Trabajo Fin de Grado

Grado en Administración y Dirección de Empresas

Facultad de Administración y Dirección de Empresas
Universitat Politècnica de València

Modelos basados en variables WEB para predecir el comportamiento exportador empresarial.

Autor:
Sara Agües Solaz
Tutores:
Ana Mª Debón Aucejo
Josep Domènech i de Soria

Valencia, septiembre 2019

Resumen

La actividad exportadora es una importante contribución al desarrollo económico de los países y, como consecuencia de la cada vez mayor globalización, esta actividad se convierte en fundamental para la dinamización de la economía de los mismos. Es por ello que, es de vital importancia el poder obtener datos y seguir su evolución con indicadores que puedan obtenerse en tiempo real.

Esto es difícil de obtener dado que las estadísticas oficiales suelen publicarse con cierta dilación de tiempo, por lo que en estos casos cobran especial importancia los modelos de aprendizaje basados en variables web.

En este Trabajo Fin de Grado, correspondiente a los estudios de Graduado en Administración y Dirección de Empresas, se analizan diferentes modelos de *Machine Learning* que nos van a permitir predecir, en tiempo real, el comportamiento exportador de un país a partir de diferentes variables que encontramos en las webs de sus empresas.

Palabras clave: Variables web, Machine Learning, exportación

Resum

L'activitat exportadora és una important contribució al desenvolupament econòmic dels països i, com a conseqüència de la cada vegada major globalització, aquesta activitat es converteix en fonamental per a la dinamització de l'economia d'aquests. És per això que, és de vital importància el poder obtenir dades i seguir la seua evolució amb indicadors que puguen obtenir-se en temps real.

Això és difícil d'obtenir atès que les estadístiques oficials solen publicar-se amb certa dilació de temps, per la qual cosa en aquests casos cobren especial importància els models d'aprenentatge basats en variables web.

En aquest Treball Fi de Grau, corresponent als estudis de Graduat en Administració i Direcció d'Empreses, s'analitzen diferents models de Machine Learning que ens permetran predir, en temps real, el comportament exportador d'un país a partir de diferents variables que trobem en les webs de les seues empreses.

Paraules clau: Variables web, Machine Learning, exportació

Abstract

Exporting activity is an important contribution to the economic development of the countries and, like consequence of every time main globalization, this activity turns into fundamental for the give push of the economy of the same. It is thus that, is of vital importance the can obtain data and follow his evolution with indicators that can obtain in real time.

This is difficult to obtain since the official statistics are used to publish with some delay of time, by what in these cases charge particular importance the models of learning based in variable web.

In east Work End of Degree, corresponding to the studios of Graduated in Administration and Direction of Companies, analyse different models of Machine Learning that they go us to allow predict, in real time, the exporting behaviour of a country from different variables that find in the webs of his companies.

Keywords: Web Variable, Machine Learning, Export

Tabla de contenidos

Índice de Tablas

Índice de Gráficos

1 Introducción

1.1 Resumen

Hoy en día la actividad exportadora es una importante contribución al desarrollo económico de los países y, como consecuencia de la cada vez mayor globalización, esta actividad se convierte en fundamental para la dinamización de la economía de los mismos. Es por ello que, es de vital importancia el poder obtener datos relevantes y significativos y, seguir su evolución con indicadores que puedan obtenerse lo más aproximado al tiempo real.

Además, y dado que actualmente el crecimiento de las Tecnologías de la Información y de la Comunicación (TIC) es exponencial comparado con otros canales de información y cada vez son más las empresas, prácticamente todas, que utilizan internet como medio para actualizar, publicar y difundir información sobre su actividad, es en este momento cuando cobra gran importancia la página WEB destinada a ello, ya que es esta es la fuente de información en tiempo real por excelencia, pudiendo encontrar día tras día gran flujo de información con una alta fiabilidad, aunque no siempre podemos pensar que lo será al cien por cien.

Las estadísticas oficiales suelen publicarse con cierta dilación de tiempo. Este retraso, variable según la fuente, puede ir de unos meses a, en ocasiones, más de un año.

Este hecho es un hándicap para llevar a cabo el estudio de cualquier variable que se pretenda, pues los datos a analizar quedan desfasados incluso antes de su publicación. Es por ello que, en estos casos, cobran especial importancia los modelos basados en variables web, los cuales pueden ayudarnos a obtener la información que nos interesa en un tiempo más aproximado al real.

Existen plataformas como la World Wide Web, coloquialmente conocida como WEB donde, día tras día, las empresas tienen la posibilidad de publicar y actualizar la información que de ellas quieren dar a conocer. En este sentido, cada día más la Web, se está convirtiendo en una gran base de datos, útil para contrarrestar la información desfasada que obtenemos con las estadísticas oficiales, debido a lo cambiante e impredecible que puede llegar a ser la economía.

En el presente Trabajo de Fin de Grado (TFG), correspondiente a los estudios de Graduado en Administración y Dirección de Empresas por la Universitat Politècnica de

València, se van a analizar diferentes modelos de Machine Learning que nos van a permitir predecir el comportamiento exportador de una empresa a partir de las variables que encontramos en las diferentes webs.

Esto abre nuevas posibilidades de investigación en el ámbito empresarial a partir del empleo de la información extraída de la estructura de enlaces de la Web.

Para ello, mediante el programa *RStudio* se ha realizado un código para obtener, si los hubiera, resultados referentes a lo explicado anteriormente.

1.2 Objetivos

El objetivo de este TFG es analizar los diferentes modelos de aprendizaje automático (Machine Learning) supervisado de clasificación, para predecir el comportamiento exportador de las empresas a partir de las variables que encontramos en sus webs corporativas. Dichos modelos, servirán para que, tanto empresas como otras entidades gubernamentales, puedan anticipar movimientos futuros y así adelantarse a los acontecimientos.

Para ello, hemos analizado a partir de una base de datos, con información de diferentes empresas, obtenida como fuente de información secundaria, si el comportamiento exportador de una empresa viene dado por alguna de las diferentes variables que podemos encontrar en su web.

Esta base de datos ha sido ligeramente modificada para facilitar su estudio con el programa estadístico R mediante su interfaz RStudio.

Con ello, se ha querido demostrar que hay información válida en la web y que solamente hay que saber hacer un buen uso de estos datos. Que dependiendo de donde obtengamos dicha información va a ser posible anteponerse al comportamiento de las empresas y así, establecer predicciones y especular en tiempo real sobre cómo puede evolucionar la economía.

1.3 Justificación y relación con las asignaturas de la titulación.

Se puede afirmar que hoy en día el mercado exterior es una gran oportunidad de negocio para las empresas en expansión, más aun, teniendo en cuenta la dificultad que supone tener acceso, lograr entrar y estabilizarse en el mismo.

Es por ello que, cuando una empresa está consolidada en dicho mercado exterior su comportamiento, sobre todo en lo que a visibilidad a través de páginas web se refiere, tiende a cambiar, intentando adaptarse a ese nuevo mercado ofreciendo la información que se demanda. De ahí, la posibilidad de realizar diferentes estudios y análisis con los datos ofrecidos para poder predecir el comportamiento de dichas empresas, y hacer un seguimiento de las ya asentadas anteriormente.

En primer lugar, es necesario ser consciente de que, para lograr adaptarse con éxito a las exigencias de este nuevo mercado, sin descuidar al resto, es de vital importancia conocer qué se exige en el mismo, ya que sin dicho conocimiento será difícil mantenerse en él.

Si para ello utilizamos diferentes modelos Machine Learning basados en variables que podemos encontrar en las páginas web de las diferentes empresas para anticiparnos a la posible evolución, y no nos quedaremos a la espera de unas estadísticas oficiales, seguramente desfasadas por el paso del tiempo, podemos tomar decisiones con mayor objetividad y seguridad.

El RD 1393/2007, por el que se establece la ordenación de las enseñanzas universitarias oficiales, modificado por el RD 861/2010 dispone, con carácter general, que todos los títulos oficiales "concluirán con la elaboración y defensa" de un Trabajo Fin de Grado (TFG) o Trabajo Fin de Máster (TFM), según el caso.

De acuerdo con la Normativa Marco de Trabajo Fin de Grado de la Universitat Politécnica de Valencia, los TFG de sus estudiantes deberán estar orientados a la aplicación y evaluación de competencias asociadas al título y consistirán en la realización de un trabajo o proyecto original en el que queden de manifiesto conocimientos, habilidades y competencias adquiridas por el estudiante a lo largo de sus estudios y, expresamente, las competencias asociadas a la materia TFG o TFM, tal y como se indique en la memoria de verificación.

En el caso del TFG que nos ocupa, estando el mismo relacionado directamente con varias de las competencias específicas reflejadas en la memoria de verificación, hay que destacar la vinculación directa con la competencia Nº 15 la cual se define como: Saber aplicar las herramientas básicas de naturaleza cuantitativa para el diagnóstico, análisis y prospección empresarial y conocer los modelos matemáticos, estadísticos, econométricos y de optimización para la toma de decisiones

Para lograrlo, ha sido de gran ayuda, siendo en algunos casos imprescindibles, los conocimientos adquiridos en las asignaturas incluidas en el Plan de Estudios y que a continuación se detallan:

Introducción a la Administración de empresas. En esta asignatura el estudiante obtiene una visión general sobre la empresa, aprendiendo algunos métodos y procedimientos para la gestión de la información, considerando la organización como un sistema.

Gestión de comercio exterior. El objetivo de esta asignatura es que el estudiante adquiera los conocimientos necesarios para enfrentarse al mercado internacional con éxito, siendo para ello necesario conocer los aspectos de la gestión de este mercado.

¿Por qué la empresa se lanza al exterior? ¿Qué factores hay que considerar en la contratación internacional? ¿A qué barreras se enfrentan? Todas estas y muchas más son las dudas que surgen al tomar esta decisión.

Como ya se ha dicho anteriormente, en esta asignatura, además de aprender a obtener y analizar datos sobre el comercio del propio país y sus relaciones con otros países para poder dar respuesta a las preguntas planteadas anteriormente, también conocemos la normativa y los diferentes pasos y procedimientos que ha de seguir una empresa para poder exportar sus productos dando el salto al mercado internacional.

Finalmente, para entender la totalidad del trabajo es necesario conocer el entorno empresarial propio y el entorno empresarial en el que la empresa se quiere desarrollar.

Los conocimientos aquí adquiridos nos van a ser muy útiles durante la elaboración de todo el TFG.

Introducción a la estadística. En esta asignatura se nos proporciona como estudiantes el conocimiento sobre el manejo básico de las técnicas más relevantes para el tratamiento de la información cuantitativa, proporcionando herramientas para ser capaces de analizar los resultados obtenidos a través de los modelos estadísticos que se puedan utilizar y con la finalidad de ayudar a las empresas a la hora de tomar decisiones.

Así mismo se facilita al estudiante las herramientas necesarias para ser capaces de interpretar de forma rigurosa la realidad a la que se enfrentan en el mundo laboral, dando un enfoque científico a la toma de decisiones, la gestión empresarial y la mejora resultados y siendo capaces de interpretar los resultados que se puedan obtener en los modelos estadísticos que se utilicen.

Métodos estadísticos para la economía. En esta asignatura se amplían los conocimientos adquiridos en Introducción a la Estadística, siendo necesario este conocimiento para el correcto desarrollo de la inferencia estadística que se pueda producir

para obtener conclusiones en el estudio. Se enseña al estudiante a tener un pensamiento crítico y a saber cómo reaccionar ante problemas estadísticos que l puedan surgir.

Así mismo, se introduce al alumno en alguno de los modelos estadísticos que se van a ver en este TFG, así como la manera de afrontar algún modelo desconocido en el caso de ser necesario.

Econometría. La asignatura se centra en diferentes modelos para describir la situación y predecir los valores futuros de una variable de interés. El objetivo es conocer diferentes modelos, saber cuándo es necesario aplicarlos, e interpretar sus resultados para cuantificar adecuadamente la situación económica de la empresa y su entorno

1.4 Estructura del trabajo.

Estructuraremos el TFG en siete capítulos distribuidos de la siguiente forma:

Capítulo 1. Introducción

Se trata de un capítulo introductorio, donde hemos detallado los objetivos del trabajo y la relación del mismo con las competencias específicas de la titulación y la vinculación con las diferentes asignaturas en las que se estructura el plan de estudios del Grado en Administración y Dirección de Empresas.

Capítulo 2. Situación actual.

En este capítulo se habla del momento actual en el que se encuentra la exportación a nivel nacional e internacional haciendo referencia a la importancia que tiene el comercio exterior y la evolución que ha tenido este en los últimos años, según el gobierno de España.

Capítulo 3. Ciencias de la información.

Explicación de la World Wide Web, así como los diferentes campos que componen las ciencias de la información, haciendo especial hincapié en la webmetría y su relación con el resto de ciencias de la información.

Capítulo 4. Metodología.

Se dedica este capítulo a la explicación de los diferentes modelos estadísticos que han tenido un peso importante en el contenido de TFG. Modelos para la predicción de variables binarias tales como “Arboles de regresión”, modelo “Random Forest”, modelo “El vecino más próximo”, y “Máquinas de Soporte Vectorial”.

Capítulo 5. Modelos y resultados

Se lleva a cabo en este capítulo una descripción detallada de la base de datos utilizada y la aplicación de los modelos basados en las variables web para predecir el comportamiento exportador de las empresas.

Capítulo 6. Conclusiones y propuestas de mejora.

Finalmente, en el sexto capítulo, se recogen las conclusiones derivadas de los resultados del análisis mediante los modelos citados con anterioridad, modelos que como ya se ha indicado analizan variables web para estudiar el comportamiento exportador de las empresas.

Como conclusión se lleva a cabo un balance de las posibles mejoras a implementar sentando las bases de un futuro trabajo que pudiera acontecer.

2 Situación actual de la exportación

En este capítulo se describe la importancia de los intercambios comerciales entre países y la manera en la que estos han ido evolucionando a lo largo de los últimos años en el entorno empresarial todo ello desde la perspectiva y el contexto de la red a través de internet.

En particular centraremos nuestra atención en la manera en que en muy poco tiempo y de forma muy rápida se ha vuelto de vital importancia para las empresas estar en todo momento conectadas y al día en internet. La red se ha convertido en una herramienta indispensable para la labor de las empresas, ya que les permite mantenerse comunicadas, tener acceso y compartir información, e incluso organizar las tareas cotidianas como pueda ser la comunicación electrónica.

Por último, se hablará de los métodos de análisis de las empresas y las variables web a analizar.

2.1 Comercio exterior

El comercio exterior es aquel que se refiere al intercambio de bienes y servicios que se realiza fuera de las fronteras geográficas de un país con otros países y sus mercados, y contiene regulaciones adicionales que establecen los participantes en el intercambio y los gobiernos de sus países de origen.

Según el Fondo Monetario internacional, ningún país es autosuficiente por él mismo, y es aquí cuando cobra realmente importancia el comercio exterior, que busca generar bienestar y supervivencia. En un país, el comercio exterior de sus empresas es una pieza fundamental de la estabilidad financiera, ya que los países obtienen ingresos por los intercambios realizados con el exterior que contribuyen a acrecentar la riqueza del mismo y, por tanto, favorecen el aumento de su producto interior bruto (PIB).

Englobado en el concepto de comercio exterior, podemos distinguir entre “exportación” e “importación”.

En el contexto de la exportación, a través de estudios realizados por el Ministerio de comercio exterior, podemos concluir que, si una empresa está dentro de lo denominadas “exportadoras” de un país y además se trata de una Pequeña y Mediana Empresa (PYME), es un buen indicador de que la economía de dicha empresa va por buen camino, ya que normalmente cuando una empresa expande sus horizontes al exterior será debido a que ya está asentada y tiene unos beneficios estables en su país de origen.

Estudios de la Dirección General de Estudios Económicos, Evaluación y Competitividad Territorial concluyen en que las PYMES exportadoras son de mayor tamaño, tienen una mayor productividad, mayor volumen de facturación, más y mejores ingresos y, muy importante, un desarrollo y un desempeño tecnológico mayor que las que no lo son.

2.1.1 La evolución de la exportación y su importancia.

En la web de la Agencia Tributaria del gobierno de España encontramos publicados los datos estadísticos de transacciones comerciales realizadas por las empresas españolas con operadores de otros países, proporcionándonos dichas estadísticas información multidimensional en términos de valor y de cantidades físicas de las mercancías objeto de las transacciones comerciales así como variables de carácter general tales como país de contrapartida, código de la mercancía, mes de referencia, naturaleza de la transacción, provincia de origen/destino, condiciones de entrega, medio de transporte, etc.

Con datos actualizados a finales de 2018 podemos comprobar que, en España, la exportación de mercaderías ha crecido un 2,57% en 2016 respecto al año anterior 2015, un 7,15 % en 2017 respecto a 2016 y se llegó a alcanzar un crecimiento del 3,11% en el último año objeto de estudios respecto a 2017. Es por esto que, en 2018, las ventas al extranjero representan un 24,17% del PIB español.

A la vista de los datos declarados se confirma la importancia de la exportación para el desarrollo de un país, siendo esta importancia creciente, habiéndose alcanzado en el año 2018 cifras récord de exportación de mercaderías en nuestro país.

Esto datos nos llevan a la conclusión de que la exportación en nuestro país cada año cobra más importancia, siendo ello consecuencia de que la actividad exportadora de las empresas es cada vez mayor.

Los sectores que más aportan a la exportación en España, y que son de los que más contribuyen al aumento de las exportaciones cada año, son el sector de bienes de equipo (19,99%), sector del automóvil (15,61%), alimentación, bebidas y tabacos (16,1%) y productos químicos (14,31%).

A continuación, podemos ver la tabla 1 y dos gráficas en las que se muestra resumida la información ofrecida por el Gobierno de España “informe sobre las exportaciones de Mercadería” (2018).

En la tabla 1 se muestra vemos los datos correspondientes a 2018 diferenciados por sectores y cuantificados, tanto en millones de euros como en porcentaje.

Sector	Exportacion (millones €)	% del total
Alimentación, bebidas y tabaco	**45.877,3**	**16,10%**
Productos energéticos	**22.581,3**	**7,92%**
Materias primas	**7.696,6**	**2,70%**
Semimanufacturas no químicas	**29.743,4**	**10,44%**
Productos químicos	**40.789,1**	**14,31%**
Bienes de equipo	**56.981,0**	**19,99%**
Sector automóvil	**44.490,4**	**15,61%**
Bienes de consumo duradero	**4.530,8**	**1,59%**
Manufacturas de consumo	**28.416,2**	**9,97%**
Otras mercancías	**3.917,9**	**1,37%**
Total	**285.023,9**	**100,00%**

Tabla 1: . Datos exportación en España Enero-diciembre 2018. Fuente: Elaboración propia con datos del Informe de Comercio Exterior del Ministerio de Industria, comercio y turismo.

En gráfico 1 podemos ver la evolución del crecimiento de las exportaciones, podemos ver que, aunque el crecimiento se ha ralentizado comparado con el año anterior, siguen creciendo año tras año (3,1%) en 2018.

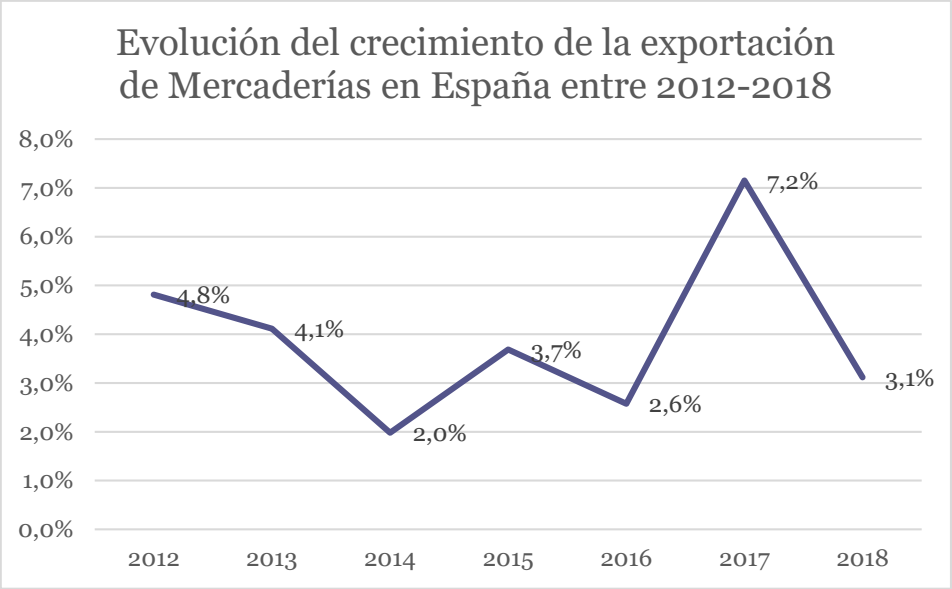


Gráfico 1: Evolución del crecimiento de la exportación de Mercaderías en España entre 2012-2018. Fuente: Elaboración propia con datos del Informe de Comercio Exterior del Ministerio de Industria, comercio y turismo.

Por último, la evolución de dicha importación en millones de euros.

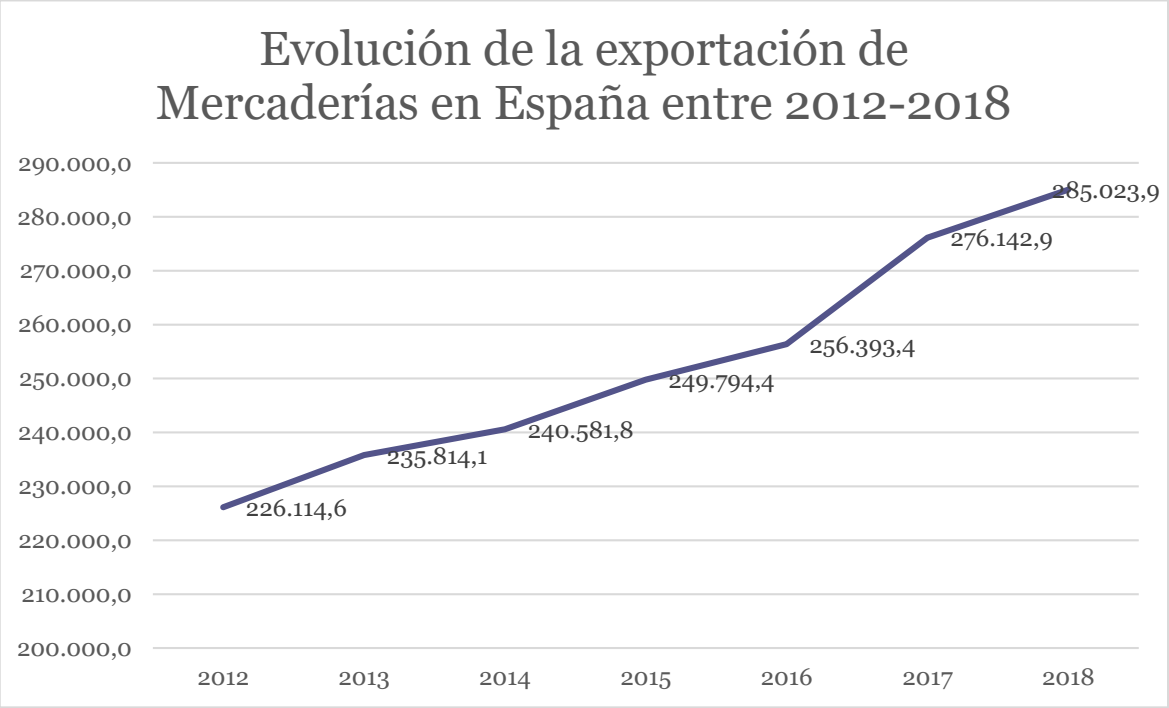


Gráfico 2: Evolución de la exportación de Mercaderías en España entre 2012-2018. Fuente: Propia con datos del Informe de Comercio Exterior del Ministerio de Industria, comercio y turismo.

Y en la tabla 2 un resumen con todos los datos mostrados anteriormente:

AÑO	2012	2013	2014	2015	2016	2017	2018
MILLONES €	226.114,6	235.814,1	240.581,8	249.794,4	256.393,4	276.142,9	285.023,9
%CRECIMIENTO	4,81%	4,11%	1,98%	3,69%	2,57%	7,15%	3,12%

Tabla 2: Datos exportación en España Enero-Diciembre 2018. Fuente: Propia con datos del Informe de Comercio Exterior del Ministerio de Industria, comercio y turismo.

2.2 Gestión del conocimiento

Desde la década de los años ochenta en el siglo XX, la llamada gestión de información o gerencia de información ha ganado un espacio importante en la vida de las instituciones en general y, en particular, en aquellas que tienen como misión el desarrollo de servicios y productos de información,

Según Paul Watzalawick y su estudio de la teoría de la comunicación humana (2018), cuando se habla de organización es casi imposible no hablar de información, sin información no hay organización posible. Así, se entiende la Gestión de Información

como una función, actividad o proceso estratégico que se desarrolla en una organización de cualquier tipo con la finalidad de conseguir el desarrollo de servicios y productos.

Es un proceso que afecta e implica a todos las operaciones, gestiones y actividades que tengan lugar en dicha organización, así como a sus componentes, por lo que necesariamente se establece una estrecha relación con el sistema que lo rige.

Es importante, para la compresión de este trabajo, diferenciar entre datos, información y conocimiento.

Para establecer estas diferencias hemos seguido la teoría de Ponjuan.

Según Ponjuán los datos se pueden definir como "conjunto de hechos discretos y objetivos sobre acontecimientos. En el contexto de una organización, los datos son descritos como registros estructurados de transacciones.

Constituyen la materia prima para la creación de información." (Ponjuán, 2004).

Ponjuan nos dice que los datos solo son puntos en un espacio y un tiempo determinado.

Un conjunto de datos no llega a ser información sin un grado de relación y cohesión entre ellos. Un dato pasa a ser información cuando adquiere significación para su receptor. Es por ello que, los datos han de estar contextualizados y tener significación para que se conviertan en información y tengan valor para quien los interpreta.

Dicha significación es la que nos da las claves para entender los datos. Hasta que estos no son procesados, no se convierten en información.

Solamente cuando la información se pone en el contexto de una persona, y esta información le va a permitir actuar y tomar decisiones pasamos al conocimiento. La información se transforma en conocimiento con la percepción de uno mismo. En este sentido el conocimiento se encuentra más relacionado con las acciones que se lleven a cabo con la información que se tiene que con los datos en sí.

El individuo puede establecer conexiones entre la información obtenida y sus prácticas acumuladas a lo largo del tiempo, contextualizarla a través de experiencias, y será entonces cuando se podrá decir que la información fue interpretada y comprendida, transformándose en conocimiento.

Ponjuan (2004) considera que "tanto la información como el conocimiento tienen que ver con las personas, pero en diferentes niveles o dimensiones: la información depende de los datos que se convierten en información al tener significado a partir de diferentes procesos de agregación de valor, y de una determinada contextualización.

El conocimiento es información transformada en creencias, conceptos y modelos mentales mediante razonamiento y reflexiones".

Estos tres conceptos, lejos de ser excluyentes entre sí, se complementan e interrelacionan, y ninguno tuviera sentido sin la presencia implícita del otro.

De esta manera el conocimiento se vierte a posteriori como la evidencia concreta de la cadena de valor.

El objetivo de las empresas exportadoras, será lograr el máximo conocimiento, para poder predecir el comportamiento del mercado exterior, y el comportamiento del resto de empresas.

Gráfico 3: Piramide informacional. Fuente:Ponjuan 1998

El siguiente estudio es un claro ejemplo, teniendo en cuenta modelos basados en variables web para predecir el comportamiento exportador de las empresas, de ámbito práctico, para el que se utiliza información a partir de datos publicados para generar un conocimiento.

En él vamos a utilizar una serie de datos organizados como información para transmitir un cierto conocimiento a las personas que asimilen y comprendan el estudio.

3 Ciencias de la información

3.1 Indicadores económicos

Los indicadores económicos son una herramienta fundamental para conocer el estado de la situación actual respecto a la economía de un país. Se trata de datos estadísticos que nos van a permitir realizar un análisis de la situación económica tanto pasada como presente y nos facilitarán información para poder previsiones de cómo evolucionara la economía en el futuro con los datos de los que disponemos a día de hoy.

Según un estudio del BBVA (2015), existen diferentes tipos de indicadores económicos, los tres principales grupos en los que se dividen son;

Indicadores económicos adelantados, que son aquellos que por lo general cambian antes que la economía en su conjunto. Cumplen una función de detectores de tendencias económicas. Son útiles para predecir el comportamiento de la economía a corto plazo.

Indicadores económicos coincidentes, son indicadores que generalmente realizan un cambio de tendencia aproximadamente al mismo tiempo que la economía realiza un cambio en el ciclo económico. Al cambiar al mismo tiempo que la economía, nos ayudan a entender el estado actual de la misma.

Por último, están los **Indicadores económicos retardados** que son aquellos indicadores que por lo general cambian después que la economía en su conjunto. Se usan para confirmar previsiones económicas. Normalmente el retraso suele ser de uno o dos trimestres.

Entre los principales indicadores económicos encontramos el producto interior bruto (PIB), la inflación, tasa de desempleo, tasa de interés, prima de riesgo, etc.

Estos indicadores podemos encontrarlos a través de páginas oficiales en Internet, aunque tienen algunas desventajas, principalmente tenemos que destacar que los datos suelen salir con un retardo importante sobre todo si queremos obtener una predicción inmediata. Para evitar esto, tenemos alternativas como el uso de la World Wide Web.

La World Wide Web, a la que nos referiremos como Web es un sistema de documentos de hipertexto vinculados entre sí, accesibles desde internet, usando lo que conocemos como navegador Web. La Web

La Web se basa en tres estándares para funcionar:

- URL (Uniform Resource Locator):

El localizador de recursos uniformes, es una cadena de caracteres, que se utiliza para asignar una dirección que será única a cada uno de los recursos de la información que aparecen en internet.

Cada uno de los documentos que encontramos en la World Wide Web tienen un URL diferente

- HTTP (Hyper Text Transfer Protocol)

El HTTP establece el protocolo para el intercambio de documentos en la web.

Una variante del HTTP es el protocolo HTTPS, se trata de una variante cifrada por medio de TLS (Transport Layer Security). Se trata de un protocolo que nos ofrece una comunicación segura en internet mediante la encriptación.

HTTP utiliza un modelo cliente-servidor donde el navegador (cliente) solicita información mediante una petición al servidor y espera a que este le envíe la misma mediante una serie de intercambios.

- HTML (Hyper Text Markup Language)

Es un estándar a cargo del World Wide Web Consortium que sirve de referencia del software que conecta con la elaboración de páginas web en sus diferentes versiones, definiendo la estructura básica y el código HTML para los contenidos de una página web tales como texto, imágenes, videos, etc.

Para entender el estudio y su contexto, tenemos que tener claros algunos aspectos de las ciencias de la información.

3.2 Ciencias de la información

- INFORMETRÍA

Es una de las disciplinas de las ciencias de la información siendo un término que empezó a utilizarse en los años 80. Según Brookes (1990), la informetría se define como la ciencia que estudia los aspectos cuantitativos de la información, sin tener en cuenta el modo en el que esta se genere.

La informetría también estudia los aspectos cuantitativos de la comunicación hablada o informal de la misma forma que los de la información si registrada.

- BIBLIOMETRÍA

La Bibliometria, como disciplina instrumental, según Brookes (1990) se la define como la técnica de investigación bibliológica que aplica los diferentes métodos matemáticos y estadísticos a la literatura de carácter científico con el objetivo de, por un lado, analizar el tamaño, crecimiento y distribución de la bibliografía en un campo determinado, y, por otro, estudiar la estructura social de los grupos que la producen y la utilizan.

Los indicadores que se utilizan para la medición de estos aspectos son los indicadores bibliométricos, que proporcionan información sobre los resultados de la actividad científica.

El Dr. Melvin Morales et al (1998), la bibliometría:

"como disciplina métrica que aplica métodos y modelos matemáticos al objeto de estudio de la bibliotecología, biblioteca, documento y lector, con el propósito de cuantificar el desarrollo de los procesos relacionados con las bibliotecas como fenómenos sociales, vinculados a la utilización de las riquezas literarias en interés de la sociedad, es decir, se ocupa del análisis de la teoría y regularidades, tanto del documento como de los procesos y actividades bibliotecarias (teoría de la circulación, uso en biblioteca, de las fuentes documentales, de bases de datos, modelos de redes de bibliotecas y solapamiento, etc.) para contribuir a la organización y dirección de las bibliotecas."

Además, el autor español Pedro López López (1996) planteó que:

"dicha ciencia es simplemente una herramienta metodológica que parte de la necesidad de cuantificar ciertos aspectos de la ciencia y que una de las facetas de la cienciometría sería la bibliometría, entendida como el cómputo de diversos indicadores de publicaciones que los científicos producen."

- CIENCIOMETRÍA

La cienciometría es la ciencia que nos permite el estudio cuantitativo de la actividad científica con el fin de medir y analizar la misma desde una visión económica y social.

A través de técnicas métricas, la cienciometría permite el desarrollo de las políticas científicas de un país u organización, al examinar el crecimiento cuantitativo, el desarrollo de las disciplinas, la vigencia de paradigmas científicos, estructuras de comunicación, productividad, innovación, desarrollo científico, crecimiento económico, etc.

En la práctica, existe una superposición significativa entre la cienciometría y otros campos científicos como la bibliometría, y los sistemas de información,

- CIBERMETRÍA

Cibermetría es la disciplina dedicada a la descripción cuantitativa de los contenidos y procesos de comunicación que se producen en el ciberespacio, considerando ciberespacio como el conjunto de contenidos accesibles en formato electrónico.

Fue definida por Ali Ashgar Shiri en 1998 como la medición, estudio y análisis de toda clase de información y medios de información que podemos encontrar en el ciberespacio y que emplean técnicas bibliométricas, cienciométricas e infométricas.

Está considerada como un subcampo dentro de la informetría, su cometido es el análisis de la información electrónica, aquella que circula por la web.

La cibermetría suele utilizarse para, por ejemplo, analizar la presencia de un país en las diferentes herramientas de internet (webs, foros etc.)

- WEBMETRÍA

Por último, a finales del siglo XX empieza a desarrollarse una nueva métrica paralela al acelerado avance en el mundo de la información denominada Webmetría.

Esta nueva técnica va a permitir estudiar la Web y obtener datos de la misma desde un punto de vista cuantitativo

De acuerdo con Björneborn (2004), se define como:

"*El estudio de los aspectos cuantitativos de la construcción y uso de los recursos de información, estructuras y tecnologías de una parte concreta de Internet, por regla general a una web o portal, desde perspectivas bibliométricas e informétricas* "

Un aspecto importante en el estudio de las ciencias de la información es el uso de indicadores métricos los cuales son una gran ayuda, una vez analizados, para predecir ciertos fenómenos como, por ejemplo:

- Crecimiento de algún campo de la ciencia
- Evolución cronológica de la producción
- Productividad de ciertas instituciones

La webmetria está basada en la bibliometría, que es la parte de la bibliología que estudia la producción científica editada en los libros a través de métodos estadísticos.

Aunque la webmetría esté basada en la bibliometría, tiene que ser capaz por ella misma de desarrollar técnicas para identificar y cuantificar los continuos cambios que hay en la web.

La Webmetría se encuentra en el contexto de las ciencias de la información. En este contexto encontramos todos los términos definidos en este capítulo: informetría, bibliometría, cienciametria, cibermetria y webmetría.

La Webmetria se limita a la información en la Web, que cada día más se trata del espacio más relevante para la obtención de información aplicadas a las ciencias sociales, y que cada vez son más importantes para estudios relacionados con la economía y las empresas como es el caso que nos ocupa.

En el gráfico 4 podemos ver la relación entre las diferentes ciencias de la información y que posición ocupa la Webmetria en ellas.

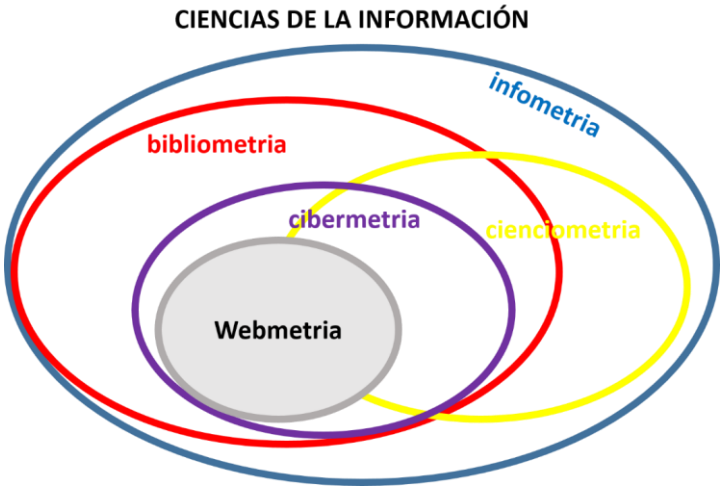


Gráfico 4: Esquema de la situación de la Webmetria en el contexto de las ciencias de la información. Fuente: Adaptado a partir de Björneborn e Ingwersen(2004).

Dada la cantidad de información almacenada en la web las posibilidades para medirla son infinitas. Para llevar a cabo este análisis una de las herramientas utilizadas son los llamados webcrawler, también conocidos como arañas, robot de búsqueda, spider, bot, etc.

El propósito de estas herramientas es seleccionar aquellos elementos de la web que nos puedan interesar y hacer visible dicha información. Entre las principales variables que a través de esta técnica vamos a buscar en la Web podemos destacar:

- El estudio de palabras clave, realizando de esta manera el análisis del impacto que pueden tener marcas, ideas, empresas, todo esto a partir de la información que se obtiene de las páginas web.
- La investigación basada en hiperenlaces, destacando las relaciones a través de enlaces.

- La manera en la que se comportan los usuarios que realizan búsquedas en páginas web.

La importancia de la web, así como de las mediciones que de ella se hagan, radica, como ya se ha indicado anteriormente, en que podemos establecer pronósticos y tendencias a partir de ciertas variables y tomar decisiones de acuerdo a un contexto dado.

4 Metodología

Es en este capítulo vamos a proceder a identificar y detallar los modelos y métodos utilizados en el desarrollo de este trabajo.

En primer lugar, definiremos los objetivos del trabajo, y posteriormente pasaremos a explicar los diferentes métodos de aprendizaje supervisado que hemos estudiado y utilizado para el estudio.

4.1 Definición de los objetivos

El objetivo principal en los modelos es que, mediante las variables creadas, a partir del contenido de las páginas web corporativas de una empresa se pueda predecir si dicha empresa es o no exportadora.

Una vez definido este objetivo en los diferentes modelos, alineado con el objetivo principal de este trabajo, tenemos que tener en cuenta otros aspectos importantes tales como el hecho de que la información extraída de la base de datos que utilizaremos provienen de una fuente secundaria.

Estos datos se han ordenado en un fichero Excel cuyas columnas y filas son datos de 300 empresas en el año 2013. Utilizando estos datos, vamos a construir modelos a partir de los cuales podamos medir el comportamiento exportador de las empresas de nuestra economía.

En este capítulo también definiremos los diferentes métodos estadísticos que vamos a usar durante el estudio.

Los modelos utilizados son de predicción de variables binarias, aunque también utilizaremos procedimientos de validación de modelos.

El software estadístico que vamos a utilizar para hacer todos estos análisis será R.

R es un lenguaje y entorno de programación que generalmente, no se utiliza para programar; sino que más bien se utiliza interactivamente enfocado al análisis estadístico y está formado por un conjunto de herramientas muy flexibles que pueden ampliarse fácilmente mediante paquetes, librerías o definiendo nuestras propias funciones.

Además, es gratuito y de código abierto, lo que es una de sus principales ventajas.

Cualquier usuario puede descargar y crear su propio código de manera gratuita, sin restricciones de uso, la única regla es que la distribución sea siempre libre (GPL). Ello permite la replicabilidad y reproducibilidad de los resultados de este trabajo.

4.2 Aprendizaje supervisado

El aprendizaje automático, también conocido como machine learning, fue definido por Samuel Arthur (1959) como *"la capacidad de aprender sin programación explicita",* es por esto, que podemos decir con más exactitud que el aprendizaje automático explora el estudio y la construcción de algoritmos que son capaces de aprender y hacer predicciones sobre datos, si se siguen unas instrucciones estrictamente estáticas del programa al hacer predicciones o decisiones basadas en datos, mediante la construcción de un modelo a partir de entradas de la muestra.

Para desarrollar estos algoritmos, hablaremos de aprendizaje supervisado.

En el aprendizaje supervisado se entrena al algoritmo con un histórico de datos definiendo las preguntas, denominadas "características o atributos", y las respuestas a las mismas que nos ayudarán a definir las denominadas "etiquetas". De esta forma "aprende" a asignar la etiqueta de salida adecuada a un nuevo valor no incluido en la muestra de datos inicial.

Esto se hace con la finalidad de que el algoritmo las combine y pueda hacer predicciones.

Existen, a su vez, dos tipos de aprendizaje supervisado:

- **Regresión:** tiene como resultado un número específico, dado que la variable es un resultado concreto.
- **Clasificación:** en este tipo, el algoritmo encuentra diferentes patrones y tiene por objetivo clasificar los elementos en diferentes grupos, siendo en este caso la variable de tipo categórico.

Mediante este sistema se va a intentar encontrar alguna función que nos permita asignar a situaciones no estudiadas con anterioridad un resultado con valor objetivo.

Uno de los algoritmos de clasificación más conocidos y un ejemplo de aprendizaje supervisado es el "Algoritmos de vecinos más próximo" (KNN), que se explicará con más profundidad más adelante.

En resumen, el algoritmo del vecino más próximo se utiliza un conjunto de entrenamiento para clasificar los nuevos ejemplos. Este es un algoritmo de vecindad, que se basa en la idea de que si existe un nuevo ejemplo que todavía no está clasificado, este pertenecerá a aquella clase que comparta el mayor número de sus vecinos más próximos a él.

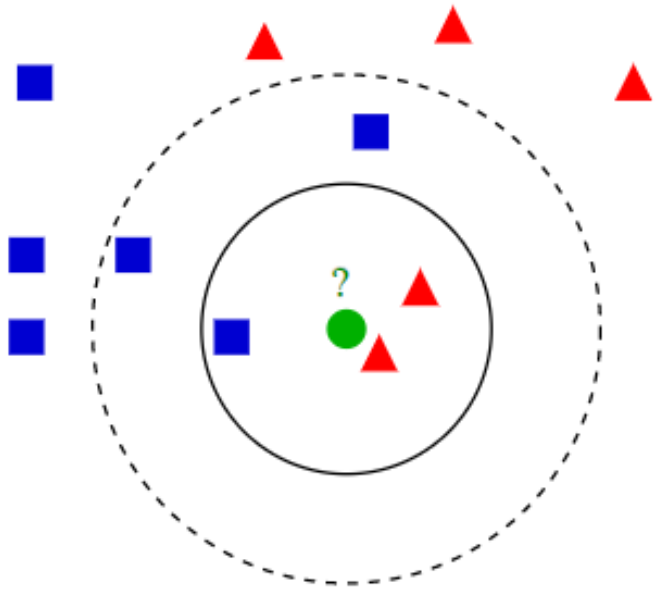


Gráfico 5: Algoritmo del vecino más próximo. Fuente: Sancho Caparrini, F (2018).

Cuando tenemos ya solucionado el problema de cómo desempeñar la tarea a realizar, tenemos que medir la **eficiencia de la máquina**, es decir, conseguir la extracción de alguna medida que nos informe de cómo está trabajando el modelo.

En el caso del **Aprendizaje supervisado,** como los algoritmos que se utilizan se retroalimentan de los datos de los diferentes modelos para ajustarse y poder sacar así alguna conclusión correcta, no tiene mucho sentido medir la eficacia de esta máquina volviendo a pasar los datos que ya conoce, ya que nos saldrían unas conclusiones mucho más optimistas de lo que realmente sería lo real.

Es por esto que lo que se busca es averiguar si la máquina encuentra la forma a partir de los ejemplos entrenados, de generalizar el comportamiento que anteriormente se ha aprendido, de forma que la máquina sea lo suficientemente buena sobre los datos que todavía no ha estudiado.

Si esto ocurre, es cuando decimos que el modelo/algoritmo generaliza de forma correcta.

La forma más habitual en la que se mide esta capacidad de generalización es guardar varios ejemplos iniciales para usarlos con posterioridad como validación del modelo/algoritmo.

En el caso de que hablemos del aprendizaje NO supervisado, el problema que surge es la no disposición de una respuesta verdadera desde el principio, por lo que se vuelve más difícil dar medidas fiables de eficiencia.

En el ejemplo de clusterización que hemos explicado antes, por ejemplo, lo que se mide es un potencial de estrés de la agrupación que hemos conseguido.

4.2.1 Modelos lineales generalizados

Los modelos lineales generalizados (GLM) son una extensión de los modelos de regresión lineal ordinarios que nos permiten utilizar distribuciones no normales y varianzas no constantes. Éstos investigan la relación entre uno o más predictores y una variable respuesta, utilizando variables de respuesta categórica.

La técnica de modelo lineal generalizado, al igual que la técnica de mínimos cuadrados, estiman los parámetros del modelo de forma que se busca optimizar el ajuste del mismo.

En los modelos lineales generalizados podemos encontrar tres componentes básicas:

- Componente aleatoria, que nos identifica la variable respuesta y la distribución de probabilidad de la misma.

  Consiste en una variable aleatoria Y con observaciones independientes (Y1> YN). Existen ocasiones en las que las observaciones son binarias y solamente se identifican como éxitos o fracaso, aunque generalmente cada Y1 nos indica el número de éxitos que encontramos en un número fijo de ensayos.
  Estos modelos los podemos incluir en la llamada “familia exponencial” de distribuciones:

  $$f(y_i|\theta_i) = a(\theta_i)b(y_i)\exp[y_i Q(\theta_i)]$$

- Componente sistemática, que nos especifica las diferentes variables explicativas ya sean independientes o predictoras que se han utilizado en la función predictora lineal.
  La componente sistemática nos especifica las variables explicativas que forman parte de efectos fijos en un modelo lineal, esto quiere decir, las variables Xj se relacionan mediante:

  $$\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

Esta combinación lineal de las diferentes variables explicativas lo denominamos predictor lineal.

- Función link. Mediante esta función relacionamos las componentes aleatoria y sistemática. Se trata de una función de valor esperado de Y, E(Y), como una combinación lineal de las variables predictoras.

4.3 Medición de variables con métodos basados en árboles

El enfoque de clasificación y regresión (CART) fue desarrollado por Breiman et al.(1984).

Los métodos basados en árboles para la regresión son un tipo de algoritmos de aprendizaje supervisado, donde existe una variable objetivo predefinida. Estos métodos dividen el espacio de predictores, que serían las variables independientes en regiones distintas y sencillas, no sobrepuestas. Para obtener las predicciones se suele usar la medida o moda de los entrenamientos en la zona en la que cada observación que se quiere predecir pertenece.

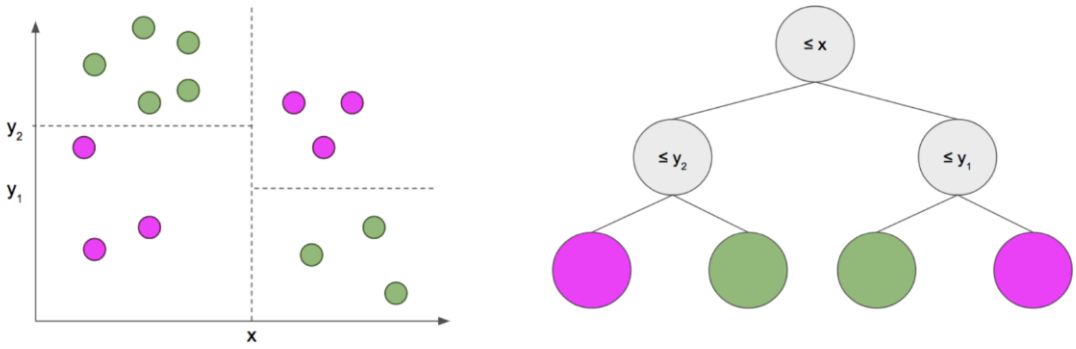


Gráfico 6: Árboles de decisión. Fuente: Orellana Alvear Johanna

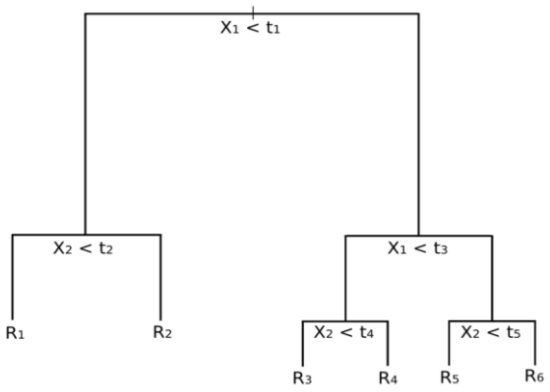


Gráfico 7: Árbol de decisión con seis regiones. Fuente: Orellana Alvear Johanna

Los métodos basados en árboles son sencillos y útiles para la interpretación, sin embargo, normalmente no son competitivos con los mejores enfoques de aprendizaje supervisados en términos de exactitud de predicción, es por esto que también se utilizan Existen métodos los cuales producen múltiples árboles que se combinan para mejorar la predicción, a expensar de una interpretación más complicada. A continuación, se explican sin entrar en detalle algunos de ellos.

- Bagging: Esta técnica consiste en crear diferentes modelos usando muestras aleatorias con reemplazo y luego combinar o ensamblar los resultados.
  El Bagging divide el set de entrenamiento en distintos sub set de datos, obteniendo como resultado diferentes muestras aleatorias
  Este sistema mejora la predicción, ya que lo que no detecta un modelo lo detectan los otros.
  Además, los árboles de decisión suelen sufrir de alta varianza, por lo que, si dividimos al azar los datos en dos o más grupos y ajustamos un árbol de decisión a cada uno de ellos, los resultados que obtendremos podrán ser muy diferentes.
  Es por ello que el método de bagging o bootstrap aggregation es un procedimiento utilizado para reducir la varianza de un método de aprendizaje estadístico, usado muy frecuentemente con árboles de decisión.

- Boosting*: Boosting* funciona de manera parecida al *bagging* en cuanto a que combina un gran número de árboles, a excepción de que los árboles se construyen de manera secuencial. Cada árbol se genera usando información, concretamente los residuos, de árboles previamente generados, en lugar de utilizar la variable respuesta (por ello suelen ser suficientes árboles más pequeños, en lugar de un gran árbol que pueda sobreajustarse a los datos).
  Boosting no utiliza remuestreo por *bootstrapping*, sino que cada árbol se genera utilizando una versión modificada del set de datos original.

Está basado en la idea de crear una regla de predicción altamente precisa combinando muchas reglas relativamente débiles.

Todos estos métodos construyen varios árboles que luego se combinan para producir una única precondición de consenso.

Normalmente, la combinación de un gran número de árboles puede conducir a mejorar la exactitud de la predicción.

Los árboles de decisión se pueden aplicar a los problemas de regresión y clasificación. Los primeros que veremos serán los de clasificación y, más tarde, los de regresión.

4.3.1 Terminología para árboles

Para poder definir la terminología de los árboles, utilizaremos el gráfico 8:

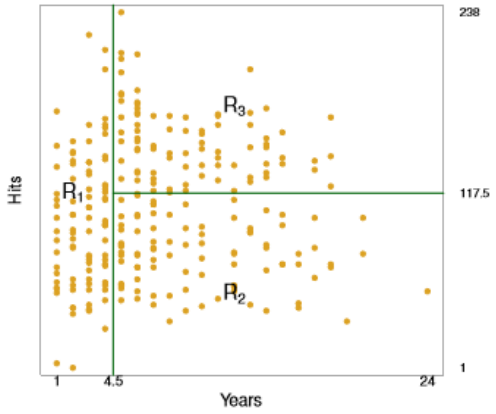


Gráfico 8:ISLR decisión tree Fuente: Mathur, Sameer

- De acuerdo con la analogía del árbol, las regiones R1, R2, R3, se conocen como nodos terminales.
- Los árboles de decisión típicamente se dibujan al revés, esto quiere decir que las hojas las encontraremos en la parte inferior del árbol.
- Denominamos nodos internos a los puntos a lo largo del árbol donde se divide el espacio del predictor.

4.3.2 Árboles de clasificación

Los árboles de clasificación y regresión (CART = Classification and Regresion Trees) son una alternativa a la tradicional regresión.

Las ventajas que los árboles CART tienen son, principalmente, su robustez hacia los outliers, la invarianza en la estructura de sus árboles de clasificación o de regresión a las transformaciones monótonas de las variables independientes y, sobre todo, su interpretabilidad.

Hablamos de árboles de regresión cuando la variable dependiente es continua y, árboles de clasificación cuando la variable dependiente es de tipo cualitativo. Estos métodos realizan de forma repetida particiones binarias del conjunto de observaciones que conforman la variable respuesta. En definitiva, se busca en ambos casos que, siguiendo las diferentes bifurcaciones, podamos obtener una predicción para la clasificación o para el valor que toman los individuos que cumplen con las propiedades que se han ido exigiendo en las distintas bifurcaciones.

Los árboles de decisión se construyen mediante un algoritmo conocido como segmentación recursiva, que es el proceso paso a paso para dicha construcción.

Existen 3 procedimientos principales:

- CHAID (Chi-Square Automatic Interaction Detector)
- QUEST (Quick unbiased Efficient Statistical Tree)
- CART (Classification an Regression Trees)

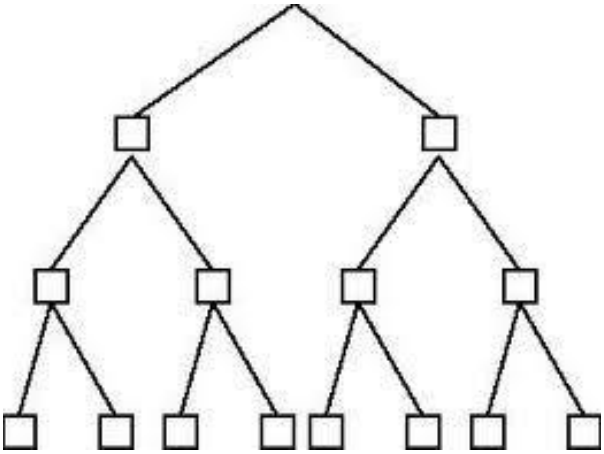


Gráfico 9: Estructura árbol de clasificación Fuente:Reed, Gina.

Método CART

La metodología CART (Classification And Regression Trees) consiste en tres pasos:

- Construcción del árbol saturado,
- Elección del tamaño correcto.

- Clasificación de nuevos datos a partir del árbol construido.

El modelo de regresión que se utiliza para predecir la variable respuesta que nos genera el algoritmo CART se puede definir de la siguiente manera

$$\hat{f}(x) = \sum_{i=1}^{n} c_m I(X_i \dots X_{i+n}) \epsilon R_m$$

Donde cm sería la constante que utilizamos en la región Rm

El método CART es una técnica de aprendizaje supervisado. Se tiene una variable objetivo y la finalidad es encontrar una función que nos permita predecir, a partir de variables independientes, el valor de nuestra variable objetivo, en un caso que es normalmente desconocido.

Una de las características más ventajosas de esta técnica, comparándola con las técnicas tradicionales de análisis de datos multivariables, es su mejor comportamiento ante situaciones de estructura discriminante, muy alejadas de la linealidad (OUTLIERS).

Como su método indica, el método CART es una técnica con la que se obtienen árboles de regresión o clasificación. Utilizaremos regresión cuando nuestra variable es continua y clasificación cuando ésta sea discreta.

En el software RStudio se implementará este método con RPART (Recursive Partitioning and Regression Trees.)

De forma general, lo que hace este algoritmo es, buscar la variable independiente que hace una mejor separación de los datos en distintos grupos, que se corresponden con las categorías de nuestra variable objetivo. Esto se expresa mediante una regla, y a cada regla, le correspondería un nodo.

La ventaja principal de este método es su interpretabilidad, ya que nos proporciona un conjunto de reglas, a partir de las cuales se deben tomar decisiones.

Hablamos de un método que no demanda en exceso una potencia de cálculo, pero aun así nos devuelve muy buenos resultados, para muchos tipos de datos y muy diferentes.

Por otro lado, existen desventajas. Hablaremos de que el método CART es un tipo de clasificación débil, pues sus resultados varían mucho dependiendo de la muestra de datos que utilicemos para entrenar al modelo, además resulta más fácil ajustar los modelos, y esto a veces se puede volver una desventaja.

No obstante, hemos decidido utilizar este método ya que interpretar los resultados resulta muy intuitivo.

4.3.3 Método Random Forest

El método *Random Forest* es una técnica de agregación, que mejora la precisión en la clasificación mediante la incorporación de aleatoriedad en los diferentes clasificadores individuales.

La aleatoriedad les algo que podemos introducir tanto en la construcción del árbol como en la muestra de entrenamiento.

El R*andom Forest* nace de la necesidad de resolver uno de los problemas más habituales que aparecían al construir un árbol. Cuando el árbol de decisión conseguía una profundidad suficiente, el árbol tendía a memorizar las soluciones en lugar de generalizar el aprendizaje.

Para solucionar esto, se creó el método *Random Forest*, que básicamente consiste en la creación de muchos árboles que trabajen en conjunto.

La forma en la que funciona el *Random Forest* es la siguiente:

- Se seleccionan k columnas de las totales n (será aproximadamente el 66% del conjunto total), y se crea un árbol de decisión con esas características.
- Se crean árboles ("*bootstrap* simple"),
- Se seleccionan cada uno de los diferentes árboles y se les pide que hagan una misma clasificación.
- Se calcularían los votos obtenidos para cada una de las clases y se consideraría la más votada como la clasificación final que obtendríamos de nuestro bosque.

Podemos destacar como principales características del método *Random For*est las siguientes:

- Es uno de los modelos más precisos que encontramos en la actualidad.
- Funciona muy bien en grandes bases de datos.
- Funciona sin borrado de variables, independientemente de la cantidad de variables de entrada que se maneje.
- A medida que avanza la construcción del bosque de árboles, se genera una estimación objetiva interna de la generalización de error.
- Tiene un método eficaz de estimación de datos faltantes.
- Las características citadas anteriormente, se pueden extender también a datos no etiquetados, que conducen a agrupamiento no supervisado, vistas de datos y la detección de valores atípicos.

Algunas de las ventajas y las desventajas más significativas pueden ser:

Ventajas

- El *Random Forest* funciona bien tanto en problemas de clasificación como de regresión.
- Al utilizar múltiples árboles se reduce considerablemente el riesgo de *overfiting*.
- Se mantiene estable con nuevas muestras.

Inconvenientes

- *Random Forest* puede llegar a caer en *overfiting* con la entrada de algunos datos "particulares".
- Tiene un mayor coste que la creación y ejecución de un solo árbol.
- No funciona bien con datasets pequeños.
- Es difícil de interpretar si obtenemos muchos árboles.

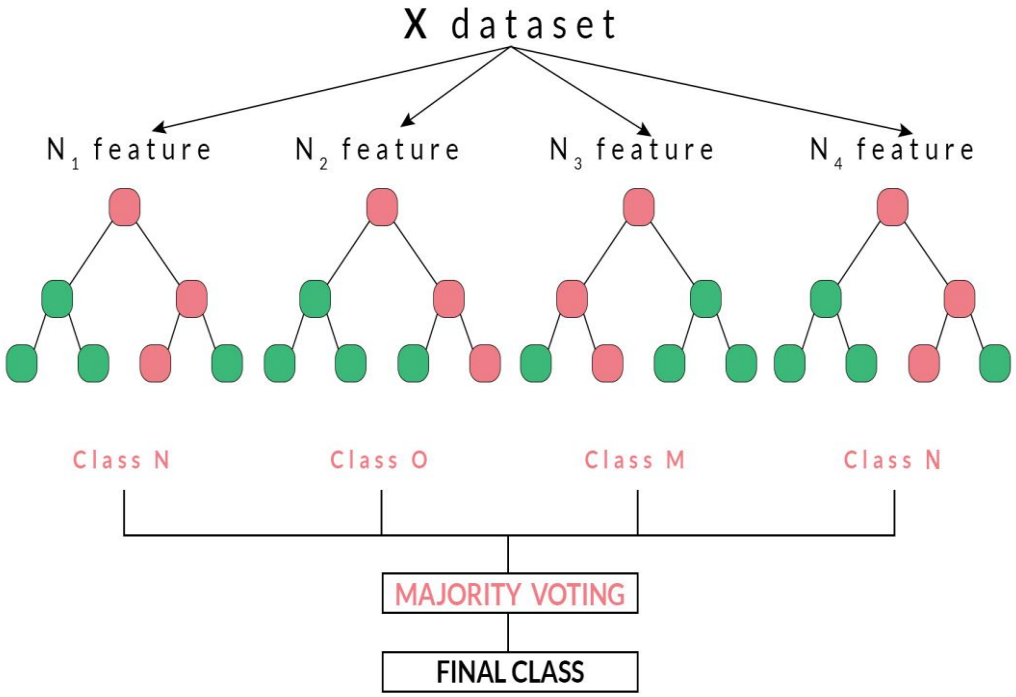


Gráfico 10:Algoritmo Random Forest Fuente: Chen, H.

Por último, para finalizar este capítulo, tenemos que conocer que la implementación del método Random Forest en no es posible si las variables que utilizamos y queremos analizar son cuantitativas

4.3.4 Árboles de decisión vs modelos lineales

Mientras que la regresión lineal asume un modelo con la siguiente forma:

$$f(x) = \beta_0 \sum_{j=1}^{p} X_j \beta_j$$

Los árboles de regresión asumen un modelo con la forma:

$$f(x) = \sum_{m=1}^{M} c_m . 1_{(X \epsilon R_m)}$$

donde $R_1 \dots R_m$ representa la partición del espacio del predictor.

El modelo a utilizar y que previsiblemente nos ofrecerá mejores resultados estará en función del problema en cuestión y de lo lineal que sean los predictores y la variable respuesta.

Si se trata de una relación bastante lineal, nos darán buen resultado los modelos basados en regresión lineal. Si por el contrario no tenemos una relación altamente lineal, los mejores modelos para utilizar son aquellos basados en árboles.

En ambos casos, el rendimiento puede obtenerse con la validación cruzada o la validación simple, aunque podemos tener otras consideraciones en cuenta.

4.3.5 Máquinas de vectores de soporte

Las máquinas de vector de soporte o *Support Vector Machines* (SVM) son un conjunto de algoritmos de aprendizaje supervisados, que están directamente relacionados con problemas de clasificación binaria y regresión.

Este tipo de máquinas son muy utilizadas en aplicaciones como el procesamiento del lenguaje natural, el habla, reconocimiento de imágenes y visión artificial.

La máquina de vector de soporte está dentro de una clase de algoritmos de "*Machine Learning*" a la que se conoce como métodos *kernel*.

Con las SVM se aborda el problema de la clasificación de dos clases, de dos formas diferentes:

Si tenemos un conjunto de muestras, podemos etiquetar las clases y entrenar una Máquina de vectores de soporte para construir un modelo que intente predecir la "clase" de una nueva muestra.

De forma intuitiva, una máquina de vectores de soporte separa las clases en dos espacios distintivos, lo más amplios posibles, mediante un hiperplano, que estará definido mediante un vector entre dos puntos, dentro de las dos clases, a los cuales se les llama "Vector soporte". Si no es posible, tenemos dos opciones:

- Podemos suavizar lo que entendemos por espacios distintivos.
- O, enriquecer y ampliar dicho espacio para lograr al final la separación deseada.

En definitiva, con el hiperplano, se conseguirá una dimensionalidad que puede ser utilizada para problemas de clasificación y regresión.

Si existe una buena separación entre las clases, se podrá lograr una correcta clasificación de clases.

Como en la gran mayoría de los métodos de clasificación supervisada, los datos de entrada se verán como un vector p-dimensional.

La finalidad de la SVM es buscar un hiperplano que nos separe de manera óptima los puntos de una clase de la otra.

Es aquí donde encontramos la mayor característica de las SVM. Este método busca el hiperplano que tenga una distancia mayor con los puntos que estén más cerca de él mismo.

De esta forma obtendremos dos categorías, cada una diferente, y cada una a un lado del hiperplano.

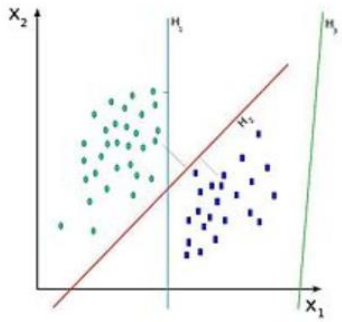


Gráfico 11:Máquinas de vectores de soporte. Fuente: Alba Castro, José Luis

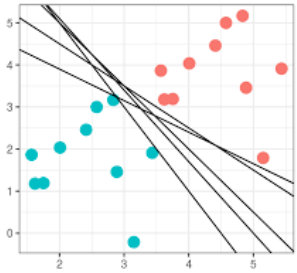


Gráfico 12:Ejemplo de posibles hiperplanos de separación de SVM Fuente: Amat Rodrigo, Joaquín.

En conclusión, podemos decir que los algoritmos SVM pertenecen a la familia de los clasificadores lineales.

Conceptualmente, en el modelo SVM se le llama atributo a la variable predictora y característica a un atributo transformado, que utilizamos para definir el hiperplano.

4.3.6 El vecino más próximo

El algoritmo del vecino más próximo es un método que nos permite clasificar casos basándonos en su similitud a otros.

Dentro del aprendizaje automático, este análisis se desarrolló como una forma de reconocer diferentes patrones de datos sin la necesidad de que coincidieran de forma exacta con los ya almacenados.

Los casos que son parecidos se clasifican como próximos. La distancia que hay entre dos casos se conoce como disimilaridad. Para clasificar adecuadamente los individuos deberemos determinar lo similares o disimilares (divergentes) que son entre sí, en función de lo diferentes que resulten ser sus representaciones en el espacio de las variables.

Los casos próximos entre sí se denominan casos “vecinos”. Cuando se presenta un nuevo caso, denominado caso reserva, se procede a calcular su distancia con respecto a otros casos ya existentes del modelo. Las clasificaciones de los casos más parecidos, lo llamados vecinos más próximos, se cuadran entre ellos, y el nuevo caso se incluye en la categoría que contiene el mayor número de vecinos más próximos.

Con un valor al que se le denomina K, podemos especificar el número de vecinos más próximos que deben examinarse.

También podemos utilizar el método del vecino más próximo para calcular valores para un destino continuo. En esta situación, el valor objetivo medio de los vecinos más próximos se utiliza para obtener el valor citado del nuevo caso.

A continuación, vamos a considerar realizar algunas consideraciones sobre los datos con los que se va a llevar a cabo el análisis:

- Nominal: Se habla de una variable nominal cuando sus valores representan categorías que no obedecen a una clasificación intrínseca. Podemos citar algunos ejemplos de variable nominal como: región, código postal o confesión religiosa.

- Ordinal: Una variable puede considerarse nominal cuando sus valores representan categorías con alguna clasificación intrínseca. Un ejemplo de variable ordinal serían las escales de actitud que representan el grado de satisfacción o confianza y las puntuaciones evalúan las preferencias en un servicio.

- Escalas: Una variable puede tratarse como escala cuando sus valores representan categorías ordenadas con una métrica con significado, por lo que son adecuadas las comparaciones de distancia entre valores. Un ejemplo de variables escala son: la edad en años o los ingresos en dólares.

El análisis de vecinos más próximos se comporta de la misma forma con las variables nominales u ordinales. Con el procedimiento sabemos que se le ha asignado a cada variable el nivel adecuado de medición.

La implementación del vecino más próximo tomando las variables como factores no es posible hacerla en el programa R

5 Aplicación de los modelos estadísticos descritos para la predicción del comportamiento exportador de una empresa.

En este capítulo vamos a profundizar en la descripción de los datos con los que se ha trabajado en este TFG y sobre los que aplicaremos los modelos estadísticos descritos con anterioridad con objeto de obtener resultados que nos permitan realizar el análisis que nos ayude a concluir si existe o no alguna variable web con la que se pueda predecir el comportamiento exportador de las empresas.

5.1 Descripción de la base de datos

En esta sección, vamos a describir la base de datos que se ha utilizado para el estudio. Se trata de una fuente secundaria. La base de datos ha sido obtenida del estudio "Ajuste y Análisis de Modelos Basados en Variables Web para predecir el Comportamiento Exportador Empresarial" que, a su vez, para realizar el estudio se acudió al Sistema de Análisis de Balances Ibéricos (SABI), siendo este fichero adaptado para facilitar su estudio posterior, agrupando diferentes columnas con la misma información en una sola y obviando información que no nos aportaba información útil.

Las empresas utilizadas para el estudio tienen domicilio social en la Comunidad Valenciana, y todas ellas tienen sitio Web corporativo.

En la base de datos podemos encontrar información acerca de 2 variables diferentes: Web manuales y Económicas.

Las variables económicas se han obtenido de la base de datos del SABI, de los directorios de Exportadores del Instituto Español de Exportación (ICEX) y del Consejo Superior de Cámaras de Comercio de España.

En este caso la base de datos objeto de estudios consta de 359 registros referidos a diferentes empresas de la Comunidad Valenciana con 16 datos en cada registro.

5.1.1 Variables económicas

Hemos obtenido estas variables de SABI y del ICEX.

Identificador (Id):

Código con el que se identifica a la empresa en el SABI.

Nombre (Nombre):

Nombre social de la empresa.

Estado (Est.):

Registro que contiene los datos referidos a la situación en la que en 2014 se encontraba la empresa. Puede adoptar los siguientes valores: activa, disuelta, extinguida, en concurso o ilocalizable.

Antigüedad de la empresa (Antigüedad):

Variable continua de tiempo medida en años. Se trata de cuantificar de forma aproximada la experiencia de la empresa. Se ha tomado como punto de referencia inicial el año de fundación de la empresa y para computar la antigüedad total se ha calculado el tiempo transcurrido desde la fecha de referencia inicial a la fecha de referencia final, siendo esta en todos los casos el día 31 de diciembre del año 2013.

Comportamiento Exportador (Exportador):

Variable binaria con dos niveles, los cuales cubren todos los resultados posibles, no existiendo ningún otro resultado viable.

Si ocurre uno de los resultados, es imposible que se dé el otro. En este caso, los valores binarios son 1 y 0.

Si la variable toma el valor 1 significa que la empresa es exportadora, adoptando el valor 0 en caso de que no lo sea.

Es la variable dependiente objeto de estudio de este TFG.

5.1.2 Variables web manuales

Se trata de diferentes variables que se encuentran al estudiar el sitio web.

En concreto, en la base de datos sobre la que se ha trabajado, se encuentran las variables con las características de estudiar la web de las diferentes empresas hasta el 2013 inclusive:

Palabras Clave (PClave):

Variable binaria con valor 1 si el sitio web que se analiza contiene algún término que esté relacionado con comercio y 0 en caso contrario. Con el objetivo de acotar los términos de búsqueda, se confecciono un listado de palabras relacionadas con el comercio[1]. Cada palabra encontrada en la lista, se ha buscado manualmente en los sitios web corporativos, mediante la búsqueda avanzada de Google.

Versión WEB en español (Español):

Variable binaria que toma el valor 1 en los casos en los que el sitio web de la empresa dispone de una versión actualizada y con todas sus opciones activas en español. En caso contrario, adopta el valor 0.

Versión WEB en inglés (Inglés):

Variable binaria que toma el valor 1 en los casos en los que el sitio web de la empresa dispone de una versión actualizada y con todas sus opciones activas en inglés. En caso contrario, adopta el valor 0.

Versión WEB en francés (Francés):

Variable binaria que toma el valor 1 en los casos en los que el sitio web de la empresa dispone de una versión actualizada y con todas sus opciones activas en francés. En caso contrario, adopta el valor 0.

Versión WEB en alemán (Alemán):

Variable binaria que toma el valor 1 en los casos en los que el sitio web de la empresa dispone de una versión actualizada y con todas sus opciones activas en alemán. En caso contrario, adopta el valor 0.

Versión WEB en italiano (Italiano):

Variable binaria que toma el valor 1 en los casos en los que el sitio web de la empresa dispone de una versión actualizada y con todas sus opciones activas en italiano. En caso contrario, adopta el valor 0.

[1] En la lista encontramos: continental; continente; continentes; export; exporta; exportación; exportaciones; exportamos; exportando; exporter; extranjero; globalización; internacional; internacionales; internacionalización; mundial; países.

Facebook Activo (Facebookactivo13):

Variable binaria que toma el valor 1 en los casos en los que a fecha 31 de diciembre de 2013 la empresa disponía de registro activo en la red social Facebook. En caso contrario, adopta el valor 0.

Twitter Activo 2013 (Twitteractivo13):

Variable binaria que toma el valor 1 en los casos en los que a fecha 31 de diciembre de 2013 la empresa disponía de registro activo en la red social Twitter. En caso contrario, adopta el valor 0.

Dirección Web:

Dirección Web de la empresa sobre la que se han obtenido los datos reflejados en las variables objeto de estudio.

Dominio:

Denominación que identifica al sitio web, que indica que pertenece a una categoría determinada.

5.1.2.1 Justificación

Como dijimos anteriormente, la base de datos fue modificada para facilitar el estudio y la obtención de conclusiones. Las variables que seleccionamos al final, son las definidas con anterioridad.

La razón por la que hemos elegido estas variables es el potencial que puede llegar a tener su relación con el comportamiento exportador de las empresas.

- **Palabras clave:** Mediante el marketing digital, las empresas están cada vez más actualizadas en la web. En la actualidad, si una empresa no está al día tecnológicamente hablando, puede quedar retrasada en relación a sus competidores. Es por esto que, a través de las páginas web de las empresas, cada vez es más común encontrar información que nos ayuda mediante el estudio de palabras clave, a saber, cuál es el carácter exportador que tiene una empresa determinada. En este caso, tenemos la existencia de una lista que contiene palabras que estarían en la web corporativa de una empresa cuyo carácter es más bien exportador.

- **Versión de la web en [idioma]:** Para lograr una buena posición en el mercado internacional, lo primero es estar preparado.

Un sitio web en otro idioma siempre ayuda en la estrategia de penetración a un nuevo país en el que podría la empresa llegar a alcanzar su mercado objetivo. Además, la comunicación con clientes y proveedores siempre será más fluida, y los clientes logran sentirse más cómodos y seguros si pueden consultar información sobre la empresa con la que quieren comunicarse, y el idioma en el que lo pueden hacer les resulta familiar.

Esto nos hace pensar que existe una relación entre que la web esté disponible en varios idiomas y que dicha empresa pueda operar en mercados extranjeros.

En el trabajo se estudia la relación que puede tener la versión de la web en español, francés, inglés, alemán e italiano, descartando el español, ya que como hemos dicho con anterioridad, la base de datos consta de empresas españolas, El idioma que parece más obvio es el inglés para las empresas exportadoras localizadas en países de habla no inglesa, ya que de todos estos es el idioma más hablado y el más utilizado en el mundo empresarial para la comunicación entre empresas que no comparten un mismo idioma.

- **Facebook/Twitter activo:** En la era digital, es cada vez más importante la comunicación a través de las redes sociales.

  Las empresas que se han sabido adaptar a esta nueva era y han podido identificar tanto nuevas amenazas como oportunidades de negocio, como por ejemplo entrar en un mercado extranjero, suelen ser aquellas que han desarrollado métodos de relación y captación de clientes, muchas veces desde las redes sociales.

  Las redes sociales, hoy en día se consideran un canal más de comunicación e iteración de las empresas, tanto con clientes, y también de manera interna.

5.2 Dependencia entre variables

Para el análisis de dependencias entre variables, se empezó por analizar, la dependencia por pares.

En primer lugar, se analizaron las dependencias entre variables web manuales, de esta forma se podía detectar una posible multicolinealidad.

Variable	Inglés	Pclave	facebookactivo13
Inglés	1	-	-
Pclave	0,37051469	1	-
facebookactivo13	0,06057158	0,23515128	1

Tabla 3: Dependencias entre variables, coeficiente de correlación. Fuente: Elaboración propia.

Posteriormente hicimos lo mismo con las dependencias entre las variables económicas independientes.

La multicolinealidad es una condición que sucede cuando algunas variables predictoras del modelo, están correlacionadas con otras variables predictoras.

La multicolinealidad puede llegar a ser problemática, ya que puede incrementar la varianza de los coeficientes.

En este caso, las variables no tienen una fuerte relación entre ellas, por lo tanto, podemos decir que el riesgo de que exista multicolinealidad es despreciable.

5.3 Proceso seguido en el trabajo y resultados según método

En este capítulo explicaremos el proceso que hemos seguido para la obtención de resultado mediante diversos métodos utilizando la interface RStudio del software estadístico R y los resultados que podemos extraer de cada uno de ellos.

En primer lugar, tenemos que cargar y leer la base de datos que utilizaremos, y las diferentes librerías que nos serán útiles.

En este caso cargamos un fichero Excel llamado “abreviado”, donde encontramos la base de datos modificada con la que trabajaremos.

```
library(rpart)
library(DMwR2)
library(randomForest)
library(kknn)
library(e1071)
library(kernlab)
library(rpart.plot)
library(readxl)


datos<-read_excel("abreviado.xlsx")
```

Gráfico 13: Librerías utilizadas en el código. Fuente: Elaboración propia, desde el software R.

```
datos$Exportador<-as.factor(datos$Exportador)
datos$PClave<-as.factor(datos$PClave)
datos$esp<-as.factor(datos$esp)
datos$Ingles<-as.factor(datos$Ingles)
datos$fr<-as.factor(datos$fr)
datos$de<-as.factor(datos$de)
datos$it<-as.factor(datos$it)
datos$facebookactivo13<-as.factor(datos$facebookactivo13)
datos$twitteractivo13<-as.factor(datos$twitteractivo13)
summary(datos)
```

Gráfico 14: Variables estudiadas como factores. Fuente Elaboración propia.

Posteriormente obtendremos un resumen de la base de datos, mediante la función "summary" y hacer que el software entienda nuestras variables binarias como factor:

Edad		Exportador		Pclave		esp		Ingles		fr	
Min	3.395	0	183	0	202	0	7	0	214	0	291
1st Qu.	13.542	1	175	1	156	1	351	1	144	1	67
Median	20.670										
Mean	21.446	de		it		facebookactivo13		twitteractivo13		Dominio	
3rd Qu.	27.697	0	344	0	351	0	319	0	338	Length	358
Max	87.058	1	14	1	7	1	39	1	20	Class	character
										Mode	character

Tabla 4: Resumen de los datos de la base de datos estudiada. Fuente: Elaboración propia.

Ya que:

- **Exportador** es una variable binaria, 0 no exportadora, 1 exportadora.
- **PClave** es una variable binaria 0 no existen palabras clave, 1 si existen.
- **esp** es una variable binaria, 0 la web no está en español, 1 sí que lo está.
- **Inglés** es una variable binaria, 0 la web no está en inglés, 1 sí que lo está.
- **fr** es una variable binaria, 0 la web no está en francés, 1 sí que lo está.
- **de** es una variable binaria, 0 la web no está en alemán, 1 sí que lo está.
- **it** es una variable binaria, 0 la web no está en italiano, 1 sí que lo está.
- **twitteractivo13** es una variable binaria, 0 la empresa no tiene twitter activo 1 sí que lo tiene.
- **facebookactivo13** es una variable binaria, 0 la empresa no tiene Facebook activo 1 sí que lo tiene.

Una vez explicado esto, pasaremos a ver los modelos utilizados:

5.3.1 Árbol de Clasificación

En primer lugar, debemos entrenar a nuestro modelo, y para ello utilizaremos la función rpart. Esta función nos pide una fórmula para especificar la variable objetivo de la clasificación.

La fórmula que hemos utilizado en nuestro código ha sido del tipo **Exportador~.**, esto nos indica que vamos a intentar clasificar **Exportador** usando las demás variables como predictoras.

El cp, que corresponde al parámetro de complejidad utilizado debe ser un valor pequeño, no negativo que esté cercano a 0. En este caso hemos utilizado cp=0.001

```
treefull <- rpart(Exportador~.,data=datos, method="class", cp=0.001)
```

A continuación, mediante la función plotcp, sacaremos la gráfica que nos muestra el conjunto de posibles podas de complejidad de costes de un árbol de un conjunto nidificado.

Para los medios geométricos de los intervalos de valores de los cp para los que una poda es óptima en la construcción inicial, se ha hecho una validación cruzada rpart.

La imagen que vemos a continuación, obtenida mediante la función **plotcp** contiene el ajuste de la media y la desviación estándar de los errores en la predicción de cada una de

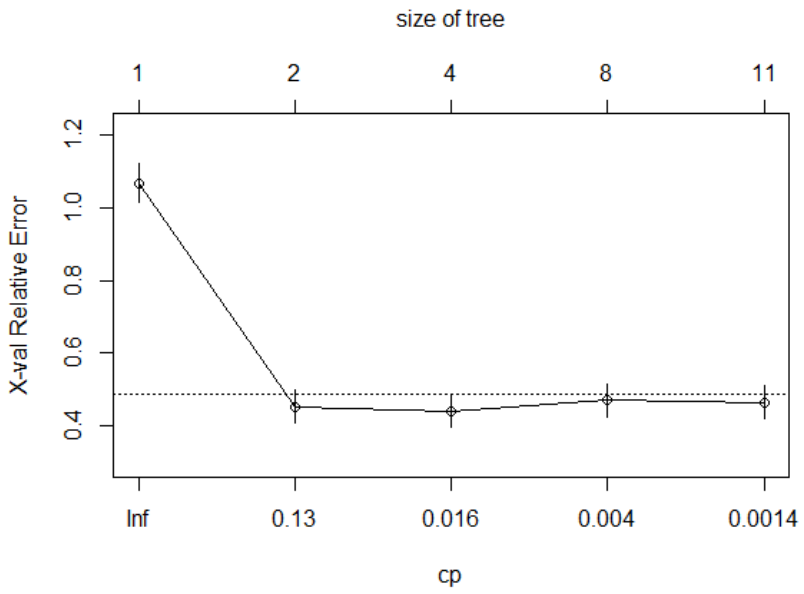


Gráfico 15: Gráfico número de árboles posibles para la poda. Fuente Elaboración propia.

las medias geométricas. Una buena opción **cp** para podar es normalmente el valor más izquierdo para el que la media se sitúa por debajo de la línea horizontal.

A continuación, y para seguir con nuestro modelo, y después de haber obtenido la tabla 5:

	CP	nsplit	rel error	xerror	xstd
1	0.5485714	0	1.00000	1.08000	0.053975
2	0.0285714	1	0.45143	0.45143	0.044837
3	0.0085714	3	0.39429	0.45143	0.044837
4	0.0019048	7	0.36000	0.46857	0.045434
5	0.0010000	10	0.35429	0.48571	0.046006

Tabla 5: Posibles árboles para la poda : Fuente: Elaboración propia.

Observando la tabla 5, cada una de las filas es un posible árbol que podemos construir, dependiendo de si queremos que el árbol que obtengamos nos sirva solamente para nuestros datos en concreto o no.

En este caso, podemos decir que el árbol inicial tiene 10 particiones, ya que es el valor más alto que podemos encontrar en la columna **"nsplit",** (este sería el número de particiones que se haces ya que se trataría de un árbol sin podar) y que se trataría de un árbol estrictamente construido para la base de datos concreta que nosotros le hemos pasado.

La primera columna "**CP**" se trata de la "Capacidad del proceso". La capacidad es la que tiene el proceso para producir piezas de acuerdo con las especificaciones, es decir, dentro de los límites establecidos. Para evaluar la capacidad de un proceso, es necesario contar con suficientes muestras, es por esto que el cálculo del CP se encuadra en el marco de un estudio estadístico.

La segunda columna **nsplit** se trata del número de niveles de la estructura del árbol.

La tercera donde encontramos el **error relativo** que es el cociente entre el error absoluto y el valor que se considera exacto (la media). Puede ser que este error sea positivo o negativo, ya que se podría producir por exceso o por defecto.

La cuarta columna, donde vemos **el xerror**, que se trata de mirar el error estimado para la validación cruzada.

Y por último la columna **xstd**, donde encontramos la desviación estándar del error estimado para la validación cruzada

Para poder podar el árbol, tenemos que tener en cuenta la columna xerror, el error de precisión, y escoger la fila que contenga el menor de ellos.

En este caso elegiríamos el 2 o el 3, como tienen el mismo xerror, nos quedaremos con el 3, cuyo cp = 0.0085714, ya que la partición sería menor (nsplit=3).

Una vez seleccionado el **cp**, utilizaremos la función:

```
tree1<-prune.rpart(treefull,0.0285714)#cp del arbol con min xerror
prp(tree1)
```

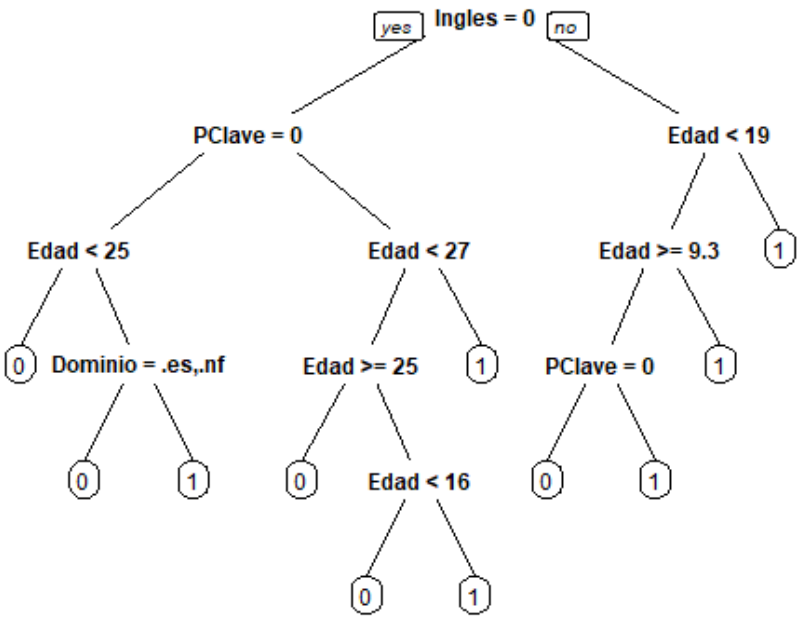


Gráfico 16: Árbol de Regresión Fuente Elaboración propia.

El árbol anteriormente mostrado en la Figura 16 se interpreta de la siguiente forma. Para realizar la primera partición, partimos del nodo inicial que contiene todas las empresas de la base de datos que vamos a analizar, y se comprueba si la variable "inglés" es igual a 0, teniendo en cuenta si las empresas son o no exportadoras.

La variable utilizada para ramificar por primera vez ha sido "inglés". El modelo ha utilizado esta variable ya que es la que genera una ganancia predictiva mayor al dividir.

Las empresas que no cuentan con sitio web en inglés se encuentran en la rama de la derecha mientras que aquellas empresas que sí que cuentan con web en inglés se clasificarán a la izquierda.

A partir del nodo de la derecha, la variable "Edad < 19" genera otra partición. En este caso, la partición de la derecha engloba aquellas empresas cuya edad es menor a 19 años y la rama de la derecha, aquellas cuya edad es mayor. En cuanto a la partición de la izquierda, acoge todas aquellas empresas cuya edad es mayor de 19 años. Dentro de esta rama, podemos encontrar otra partición, en la que se verá la segmentación dependiendo de si existen o no palabras clave (PClave).

En cuanto a la partición de la izquierda desde el nodo inicial y donde se encontrarían las empresas exportadoras, vemos que se genera una partición nueva, dependiendo de la variable "PClave", si la empresa es exportadora y existen palabras clave generaríamos otra partición dependiendo de la edad de la empresa "Edad <25" y si esto se cumpliera también habría otra partición "Dominio = .es, .info".

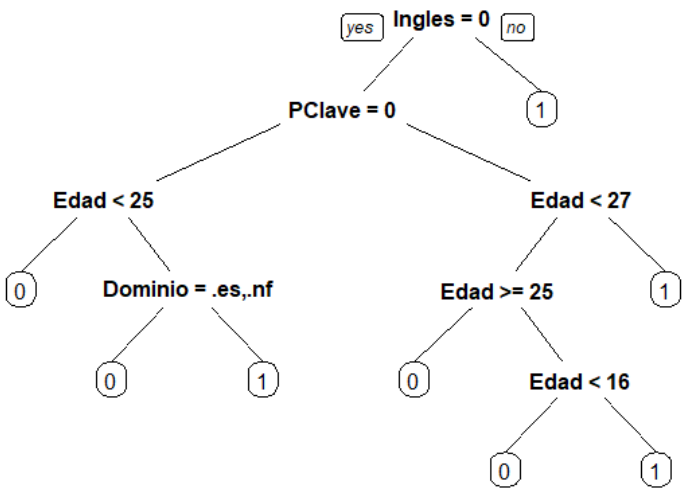


Gráfico 17: Árbol regresión segunda poda Fuente: Elaboración propia.

En cuanto a la capacidad predictiva del modelo generado, la tasa de acierto obtenida es 82,4%, por lo que podemos decir que este modelo predice correctamente el 82,4% de los casos.

		Comportamiento exportador real	
		0	1
Comportamiento exportador predicho	0	140	20
	1	43	155

Tabla 6: Capacidad predictiva del modelo Fuente: Elaboración propia.

5.3.2 Máquinas de vectores de soporte

Las máquinas de vectores de soporte son una herramienta excelente para la clasificación, detección de novedades y regresión.

La función ksvm del software R, consta de:

ksvm(x, data = NULL, ..., subset, na.action = na.omit, scaled = TRUE)

Comenzamos pidiéndole al programa un resumen de la clase "Ksvm", del que podemos concluir que es de largaría 1, y que el modo que utilizamos es el s4. Una clase S4 contiene la salida del ksvm.

Sumary(svp)

```
Length  Class   Mode
     1   ksvm     S4
```

Para la capacidad predictiva del modelo podemos ver que el modelo predice correctamente el 82,12% de las veces.

Tasa de acierto

0.8212291

Tasa de fallo:

0.1787709

Por último, si aplicamos el modelo svm, para estudiar la variable "Exportadora", obtenemos un "summary" de los datos:

SVM-Type: C-classification

SVM-Kernel: radial

Coste: 10

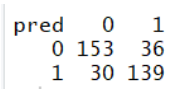
Gama: 0.1

Y el número de soportes vectoriales son 171, separados en dos espacios:

(90-81).

Obteniendo una buena separación entre las clases, nos permitirá hacer una correcta clasificación.

Con este modelo, con un dato nuevo, podemos predecir, a que categoría pertenece, a aquella que es exportadora o a la que no lo es.



pred	0	1
0	153	36
1	30	139

La tasa de acierto del SVM es de 81,56% esto quiere decir que el 81,56% de las veces nos da un resultado correcto, por consecuente la tasa de fallo sería de 18,43%

6 Conclusiones y próximos pasos

Es en este capítulo donde vamos a sintetizar las principales conclusiones obtenidas a partir de todo el trabajo realizado y cuáles podrían ser los posibles campos o futuros trabajos a estudiar

6.1 Conclusiones

Para poder sacar conclusiones y predecir diferentes variables en la economía de la manera lo más actualizadas posible, es necesario obtener la información de la forma más rápida y lo más "in time" posible.

Tradicionalmente, los datos que nos da el gobierno de España son fiables, pero se proporcionan con un destiempo de 3-4 meses, por lo que no nos sirven si el objetivo es predecir lo más inmediato posible al momento real, o incluso anticiparse.

Es por ello que los avances que encontramos en las Tecnologías de la Información y la Comunicación, principalmente en el esfuerzo de las empresas por tener actualizadas las diferentes webs corporativas, nos facilita el trabajo de la obtención de dichos datos en tiempo real.

Este trabajo, se ha centrado en:

- Mostrar de qué manera es posible predecir el comportamiento exportador de las empresas de una economía, mediante la información que obtenemos desde sus páginas webs corporativas.

Para hacerlo, se han construido varios modelos de predicción con variables basadas en la web, obtenidas mediante un fichero .xlsx como base de datos secundaria, que posteriormente modificaríamos para facilitar su estudio, donde se muestra información del sitio web corporativo de 358 empresas.

En concreto los modelos utilizados han sido un árbol de clasificación, Random Forest, el vecino más próximo y las máquinas de soporte vectorial.

De estos modelos estudiados, tanto el Random Forest por problemas de la implementación en RStudio, como el vecino más próximo, por problemas del método, tuvieron que ser desechados, ya que no admitían algunas de las variables, al usar los datos como factor.

Con los modelos estudiados, se ha podido demostrar que sí que existen algunas características de los sitios web corporativos que pueden hacernos deducir el comportamiento exportador o no de una empresa.

Lo tasa de acierto de los modelos es > 80% que, si bien no podemos decir que sea del todo fiable, es un valor importante.

Por lo tanto, podemos decir que la información que encontramos en las webs corporativas de las empresas nos proporcionan una fuente barata, fiable y actualizada para algún posible estudio que nos pueda interesar, y en concreto sobre el comportamiento exportador de las empresas de forma individual. También, los indicadores web, pueden funcionar como información adicional o complementar otro tipo de fuentes.

Por otro lado, al haber demostrado que hay variables que obtenemos de los sitios web a partir de las cuales, hemos podido construir un indicador del comportamiento exportador de las empresas, se ha encontrado una forma de hacerlo más sencilla y barata.

6.2 Próximos pasos

En cuanto a los próximos pasos previstos, la línea del tiempo de los mismos se encuentra a un año vista. La idea es mejorar la obtención de datos y hacerlo de una forma más rápida y actualizada.

En estos momentos me encuentro realizando el Grado en Ingeniería Informática y en la realización de este TFG, la idea es conseguir la información de la web desde la URL de los diferentes sitios web corporativos, con el programa “crawler” y la técnica “scrapping”, que se trata de una técnica para extraer información de los sitios web corporativos.

En resumen, este programa simula lo que sería la exploración humana de la World Wide Web. De esta manera obtendríamos un modelo con los datos actualizados, en el momento del análisis y ver como evoluciona en el tiempo.

Bibliografía

Amat Rodrígo, Joaquin, "Correlación lineal y regression lineal simple". <https://rpubs.com/Joaquin_AR/223351 (4,2)>. Jun 2016.

Bagnato, Juan Ignacio, "Aprende Machine Learning antes de que sea demasiado tarde". <https://www.aprendemachinelearning.com/que-es-overfitting-y-underfitting-y-como-solucionarlo/ (4.4.2)>.

Bagnato, Juan Ignacio, "Random Forest, el poder del ensamble". <https://www.aprendemachinelearning.com/random-forest-el-poder-del-ensamble/

Barud, Samia, "¿Qué hacer si tu público habla varios idiomas en Facebook?" https://www.agorapulse.com/es/blog/varios-idiomas-en-facebook. Octubre 2017.

Blázquez Soriano, María Amparo, "Ajuste y Análisis de Modelos Basados en Variables Web para Predecir el Comportamiento Exportador Empresarial" (2015).

Bravo, Andreu, "La importancia de las redes sociales en las empresas y su gestión" <https://www2.deloitte.com/es/es/pages/governance-risk-and-compliance/articles/importancia-redes-sociales-empresas-gestion.html>.

Carmona Suárez, Enrique J., "Tutorial sobre Máquinas de Vectores Soporte (SVM)" 11. <http://www.ia.uned.es/~ejcarmona/publicaciones/[2013-Carmona]%20SVM.pdf>. Julio 2014.

"Comercio exterior", < https://www.importancia.org/comercio-exterior.php>, Mar 2013.

Delgado, Hugo, "World Wide Web – WWW significado, historia y origen", <https://disenowebakus.net/world-wide-web-www.php (2.3)>. Nov 2018.

Durban, María, “Modelos Lineales Generalizados”, <
http://halweb.uc3m.es/esp/Personal/personas/durban/esp/web/GLM/curso_GLM.pdf.>

Ecured, “Bibliometría”. <https://www.ecured.cu/Bibliometr%C3%ADa (3)>.

“El comercio exterior y su importancia”. <http://informaciocomercioexterior.blogspot.com/2012/06/el-comercio-exterior-y-su-importancia.html (2.1)> Jun 2012.

El vigia, “España cierra 2018 con un aumento de las exportaciones del 3%”, <http://elvigia.com/espana-cierra-2018-con-un-aumento-de-las-exportaciones-del-3/ (2.1.1)> Feb 2019.

Gil, Cristina, “RPubs brought to you by RStudio”. <https://rpubs.com/Cristinina_Gil/arboles_ensemble (2)>.

“La importancia de las redes sociales en el marketing digital”. <http://comunicacionyproyeccion.com/blog/2018/09/20/redes-sociales-marketing/>. Septiembre 2018.

Marta, “Correlación”. <https://www.superprof.es/apuntes/escolar/matematicas/estadistica/disbidimension/correlacion.html#tema_grado-de-correlacion>.

Mathworks, “Algoritmos de Machine Learning para clasificación (svm)”.<https://la.mathworks.com/discovery/support-vector-machine.html (4,6)>

McCullagh,P y Nelder, J. A. “Generalized Linear Models”1992. <https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/regression/supporting-topics/logistic-regression/what-is-a-generalized-linear-model/#fntarg_1>.

Mendoza Vega, Juan "Árboles de decision con R-Clasificación". <https://rpubs.com/jboscomendoza/arboles_decision_clasificacion>. Abr 2018.

Ministerio de industria comercio y turismo, <https://www.mincotur.gob.es/es-ES/Paginas/index.aspx>.

"Model training and tuning". <https://topepo.github.io/caret/model-training-and-tuning.html (2)>.

Na8, "Random Forest, el poder del ensamble" <https://www.aprendemachinelearning.com/random-forest-el-poder-del-ensamble/ (4.7.4)>. Jun 2017.

Nicole Roldán, Paula. "El comercio exterior consiste en el intercambio de bienes y servicios entre dos o más países". <https://economipedia.com/definiciones/comercio-exterior.html (2.1)>, oct. 2018.

Numerictron, "Regresión por mínimos cuadrados lineal y numérica". <

https://sites.google.com/site/numerictron/unidad-4/4-3-regresion-por-minimos-cuadrados-lineal-y-cuadratica>.

OEC "The Observatory of Economic Complexity". <https://oec.world/es/profile/country/esp/>.

Ordoñez, Laia, "La importancia de Hreflang en la estrategia de exportación", https://www.oleoshop.com/blog/la-importancia-del-hreflang-en-la-estrategia-de-exportacion. Agosto 2016.

Parra, Francisco, "Estadística y Machine Learning con R" <https://rstudio-pubs-static.s3.amazonaws.com/305959_ad667593ada14c0d892f902eaa2b7fc8.html>. Junio 2017.

Piccini Juan, "Árboles de Clasificación y Regresión basados en atributos funcionales y su utilización en el contexto de procesos epidémicos". <http://premat.fing.edu.uy/ingenieriamatematica/archivos/tesis_juan_piccini.pdf>. Sept 2009.

R Core Team. "R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria". <https://www.R-project.org/>. 2018

Rios, Andrés, "Máquinas de vectores de soporte (clasificación y regression)". <https://platzi.com/tutoriales/1180-redes-neuronales/2677-maquinas-de-vectores-de-soporte-clasificacion-y-regresion/>. 2018.

Rivera Sánchez, Claudia, "Un acercamiento a la Webmetría",

<http://www.infotecarios.com/un-acercamiento-la-webmetria/#.XROiJOgzY2w (3)>. Abr 2015.

Romero-Frias, Esteban, "El empleo de la Webmetría para el análisis de los indicadores de desempeño y posición financiera de la empresa" <https://revistas.unal.edu.co/index.php/innovar/article/view/48993/50688 (2.3)>. Jun, 2013

RStudio Team. "RStudio: Integrated Development for R. RStudio, Inc., Boston, MA"<http://www.rstudio.com/> 2016

Sancho Caparrini, Fernando, "Clasificación Supervisada y no supervisada". <http://www.cs.us.es/~fsancho/?e=77 (4,7,1)>. Dic 2018

Shagufta Tahsildar, "Random Forest Algorithm in trading using Python". <https://d1rwhvwstyk9gu.cloudfront.net/2019/03/Random-Forest-Algorithm.jpg (4,7,5)>. 12 marzo 2019.

Universitat de Valencia, “Árboles de clasificación y regresión” <https://www.uv.es/mlejarza/actuariales/tam/arbolesdecision.pdf (2.4.1)>.

ANEXO 1- Código completo RStudio

```
library(rpart)
library(DMwR2)
library(randomForest)
library(kknn)
library(e1071)
library(kernlab)
library(rpart.plot)
library(readxl)


datos<-read_excel("abreviado.xlsx")
summary(datos)

datos$Exportador<-as.factor(datos$Exportador)
datos$PClave<-as.factor(datos$PClave)
datos$esp<-as.factor(datos$esp)
datos$Ingles<-as.factor(datos$Ingles)
datos$fr<-as.factor(datos$fr)
datos$de<-as.factor(datos$de)
datos$it<-as.factor(datos$it)
datos$facebookactivo13<-as.factor(datos$facebookactivo13)
datos$twitteractivo13<-as.factor(datos$twitteractivo13)
summary(datos)
#Exportador es una variable binaria, 0 no exportadora, 1 exportadora
#Inglés es una variable binaria, 0 la web no está en inglés, 1 si que lo está
#facebook/twitter variable binaria que nos dice 0 no está activo 1 si que lo está
datos<-datos[,4:14]


      ################
      #ARBOL
treefull <- rpart(Exportador~.,data=datos, method="class", cp=0.001)
plotcp(treefull)
printcp(treefull)
prp(treefull)
tree1<-prune.rpart(treefull,0.0285714)#cp del arbol con min xerror
prp(tree1)
tree1<-prune.rpart(treefull,0.0085714)#cp del arbol con min xerror
prp(tree1)
#podar el arbol con la regla SE
tree2<-rpartXse(Exportador~.,data=datos, se=0.5,model=TRUE)
prp(tree2)
pred <- predict(tree1, datos[,-2])
#para predecir hemos de quitar la exportación que ocupa el lugar 2
dim(pred)# predice la prob de 0 y de 1

clase<-pred[,2]
clase[pred[,2]>0.5]<-1 #este punto de corte puede variarse como luego veremos en la ROC
clase[pred[,2]<=0.5]<-0
#matriz de confusión
tab <- table(clase, datos$Exportador)
#tabla cruzada de predicciones y clasificación real, MATRIZ DE CONFUSION
tab
#tasa de acierto
sum(tab[row(tab)==col(tab)])/sum(tab)
```

```
#obtener la tasa de fallo
sum(tab[row(tab)!=col(tab)])/sum(tab)

    ###############
    #RAMDOMFOREST
dim(datos)

head(datos)
#notar que el random forest he de ponerle que es factor la variable dep
rf.Exportador<-randomForest(factor(Exportador)~ Edad, data=datos,mtry=3,method="class",
#mtry=3 raiz(10) donde 3 es el num de var indep
print(rf.Exportador)   #en este caso no se puede poner factores como var indep.
importance(rf.Exportador)
# Plot variable importance
varImpPlot(rf.Exportador, main="",col="dark blue")

                          ###############
                            #ELVECINOMASPROXIMO


library(class)
vecino<-knn(datos[,-2], factor(datos$Exportador), k = 3, prob = TRUE)
vecino.cv<-knn.cv(datos[,-2], factor(datos$Exportador), k = 3, prob = TRUE)
summary(vecino)
summary(vecino.cv)

#matriz de confusion
#pred <- predict(vecino, datos[,-2]), no podemos aplicar la funci?n predict
tab<-table(factor(datos$Exportador),vecino[1:354])
sum(tab[row(tab)==col(tab)])/sum(tab)
tab<-table(factor(datos$Exportador),vecino.cv[1:354])
sum(tab[row(tab)==col(tab)])/sum(tab)
```

```
#algoritmo de b?squeda de vecinos
library(FNN)
vecino.get<-get.knn(datos[,-6], k=3, algorithm=c("kd_tree", "cover_tree", "CR", "brute"))
vecino.get<-get.knn(datos[,-6], k=3, algorithm=c("CR"))
#vecino.get$nn.index, proporciona los indices de los tres vecinos m?s pr?ximos


                        ###############
                          #SMV


svp <- ksvm(factor(Exportador)~ ., data=datos, type = "C-svc", kernel = "rbfdot",kpar = "automatic")
summary(svp)
pred <- predict(svp, datos[,-2])
tab <- table(pred, datos$Exportador)
#obtener la tasa de acierto
sum(tab[row(tab)==col(tab)])/sum(tab)
#obtener la tasa de fallo
sum(tab[row(tab)!=col(tab)])/sum(tab)
model <- svm(factor(Exportador)~ ., data = datos, method = "C-classification", kernel = "radial",cost = 10, gamma = 0.1)
summary(model)
summary(datos)
#plot(model, datos, Exportador ~ .)
pred <- predict(model, datos[,-2])
tab <- table(pred, datos$Exportador)
tab
#obtener la tasa de acierto
sum(tab[row(tab)==col(tab)])/sum(tab)
sum(tab[row(tab)!=col(tab)])/sum(tab)
citation()
```