar si el retraso de la salida de un avión afecta en tiempo la entrega de la mercancía.

El objetivo que se siguió con la aerolínea fue de obtener, los diferentes tiempos en retrasos que van desde los 15 minutos hasta más de 4 horas o la cancelación total del mismo.

1.1 Antecedentes del problema

Zapotlán el Grande, Jalisco, es un municipio que está creciendo mucho en el ramo agrícola, debido a ello los productores buscan la manera de ampliar sus mercados, cada vez es más frecuente que realicen exportaciones a los Estados Unidos y la Unión Europea.

MCP Partner, es una empresa que se dedica actualmente a la exportación de diferentes productos perecederos y no perecederos cuyo destino es Internacional.

La empresa es joven, tiene casi cuatro años brindando este servicio, en su actual crecimiento quiere expandir su mercado a niveles internacionales.

El procedimiento de una exportación y/o importación, la empresa MCP Partner contacta al cliente, al obtener la información de la mercancía que se va enviar, se pregunta qué medio

le interesa al cliente (terrestre, aéreo o marítimo), se hace la reservación del transporte en el que se va a exportar y/o importar la mercancía, ya obtenida la reservación el cliente hace el depósito correspondiente, y se le da aviso a la empresa mandando la ficha de depósito o transferencia bancaria.

Ya realizados los pasos anteriores la empresa MCP Partner confirma la reservación y continúa con el proceso de la exportación, que consta de contactar al agente aduanal para verificar el producto que se va enviar. Así es como funciona la empresa MCP Partner para realizar una exportación.

Actualmente el internet es una herramienta que ha venido a simplificar el manejo de la información, es más rápido buscar información de cualquier tipo o contactar un servicio de una empresa sin ir a su domicilio fiscal (Martínez, 2016).

Uno de los servicios que se puede contactar a través de internet, es el servicio de exportación de productos a cualquier parte del mundo a través de empresas especializadas en la logística necesaria para este proceso, el cliente puede contactarlos sin necesidad de ir físicamente a las instalaciones.

A la empresa se le desarrolló el sistema web Mundi, para poder llevar a cabo las exportaciones, y con ello darse a conocer internacionalmente. Se analizó de manera particular el medio de transporte aéreo utilizando una base de datos de la plataforma Kaggle con los registros de todas la aerolíneas de E.E. U.U., para conocer cuales son los retrasos que se pueden presentar al momento de realizar la exportación, todo esto será detallando en los siguientes capítulos hasta llegar al resultado para conocer cuales fueron los tiempos de retraso que se pueden presentar y pueden llegar afectar la logística de una exportación.

En el capítulo II se ilustran las páginas que realizan el servicio de exportación y las definiciones teóricas que servirán de guía para llevar a cabo el proyecto.

En el capítulo III se trata del tipo de investigación se realizó, la muestra, instrumentos y aparatos utilizados.

En el capítulo IV. Se muestran la principales ventanas del sistema web Mundi, la analítica de la base de datos que se analizó y los resultados obtenidos.

Fuentes Consultadas. Se anexan las fuentes bibliográficas y sitios web consultados.

Glosario. Descripción de términos utilizados en el proyecto que son poco usuales.

1.2 Descripción del trabajo de investigación

Este proyecto surge de la necesidad de la empresa denominada Partner, la cual se dedica a la exportación de productos perecederos y no perecederos. En el proceso de logística se realizaba de forma manual lo que hace que se pierda mucho tiempo de respuesta entre la cotización de la naviera y tener una respuesta para el cliente, con el proyecto se pretende relizar el sistema para poder agilizar los procesos y no haya tanto tiempo de holgura y poder aumentar la cartera de clientes, además de contar con la analítica de datos para poder ofrecer al cliente un servicio de calidad y que la empresa sea reconocida a nivel mundial.

Las exportaciones se hacen por tres medios de transporte que son: aéreo, marítimo y terrestre. En este proyecto se decidió analizar el medio de trasporte aéreo y con ello todos las inconvenientes que pueden surgir al momento que se envíe el producto por este medio tales como los diferentes tiempos de retraso que van desde 15 minutos hasta 4 horas o puede ser la cancelación total del vuelo.

La solución que se obtuvo fue a través de Big Data con Analítica de la base de datos de 1998 al 2008 cuantos vuelos salieron a tiempo, cuantos con retraso y que tiempo fue lo que se retrasó, para poder considerar los tiempos de demora que existen al momento de hacer las exportaciones y qué tanto puede afectar para envíar los productos sobre todo tratándose de productos perecederos.

1.3 Definición del problema

La empresa MCP Partner, se dedica a los procesos de logística para llevar a cabo una exportación. Los procesos admistrativos se realizan de forma manual y uno de los objetivos

de la empresa que tiene es darse a conocer internacionalmente, para ello necesita contar con un sistema web que ayude a controlar los procesos de una manera más eficiente. Para que el cliente cada vez que realice una exportación, no tenga que estar dando información básica cuando se realice un trámite, que este sea más sencillo y rápido. A través del sistema el cliente podrá solicitar la información necesaria para realizar la exportación, que conozca cómo va el proceso en tiempo real, que conozca a través de un reporte cuáles son las exportaciones que ha realizado y qué mercancía ha enviado.

Por otro lado se realizó la analítica de datos en el transporte aéreo, teniendo el conocimento de los tiempos de retraso que pueden llegar a ocurrir ya sea por mal tiempo o por servicio de la aerolínea, eso da como resultado un tiempo de holgura en los procesos de logistíca que se deben de tomar en cuenta al momento de planear una exportación.

Para poder llevar a cabo la analítica, se utilizó una base de datos externa a la empresa con más de un millón de datos que cuenta con las característica de Big Data. Se tomó esta decisión debido a que la empresa no pudo solventar el acceso a las bases de datos de las empresas dedidacas al transporte, a demás, los registros que se generaron en el sistema Mundi son muy pocos para poder realizar la analítica y saber si las salidas con retraso de los aviones afecta la entrega de la mercancia.

1.4 Justificación

Según el estudio del MIT (Instituto Tecnológico de Massachusetts), durante los últimos 5 años las exportaciones en México se han incrementado un 5% anual, siendo los principales destinos: Estados Unidos de América, Canadá, Alemania, China, Japón y países bajos.

México es uno de los países con más Tratados de Libre Comercio, con un total de 12 acuerdos, que engloban a 46 países, así que la oportunidad de intercambio comercial es enorme.

No obstante la incertidumbre económica que se generó a principios del año 2017, y pese al inicio de la renegociación del Tratado de Libre Comercio de América del Norte (TLCAN), Jalisco proyecta mantener cifras positivas tanto en exportaciones como en Inversión

Extranjera Directa (IED) fortaleciendo las exportaciones hacia otros mercados internacionales, aprovechando que Jalisco “es el único estado del país que tiene 21 sectores exportadores, y son 181 países a los que están llegando” (IMCP.ORG.MX, 2018).

De acuerdo con el Instituto de Información Estadística y Geográfica del estado de Jalisco (IIEG), en el primer bimestre del 2017 la entidad exportó 8,036 millones de dólares, cifra que supone un crecimiento de 10.2% a la tasa anual, y Estados Unidos se mantuvo como el principal comprador de los productos jaliscienses al captar 64.5% del total de las ventas al exterior.

La empresa MCP Partner se dedica a la logística de las exportaciones de la región Sur de Jalisco, la mayor parte de este proceso se realizaba en forma manual con ello solicitando información que se repetía en diferentes formularios, los datos eran vulnerables al estar almacenados en hojas de Excel y esto producía pérdidas de información. A través de vía teléfonica se contactaban a los productores con las empresas que realizan el traslado de la mercancía a otros países , provocando que el proceso fuera más lento. El software solicitado por la empresa permite tener una mayor cartera de clientes tanto nacionales como internacionales así mismo reducir el tiempo de los trámites de la reservación y los trámites administrativos que conlleva una exportación.

Debido a que la empresa no continúo con el proyecto por falta de liquidez, ya que conectarse con las grandes navieras resultaba muy oneroso para la misma, se entregó un sistema Web que cumple con las necesidades administrativas solicitadas y la analítica de datos, se realizó con una base de datos libre, que estuviera relacionada con un medio de transporte utilizado en las exportaciones y el más idóneo al que se tuvo acceso fue a una base de datos en la plataforma Kaggle que contiene registrados los vuelos del año 1998 al 2008 lo cual permite realizar un buen análisis del comportamiento de las aerolíneas.

1.5 Objetivos de la investigación

1.5.1 Objetivo general

Desarrollar la analítica de datos para una Base de Datos de la plataforma de Kaggle llamada DelayedFlights, para conocer el impacto en tiempo real de llegada de los vuelos, a través de los intervalos de tolerancia basados en la hora programada con respecto a la hora de salida, con el fin de evaluar si se afecta la entrega a tiempo de la mercancía exportada.

1.5.2 Objetivos específicos

1. Buscar la base de datos del medio de transporte para poder realizar las exportaciones en el repositorio Kaggle.
2. Localizar la base de datos con las características más idóneas.
3. Según las características de la base de datos seleccionar el algoritmo para realizar la analítica de datos.
4. Realizar pruebas con una porción de datos en Weka para ver si el algoritmo es el apropiado.
5. Programar en R o Python el algoritmo de aprendizaje autómatico y determinar si los retrasos de los vuelos afectan los tiempo de entrega de la mercancía exportada.
6. Documentar los resultados obtenidos.

1.6 Hipótesis

Con analítica de datos se puede conocer el impacto en tiempo real de llegada de los vuelos, a través de los intervalos de tolerancia basados en la hora programada con respecto a la hora de salida.

1.7 Motivación

La principal motivación para realizar este proyecto fue un reto profesional y personal. Los trabajos que se realizaron anteriormente se habían enfocado en lo administrativo, esto fue de gran ayuda para tener una noción al momento de recabar la información, debido a que había procesos que resultaban familiares, gracias a la experiencia obtenida.

Las exportaciones era un tema totalmente desconocido, con este trabajo se despejaron muchas dudas y con ello se abrirán más puertas de trabajo al momento de egresar.

Realizar el proceso de analítica sirvió para reafirmar el concepto que, de cualquier dato se puede obtener información.

Capítulo II. Fundamento teórico

2.1 Estado del arte

Se revisaron diferentes empresas que se dedican a la logística de la exportación e impotación de mercancía al extranjero, entre las características principales que tienen estas empresas son: los tipos de empaques en el que se va a transportar la mercancía, cantidad de mercancía ya sea por kilo o libra, la fecha de envío y destino. Las empresas más reconocidas son las siguientes:

- Inttra.
- Icontainers.
- Gurucargo.
- 45hc.
- Flexport.
- Lotebox.
- Kontainers.

A continuación se muestran las principales ventanas del paso a paso como se puede realizar la reservación para poder llevar a cabo una exportación. Debido al sistema que se desarrolló se enfoca a exportaciones.

Inttra

Inttra (Inttra, 2018) es un portal donde las principales navieras del mundo pueden ofrecer sus servicios al cliente brindando una gama de empresas marítimas que le ayuden a optimizar tiempo y costos para la exportación de los productos.

La fortaleza de esta plataforma es que se pueden comparar precios y tiempos de entrega para brindar una mejor opción al cliente.

A continuación se muestran las ventanas con la información de los servicios que ofrecen y los pasos a seguir para hacer una importación:

Al navegar en la página de Inttra se aprecia el menú principal donde se indican las opciones con los servicios que se ofertan como se muestran en la figura 2.1:

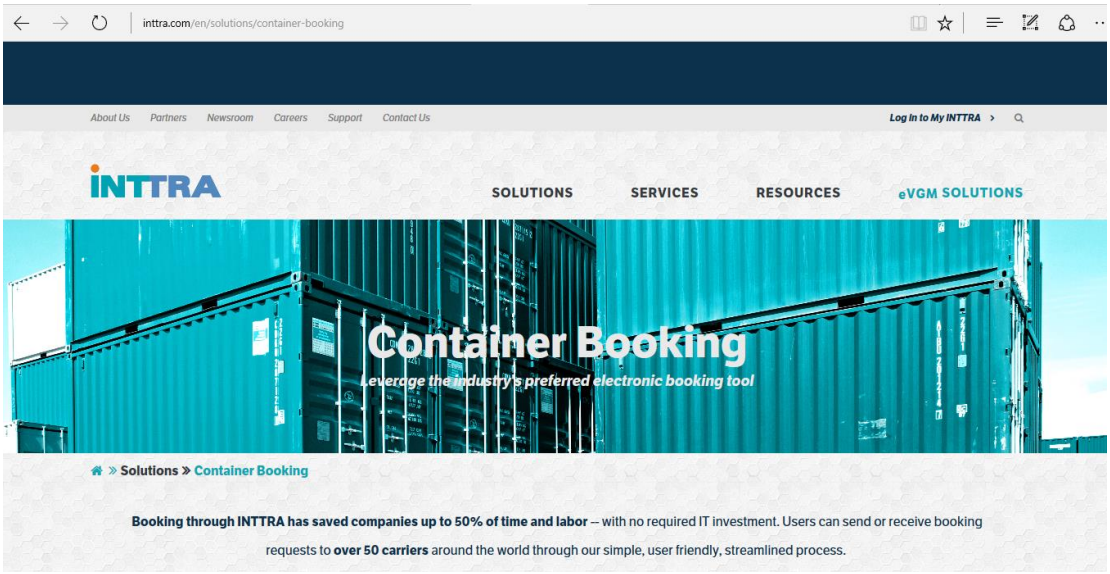


Figura 2.1. *Interfaz principal de Inttra.*

A continuación se muestra una serie de ventanas donde se puede visualizar el formulario con los requisitos para dar de alta un nuevo usuario:

Para que un nuevo usuario pueda darse de alta se cuenta con diferentes idiomas tal como se muestra en la figura 2.2, cabe señalar que las funciones están deshabilitadas, los idiomas son: inglés, chino, chino mandarín, francés, japonés, portugués, español y turco.

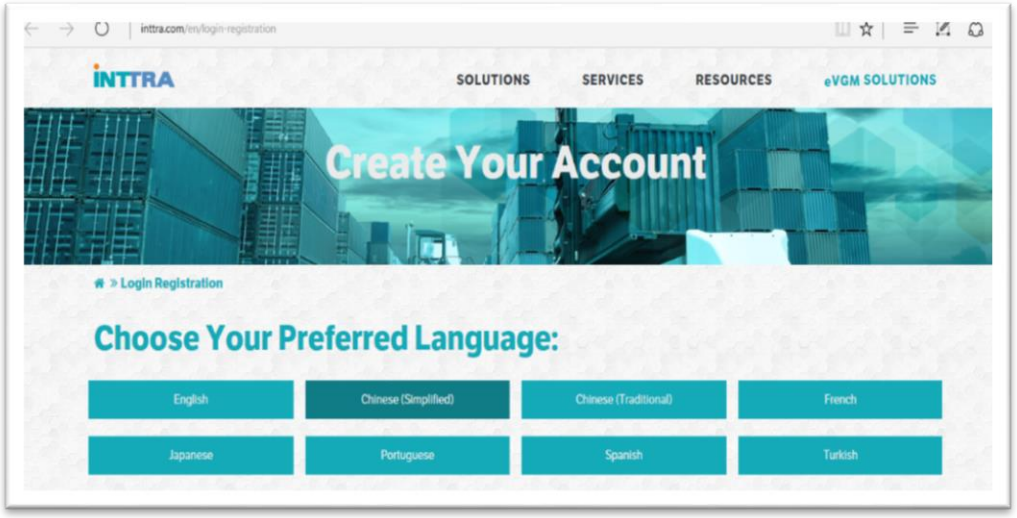


Figura 2.2 *Selección de idioma.*

En la figura 2.3 se muestran los datos a llenar por parte de la empresa para darse de alta por primera vez, los campos que contienen un asterisco son campos obligatorios a llenar, en el registro de industria se despliega un menú con los diferentes tipos de giro industrial por ejemplo: agricultura, automotriz, entretenimiento, etc., debe incluir el volumen que exporta y Tax Id (clave para pagar los impuestos o RFC).

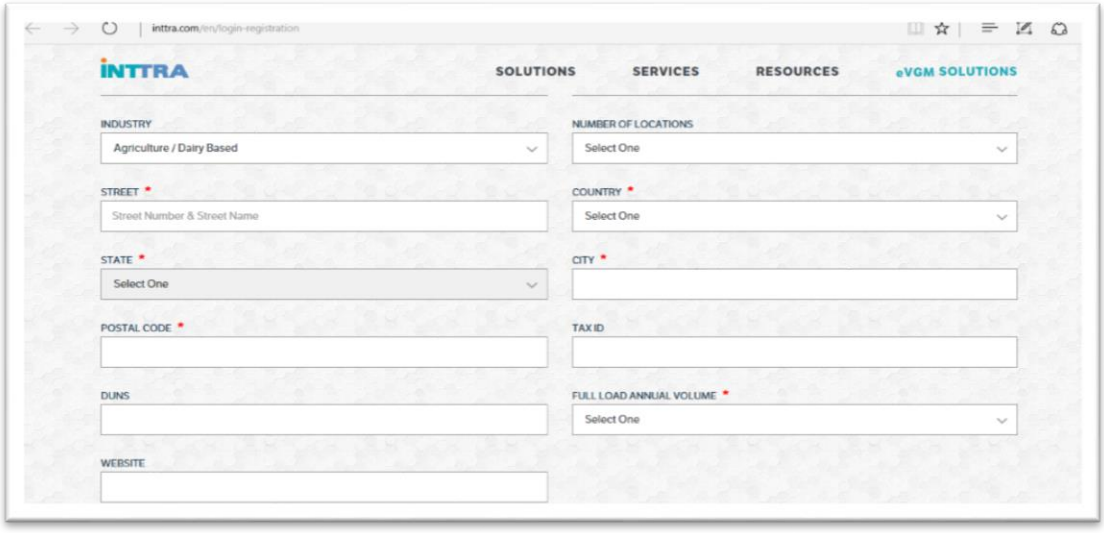
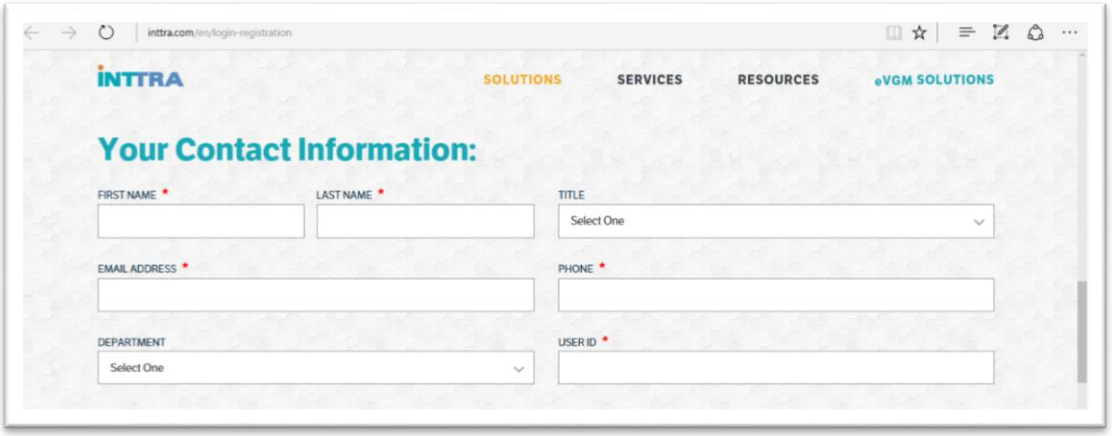


Figura 2.3 *Formulario para dar de alta un nuevo usuario.*

Como se puede observar en la figura 2.4 los campos a llenar son de la persona que va estar en contacto directo para realizar la exportación y/o importación, por eso es muy importante llenar los datos marcados con un asterisco ya que son esenciales para tener una formalidad y atención rápida con el cliente los cuales son el nombre, teléfono, correo, etc.

Figura 2.4. *Continuación del formulario.*



Una vez que se hayan llenado los registros correspondientes de las figuras 2.3 y 2.4, se tienen que aceptar los términos y condiciones de la empresa tal cual se muestran en la figura 2.5:

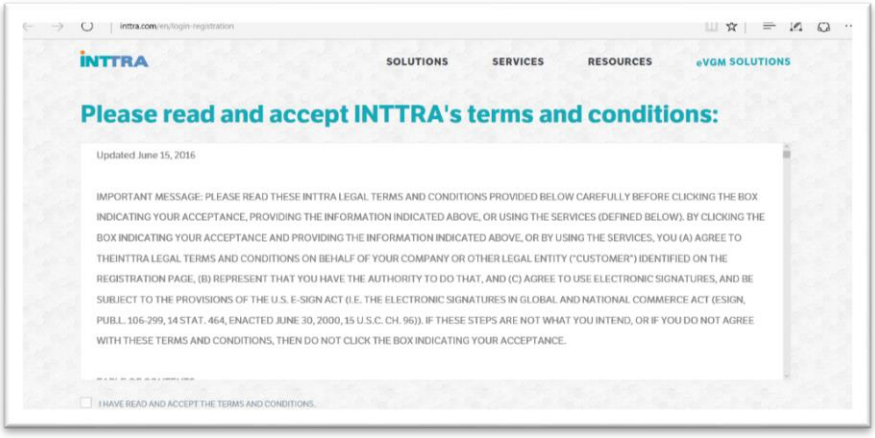


Figura 2.5. *Términos y condiciones.*

Después de llenar los datos del usuario, en la figura 2.6 se observa la ventana de Inttra, donde el cliente introduce su usuario y contraseña para poder entrar a hacer la reservación del contenedor para exportar y/o importar la mercancía, u otros movimientos.

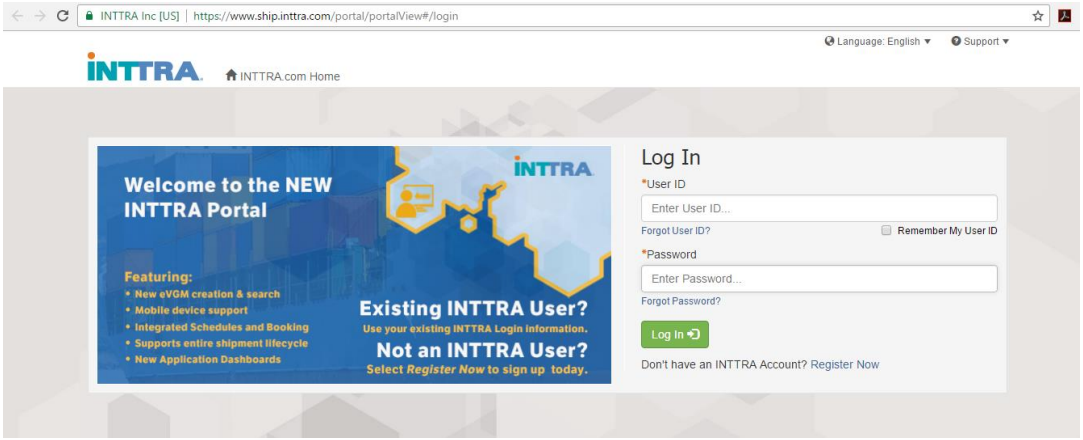


Figura 2.6. *Ventana de entrada al sistema.*

Una vez que se introducen los datos y se ingresa al sistema, se despliega el menú con las opciones que se pueden realizar, ya sea una nueva reservación, revisar el estatus de la importación y/o exportación tal como lo muestra la figura 2.7:

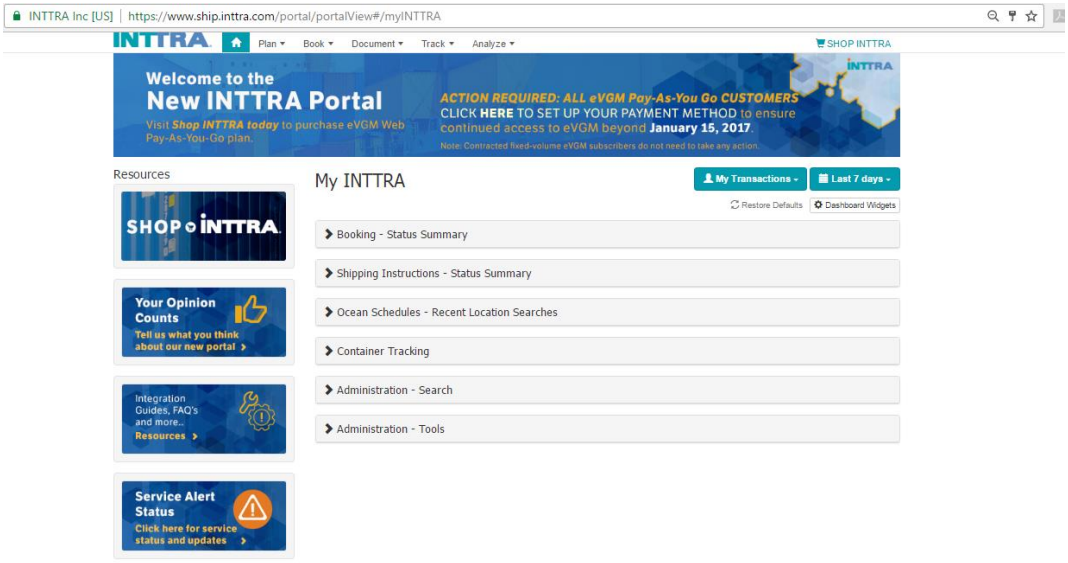


Figura 2.7*. Ventana menú de opciones de movimientos a realizar.*

En el menú principal del sistema tiene las opciones de Plan, Book, Document, Track, Analyze Para hacer una nueva reservación para exportar, seleccionar **book** y posteriormente **create new** (crear nuevo) tal como se muestra en la figura 2.8:

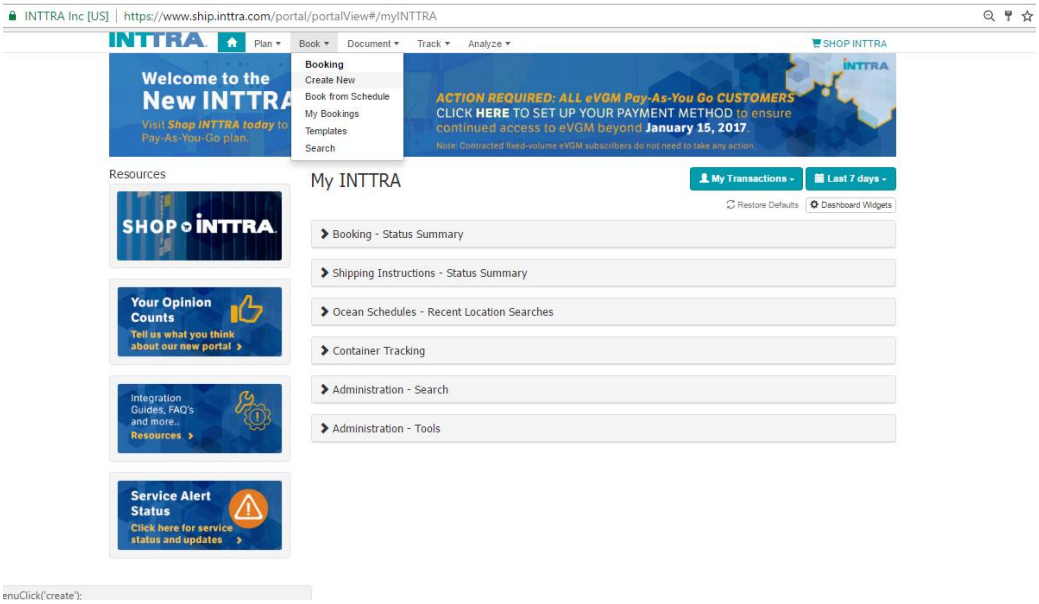


Figura 2.8. *Ventana para seleccionar el booking.*

En la figura 2.9 se muestra la ventana para hacer la reservación para una nueva exportación, los campos con asterisco son obligatorios a llenar, en la opción **carrier** son las navieras con las que actualmente trabaja con la plataforma.

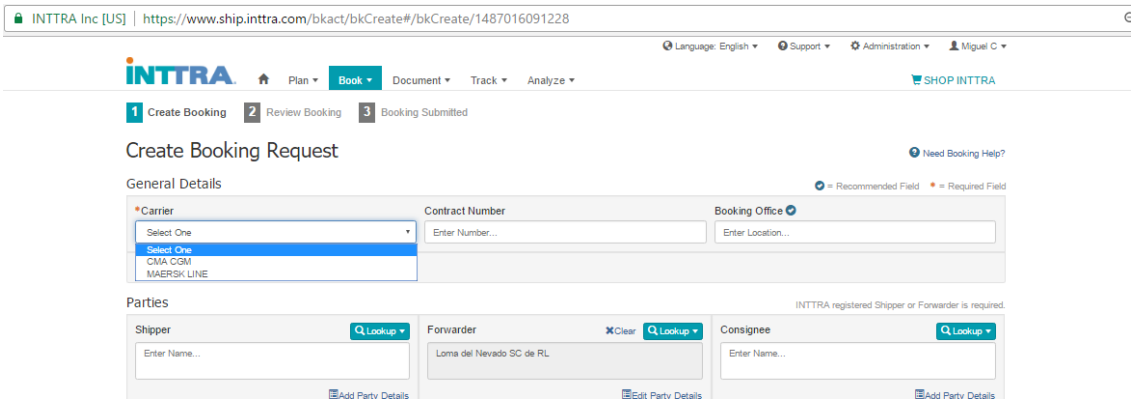


Figura 2.9. *Ventana para realizar una nueva reservación.*

Una vez seleccionada la naviera con la que se va a trabajar, para hacer el envío, en este caso ya lo da por Default, tal como lo muestra en la figura 2.10, se puede hacer algunas anotaciones adicionales.

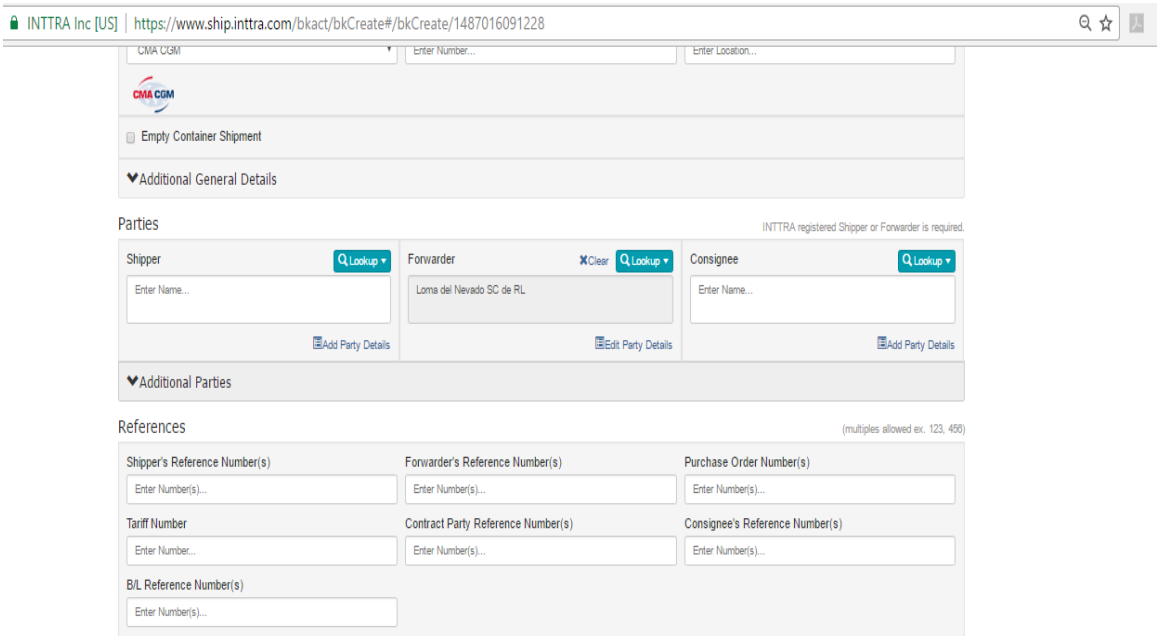


Figura 2.10. Ventana para indicar la empresa que va hacer el envío.

En la figura 2.11 se muestra la ventana para llenar los datos del transporte, en la casilla de **Move Type** o tipo de movimiento es donde se va a seleccionar si es importación o exportación, en la opción de **Pre carriage** se indica si se necesita un arrastre terrestre hacía el puerto marítimo.

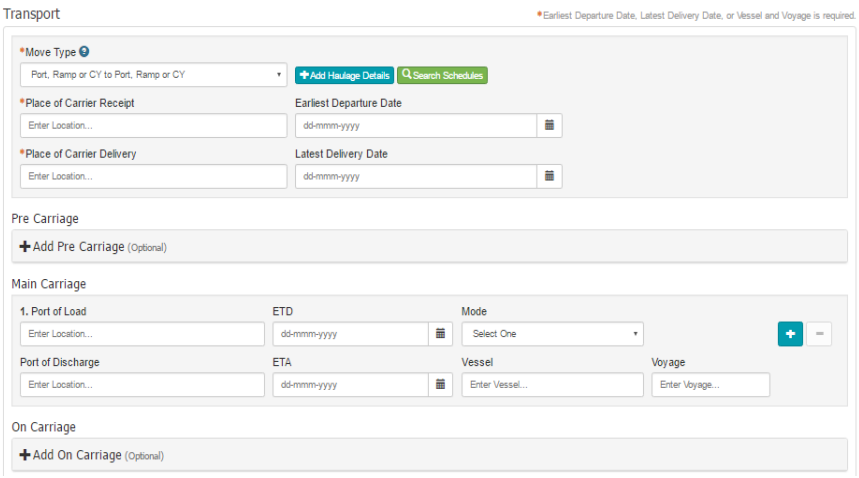


Figura 2.11. *Ventana para indicar si es importación o exportación.*

En la figura 2.12 en la opción de **container** (contenedor) el usuario selecciona la capacidad del contenedor que necesita y si requiere que tenga condiciones especiales como oxígeno, la temperatura debe de ir ciertos grados para que el producto llegue en buenas condiciones al destino, si es un material peligroso.

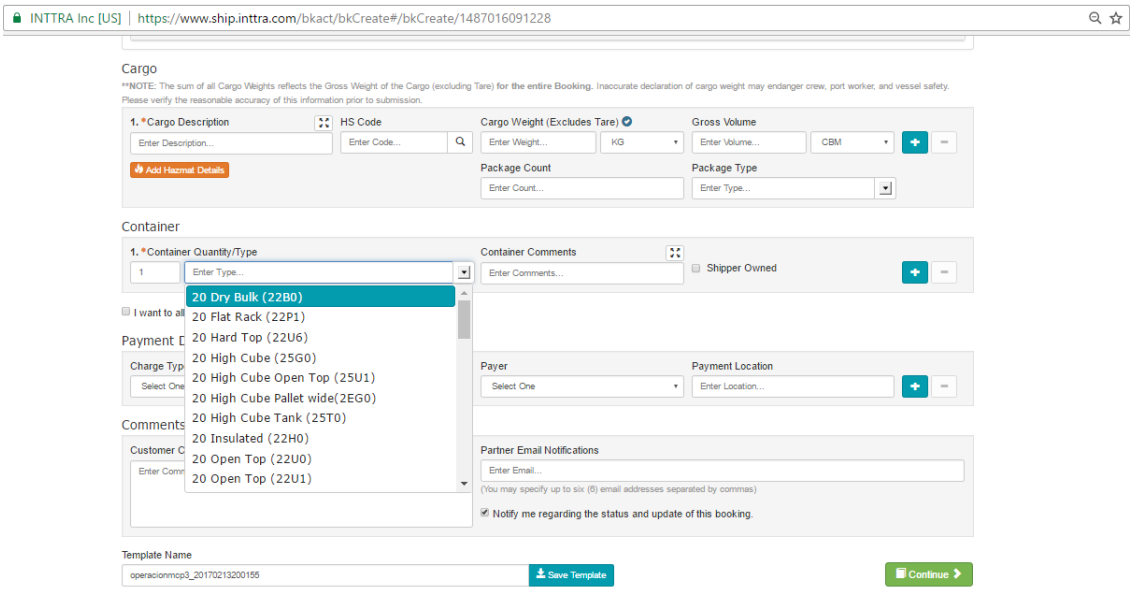


Figura 2.12. *Ventana para seleccionar el contenedor.*

Una vez completados los pasos para hacer el booking, en la parte inferior de la ventana en **Template Name** se indica un registro con la clave de usuario, la fecha, los dígitos para hacer referencia al booking. Esa clave es la referencia para rastrear la mercancía y conocer el estatus del envío, tal como se muestra en la figura 2.13.

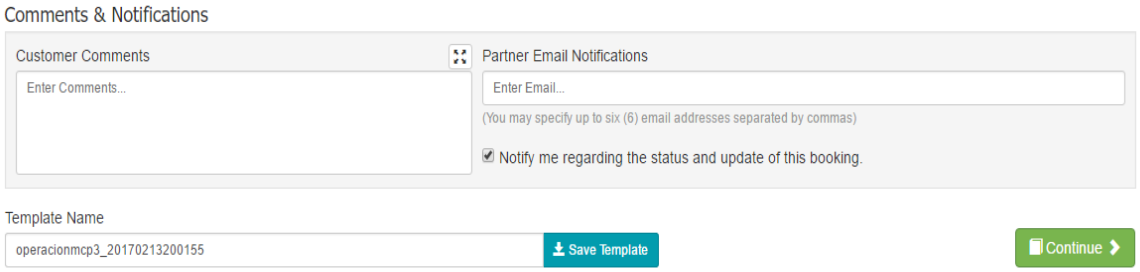


***Figura 2.13**. Ventana con el registro del booking*.

Conclusiones:

La empresa Inttra es una empresa muy completa para hacer importaciones y/o exportaciones, trabaja con todas las navieras, se pueden contactar para poder hacer envíos a cualquier parte del mundo, sin importar donde se encuentre el cliente, actualmente es la empresa intermediaria con la que trabaja MCP The World Partner.

La desventaja: el menú principal está muy saturado de información y la interfaz es poco amigable, confunde al usuario, tiene diferentes tipos de idiomas y no funcionan. Como es tanta información que necesitan mostrar al cliente, su página es poco amigable con el usuario, es mucha información la que se le presenta que se puede llegar a confundir.

ICONTAINERS

De acuerdo al sitio Web Icontaniners (2018), este es una empresa internacional dedicada a la exportación e importación ofrece variedad de embarques sin tener sucursal en el país, su interfaz es más amigable y es fácil de entender.

En las siguientes ventanas se muestra paso a paso cómo realizar una reservación ya sea para importar y/o exportar la mercancía. En la página no se necesita que el usuario esté registrado para poder cotizar la reservación para realizar una exportación y/o importación.

En la figura 2.14 se muestra la interfaz principal donde se puede seleccionar el idioma, si se quiere registrar o simplemente iniciar con la cotización.

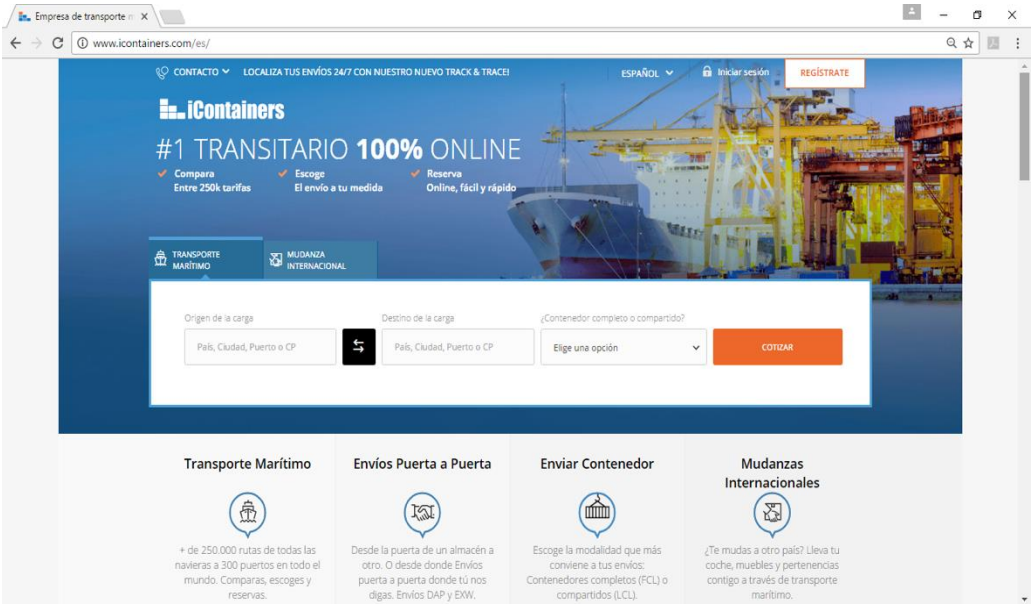


Figura 2.14. *Interfaz principal Icontainers.*

En la figura 2.15 se observa que al escribir el nombre de México en la opción origen de carga se despliega un menú con los puertos que hay en el país, así como un mapa donde se encuentra el puerto a seleccionar

.

.

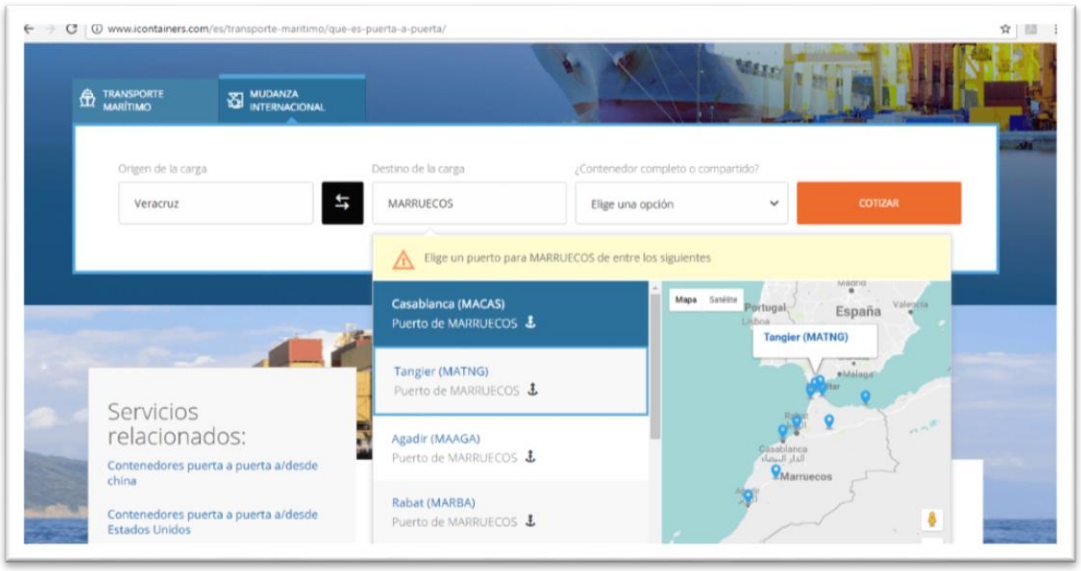


Figura 2.15. *Interfaz para seleccionar el origen de la carga*

Una vez que ya se seleccionó el origen de la carga el siguiente paso es seleccionar el destino, tal como se indica en la figura 2.16

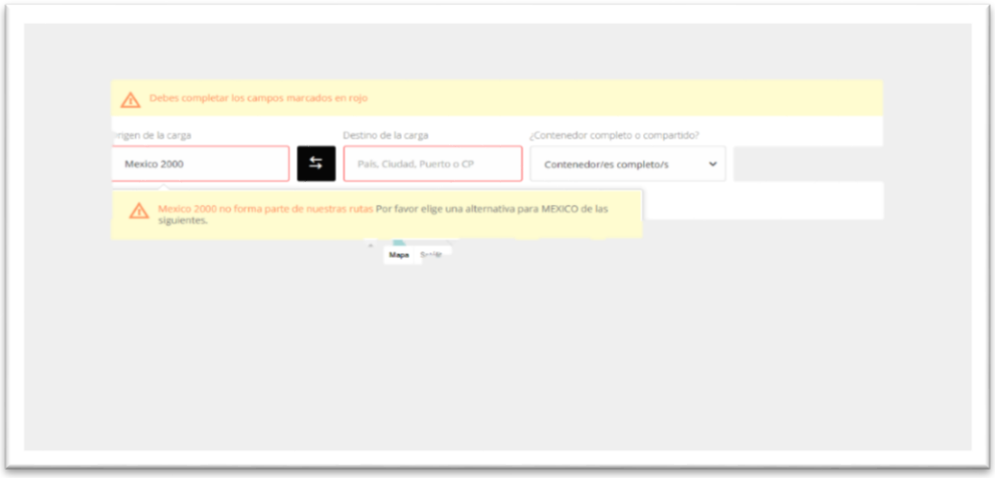


Figura 2.16. *Interfaz para seleccionar el destino de la carga***.**

.

Una vez que ya se eligieron los puertos de origen y destino, el siguiente paso es seleccionar el tipo de contenedor que se necesita para seguir con la reservación, en la figura 2.17 se hará la selección de un contenedor simple

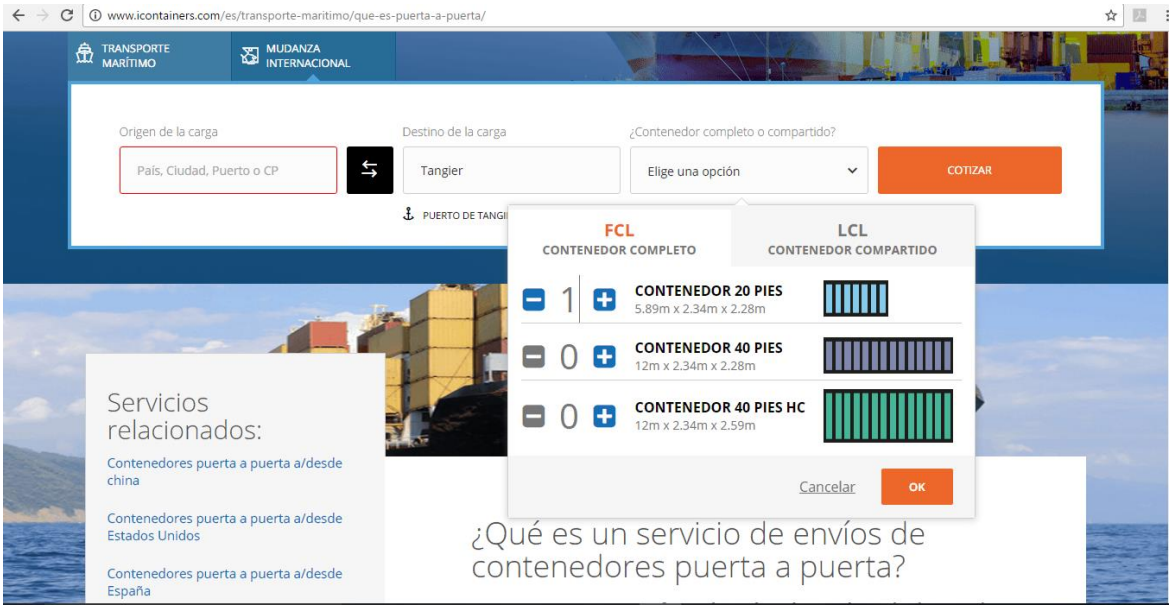


Figura 2.17 *Selección del contenedor.*

Al oprimir la tecla de **cotizar** aparece un mensaje tal como se muestra en la figura 2.18 donde indica que no existe la ruta seleccionada.

Figura 2.18*. Mensaje de la búsqueda del contenedor***.**

Se hicieron varias pruebas para seleccionar una ruta donde se pudieran cotizar el booking tal como lo muestra la figura 2.19.

Figura 2.19. *Ventana con la cotización del booking.*

En la figura 2.20 la interfaz muestra las fechas y horarios de las salidas en puerto así como la fecha límite para entregar documentación.

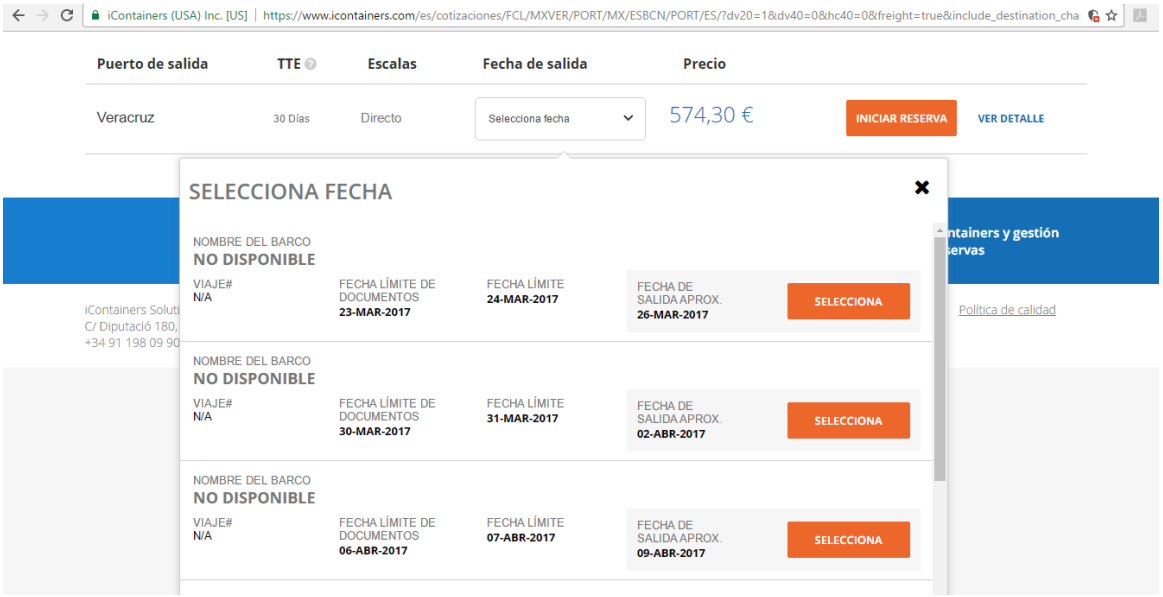


Figura 2.20. *Horarios y fechas de salida de la exportación.*

Una vez realizados los pasos anteriores, para seguir con la reservación se solicita al cliente que se registre o si ya es cliente ingrese con sus datos para darse de alta como nuevo usuario, tal como se muestra en la figura 2.21:

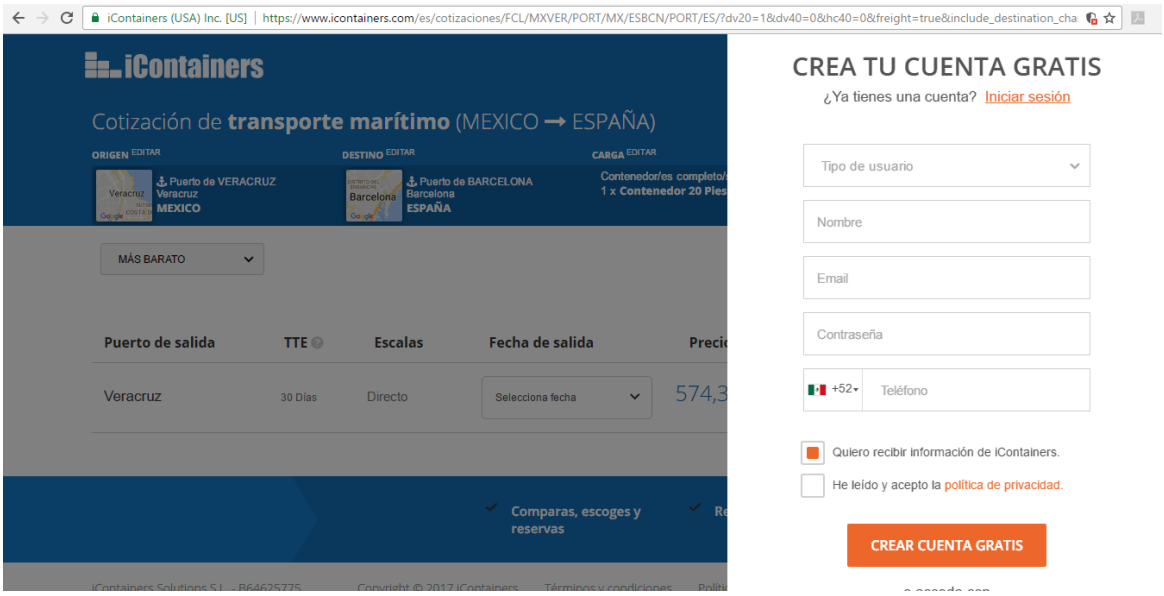


Figura 2.21. *Crear nueva cuenta o ingresar como usuario.*

En la figura 2.22 se muestra la impresión de la reservación en resumen con los datos básicos para la reservación:

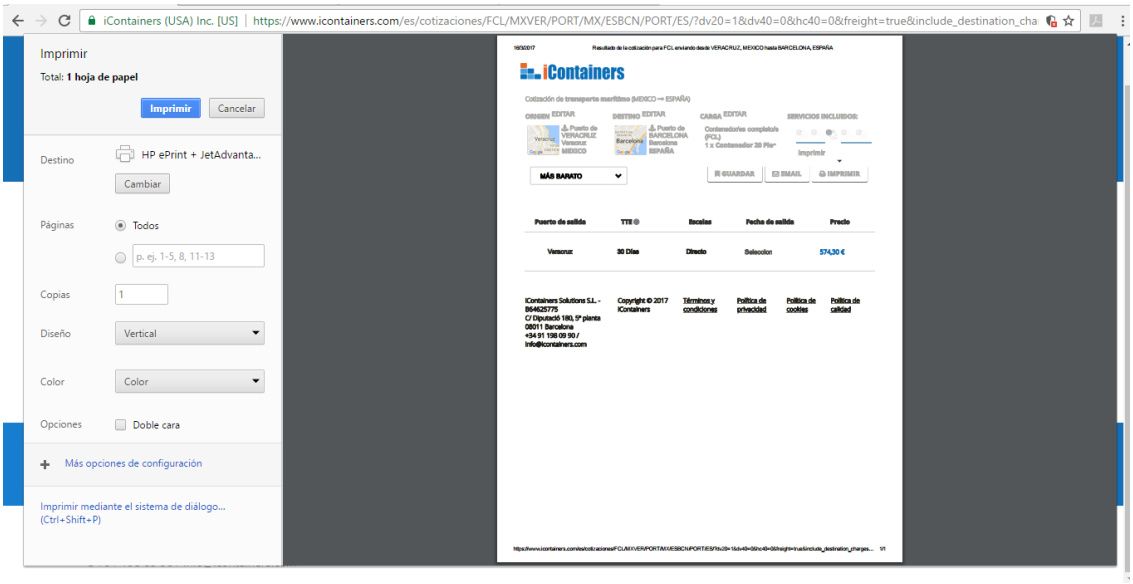


Figura 2.22. *Impresión de la reservación.*

Conclusiones:

La plataforma es amigable con usuario, no se necesita ser cliente para poder realizar la cotización de la exportación del producto, una vez que se hace el procedimiento del booking muestra toda la información detallada de la exportación.

La desventaja es que no cuenta con muchos destinos, no es muy fácil reservar de un puerto de México hacia el extranjero. Se tuvieron que hacer muchas pruebas para conocer los destinos y poder obtener la información deseada.

GURUCARGO

Por otra parte Gurucargo (2018), es una empresa brasileña dedicada a la logística para llevar a cabo la exportación de productos.

En la figura 2.23 se muestra parte de la interfaz de la ventana principal, es una interfaz que cuenta con diferentes opciones para tener comunicación con ellos por si se requiere de ayuda para realizar la operación.

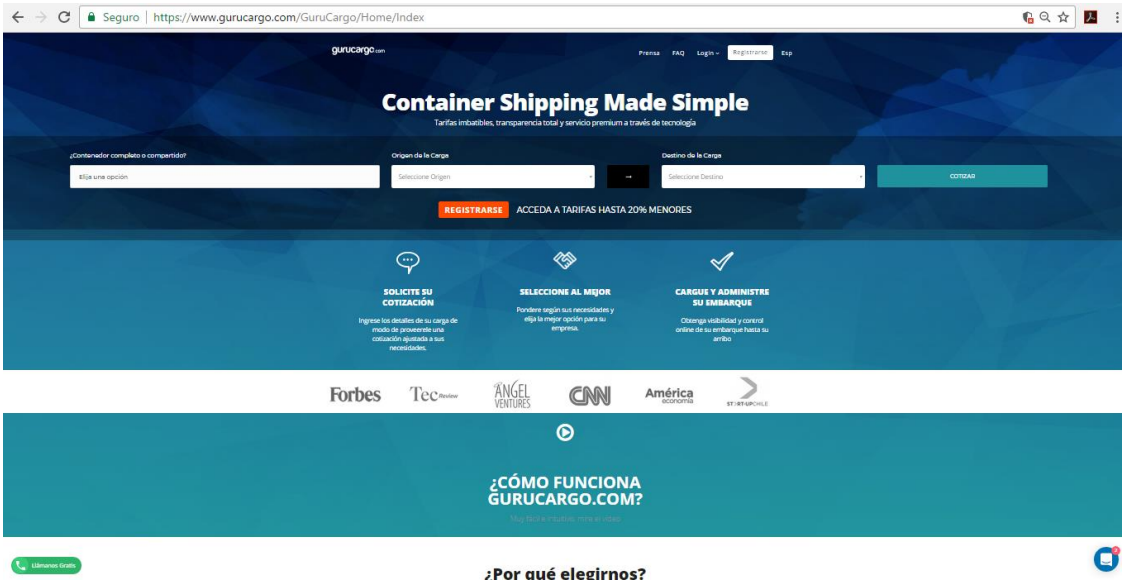


Figura 2.23. *Página principal de Gurucargo, parte I.*

En la parte dos de la ventana principal se muestran los beneficios de elegir a la compañía para hacer la exportación e importación con ellos, figura 2.24.

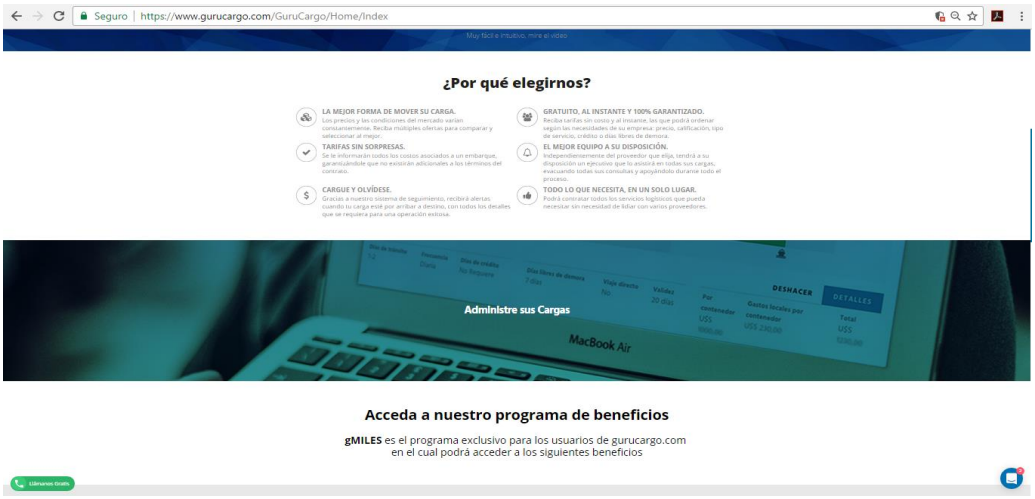


Figura 2.24. *Página principal de Gurucargo, parte II.*

La tercera parte de la interfaz principal muestra una clasificación del cliente y los descuentos que se obtienen a partir de las exportaciones e importaciones que se realizan por mes tal como lo muestra la figura 2.25.

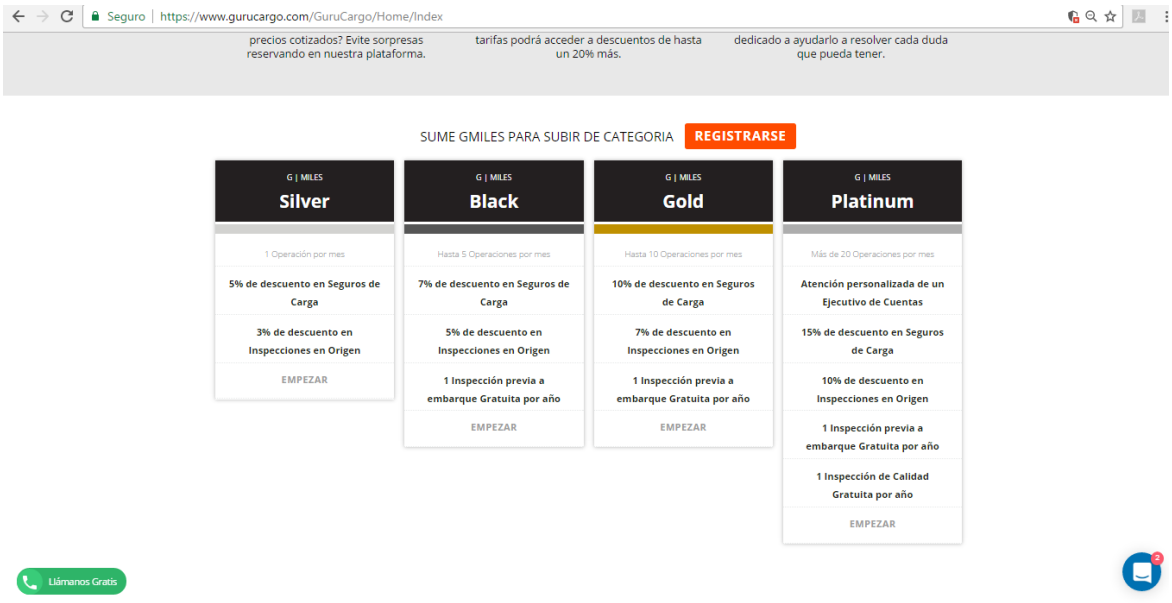


Figura 2.25. *Página principal de Gurucargo, parte III.*

Una vez que se tiene la información principal de la empresa y los beneficios que ofrece por ser cliente, se inicia con el proceso de la cotización del booking, en esta página primero se selecciona el contenedor, tal como lo muestra la figura 2.26.

Figura 2.26. Selección del contenedor.

Una vez seleccionado el contenedor se procede a indicar el origen de la carga como se muestra en la figura 2.27.

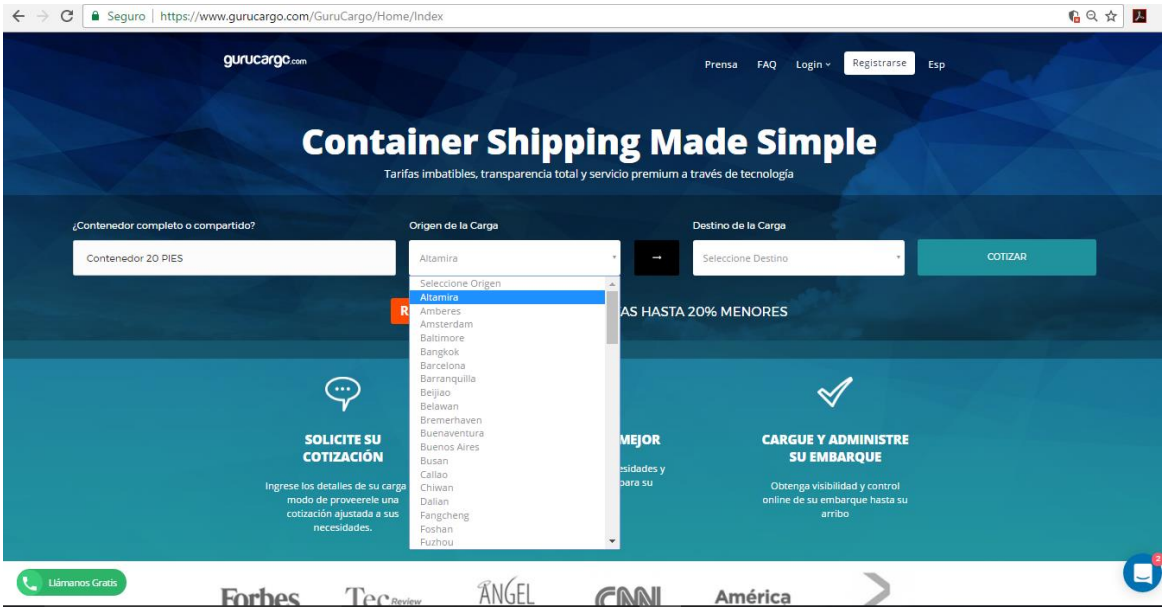


Figura 2.27. *Seleccionar origen de la carga.*

Cuando ya se ha seleccionado el contenedor, el origen de la carga, se indica el destino de la misma. Como se escogió el puerto de Altamira sólo tiene un destino así como se muestra en la figura 2.28.

Figura 2.28. Selección del contenedor.

Una vez completados los pasos anteriores, en la ventana se despliegan los datos de la naviera, tiempo que se tarda en llegar al destino y el costo, se procede a solicitar la cotización en firme tal como lo muestra la figura 2.29.

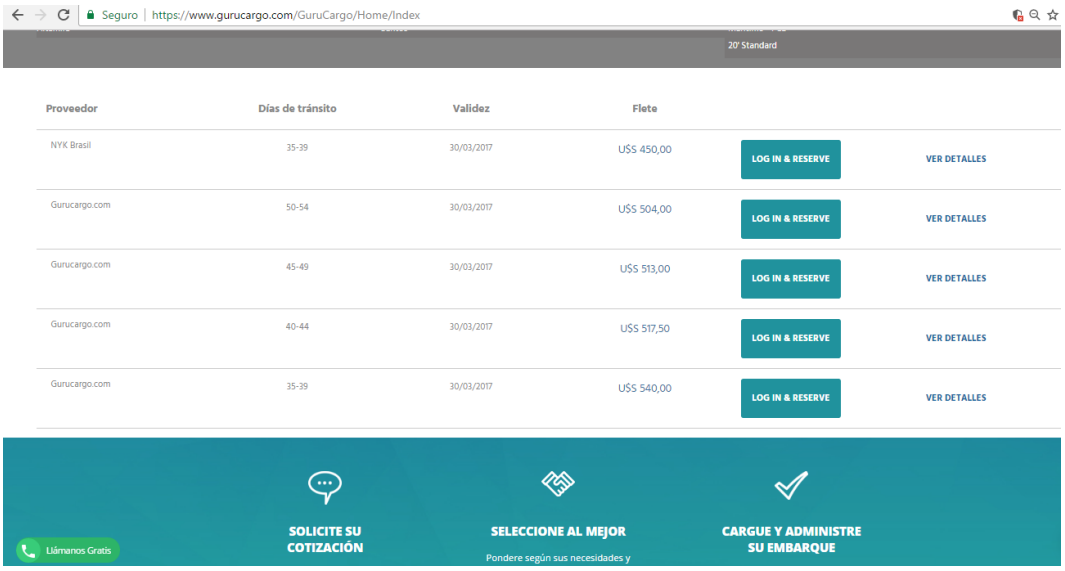


***Figura 2.29**. Cotización del envío.*

Una vez obtenida la cotización del envío, para completar el booking es necesario ingresar como usuario o darse de alta, así como lo muestra la figura 2.30.

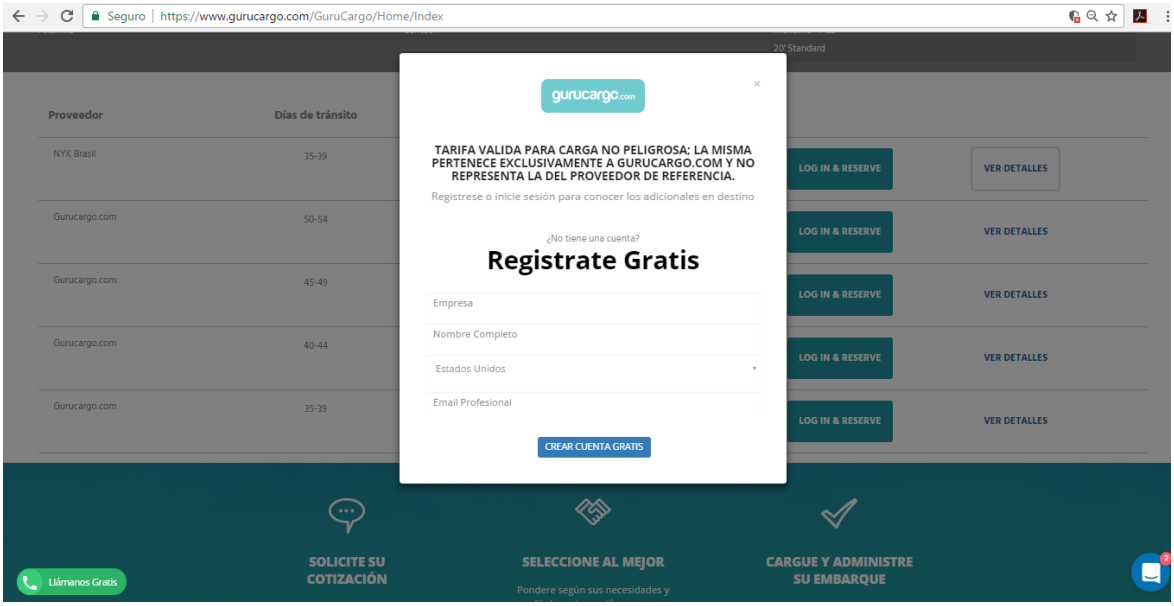


Figura 2.30. *Registro como nuevo usuario.*

Conclusiones:

El sistema en pocos pasos brinda la información necesaria para poder hacer la reservación del contenedor para hacer la exportación.

La desventaja es que es muy limitada con los destinos para la entrega de la mercancía.

La continuación del proceso ya no se llevó a cabo porque es la reservación en firme de la exportación, ya no se podía realizar debido a que no se necesitaba el servicio.

45HC

Otra página Web visitada es 45HC (2018), es la de la empresa que tiene por nombre 45HC, es una empresa de origen asiático por lo tanto su mercado está muy cerrado, no tiene muchas opciones para hacer la exportación, se hizo el intento y no se tuvo éxito con ningún campo ni ciudad, ni país.

A continuación se describe paso a paso lo que se visualizó en la página. En la figura 2.31 se muestra la interfaz principal de la plataforma 45HC, se aprecia que es una interfaz muy sencilla, cuenta con los elementos básicos para hacer la cotización del *booking*.

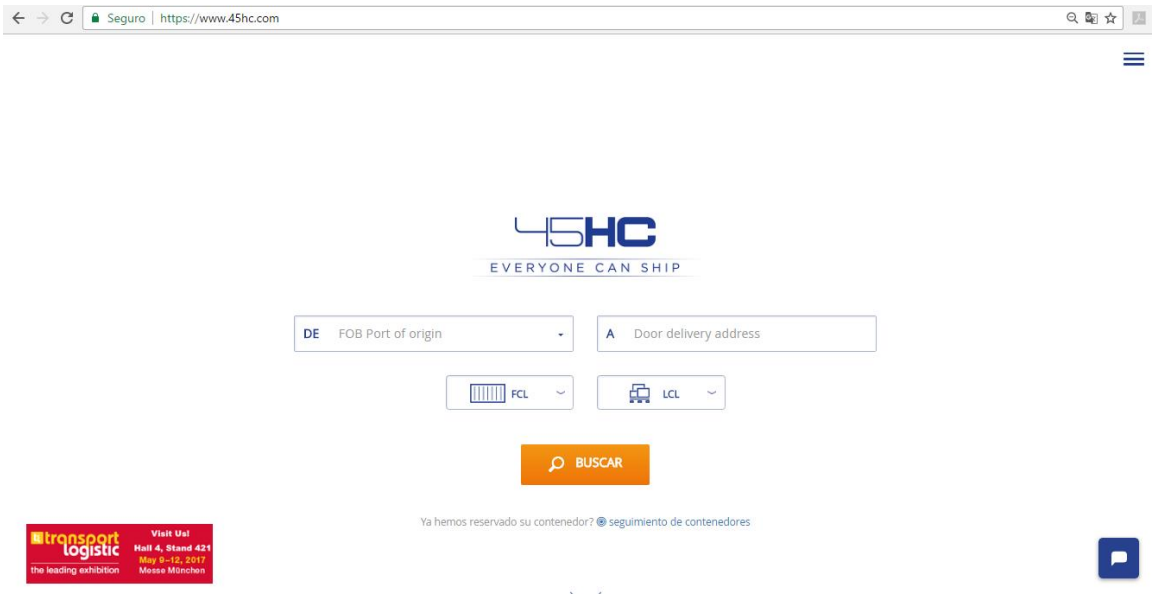


Figura 2.31. *Interfaz principal de 45hc.*

En la figura 2.32 se muestran las ciudades de donde se recogerá la mercancía a exportar, es importante mencionar que al seleccionar la ciudad aparece una ventana para iniciar el chat, donde preguntan:"¿Cómo podemos ayudar? ¡Estamos aquí por tí!"

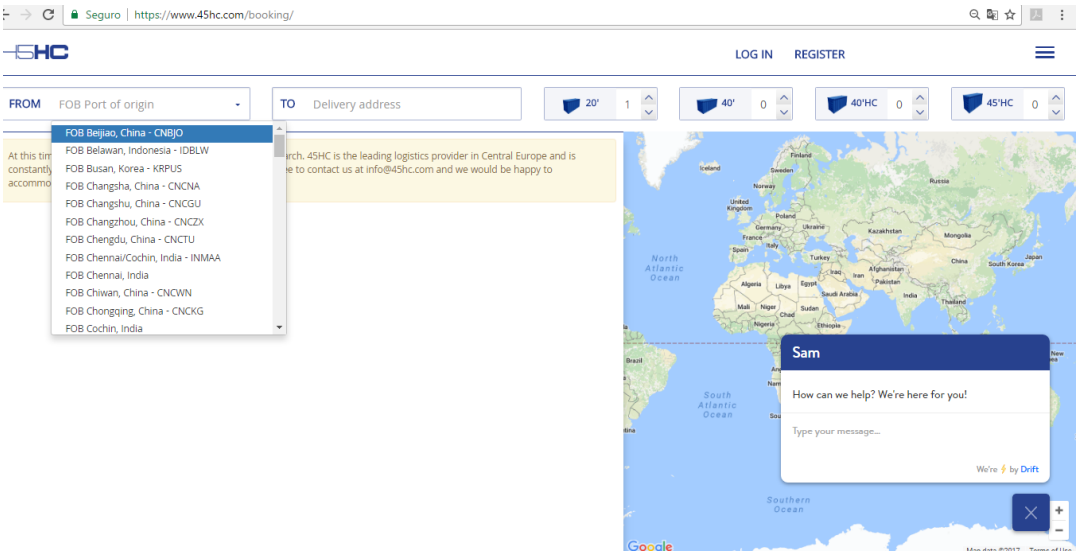


Figura 2.32*. Selección de origen de la carga.*

En la figura 2.33 se observan algunas de las ciudades a las cuales se puede entregar la mercancía que se exportará.

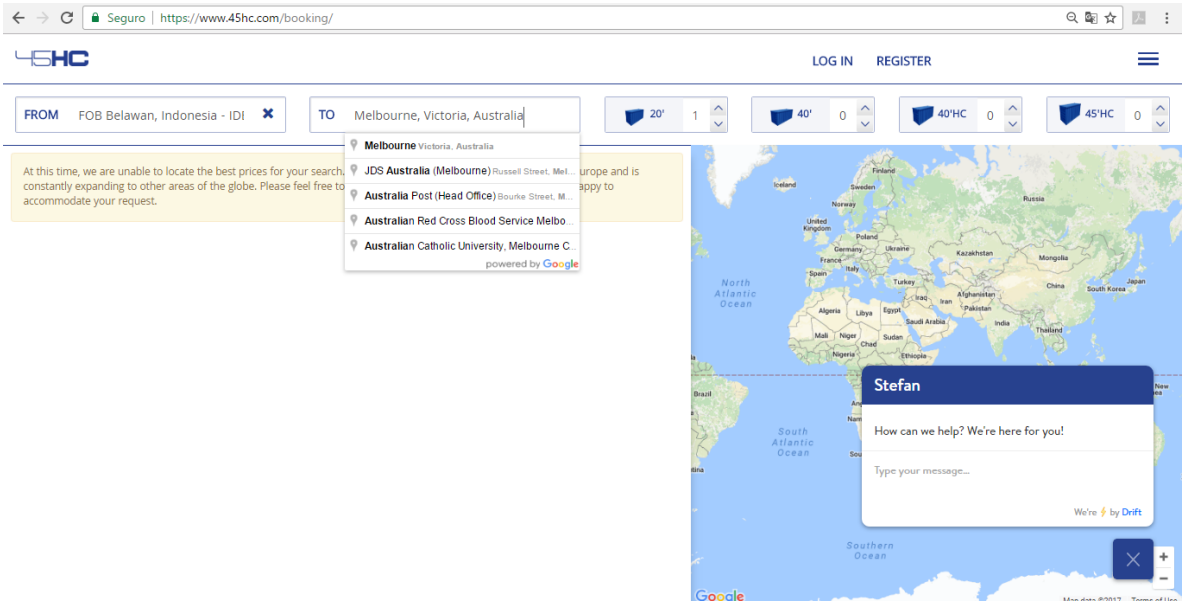


Figura 2.33. *Selección de destino de la carga***.**

En la figura 2.34 se muestra que no hubo éxito en la exportación, se intentó con diferentes ciudades y con todas se obtuvieron el mismo resultado, el mensaje es el siguiente:

“En este momento, no podemos encontrar los mejores precios para tu búsqueda. 45HC es el proveedor líder de logística en Europa Central y se está expandiendo constantemente a otras áreas del mundo. Por favor, no dude en contactar con nosotros en info@45hc.com y estaremos encantados de satisfacer su solicitud”.

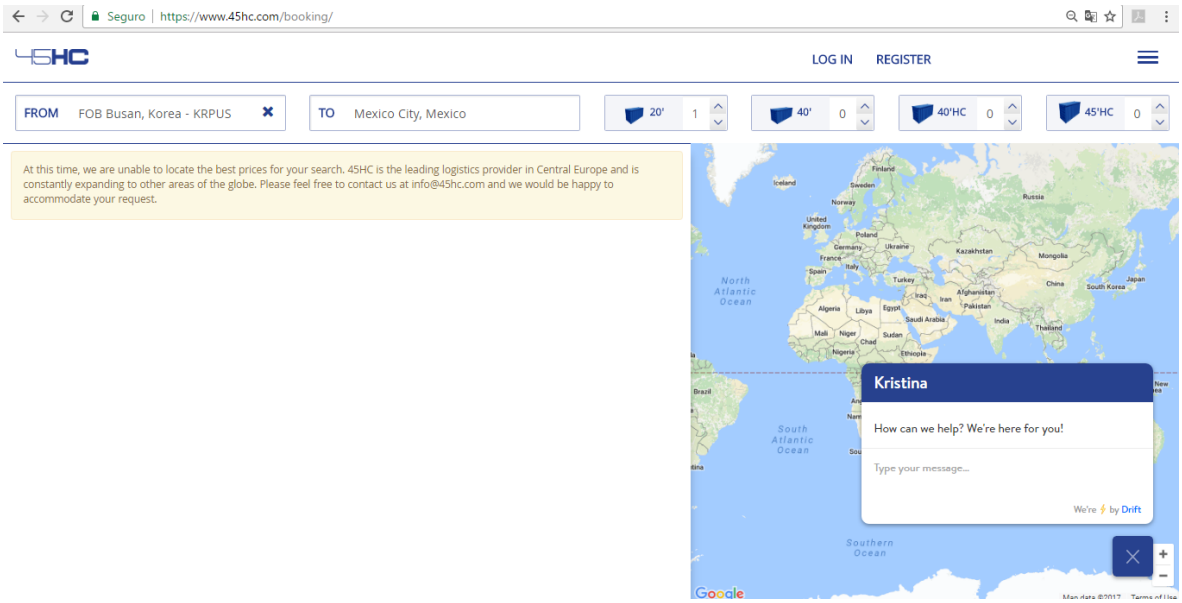


Figura 2.34. Mensaje que no se pudo realizar cotización.

Conclusiones:

La plataforma es muy limitada sólo es para el continente asiático, el lenguaje es inglés o chino mandarín, si no es usuario de ellos no da la opción de poder cotizar la reservación del booking, al seleccionar una opción aparece el nombre de un empleado preguntando que sí puede ayudar.

FLEXPORT

La empresa Flexport (2017), se dedica a la logística de las exportaciones e importaciones, sus principales oficina están en Europa y una en los EE.UU. A continuación en la figura 2.35 se muestra la interfaz principal de la empresa.

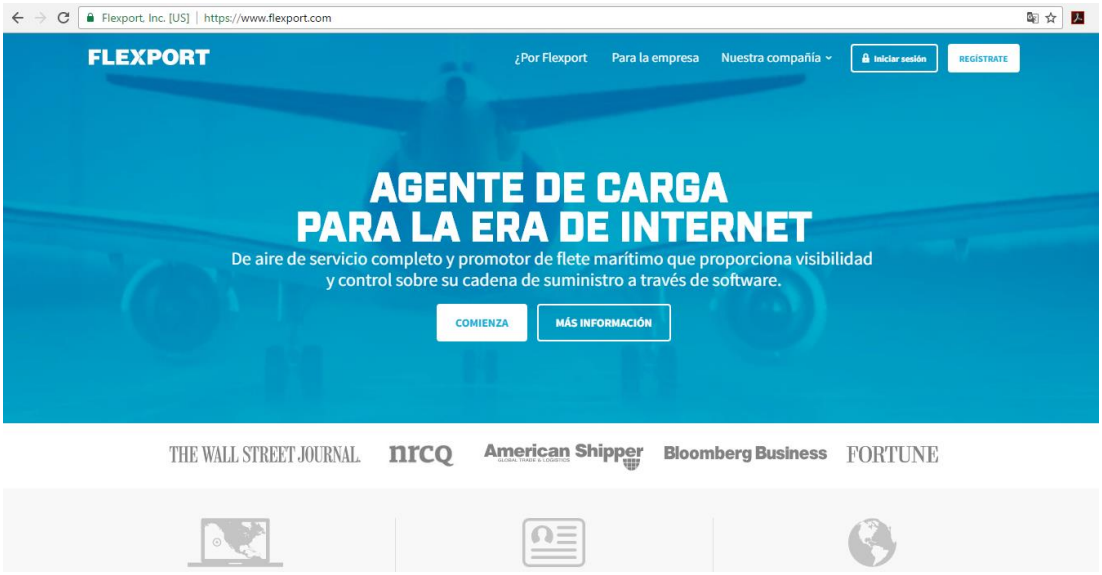


Figura 2.35. *Interfaz principal de Flexport.*

Para comenzar a darse de alta como nuevo usuario es necesario llenar el siguiente formulario tal como se muestra en la figura 2.36.

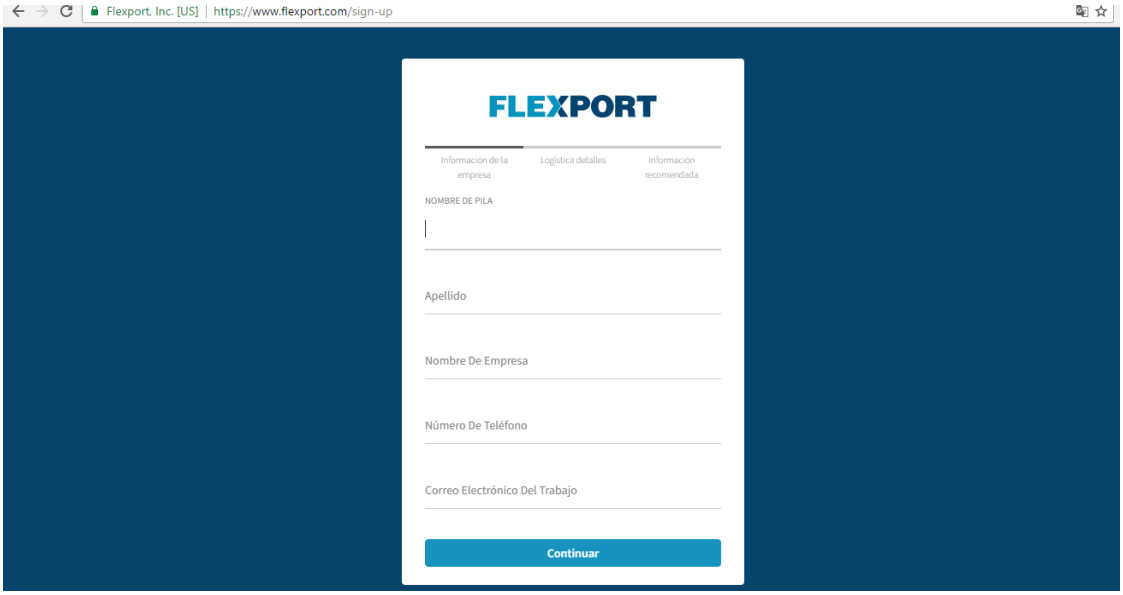


Figura 2.36. *Formulario para darse de alta.*

Cuando el usuario ya está registrado en la base de datos, se da click en la opción login y se solicita el correo y contraseña como se muestra en la figura 2.37.

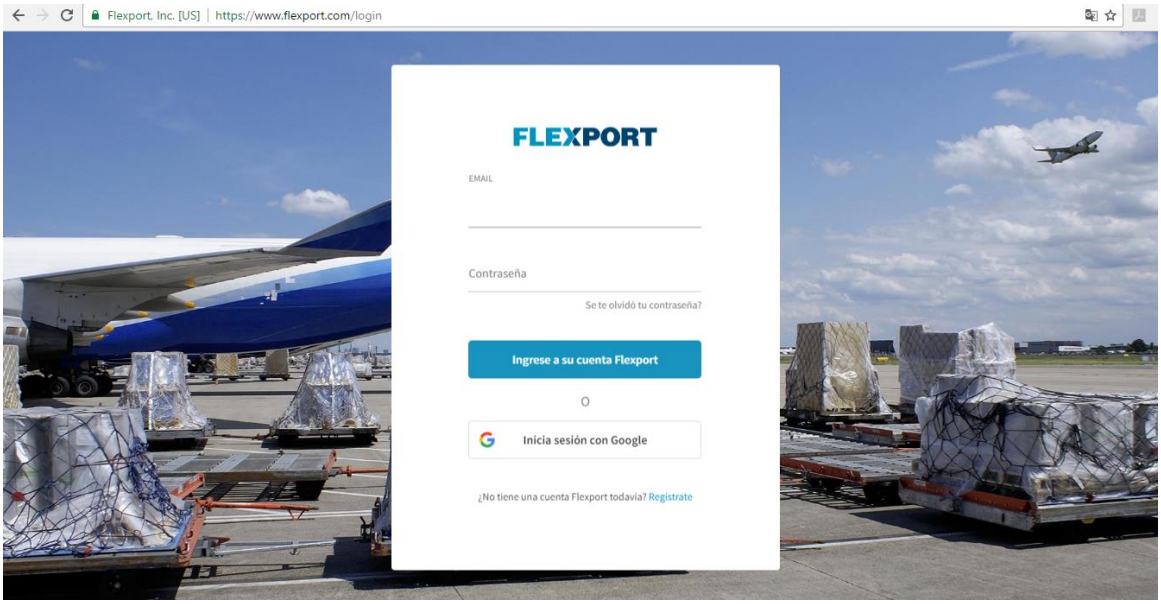


Figura 2.37. *Formulario para ingresar.*

Como no se cuenta con un correo para poder realizar el booking, fue la única información que se pudo obtener, se envió el correo para ver si se podía obtener más información con respecto a la reservación del booking, pero no se obtuvo una respuesta de parte de Flexport.

Conclusiones:

Para poder acceder a los servicios que ofrece, se tiene que ser cliente, de lo contrario sólo muestra el formulario para darse de alta, visualmente es una plataforma atractiva pero desgraciadamente para un nuevo usuario no tiene muchas opciones para conocer la empresa.

Lotebox

Por otra parte Lotebox (2018), la empresa es europea que se dedica al igual que las anteriores a importar y exportar, su interfaz principal muestra el objetivo que tienen hacía los clientes. Tiene la opción para hacer una demostración de los pasos que se deben seguir como se puede observar en la figura 2.38.

Figura 2.38. *Página principal de Lotebox.*

Para poder realizar la reservación y/o demostración gratuita se proporcionó un correo electrónico y sólo apareció el letrero que se pondrán en contacto, figura 2.39:

Figura 2.39. *Respuesta de la plataforma.*

Conclusiones:

La ventaja de le empresa es que su información está protegida, no la muestra si no se es cliente.

La desventaja es que para poder acceder se tiene que enviar correo y hasta que se pongan en contacto con el cliente pueden pasar muchos días e incluso meses o no se tenga respuesta por parte de la empresa.

KONTAINERS

La página de internet Kontainers (2018) es la siguiente plataforma a analizar, la cual es una empresa originaria del Reino Unido que sólo permite exportaciones muy limitadas, su interfaz es muy amigable con el usuario, fácil de interactuar, con animaciones esto lo hace más atractivo visualmente.

En la figura 2.40 se muestra la interfaz principal de la plataforma, tiene la opción de realizar el demo para saber cuáles son los pasos a seguir para poder hacer una exportación.

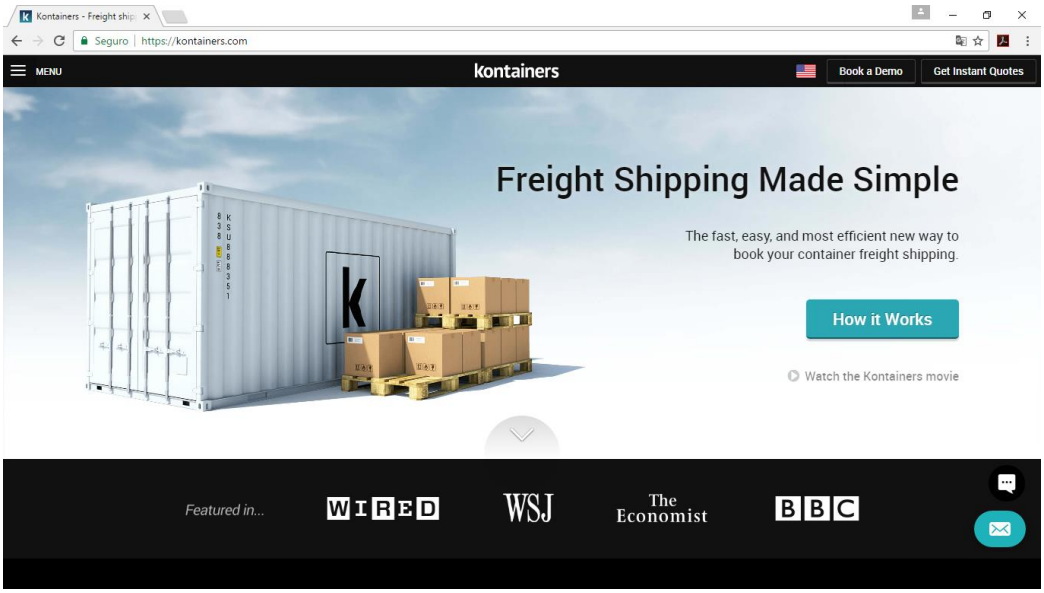


Figura 2.40. *Interfaz principal.*

En la figura 2.41 se muestra la ventana para seleccionar qué es lo que se quiere hacer si exportar o importar, todo se selecciona con un click.

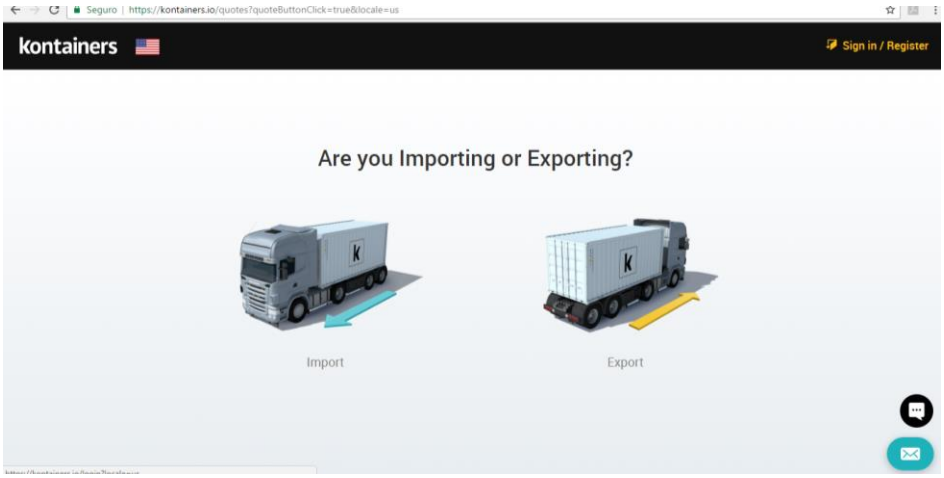


Figura 2.41. *Selección de proceso.*

Una vez seleccionado el proceso que se va a realizar, en este caso fue exportación, en la figura 2.42 se muestra que se está enviando, si necesita en contenedor o pallet, en la parte izquierda se observa que aparece la palabra **reserva**, ahí muestra lo que se está seleccionando.

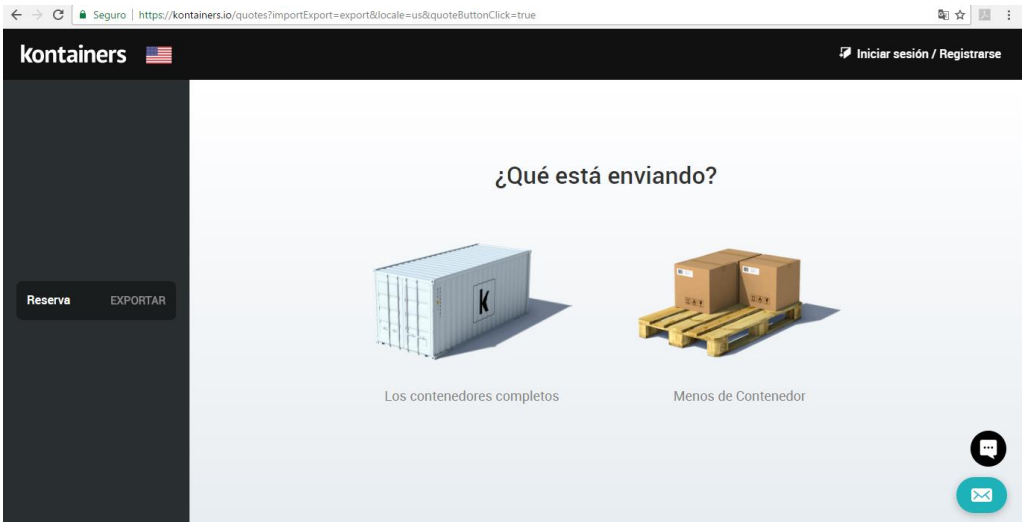


Figura 2.42. *Selección contenedor o pallet.*

Una vez seleccionado el tipo de contenedor en este caso fue un pallet, se muestra el siguiente menú, figura 2.43.

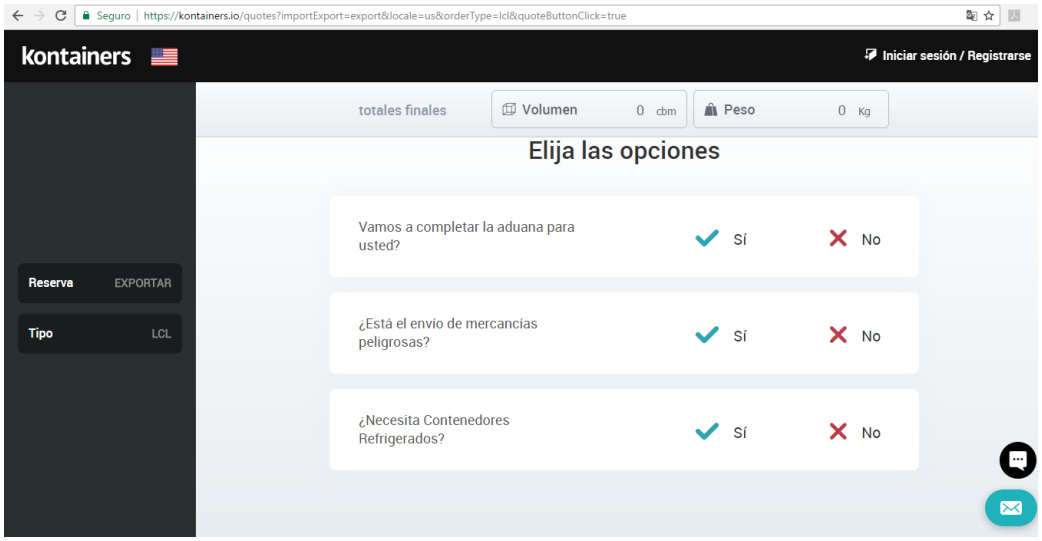


Figura 2.43*Características del contenedor.*

Una vez seleccionado el contenedor, la siguiente interfaz es para indicar el lugar de origen de la carga, desafortunadamente no se despliegan los destinos que manejan y es complicado para el usuario (figura 2.44).

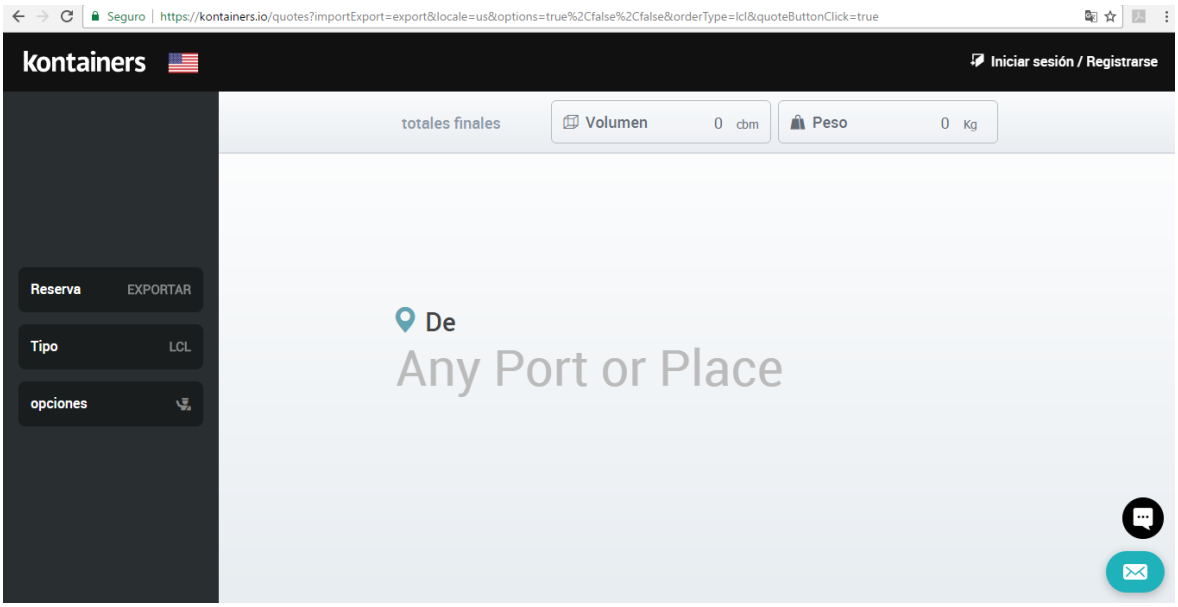


Figura 2.44. *Lugar de origen de carga.*

Cuando se tiene el lugar de carga aparece el calendario para indicarle la fecha que se necesita para que se recoja la mercancía figura (2.45).

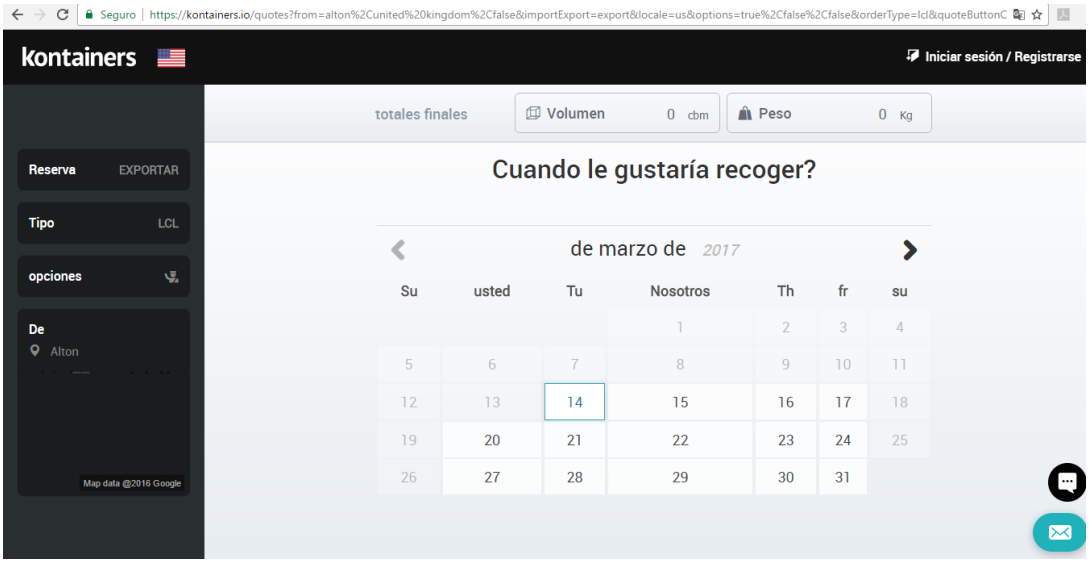


Figura 2.45. *Calendario para recoger la carga.*

Cuando ya se tiene el dato del lugar y fecha, se procede a elegir el destino para la entrega de la mercancía, tampoco se despliega un menú con los destinos que se manejan. Figura 2.46.

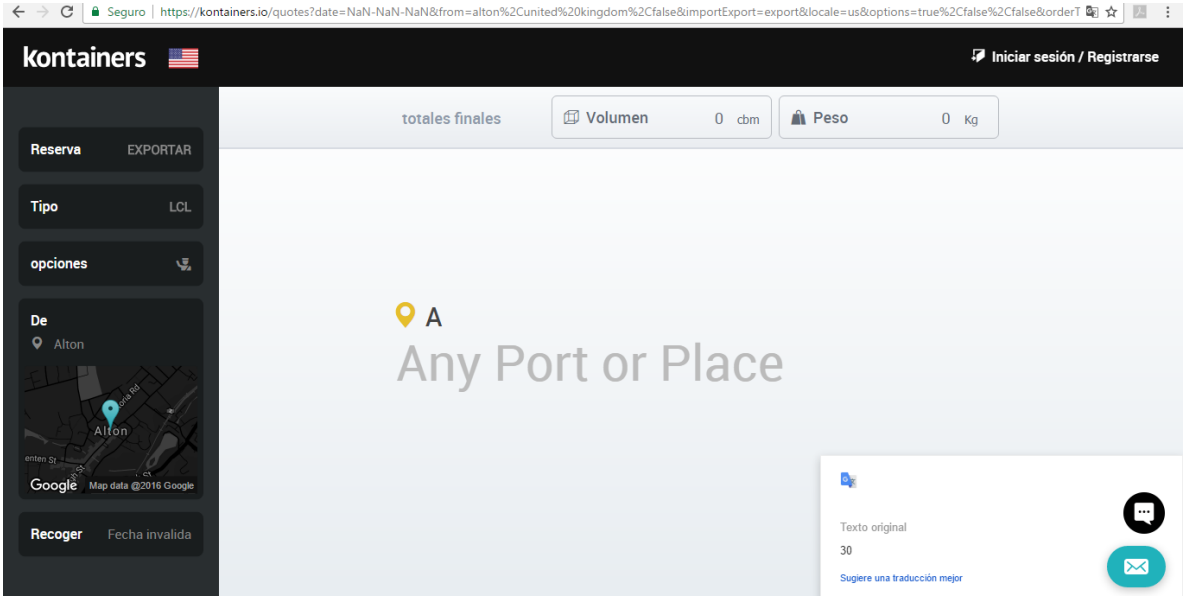


***Figura 2.46**. Selección del destino de la carga.*

En la figura 2.47 después de haber intentado diferentes ciudades para el destino de la carga se bloqueó, ya no dio otra opción para continuar.

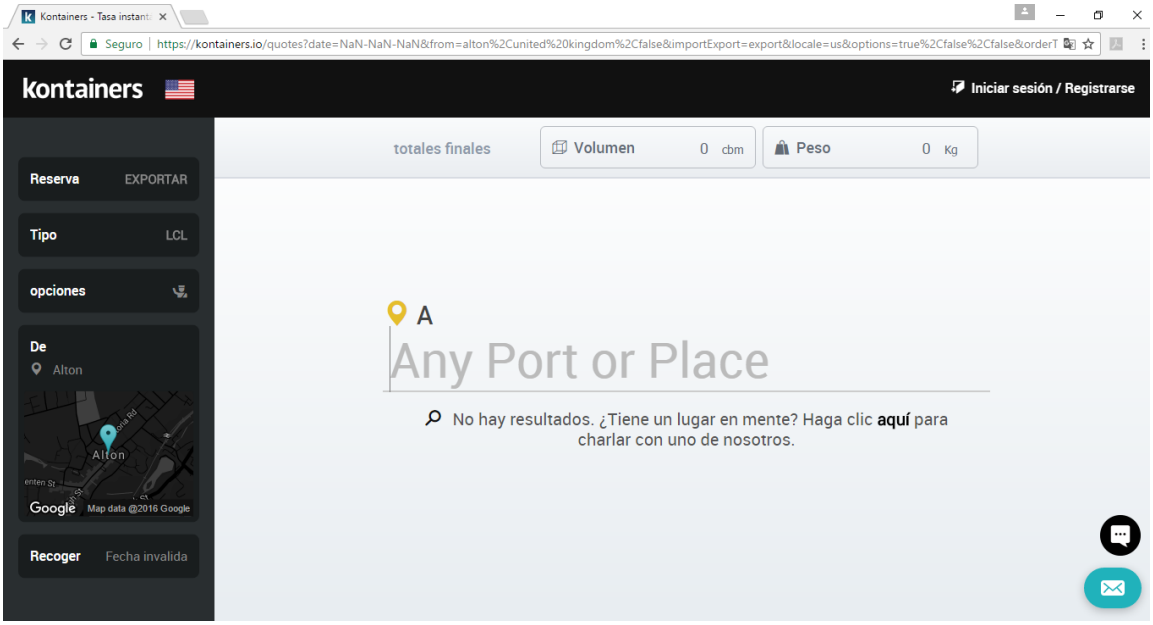


***Figura 2.47.** Ventana de aviso de no encontrado.*

Conclusiones:

El sistema es muy fácil de utilizar, muy amigable hay mucha animaciones para que sea más práctico, el problema es que cuando se va a seleccionar el lugar de origen no se tiene con un listado de los puertos donde pueden llegar las exportaciones.

Comparación de la plataforma de MCP the World Parther contra los programas analizados

El propósito principal de las plataformas es dar el servicio a través de Internet, tener un mayor número de clientes y ventas (exportaciones e importaciones), todo esto, en cualquier parte del mundo.

Las plataformas Icontainers y gurucargo, tienen la opción de poder cotizar el booking sin ser cliente, solicitan la información básica como: lugar de origen, tipo de contenedor, capacidad, lugar destino y fecha; para reservar se ingresa al sistema o en caso de ser la primera vez que se consultan los servicios pues se tiene la opción para darse de alta y continuar con el proceso que conlleva para la exportación.

La plataforma 45hc a simple vista es una interfaz muy limpia no tiene animaciones, el problema que se encontró que al momento de seleccionar el lugar de origen en lista ciudades del medio oriente y para seleccionar el destino no da opciones y en cualquier destino aparece el siguiente texto: “En este momento, no podemos encontrar los mejores precios para tu búsqueda. 45HC es el proveedor líder de logística en Europa Central y se está expandiendo constantemente a otras áreas del mundo. Por favor, no dude en contactar con nosotros en info@45hc.com estarán atentos de atender su solicitud, también muestra un cuadro del chat para solicitar ayuda”.

La plataforma Flexport y lotebox son una plataforma que para poder acceder y conocer sus servicios, se necesita ser cliente de ellos.

El sistema Web que se diseñó solicita ser cliente o darse de alta en la platafrma para poder tener acceso, soporta 50 usuarios únicos conectados al mismo tiempo, el formulario es de auto completar y muestra los productos que más se han enviado el usuario para que sea menos tediosos al momento de completar la información para relizar la exportación, el sistema se

despliega en diferentes menús para seleccionar la naviera, se ofrece una lista con las navieras dadas de alta en la plataforma, mostrando tarifas y tiempo de tránsito.
El administrador puede realizar diferentes consultas como conocer cuáles artículos son los que más se han enviado, búsquedas de booking, con los privilegios con que se cuenta puede agregar nuevas navieras.
Un plus con el que cuenta el sistema es la *landing page* que muestra una descripción breve de los beneficios, servicios y socios principales del proyecto, tiene una sección que permite a los visitantes enviar un mensaje sobre alguna duda, sugerencia, etc.
Si el sistema Web se hubiera podido conectar directamente con la base de datos de la naviera, el proceso del booking sería más rápido porque directamente se estaria reservado el espacio en el medio de transporte donde se va envíar la mercancía, así como todo el proceso que se necesita para realizar una exportación, pero como no se pagó el permiso para accesar, el sistema quedó solamente como un sistema administrativo.

Artículos

Como parte del marco teórico se leyeron artículos referentes a Big Data con analítica y las formas de presentar la información. A continuación se hace mención de algunos de los artículos leídos indicando el título y los autores, describiendo la idea principal del mismo:

En el artículo Topic- and Time-Oriented Visual Text Analysis, **Wenwen** Dou y **Shixia** Liu hablan de los métodos y técnicas que utilizaron para presentar el análisis de texto visual.

El análisis de texto visual se ha convertido en un tema popular en los últimos años, y como el desafiante futuro de las áreas de investigación, es importante saber que se va hacer con la información y de qué manera se puede presentar al usuario de una forma clara, precisa y ordenada. Se han desarrollado muchos algoritmos con minería de datos para resumir y analizar grandes cantidades de datos textuales, estas técnicas han sido adaptadas y aplicadas de acuerdo a la relación de la información donde se contienen coordenadas paralelas y matrices para relevar la relación que tiene el documento con el tema.

Se habla de técnicas para el análisis de texto visual, la primera es el cepillado que permite filtrar a un conjunto de documentos que son altamente relevantes para el tema de interés, los usuarios pueden colapsar, expandir o modificar la jerarquía de temas. La segunda técnica es

la metáfora del rio que representa las tendencias de los temas a lo largo del tiempo, donde cada cinta representa un tema por lo que los usuarios pueden interactuar con las tendencias actuales. El sistema de análisis visual de Leadline también para explorar y analizar eventos que se describen en colecciones de texto. Las palabras clave del tema se vuelven a agrupar para identificar palabras que describen con mayor exactitud los eventos específicos, como resultado, los usuarios pueden analizar y explorar sucesos que pasan en las colecciones de texto, con el fin de poder filtrar los temas que son de interés para el usuario.

El sistema integra un algoritmo de corte de árbol en transmisión y una nueva representación que apoya el análisis interactivo de la transmisión de textos. La visualización integra una sedimentación en el flujo de la corriente de información para ilustrar cómo se agregan nuevos documentos y documentos antiguos, se puede hacer una comparación de que se tenía y que es lo que se ha estado integrando.

En la figura 2.48 se expresan sistemas de análisis de los diferentes tipos de datos de texto visual

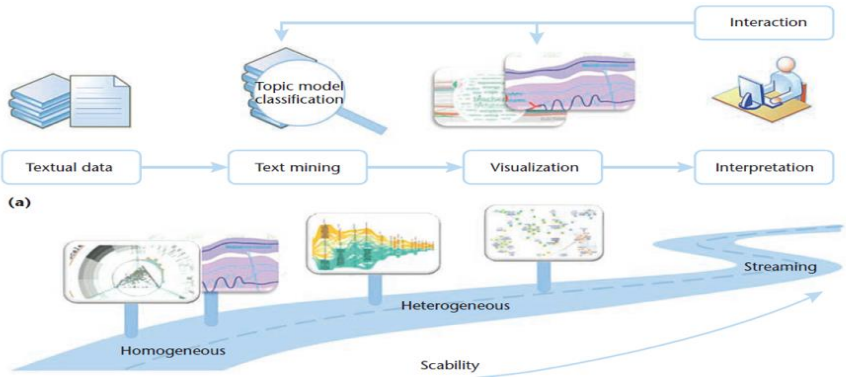


Figura 2.48. *Representación de las técnicas*

En la visualización, la actualización debería ser gradual para que los usuarios no se sienten abrumados por múltiples patrones todos cambiando a la vez.

Dentro de los datos que se tienen que analizar son los datos heterogéneos, en conjunto se encuentra en la identificación de un espacio de características comunes que pueden representar varios tipos de datos ya que provienen de diferentes fuentes y diferentes textos, lo que hace que la visualización de datos se convierta en un desafío al momento de presentárselos al usuario. Con la tendencia de Big Data uno de los desafíos es como poder presentar al usuario toda la información, de una manera que sea fácil de entender y con los

datos que se necesitan tener sin pérdidas de los mismos, con las palabras claves se pueden filtrar una o varias veces la información, al momento de hacer un filtrado el resto de la información parece irrelevante porque en ese momento no es necesario para ese usuario y pero para otros la información es de suma importancia.

En el artículo ***Multimedia Big Data Computing*** los autores Zhu, Cui, Tsinghua y Hua hablan del desafío de cómo hacer para que el usuario visualice los datos através de Big Data con Multimedia, el cual brinda un ábaco de oportunidades para poder visualizar la información en diferentes medios, hacer recomendaciones, anuncios, etc.

En comparación con los enfoques de computación de datos grandes, computación de datos multimedia grande enfrenta una compresión adicional, almacenamiento, transmisión y análisis en términos de organización heterogénea y no estructurada de datos, tratar con la cognición, la comprensión, complejidad, servicios en tiempo real y calidad del servicio.

Xindong Wu y sus colegas presentaron un teorema Heterogéneo, Autónomo, Complejo, modelo de procesamiento de datos a través de la minería de datos. Los grandes datos de la multimedia no estructurados, heterogéneos y multimodal, hace que la multimedialidad de la representación de datos sea difícil

Las aplicaciones multimedia de gran tamaño y los servicios suelen ser en tiempo real, requieren de *streamed/online*, procesamiento paralelo/distribuido para el análisis, la minería y el aprendizaje.

La reducción de datos del volumen de grandes datos multimedia debe ser reducido para un almacenamiento y comunicación eficaz. Esto se refiere al muestreo del conjunto de datos (masivo) para que pueda ser calculado con recursos informáticos limitados. La compresión de datos multimedia se refiere a la reducción del tamaño de datos sin procesar para el almacenamiento o la comunicación.

Debido a que los datos multimedia vienen de múltiples fuentes, tienden a tener representaciones diversas para cada fuente, o a veces se necesita una representación común para el análisis multimodal. Estas interpretaciones se describen habitualmente tanto a nivel estructural como metadatos descriptivos.

En el análisis de datos multimedia se ha centrado en cómo fusionar la información de las diferentes modalidades de los medios para formar una decisión. Sin embargo, surgen problemas cuando las entradas de datos carecen de modalidad de datos.

La fusión del motor de búsqueda y de las redes sociales son claramente una tendencia común en la industria. Por ejemplo, Google adquirió YouTube y lanzó Google Plus; Yahoo! adquirió Flickr y Facebook hicieron esfuerzos para desarrollar búsqueda de servicios con un alcance de Facebook-externo.

Se podría aprovechar mucho la integración de plataformas multimedia con sistemas de búsqueda multimedia, como se muestra en la Figura 2.49. Cómo descubrir y representar la intención de búsqueda del usuario medios de comunicación y unir fácilmente estas intenciones sistemas de búsqueda multimedia es un tema de investigación que requiere una atención seria.

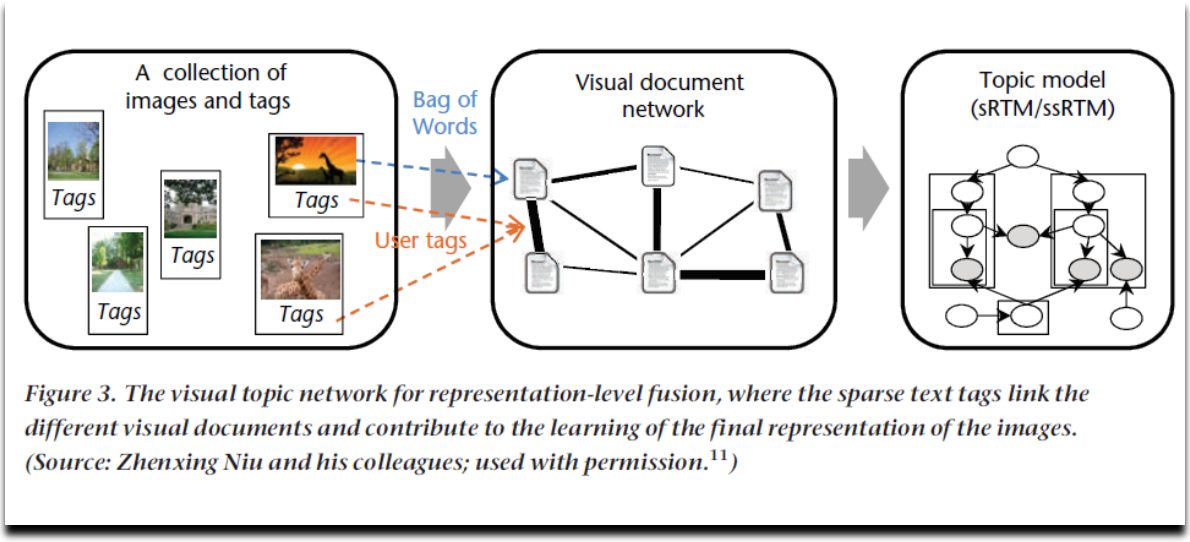


Figure 3. The visual topic network for representation-level fusion, where the sparse text tags link the different visual documents and contribute to the learning of the final representation of the images. (Source: Zhenxing Niu and his colleagues; used with permission.[11])

Figura 2.49 *Tipos de datos multimedia.*

Toda la información antes mencionada sirve de referencia para poder trabajar con los métodos, procesos para obtener la información que se necesita, una vez que se tiene la base de datos de gran volumen se tiene que hacer un análisis de toda la información, para utilizar la mayor parte de la información y que no tenga información basura en lo menor posible.

Otro de los retos que se tiene con la plataforma es cómo presentar al usuario toda la información que se tiene almacenada, de una manera que sea muy fácil de digerir y visual,

de tal manera que no lo abrumen tantos datos y que la información presentada sea de interés y lo que el usuario está solicitando. Por eso es importante que los datos estén disponibles, abiertos y accesibles, para facilitar a la toma de decisiones.

2.2 Marco teórico

La Minería de Datos (Data Mining)

Es un proceso que utiliza técnicas estadísticas, matemáticas, inteligencia artificial y de aprendizaje de máquinas para extraer e identificar información útil que convierte en conocimiento a partir de grandes bases de datos, data warehouses o data mart (Aguilar, 2013).

El Análisis de Datos (*Data Analysis, o DA*) es la ciencia que examina datos en bruto con el propósito de sacar conclusiones sobre la información. El análisis de datos es usado en varias industrias para permitir que las compañías y las organizaciones tomen mejores decisiones empresariales y también es usado en las ciencias para verificar o reprobar modelos o teorías existentes.

El análisis de datos se distingue de la extracción de datos por su alcance, su propósito y su enfoque sobre el análisis. Los extractores de datos clasifican inmensos conjuntos de datos usando software sofisticado para identificar patrones no descubiertos y establecer relaciones escondidas. El análisis de datos se centra en la inferencia, el proceso de derivar una conclusión basándose solamente en lo que conoce el investigador (Aguilar, 2013).

Analítica de datos (data analytics):

Implica los procesos y las actividades diseñadas para obtener y evaluar datos para extraer información útil. Los resultados de la AD (DA) se pueden utilizar para: áreas clave de riesgos, fraudes, errores o mal uso; mejorar los procesos de negocios; verificar la efectividad de los procesos e influir en las decisiones del negocio.

La Analítica de Datos intenta responder preguntas del tipo:

- ¿Qué sucederá con el volumen de ventas si continúan estas mismas condiciones económicas?
- ¿Bajo las actuales condiciones del mercado cuál debería ser el precio óptimo del producto?

Tipos de analítica:

Descriptiva:

El objetivo es describir lo que está sucediendo en una determinada situación o escenario en un determinado periodo de tiempo. Por ejemplo: Las ventas y el valor tomado por los indicadores el trimestre anterior.

Predictiva:

El objetivo es pronosticar lo que sucederá en el futuro a partir del análisis de datos históricos.

Prescriptiva:

Va más allá de la analítica descriptiva y predictiva. Esta busca dar recomendaciones o cursos de acción mostrando la probabilidad de ocurrencia de cada decisión y sus posibles consecuencias (Aguilar, 2013).

La analítica de Big Data

Es la utilización de técnicas analíticas avanzadas en conjuntos de Big Data. Por consiguiente, analítica de Big Data se compone de dos teorías: analítica y Big Data. Las organizaciones necesitan recurrir a la analítica de Big Data para tomar decisiones de negocio de lo más acertada posible. Las herramientas de analítica deben completar: reporting, query y visualización, analítica predictiva, analítica web, analítica social, y social listening, analítica especializada para Big Data procedentes de fuentes M2M o internet de las cosas entre otras cosas (Aguilar, 2013).

¿Qué es Machine Learning?

Es una disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente. Aprender en este contexto quiere decir identificar patrones complejos en millones de datos**. La máquina que realmente aprende es un algoritmo** que revisa los datos y es capaz de predecir comportamientos futuros automáticamente, también en este contexto, implica que estos sistemas se mejoran de forma autónoma con el tiempo, sin intervención humana (González, 2018).

Machine Learning Supervisado y Machine Learning no Supervisado

El Machine Learning es un campo muy amplio. Su rápida expansión está haciendo, además, que éste se vea continuamente dividido en diferentes especialidades entre las que cabe destacar:

- **Machine Learning Supervisado.** Es el más utilizado y requiere de intervención humana para la creación de etiquetas en el histórico de datos de manera que la máquina pueda predecir un resultado probable a partir de las mismas. Este método se utiliza, por ejemplo, para la predicción de posibles reclamaciones en sistemas de atención al cliente.

- **Machine Learning no Supervisado.** El aprendizaje no supervisado es menos común y utiliza datos históricos que no han sido etiquetados. El objetivo es encontrar patrones a partir del propio análisis de datos. Un uso muy frecuente es el de segmentación de clientes con atributos similares para campañas de marketing (González, 2018).

PHP

Es un lenguaje de programación que permite entre otras cosas, la generación dinámica de contenidos en un servidor Web. Su nombre "oficial" es PHP: Hypertext Preprocessor.

Es una herramienta de desarrollo para los programadores Web, ya que proporcionan de elementos que permiten generar de manera rápida y sencilla sitios Web dinámicos.

Entre sus principales características, se pueden destacar su potencia, alto rendimiento y su facilidad de aprendizaje (Aprender a programar , 2018).

LARAVEL

Es un framework de PHP de código abierto para el desarrollo de aplicaciones y los servicios Web del PHP 5 al 7.

Laravel tiene como objetivo ser un framework que permita el uso de una sintaxis elegante y expresiva para crear código de forma sencilla y permitiendo multitud de funcionalidades. Intenta aprovechar lo mejor de otros frameworks y las características de las últimas versiones de PHP (Aprender a programar , 2018).

Lenguaje de programación R

R es un lenguaje de programación y entorno de software de código abierto para computación y gráficos estadísticos. Proporciona múltiples técnicas para simulación, modelado lineal y no lineal, análisis de series temporales, pruebas estadísticas clásicas, clasificación, agrupación en clústeres, etc. (Data Tons by easy admin, 2018).

Características y ventajas de la programación en R:

- Es un lenguaje interpretado, funciona mediante comandos.
- R proporciona una amplia gama de herramientas estadísticas que incluyen análisis de datos y generación de gráficos. Este lenguaje tiene capacidad de generar gráficos de alta calidad. Estas características lo convierten en una potente herramienta de cálculo.
- Gracias a este lenguaje de programación los Científicos de datos pueden manejar grandes volúmenes.
- Puede integrarse con distintas bases de datos. Una de las ventajas más importantes de R es que funciona con diferentes tipos de hardware y software (Windows, Unix, Linux, etc.).
- El lenguaje R ofrece la posibilidad de cargar bibliotecas y paquetes con diversas funcionalidades lo que permite a los usuarios extender su configuración básica.

Existen varios paquetes en R para construir árboles de decisión. De entre todos ellos, se utiliza la librería party.

Árboles de decisión

Estos diagramas de flujo en forma de árbol usan ramificaciones para ilustrar cualquier posible resultado de una decisión. Muchos de estos diagramas de árbol usan ramificación binaria (dos opciones) basados en valores actuales o atributos de los datos. Para grandes volúmenes de datos, se pueden crear muchísimos de estos árboles de decisión múltiple, los cuales juntos forman una decisión consensuada sobre los resultados. Los arboles de decisión se pueden usar para problemas de clasificación y también de regresión (UTM.MX, 2018).

Los árboles de decisión son útiles para entender la estructura de un conjunto de datos. Sirven para resolver problemas tanto de clasificación (predecir una variable discreta, típicamente binaria) como de regresión (predecir una variable continua). Se trata de modelos excesivamente simples pero, y ahí reside fundamentalmente su interés, fácilmente interpretables.

Particularidades de los Arboles de Decisión:

- Las características con valores contínuos se pueden partir en dos o más rangos, mediante un umbral (p.e. longitud < 3 y longitud ≥ 3).
- Existen métodos para tratar datos faltantes.
- Los árboles de clasificación tienen etiquetas de clase discretas en las hojas, mientras que los árboles de regresión tienen valores contínuos.
- Los algoritmos para construir árboles son eficientes para el procesamiento de grandes cantidades de datos.
- Existen métodos para tratar datos de entrenamiento ruidosos, con errores tanto en en las características como en la clase.
- Los árboles pueden representar cualquier función de clasificación.

Algoritmo C4.5

Es un algoritmo usado para generar un árbol de decisión y es una extensión del algoritmo ID3 (es un algoritmo de aprendizaje que pretende modelar los datos mediante un árbol, llamado árbol de decisión) desarrollado también por Quinlan previamente. Los árboles de decisión generados con C4.5 se pueden usar para clasificación, por ello es conocido como un clasificador estadístico (López Takeyas, 2005).

El algoritmo C4.5 genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente. El árbol se construye mediante la estrategia de profundidad-primero (depth-first). El algoritmo considera todas las pruebas posibles que pueden dividir el conjunto de datos y selecciona la prueba que resulta en la mayor ganancia de información. Para cada atributo discreto, se considera una prueba con **n** resultados, siendo **n** el número de valores posibles que puede tomar el atributo. Para cada atributo continuo, se realiza una prueba binaria sobre cada uno de los valores que toma el atributo en los datos. En cada nodo, el sistema debe decidir cuál prueba escoge para dividir los datos.

Características del algoritmo C4.5

- Permite trabajar con valores continuos para los atributos, separando los posibles resultados en 2 ramas Ai<=N y Ai>N.
- Los árboles son menos frondosos, ya que cada hoja cubre una distribución de clases no una clase en particular.
- Utiliza el método "divide y vencerás" para generar el árbol de decisión inicial a partir de un conjunto de datos de entrenamiento.
- Se basa en la utilización del criterio de proporción de ganancia (gain ratio), definido como I(Xi,C)/H(Xi). De esta manera se consigue evitar que las variables con mayor número de posibles valores salgan beneficiadas en la selección.
- Es Recursivo.

Estructuras utilizadas en el algoritmo C4.5

- El C4.5 forma parte de la familia de los TDIDT (*Top Down Induction Trees*), junto con antecesor el ID3.
- El C4.5 se basa en el ID3, por lo tanto, la estructura principal de ambos métodos es la misma.
- El C4.5 construye un árbol de decisión mediante el algoritmo "divide y vencerás" y evalúa la información en cada caso utilizando los criterios de Entropía, Ganancia o proporción de ganancia, según sea el caso.

Por otra parte, los árboles de decisión pueden entenderse como una representación de los procesos involucrados en las tareas de clasificación. Están formados por:

- Nodos: nombres o identificadores de los atributos.
- Ramas: posibles valores del atributo asociado al nodo.

- Hojas: conjuntos ya clasificados de ejemplos y etiquetados con el nombre de una clase.

Heurística

Utiliza una técnica conocida como Gain Ratio (proporción de ganancia). Es una medida basada en información que considera diferentes números (y diferentes probabilidades) de los resultados de las pruebas.

Algoritmo J48

Es una implementación open source en lenguaje de programación Java del algoritmo C4.5 en la herramienta weka de minería de datos (itnuevolaredo, 2005).

Weka

Acrónimo de Waikato Environment for Knowledge Analysis, es un entorno para experimentación de análisis de datos que permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático, sobre cualquier conjunto de datos del usuario. Contiene una colección de algoritmos para realizar análisis de datos y modelado predictivo, también tiene herramientas para la visualización de estos datos, además provee una interfaz gráfica que unifica las herramientas para que estén a una mejor disposición (Fallas., 2013).

Kaggle

Es una plataforma online para el modelado predictivo y competencias de análisis en el que estadísticos y mineros de datos compiten para producir los mejores modelos para predecir y describir los conjuntos de datos cargados por empresas y usuarios, para realizar competiciones de Data Mining, la cual proporciona un repositorio para que las compañías publiquen sus datos y otras puedan hacer pruebas con las bases de datos libres que ahí proporcionan (Gracia, 2018).

Regresión lineal y polinómica

La regresión se ocupa de modelar la relación entre variables numéricas que están usando una medida de error en las predicciones hechas por el modelo. La suposición básica es que la

variable de salida (un valor numérico) pueda ser expresado como una combinación (suma ponderada) de un conjunto de variables de salida numéricas (González, 2018).

Redes neuronales

Este concepto está inspirado en la manera en que funciona el sistema nervioso, así como el cerebro para procesar información. Un gran número de elementos de procesamiento altamente interconectados trabajan al unísono para resolver problemas específicos, usualmente de clasificación o de coincidencia de patrones. Cada neurona "vota" sobre el resultado de la decisión, lo cual podría instar a otras neuronas a votar, entonces los votos se contabilizan creando una clasificación de los resultados dependiendo del apoyo que cada uno haya recibido (González, 2018).

Red Bayesiana

Estas estructuras gráficas, también conocidas como red de creencias, son usadas para representar el conocimiento sobre un dominio incierto. La gráfica es un mapa probabilístico de causas y efectos donde cada nodo representa una variable aleatoria, mientras los bordes entre los nodos representan dependencias probabilísticas. Por ejemplo "cielo rojo en la noche" podría conducir a un 75% de probabilidades de "buen clima". Estas dependencias condicionales son estimadas frecuentemente usando métodos estadísticos y computacionales (González, 2018).

Cadenas de Markov

Son sistemas matemáticos que van de un "estado" (situación o grupo de valores) a otro: se supone que los futuros estados depende sólo del estado presente y no de la secuencia de eventos que lo preceden. Por ejemplo, si se hace un modelo de la cadena de Markov sobre el comportamiento de un bebé, podrías incluir "jugar", "comer", "llorar" y "dormir" como estados, los cuales junto con otros comportamientos podrían formar un "espacio de estados" que sería una lista de todos los estados posibles. Adicionalmente, una cadena de Markov te dice cuál es la probabilidad de saltar o "trascender" de un estado a otro, por ejemplo la probabilidad de que un bebé que está jugando se quede dormido en los próximos cinco minutos sin llorar antes (González, 2018).

Scrum

Esta metodología es un proceso que incentiva o ayuda a que el trabajo en equipo sea regular para tener el mejor resultado posible, en Scrum se realizan las entregas de manera parcial y regular del producto final, priorizando al cliente como beneficioso en la que está indicado especialmente cuando los proyectos son complejos y se necesitan los resultados pronto, donde los requisitos son cambiantes o muy poco definidos. (Pressman, 2010)

Ventajas:

- Reuniones recurrentes.
- Se presentan avances periódicamente.
- Permite a un equipo de desarrollo trabajar colaborativamente.

Desventajas:

- Posibles retrasos por parte del cliente ya que los requerimientos pueden cambiarse periódicamente.
- No están establecidos todos los puntos de requerimientos.

Características:

- Entregas parciales y regulares del producto final.
- Define un conjunto de prácticas y roles.
- Las fechas de entrega se programan con mucha antelación.

Incremental

Este modelo combina elementos del modelo de cascada con la filosofía interactiva de construcción de prototipos. Este modelo aplica secuencias lineales de forma escalonada mientras se progresa el tiempo en el calendario. Cada secuencia lineal produce un incremento en el *software*. El modelo se centra en la entrega de un producto final, pero proporcionan al usuario la funcionalidad que precisa y también una plataforma para la evaluación. (Ortiz, n.d.)

Ventajas:

- Se genera *software* operativo de forma rápida y en etapas tempranas del ciclo de vida del *software*.
- Más flexible, por lo cual reduce el costo en el cambio de alcance y requisitos.
- Es más fácil de probar y depurar.
- Es más sencillo gestionar riesgos.

Desventajas:

- Cada fase de iteración es rígida y no se superponen con otras.
- Pueden surgir problemas de arquitectura del sistema ya que no todos los requisitos no han sido reunidos.
- Dificíl evaluar costo total.
- Requiere gestores experimentados.

Características:

- Incrementos pequeños.
- Permite fácil administración de las tareas en cada iteración.
- La inversión se materializa a corto plazo.
- Se adapta a las necesidades que surjan.

CRISPDM. CRISP-DM (Cross Industry Standard Process for Data Mining)

CRISP–DM es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de Data Mining, como se puede constatar en la gráfica presentada en la figura 2.50. Esta gráfica, publicada el año 2007 por kdnuggets.com, representa el resultado obtenido en sucesivas encuestas efectuadas durante los últimos años, respecto del grado de utilización de las principales guías de desarrollo de proyectos de Data Mining. En ella se puede observar, que es la guía de referencia más ampliamente utilizada.

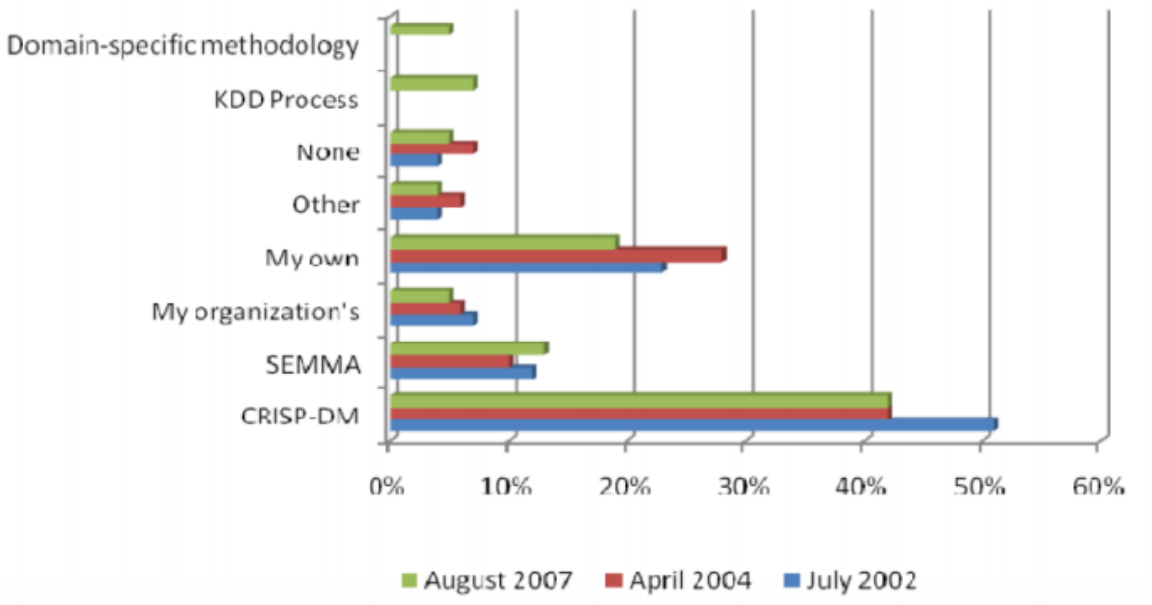


Figura 2.50. *Metodologías utilizadas en Data Mining ([kdnuggets, 2007]).*

CRISP-DM, está dividida en 4 niveles de abstracción organizados de forma jerárquica (figura 2.51) en tareas que van desde el nivel más general, hasta los casos más específicos y organiza el desarrollo de un proyecto de Data Mining, en una serie de seis fases (figura 2.52):

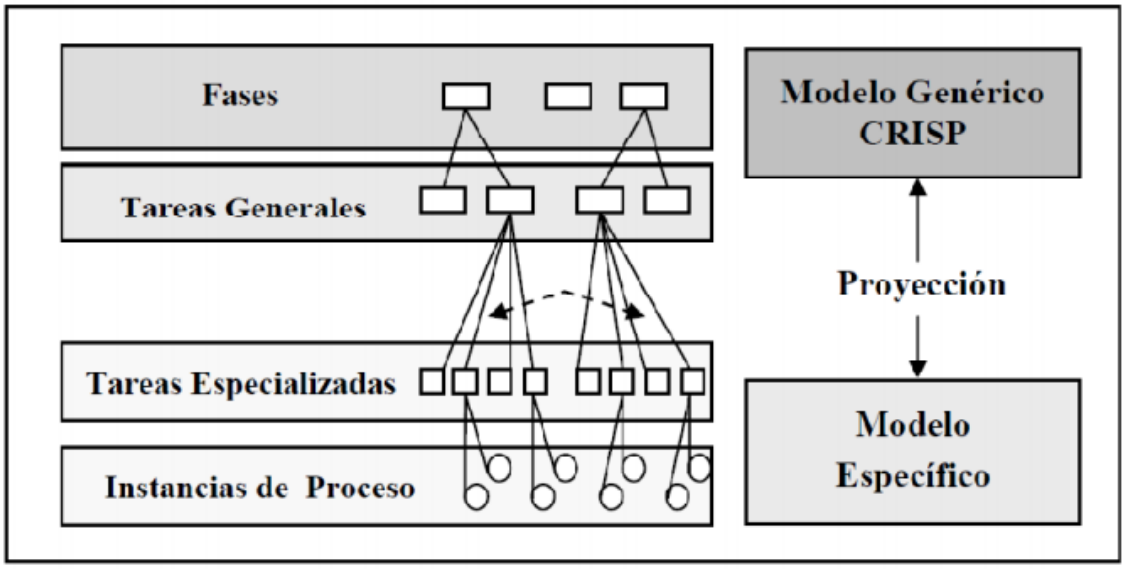


Figura 2.51. *Esquema de los 4 niveles de CRISP-DM ([CRISP-DM, 2000])*

La sucesión de fases no es necesariamente rígida. Cada fase es estructurada en varias tareas generales de segundo nivel. Las tareas generales se proyectan a tareas específicas, donde finalmente se describen las acciones que deben ser desarrolladas para situaciones específicas, pero en ningún momento se propone como realizarlas.

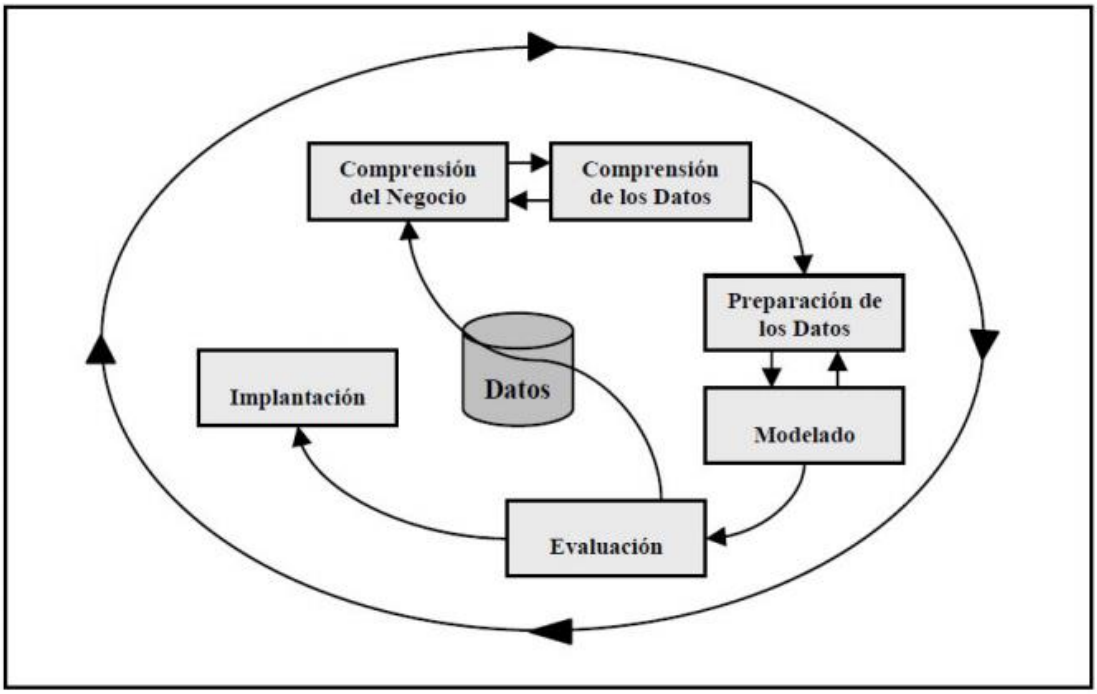


Figura 2.52. *Modelo de proceso CRISP–DM ([CRISP-DM, 2000]).*

Capítulo III. Marco Metodológico

3.1 Tipo de Investigación

El enfoque de la investigación que se realizó es cuantitativo, debido a que se obtendrá un número de retraso y cancelación de vuelos a través de la analítica de datos, para comprobar la hipótesis planteada al inicio del trabajo.

El tipo de estudio es correlacional; se eligió por qué "asocia variables mediante un patrón predecible para un grupo o población" (Sampieri, Fernández, & Baptista, 2014) ya que se comprobó la relación entre la variable independiente: Con analítica de datos se puede conocer el impacto en tiempo real de llegada de los vuelos y la variable dependiente: a través de los intervalos de tolerancia basados en la hora programada con respecto a la hora de salida. El patrón se buscó a través de aprendizaje automático, con un algoritmo de árboles de decisión en una base de datos proporcionada por el repositorio Kaggle denominada DelayedFlights, que contiene los datos de las aerolíneas y los tiempos necesarios para conocer si el vuelo está a tiempo o retrasado.

3.2 Universo, población o unidades de análisis

La población a considerar son las diferentes aerolíneas comerciales que salen de Estados Unidos. No se toman las aerolíneas que salen de México, por que no existen datos de libre acceso hasta la fecha.

3.3 Criterios de inclusión/exclusión

Los criterios de inclusión que se van a tomar en cuenta son la hora de llegada real y programada, la hora de salida real y programada de cada una de las aerolíneas que están registradas en la base de datos.

Los criterios de exclusión son los datos registrados en la base de datos que no se van analizar como el número de avión, destino, kilometraje, etc., lo que únicamente importa saber es si el vuelo sale a tiempo o cuenta con algún retraso.

3.4 Muestra

La muestra es la base de datos que contiene registrados los vuelos de aviones comerciales comprendidos entre 1998 al 2008.

3.5 Instrumentos

- Plataforma de software para el aprendizaje automático llamado Weka.
- Lenguaje de programación R.
- Algoritmo de árbol de decisiones C4.5 y random Forest.

3.6 Aparatos

- OS X Yosemite/ Versión 10, 10.5.
  IMac (21.5, pulgadas).
  Procesador 2.9 GHz Intel Core i5.
  Memoria 8GB 1600 MHz DDR3.
  Gráficos NVIDIA GeForce GT 750 M 1024 MB.

- HP Laptop 5N3D0ECL.
  Procesador AMD A8-7410 APU with AMD Radeon R5 Graphics 2.20 GHz.
  RAM 8 GB.
  S.O. 64 bits, procesador x64.

3.7 Procedimiento

Todo software requiere una serie de pasos específicos para que se desarrollE de forma correcta y además que se pueda cumplir cada uno de los objetivos planteados. Para llevarlo a cabo, es necesario hacer la implantación de algunas de las metodologías de desarrollo de software para obtener un producto de alta calidad.

Para realizar el proyecto y cubrir los objetivos planteados, se contó con la colaboración de alumnos de licenciatura para programar la plataforma Web que lleva el control administrativo de la empresa, mientras que la parte de analítica fue desarrollada como tema de Tesis de maestría. El proyecto se realizó con diferentes metodologías debido a la naturaleza del mismo, primero se describe el procedimiento que se realizó para la plataforma Web para la empresa MCP The World Parther y después el realizado para la base de datos de Kaggle.

3.7.1 Metodología para la plataforma Web

Antes de decidir el tipo de metodología para poder desarrollar la plataforma, se realizaron varias juntas con el cliente para conocer a fondo los procesos de exportación y el manejo de información, para con ello poder comprender la magnitud, alcance y viabilidad del proyecto.

Para realizar el software se estableció una metodología hibrída donde se conjuntan Scrum y el modelo incremental. A continuación se describen las fases que se siguieron para el desarrollo de la plataforma Web.

Análisis

En esta sección se especificaron los aspectos que intervienen para el desarrollo de dicha plataforma. Por lo cual se hace un análisis de requerimientos, que constó de varias partes que son: definiciones, descripción de funciones del sistema, características de usuarios y lista con todos los requerimientos.

Descripción de las funciones del sistema.

La plataforma realiza las siguientes funciones:

- Registro de usuarios: los usuarios pueden darse de alta en la plataforma ingresando los datos necesarios para su registro.

- Inicio de sesión a la plataforma: cada usuario tiene una cuenta única con la información que se agregó al registro, así como también toda información que se genere durante el tiempo que se utiliza la plataforma.
- Administrar productos: dentro de la plataforma, se encuentra una sección donde el usuario podrá agregar productos que frecuentemente exporta.
- Administrar contactos: en la plataforma, se encuentra una sección en la cual el usuario puede agregar datos de los contactos que utiliza, ya sea *consignatario*, *embarcador*, *agente aduanal*.
- Exportaciones o *bookings*: la parte más importante es en este punto, es en la cual el usuario puede hacer el proceso para poder realizar una exportación de manera exitosa.

Características de los usuarios

La plataforma se realizó para 2 tipos de usuarios:

- Administrador: es el encargado de agregar, navieras, contenedores, costos, etc.
- Usuario: persona que hará uso de la plataforma.

Requerimientos

En la tabla 3.1 se muestran los requerimientos necesarios para la realización del proyecto.

Abrv.	**Requerimiento no funcional**	**Dependencia**
R1	Diseño de base de datos	-
R2	Crear base de datos	R1
R3	Conectar base de datos	R2
R4	Validar Datos	-
R5	Diseño de las interfaces	-
R6	Mantenimiento de la base de datos	R1, R2, R3
R7	Diseño de *landing page*	-
R8	Crear nuevo usuario	R1,R2,R3,R4,R5
R9	Iniciar sesión	R1,R2,R3,R4
R10	Crear nuevo *booking*	R1,R2, R3,R4,R5,R8,R10
R11	Consultar *booking*	R1,R2,R3, R4,R5, R8,R9,R10
R12	Exportar productos	R1,R2,R3,R4,R5R10
R13	Proceso de exportación	R1,R2,R3,R4,R5R10

Tabla 3.1 *Requerimientos*

Diseño

En esta sección se encuentra desglosado el diseño de la base de datos, el diagrama Entidad-Relación, descripción de entidades y relaciones, limitantes de mapeo, diccionario de datos, entre otros. En la figura 3.1 se muestra como quedó el diagrama Entidad-Relación de la base de datos después de ser normalizada.

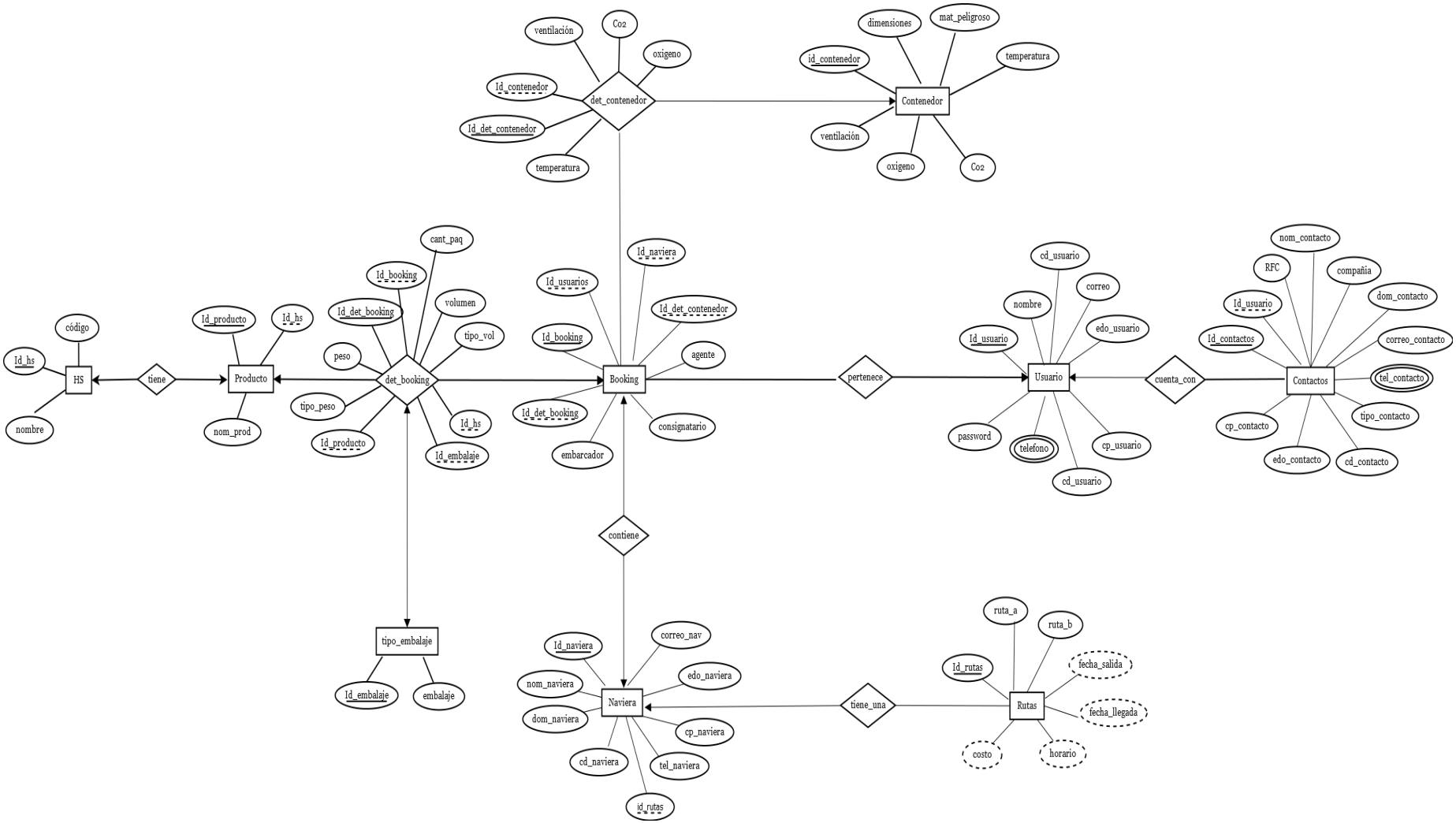


Figura 3.1. *Diagrama Entidad-Relación (E-R) Partner.*

Descripción de entidades y relaciones

En la tabla 3.2 se muestran las entidades que contiene la base de datos y la descripción de su función.

Entidad	Descripción
booking	Detalla datos principales para hacer una exportación (Como datos del *agente aduanal*, *consignatario*, naviera, entre otros).
producto	Describe los datos básicos del producto.
hs	Describe brevemente el código de cada producto arancelario
contenedor	Detalla características que tiene cada contenedor disponible.
usuario	Contiene los datos personales del usuario.
contactos	Contiene los datos que ingreso cada usuario (Estos datos solo los podrá ver el).
naviera	Describe los datos de la naviera
rutas	Detalla las rutas existentes.
tipo_embalaje	Detalla el tipo de tipo_embalaje que contienen un booking.

Tabla 3.2. *Descripción de entidades.*

En la tabla 3.3 muestran las relaciones del diagrama Entidad-Relación del sistema Partner.

Relación	Descripción
det_contenedor	Describe las variables que tendrá el contenedor para el producto.
det_booking	Detalla datos del producto que se va a exportar

Tabla 3.3. *Descripción de relaciones*

Limitantes de mapeo

En la tabla 3.4 se muestran las limitantes de mapeo, así como su descripción.

LIMITANTE	DESCRIPCIÓN
hs ↔ producto	Un hs pertenece a un producto.
producto← det_booking	Un producto tiene muchos det_booking.
det_bookin − tipo_embalaje	Muchos det_booking tienen diferentes tipo_embalaje.
booking ← det_booking	Un booking tiene muchos det_booking.
booking ↔ naviera	Un booking tiene una naviera.
naviera ← rutas	Una naviera tiene muchas rutas.
bookin − det_contenedor	Muchos booking tienen muchos det_contenedor.
contenedor← det_contenedor	Un contenedor tiene muchos det_contenedor.

usuario ← booking	Un usuario tiene muchos booking.
usuario ← contacto	Un usuario tiene muchos contactos.

***Tabla 3.4**. Limitantes de mapeo.*

En la tabla 3.5 se muestra el diccionario de datos, que contiene una breve descripción de los atributos que integran el diagrama de Entidad-Relación.

ATRIBUTO	**DESCRIPCIÓN**	**DOMINIO**
agente	Empresa que se encarga de hacer el despacho de mercancía.	Cadena válida para atributo agente, compuesta de 100 caracteres de la A-Z, a-z, espacios en blanco y acentos.
cant_paquete	Almacena el número de paquetes enviados.	Variable de tipo entero del 1-n.
celular_usu	Número telefónico móvil del usuario.	Conjunto de enteros compuesto por 10 dígitos, cada uno del 0-9.
ciudad_con	Ciudad de procedencia del contacto.	Cadena válida para atributo ciudad_con compuesta por 100 caracteres de la a-z, A-Z, espacios en blanco y acentos.
ciudad_nav	Almacena la ciudad donde se encuentra la naviera.	Cadena válida para atributo ciudad_nav, compuesta de 50 caracteres de la A-Z, a-z y acentos
ciudad_usu	Ciudad en la que procede el usuario.	Cadena válida para atributo ciudad_usu compuesta por 100 caracteres de la a-z, A-Z, espacios en blanco y acentos.
co2_conte	Si el contenedor contendrá dióxido de carbono regulado.	Conjunto de caracteres válidos para atributo co2_conte compuesto de 2 que solo pueden ser: false o true.
co2_detc	Porcentaje de dióxido de carbono en el que debe estar regulado el contenedor del booking.	Conjunto de dígitos flotantes solo acepta números dentro del siguiente rango: 0.04-21.
codigo	Almacena el código arancelario.	Variable tipo string, este formato se realizó de acuerdo al código arancelario. Ejemplo: 10.1a2b.125.
compania	Nombre de la compañía.	Cadena válida para atributo compania, compuesta de 100 caracteres de la A-Z, a-z, espacios en blanco y acentos.

consignatario	Persona quien recibe en destino (cliente).	Cadena válida para atributo consignatario, compuesta de 100 caracteres de la A-Z, a-z, espacios en blanco y acentos.
correo_con	Almacena el correo del contacto de la empresa.	Conjunto de caracteres válidos para atributo correo_con compuesta de 100 caracteres de la a-z, 1 punto o guion medio, @ y dominio valido como: gmail.com, Hotmail.com.
correo_nav	Guarda el correo oficial de la naviera	Conjunto de caracteres válidos para atributo correo_nav compuesta de 100 de la a-z, 1 punto o guion medio, @ y dominio valido como: gmail.com, Hotmail.com.
costo	Costo total del booking	Conjunto de dígitos flotantes del 0.01-99999.99.
cp_con	Guarda el número postal del contacto.	Conjunto de enteros del 0-9 compuesta por 5 dígitos, se entiende que 00000 no será válido.
cp_nav	Número postal de la ubicación física de la naviera.	Conjunto de enteros del 0-9 compuesta por 5 dígitos, se entiende que 00000 no será válido.
cp_usu	Número postal en la zona que se encuentra.	Conjunto de enteros del 0-9 compuesta por 5 dígitos, se entiende que 00000 no será válido.
dimensiones	Tamaño del contenedor.	Variable de tipo entero, solo acepta valores de 20 o 40.
domicilio_con	Guarda el domicilio del contacto.	Cadena válida para atributo domicilio_con, compuesta de 50 caracteres de la A-Z, a-z y acentos.
domicilio_nav	Guarda el domicilio de la naviera.	Cadena válida para atributo domicilio_nav, compuesta de 50 caracteres de la A-Z, a-z y acentos.
domicilio_usu	Guarda el domicilio de los usuarios.	Cadena válida para atributo domicilio_usu, compuesta de 50 caracteres de la A-Z, a-z y acentos.

duracion	Tiempo para llegar a su destino.	Conjunto de caracteres válidos para atributo fecha_salida compuesto de 8 caracteres cumpliendo la siguiente restricción: {0-31}/{0-12}/{17-99}.
embalaje	Tipo de envío del producto puede ser, caja, palet, bolsa, etc.	Cadena válida para atributo embalaje, compuesta de 30 caracteres de la A-Z, a-z, espacios en blanco y acentos.
embarcador	Empresa que se dedica a embarcar los contenedores o recibirlos.	Cadena válida para atributo embarcador, compuesta de 100 caracteres de la A-Z, a-z, espacios en blanco y acentos.
estado_con	Estado de procedencia del contacto.	Cadena válida para atributo estado_con compuesta por 100 caracteres de la a-z, A-Z, espacios en blanco y acentos.
estado_nav	Guarda el estado de la naviera	Cadena válida para atributo estado_nav, compuesta de 50 caracteres de la A-Z, a-z y acentos.
estado_usu	Estado en la que procede el usuario.	Cadena válida para atributo estado_usu compuesta por 100 caracteres de la a-z, A-Z, espacios en blanco y acentos.
fecha_salida	Fecha de salida del embarque o naviera.	Conjunto de caracteres válidos compuesto de 8 de ellos cumpliendo la siguiente restricción: {0-31}/{0-12}/{17-99}, respetando al menos 1 día mayor al día que se haga el booking.
id_booking	Identificador del booking.	Variable de tipo entero auto-incrementable del 1-n.
id_contacto	Identificador del contacto.	Variable de tipo entero auto-incrementable del 1-n.
id_contenedor	Identificador de los contenedores existentes.	Variable de tipo entero auto-incrementable del 1-n.
id_det_booking	Identificador de los detalles del booking.	Variable de tipo entero auto-incrementable del 1-n.
id_det_contenedor	Identificador de los detalles del contenedor.	Variable de tipo entero auto-incrementable del 1-n.
id_embalaje	Identificador de los embalajes existentes.	Variable de tipo entero auto-incrementable del 1-n.
id_hs	Clave para identificar el código arancelario.	Variable de tipo entero auto-incrementable del 1-n.

id_naviera	Identificador de las navieras existentes.	Variable de tipo entero auto-incrementable del 1-n.
id_producto	Identifica el producto registrado.	Variable de tipo entero auto-incrementable del 1-n.
id_producto	Identificador usado para ubicar el producto.	Variable de tipo entero auto-incrementable del 1-n.
id_rutas	Identificador usado para ubicar la ruta.	Variable de tipo entero auto-incrementable del 1-n.
id_usuario	Identificador de usuario registrado.	Variable de tipo entero auto-incrementable del 1-n.
mat_peligroso	Si el contenedor tendrá material peligroso o no.	Conjunto de caracteres válidos para atributo mat_peligroso compuesto de 2 de ellos que solo pueden ser: si o no.
nombre_con	Nombre del contacto.	Cadena válida para atributo nombre_con, compuesta de 100 caracteres de la A-Z, a-z, espacios en blanco y acentos.
nombre_hs	Nombre del código arancelario.	Cadena válida para atributo nombre_hs, compuesta de 50 caracteres de la A-Z, a-z y acentos.
nombre_nav	Nombre de la naviera.	Cadena válida para atributo nombre_nav, compuesta de 50 caracteres de la A-Z, a-z y acentos
nombre_prod	Nombre del producto.	Cadena válida para atributo nombre_prod, compuesta de 100 de la A-Z, a-z y acentos.
nombre_usu	Nombre del usuario.	Cadena válida para atributo nombre_usu, compuesta de 100 caracteres de la A-Z, a-z, espacios en blanco y acentos.
oxigeno_conte	Si el contenedor contendrá oxigeno regulado.	Conjunto de caracteres válidos para atributo oxigeno_conte compuesto de 2 de ellos que solo pueden ser: true o false.
oxigeno_detc	Porcentaje de oxígeno en el que debe estar regulado el contenedor del booking.	Conjunto de dígitos flotantes solo acepta números dentro del siguiente rango: 0.01-99.99.
password	Contraseña del usuario.	Cadena válida para el atributo password, compuesta por 20 caracteres de la A-Z, a-z, números y caracteres especiales.

peso	Guarda el peso de la mercancía enviada.	Variable de tipo flotante del 1.00-10.00
ruta_a	Almacena la ruta de puerto inicio o donde saldrá el embarque.	Cadena válida para atributo ruta_a, compuesta de 50 caracteres de la A-Z, a-z y acentos.
ruta_b	Almacena la ruta destino, al lugar donde llegara el embarque.	Cadena válida para atributo ruta_b, compuesta de 50 caracteres de la A-Z, a-z y acentos.
telefono_con	Guarda el teléfono del contacto.	Conjunto de enteros compuesto por 10 dígitos, cada uno del 0-9.
telefono_usu	Número telefónico del usuario incluyendo la lada.	Conjunto de enteros compuesto por 10 dígitos, cada uno del 0-9.
temperatura_conte	Si tendrá o no temperatura el contenedor.	Conjunto de caracteres válidos para atributo temperatura_conte compuesto de 2 que solo pueden ser: true o false.
temperatura_detc	Temperatura en el que debe estar regulado el contenedor del booking.	Conjunto de dígitos flotantes solo acepta números dentro del siguiente rango: 0.01-99.99.
tipo	Especifica si es agente aduanal, embarcador o consignatario.	Cadena válida para atributo tipo, compuesta de 100 de la A-Z, a-z, espacios en blanco y acentos.
tipo_peso	Almacena el tipo de peso, si es kilogramo o libra.	Variable de tipo carácter, solo acepta 2 valores que son: kg ó lb.
tipo_volumen	Índica si es pulgada o metro cubico.	Variable de tipo carácter, solo acepta 4 valores que son: $pulg$ ó $m3$.
ventilacion_conte	Si el contenedor contendrá ventilación regulada.	Conjunto de caracteres válidos para atributo ventilación compuesto de 2 valores que solo pueden ser: true ó false.
ventilacion_detc	Nivel de ventilación que debe estar regulado el contenedor del booking.	Conjunto de dígitos flotantes solo acepta números dentro del siguiente rango: 0.00-205.00.
volumen	Guarda el volumen de mercancía enviada.	Variable tipo flotante del 1.00-100.00

Tabla 3.5 *Diccionario de datos.*

Nota: en la residencias denominadas “Base de datos para un sistema de automatización de la logística en los procesos de exportación (Partner)” de César Aarón Sosa Patricio se realizó la normalización y los diagramas de UML del sistema y sus narrativas. A continuación se muestra solamente el diagrama de casos de usos general (figura 3.2), la narrativa de introducir login (cuadro 1.1) y el correspondiente diagrama de actividades (figura 3.3).

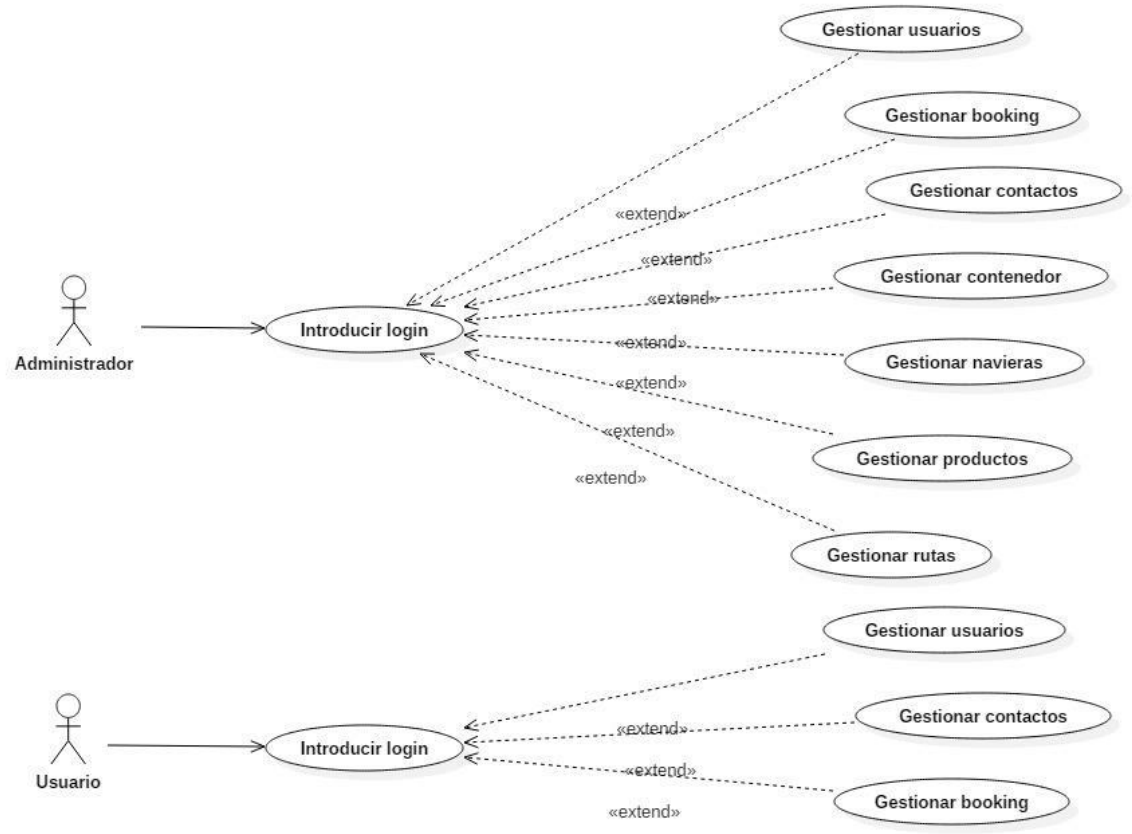


Figura 3.2. *Diagrama general de casos de uso del sistema.*

En el cuadro 1.1 se muestra la narrativa del caso de uso introducir login

de uso	Introducir login.
Meta en el contexto	Que el usuario pueda ingresar al sistema.
Alcance y nivel	Es una actividad primaria, porque se necesita acceder al sistema para poder elegir una opción. Además, de que otras actividades dependen de ella.
Precondiciones	Ninguna.
Condición final de éxito	Que se pueda introducir el correo y contraseña e ingresar al sistema.
Condición final de fallo	Que no se pueda introducir el correo y la contraseña e ingresar al sistema.
Actor primario	Administrador.

Actor secundario	Usuario.	
Lanzador	Que el Usuario necesite ingresar al sistema.	
Escenario de éxito principal	**Flujos alternativos**	
Acciones del Usuario/Administrador	**Acciones del Sistema**	
1. Inicio. 4. Introduce correo y contraseña.	2. Muestra interfaz de inicio. 3. Solicita de usuario/administrador y contraseña. 5. Recibe datos. 6. Revisa datos. 7. Envía datos para su búsqueda. 8. Permite el acceso al sistema. 9. Fin.	6.1 Si los datos son incorrectos, regresa al paso 3. 7.1 Si el usuario no se encuentra registrado, ir al paso 9.

Cuadro 3.1. *Narrativa del caso de uso de introduir login*

En la figura 3.3 se muestra el diagrama de actividades del caso de uso Introducir login

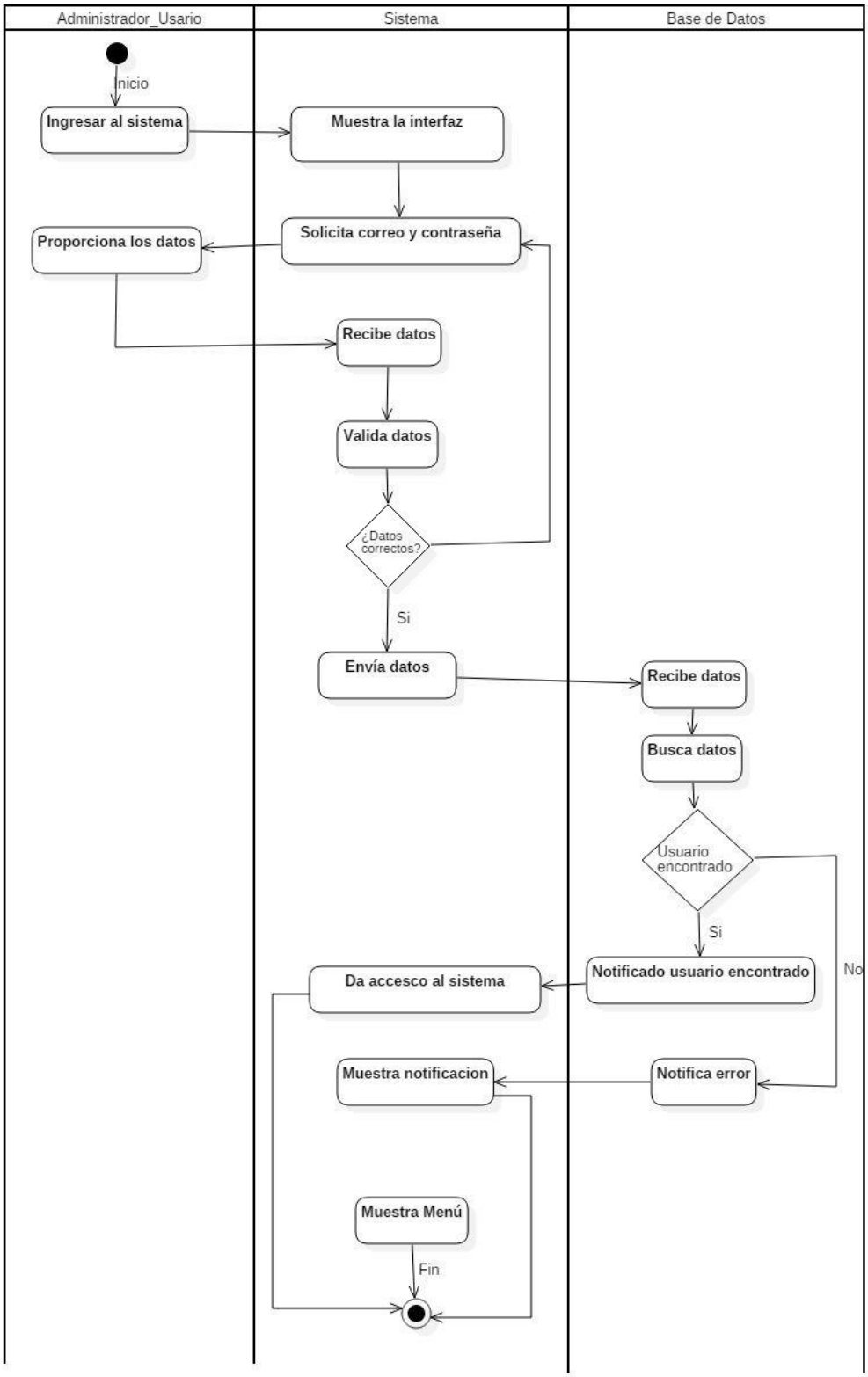


Figura 3.3. *Diagrama de actividades del caso de uso introducir login.*

En la figura 3.4 se muestra el diagrama de componentes del sistema Partner.

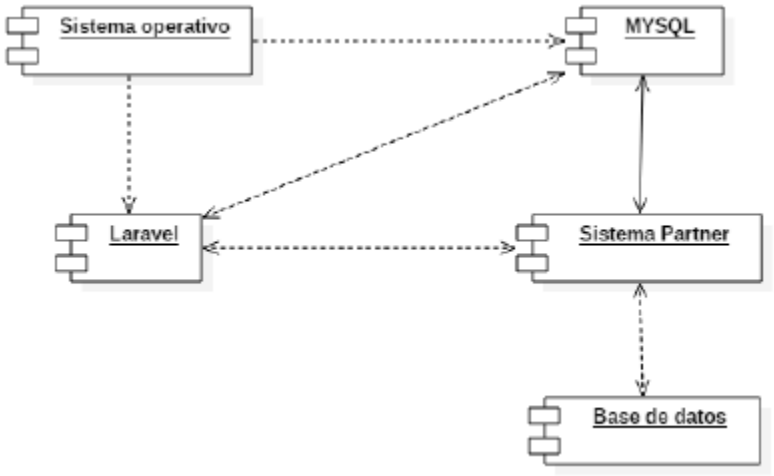


Figura 3.4. *Diagrama de componentes del sistema Partner.*

En la figura 3.5. se muestra el diagrama de despliegue del sistema Partner.

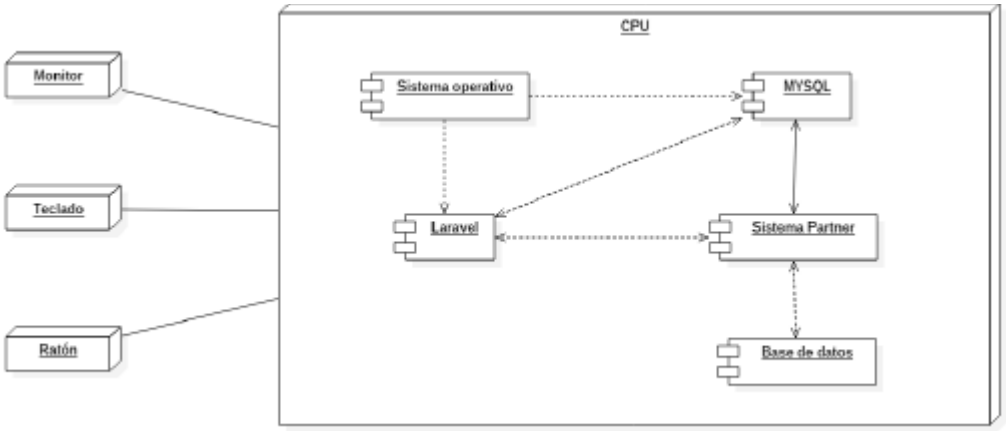


Figura 3.5. *Diagrama de despliegue del sistema Partner.*

En la figura 3.6 se muestra el diagrama de clases, se observa la estructura interna del sistema y las relaciones que poseen las clases.

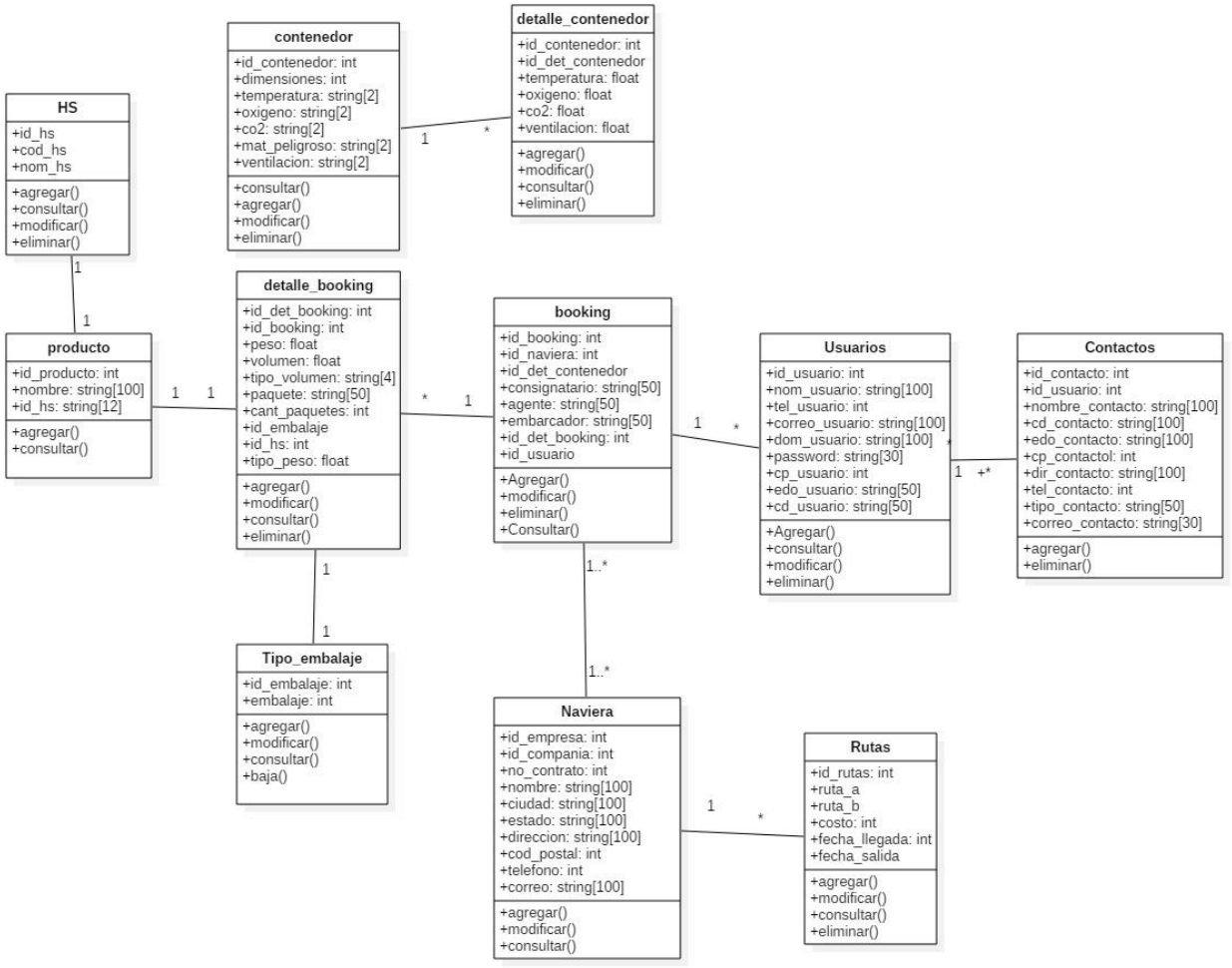


Figura 3.6. *Diagrama de clases de sistema Partner.*

Codificación

Para el manejo de la base de datos se usó un *framework* llamado Laravel que trabaja con el lenguaje PHP, que contiene un sistema de mapeo llamado Eloquent ORM para realizar el mapeo de la base de datos, la cual está realizada en MySQL.

Después de haber establecido el lenguaje y *framework* en el cual se realizó la base de datos, se hizo la generación del código en el manejador Eloquent ORM, tomando en cuenta que cada uno de los módulos debe contemplar:

1- Los que contengan llaves primarias y no dependan de otros módulos.

2- Los que contengan llaves primarias y dependan de los módulos anteriores.

Este código se generará automáticamente una sólo una vez al iniciar la plataforma.
Después de generar la base de datos, se envía informacióna través de un CRUD (Create, Read, Update y Delete/ Crear, leer, guardar y borrar), denominado Model, que es una forma en que ayuda a la plataforma a enviar y recibir datos de manera rápida, por medio de la generación de archivos Json para él envió de información. Estos Models, por defecto generan funciones CRUD para cada uno de los módulos dichos a continuación:

- hs.
- embalaje.
- producto.
- det_booking.
- booking.
- contenedor.
- det_contenedor.
- naviera.
- rutas.
- usuario.
- contactos.

Después de haber hecho el análisis y el diseño de la plataforma Web, se dividió tanto la codificación como las pruebas en tres residencias, La parte que documenta y codifica la base de datos quedó a cargo del actual ISC César Aarón Sosa Patricio y su trabajo se denominó : "Base de datos para un sistema de automatización de la logística en los procesos de exportación (PARTNER)". El desarrollo y comunicación de las interfaces y la Landing page, fue programado por la ISC Erika Lizbeth Gómez Ramos, cuyas residencias se denominaron: "Diseño y comunicación de interfaces para el sistema de automatización de la logística de los procesos de exportación". Siguiendo con el enfoque de programación MVC (Modelo, Vista, Controlador) la parte del controlador y la configuración del servidor fue desarrollado por el ahora ISC Antonio Caro Guerrero, cuya residencia se denominó : " Diseño de la API (Interfaz de Programación de Aplicaciones) para comunicación con plataforma de terceros (PARTNER)". La documentación que respalda el trabajo realizado por los entonces estudiantes de la Ingeniería en Sistemas Computacionales se encuentra en la hemeroteca de la biblioteca del Tecnológico de Cd. Guzmán.

3.7.2 Metodología para la analítica de datos

Existen diferentes metodologías donde todo depende de qué se quiera analizar. En este caso, Big Data es el proceso de recolección de datos en grandes cantidades y su tratamiento para encontrar patrones y correlaciones. Los datos que van a ser analizados, deben cumplir mínimo con las siguientes características:

Volumen: hace referencia a las cantidades masivas de datos que se almacenan con la finalidad de procesar dicha información, transformando los datos en acciones.

Velocidad: se refiere a los datos en movimiento por las constantes interconexiones que se realizan, es decir, a la rapidez en la que son creados, almacenados y procesados en tiempo real.

Variedad: se refiere a las formas, tipos y fuentes en las que se registran los datos.

Veracidad: se hace referencia a la incertidumbre de los datos, es decir, al grado de fiabilidad de la información recibida.

Valor*:* el valor se obtiene de datos que se transforman en información; esta a su vez se convierte en conocimiento, y este en acción o en decisión. El valor de los datos está en que sean accionables, es decir, que los responsables de las empresas puedan tomar una decisión (la mejor decisión) en base a estos datos ("Las 7 V del Big data: Características más importantes - IIC", 2018).

Aunque existen grandes bases de datos que manejan petabytes, estas no son consideradas Big Data por proceder de una misma fuente. Existen diferentes plataformas que ofrecen bases de datos que cumplen con estos requisitos, etre ellas Twitter, que cuenta con el servicio y cobra con base al tiempo que se accese y se utilicen sus bases de datos. ("BASES DE DATOS (@basesdedatos) on Twitter", 2018).

Al ser esta investigación de carácter educativo y no contar con los fondos suficientes para pagar por el uso de la información, se optó por una plataforma de carácter científico que

ofrece bases de datos libres, con la finalidad de que los estudiantes hagan aportaciones a las empresas que liberan sus datos de forma gratuita. Kaggle es un sitio seguro que maneja información real y confiable, con lo que se cubre el requisito de veracidad.

El valor de la base de datos seleccionada radica en que las diferentes aerolíneas toman y continuarán tomando decisiones con base a los diferentes resultados obtenidos por cada uno de los analistas que suben sus trabajos a la plataforma y lograr ver de diferente forma la información, buscando patrones en sus vuelos, demoras, cancelaciones y llegadas al destino, que les sirva a los ejecutivos de las aerolíneas, donde encuentren nuevas áreas de oportunidad.

La característica de velocidad se cubre con que la información se genera en tiempo real y cada minuto se está generando nueva información, mientras la variedad proviene de las bases de datos de diferentes aerolíneas donde cada una cuenta con una gran cantidad de vuelos y destinos.

Otro reto enfrentado, fue el encontrar una metodología adecuada para realizar la analítica de datos, como el desarrollo de esta actividad se encuentra inmersa dentro de metodologías ágiles que utilizan las empresas, se decidió seguir un proceso utilizado en la minería de datos, denominado CRISP-DM, cuyas fases se adaptan perfectamente al desarrollo de la analítica para Big Data, inclusive IBM utiliza esta metodología como base de su propia analítica ("s|ngular - CRISP-DM: La metodología para poner orden en los proyectos de Data Science", 2018). A continuación se detalla cómo se implementó esta metodología en el desarrollo del proyecto.

1. Fase de comprensión del negocio o problema

Esta fase cuenta con diferentes tareas que son:

1. Determinar los objetivos del problema: desarrollar la analítica de datos para una Base de Datos de la plataforma de Kaggle llamada DelayedFlights, para conocer el impacto en tiempo real de llegada de los vuelos, a través de los intervalos de tolerancia basados

en la hora programada con respecto a la hora de salida, con el fin de evaluar si se afecta la entrega a tiempo de la mercancía exportada.

- Evaluación de la situación: en esta tarea se analiza la situación y requerimientos que se necesitan. Se investigaron cuáles son los rangos de tiempo que se consideran como un retaso en la salida de vuelo y se encontró que debe de haber transcurrido entre 15 minutos y hasta menos de 8 horas, mientras que después de este tiempo, el vuelo se considera como cancelado. Estos factores denominados intervalos de tolerancia permen definir las decisiones que toma el modelo de clasificación seleccionado.
- Determinación de los objetivos: tiene como función representar los objetivos en términos de las metas del proyecto, los objetivos a cumplir son:

1. Buscar la base de datos del medio de transporte para poder realizar las exportaciones en el repositorio Kaggle.
2. Localizar la base de datos con las características más idóneas.
3. Según las características de la base de datos, seleccionar el algoritmo para realizar la analítica de datos.
4. Realizar pruebas con una porción de datos en Weka para ver si el algoritmo es el apropiado.
5. Programar en R o Python el algoritmo de aprendizaje automático y determinar si los retrasos de los vuelos afectan los tiempos de entrega de la mercancía exportada.
6. Documentar los resultados obtenidos.

2. Fase de comprensión de los datos

En esta fase es donde la base de datos sufre cambios ya sea que se le agreguen o eliminen campos dependiendo de la necesidad. La base de datos DelayedFligths cuenta con veintinueve registros de los cuales, después de hacer un análisis de cada uno de los campos se seleccionaron los siguientes: hora de salida programada y hora real programada, hora de llegada programada y hora real de llegada para poder efectuar la analítica y con ello por cumplir con la hipótesis planteada.

La fase de comprensión cuenta con diferentes tareas que son:

- Recolección de datos iniciales. La plataforma Kaggle realizó la recolección de los datos y los pone en un *archivo CSV* denominado en la página donde están a disposición de la comunidad científica. La base de datos se llama Delayerflith y cuenta con 1,000,000 de registros. Esta recolección abarca los vuelos desde el 1998 hasta el 2008.
- Descripción de los datos: a continuación se describen los campos que contiene la base de datos.

No.	Name	Descripción
1	Year	Año de registro.
2	Month	Mes.
3	DayofMonth	Día del mes.
4	DayOfWeek	1 (lunes) - 7 (domingo).
5	DepTime	tiempo de salida real (local, hhmm).
6	CRSDepTime	hora de salida programada (local, hhmm).
7	ArrTime	hora de llegada real (local, hhmm).
8	CRSArrTime	hora de llegada programada (local, hhmm).
9	UniqueCarrier	Código único de transportista.
10	FlightNum	Número de vuelo.
11	TailNum	registro de la aeronave, identificador único de la aeronave.
12	ActualElapsedTime	Tiempo actual transcurrido en minutos.
13	CRSElapsedTime	Tiempo transcurrido en minutos.
14	AirTime	Tiempo en antena.
15	ArrDelay	Tiempo del vuelo.
16	DepDelay	Demora del vuelo.
17	Origin	Origen.
18	Dest	Destino.
19	Distance	Distancia en millas.
20	TaxiIn	Tiempo de entrada.
21	TaxiOut	Tiempo de salida.
22	Cancelled	Cancelado.
23	CancellationCode	motivo de cancelación (A = transportista, B = clima, C = NAS, D = seguridad).
24	Diverted	Desviado 1 = sí, 0 = no.

25	CarrierDelay	Retraso del operador.
26	WeatherDelay	Retraso meteorológico
27	NASDelay	Retraso que está dentro del control del Sistema Nacional de Espacio Aéreo (NAS).
28	SecurityDelay	Retraso de seguridad.
29	LateAircraftDelay	Retraso de llegada al aeropuerto.

Tabla 3.6. *Descripción de datos*

Como se puede observar, todos los campos están relacionados tanto con retrasos como con cancelaciones de los vuelos y los motivos existentes. Como en otros trabajos consultados en la plataforma se enfocan en la cancelación de los vuelos y el porqué, se decidió que en este trabajo se dirige hacia la posible existencia de los retrasos en las llegadas de los vuelos a su destino, debido a que es un factor importante para decidir si el producto que se va a exportar se envía por medio de un vuelo o se decide por otro medio de trasporte.

Exploración de datos: para realizar la exploración de datos se utilizó la plataforma Weka para para ver si la información que se está tomando permitirá conocer los tiempos de holgura que se tiene en la llegada como en la salida de los vuelos. A continuación se muestra la ventana de la plataforma (figura 3.7) donde se muestra la información procesada con respecto a los tiempos de holgura de la base de datos.

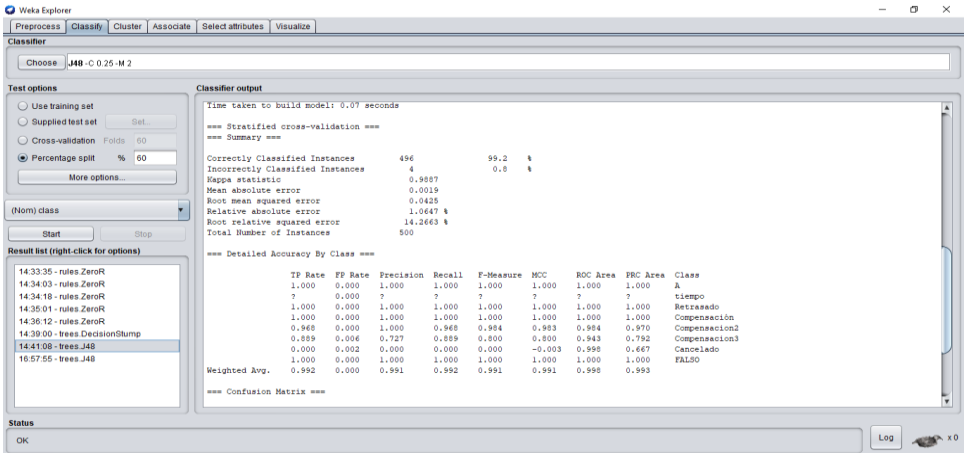


Figura 3.7*. Base de datos procesada por Weka.*

en esta fase se hicieron pruebas en la plataforma Weka con otra técnica que es para problemas estadísticos que son las Redes bayesianas, comparando los métodos se

obtuvo que los árboles de decisión fue el que arrojó mejor resultado, a continuación se describe el procedimiento de las pruebas.

1. Se uso una plataforma que valida la confiabilidad de los datos.
2. Se utilizaron dos diferentes métodos redes bayesianas y árboles de decisión
3. Se decicidio por arboles por el que clasifica mayor información de la base de datos.
4. Los resultados de las pruebas realizadas en Weka es tan en el apartado de resultados en tema “4.1 pruebas realizadas”

3. Fase de preparación de los datos

La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato. Esta fase se encuentra relacionada con la fase de modelado, puesto que en función de la técnica de modelado elegida para este proyecto se decidió elegirla técnica de árboles de decisión.

Una descripción de las tareas involucradas en esta fase son las siguientes:

1. Selección de datos: en esta tarea como su nombre lo indica se va a seleccionar los datos con los que se va a trabajar, apoyándose en las tareas anteriores para elegir los correctos. Se hizo la selección de las columnas DepTime, CRSTime, ArrTime, CRSArrTime, los datos de estas cuatro columnas contienen la información tanto la hora de llegada y de salida programada como la hora de llegada y de salida real, con esta información se puede conocer cuáles fueron los tiempos de retrasos para cumplir con el objetivo propuesto al inicio del proyecto.
2. Limpieza de los datos: **data cleansing o scrubbing** es un proceso necesario para asegurar la calidad de los datos que se emplearán para analítica. Este paso es fundamental para minimizar el riesgo que supondría el basar la toma de decisiones en información poco precisa, errónea o incompleta ("Data cleansing y sus fases: contra los problemas de calidad de datos", 2018), con la limpieza de los datos la base se hace menos pesada para poder accesarla.

Para hacer la limpieza de los datos se utilizó el lenguaje de programación R, una vez que se lee la base de datos y se asigna a una variable **vueretraso** para que cuando se le mande llamar sea más fácil y rápido por si se necesita volver a llamar a la base de datos, en la línea de código se le indica hacer la limpieza con la sintaxis: vueretraso <- vuelos[,c(5,6,7,8)]

Con esta línea se le está indicando a la base de datos que haga una limpieza de la base de datos que nada más se van a utilizar las cuatro columnas de:

DepTime: tiempo de salida real (hh:mm)
CRSDepTime: hora de salida programada (hh:mm)
ArrTime: hora de llegada real (hh:mm)
CRSArrTime: hora de llegada programada (hh:mm)

Con la finalidad de hacer más ligera la base de datos y no utilizar datos de otra información que no se necesita. En las dos líneas de código se indica que la que a la variable **vue** se le va a signar la base de datos con las cuatro columnas seleccionadas : vue<-(vueretraso).

4. Fase de modelado

En esta fase de CRISP-DM, se seleccionan las técnicas de modelado más apropiadas para el proyecto, aquí es donde se decide con cuál de las técnicas mencionadas en el marco teórico es la ideal para la cumplir con los objetivos planteados.

Las tareas de esta fase son las siguientes:

1. Selección de la técnica de modelado: Cómo primer paso, se decidió utilizar un algoritmo de aprendizaje automático debido a la capacidad que tiene ésta técnica de generar nuevo conocimiento (Martínez, 18). Existen 10 algoritmos esenciales de Machine learning ("Los 10 Algoritmos esenciales en Machine Learning - Raona", 2018), a continuación se enlistará los nombres de los algoritmos:
    a) **Árboles de decisión**: un árbol de decisiones es una herramienta de apoyo a la decisión que utiliza un gráfico o un modelo similar a un árbol de decisiones.
    b) **Naïve Bayes Clasification:** los claseificadores Naïve Bayes son una familia de simples clasificadores probabilísticos basado en la aplicación de Bayes

'teorema con fuertes (Naïve) supuestos de independencia entre las características'.

c) **Regresión lineal**: la tarea de ajustar una línea recta a través de un conjunto de puntos. Hay varias estrategias posibles para hacer esto, y la estrategia de "mínimos cuadrados ordinarios" va así: puede dibujar una línea y luego, para cada uno de los puntos de datos, medir la distancia vertical entre el punto y la línea y sumarlos.

d) **La regresión logística:** mide la relación entre la variable dependiente categórica y una o más variables independientes estimando las probabilidades utilizando una función logística, que es la distribución logística acumulativa.

e) **Support Vector Machines:** SVM es un algoritmo de clasificación binario, Digamos que se tienen algunos puntos de 2 tipos en un papel que son linealmente separables. SVM encontrará una línea recta que separa esos puntos en 2 tipos y situados lo más lejos posible de todos esos puntos.

f) **Métodos Ensemble:** los métodos Ensemble son algoritmos de aprendizaje que construyen un conjunto de clasificadores y luego clasifican nuevos puntos de datos tomando un voto ponderado de sus predicciones.

g) **Algoritmos Clustering:** Clustering es la tarea de agrupar un conjunto de objetos tales que los objetos en el mismo grupo (cluster) son más similares entre sí que a los de otros grupos.

h) **Análisis de Componentes Principales**: PCA es un procedimiento estadístico que usa una transformación ortogonal para convertir un conjunto de observaciones de variables posiblemente correlacionadas en un conjunto de valores de variables linealmente no correlacionadas llamadas componentes principales.

i) **Singular Value Decomposition:** en el álgebra lineal, SVD es una factorización de una matriz compleja real. Para una matriz M * n dada, existe una descomposición tal que M = UΣV, donde U y V son matrices unitarias y Σ es una matriz diagonal.

j) **Análisis de Componentes Independientes:** ICA es una técnica estadística para revelar los factores ocultos que subyacen a conjuntos de variables,

mediciones o señales aleatorias. ICA define un modelo generativo para los datos multivariados observados, que se suele dar como una gran base de datos de muestras

De los cuales se decidió utilizar el de árboles de decisión ya que son específicos para realizar clasificación, funciona para los datos numéricos o categóricos, usa un modelo de caja blanca (lo que hace que los resultados sean fáciles de exponer), explican el comportamiento respecto a una determinada tarea de decisión, reduce el número de variables independientes y se analizan todas las posibles consecuencias de tomar una decisión.

De los diferentes árboles de decisión que existen R implementa un algoritmo denominado random Forest, que es un método que combina una cantidad grande de árboles de decisión independientes probados sobre conjuntos de datos aleatorios con igual distribución y los árboles de decisión sencillos sólo se construye mediante un proceso de clasificación, es decir el árbol de decisión sencillo nada más clasifica una sola desición y el random forest clasifica varías decisiones. Es una técnica de aprendizaje automático supervisado, por lo tanto se entrenó el algoritmo con el 60% de la información de la base de datos, y después se "aplicó lo aprendido" con el 40% restante de los datos.

En la instrucción: **Ind <- sample (2, nrow(vueretraso),remplace=TRUE, prob=c(0.6,0.4))**, se crea un arreglo que será utilizado para decirle al modelo que utilice el 60 % para entrenamiento y el 40% para probar lo aprendido por el árbol.

2. Generación del plan de prueba*:* Una vez construido un modelo, se debe generar un procedimiento destinado a probar la calidad y validez del mismo. Entonces, típicamente se separan los datos en dos conjuntos, uno de entrenamiento y otro de prueba, para luego construir el modelo basado en el conjunto de entrenamiento y medir la calidad del modelo generado con el conjunto de prueba.

De los diferentes árboles de decisión que existen R implementa un algoritmo denominado random Forest, que es un método que combina una cantidad grande de árboles de decisión independientes probados sobre conjuntos de datos aleatorios con igual distribución y los árboles de decisión sencillos sólo se construye mediante un proceso de clasificación, es decir

el árbol de decisión sencillo nada más clasifica una sola desición y el random forest clasifica varías decisiones. Es una técnica de aprendizaje automático supervisado, por lo tanto se entrenó el algoritmo con el 60% de la información de la base de datos, y después se "aplicó lo aprendido" con el 40% restante de los datos.

En la instrucción: **Ind <- sample (2, nrow(vueretraso),remplace=TRUE, prob=c(0.6,0.4))**, se crea un arreglo que será utilizado para decirle al modelo que utilice el 60 % para entrenamiento y el 40% para probar lo aprendido por el árbol.

3. Construcción y evaluación del modelo: a continuación se describen los pasos a grandes rasgos que se hicieron para llevar a cabo la programación del modelo:

    - Cargar las librerías necesarias para hacer el árbol.
    - Leer la base de datos y asignar a una variable.
    - Cargar la variable e informarle con que campos se va a trabajar de esa base de datos.
    - Realizar árbol de decisión.
    - Gráficar el árbol de decisión.
    - Mostrar las estadísticas del modelo.

El uso de los parámetros utilizados en el modelo, los resultados, la interpretación y el rendimiento de la técnica seleccionada se encuentran el el capítulo de resultados bajo el nombre de "4.3 Resultados obtenidos".

5. Fase de evaluación

En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. De acuerdo a los objetivos planteados se evalúan los resultados arrojadas por el modelo y si es necesario corregir algo del procedimiento para mejorar los resultados, en el caso de la base de datos el programa arrojo buen resultado más del 80% de los datos son clasificados de manera correcta, que se puede confiar para tomar una decisión con respecto a los intervalos de tiempo que se tienen en las llegadas y salidas de los aviones. Se optó por dibujar en un árbol de decisones los resultados arrojados por el modelo. Esta figura es la 4.23

denominada ábol de decisión y la interpretación de los datos se encuentra explicada en el mismo apartado.

6. Implementación

Los resultados obtenidos en la presente tesis, específicamente a la programación del árbol de decisión y la interpretación de los resultados junto con las conclusiones serán traducidos al idioma inglés para poder subirlos a la plataforma de Kaggle, para dar a conocer el impacto en tiempo real de llegada de los vuelos, a través de los intervalos de tolerancia basados en la hora programada con respecto a la hora de salida, con el fin de evaluar si se afecta la entrega a tiempo de la mercancía exportada, esto con la finalidad de cubrir uno de los requisitos del uso de ésta base de datos, el dar a conocer a la comunidad científica y de negocios el resultado de utilizar una de sus bases de datos y ser parte de la competencia convocada en la plataforma.

Capítulo IV. Resultados

A manera de resultados se puede interpretar que el sistema está elaborado en dos partes, la primera es el sistema Web montado en un host, realizado para poder llevar los procesos administrativos y con ello el objetivo de la empresa es darse a conocer de manera internacional, la segunda parte es la analítica que se realizó con la base de datos DelayFligths, ya que como la base de datos de la empresa contaba con muy pocos registros lo cual no cumplía con los requisitos por eso se optó por buscar la base de datos que cumpliera con las características de Big Data para desempeñar el proyecto.

Cuando se inició el proyecto, la empresa contaba con un diseño del sistema aunque sólo se tenía la propuesta de cómo sería la interfaz para que el usuario realice una reservación para la exportación.

La figura 4.1 muestra la interfaz principal para realizar la reservación, los datos a capturar es el origen de donde parte la mercancía, destino hacia dónde va, y la fecha de embarque de la mercancía.

Figura 4.1. *Ventana para reservación***.**

Una vez que se ha seleccionado origen-destino, el siguiente paso es seleccionar el producto que se va a importar, tal como se muestra en la figura 4.2

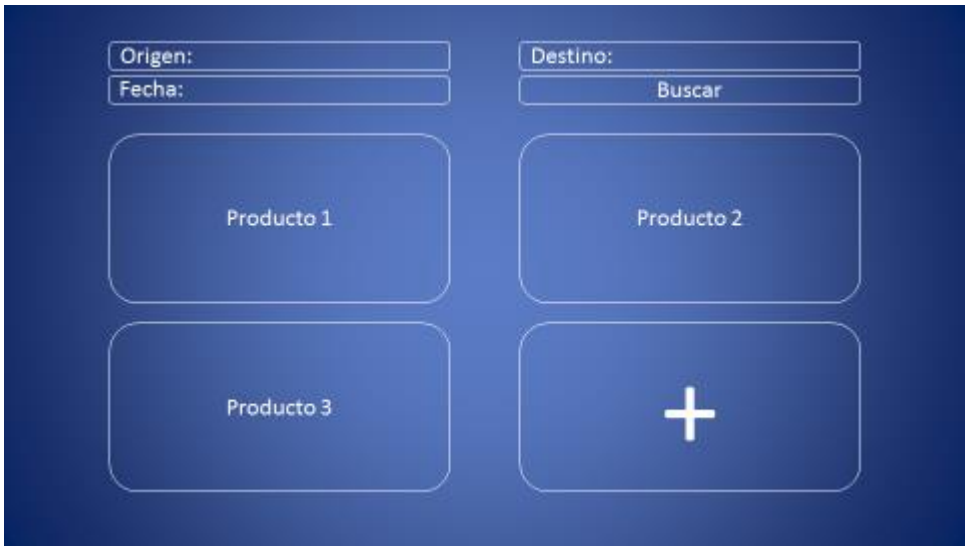


Figura 4.2. *Ventana para agregar los productos a exportar.*

Una vez que se agregó el producto o los productos, se agregan los clientes tal como se muestra en la figura 4.3

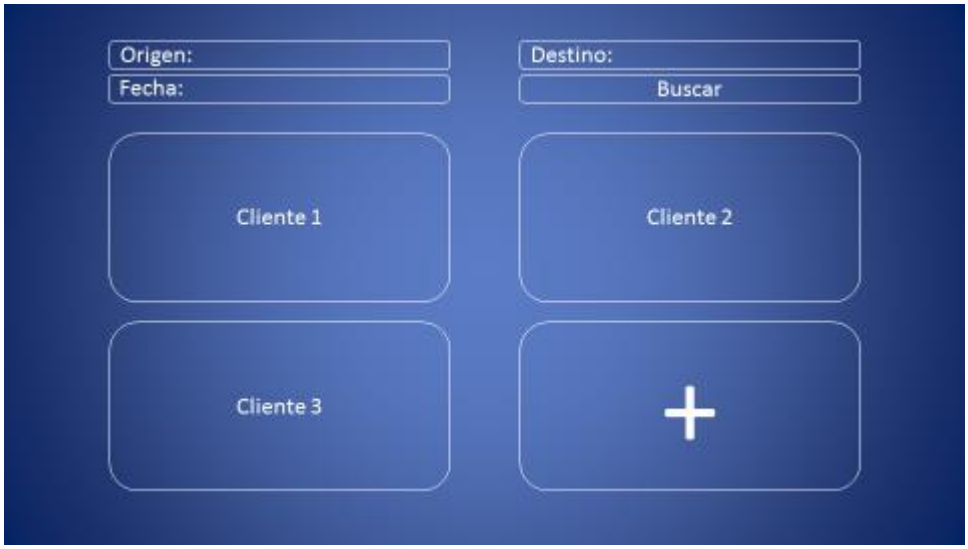


Figura 4.3 *Ventana para agregar clientes.*

En la figura 4.4 muestra las diferentes navieras, los costos y el tiempo de entrega de exportación. Aquí el cliente selecciona su mejor opción.

Figura 4.4. *Ventana para seleccionar la naviera.*

Una vez seleccionada la naviera es importante conocer quien se va encargar del arrastre terrestre de donde está la mercancía al puerto, puede ser Partner o directamente la empresa naviera tal como se muestra en la figura 4.5:

Figura 4.5*. Ventana para hacer el arrastre****.***

.

A continuación se muestran algunas de las imágenes de cómo quedó el sistema que se entregó a la empresa MCP Partner, se pondrán algunas de las principales ventanas.

Cuando el usuario entra a la plataforma lo primero que aparece es la figura 4.6, esta contiene la información básica, así como el estado de los Bookings que el usuario haya creado y los que ya han sido completados.

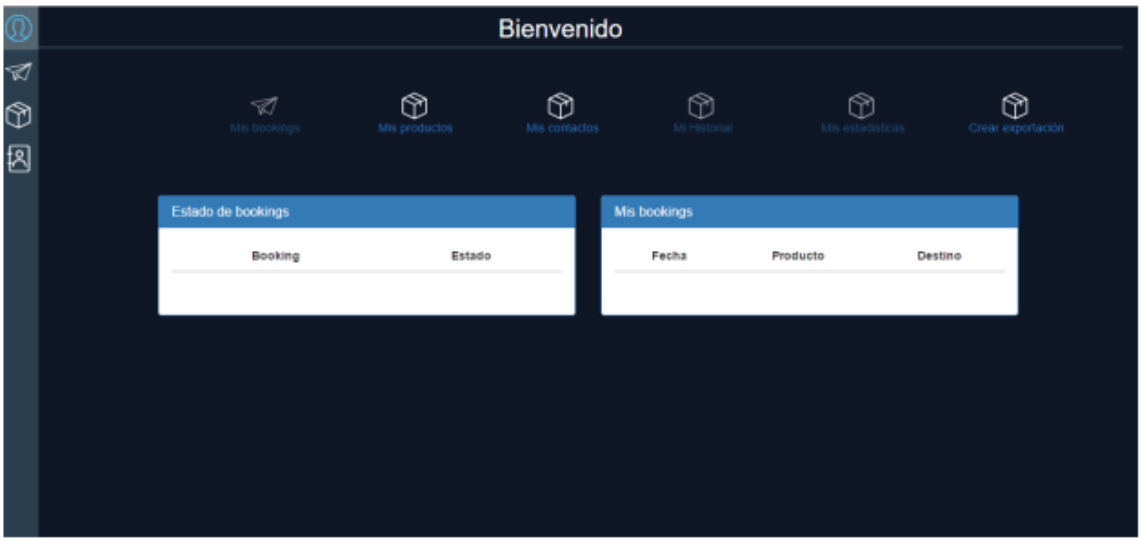


Figura 4.6. *Ventana de inicio*

Para poder realizar un boking se despliegan diferentes menús, cuando el usuario completa uno le da la opción de rectificar los datos, anterior o el siguiente menú hasta terminar con todos los datos necesarios para poder realizar la exportación con la finalidad de no saturarlo con tanta información, la figura 4.7 muestra el primer menú que es la selección de origen y destino de la exportación.

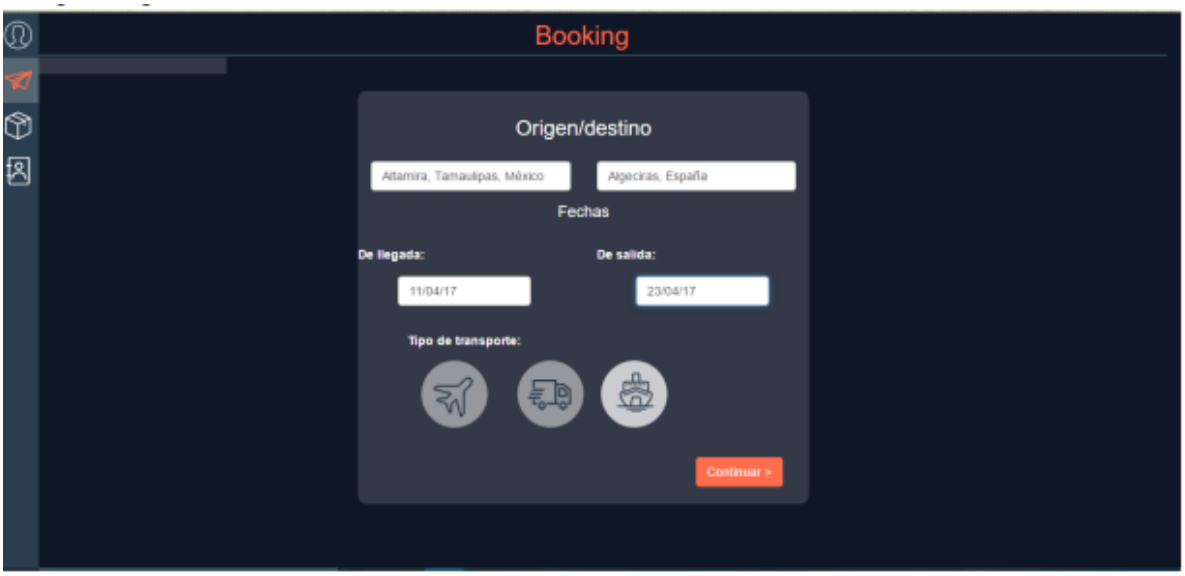


Figura 4.7. *Ventana origen/destino***.**

Una vez completado el origen/destino, se procede a seleccionar el producto, la naviera por la cual se va a enviar, tipo de contenedor, contrato, contenedor, producto a enviar, estos son los datos necesarios que se tienen que completar para poder realizar el booking.

Se realizó una landing page con la información de la empresa con una breve descripción de la plataforma y los servicios que ofrece, las navieras con las que se trabaja (figura 4.8).

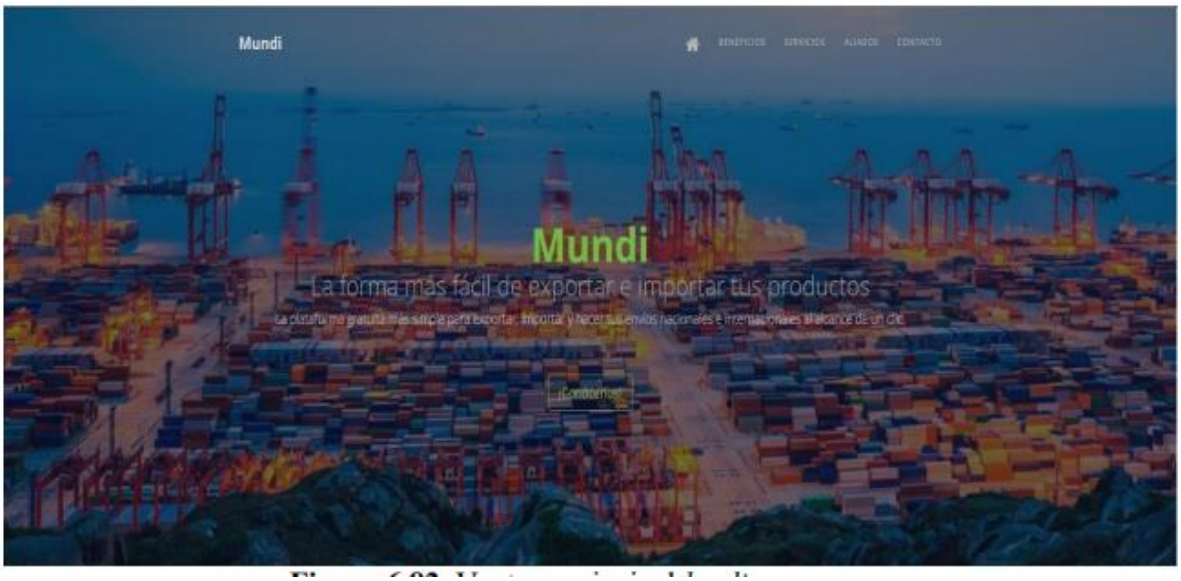


Figura 4.8. *Landing page***.**

4.1 Pruebas realizadas

Antes de trabajar con toda la base de datos, se realizaron pruebas a través de la plataforma Weka para probar diferentes algoritmos y corroborar que la base de datos con la que se trabaja tiene la información necesaria para poder cumplir con el objetivo del proyecto y saber que tan confiable es, para al momento de clasificar saber cuánto es el porcentaje de acierto o error en los datos.En seguida se describen los pasos que se siguieron.

En la plataforma como entrenamiento para realizar pruebas con la base datos se tomó una muestra de 500 datos, se arrojó el siguiente resultado: entrenando el 60% de la base de datos y el 40% para pruebas, se arroja el resultado de 99.2 % de los casos se han clasificado bien, mientras que el 0.8% lo han hecho incorrectamente, a continuación se muestra la figura 4.9 con estos datos.

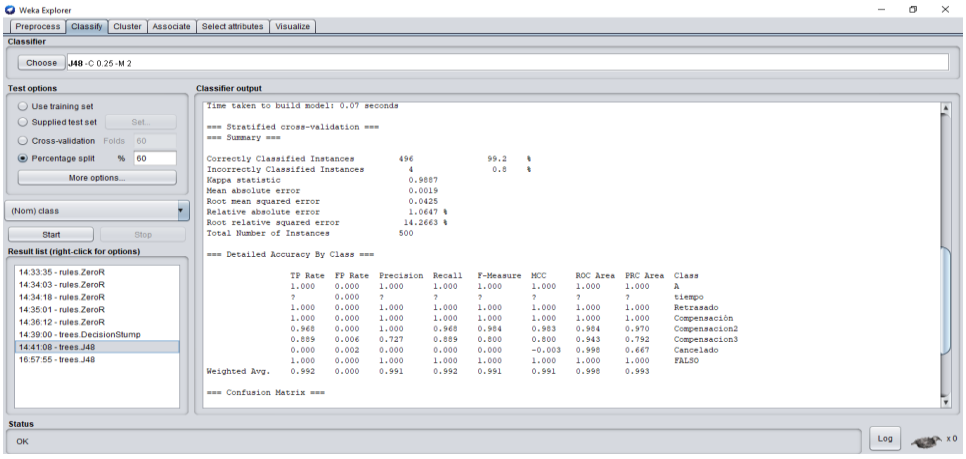


Figura 4.9. *Muestra de resultados de la prueba.*

En la figura 4.10 se muestra la información necesaria como la cantidad de atributos, se tomaron 4 columnas de la base de datos para analizar, el algoritmo que se va a utilizar para esta ocación será el J48 para formar el árbol de decisión, también se muestran los tiempos que se tomaron en cuenta para considerarse retraso, mayor a los 15 minutos y menor a una hora, mayor de una 1 y menor de 2 horas, etc., hasta llegar a 4 horas, si se pasa de 8 horas se considera un vuelo cancelado.

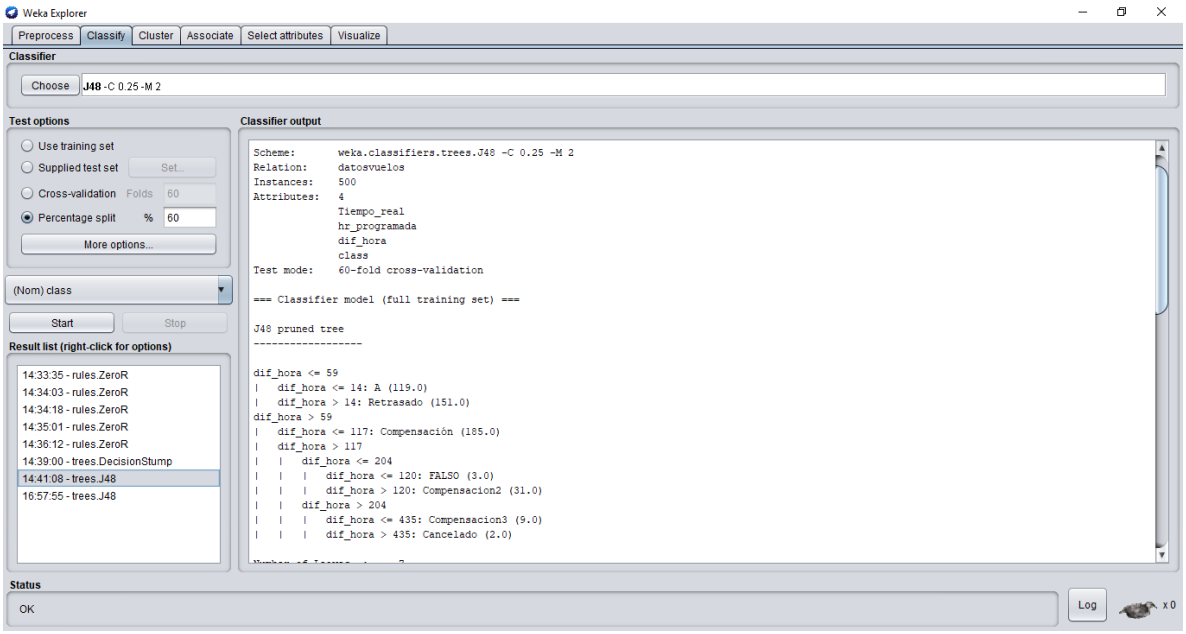


Figura 4.10 *Datos para formar el árbol de decisión.*

En la figura 4.11 se muestra el árbol de decisión que se realizó con la plataforma Weka con la información siguiente: el árbol se divide en dos nodos de los cuales uno es <= 59 min y en ese nodo se conforma de otros dos nodos, si es <= a 14 min salió a tiempo, si es >14 min el vuelo salió con retraso, en el nodo que índica que es >59 min se va abriendo o aparecen más nodos, eso revela que va ir indicando los tiempos de retraso de una hora, dos horas, etc. hasta llegar a la cancelación por completo del vuelo.

Se hicieron otras pruebas con el método de Redes Bayesianas al igual que los árboles de decisión son algoritmos probabilísticos, las redes bayesianas dan probabibidades y los árboles de decisión son más determinantes, comparando los resultados se obtuvo:

En la figura 4.12 se muestran los datos con los que se trabajaron en una red bayesiana, para este caso se utilizó una muestra de 500 registros, con 4 atributos al igual que con el árbol de decisión, se utilizó el 60% de los datos para entrenar y el 40 % de prueba.

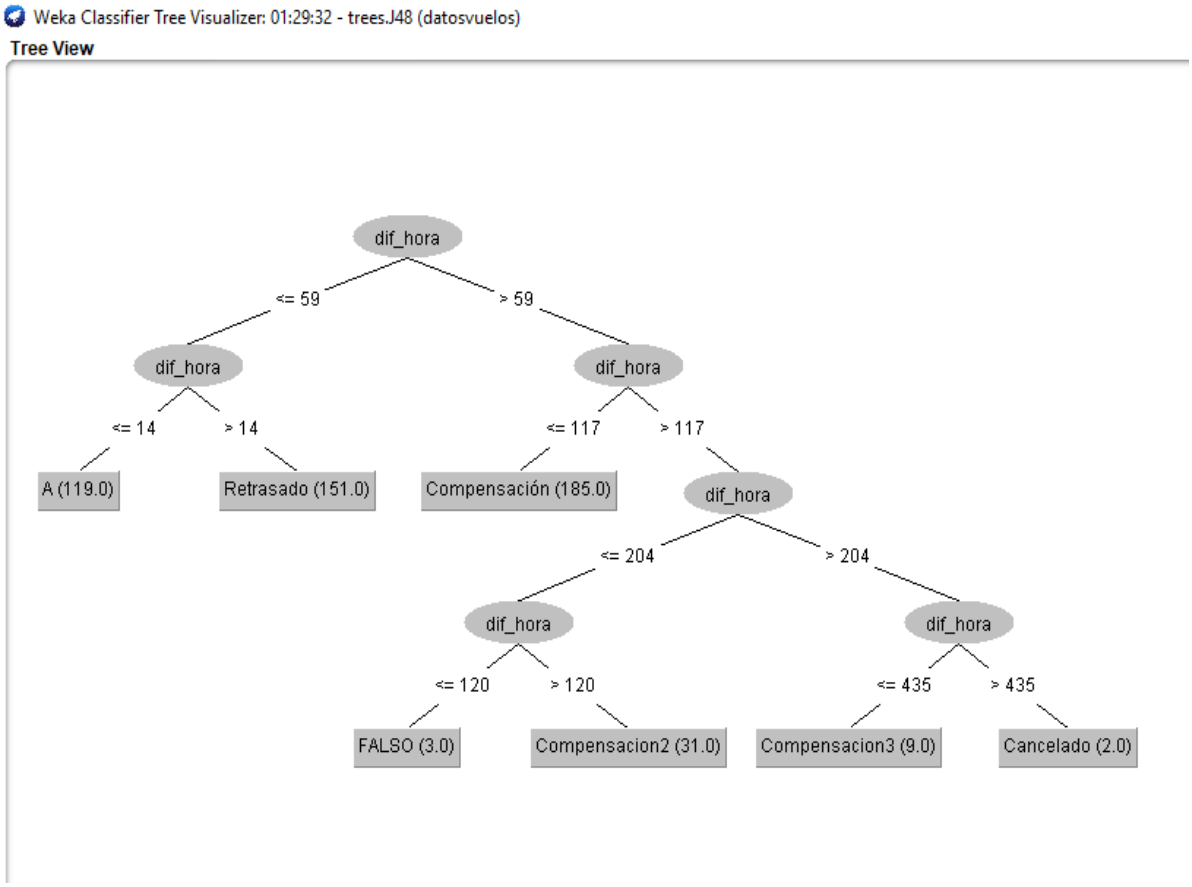


Figura 4.11. *Árbol de decisión*

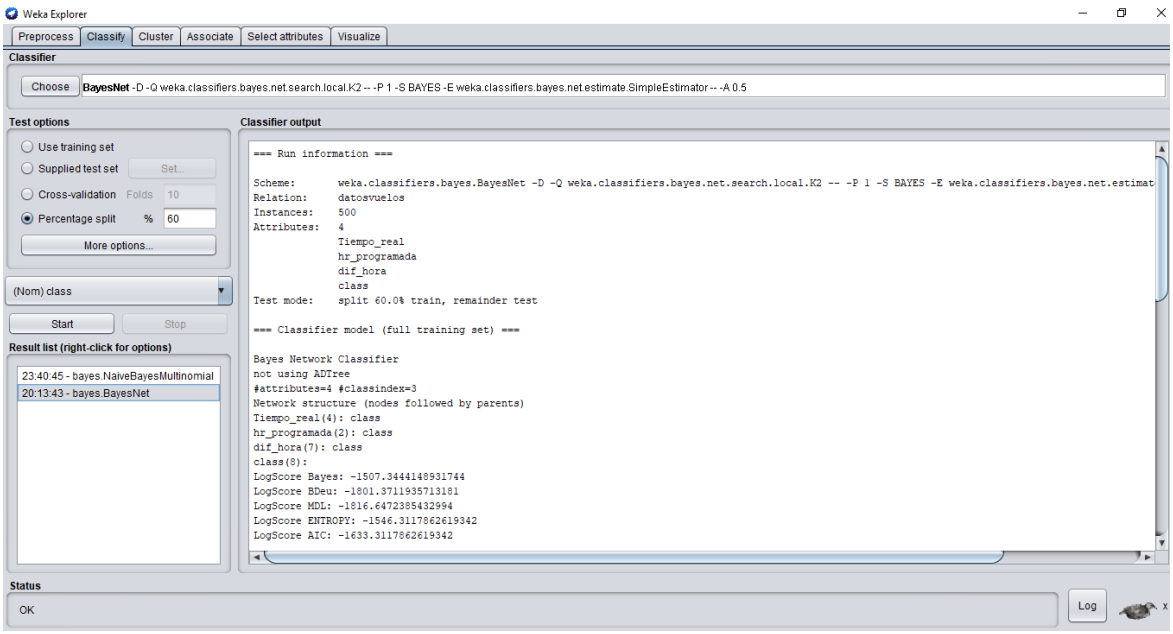


Figura.4.12. *Red bayesianas.*

La ventana siguiente (figura 4.13) muestra la clasificación informando que el 98% de los datos es confiable e indica que el 2% es error de los datos, tardando sólo 5 segundos en procesar la información, se puede comparar que es más eficiente el árbol de decisión que la red bayesiana porque la probababilidad de error es mínima, en la toma de decisión para poder conocer qué tanto puede afectar al momento de hacer la exportación de los productos los retrasos de los vuelos para que no afecte el proceso de logística y tenga secuelas secuandarias.

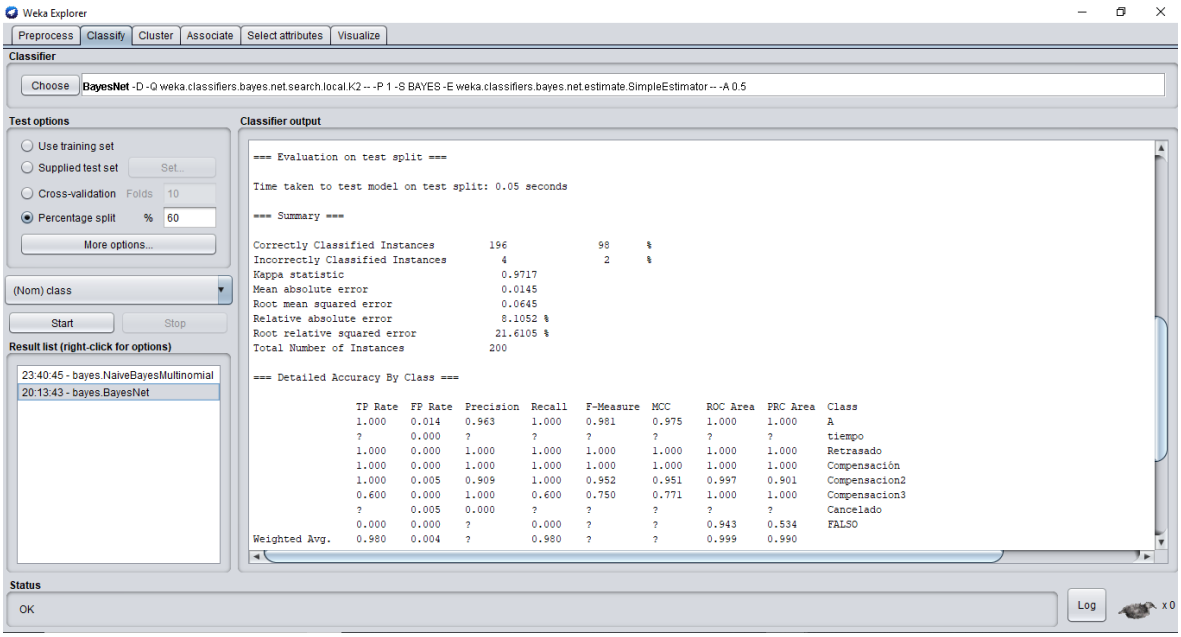


Figura 4.13. *Confiablidad de la Red Bayesiana*

En la figura 4.14 que a continuación se muestra, se visualiza el gráfo de la Red Bayesiana con los parámetros obtenidos de la base de datos donde los tres nodos que conforman la red indican hacia que gráfo se va a dirigir de acuerdo a la clasificación:

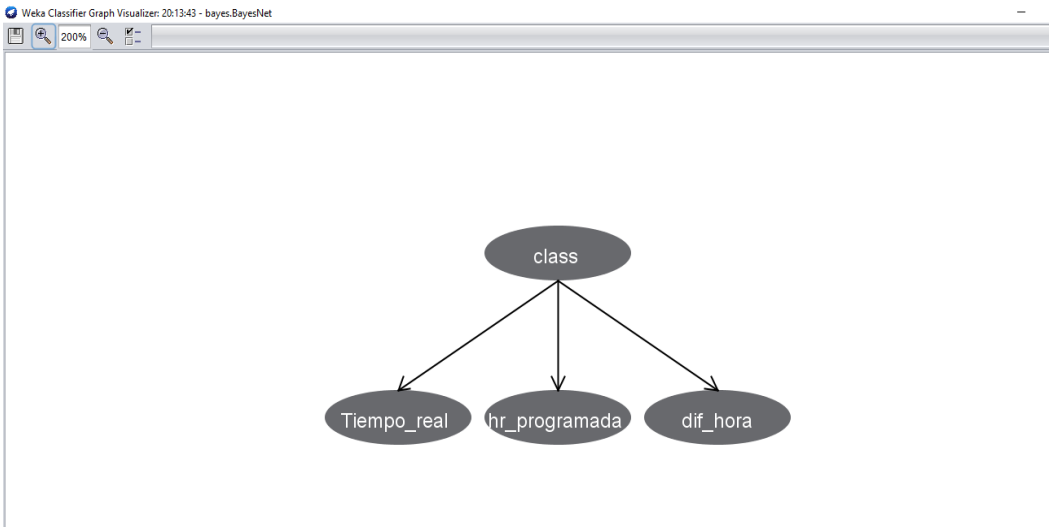


Figura 4.14. *Gráfo Red Bayesiana.*

Conclusión: el gráfo de la red bayesiana no servirá para poder tomar una desición, sólo presenta los atributos que se tomaron en cuenta pero no da estadísticas para conocer el resto de la información como los retrasos de los vuelos.

4.2 Recolección y procesamiento de datos

Como parte de la recolección de los datos se hizo una búsqueda exhaustiva de una base de datos libre que estuviera relacionada con el tema de exportaciones, en la búsqueda se encontró la plataforma de nombre Kaggle, en la cual se encontraron diferentes bases de datos como: Tweets.csv, Utterance-Flights.csv, airports.csv por mencionar algunas, ninguna de estas sirvió para el proyecto ya que no contaban con los registros suficientes para cumplir con las características de Big Data.

Tweets.csv:

Se muestra una parte de la base de datos donde algunos de los encabezados son: tweet_id: es el número de tweets, airline_sentiment: indica si el comentario es positivo, negativo o neutral, airline_sentiment_confidence: la confianza para la aerolínea, negativereason: negativas para la aerolíneas de como estuvo el vuelo, airline: la aerolínea, name: nombre de quién envía el tweets, tex: el comentario del usuario, tweet_created: fecha de cuándo se hizo el tweets, tweet_location: donde tomaron el avión, user_timezone: la zona horaria del usuario. La base

de datos con los tweet sirve para saber cómo estuvo el servicio: positivo, neutral o negativo (tabla 4.1).

tweet_id	airline_senti	airline_senti	negativerea	negativerea	airline	airline_senti	name	negativerea	retweet_cou	text	tweet_coord	tweet_created	tweet_locati	user_timezone
5.7031E+17	neutral	1			Virgin America		cairdin		0	@VirginAmerica What @dhepburn said.		24/02/2015 11:35		Eastern Time (US & Canada)
5.703E+17	positive	0.3486		0	Virgin America		jnardino		0	@VirginAmerica plus you've added commercia		24/02/2015 11:15		Pacific Time (US & Canada)
5.703E+17	neutral	0.6837			Virgin America		yvonnalynn		0	@VirginAmerica I didn't today... Must mean I r		24/02/2015 11:15	Lets Play	Central Time (US & Canada)
5.703E+17	negative	1	Bad Flight	0.7033	Virgin America		jnardino		0	@VirginAmerica it's really aggressive to blast (		24/02/2015 11:15		Pacific Time (US & Canada)
5.703E+17	negative	1	Can't Tell	1	Virgin America		jnardino		0	@VirginAmerica and it's a really big bad thing		24/02/2015 11:14		Pacific Time (US & Canada)
5.703E+17	negative	1	Can't Tell	0.6842	Virgin America		jnardino		0	@VirginAmerica seriously would pay $30 a		24/02/2015 11:14		Pacific Time (US & Canada)
5.703E+17	positive	0.6745		0	Virgin America		cjmcginnis		0	@VirginAmerica yes, nearly every time I fly VX		24/02/2015 11:13	San Francisc	Pacific Time (US & Canada)
5.703E+17	neutral	0.634			Virgin America		pilot		0	@VirginAmerica Really missed a prime opport		24/02/2015 11:12	Los Angeles	Pacific Time (US & Canada)
5.703E+17	positive	0.6559			Virgin America		dhepburn		0	@virginamerica Well, I didn't…but NOW I DC		24/02/2015 11:11	San Diego	Pacific Time (US & Canada)
5.703E+17	positive	1			Virgin America		YupitsTate		0	@VirginAmerica it was amazing, and arrived a		24/02/2015 10:53	Los Angeles	Eastern Time (US & Canada)
5.7029E+17	neutral	0.6769		0	Virgin America		idk_but_youtube		0	@VirginAmerica did you know that suicide is t		24/02/2015 10:48	1/1 loner squ	Eastern Time (US & Canada)
5.7029E+17	positive	1			Virgin America		HyperCamiLax		0	@VirginAmerica I <3 pretty graphics. so muc		24/02/2015 10:30	NYC	America/New_York
5.7029E+17	positive	1			Virgin America		HyperCamiLax		0	@VirginAmerica This is such a great deal! Alre		24/02/2015 10:30	NYC	America/New_York
5.7029E+17	positive	0.6451			Virgin America		mollanderson		0	@VirginAmerica @virginmedia I'm flying your		24/02/2015 10:21		Eastern Time (US & Canada)

Tabla 4.1. *Base de datos Tweets.csv*

Utterance-Flights.csv:

Es otra de las bases de datos que se encontró y tiene datos muy parecidos a la anterior. Los datos que incluye son: unit_id: que es el número de usuario, created_at: el dia que envío el twits, work_id: número de empleado, city: lugar de donde viajaron, por mencinar algunos de los encabezados de la base de datos. En esta base de datos en el campo escenario comparten su experiencia y si volverían a utilizar el servicio, tabla 4.2.

_unit_id	_created_at	_id	_started_at	_tainted	_channel	_trust	_worker_id	_country	_region	_city	_ip	response_1	response_2	response_3	scenario
1403461970	10/11/2017 21:21	2909637246	##########	false	clixsense	1	10422922	PAK	5	Karachi	39.48.57.27	Your service	I won't trave	Your service	You are threatening to never to use this airline again
1403461970	10/11/2017 23:06	2909795516	##########	false	clixsense	1	21764533	HRV	12	Rijeka	78.2.113.83	I will never u	Not only tha	You are so ba	You are threatening to never to use this airline again
1403461970	10/11/2017 22:31	2909748520	##########	false	clixsense	1	37780823	VEN	25	Caracas	190.39.8.228	BECAUSE THI	THEY TREAT	YOUR DELAY	You are threatening to never to use this airline again
1403461970	10/11/2017 21:25	2909643979	##########	false	neodev	1	38372558	BRA	29	Goiania	200.163.232.	If I have to p	Why do you	I'll get in cou	You are threatening to never to use this airline again
1403461970	10/11/2017 22:31	2909748981	##########	false	clixsense	1	39078729	EGY	11	Cairo	41.237.4.122	If you do not	I will book o	I will not boo	You are threatening to never to use this airline again
1403461970	10/11/2017 21:36	2909668685	##########	false	neodev	1	39257488	VEN	25	Caracas	186.167.242..	not only that	they could s	should impr	You are threatening to never to use this airline again
1403461970	10/11/2017 22:24	2909740868	##########	false	clixsense	1	39517779	GRC	13	ThessalonÃ	46.12.194.97	I will never u	I'm done wit	Never again	You are threatening to never to use this airline again
1403461970	10/11/2017 22:53	2909778130	##########	false	clixsense	1	39767721	RUS	73	Kazan	188.162.195..	I this flight d	On this fligh	Give me a co	You are threatening to never to use this airline again
1403461970	10/11/2017 21:58	2909706944	##########	false	neodev	1	40214667	VEN	18	Acarigua	186.91.118.1	No volvere a	No volverÃ€	no recomen	You are threatening to never to use this airline again
1403461970	10/11/2017 21:53	2909696415	##########	false	clixsense	1	40597878	ESP	59	Leioa	85.86.226.14	What other a	Please chang	I hope this n	You are threatening to never to use this airline again
1403461970	10/11/2017 21:15	2909629212	##########	false	neodev	1	40956186	EGY			62.114.130.5	there are an	there are an	there are an	You are threatening to never to use this airline again
1403461970	10/11/2017 21:31	2909657203	##########	false	neodev	1	41034016	VEN	25	Caracas	186.167.250.	This is a reall	I am so late (	Do you reco	You are threatening to never to use this airline again
1403461970	10/11/2017 22:01	2909712022	##########	false	neodev	1	41429888	VEN	23	Cabimas	201.209.201.	Yes, because	no	no	You are threatening to never to use this airline again

Tabla 4.2*. Base de datos Utterance-Flights.csv*

Airports:

Esta base de datos a diferencia de las anteriores trae datos muy técnicos por asi llamarlos, habla de iata_code: son las siglas del aeropuerto, Airport: es el nombre del aeropuerto, City:

Ciudad donde se tomó el vuelo, Country: el país donde se tomó el vuelo, por mencionar algunos de los encabezados (tabla 4.3).

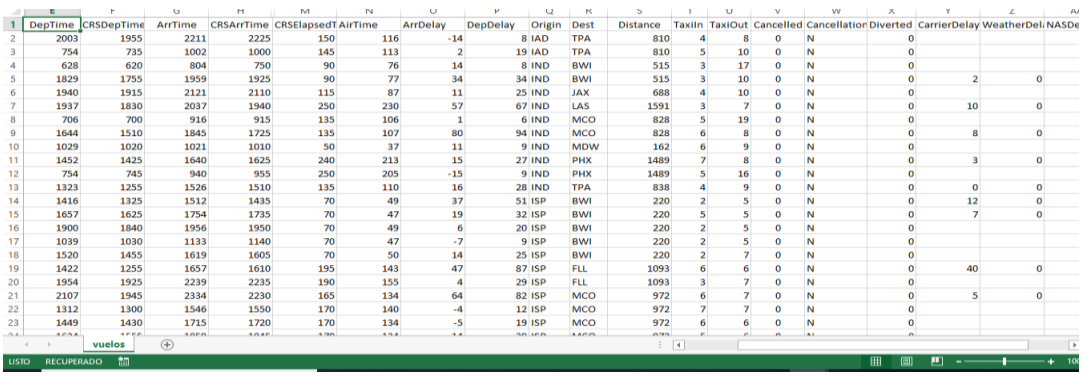
IATA_CODE	AIRPORT	CITY	STATE	COUNTRY	LATITUDE	LONGITUDE
ABE	Lehigh Valley International Air	Allentown	PA	USA	40.65236	-75.4404
ABI	Abilene Regional Airport	Abilene	TX	USA	32.41132	-99.6819
ABQ	Albuquerque International Sur	Albuquerque	NM	USA	35.04022	-106.60919
ABR	Aberdeen Regional Airport	Aberdeen	SD	USA	45.44906	-98.42183
ABY	Southwest Georgia Regional A	Albany	GA	USA	31.53552	-84.19447
ACK	Nantucket Memorial Airport	Nantucket	MA	USA	41.25305	-70.06018
ACT	Waco Regional Airport	Waco	TX	USA	31.61129	-97.23052
ACV	Arcata Airport	Arcata/Eurek	CA	USA	40.97812	-124.10862
ACY	Atlantic City International Airp	Atlantic City	NJ	USA	39.45758	-74.57717
ADK	Adak Airport	Adak	AK	USA	51.87796	-176.64603
ADQ	Kodiak Airport	Kodiak	AK	USA	57.74997	-152.49386
AEX	Alexandria International Airpo	Alexandria	LA	USA	31.32737	-92.54856
AGS	Augusta Regional AirportÂ (Bu	Augusta	GA	USA	33.36996	-81.9645
AKN	King Salmon Airport	King Salmon	AK	USA	58.6768	-156.64922
ALB	Albany International Airport	Albany	NY	USA	42.74812	-73.80298

Tabla 4.3. *Base de datos Airports.csv*

Las bases de datos antes mencionadas no cumplían con los datos para poder comprobar la hipótesis, dos de ellas hablaban de twits para calificar el servicio de la aerolínea, que necesitaban encontrar un vuelo de urgencia, entre otros comentarios, la otra base de datos muestra datos como el aeropuerto y donde tomaron el vuelo, con esto se comprueba que no tienen relación con lo que se está necesitando para ver qué tanto afectan los retrasos para entregar la mercancía exportada, por eso se decidió trabajar con la base de datos DelayedFligths ya que contiene los datos necesarios para comprobar la hipótesis.

DelayedFligths:

Es la base de datos que se decidió trabajar por la información que contiene y cumple los datos necesarios realizar la analítica y comprobar la hipótesis antes mencionada. A continuación se muestra imagen de la base de datos en la figura 4.15



DepTime	CRSDepTime	ArrTime	CRSArrTime	CRSElapsedT	AirTime	ArrDelay	DepDelay	Origin	Dest	Distance	TaxiIn	TaxiOut	Cancelled	Cancellation	Diverted	CarrierDelay	WeatherDel	NASDe
2003	1955	2211	2225	150	116	-14	8	IAD	TPA	810	4	8	0	N	0			
754	735	1002	1000	145	113	2	19	IAD	TPA	810	5	10	0	N	0			
628	620	804	750	90	76	14	8	IND	BWI	515	3	17	0	N	0			
1829	1755	1959	1925	90	77	34	34	IND	BWI	515	3	10	0	N	0	2	0	
1940	1915	2121	2110	115	87	11	25	IND	JAX	688	4	10	0	N	0			
1937	1830	2037	1940	250	230	57	67	IND	LAS	1591	3	7	0	N	0	10	0	
706	700	916	915	135	106	1	6	IND	MCO	828	5	19	0	N	0			
1644	1510	1845	1725	135	107	80	94	IND	MCO	828	6	8	0	N	0	8	0	
1029	1020	1021	1010	50	37	11	9	IND	MDW	162	6	9	0	N	0			
1452	1425	1640	1625	240	213	15	27	IND	PHX	1489	7	8	0	N	0	3	0	
754	745	940	955	250	205	-15	9	IND	PHX	1489	5	16	0	N	0			
1323	1255	1526	1510	135	110	16	28	IND	TPA	838	4	9	0	N	0	0	0	
1416	1325	1512	1435	70	49	37	51	ISP	BWI	220	2	5	0	N	0	12	0	
1657	1625	1754	1735	70	47	19	32	ISP	BWI	220	5	5	0	N	0	7	0	
1900	1840	1956	1950	70	49	6	20	ISP	BWI	220	2	5	0	N	0			
1039	1030	1133	1140	70	47	-7	9	ISP	BWI	220	2	5	0	N	0			
1520	1455	1619	1605	70	50	14	25	ISP	BWI	220	2	7	0	N	0			
1422	1255	1657	1610	195	143	47	87	ISP	FLL	1093	6	6	0	N	0	40	0	
1954	1925	2239	2235	190	155	4	29	ISP	FLL	1093	3	7	0	N	0			
2107	1945	2334	2230	165	134	64	82	ISP	MCO	972	6	7	0	N	0	5	0	
1312	1300	1546	1550	170	140	-4	12	ISP	MCO	972	7	7	0	N	0			
1449	1430	1715	1720	170	134	-5	19	ISP	MCO	972	6	6	0	N	0			

Figura 4.15. *Base de datos seleccionada*

En la tabla 4.4 se describen los nombres y la descripción de los campos que contiene la base de datos.

No.	Name	Descripción
1	Year	Año de registro.
2	Month	Mes.
3	DayofMonth	Día del mes.
4	DayOfWeek	1 (lunes) - 7 (domingo).
5	DepTime	tiempo de salida real (local, hhmm).
6	CRSDepTime	hora de salida programada (local, hhmm).
7	ArrTime	hora de llegada real (local, hhmm).
8	CRSArrTime	hora de llegada programada (local, hhmm).
9	UniqueCarrier	Código único de transportista.
10	FlightNum	Número de vuelo.
11	TailNum	registro de la aeronave, identificador único de la aeronave.
12	ActualElapsedTime	Tiempo actual transcurrido en minutos.
13	CRSElapsedTime	Tiempo transcurrido en minutos.
14	AirTime	Tiempo en antena.
15	ArrDelay	Tiempo del vuelo.
16	DepDelay	Demora del vuelo.
17	Origin	Origen.
18	Dest	Destino.
19	Distance	Distancia en millas.
20	TaxiIn	Tiempo de entrada.
21	TaxiOut	Tiempo de salida.
22	Cancelled	Cancelado.
23	CancellationCode	motivo de cancelación (A = transportista, B = clima, C = NAS, D = seguridad).
24	Diverted	Desviado 1 = sí, 0 = no.
25	CarrierDelay	Retraso del operador.
26	WeatherDelay	Retraso meteorológico
27	NASDelay	Retraso que está dentro del control del Sistema Nacional de Espacio Aéreo (NAS).
28	SecurityDelay	Retraso de seguridad.
29	LateAircraftDelay	Retraso de llegada al aeropuerto.

Tabla4.4. *Descripción de los campos de la base de datos.*

Las bases de datos de Kaggle son proporcionadas por diferentes empresas con la finalidad de realizar competencias para que la comunidad científica aporte soluciones y predicciones con los datos que se proporcionan. La base de datos DelayedFligths cuenta con 14 diferentes análisis, donde la mayoría ha utilizado los datos para determinar cancelaciones de los vuelos,utilizando Python. A diferencia del analísis que se hizo para este proyecto, se utilizó el lenguaje de programación R, los registros de hora de salida programada, hora real de salida, hora de llegada programada, hora de llegada real, sirvieron para conocer que tanto afecta los retrasos de los vuelos para hacer el envío de la mercancía a exportar.

A continuación se muestran un resumen de tres analistas sobresalientes que publicaron sus resultados en la plataforma de Kaggle (Kaggle, 2018).

De acuerdo al análisis de datos que realizó el programador con el pseudónimo Dan, el clima es la razón más común para la cancelación de un vuelo, y no hay casos de cancelación relacionada con la seguridad en el conjunto de datos. La mayoría de las cancelaciones son en noviembre y diciembre. Teniendo en cuenta el momento de las cancelaciones, parece que algunas cancelaciones relacionadas con el operador. A pesar de que hay sustancialmente más retrasos climáticos en diciembre, la demora promedio fue mucho más larga en junio. Cinco de los seis meses que tienen los retrasos promedio más largos no suelen ver nieve en los EE. UU., lo que podría significar que las demoras causadas por las condiciones de nieve no duran tanto. El monto de las cancelaciones en noviembre y diciembre también podría sugerir que cuando las condiciones climáticas se vuelvan demasiado pobres en el invierno, los vuelos se cancelan en lugar de esperar a que mejoren las condiciones climáticas, lo que podría disminuir el tiempo promedio de demora en esos meses en la imagen 4.16 se muestra la gráfica de como se publicó la información ("Exploratory Analysis of Flight Cancellations. | Kaggle", 2018).

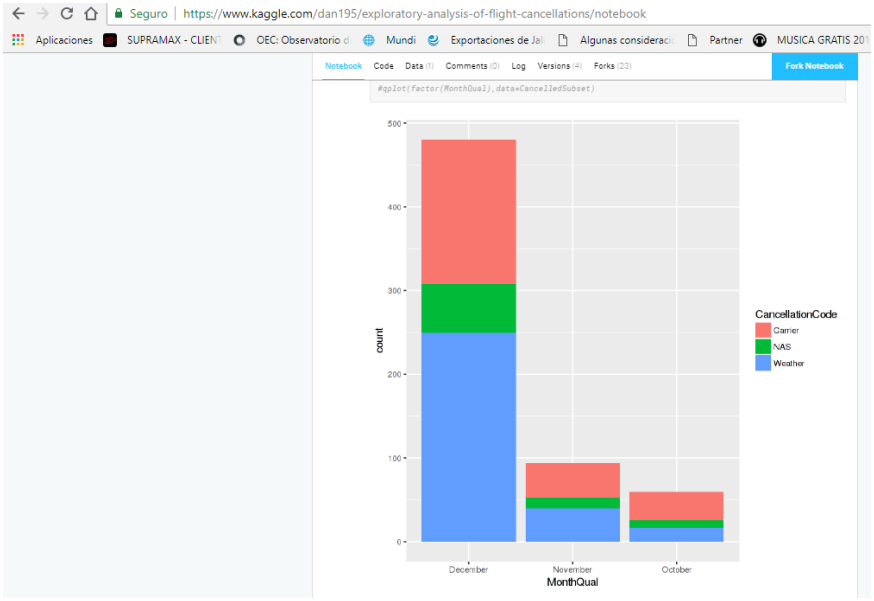


Figura 4.16 *Cancelaciones por el clima.*

Otros de los analistas de nombre Adrián Vera, llegó a la siguiente conclusión: las demoras se centraron en febrero, junio y diciembre, con un pico en las demoras promedio en julio de 2008.

Con respecto a la hora del día en que se programó el despegue, se puede ver en la parte superior del diagrama de dispersión cómo los retrasos se concentran en una hora. A medida que avanza el día, hay más y más retrasos, pero como se muestra en el centro del diagrama de dispersión(figura 4.17), los vuelos retrasados se dividen en dos grupos: uno con retrasos más largos y otro con más cortos. Una posible interpretación es que los retrasos generados por los vuelos anteriores aumentan o disminuyen en cada uno de los siguientes viajes ("Flight Delay EDA (Exploratory Data Analysis) | Kaggle", 2018).

```
k = 10 #number of variables for heatmap
f, ax = plt.subplots(figsize=(12, 9))
cols = delcorrmat.nlargest(k, 'ArrDelay')['ArrDelay'].index
cm = np.corrcoef(Delayedflights[cols].values.T)
sns.set(font_scale=1.25)
hm = sns.heatmap(cm, cbar=True, annot=True, square=True, fmt='.2f', annot_kws={'size': 15}, yti
cklabels=cols.values, xticklabels=cols.values)
plt.show()
```

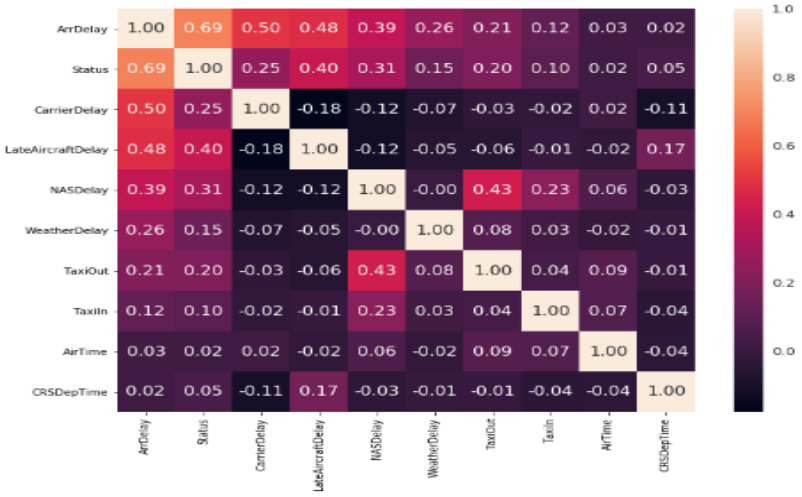


Figura 4.17. *Demoras de los vuelos*

El científico de datos Jared Brook comenta que las predicciones de cancelación no son factibles con los datos proporcionados. Combinar esto con el clima actual y los datos de mantenimiento de aeronaves / aeropuertos podría hacer que las predicciones de cancelación sean algo útiles. La predicción de los tiempos de demora del vuelo es igualmente difícil y también podría mejorarse con datos relevantes en tiempo real. A pesar de que este conjunto de datos no es particularmente útil para las predicciones, no obstante, se pueden mostrar relaciones y distribuciones interesantes. Agrupó ambos dataframes por DayOfWeek, y calculó el porcentaje de vuelos cancelados para cada día de la semana. Esto muestra que el retraso promedio del vuelo está correlacionado positivamente con el tiempo promedio de rodaje, y que el porcentaje de vuelos cancelados se correlaciona negativamente con los tiempos promedio de vuelo, cuando los promedios se calculan entre los operadores. Sin embargo, ninguna de estas correlaciones es particularmente fuerte. En la figura 4.18 se muestra que día de la semana es cuando ocurren más cancelaciones de los vuelos ("Airlines Delay and Cancellation Analysis | Kaggle", 2018).

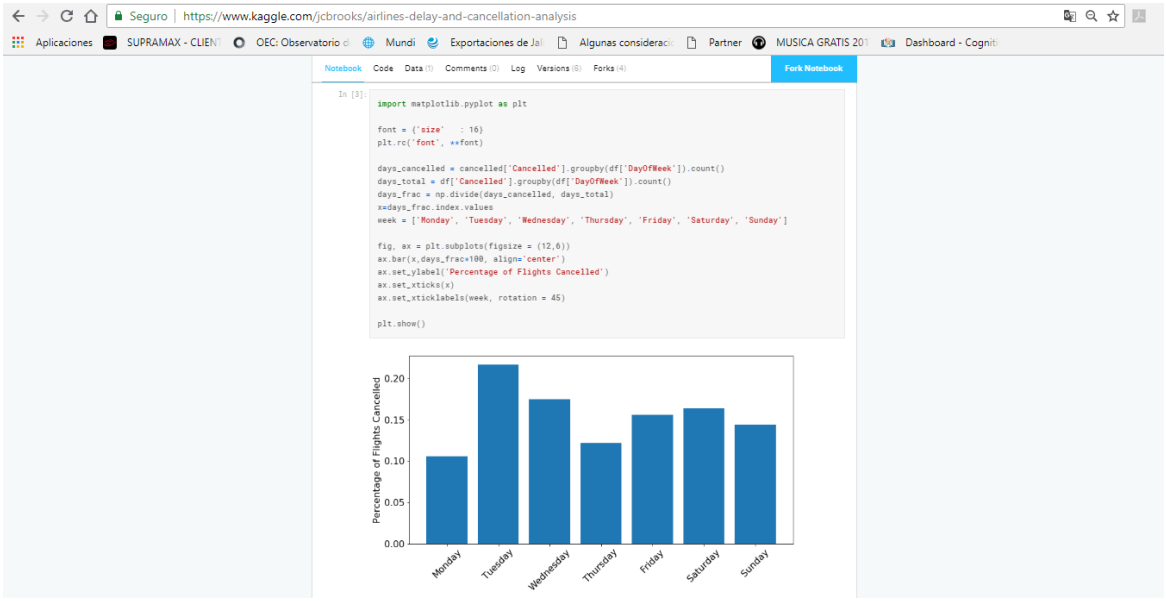


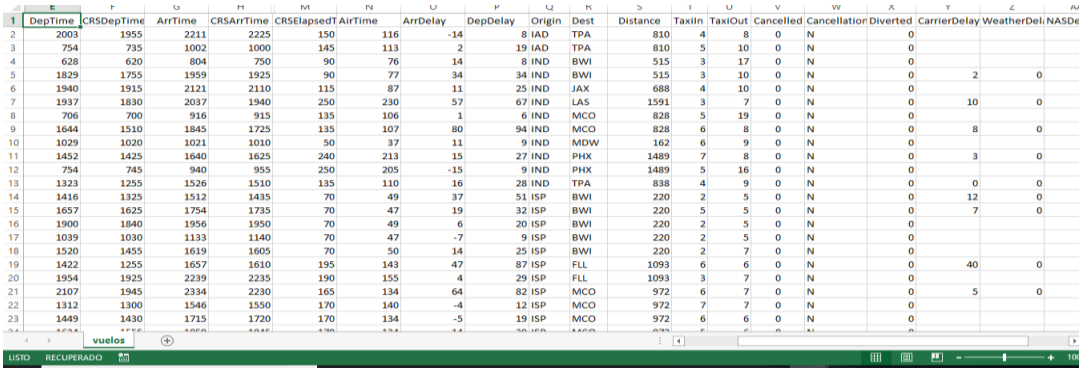
Figura 4.18. *Cancelaciones por día*

Como se puede observar con un fragmento de las conclusiones que llegaron algunos de los analistas que trabajaron con la base de datos, ellos se centraron en cancelaciones y retrasos, pero no verificaron si afectaba la llegada a tiempo de los vuelos, lo que abrió una oprtunidad de análisis para el presente trabajo. Se ha llegando a la conclusión que teniendo la misma base de datos cada uno de los científicos de datos puede analizar los mismos campos y darle un enfoque diferente a la información que contiene la base de datos, sin que ningún trabajo se vea duplicado. Para este proyecto se requiere saber qué tanto afectan los retrasos, aunque es de gran utilidad toda la información presentada de los diferentes científicos para conocer cuales son los factores para que se cancele un vuelo, que día es cuando más cancelaciones se tienen y los horarios, todas esas inconstantes sirve para tomar en cuenta los retrasos, lo que realmente importa para poder llevar una exportación son los tiempos de retraso para poder entregar la mercancía en tiempo.

4.3 Resultados obtenidos

En esta sección se va a presentar de donde se obtuvo la Base de Datos y de qué manera se procesó en el lenguaje R para poder obtener el resultado al cual se llegó.

En la figura 4.18 siguiente se presenta un parte de la base de datos con extensión .csv, cabe mencionar que la base cuenta con 1´048,576 de datos, en ella se presentan los encabezados de las columnas, la cual se decidió trabajar con 4 de ellas que son las siguientes: DepTime, CRSTime, ArrTime, CRSArrTime.



DepTime	CRSDepTime	ArrTime	CRSArrTime	CRSElapsedT	AirTime	ArrDelay	DepDelay	Origin	Dest	Distance	TaxiIn	TaxiOut	Cancelled	Cancellation	Diverted	CarrierDelay	WeatherDeli	NASDe
2003	1955	2211	2225	150	116	-14	8	IAD	TPA	810	4	8	0	N	0			
754	735	1002	1000	145	113	2	19	IAD	TPA	810	5	10	0	N	0			
628	620	804	750	90	76	14	8	IND	BWI	515	3	17	0	N	0			
1829	1755	1959	1925	90	77	34	34	IND	BWI	515	3	10	0	N	0	2	0	
1940	1915	2121	2110	115	87	11	25	IND	JAX	688	4	10	0	N	0			
1937	1830	2037	1940	250	230	57	67	IND	LAS	1591	3	7	0	N	0	10	0	
706	700	916	915	135	106	1	6	IND	MCO	828	5	19	0	N	0			
1644	1510	1845	1725	135	107	80	94	IND	MCO	828	6	8	0	N	0	8	0	
1029	1020	1021	1010	50	37	11	9	IND	MDW	162	6	9	0	N	0			
1452	1425	1640	1625	240	213	15	27	IND	PHX	1489	7	8	0	N	0	3	0	
754	745	940	955	250	205	-15	9	IND	PHX	1489	5	16	0	N	0			
1323	1255	1526	1510	135	110	16	28	IND	TPA	838	4	9	0	N	0	0	0	
1416	1325	1512	1435	70	49	37	51	ISP	BWI	220	2	5	0	N	0	12	0	
1657	1625	1754	1735	70	47	19	32	ISP	BWI	220	5	5	0	N	0	7	0	
1900	1840	1956	1950	70	49	6	20	ISP	BWI	220	2	5	0	N	0			
1039	1030	1133	1140	70	47	-7	9	ISP	BWI	220	2	5	0	N	0			
1520	1455	1619	1605	70	50	14	25	ISP	BWI	220	2	7	0	N	0			
1422	1255	1657	1610	195	143	47	87	ISP	FLL	1093	6	6	0	N	0	40	0	
1954	1925	2239	2235	190	155	4	29	ISP	FLL	1093	3	7	0	N	0			
2107	1945	2334	2230	165	134	64	82	ISP	MCO	972	6	7	0	N	0	5	0	
1312	1300	1546	1550	170	140	-4	12	ISP	MCO	972	7	7	0	N	0			
1449	1430	1715	1720	170	134	-5	19	ISP	MCO	972	6	6	0	N	0			

Figura 4.19. *Base de datos DelayedFligts.cvs*

Para procesar los datos, se utilizó el lenguaje de programación R para poder visualizar la información de una manera más fácil, a continuación en las diferentes ventanas se va a explicar el código para ir procesando la información. Las librerías utilizadas y sus funciones son:

- tree y rpart: contiene funciones para la creación y representación de árboles de regresión y clasificación.
- rpart.plot: permite crear representaciones detalladas de modelos creados con rpart.
- C50: contiene los algoritmos *C5.0* para árboles de clasificación.

En la figura 4.20 el primer paso que se debe hacer es cargar las librerías que se van a utilizar. **readr** es para poder leer los datos, **rpart** es para particionamiento recursivo y **rpart.plot** ayudará a graficar el árbol de decisión, la base de datos se asigna a la variable **vueretraso** para poderla estar llamando cuando se requiera.

La variable **vueretraso** se encarga de leer toda la base de datos, para visualizar que se haya cargado bien se asigna a **vue** todo lo que esté en **vueretraso**, en la figura 4.19 muestran algunos datos y en la parte inferior dice los campos que faltaron por mostrar.

```
> library("readr")
> library(rpart)
> library(rpart.plot) # Graficar ??rbol de decisi??n
> library(C50)
>
> vueretraso <- read_csv(file.choose('C:/Users/blanc/Documents/PROYECTO MAESTR?A/Proyecto MCP/prueba6.csv'),skip = 0,col_nar
es = TRUE)
Parsed with column specification:
cols(
  .default = col_integer(),
  UniqueCarrier = col_character(),
  TailNum = col_character(),
  Origin = col_character(),
  Dest = col_character(),
  CancellationCode = col_character(),
  CarrierDelay = col_double(),
  WeatherDelay = col_double(),
  NASDelay = col_double(),
  SecurityDelay = col_double(),
  LateAircraftDelay = col_double()
)
See spec(...) for full column specifications.
>
> vue<-(vueretraso)
> vue
# A tibble: 364 x 29
    Year Month DayofMonth DayOfWeek DepTime CRSDepTime ArrTime CRSArrTime UniqueCarrier FlightNum TailNum
   <int> <int>      <int>     <int>   <int>      <int>   <int>      <int>         <chr>     <int>   <chr>
 1  2008     1          3         4    2003       1955    2211       2225            WN       335  N712SW
 2  2008     1          3         4     754        735    1002       1000            WN      3231  N772SW
 3  2008     1          3         4     628        620     804        750            WN       448  N428WN
 4  2008     1          3         4    1829       1755    1959       1925            WN      3920  N464WN
 5  2008     1          3         4    1940       1915    2121       2110            WN       378  N726SW
 6  2008     1          3         4    1937       1830    2037       1940            WN       509  N763SW
 7  2008     1          3         4     706        700     916        915            WN       100  N690SW
 8  2008     1          3         4    1644       1510    1845       1725            WN      1333  N334SW
 9  2008     1          3         4    1029       1020    1021       1010            WN      2272  N263WN
10  2008     1          3         4    1452       1425    1640       1625            WN       675  N286WN
# ... with 354 more rows, and 18 more variables: ActualElapsedTime <int>, CRSElapsedTime <int>, AirTime <int>,
#   ArrDelay <int>, DepDelay <int>, Origin <chr>, Dest <chr>, Distance <int>, TaxiIn <int>, TaxiOut <int>,
#   Cancelled <int>, CancellationCode <chr>, Diverted <int>, CarrierDelay <dbl>, WeatherDelay <dbl>, NASDelay <dbl>,
#   SecurityDelay <dbl>, LateAircraftDelay <dbl>
> data(list=vueretraso)
There were 50 or more warnings (use warnings() to see the first 50)
> ind        <- sample(2,nrow(vueretraso), replace=TRUE, prob=c(0.6, 0.4)) # 60% entrenamiento, 40 % test<
> vuelosTrain  <- vueretraso[ind==1, ] # Entrenamiento
```

Figura 4.20. *Programación código R*

El proceso de construcción de árboles descrito en las secciones anteriores tiende a reducir rápidamente el error de entrenamiento, es decir, el modelo se ajusta muy bien a las observaciones empleadas como entrenamiento. Como consecuencia, se genera un *overfitting* que reduce su capacidad predictiva al aplicarlo a nuevos datos. La razón de este comportamiento radica en la facilidad con la que los árboles se ramifican adquiriendo estructuras complejas. De hecho, si no se limitan las divisiones, todo árbol termina ajustándose perfectamente a las observaciones de entrenamiento creando un nodo terminal por observación. Existen dos estrategias para prevenir el problema de *overfitting* de los árboles: limitar el tamaño del árbol y el proceso de podado (*pruning*).

Limitar el tamaño del árbol: el tamaño final que adquiere un árbol puede controlarse mediante reglas de parada que detengan la división de los nodos dependiendo de si se cumplen o no determinadas condiciones.

Pruning: la estrategia de controlar el tamaño del árbol mediante reglas de parada tiene un inconveniente, el árbol se crece seleccionando la mejor división en cada momento hasta alcanzar una condición de parada.

La información antes mencionada sirve para conocer el overfitting que se produce en el árbol que son los datos. En la figura 4.20 muestra las clases: ArrDelay, CRSArrtime, DepDelay, DepTime que se van a tomar para realizar la analítica y posteriormente realizar el árbol de decisión, donde n= No. de observaciones, Split= número de cortes, xerror= árbol deja descender su error. Después de la información presentada, se programa el árbol de decisión con la función plotcp (ArbolRpart) y se imprime con printcp (pArbolRpart). Con el corte que se realiza es para que la mayor parte de la información sea clasificada.

```
Variables actually used in tree construction:
[1] ArrDelay   CRSArrTime DepDelay   DepTime

Root node error: 194/299 = 0.64883

n= 299

        CP nsplit rel error  xerror     xstd
1 0.412371      0   1.00000 1.00000 0.042546
2 0.201031      1   0.58763 0.58763 0.043291
3 0.061856      2   0.38660 0.40206 0.039139
4 0.036082      3   0.32474 0.35052 0.037361
5 0.016753      4   0.28866 0.32990 0.036558
6 0.010000      8   0.22165 0.34536 0.037166
> plotcp(ArbolRpart)
> pArbolRpart<- prune(ArbolRpart, cp= ArbolRpart$cptable[which.min(ArbolRpart$cptable[,"xerror"]),"CP"])
> pArbolRpart<- prune(ArbolRpart, cp= 0.016753)
> printcp(pArbolRpart)

Classification tree:
rpart(formula = tipoRetraso ~ ., data = vuelosTrain, method = "class")

Variables actually used in tree construction:
[1] ArrDelay DepDelay

Root node error: 194/299 = 0.64883

n= 299

        CP nsplit rel error  xerror     xstd
1 0.412371      0   1.00000 1.00000 0.042546
2 0.201031      1   0.58763 0.58763 0.043291
3 0.061856      2   0.38660 0.40206 0.039139
4 0.036082      3   0.32474 0.35052 0.037361
5 0.016753      4   0.28866 0.32990 0.036558
>
```

Figura 4.21. *Código para realizar el árbol de decisión*

En esta parte del código se imprime la tabla con la información de dónde se van a ha hacer los cortes para los nodos, donde **n** es el No. de observaciones, **nsplit** es el número de cortes que va a tener, con esto se obtiene como resultado que el 64% de los vuelos clasificados tienen retrasos.

En la figura 4.21 se muestra el informe de los vuelos a tiempo un total de 121,751, anticipados 281, cancelados 0, y las diferentes tipos de compensaciones que son los vuelos retrasados desde una hora hasta cuatro horas con un total de 19,504.

Por último se utiliza la función **Sum** para conocer la efectividad de aciertos del modelo aplicado a la base de datos, lo cual demuestra que el 80% del resultado es confiable, considerando un valor confiable para tomar la decisión y conocimiento que la mayoría de los vuelos tendrán algún tipo de retraso.

```
> testPredRpart <- predict(ArbolRpart, newdata = vuelosTest, type = "class")
> table(testPredRpart, vuelosTest$tipoRetraso)

testPredRpart    A tiempo Anticipado cancelacion compensacion compensacion 2 compensacion 3 retrasado
  A tiempo         121751        281           0            0              0              0      19504
  Anticipado            0          0           0            0              0              0          0
  cancelacion           0       1452       14536            0              1            692          0
  compensacion          0        602           0        86719           4021              0      30169
  compensacion 2        0        186           0            0          18474              0          0
  compensacion 3        0        506        2812            0           2554          19121          0
  retrasado             0        307           0        19989              0              0      75990
> sum(testPredRpart == vuelosTest$tipoRetraso) / length(vuelosTest$tipoRetraso)*100
[1] 80.2043
>
```

Figura.4.22. *Tabla de información*

En la 4.22 figura se muestra el árbol de decisión de una forma gráfica para que sea más fácil entender toda la información. A continuación se va explicar un poco la información de que es lo que contiene cada nodo: en el nodo raíz dice que el 29% de los vuelos salieron a tiempo, el 1% fue anticipado, el 4% cancelados, el 25 % obtuvieron compesación tipo 1, el 6% compensación de tipo 2, 5% compensación tipo 3, 30% compensación tipo 3, asi es como se conforma el 100% de los datos procesados.

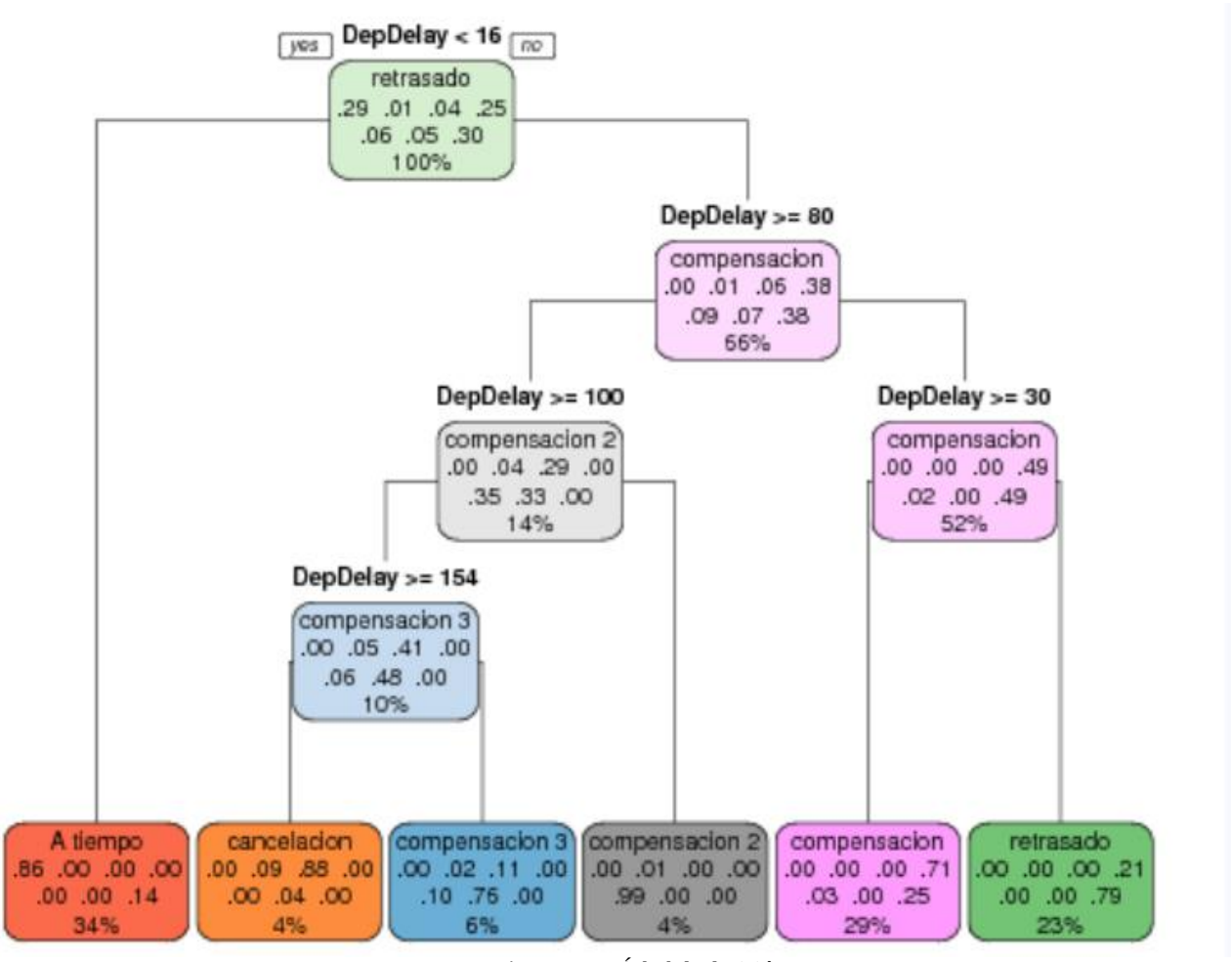


Figura 4.23 *Árbol de decisión*

Con las pruebas que se hicieron al comparar el resultado del árbol de desición con el programa Weka dio un mayor porcentaje de resultado de 99.2 % de los casos se clasificaron bien y con el lenguaje de programación R se obtuvo el 80%, la diferencia es por el *Overfitting* que se produce de acuerdo al algoritmo C4.5 que se utiliza en Weka este ayuda a que el mayor parte de la información sea clasificada.

4.4 Conclusiones y Recomendaciones

4.4.1 Conclusiones

México es un país lleno de riqueza, tanto natural como cultural, debido a esto, otros países están interesados en adquirir los productos que se elaboran en cada región del país. Las exportaciones mexicanas cerraron el 2017 con 409 mil 494 millones de dólares, lo que representa un monto récord, según cifras del Instituto Nacional de Estadística y Geografía y del Banco de México (Blanco, 2018). Las ventas al exterior de las fronteras nacionales vieron un aumento de 9.5 por ciento, el mayor incremento registrado desde 2011 en el rubro.

El presente proyecto desató una investigación sobre las exportaciones y la logística que implica, desde el empresario hasta el agente aduanal, pasando por el booking y la reserva de contenedores con características muy específicas, de acuerdo a la mercancía que se va a exportar.

Dentro de los diferentes medios de transporte que se utilizan, se decidió realizar la analítica con una base de datos de los Estados Unidos de vuelos registrados del año 1998 al 2008, para el proyecto resaltan los registros de la llegada y salida de los vuelos y que por diferentes circunstancias los vuelos no salen o llegan en el tiempo programado por lo tanto se indagó sobre cuánto es el tiempo que se tiene permitido que un avión no despegue o llegue a la hora programada y así conocer los intervalos de tolerancia que hay entre el despegue de los vuelos y sus llegadas.

Se tiene conocimiento que el tiempo de retraso va desde 15 minutos hasta 4 horas, ese tiempo se tendrá que considerar cuando se exporte la mercancía para que no afecten los procesos de logística para la entrega de los productos.

Como resultado de la analítica se observó que a través del árbol de decisión que el 86% de los vuelos salen a tiempo, lo que indica que es un buen porcentaje para llevar a cabo el proceso de logística, ya que no afectará al momento de hacer el proceso de exportación tomando en cuenta que los retrasos que se pueden presentar ya que existe el 14% de

posibilidades que esto suceda teniendo en cuenta cuáles son los tiempos de espera. Con estos resultados se comprueba la hipótesis planteada al inicio del proyecto e inclusive se observa que el 29% de los vuelos salieron a tiempo, el 1% fue anticipado, el 4% cancelados y sorprendentemente el 66% de los vuelos salen con retraso.

4.4.2 Recomendaciones

Probar los algoritmos con una base de datos mexicana que es de donde saldrán las exportaciones del software realizado.

Falta subir los resultados obtenidos a la plataforma de Kaggle para ser parte de la competencia, no se ha hecho porque primero hay que traducir los resultados, ya que toda la información que se publica ahí debe de ir en Inglés.

Fuentes consultadas.

Aguilar, L. J. (2013). Big Data. En L. J. Aguilar, *Big Data* (pág. 400). Marcombo.

Airlines Delay | Kaggle. (2018). Recuperado el 07 de febrero de 2018, de https://www.kaggle.com/giovamata/airlinedelaycauses

Airlines Delay and Cancellation Analysis | Kaggle. (2018). Recuperado el 07 de febrero de 2017 https://www.kaggle.com/jcbrooks/airlines-delay-and-cancellation-analysis

BIG Data SAC. (2013). Recuperado el 01 de Mayo de 2018, de http://www.bigdata.pe/web/index.php/metodología

BASES DE DATOS (@basesdedatos) on Twitter. (2018). Recuperado el 11 de Marzo de 2017, de https://twitter.com/basesdedatos

Data Tons by easy admin. (2018). *Data Tons*. Obtenido de Data Tons: https://blog.datatons.com/2016/04/08/que-es-lenguaje-programacion-r/

Data cleansing y sus fases: contra los problemas de calidad de datos. (2018). Recuperado el 07 de febrero de 2017, de https://blog.es.logicalis.com/analytics/data-cleansing-y-sus-fases-contra-los-problemas-de-calidad-de-datos

Exploratory Analysis of Flight Cancellations. | Kaggle. (2018). Recuperado el 07 de febrero de 2017, de https://www.kaggle.com/dan195/exploratory-analysis-of-flight-cancellations

Fallas., L. C. (2013). *Minerìa de Datos*. Obtenido de Minería de Datos: http://cor-mineriadedatos.blogspot.com/2011/06/weka.html

Gob.mx. (s.f.). *Datos.gob.mx*. Recuperado el 13 de Marzo de 2018, de https://www.gob.mx/promexico/prensa/mexico-alcanza-cifra-record-en-exportaciones-en-2017-144820, 2018.

Gracia, L. (2018). *Un poco de Java Y +*. Obtenido de Un poco de Java Y +: https://unpocodejava.com/2015/12/23/que-es-kaggle/

gurucargo.com. (2018). Recuperado el 30 de abril de 2017, de https://www.gurucargo.com/Intro/Home/Index

Hernández O., J. (2004). *Introducción a la minería de Datos* España: Pearson Education

IBM Institute for Business Value (2013). Analítica de datos: un proyecto de generación de valor. Mayo 09, 2017, de IBM Sitio web: https://www.ibm.com/midmarket/es/es/att/pdf/Analitica_de_datos_para_pymes.pdf

iContainers: transporte marítimo internacional de mercancías online. (2018). Recuperado el 29 de abril de 2017, de https://www.icontainers.com/es/

INEGI. (s.f.). *I*nstituto Nacional de Estadística y Geografía. Recuperado el 01 de Abril de 2018, de http://www.inegi.org.mx

Joyanes A., L. (2013). *Big Data.* México: Alfa Omega.

Kontainers. (2018). Recuperado el 28 de Abril de 2017 https://kontainers.com/

Las 7 V del Big data: Características más importantes - IIC. (2018). Recuperado el 13 de agosto de 2018, de http://www.iic.uam.es/innovacion/big-data-caracteristicas-mas-importantes-7-v/

López Takeyas, B. (2005). *Inteligencia Artificial*. Obtenido de Inteligencia Artificial: http://www.itnuevolaredo.edu.mx/takeyas/Apuntes/Inteligencia%20Artificial/Apuntes/tareas_alumnos/C4.5/C4.5(2005-II-B).pdf

Los 10 Algoritmos esenciales en Machine Learning - Raona. (2018). Recuperado el 13 de agosto de 2018, de https://www.raona.com/los-10-algoritmos-esenciales-machine-learning/

México rompe récord en exportaciones de 2017. (2018). Recuperado el 13 de agosto de 2018, de http://www.elfinanciero.com.mx/economia/mexico-rompe-record-en-exportaciones-de-2017

Ortiz, M. (s.f.). *isw-udistrital.blogspot*. Obtenido de http://isw-udistrital.blogspot.mx/2012/09/ingenieria-de-software-i.html

Ponce J., Oronia Z., Silva A., Muños J., Ornelas F. & Alvares F. (2014). 481 Incremento del Interés de Alumnos en Educación Básica en los Objetos de Aprendizaje Usando Realidad Aumentad a en las Matemáticas. noviembre 14, 2016, de LACLO Sitio web: http://laclo.org/papers/index.php/laclo/article/viewFile/268/250

QA Técnico. (12 de 12 de 2015). *QA Técnico:* . Recuperado el 29 de abril de 2018, de http://qatecnico.blogspot.mx/2015/12/big-data-introduccion-y-caracteristicas.html

Raschka, S. (2015). Python Machine Learning. En S. Raschka, *Python Machine Learning* (pág. 425). Birmingham, Alabama: Published.

RPubs - Árboles de predicción: bagging, random forest, boosting y C5.0. (2018). Recuperado el 2 de agosto de 2018, de https://rpubs.com/Joaquin_AR/255596

s|ngular - CRISP-DM: La metodología para poner orden en los proyectos de Data Science. (2018). Recuperado el 10 de agosto de 2018, de https://data.sngular.team/es/art/25/crisp-dm-la-metodologia-para-poner-orden-en-los-proyectos-de-data-science

The easiest way to manage and control all logistics operations. (2018). Recuperado el 29 de abril de 2017, de http://lotebox.com/index.html

The observatory of economic complexity(2010). The observador of economic complexity. Junio 01, 2017 de Sitio web: http://atlas.media.mit.edu/es/profile/country/mex/

45HC.com - Easy online container booking. (2018). Recuperado el 29 de abril de 2017, de https://www.45hc.com/

(2018). Recuperado el 02 de agosto de 2018, de http://www.itnuevolaredo.edu.mx/takeyas/Apuntes/Inteligencia%20Artificial/Apuntes/tareas_alumnos/C4.5/C4.5(2005-II-B).pdf

(2018). Recuperado el 02 de agosto de 2018, de http://www.utm.mx/~jahdezp/archivos%20estructuras/DESICION.pdf

Glosario

Agente aduanal: persona autorizada para promover por cuenta ajena el despacho de las exportaciones.

Archivo CSV: es un archivo de texto que almacena los datos en forma de columnas, separadas por coma y las filas se distinguen por saltos de línea.

Book: reservación.

Booking: reservación del espacio a utilizar para hacer la exportación en el medio de transporte a utilizar.

Consignatario: es el responsable de la mercancía que recibe.

Embarcador: persona que opera el transporte internacional, cuando la mercancía tiene como destino una exportación.

Framework: patrón, esqueleto para el desarrollo o implementación de alguna aplicación.

Landing page: la landing page o Página de aterrizaje su objetivo es conseguir usuarios quienes las visitan realizar una acción, puede ser desde comprar un producto hasta enviar un formulario de cualquier tema.

Overfitting: proceso que reduce la capacidad predictiva al aplicarlo a nuevos datos.

Pruning: estrategia para controlar el tamaño del árbol, seleccionando la mejor condición de parada.

Streamed/online: tecnología que permite ver un archivo de audio o video directamente desde Internet.