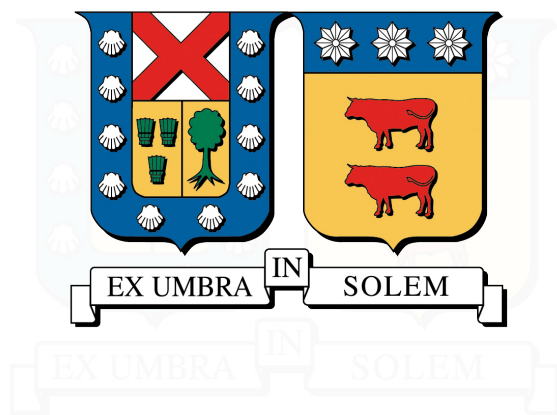


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INDUSTRIAS
VALPARAÍSO - CHILE



**MÉTODOS DE ANÁLISIS PARA BIG DATA Y SU PARTICIPACIÓN EN LA
INDUSTRIA:
ESTUDIO APLICADO A LA PREVENCIÓN DE FALLOS EN EMPRESAS
FERROVIARIAS**

TOMÁS MARCHANT MORALES

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL

PROFESOR GUÍA : DRA. MÓNICA LOPEZ.
PROFESOR CORREFERENTE : DR. PABLO ESCALONA.

JULIO 2018



A mi familia . . .

“You can’t manage what you dont measure.”

-Peter Drucker.

AGRADECIMIENTOS

Para comenzar deseo agradecer a mi profesora guía, Mónica López por la buena disposición y el genuino interés que ha mostrado por el avance de este trabajo. En adhesión a esto su buen humor y las palabras de alientos también motivaron a que la realización de esta tesis fuera de la manera más prolija posible. De la misma forma incluyo al profesor Pablo Escalona por sus correcciones y opiniones sobre el tema. Estas me motivaron para estudiar y aprender más.

Dedico este trabajo a mi familia, que siempre ha estado preocupada por mi avance curricular y también por mi formación como persona. Deseo incluir también a mis amigos, tanto a los que me aconsejaron concentrarme, como a los que ayudaron a distraerme. Todos ellos hicieron que este periodo fuera muy grato y muy bien disfrutado. Son muchos para ser nombrados, y no quisiera caer en comparaciones, pero sin duda son lo que mejor me llevo de la universidad.

RESUMEN EJECUTIVO

Big Data se refiere a una base de datos muy grande para ser procesada por los métodos comunes de lectura de datos tal que se necesita un sistema de computación distribuida, o bien se define simplemente como una base de datos de gran volumen, velocidad y variedad. Existen varias herramientas para extraer información de estas. Como las cartas de control, Visualización de datos, Hadoop y la minería de datos. De esta última se usan herramientas conocidas como algoritmos de aprendizajes, la cual se aplicarán a una base de datos que se emulará como Big Data para realizar un modelo predictivo que permita minimizar las fallas de asignación de carga en una empresa ferroviaria. Esta base de datos proviene de lecturas reales provisto por el Instituto de Investigación Operativa y Ciencias de la Gestión, llegando a sumar más de 10 millones de muestras.

Dentro de este estudio también se destaca la influencia que ha tenido Big Data dentro de el área de la ingeniería industrial, perfeccionando el área de Supply Chain Management, Finanzas y empresas de retail. Tal ha sido la consideración de las empresas por esta nueva tecnología que las inversiones sobrepasan los 30 billones de dólares en Estados Unidos sólo para el año 2017. Esto corresponde a un incremento del 300 % con respecto al año 2014. Por ello es menester y objetivo de este estudio concientizar del recurso inexplorado de esta área en Chile.

Las conclusiones obtenidas del caso aplicado muestran que se obtiene un modelo categorizado como bueno y/o aceptable para la predicción de datos. Teniendo en cuenta que la distribución de los datos es altamente desequilibrada estos resultados se deben interpretar según un criterio evaluador que se rija por los costos asociados a fallas, y no simplemente a una categorización generalizada de métricas.



ABSTRACT

Big Data refers to a very large database to be processed by common methods of reading data such that a distributed computing system is needed, or it is simply defined as a database of great volume, speed and variety. There are several tools to extract information from these databases. Such as control charts, data visualization, Hadoop and data mining. The latter uses tools known as learning algorithms, which will be applied to a database that will be emulated as Big Data to make a predictive model that minimizes load allocation failures in a railway company. This database comes from real readings provided by the Institute of Operations Research and Management Sciences, reaching more than 10 million samples

This study also highlights the influence that Big Data has had within the area of industrial engineering, perfecting the area of Supply Chain Management, Finance and retail companies. Such has been the consideration of the companies for this new technology that the investments surpass the 30 trillions of dollars in the United States only for the year 2017. This corresponds to an increase of 300 % with respect to the year 2014. Therefore it is necessary and objective of this study to raise awareness of the unexploited resource of this area in Chile.

The conclusions obtained from the applied case show that a model categorized as good and / or acceptable for the prediction of data is obtained. Bearing in mind that the distribution of data is highly unbalanced, these results should be interpreted according to an evaluating criterion that is governed by the costs associated with failures, and not simply a generalized categorization of metrics.



Índice de Contenidos

1. Problema de investigación	1
1.1. Descripción de la problemática	1
1.2. Contexto y antecedentes preliminares	2
2. Objetivos	5
2.1. Objetivo General	5
2.2. Objetivos Específicos	5
3. Metodología	7
4. Marco Teórico	11
4.1. Antecedentes	11
4.2. ¿Qué es Big Data?	12
4.3. El valor del Big Data	13
4.4. Herramienta de análisis para Big Data	14
4.4.1. Inteligencia Artificial	15
4.4.2. Cartas de control	16
4.4.3. Minería de datos	20
4.4.4. Machine Learning	21
4.4.4.1. Formas de aprendizaje	22
4.4.4.2. Errores de ajustes en la creación de modelo	23
4.4.4.3. Algoritmos de Lenguaje supervisado	24
4.4.4.4. Validación de modelo	31
4.4.4.5. Potencial de Machine Learning	35
4.4.5. Hadoop	40
5. Big Data en la Industria	43
5.1. Teorías de la sostenibilidad en una cadena de suministros	43
5.2. Nuevo enfoque a la administración	45
5.2.1. Tomadores de decisiones	46
5.2.2. 5 Desafíos en la administración	48
5.3. Estado del Arte	49
5.3.1. Inversiones y utilidades dentro de las empresas digitales	49
5.3.2. Impacto de inteligencia artificial en la industria del Retail	50
5.3.2.1. Mejoras en el área de operaciones	51
5.3.2.2. Proyección del retail a futuro	52
5.3.3. Big Data en Finanzas	54
5.3.3.1. Puntuación de crédito	54
5.3.3.2. Trading algorítmico	55
5.3.3.3. Gestión de riesgos	55
6. Caso de estudio aplicado	57

6.1. Introducción al desafío	57
6.2. Descripción del problema	58
6.3. Resolución del desafío	60
6.3.1. Canalización de datos	61
6.3.1.1. Preparación de datos	61
6.3.1.2. Creación de modelos	66
6.3.1.3. Evaluación	67
6.3.1.4. Afinación de modelo	68
6.3.1.5. Predicción y Resultados	70
7. Conclusiones y perspectivas	73
Bibliografía	75
A. Anexo	77
A.1. Código en python	77

Índice de Figuras

4.1. Avance desde Inteligencia Artificial a Machine Learning	15
4.2. Ejemplo de Carta de Control	16
4.3. Caso práctico de observaciones de paquetes de datos Ethernet	20
4.4. Workflow de la metodología usual usada en Machine Learning	21
4.5. Ejemplo de ajustes en un modelo	23
4.6. Diagrama básico de red neuronal	27
4.7. Diagrama de sólo 1 neurona	27
4.8. Ejemplo de árbol de decisión con su base de datos	30
4.9. Imagen ejemplificadora de una validación cruzada	32
4.10. Ejemplo de una matriz de confusión binaria	33
4.11. Infografía sobre las capacidades de Machine Learning	36
4.12. Flujo de proceso de información en algoritmo MapReduce	40
5.1. Crecimiento de interés sobre ciencia de dato en las búsquedas de Google	47
6.1. Representación de la relación entre peak kips, kips promedios y kips dinámicos	59
6.2. Peak kips de una rueda transicionando entre una estado <i>vacío</i> y <i>lleno</i>	59
6.3. Ejemplo simplificado de la base de datos	63
6.4. Ejemplo simplificado de base de datos con presión en tiempo $t + 1$	64
6.5. Ejemplo simplificado de base de datos con presión en tiempo $t + 1$	64
6.6. Ejemplo ilustrativo de transformación de la base antes y después de OneHotEncoding	65



1 | Problema de investigación

1.1. Descripción de la problemática

Durante estos últimos años las palabras BIG DATA han estado en el radar para cualquier persona que se dedique al análisis de datos, inteligencia de negocios o esté consciente de las últimas actualizaciones en la tecnología, pero la verdad es que este concepto, que refiere sobre una base de datos masiva que va más allá del procesamiento normal de una computadora, es tan transversal que prácticamente cualquier persona que sea perteneciente a una red social, o use servicios de internet ya sea Uber, Netflix o Spotify está ligado a ella (big data perspective). Inclusive, basta con realizar una compra en cualquier tienda de retail para estar catalogado y guardado en una base de datos sin nuestra autorización y poder ser estudiado para beneficio personal o privado de alguna identidad, ergo, se puede presumir que toda la población es material de estudio para poder predecir ciertas conductas y estimar la mejor forma de realizar una oferta o promoción.

Es por esto que se hace preocupante que estando ya hace algunos años en conocimiento público esta área del conocimiento, en las universidades no esté el interés que merece esta ciencia para ser estudiada y aplicada. En el caso de ingeniería industrial un conocimiento técnico de esta área puede significar una característica competitiva que permita sobresalir en el mercado [25]. Las formas en que se aplican son versátiles, pues sus herramientas son tan variadas que entre ellas permiten la estimación de demanda de un producto, hacer un análisis predictivo e incluso escapa del ámbito económico y financiero para ser usada en seguridad empresarial, como la detección de fraude, intrusos o situaciones inusuales. Todas las grandes entidades privadas usan estas técnicas, como las empresas de seguro, Bancos, grandes tiendas de retail y empresas contratistas de minería de datos.

El ejemplo más cercano para el común de la gente, es la aplicación de algoritmos de clasificación llamados análisis de sentimiento. La razón de que estos sean tan cercanos es que se usan en las redes sociales como Facebook y Twitter y pueden distinguir la reacción de ciertos comentarios, catalogándolos si han sido de connotación positiva, negativa o neutra. Esto es muy usado para cuando una empresa lanza productos nuevos al mercado. e.g. una empresa de elaboración de alimentos que saque una nueva línea de algún producto, puede estimar que como fue percibido el producto, para ver si lo retira del mercado, o al contrario, acelera

y aumenta la producción de este. De hecho, de esta bien detallado podría indicar la cantidad y plazos de producción.

Si bien este avance también se puede usar en cosas fructíferas y que contribuyan al desarrollo de estudios que mejoren la calidad de vida, como aplicaciones en el área de salud para poder verificar enfermedades a través de síntomas, o la detección de células anómalas que puedan predecir el comportamiento de cáncer o algún agravio de salud. Desafortunadamente se está haciendo conocido por la malversación de los datos de las personas, como fue el escándalo de la empresa Cambridge's Analytica ¹. Este agravio consistió en la adquisición ilícita de información privada de millones de personas en Facebook para la creación de un software que pudiera apuntar a los posibles votantes indecisos en campañas políticas para poder inducirlos y ganar su voto. Inclusive hay reportes, como el de Martin Hilbert, experto en redes digitales, que indican que esta fue la jugada maestra del actual presidente de Estados Unidos, Donald Trump para ganar las elecciones presidenciales del 2016. Se estima que el presupuesto en análisis de datos que invirtió Donald Trump fue de más de mil millones de dólares y las personas analizadas ascienden más de 86 millones.

Esto es sólo una demostración de lo poderosa que puede ser esta herramienta, por lo que es menester, tanto del alumnado como los profesores, desarrollar más el estudio de esta área, ya que además de poderosa, es transversal en casi todas las áreas de ciencias y negocios, pues en este mismo estudio se pretende realizar una investigación aplicada de esta ciencia a la ingeniería industrial de operaciones con enfoque a la mantención y prevención de un sistema de transporte ferroviario. Hoy en día el no tener conocimiento de Big Data en una empresa, puede llevar al fracaso si se pretende estar dentro del top de los negocios.

1.2. Contexto y antecedentes preliminares

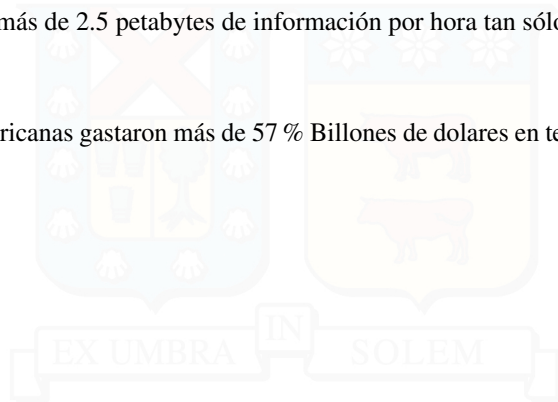
Es un hecho que Big Data se ha vuelto parte de la vida común para todas las personas que participen de alguna red, o sean usuarios de algún servicio de comunicación o compra. Al haberse abaratado los costos de captura y respaldo de información se ha dado pie a eventos de grandes magnitudes en cuanto a la ciencia de los datos. La siguiente lista recopilada por el sitio Zephoria y Forbes ² muestra un ejemplo actualizado al 2018 de como se está llevando en la vida cotidiana y profesional lo que es Big Data:

1. Los datos generados en el 2014 y 2015 superaron con creces los datos generados por toda la humanidad desde el principio de la escritura hasta el 2013. Esta tendencia sigue en alza.
2. En todo el mundo, hay más de 2.20 mil millones de usuarios activos mensuales de Facebook para el 2018 (Facebook MAUs), que es un aumento del 13 % año tras año.
3. Para el 2015, menos del 0,5 % de todos los datos en la humanidad han sido analizados y trabajados

¹<https://www.aljazeera.com/news/2018/03/cambridge-analytica-facebook-scandal>

²<https://zephoria.com/top-15-valuable-facebook-statistics/>

4. Se prevé que Big Data podría proveer más de 6 millones de puestos de trabajos en la economía de Estados Unidos
5. Los datos deficientes entre las empresas y el gobierno le cuestan a la economía estadounidense \$ 3.1 billones de dólares al año
6. Walmart colecciona más de 2.5 petabytes de información por hora tan sólo en la recopilación de datos de sus clientes
7. Compañías norteamericanas gastaron más de 57 % Billones de dolares en tecnología para el tratamiento de Big Data





2 | Objetivos

2.1. Objetivo General

Explorar métodos de análisis para Big Data y su aplicación en problemas de ingeniería con un caso de estudio aplicado a la prevención de fallos en una empresa ferroviaria.

2.2. Objetivos Específicos

1. Conocer, comprender y definir diversos métodos de estudio de Big Data relacionados al análisis de datos.
2. Clasificar y crear un panorama general de la aplicación de los métodos de tratamiento de Big Data en la ingeniería industrial.
3. Desarrollar como caso aplicado la solución a un problema propuesto sobre prevención de fallos en ingeniería industrial, enfocado en la predicción de carga que soporta un riel y clasificación de vagones dañados según la carga ejercida.
4. Generar conclusiones y recomendaciones a través de un análisis técnico.



3 | Metodología

A continuación se señala la metodología usada para la elaboración de esta tesis, incluyendo la resolución del conflicto: Big Data en la industria y el estudio de un caso aplicado a la prevención de fallas en una empresa ferroviaria.

1. **Definición del problema:** Se plantea el problema de como está afectando la aplicación de análisis de Big Data en la industria. A este se le da un enfoque de como se relaciona directamente con la ingeniería industrial. En adhesión se plantea también un desafío propuesto por El Instituto de Investigación Operativa y Ciencias de la Gestión, INFORMS. Este consta de la lectura de una base de datos masiva con 10 millones de muestra donde se debe extraer información para predecir y evaluar eventos para generar conclusiones y toma de decisiones.

2. **Búsqueda bibliográfica:** Se recopila información de paper, libros y artículos de información digital. Para la recopilación de estos, específicamente se uso:

Papers: Scholar Google – Springer – Elsevier – Google

Libros: Scholar Google – Biblioteca UTFSM – Libgen

Revistas : Google – Scholar Google

Esta información fue usada tanto para hacer análisis sobre el enfoque que tiene Big Data en la industria, tanto como en las técnicas de aprendizaje para aplicar minería de datos en la propuesta realizada por INFORMS.

3. **Enfoque industrial:** Una vez recopilada la información, se procede a realizar el análisis de la relación entre esta ciencia y la industria en general. Para esto se abordan temas de intereses como la inversión actual que se está utilizando, enfoque económico-financiero de las herramientas de Big Data, y la posible relación de causa que hay entre ser una empresa que toma decisiones por Big Data y su aumento de beneficios económicos.

4. **Desarrollo del problema:** En esta fase se decide las herramientas a ocupar. En caso del Hardware se cuenta con un laptop de oficina común, específicamente un Dell Latitude 3450. Para Software se

usa el lenguaje de programación Python. Esto se decide por la cantidad de documentación y paquetes existentes que hay para relacionar herramientas de Machine Learning con el manejo de base de datos gigantes.

5. **Metodología KDD:** Para la resolución del problema de investigación, se aborda a través de la metodología KDD. Esta consiste en un manual que consta de 5 pasos para poder extraer relaciones y patrones en una base de datos: Estos 5 pasos son Preparación de datos – Creación de modelo – Evaluación – Mejora de modelo – Predicción
6. **Conclusión:** Esta fase se realiza a través de reuniones con la profesora guía y el alumno, donde a través de lluvias de ideas y análisis de los resultados se llegan a conclusiones, tanto como al rendimiento de modelos para predecir y resolver la propuesta de INFORMS, como para entender y relacionar los eventos que ha tenido Big Data y la Ingeniería Industrial. Llegando a la idea de que es una revolución de la cual Chile aún no es parte, pero se puede integrar.

A continuación se añade un resumen de la metodología, donde se visualiza cual fue el paso a paso para la resolución de este tema de estudio:





4 | Marco Teórico

4.1. Antecedentes

Para el año 2013 un reporte de McKinsey, una compañía influyente en el mundo de los negocios, pronosticó que el área de la ciencia de los datos será el impulsor más importante para el crecimiento económico. McKinsey identificó que una de las oportunidades que contribuyó al lanzamiento de la era del big data fue "*torrente creciente de datos*". Este concepto hace referencia a que los datos se almacenan continuamente a una tasa de velocidad muy alta [7].

Otro factor importante son los bajos precios que se han adquirido para almacenamiento, e.g toda la música en el mundo puede ser almacenada en un dispositivo de disco duro que tan solo cuesta 600 Dólares[9]. Esto es una capacidad de almacenamiento impensada hace pocos años atrás. Un caso cercano para la mayoría de las personas, sobre el fenómeno de "torrente creciente de datos", se puede apreciar en la red social Facebook, pues para el año 2013 según el mismo reporte de la empresa, se tenía que había 4.75 billones de publicaciones compartidas diariamente para el mes de mayo, lo que significaba un aumento de 94 % sobre el mes de agosto del año 2012³.

Por otra parte, Cloud Computing también es considerado otro de las aristas fundamentales que contribuyen al poder catalizador de la ciencia de los datos [14]. Esta tecnología es una de las formas en que la informática se ha convertido en algo que se puede realizar en cualquier momento y lugar. Se define como una computación basada en internet que proporciona recursos compartidos de procesamiento entre computadoras y datos a computadoras u otros dispositivo en demanda. Esta habilidad combinada con el "*torrente creciente de dato*" da la oportunidad de realizar análisis dinámico y novedosos, por lo que se considera el nucleo del lanzamiento de la era de big data.

³<https://zephoria.com/top-15-valuable-facebook-statistics/>

4.2. ¿Qué es Big Data?

El big data y su análisis son el centro de la ciencia moderna y de negocios. Los datos actualmente se consiguen de diversas formas tales como transacciones en línea, correos electrónicos, videos, audios, imágenes, secuencias de clics, registros, publicaciones, consultas de búsqueda, registros de salud, interacciones de redes sociales, datos científicos, sensores y teléfonos móviles y sus aplicaciones [23]. Estos se almacenan en bases de datos que crecen masivamente y se vuelven difíciles de capturar, formar, almacenar, administrar, compartir, analizar y visualizar a través de herramientas de software de base de datos típicas.

En cuanto a la definición técnica de Big Data no se tiene una idea clara y consensuada de lo que este constituye, sin embargo hay una definición recurrente en el área de investigación que está de acuerdo con la apreciación que en términos de datos, es tan colosal que se necesita de una forma de computación distribuida para procesar y analizar, por lo que los software comunes como hojas de cálculos no sirven para tratar la información en un tiempo afortunado [9]. Se tiene que tener en cuenta, que los desafíos de procesamiento y análisis, no solo se abarcan en la cantidad de datos, pues además del **volumen**, se tiene que tener en consideración la **variedad** y la **velocidad** de recopilación, dando pie a otra definición propuesta en Machine Learning Real World Problem [6] conocida como las **3 V's**.

Volumen

Unos de los factores claves del volumen se menciona en el estudio de Mckinsey, este refiere a que los bajos costos de almacenamiento han hecho posible que sea económicamente factible guardar gran cantidad de información, relativamente a un precio despreciable. El otro factor ha sido los avances en la tecnología de sensores y medición han hecho posible el acopio de información masiva en el área de la medicina, telecomunicaciones y aplicaciones militares. Se ha llegado a tal punto que para el 2010 se creaba tanto información como lo hacíamos desde el principio de los tiempo hasta el 2003 ⁴.

Variedad

Hoy en día las empresas están enfocadas en guardar tanta información del cliente como sea posible, y debido al alza de tecnología en captación, medición, y la explosión de las redes sociales, el big data permite oír la voz de cada consumidor, en vez de una segmentación en gran escala. En otras palabras big data da paso para una revolución en el marketing, personalizando la experiencia para el cliente, lo que se resume en clientes más satisfechos. Los consumidores generan copiosamente datos de acceso público a través de sitios de redes sociales. A través de esto las compañías aprovechan de ver sus historiales de compra, de búsqueda, lo que ven, donde han estado y sobre que están interesados a través de "likes" y publicaciones

⁴<https://techcrunch.com/2010/08/04/schmidt-data/>

compartidas.

En cuanto a la manufactura, a través de la variedad de sensores ya no están limitados por la medición tradicional de una sola dimensión. Sumado a la situación mencionada a las redes sociales, se permite tener una visión de calidad de manufactura en todas las direcciones. Para estos casos se tienen limitados métodos estadísticos para monitorear procesos multivariados con una mixtura de datos continuos y categóricos. Por lo que se suelen usar “Análisis de texto”. Los cuales se verán con más acabamiento más adelante.

Velocidad

Por convención se tiene que la noción de velocidad es la tasa en cual los datos se reciben, se almacenan y luego se recuperan para su procesamiento. Pero nuevos investigadores de la materia dicen que esta definición se debe expandir a considerar la rapidez con la que fluye la información [31], uno de los motivos es la llamada "Manguera de incendios" que ocurre en Twitter o tiendas de Retail. Uno de los principales desafíos que se tienen con respecto a la velocidad es cuán rápido una organización puede convertir los datos de alta velocidad en conocimiento. Una mejora en este punto podría aportar grandes aportes a las situaciones de stock-out en una empresa manufacturera o ser de importancia vital en el contexto de un servicio médico.

Otro desafío relacionado a la velocidad, es el hecho de que las redes sociales pueden producir flujos muy variables, con peaks muy altos y valles cercanos a 0. Un ejemplo claro en el contexto de las redes sociales es la frase "se ha viralizado", que indica cuando un video, blog o imagen se ha compartido en gran escala a una alta rapidez.

Estos puntos hacen que los flujos de ingreso colapsen los sistemas, por lo que fomenta el mercado de Infraestructuras como servicios (IaaS), donde los clientes obtienen el beneficio de computación a gran escala , almacenamiento y acceso a base de datos.

4.3. El valor del Big Data

Tal como se vio en la Sección 1.1 big data tiene una gran incidencia en el área del marketing a través de la atención casi personalizada hacia el cliente. Uno de los ejemplos que ya es familiar en productos que funcionan base a internet, son los motores de recomendaciones. Estos usan patrones y las características de los productos para enriquecer la experiencia de los usuarios. La tienda online Amazon usa como estrategia principal de marketing este servicio, donde ofrece ciertas recomendaciones según el historial de compra. En Chile tiendas de retail como Falabella, Paris y Ripley ya han empezado a usar esta estrategia. Similarmente Netflix usa estos motores de recomendaciones para ofrecer un catálogo de películas a gustos del consumidor.

Las grandes marcas usan una técnica llamada Sentiment analysis, que consiste en un proceso de

identificación y categorización computacional de las palabras empleadas en texto de opiniones como publicaciones y Twitter. Bajo este sistema, permiten dilucidar la opinión general del texto, teniendo como output si la satisfacción hacia el producto ha sido positiva, negativa o neutra.

Los teléfonos móviles no están exentos de estas tecnologías. Las marcas aprovechan de usar los sensores integrados a este, como el GPS y obtienen la localización exacta en tiempo real. Basado en esto ofrecen anuncios y ofertas a los usuarios que estén dentro de un radio cercano.

Las empresas de seguros en Chile están empezando a tomar conciencia de cómo podría afectar el big data en sus negocios. Se espera que el 60 % de la compañías van a integrar tecnologías de información y de operación para el año 2018 ⁵. Esto no quiere decir que sólo se modernicen las herramientas para mejorar procesos internos, si no de captar el potencial que el análisis de datos tiene en un negocio basado en la evaluación de los riesgos. Estos grandes avances provocarán una auténtica revolución, al permitir un mayor conocimiento del cliente y de su comportamiento, facilitando el estudio de la previsión de los riesgos y el análisis de probabilidades de ocurrencia de siniestros.

La imagen representa el cómo los conceptos están interrelacionados dentro del otro, siendo así la inteligencia Artificial el concepto global, del cual nace la motivación por desarrollar técnicas de análisis predictivos y de categorización, explicando así el por qué resulta complicado separar un tema del otro.

Para términos prácticos del tema de estudio, se hará un detalle exhaustivo de las técnicas usadas en Machine Learning, debido a que serán las herramientas utilizadas en la metodología para tratar el posterior caso de estudio. En cuanto a los otros conceptos estos se definirán de manera breve para tener una idea integral de Data Science, y así no perder el enfoque principal.

4.4. Herramienta de análisis para Big Data

Cuando se habla de las herramientas que se usan para tratar big data, se hace referencia al campo de ingeniería de computación- industrial; Data Science. Resulta difícil de separar las metodologías existentes en este campo, tales como Data Mining, Artificial Intelligence, Machine Learning o Deep Learning. Existe una superposición considerable de los conceptos y técnicas usadas en cada una de estas ramas. En la Figura 4.1 se puede apreciar la relación temporal que existe entre estas:

⁵<http://web.sofofa.cl/noticia/el-2018-el-60-de-las-companias-van-a-integrar-ti>

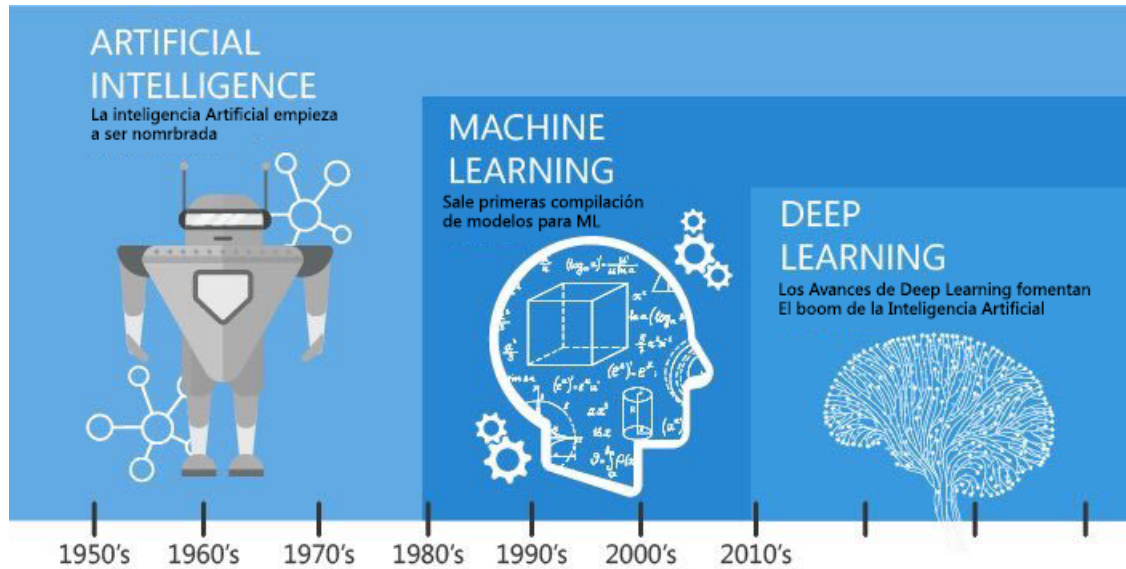


Figura 4.1: Avance desde Inteligencia Artificial a Machine Learning

Fuente: www.cleverdata.io

4.4.1. Inteligencia Artificial

Es el estudio del diseño de agentes inteligentes. Teniendo en cuenta que un agente puede ser cualquier cosa que se relacione con su entorno, para que este se considere inteligente se debe cumplir que sus acciones sean motivadas y apropiadas a sus objetivos, sea flexible a cambios en el entorno y sus metas, y que tome decisiones correctas dadas limitaciones perceptivas (inputs) teniendo en cuenta su computación finita. Esta definición cabe dentro de la categoría del pensamiento racional [18].

En cuanto al enfoque de la inteligencia artificial como pensamiento humano. Se tiene como piedra angular el Test de Turing, propuesto por Alan Turing en el año 1950. Este fue diseñado para proveer una definición operacional satisfactoria de la inteligencia.

En este test, un computador aprueba el examen si el interrogador, que ha de ser humano, tras una serie de preguntas escritas, no logra deducir si la fuente de las respuestas proviene de una máquina o una persona. Para que el computador pueda ejercer esta acción ha de tener:

- **procesamiento natural del lenguaje** que le permita comunicarse en el mismo idioma.
- **Representación del conocimiento.**
- **Razonamiento automatizado** que lo faculta de poder usar la información almacenada para sacar nuevas conclusiones
- **Machine Learning** que lo adapta a nuevas circunstancias, además de detectar y extrapolar patrones.

[20]

Dado esto la inteligencia artificial sienta las bases de las técnicas usadas en el análisis de big data. Sería redundante explayarse en el estado del arte y las herramientas de análisis usadas, pues como se indica en la Figura 4.1, la AI comprende todas las ramas de data analysis, y estas serán vistas con más detalles en su debida sección.

4.4.2. Cartas de control

Big data a menudo toma la forma de flujo de datos longitudinal, o sea observaciones de ciertos procesos coleccionados secuencialmente a través del tiempo (conservando el mismo sujeto o característica de medición). e.g la medición del peso de un objeto en una línea de producción, secuencia de imágenes satelitales y datos del clima. Uno de los muchos propósitos de coleccionar y analizar Big data puede ser el monitoreo del performance longitudinal de algún proceso. A este fin, el proceso de control estadístico (SPC) puede ser una herramienta muy útil. [19].

Existen herramientas básicas pero poderosas en el SPC, las cuales tienen por nombre Tablas de control. Estos son estadísticos, que se apoyan en gráficos, estos generalmente contienen una línea central que representa el valor medio del proceso en control. Otras dos líneas horizontales, llamadas límite de control superior (UCL) y límite de control inferior (LCL). La cuales determinan el estado de control del proceso, indicando anomalías en caso que estos límites sean superados. En la Figura 4.2 se tiene el ejemplo de una carta de control en uso, donde se aprecian los límites de control y se marcan en rojo los elementos que detonan la alerta de comportamiento erróneo.

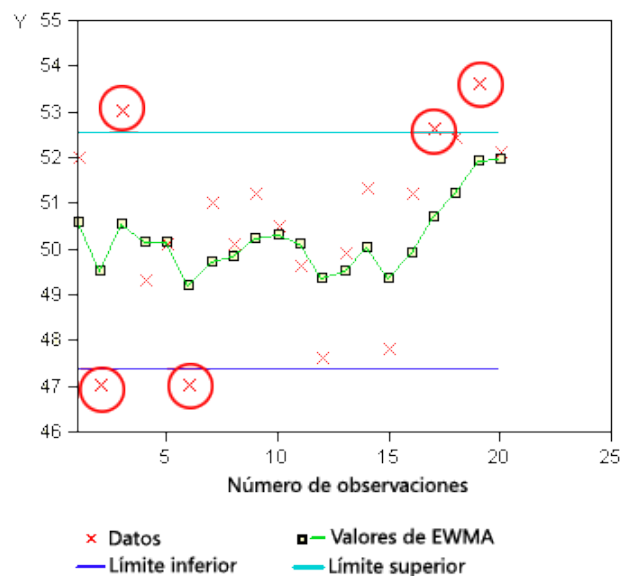


Figura 4.2: Ejemplo de Carta de Control

Fuente: Elaboración propia

La forma en que se fijan estos límites van variando según la carta de control a usar. En esta sección se detallará la carta de control EWMA y CUSUM, ya que han sido testeados en uso real de Big Data.

Carta de control EWMA

La carta de promedio móvil ponderado exponencialmente (EWMA en inglés) es un estadístico para monitorear procesos, la cual promedia los datos de una forma que le otorga menos peso e importancia a los datos a medida que estos van siendo renovados en el tiempo.

Para la carta de control de Stewart, que es la técnica más básica dentro de esta área, la decisión sobre el estado de control del proceso en cualquier momento, t , depende únicamente de la medición más reciente, y por supuesto, del grado de “veracidad” de los estimadores de los límites de control a partir de los datos históricos. Para la técnica de control EWMA, depende de su estadístico EWMA, el cual es un promedio exponencialmente ponderado de todos los datos anteriores, incluyendo la medición más reciente en el momento t .

El estadístico se calcula de la siguiente manera:

$$EWMA_t = \lambda Y_t + (1 - \lambda)EWMA_{t-1} \quad \text{para } t = 1, 2, \dots, n.$$

donde

- $EWMA_0$ es el promedio de los datos históricos(target)
- Y_t es la observación en el momento t .
- n es el número de observaciones a monotrear, incluyendo $EWMA_0$
- $0 < \lambda \leq 1$ es una constante que determina la profundidad de la memoria de EWMA.

Esta ecuación corresponde a Roberts H.S [22]

Determinando λ y Límites de control

El parámetro λ determina la tasa en la cual los datos antiguos entran al cálculo del estadístico EWMA. Un valor de $\lambda = 1$ indica que sólo los datos medidos más recientemente influyen en el EWMA(A diferencia de la carta de Stewart). A esto, un valor largo de λ aproximado a 1 da más peso a los datos más recientes y menos a los datos más antiguos. Un valor pequeño de λ (cercano a 0) da más peso a los datos antiguos. El valor habitual que se le otorga al parámetro λ es entre 0.2 y 0.3. Aunque esta elección es un tanto arbitraria.

Lucas y Sacucci entregan una tabla que ayuda a seleccionar según la naturaleza de los parámetros de estudio [11] λ

Para determinar los límites de control se necesita primero obtener la varianza estimada del estadístico EWMA, este se obtiene:

$$s_{ewma}^2 = \frac{\lambda}{2 - \lambda} s^2$$

Donde s corresponde a la desviación estándar de los datos históricos

La línea central de la tabla de control es el valor objetivo o $EWMA_0$. Los Límites de control superior(UCL) e inferior(LCL) son:

$$UCL = EWMA_0 + k s_{ewma}$$

$$LCL = EWMA_0 - k s_{ewma}$$

donde el factor k se obtiene de las tablas mencionadas

Cartas de Control CUSUM

En SPC, CUSUM(o Carta de control de suma acumulativa) es una alternativa a la carta de control EWMA o de Stewart. Aunque esta no es tan simple e intuitiva como esta última, se ha demostrado que actúa de manera más eficiente en la detección de pequeños cambios en la media del proceso.

De hecho, en los estudios del ARL (Longitud de recorrido promedio) de la carta CUSUM ⁶ se ha intuído que puede ser más práctica cuando se desea detectar cambios en la media que sean 2σ o menos.

Esta carta se emplea de la siguiente forma: Se debe recolectar m muestras, cada una de igual tamaño n , y luego calcular la media de cada muestra. Entonces la carta de control de sumas acumulativas se obtiene trazando una de las siguientes cantidades:

⁶<http://www.itl.nist.gov>

$$S_m = \sum_{i=1}^m (\bar{x}_i - \hat{\mu}_0) \quad \text{o} \quad S'_m = \frac{1}{\sigma_{\bar{x}}} \sum_{i=1}^m (\bar{x}_i - \hat{\mu}_0)$$

Donde

- m : Número de muestras
- $\hat{\mu}_0$: es la media estimada en control
- $\sigma_{\bar{x}}$: es la media conocida
- $\bar{x}_i$: es el promedio de la muestra i

La decisión de cual de estas dos ecuaciones son graficadas está usualmente determinada por algún software de optimización. De todos modos, mientras el proceso se mantenga centrado en $\hat{\mu}_0$, el gráfico de CUSUM mostrará la variación en un patrón aleatorio centrado en 0. Si la media del proceso se desplaza hacia arriba, los puntos trazados de CUSUM se desplazarán de igual forma hacia arriba, y viceversa en caso de que decrezcan.

Cartas de control aplicada a Big Data

Generalmente las tablas de control, sobre todo las mencionadas, se han graficado para monitorear líneas de producción en la industria manufacturera. Pero últimamente se han estado probando en distintas aplicaciones como vigilancia de enfermedades infecciosas, monitoreo ambiental y otras áreas.

Cuando se instala por vez primera una línea de producción, los gráficos de SPC se pueden usar para verificar si la calidad de una cantidad relativamente pequeña de productos cumple con los requisitos de calidad. Si se detectan que algunos productos son defectuosos la línea de producción se va actualizando y mejorando. Este ejercicio de ensayo y error continúa hasta que la línea de producción corre estable. A esta etapa del proceso se le llama Fase I de SPC. En cambio el monitoreo online es llamado phase II SPC. Su objetivo es asegurar que los datos estén en control, o sea dentro de los límites, y lanzar una señal de alerta lo antes posible para actuar rápido.

El siguiente ejemplo trata de un análisis de Big Data que ha sido trabajo por en el estudio Quality and Reliability Engineering International [26]. Donde se tienen paquetes de datos y/o metadatos de Ethernet, encargados de llevar la información adicional en las señales de internet. Se aprecia en la Figura 4.3 el ejemplo llevado en vida real, donde se tiene el comportamiento de la tabla CUSUM y EWMA respectivamente. Donde efectivamente hubo un comportamiento parecido, y además permitió gatillar de forma oportuna la señal para datos muy outliers. En el eje de las ordenadas se tiene el tamaño del paquete en escala logarítmica y en el eje de las abscisas el tiempo de medición.

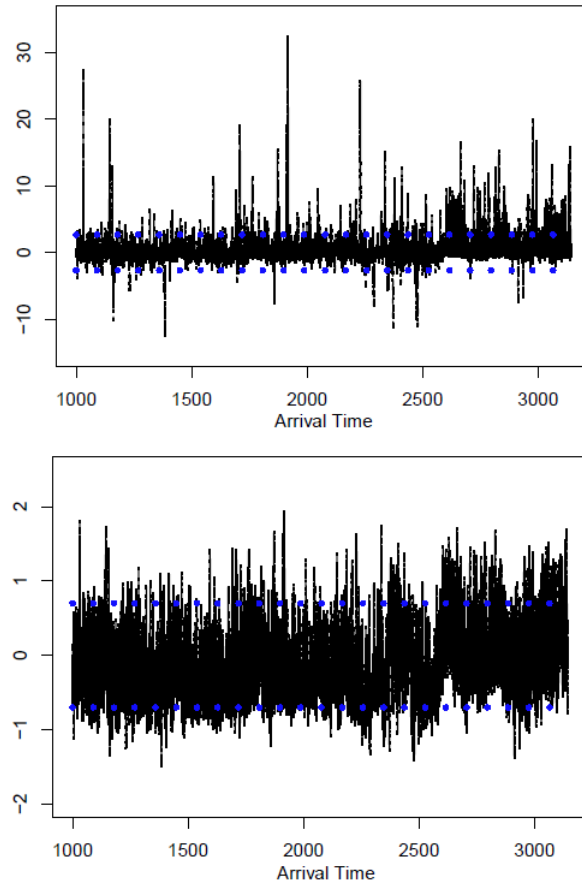


Figura 4.3: Caso práctico de observaciones de paquetes de datos Ethernet

Fuente: [19]

4.4.3. Minería de datos

Teóricamente hablando, el objetivo de esta ciencia es el descubrir patrones y relaciones en datos y usar estos descubrimientos para beneficio, ya sea personal o de una empresa. Existen varios enfoques que se han ido desarrollando durante estos últimos 30 años pero todos tienen un objetivo en común, estimar la función relacional de las característica de entrada y la variable objetivo. Estos enfoques usan un workflow en común, se que aprecia en la Figura 4.4. Esta es una representación de la metodología *Knowledge Discovery in Databases*, KDD. Esta se refiere al proceso de descubrir conocimiento útil de una colección de datos[27]. Se verá con detalles la ejecución de esta herramienta en la sección de desarrollo. La siguiente ilustración es una forma estandarizada y resumida de toda la metodología.

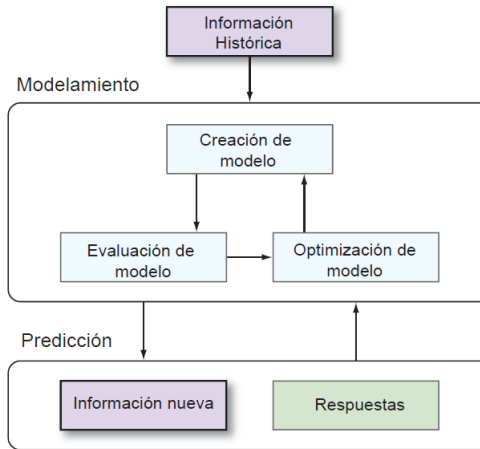


Figura 4.4: Workflow de la metodología usual usada en Machine Learning

Fuente: Elaboración propia

De este workflow se desprenden 5 etapas remarcables:

- **Preparación de datos:** Generalmente implica obtener los datos en un formato ordenado y comprensible por la máquina, habitualmente es tabular. Además se identifica que valores son continuos y discretos, y se numeran de ser necesario. e .g. Para identificar si el día está [*despejado, nublado, lluvioso*] se podría categorizar como [1,2,3].
- **Creación de modelo:** Para esta etapa se deben hacer las preguntas que pueden ser respondidas por los datos. El trabajo del algoritmo de ML será predecir con éxito el objetivo a través del conjunto de inputs. La Figura 4.11 puede ser usada de guía para ver que modelo implementar. Estos se explicarán con más detalle en la sección de aprendizaje de máquina.
- **Evaluación:** Aunque no siempre se realiza, suele ser una fase común de la metodología. Una forma de realizar la evaluación es probar la función de hipótesis con datos reales ya medurado en la data histórica. O sea viendo el error del objetivo predicho con el real.
- **Mejora de modelo:** 3 formas comunes son perfeccionar los parámetros de la hipótesis, filtrar el conjunto de los inputs y reprocesamiento y limpieza de los datos.
- **Predicción:** Es la última etapa del modelo, y el objetivo final, donde se realiza la predicción o tarea asignada al agente de Machine Learning.

4.4.4. Machine Learning

Al ser un estudio tan distendido y nuevo, perteneciente al último siglo, existen variadas definiciones. Entre estas destacan la propuesta por Arthur Samuel [6] , “*Es el campo de estudio que le otorga a las*

computadoras la habilidad para aprender sin ser explícitamente programadas para eso”, en la actualidad esta definición se considera un tanto informal y antigua.

La definición más usada de Machine Learning hoy en día es la del profesor Tom Mitchell de la Universidad de Carnegie Mellon: “ *Un computador se dice que aprende de la experiencia ‘E’, con respecto alguna clase de tarea ‘T’ y un rendimiento medido ‘P’ si es que el rendimiento de la tarea T medido en P mejora con la experiencia E* “. Dicho esto en palabras más simple, si un software (Agente inteligente) mejora a través de la experiencia, se puede decir que ha aprendido.

Machine Learning puede ser solución para una gran cartera de problemas en el sector de los negocios, desde detección de fraude, targeting de consumidores y recomendación de productos hasta monitoreo en tiempo-real de la industria, diagnósticos médicos, y análisis de texto. Puede tomar problemas que no pueden ser solucionados manualmente por la inconmensurable cantidad de datos a analizar. [6].

4.4.4.1. Formas de aprendizaje

De los aprendizajes más comunes existen 3 tipos que se basan en la naturaleza de los datos con lo que se trabaja. i.e. cómo se relacionan los inputs y outputs, entre ellos está el **Unsupervised learning** que permite abordar problemas con poca o ninguna idea de como deberían ser los resultados, pues aprende patrones del input sin tener ninguna relación explícita. Este tipo de aprendizaje permite definir una estructura de datos sin saber los efectos existentes entre las variables. La tarea más habitual dentro del Unsupervised learning es el “*Clustering*” que permite separar datos similares en agrupaciones o clusters intuitivos según sus variables, e.g. de una muestra de 1.000.000 genes distintos, se podría encontrar una forma automática de agrupar en subconjuntos de acuerdo a sus distintas variables, como la esperanza de vida, la ubicación, sus roles. También tiene algoritmos que difieren al cluster, como el “*Cocktail Party Algorithm*” que permite encontrar una estructura en un ambiente caótico, e.g. identificar voces individuales de una mezcla de sonidos como una fiesta o un restaurante con música[20].

En el **Reinforcement Learning** el agente aprende de una serie de refuerzos – Premios o castigos que adquiere a través de su contacto con su entorno, e.g. Un computador que se le premia cada vez que gana una partida de damas, después de varias jugadas el computador debería poder identificar cuales fueron jugadas ingeniosas para después replicarlas. Actualmente sus usos más comunes son los autos sin choferes, las aspiradoras de barrido automáticos, programación inteligente de ascensores entre otros.

Por último está el **Supervised learning**, que es el aprendizaje que se ahondará en este estudio. En este el agente observa de algunos ejemplos entre los input-output (La relación viene dada por defecto) y aprende una función que mapea desde los inputs hacia los output. Dicho de otra forma, dado un conjunto de entrenamiento de N ejemplos entre pares de input-output (pueden ser varios inputs):

$$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$$

Donde cada y_j fue generado por una función desconocida $y = f(x)$, se debe descubrir una función h (hipótesis) que se aproxime a la verdadera función f .

4.4.4.2. Errores de ajustes en la creación de modelo

En machine learning la forma de aprendizaje se dice que es inductiva, es decir, se aprenden conceptos generales a través de ejemplos específicos, el supervised learning es un claro ejemplo de esto. Este concepto da paso a la generalización de los datos, que refiere sobre que tan bien el modelo interpreta los datos nuevos de entrada que no han sido estudiado por el modelo en su fase de aprendizaje. Existen varios errores en cuanto a la generalización del modelo, tales como bajo ajuste y sobre ajuste, estos se pueden visualizar de manera gráfica en la Figura 4.5 expuesta más abajo:

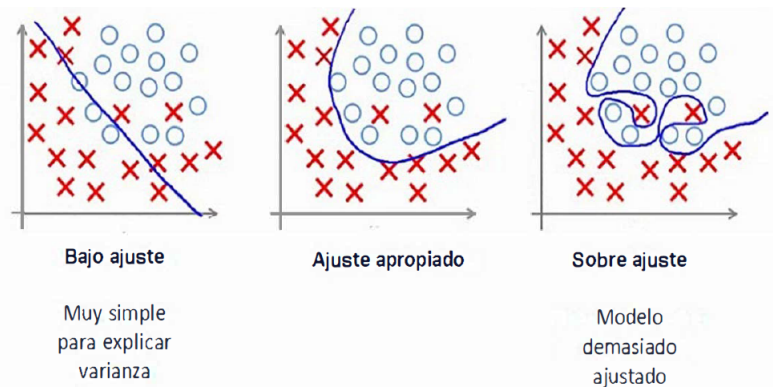


Figura 4.5: Ejemplo de ajustes en un modelo

Fuente: <http://geeksforgeeks.org>

Bajo ajuste: Se dice que un modelo estadístico de machine learning se encuentra con bajo ajuste (Figura 4.5 primer recuadro) cuando no puede capturar la tendencia subyacente de los datos, en otras palabras, el modelo no se ajusta bien a las especificaciones. Generalmente ocurre cuando se tienen pocas muestras en la base de entrenamiento o cuando se intenta construir un modelo lineal con datos de relaciones no lineales. En tales casos, las reglas de aprendizaje son demasiado básicas y flexibles para ser aplicadas en una base tan mínima, por lo tanto se tiende a producir muchos errores de predicción. Las formas que se tienen para mejorar un modelo que sufre de bajo ajuste pueden ser la captura de más datos, reducir la cantidad de atributos o simplemente cambiar y probar otros modelos que hagan otra relación de lo datos.

Sobre ajuste: Al contrario del bajo ajuste, un modelo de machine learning se dice estar en sobre ajuste cuando se entrena un modelo con muchos datos. Si bien esto al principio parece ideal que mientras más

grande la base de datos, mejor será la predicción, usualmente puede conllevar a problemas debido a que al utilizar base de datos masivas el algoritmo empieza a aprender de los ruidos y entrada de datos inexactas en el conjunto de datos. Entonces el modelo no generaliza bien los datos debido a muchos detalles y ruidos. Las causas que producen sobre ajuste suelen ocurrir en modelos no-paramétricos o no-lineales porque estos tipos de algoritmos tienen más libertad en la construcción del modelo, siendo a veces modelos irrealistas. Otra causa pueden ser las base de datos no balanceadas, ya que el modelo al ser optimizado por la minimización del error, va a tender a predecir la clase mayoritaria, así asegura una exactitud del casi 100 % de las predicciones, pero en el fondo no es capaz de realizar predicciones para la clase minoritaria. Las formas que se tienen para evitar este problema son la utilización de modelos lineales para datos lineales, reducir los outliers a través de una buena limpieza de datos y para el caso de los datos no balanceados, se puede hacer un resample (remuestreo), que consiste en un aumento de los datos minoritarios, ya sea por fabricación sintéticas de datos, el duplicado de estos o la reducción aleatoria de los datos mayoritario, todo esto enfocado a obtener una proporción más igualitaria entre los datos.

Los métodos para evaluar el modelo y revisar si este sufre de algunos de estos problemas, se pueden revisar en la Subsubsección 4.4.4.4 validación de datos,

4.4.4.3. Algoritmos de Lenguaje supervisado

En esta sección se detallan sólo algunos de los algoritmos más conocidos del Supervised learning. En la Figura 4.11 se aprecia una infografía de algoritmos usados tanto en Supervised como unsupervised. Adicionalmente se detallan para que situaciones conviene usarlos y cuando según sus limitaciones.

4.4.4.3.1 Regresión lineal univariada

Teniendo en cuenta el último ejemplo de la Subsubsección 4.4.4.1, una función lineal univariada tiene la forma de $y = \omega_1 x + \omega_0$. Donde ω_0 y ω_1 son coeficientes que van modificando el valor de y respecto a su respectivo peso entre uno y el otro. Se define $\mathbf{w}$ como el vector que contiene $[\omega_0, \omega_1]$ y se define la función de hipótesis como:

$$h_{\mathbf{w}}(x) = \omega_1 x + \omega_0$$

La **Regresión lineal** viene siendo la tarea de encontrar $h_{\mathbf{w}}$ que mejor calze con los output de los datos. Para esto se debe encontrar los valores de $\mathbf{w}$ que minimizen la pérdida empírica. Por lo que se usa la *Función*

de pérdida cuadrática, L_2 , sumada sobre toda la base de entrenamiento.

$$L(h_{\mathbf{w}}) = \sum_{j=1}^N L_2(y_j, h_{\mathbf{w}}(x_j)) = \sum_{j=1}^N (y_j - (\omega_1 x_j + \omega_0))^2 \quad (4.1)$$

Luego, el objetivo es encontrar $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w}} L_2(h_{\mathbf{w}})$, por ende la ecuación 4.1 ha de ser minimizada, derivando parcialmente los coeficiente ω_0 ω_1 e igualándolos a cero. Estas ecuaciones tienen la siguiente solución

$$\omega_1 = \frac{N(\sum x_j y_j) - (\sum x_j)(\sum y_j)}{N(\sum x_j^2) - (\sum x_j)^2} \quad \omega_0 = \frac{(\sum y_j - \omega_1(\sum x_j))}{N} \quad (4.2)$$

Al estar limitado a sólo dos componentes el espacio de los coeficientes, y ser una una función cuadrática, esta siempre será **convexa**, lo que implica que no hay mínimos locales, ergo, en los modelos lineales basta con aplicar la ecuación Ecuación 4.2 para ajustar la línea a los datos. Para ir más allá de un modelo lineal, i.e. no continuo, se puede aplicar un algoritmo hill-climbing que sigue el *gradiente* de la función a ser optimizada. (gradiente descendente para minimizar). En este se escoge cualquier punto del dominio de $\mathbf{w}$, luego se mueve a un punto vecino menor en cuesta abajo, este procedimiento se repite hasta la convergencia. Dicho esto el seudoalgoritmo se presenta a continuación de la siguiente forma [20]:

```

w ← cualquier punto en el espacio de parámetros
loop hasta convergencia do
  for each  $\omega_i$  in w do
     $\omega_i \leftarrow \omega_i - \alpha \frac{\partial}{\partial \omega_i} L_2(\mathbf{w})$ 

```

El parámetro α es llamado Tasa de aprendizaje (Learning Rate) a medida que este es más pequeño, el proceso se vuelve más detallista pues avanza por la curva cubriendo menos distancia, lo que lo hace más lento también de compilar.

La última sentencia del algoritmo se puede generalizar de la siguiente manera para una base de entrenamiento de N mediciones:

$$\omega_0 \leftarrow \omega_0 + \alpha \sum_j (y_j - h_{\mathbf{w}}(x_j)); \quad \omega_1 \leftarrow \omega_1 + \alpha \sum_j (y_j - h_{\mathbf{w}}(x_j))x_j \quad (4.3)$$

Esta actualización del parámetro recibe el nombre de *Descenso del gradiente en lotes* (batch gradient descent) y garantiza la convergencia al mínimo local del espacio de los parámetros $\mathbf{w}$. Siempre cuando se elija un α pequeño.

4.4.4.3.2 Regresión Lineal Multivariada

Este algoritmo es una versión extendida de la regresión lineal simple, donde se cuentan con más características en las mediciones, por lo que la función de hipótesis toma la forma:

$$h_{sw}(\mathbf{x})_j = \omega_0 + \omega_1 x_{j,1} + \dots + \omega_n x_{j,n} = \omega_0 + \sum_i \omega_i x_{j,i}. \quad (4.4)$$

La regresión multivariada no cambia mucho con respecto al algoritmo precesor, el gradiente descendiente en este caso alcanza el mínimo de la función de pérdida con la siguiente actualización de parametros

$$\omega_i \leftarrow \omega_i + \alpha \sum_j x_{j,i} (y_j - h_{\mathbf{w}}(\mathbf{x}_j)) \quad (4.5)$$

en otras palabras su pseudo-algoritmo queda definido como:

$$\omega_j := \omega_j - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\omega}(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)} \quad \text{para cada } j := 0 \dots n \quad (4.6)$$

donde m es la cantidad de muestras, y n la cantidad de parámetros que tiene la regresión.

Esto hace que el algoritmo trabaje mucho más lento en cuanto a su velocidad de procesamiento, y en una función con muchos parámetros, como lo es el procesamiento de imágenes puede tomar días en tan sólo el reconocimiento de una pequeña imagen. Afortunadamente se puede optimizar este proceso con herramientas estadísticas, que ayudan a minimizar la función de pérdida y así obtener $\mathbf{w}$. Sea $\mathbf{y}$ el vector de outputs para los ejemplos de entrenamiento, y $\mathbf{X}$ la matriz de datos, i.e., la matriz de inputs con n -dimensión por cada fila, la solución queda descrita por:

$$\mathbf{w}^* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (4.7)$$

En términos de computación, la velocidad con la que se procesa el método de la Ecuación 4.7 es muy superior al de la Ecuación 4.6. Esto es debido que al trabajar con matrices las operaciones se realizan instantaneamente comparado a la velocidad por iteración [6].

4.4.4.3.3 Redes neuronales

Una red neuronal artificial es un modelo computacional que se basa en la forma que la red de neuronas del cerebro humano procesan la información. Los componentes principales de una neurona son :

- **Dendrita:** Toma los inputs desde otras neuronas
- **Cuerpo Celular:** Genera interfencia en los inputs y decide qué hacer
- **Terminales Axon:** Transmiten los output en forma de impulso eléctrico

En la Figura 4.6 se aprecia una red neuronal de inteligencia artificial básica

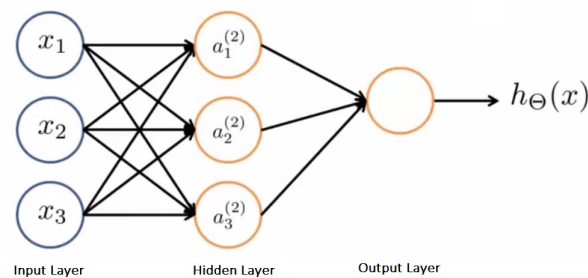


Figura 4.6: Diagrama básico de red neuronal

Fuente: tex.stackexchange.com

Se alcanzan a distinguir 3 capas principales en una red neuronal simple. Que asemejan sus funciones a las redes neuronales del cerebro.

- **Input Layer:** Las observaciones se obtienen a través de esta capa, interaccional con el entorno.
- **Hidden Layer:** Estas pueden ser más de una y son las capas encargadas de aprender las relaciones entre los datos de entrada
- **Output Layer:** El output final es donde convergen las capas anteriores, en caso de un ejemplo de clasificación con 3 clases, el output sería de 3 neuronas.

Para entender mejor la estructura de este algoritmo, se presenta este diagrama de una neurona única, mostrando además sus principales componentes.

$x_1, x_2, \dots, x_N$: Inputs de la neurona, pueden provenir del entorno o de un valor intermedio entre otras capas.

x_0 : Es una constante que trabaja similar al intercepto en las regresiones, generalmente tiene el valor de 1

$\omega_0, \omega_1, \dots, \omega_N$: Peso de cada parámetro

a :Output de la neurona, el cual es calculado de la siguiente forma:

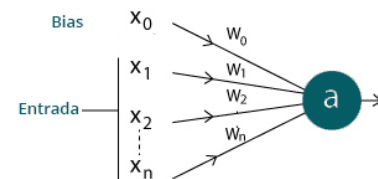


Figura 4.7: Diagrama de sólo 1 neurona

Fuente: Elaboración propia

$$a^{(j)} = g\left(\sum_{i=0}^N \omega_{i,j} x_i\right) \quad (4.8)$$

g es conocido como la función de activación. Esta hace que una red neuronal sea mucho más flexible e imparte la capacidad de estimar complejas relaciones no lineales entre los datos[28]. Cada función de activación toma un numero o input y realiza una operacion matemática fija. La función más usada y conocida es **Sigmoid** (σ):

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Teniendo en cuenta que la función sigmoid es una regresión logística, se ha de tener en cuenta que su función de pérdida o costo es la misma que el algoritmo de clasificación[10], por lo queda definida de la siguiente forma[21]:

$$L(\omega) = -\frac{1}{n} \sum_{i=1}^n [y^{(i)} \log(h_{\omega}(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\omega}(x^{(i)}))] \quad (4.9)$$

Para el ejemplo de la Figura 4.6, según la Ecuación 4.8, la función de hipótesis queda definida de la siguiente forma:

$$a_1^{(2)} = g(\omega_{10}^{(1)} x_0 + \omega_{11}^{(1)} x_1 + \omega_{12}^{(1)} x_2 + \omega_{13}^{(1)} x_3) \quad (4.10)$$

$$a_2^{(2)} = g(\omega_{20}^{(1)} x_0 + \omega_{21}^{(1)} x_1 + \omega_{22}^{(1)} x_2 + \omega_{23}^{(1)} x_3) \quad (4.11)$$

$$a_3^{(2)} = g(\omega_{30}^{(1)} x_0 + \omega_{31}^{(1)} x_1 + \omega_{32}^{(1)} x_2 + \omega_{33}^{(1)} x_3) \quad (4.12)$$

$$h_{\omega}(x) = a_1^{(3)} = g(\omega_{10}^{(2)} a_0^{(2)} + \omega_{11}^{(2)} a_1^{(2)} + \omega_{12}^{(2)} a_2^{(2)} + \omega_{13}^{(2)} a_3^{(2)}) \quad (4.13)$$

Sabiendo que se usa la función de sigmoid, y teniendo en cuenta la Ecuación 4.9, el modelo que se usa para optimizar el algoritmo de redes neuronales queda un tanto complicada, como se muestra en la siguiente ecuación [20]:

$$L(\omega) = -\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K [y_k^{(i)} \log((h_{\omega}(x^{(i)}))_k) + (1 - y_k^{(i)}) \log(1 - (h_{\omega}(x^{(i)}))_k)] + \frac{\lambda}{2m} \sum_{l=1}^{L-1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} (\omega_{j,i}^{(l)})^2 \quad (4.14)$$

Siendo

- **L:** La cantidad total de capas
- S_l : Número de unidades (a) en la capa l
- **K:** El número de outputs, o clases.

Afortunadamente existen varios algoritmos que ayudan a optimizar este proceso. Entre ellos el más conocido es el *Backpropagation*, por temas de complejidad y objetivos de la tesis no se especificarán en detalle, pero si se dará una breve explicación en la etapa de la aplicación del caso de estudio.

4.4.4.3.4 Árbol de decision y Random Forest

Los árboles de decisión son unos de los algoritmos más simples, pero a la vez exitosos dentro de machine learning. Estos representan una función que toma como entrada, un vector de atributos o características y retorna una "*decisión*"- el cual se representa por un sólo valor.

Los modelos en cuales la variable objetivo toma valores discretos son llamados árboles de clasificación, y los que toman valores continuos son llamados árboles de regresión (Técnicamente, resolver un problema de regresión es encontrar una expectativa condicional o un promedio del valor de y , porque la probabilidad de que se encuentre el valor exacto de y es 0.)

La técnica a grandes rasgos consiste en separar la población o muestra en 2 o más conjuntos homogéneos(o sub-conjuntos) bajo una estrategia codiciosa de dividir y conquistar(intenta encontrar un óptimo global en cada fase), por esto se va dividiendo según el atributo más importante, con esto la técnica divide el problema en varios sub-problemas que se pueden resolver recursivamente. Por el "*atributo más importante*", se quiere decir que es el que produce una mayor diferencia en la clasificación, y logre llegar a un nodo terminal (output) lo más rápido posible, en la Figura 4.8 se aprecia este comportamiento al tener la opción de llegar a la respuesta desde el inicio del árbol. De esta forma se espera llegar a la respuesta correcta con un número pequeño de muestras, significando que todos los caminos en el árbol serán cortos y éste como un todo será superficial, ahorrando tiempo de compilación lo cual es uno de los factores más importante al momento de entrenar base de datos de Big Data

En la siguiente figura se puede apreciar el diagrama de un árbol básico de decisión.

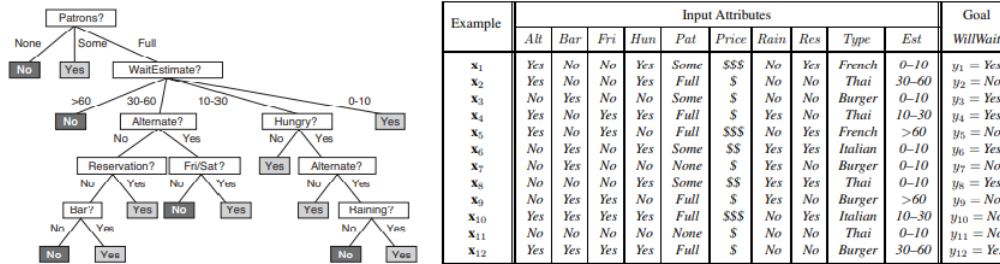


Figura 4.8: Ejemplo de árbol de decisión con su base de datos

Fuente: AI: A modern approach

Los componentes que conforman el árbol vienen definidos por:

- **Nodo raíz:** representa la población entera que será dividida en sub-conjuntos.
- **Splitting:** Es el acto de dividir el nodo
- **Nodo de decisión:** Es cuando un sub-conjunto se somete a más divisiones
- **Nodo terminal:** representa el valor del output, es decir y, siendo un valor unidimensional.

Ventajas de un árbol de decisión:

1. **Útil en exploración de datos:** Son uno de los caminos más rápidos para identificar variables significativas y la relación que se tengan entre 2 o más variables. Con la ayuda de los árboles se pueden crear nuevos atributos que tienen mayor poder de predecir la variable objetivos
2. **Fácil de entender:** No es necesario ser un experto en análisis de datos ni tener conocimiento profundo en estadísticas debido a que se pueden leer e interpretar por su representación gráfica. Suelen ser muy intuitivos y los usuarios pueden relacionar fácilmente su hipótesis.
3. **Se necesita menos limpieza de datos:** Comparado a otros modelos no está tan influenciado por los outliers y valores perdidos.
4. **Método no paramétrico:** Esto quiere decir que los árboles de decisión no tienen suposiciones sobre el espacio de distribución y la estructura de clasificación, siendo más fácil de implementar que otros modelos.

Desventajas

1. **No apto para variables continuas:** Al momento de intentar predecir una variable continua, los árboles de decisión suelen perder información cuando se utilizan variables categóricas dentro de los inputs de atributos.

2. **Overfit:** Es una de las dificultades más prácticas dentro de estos algoritmos, y tal como se menciona anteriormente, el algoritmo queda ajustado a característica muy específicas que realmente no tienen relación causal con la función objetivo.

Una de las formas en que se puede apalejar el sobre ajustes es con un algoritmo de **Random Forest**, este de modo simple consiste en la creación de múltiples árboles de decisión distintos, e.g. Suponiendo que se tienen 1000 observaciones con 10 atributos, Random Forest podría crear múltiples árboles, con tamaño de muestra 100 y tomando tan sólo 5 atributos elegido azarosamente. El proceso se repite 10 veces, para así crear una predicción final en cada observación. La predicción final total se crea con una función basada en las predicciones de cada árbol distinto, esta puede variar dependiendo de la optimización, un caso típico puede ser un promedio ponderado de cada predicción.

4.4.4.4. Validación de modelo

Como paso final dentro de la construcción y mejora de un modelo de machine learning se tienen la etapa de validación, que consiste en evaluar el actual rendimiento de los algoritmos elaborados. Con esta fase se podrá saber si un modelo sufre de los problemas de bajo y sobre ajustes, asimismo se podrá mejorar la selección de los atributos o clases para considerar en el análisis de datos dependiendo del rendimiento que se tenga en el modelo. En cuanto a los métodos de validación propuesto se hará énfasis en los de la categoría de clasificación debido a que el caso propuesto por el caso de estudio exige de manera detallada la forma en que se ha de medir el estudio de regresión, en cambio para clasificación da libertad de elección para evaluar el rendimiento del modelo propuesto.

K cross fold validation

La validación cruzada es una técnica común para evaluar el rendimiento de los modelos de aprendizaje automático, se dice ajuste de optimización al hecho de corroborar el sobre ajuste dentro del modelo. La validación sólo ocupa la base de datos de entrenamiento. Esta técnica realiza una división de la base de entrenamiento para formar una base de testeo dentro de la misma. Las particiones generadas pueden ir variando dependiendo del tipo de validación cruzada, generalmente estas van iterando según las particiones creadas. De esta forma se corrobora que tan bien se comporta el modelo para recibir datos nuevos a la base de entrenamiento. Existen varios tipos de validación cruzada y para este caso se utilizará **K cross fold Validation**, ya que se evita que los sub-conjuntos de prueba se solapen entre sí.

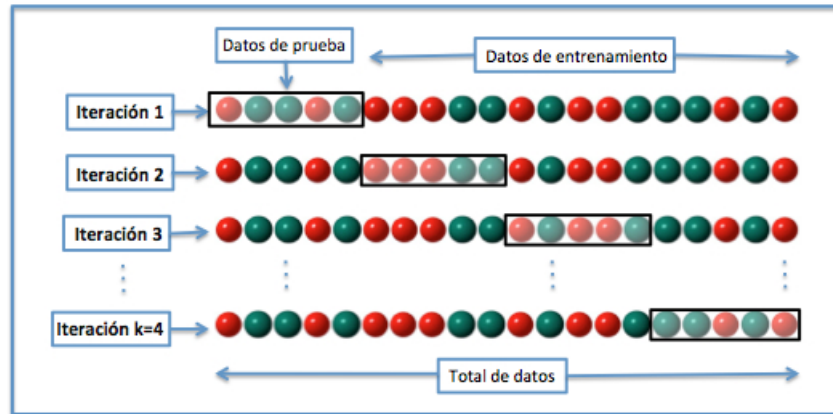


Figura 4.9: Imagen ejemplificadora de una validación cruzada

Fuente: Wikipedia

Para aplicar K cross fold validation es necesario fijar el parámetro K , que definirá el número de particiones que se aplicarán a la base de datos de entrenamiento. No existe una medida exacta para determinar este parámetro pero usualmente se usan $K = 5$ o $K = 10$, puesto que este valor se ha encontrado a través de la experimentación que tiene bajo sesgo y una variación modesta. Una vez fijado el valor, se procede a dividir la base de datos en $(K - 1)$ particiones para la base de entrenamiento y 1 partición para el testeo. Esto se repite en K iteraciones, cambiando siempre la partición destinada al testeo de los datos, para el caso de la Figura 4.9 se tiene que $k = 4$.

De esta forma se consigue que todos los datos de la base de entrenamiento sean revisados. Una vez que se tienen todos los resultados de las validaciones en cada iteración, se procede a calcular la media aritmética de cada iteración obteniéndose un único resultado. El tipo de resultado a evaluar puede variar del objetivo de cada análisis, pero usualmente en clasificación se utiliza la Exactitud, la cual indica la proporción de aciertos sobre la base de prueba, sobre la totalidad de la muestra. Para el caso de regresión se utiliza la función de costos, la cual comúnmente es la suma de residuos cuadrados o residuos de valor absoluto. Si bien lo bueno de esta revisión es que permite ocupar la totalidad de los datos en K particiones, suele ser muy costoso emplear en términos computacionales, sobre todo si se está empleando en Big Data. En cuanto a su precisión, K-Fold Cross validation puede presentar error para los datos no balanceados ya que al optimizarse por la disminución del error, el modelo se verá inclinado a predecir todos los outputs como la clase mayoritaria (desbalanceada) para así dar una exactitud cercana al 100 %.

Matriz de confusión

Esta matriz consiste en una tabla que suele medir el rendimiento de un algoritmo de clasificación, en un conjunto de datos donde los valores del output son valores reales conocidos (y). La matriz de confusión en sí es sencilla de entender, pero la terminología relacionada puede ser confusa[3].

Clase actual	Clase predicha	
	Clase = SI	Clase = No
	Clase = SI	Clase = No
	VP	FN
	FP	TN

Figura 4.10: Ejemplo de una matriz de confusión binaria
Fuente: Elaboración propia

Para un problema de clasificación binaria (problema de caso propuesto), se busca discriminar entre observaciones con un resultado específico, a partir de observaciones normales. Tales como un evento que se cumpla o no se cumpla, de esta forma, se asigna la fila del evento que se cumple como "*positiva*" y la fila sin evento como "*negativa*". A continuación, se asigna la columna de predicciones de eventos como "*verdadera*" y el no-evento como "*falsa*", obteniéndose la siguiente terminología

- "Verdadero positivo" para valores de eventos correctamente predichos.
- "Falso positivo" para valores de eventos pronosticados incorrectamente.
- "Verdadero negativo" para valores sin eventos con predicción correcta.
- "Falso negativo" para los valores sin eventos pronosticados incorrectamente.

Esta predicción resulta fácil de analizar como se muestra en la Figura 4.10, ya que basta con ver la matriz para saber cuantos datos fueron predichos correctamente. Estos se ubican en la diagonal de la matriz. Para el caso de un output no binario, si no categórico, la matriz se evalúa de la misma forma, siendo la diagonal el vector de valores predicho correctamente para cada clase. A diferencia de la exactitud, este método permite evaluar que tan bien se comporta un algoritmo de aprendizaje frente a los datos no balanceados ya que permite visualizar cuantos aciertos hubieron en los eventos en la diagonal de la matriz.

Con esta información almacenada se tienen suficientes ratios para evaluar el rendimiento de una clasificación, entre estos, lo más utilizado son [1]:

1. **Exactitud:** Indica con que frecuencia es correcto el clasificador sobre la totalidad de los datos. No es un buen medidor para base de datos no balanceadas

$$Exactitud = \frac{VP + VN}{VP + FP + FN + VN}$$

2. **Precisión:** Trabaja sobre las predicciones positivas, mostrando cuanto están correctas.

$$Precision = \frac{VP}{VP + FP}$$

3. **Recall:** También llamado Sensibilidad, trabaja sobre los eventos positivos, indicando cuantos de estos fueron predecidos en alguna clase.

$$Recall = \frac{VP}{VP + FN}$$

4. **F1-Score:** es el promedio ponderado de precisión y recall. Por lo tanto, este puntaje tiene en cuenta tanto los falsos positivos como los falsos negativos. Intuitivamente no es tan fácil de entender como la precisión, pero generalmente es más útil, sobre todo para clases desbalanceadas.

$$F1 - Score = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$$

5. **Especificidad:** Trabaja de la misma forma que Recall, pero aplicado a los eventos negativos. Por lo que se considera una medida de integridad:

$$Especificidad = \frac{VN}{VN + FP}$$

6. **G-mean:** Es el producto de las precisiones de predicción para ambas clases, es decir, sensibilidad y especificidad. Esta métrica indica el equilibrio entre los rendimientos de clasificación en la mayoría y la clase minoritaria. Un rendimiento deficiente en la predicción de los ejemplos positivos conducirá a un bajo valor del G-mean, incluso si los ejemplos negativos se clasifican correctamente según el modelo. De hecho, G-Mean es bastante importante para medir la evitación del sobreajuste a la clase negativa y el grado en que la clase positiva está marginada. Los valores que puede tomar G-mean están entre [0,1] siendo lo más cercano a 1 un buen rendimiento del modelo. Su ecuación corresponde a:

$$G - mean = \sqrt{Recall \cdot Especificidad}$$

7. **Razón de verosimilitud:** El ratio de verosimilitud (L) representa la razón entre la probabilidad de predecir un ejemplo positivo cuando verdaderamente es positivo, y la probabilidad de predecir un ejemplo positivo cuando no lo es. Se calcula de la siguiente forma:

$$L = \frac{\frac{VP}{VP+FN}}{\frac{FP}{FP+VN}} = \frac{Sensibilidad}{1 - Especificidad}$$

Para este indicador existe la Tabla 4.1 que especifica los rangos de rendimiento para evaluar el modelo.

L	Rendimiento de modelo
1	Insignificante
1-5	Pobre
5-10	Razonable
>10	Excelente

Tabla 4.1: Umbrales para la interpretación de la razón de verosimilitud

8. **Discriminante de poder:** El poder discriminante evalúa qué tan bien distingue un clasificador entre los casos positivos y negativos. Para ello usa la razón entre sensibilidad y especificidad de la siguiente forma:

$$DP = \frac{\sqrt{3}}{\pi} (\log X + \log Y)$$

Al igual que la razón de verosimilitud, el discriminante de poder cuenta con una tabla para evaluar la comprensión del modelo:

Valor DP	Rendimiento de modelo
<1	Pobre
[1,2]	Limitado
[2-3]	Razonable
>3	Excelente

Tabla 4.2: Umbrales para la interpretación del discriminante de poder

4.4.4.5. Potencial de Machine Learning

A continuación se presenta una infografía sobre las capacidades principales de Machine Learning, donde los tipos de clases de algoritmos superiores, Regresión y Clasificación pertenecen al tipo de aprendizaje supervisado, en cambio los inferiores, Detección de anomalías y cluster pertenecen a los de aprendizaje no supervisado.

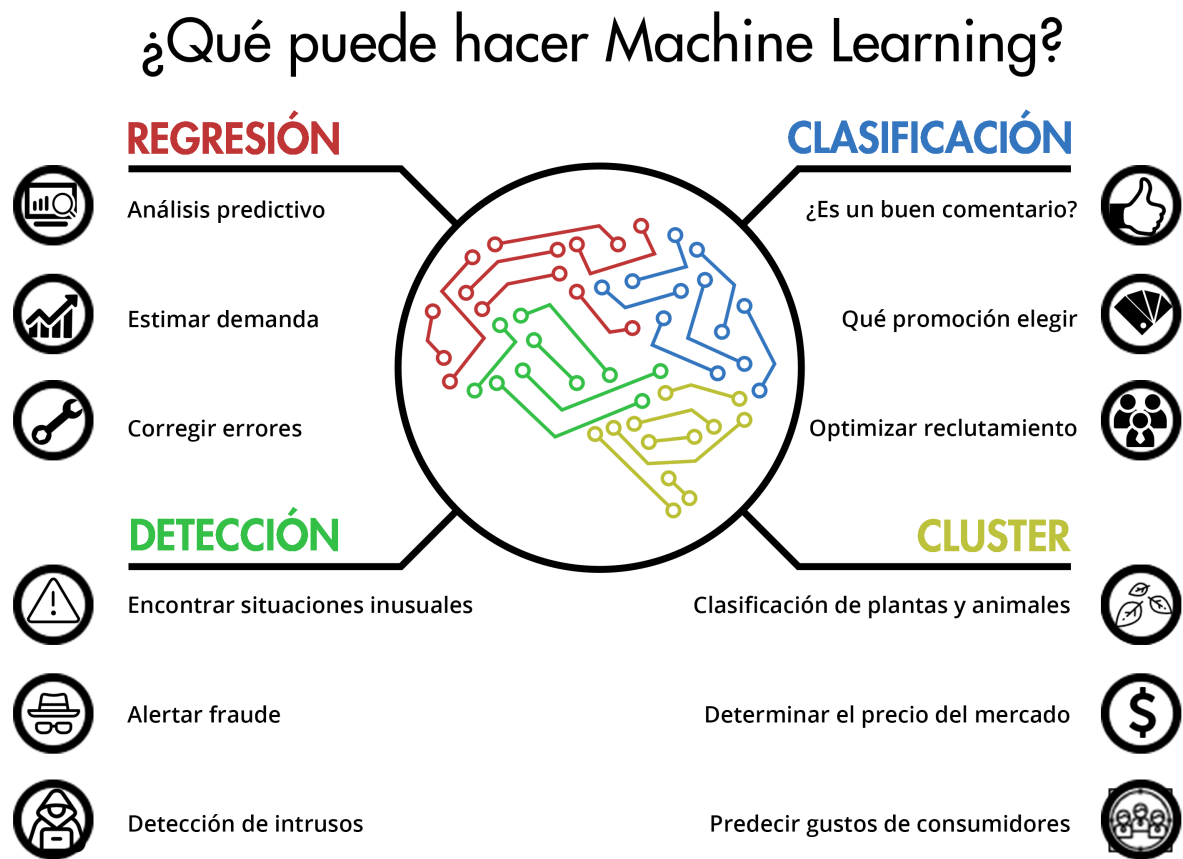


Figura 4.11: Infografía sobre las capacidades de Machine Learning

Fuente: Elaboración propia

Como se mencionó anteriormente sólo se entró en detalles en los algoritmos de aprendizaje supervisado. En el caso de redes neuronales, este se puede usar tanto en algoritmos de clasificación como regresión. A continuación también se integra una lista con los algoritmos más utilizados en Machine Learning.

A continuación se presenta una tabla a modo de resumen, para explicar los algoritmos más usados según la clasificación de la Figura 4.11. En esta tabla además se presenta los beneficios y desventaja de cada algoritmo para identificar en que tipo de base de dato se puede ocupar.

Algoritmos de Regresión

Algoritmo	Beneficio	Desventaja	Resumen
Regresión Lineal	-Uso fácil -Entrenamiento rápido	-Se ralentiza con varios parámetros. -Al estar limitados a curvas definidas, es susceptible a Overfitting.	Ajusta el comportamiento de los datos a una línea, puede ser recta o curva.
Red Neuronal	-Soporta gran cantidad de parámetros -Suele ser muy certero	-Entrenamiento lento. -No siempre sirven para entender el problema debido a su comportamiento de caja negra.	Aplica transformaciones a los datos de entrada a través de varias capas de funciones matemáticas, luego las reúne en una capa final.
Árboles de decisión	-Acertado -Entrenamiento rápido -Gran huella de memoria	-El crecimiento del cálculo se vuelve exponencial mientras aumenta el tipo de problema -Si los datos de entrenamiento son grandes comparado al número de clases, la tasa de error de clasificación es muy alta.	Descompone la base de datos en conjuntos más pequeños. De los subconjuntos crea nodos a partir de las decisiones, con estructura de datos de un árbol.
Support Vector Machine	-Construcción de modelos poderosos -Asegura optimalidad debido a su optimización cóncava.	-Trabaja mejor con base de datos pequeñas. -Se recomienda usar con pocas clases.	Realiza una segregación de los datos a través de hiperplanos. Las dimensiones dependen de la cantidad de clases que tenga el conjunto de datos

Tabla 4.3: Algoritmos típicos de regresión

Fuente: Elaboración propia

Algoritmos de Clasificación

Algoritmo	Beneficios	Desventajas	Resumen
Regresión Logística	-Uso fácil -Entrenamiento rápido	-Se ralentiza con varios parámetros. -Al estar limitados a curvas definidas, es susceptible a Overfitting.	Separa los datos según una línea, recta o curva optimizada.
Bayes de dos clases	-Gran precisión -Entrenamiento rápido	-Sólo soporta 2 clases -Se necesita de una base de datos masiva para que sea certero.	Se apoya en el teorema de Bayes. Asume que no existe relación entre una clase y la otra. Bajo este supuesto calcula las probabilidades de cada clase.
Random Forest	-Certero -Entrenamiento rápido	-Sufre los mismos problemas de árbol de decisión: Crecimiento de cálculo exponencial y alta tasa de error con base de datos pequeña.	Funciona como una iteración de varios subconjuntos aleatorios de árboles de decisiones
Perceptrón de dos clases	-Uso Fácil -Entrenamiento rápido -Permite aprendizaje en linea	-Permite sólo dos clases - Se ralentiza con varios parámetros, debido a la naturaleza de construcción del modelo.	Al igual que la regresión logística permite separar los datos por un modelo lineal, pero permite la integración de datos on-line

Tabla 4.4: Algoritmos típicos de clasificación

Fuente: Elaboración propia

Algoritmos de Clustering

Algoritmo	Beneficios	Desventajas	Resumen
K-Means	-Uso fácil -Entrenamiento rápido (Complejidad lineal) -Maneja base de datos masivas	-Al elegir primeros centroides aleatoriamente, el resultado puede diferir si se reaplica el algoritmo. -La decisión de números de cluster queda a decisión del programador.	Realiza agrupamientos iterativos, intentando encontrar máximos locales. Se repite hasta obtener el centroide de cada grupo
Agrupación Jerárquica	-Números de cluster queda determinado por la naturaleza de lo datos	-No soporta Big Data -Entrenamiento muy lento [O(n² logn)]	comienza con todos los puntos de datos asignados a un clúster propio. Luego, dos clústeres más cercanos se fusionan en el mismo clúster. Al final, este algoritmo termina cuando solo queda un solo grupo.

Tabla 4.5: Algoritmo más usados en Clustering

Fuente: Elaboración propia

Algoritmos de Detección de anomalías

Algoritmo	Beneficios	Desventajas	Resumen
SVM de una clase	-Entrenamiento rápido -Suele ser muy efectivo -Muy usado en las industrias y pruebas de calidad	-Sólo acepta 1 parámetro -Difícil de entender e interpretar el modelo final	La idea es encontrar una función que sea positiva para regiones con alta densidad de puntos y negativa para pequeñas densidades. A diferencia de algunos modelos no funciona bajo los conceptos de probabilidad.
PCA-Based Anomaly Detection	-Entrenamiento rápido -Aplicación simple	-No es fiable si los datos no están correlacionados linealmente. -Al no ser normalizado, los resultados difieren si se cambia la escala de algunos datos.	Realiza un análisis de las características disponibles para determinar qué constituye una clase "normal" la aplicación de métricas de distancia para identificar casos que representan anomalías

Tabla 4.6: Algoritmos más usados en Detección de anomalías

Fuente: Elaboración propia

4.4.5. Hadoop

Hadoop es un sistema open-source que provee un marco de trabajo para enfrentar los desafíos que producen las 3 v del Big Data, (Velocidad, volumen y variedad). Hadoop nació de varios proyectos que se basan en el algoritmo de MapReduce. Este básicamente trata en un algoritmo que permite trabajar con cálculos de varios conjuntos de datos, a través de un seccionamiento de la base de datos y mapeo a través de varios computadores o terminales [29]. Una vez que los cálculos son realizados estos se vuelven a juntar para ser llevados al conjunto de datos resultante. El siguiente diagrama representa el funcionamiento para un mejor entendimiento.

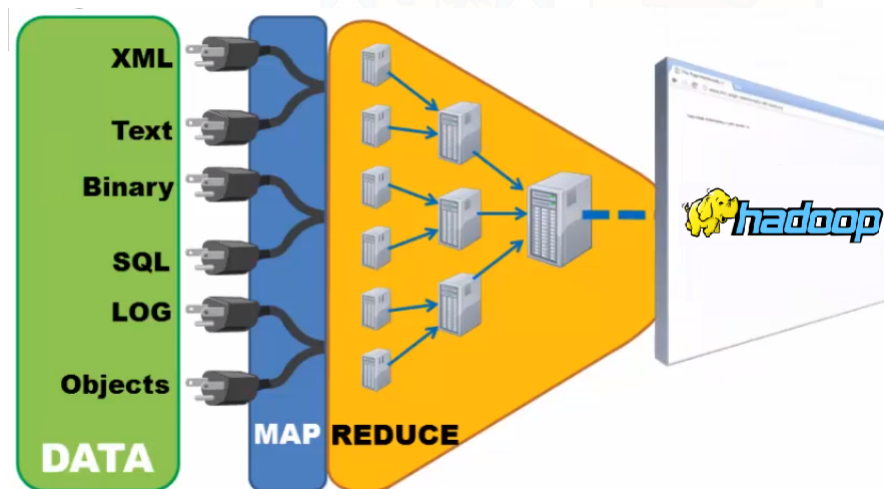


Figura 4.12: Flujo de proceso de información en algoritmo MapReduce

Fuente: youtube.com

De las complejidades que se tienen por la velocidad, variedad y volumen de la información nacen los siguientes desafíos, donde Hadoop plantea una solución para cada uno:

1. **Alta inversión de dinero en la adquisición de un servidor con alta capacidad de procesamiento:** Los clústeres de Hadoop funcionan en hardware común y corriente, estos conservan copias múltiples para garantizar la fiabilidad de los datos. Tiene un límite máximo de 4500 máquinas que se pueden conectar en paralelo.
2. **Gran cantidad de tiempo perdido en solicitar información:** Este proceso se divide en varias partes y se realiza en paralelo, lo que permite ahorrar tiempo. Consta de un límite máximo de 25 datos de Petabyte (1 PB = 1000 TB) usando Hadoop.
3. **Errores al final de una consulta de información significan horas de pérdidas:** Hadoop permite crear copias de seguridad de los conjuntos de datos en todos los niveles que se opere (capas). También ejecuta consultas en conjuntos de datos duplicados para evitar la pérdida del proceso en caso de falla individual. Estos pasos hacen que el proceso de Hadoop sea mucho más preciso.

4. **Dificultad en la aplicación de consultas de programa:** Las consultas en Hadoop son tan sencillas como la codificación en cualquier idioma, ya sea SQL o NoSQL. Solo necesita cambiar la forma de pensar en torno a la creación de una consulta para habilitar el procesamiento paralelo.





5 | Big Data en la Industria



5.1. Teorías de la sostenibilidad en una cadena de suministros

Big Data and Predictive Analysis (BDPA) Se está usando para crear ventajas en el mercado sobre empresas del mismo rubro que no aprovechen este avance [2]. Por lo que pasó de ser una opción a una necesidad donde crea nuevas capacidades estratégicas y mejora los procesos de negocios de las organizaciones. Las formas en que se avanza se observan en la creación de productos de la siguiente generación, transparencia de información y Business Intelligence entre otras.

Estas iniciativas también están marcando una fuerte tendencia en SCM (Cadena de suministros). No sólo se traducen en beneficiar los costos para SCM, si no que que abarca la creación de nuevas estrategias. Por lo tanto el objetivo de esta innovación en SCM es la mejora del resultado final triple, esto quiere decir sustentabilidad en lo **social, financiero** y las **mediciones del rendimiento del entorno**. Estas mejoras se deben aplicar en todas las áreas de la organización para obtener una ventaja en cuanto a la analítica de datos [2]. Lamentablemente no se ha hallado aún una forma de conectar la evidencia de que los conceptos de sustentabilidad en SCM y BDPA estén relacionados directamente. [5]. Por lo que investigadores han desarrollado una teoría de 8 puntos que conecta estos conceptos, y se propone temas de investigación para motivar. La siguiente tabla muestra un resumen de este estudio.

Teoría	Relación	Tema propuesto
Transacción de costos económicos	Mantiene que las transacciones son la unidad más básica de medida, por ende economizarlas es lo central. Estas transacciones se pueden reforzar, con intercambios de información y análisis aplicados por BDPA.	¿Cómo se puede medir los costos y beneficios de BDPA compartido?
Teoría de Agencia	Intenta describir las relaciones entre las entidades donde una delega o asigna trabajo a otra. Esta teoría ha sido usada en estudios de SCM resolviendo problemas de transportación, proveedores de logísticas de terceros y manejo de riesgo.	¿Cómo podría el BDPA optimizar la relación de la agencia que se requiere para la mejora de SCM?
Vista bajo Recurso	Manifiesta que las firmas logran una sustancial ventaja competitiva a través de recursos complejos, valiosos e imperfectamente imitables, por ello la información procesada por BDPA resulta ser útil.	Explorar BDPA como recurso para habilitar el desarrollo de SCM.
Modernización Ecológica	Propone que la innovación ecológica ocurre en la parte superior de la cadena de suministro. Esta provee un enfoque en el cual los estudiantes de, esta teoría podrían atacar problemas de crisis ambientales a través de, estudios de BDPA.	Explorar BDPA como un medio para promover la sustentabilidad enfocado, en las economías emergentes

Tabla 5.1: Resumen de teorías sobre BDPA y SCM

Fuente: Elaboración propia

Teoría	Relación	tema propuesto
Teoría Actor Network	La teoría de Actor-Red (ANT) propone una forma sistemática para considerar los logros tecnológicos del entorno, los cuales tratan las relaciones sociales como efectos de la red. Desde el punto de vista de SCM, ANT puede ayudar a entender la interacción y flujo de información, ya que la red provee el mecanismo para que los datos y análisis sean compartidos entre múltiples firmas y sean utilizados por uno o varios actores a través de la red (ref* borgatti cross)	¿Cómo BDPA mejora la sustentabilidad entre y más allá de los partners de SCM?
Teoría de Capital Social	Sugiere que las firmas son miembros interactivos de SCM, y están, dispuestas a aceptar las normas y valores de SCM. Estas interacciones se refieren a traspaso de información, datos, análisis para el propósito de, mejorar toda la red	¿Cómo se aplica esta teoría en condiciones de competencia entre las firmas?
Teoría de dependencia de recursos	Manifiesta que la empresa que esté falta de recursos desarrollara relaciones de dependencia con otras organizaciones en orden de complementar la falta. El intercambio de información y análisis puede convertirse en un recurso que provea beneficios a corto y largo plazo.	¿Cómo puede ser visto el BDPA como un recurso intercambiable?

Mientras que BDPA se convierta cada vez menos en una innovación y pase a ser más una necesidad, generará un gran impacto en la cadena de suministro, es por esto que este estudio fomenta el interés en aplicar esta ciencia a la ingeniería industrial. Un trabajo de estudio a futuro es necesario y altamente informativo para una comprensión holística de la naturaleza de la firma en que se emplee para saber adecuarse a las ventajas

que brinda sobre el mercado.

Las teorías estudiadas al no ser exclusiva se puede inferir que una mezcla de estos métodos puede ser altamente beneficiosa y posiblemente necesaria para tener un buen entendimiento de las características que entrelazan el desarrollo de BDPA y SCM[5].

5.2. Nuevo enfoque a la administración

La revolución que ha tenido la explosión del Big Data lo ha llevado a ser mucho más poderoso que los estudios de analítica del pasado. Un aumento en la precisión de medición conlleva a una mejor administración de los recursos. Esta hace clara alusión a la famosa frase de Deming y Trucker “No puedes administrar lo que no mides”. Las intervenciones en las empresas son más efectivas, y áreas donde prevalecía la intuición, se han visto desplazada por un método más científico gracias a los datos, información y el rigor de las predicciones echas por modelos. [23].

Las primeras empresas que han nacido digitalmente, como Netflix, Spotify y sobre todo Amazon son un claro ejemplo de que un buen manejo de sus bases de datos[24], ha contribuido a una mejor administración. Estas empresas pueden cumplir metas que ejecutivos de negocios no podrían imaginar años atrás. Después de todo es bien sabido que los negocios online compiten en que tan bien manejan sus datos [2]. Inclusive Peter Norvig, director de búsquedas de Google ha mencionado que la razón de su éxito frente a otros motores de búsqueda no se debe a la calidad de sus algoritmos, si no que ellos tienen mayor cantidad de datos, y superan al mercado en las 3 V's Sección 4.2. Uno de los objetivos de este estudio es aclarar la importancia que tiene la necesidad de transición de mentalidad en los ejecutivos, ya que es una nueva forma de liderazgo. Usando Big Data permite a los administradores decidir a través de la evidencia sobre la intuición.

En la sección anterior se menciona que no existe una clara evidencia que atache las mejoras de SCM con Big Data y pasa algo similar en cuanto a los temas administrativos. La razón que no haya evidencia que usando Big Data de manera inteligente mejore el rendimiento de un negocio se debe a que nadie ha estado investigando ese tópico a conciencia.[2]. Debido a esta razón Peter Norving, junto a funcionarios de analítica de Mckinsey lideraron a un equipo de Massachusetts Institute of Technology (MIT) especializados en Digital Business. [16]. Se partió con la hipótesis que las empresas basadas en el manejo de datos para la toma de decisiones (Data-driven) tienen mejor rendimiento. A través de entrevistas y análisis a las memorias anuales de 330 compañías públicas de Norte América, se encontró que no todas estaban familiarizadas con herramientas de Big Data. . De hecho hubo un espectro grande de las conductas tomadas por las empresas ante distintos problemas, pero una relación sobresalió por sobre todas las otras: *“Mientras más se caracterizaba una compañía por ser Data-Driven, mejor era su rendimiento en los resultados financieros y operacionales”*. Llevado esto a números, las compañías líderes que estaban en el tercio superior de su industria en el uso de Big Data para toma de decisiones, fueron en promedio un 5 % más productivos y 6 % más rentables que sus

competidores [12]. Estos estudios sobresalieron en las áreas de contribuciones de labor, capital, servicios de compras e inversión de tecnología de inversión.

5.2.1. Tomadores de decisiones

En la actualidad una forma típica de toma de decisiones son los llamados **HiPPO**(Highest-paid person's opinion), esto en español quiere decir las opiniones de las personas mejores pagadas en la oficina. HiPPOs son líderes que están tan seguros de sí mismos que no necesitan ni las ideas de los demás ni los datos para afirmar la exactitud de sus creencias instintivas. Confiando en su experiencia e inteligencia, son rápidos para derribar posiciones contradictorias y usualmente pueden desdeñar las opiniones de sus subordinados⁷. Suelen ser las personas más pagadas y caras debido a su larga trayectoria y experiencia. Son una buena alternativa para cuando los datos son escasos, caros de obtener o no esté disponible de manera digital, pero si este no fuera el caso, sería un problema el afirmarse mucho en la experiencia e intuición y no en los datos.

Una nueva forma de desafío en liderazgo nace por los **científicos de datos**, pues cualquiera que administre a un grupo de esta gente debe tener cualidades poderosas, como ser buen comunicador, un hacker de datos, analista y buen consejero[25]. Entonces el desafío para reclutarlos es aprender como identificar esos talentos, atraerlos a una empresas y hacerlo productivo. Lo que hace un científico de datos es hacer descubrimiento mientras se empapa en alguna base de datos. Ellos identifican fuentes importantes de información, las relacionan con otras fuentes, y limpian los conjuntos de datos resultantes. Pero este trabajo no sólo se queda en la parte técnica, pues mientras hacen estos descubrimientos también deben comunicar lo que han aprendido y sugerir consejos para nuevas direcciones en el negocio. Debido a que este trabajo y ciencia están un estado incipiente, generalmente ellos mismo deben trabajar con sus propias herramientas. No es de sorprender que varias herramientas usadas mundialmente hallan nacido en empresas privada, como lo es los servidores compartidos de Hadoop, que vieron sus inicio en Yahoo, o el lenguaje de programación dedicado a los proyectos de Hadoop, que fue concebido por el grupo de científicos de datos de Facebook. Esta serie de característica los hace ser muy bien pagados en grandes empresas, como se menciona en el artículo “*Data scientist: the sexiest job of the 21st century*”, pues son difíciles y caros de contratar, más aún de retener y están en un mercado muy competitivo por sus servicios. En este artículo se les compara a los “quants” de Wall Street en las décadas de los 80 y 90, los cuales eran matemáticos y científicos que se dedicaron al estudio de algoritmos y predicción para la bolsa de valores.

El origen de este trabajo puede provenir desde cualquier campo que se trabaje con base de datos y tenga un foco computacional, no es de sorprender que un científico de datos sea físico, biólogo o inclusive un economista. Actualmente las universidades no imparten alguna carrera que sea netamente sobre esta ciencia pero se prevé que a futuro se generen planes de estudio de esta índole, actualmente North Caroline State

⁷www.forbes.com

University tiene un plan de impartir un plan de estudios llamado Master of Science in Analytics, mientras que otras compañías están intentando desarrollar sus propios científicos de datos. Estos recurrentemente suelen ser confundidos con un analista de datos, pero estos últimos pueden presentar problemas al momento de enfrentarse a una base de datos masiva donde no se tenga datos estructurados, como el ejemplo de modelo de machine learning enfocado a los datos no supervisados. En el artículo *Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management* [12] se hace mención a una investigación sobre la proliferación de esta carrera a través de las búsquedas de Google. De esta fuente se decide emular la investigación, actualizándola desde Marzo del 2004 hasta Marzo del año 2018. la siguiente imagen refleja como ha subido exponencialmente la búsqueda de conceptos como Data Scientist, Data Science, Big Data y Supply Chain Management.

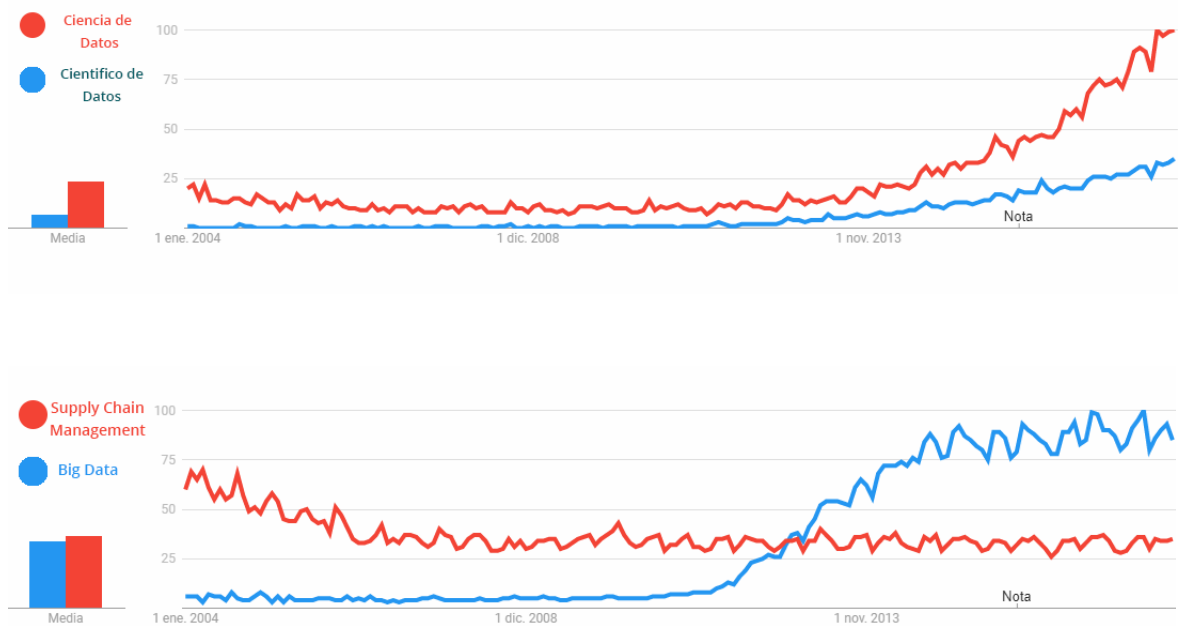


Figura 5.1: Crecimiento de interés sobre ciencia de dato en las búsquedas de Google

Fuente: Elaboración propia

En cuanto a el primer gráfico comparativo, si bien siempre ha existido conciencia de la ciencia de datos, para el caso de alguien que se aplique netamente en esto recién se empieza a notar desde el año 2012. También se aprecia que después de este boom los dos conceptos explotan en las búsquedas, alcanzando niveles que se duplican en tan sólo un año. Este fenómeno puede crear un sin fin de desafíos y oportunidades para cualquier persona que se quiera dedicar a esta ciencia. En internet ya es muy común ver sitios como Coursera, o MIT que ofrecen cursos especializados para la gente que se quiera dedicar a el análisis de datos en Big Data. El segundo gráfico también muestra el crecimiento explosivo que ha tenido este, y también la decadencia que ha sufrido Supply Chain Management en las búsquedas, aunque esto no significa que Big

Data sea más importante que SCM. Tan sólo muestra lo popular que la frase se está volviendo, y al fin de cuentas mientras incrementa el interés por esta carrera, la corriente de talento debería expandirse también. Este fenómeno se puede apreciar más adelante en la Sección 5.3 de estado de arte, a través de los papers que ha salido dedicados a esta ciencia [15].

Considerando lo difícil que es encontrar buenos científicos de datos, más aún en países no tan desarrollados como los de América Latina, se tendería a creer que una buena estrategia para contratarlos puede ser a través de empresas de consultoría, pero según varias entrevistas desarrolladas por Davenport: Estos no buscan sólo dar consejo a alguien encargado de tomar decisiones. Esto se interpreta como una Zona muerta, “*lo único que se hace para estos casos es decirle a alguien más lo que dice que hay que hacer el análisis*”. Su mejor oportunidad para agregar valor a la empresa no es creando reportes o presentaciones para ejecutivos de alto rango, sino, innovando con productos y procesos enfocados directamente al cliente. La Sección 5.3 muestra un ejemplo en la vida real aplicado en la compañía on-line LinkedIn.

5.2.2. 5 Desafíos en la administración

El artículo *Big Data: The Management Revolution*, menciona que las empresas no podrán manejar el cambio abrupto entre un enfoque de Big Data y uno más desactualizado si no pueden cubrir bien estas 5 áreas que se mencionan a continuación:

1. **Liderazgo:** Quienes triunfan no es por mejor o peor base de datos, si no por que el liderazgo de equipo que define bien como luce el éxito, y hace las preguntas correctas. Esto es algo potente pues indica que Big Data no se superpone a la necesidad de una visión humana.
2. **Gestión del talento:** La estadística es importante, pero gran parte de la clave para usar Big Data no se enseña en los cursos de estadísticas. Los mejores científicos de datos son aquellos que se sienten cómodos hablando un lenguaje de negocios y ayudan los líderes a reformular sus desafíos según los caminos del Big Data, ergo, es crucial tener un buen dominio de inteligencia emocional.
3. **Tecnología:** Las herramientas deben estar enfocadas a paliar con la teoría de las 3 V, REFERENCIA. Esto quiere decir poder manejar el volumen, la velocidad de tráfico de datos y la compleja variedad que estos puedan tener. Hadoop es un sistema ejemplar que combina hardware con software open-source.
4. **Toma de decisiones:** Una organización efectiva pone la información y las decisiones relevantes en el mismo sitio. Se ha de ser flexible para evitar el síndrome de “no soy experto en esto”.
5. **Cultura organizacional:** Se deben plantearse las preguntas ¿Qué pensamos? ¿Qué sabemos?, estas preguntas sirven para evitar pretender que se es más sofisticado en manejo de datos a lo actualmente se es. Por eso se debe tener bien concientizados de lo que se es capaz y cuánto se puede mejorar.

5.3. Estado del Arte

5.3.1. Inversiones y utilidades dentro de las empresas digitales

La empresa consultora Mckinsey, que actualmente es una de la más grande de la industria en Data Analytics ha hecho un estudio donde se han encuestado a 3.073 ejecutivos de nivel C- es decir de alto rango que toma decisiones importantes para la compañía- los cuales están conscientes de la tecnología de inteligencia artificial, para saber como están a travesando esta revolución digital hoy en día con relación a los impactos de mercado, financiero y en organizaciones de IA. Para tener un panorama en los centros de inversiones se construyó una imagen global de la actividad en innovación de IA. Cubriendo 10 países sobre 75 ciudades [8], evaluando además su ecosistema local de inteligencia artificial.

De las encuestas y análisis se tiene que Las empresas en la frontera digital -empresas nativas en línea, que vieron sus inicios en la era digital,- están apostando grandes cantidades de dinero en IA. Se estima que la inversión está entre entre **\$ 26 mil millones y \$ 39 mil millones de dólares en 2016**, En cuanto a los grandes gigantes de la industria de la tecnología como Amazon, Apple, Baidu, and Google han invertido aproximadamente entre **\$ 20 y \$ 30 mil millones de dolares** considerando un importante aumento en las fusiones y actividades de adquisición por parte de estas. En cuanto a los otros inversores privados se estima que los capitalistas de riesgo invirtieron **\$ 4 mil millones a \$ 5 mil millones en IA en 2016**, y las firmas de capital privado invirtieron **\$ 1 mil millones a \$ 3 mil millones**.⁸

Estos montos de inversiones eran impensados hace un tiempo atrás puesto que corresponden a un aumento del **300 %** con respecto al año **2013**, o sea tan sólo 3 años atrás. Una inversión adicional de **\$ 1 mil millones** provino de subvenciones y financiación inicial de programas implementados sólo en Estados Unidos.

Netflix también ha logrado resultados sorprendentes a partir de sus algoritmos que utiliza para personalizar las recomendaciones (Algoritmos de Cluster y clasificación) a sus más de 100 millones de suscriptores al rededor del mundo. Es menester ayudar a los clientes a encontrar rápidamente el contenido deseable: Según los analistas de Netflix los clientes tienden a darse por vencidos si les toma más de un minuto y medio encontrar una película o alguna serie de televisión que deseen ver. A través de mejores resultados de búsqueda, Netflix estima que está evitando suscripciones canceladas que reducirían sus ingresos en **\$ 1 mil millones** anualmente. Esto considerando que según los estudios de la compañía digital 20 % de las elecciones de películas viene por las búsquedas mientras que el restante 80 % viene de las recomendaciones[13].

Las empresas que adoptan IA a gran escala o en una parte central de su negocio ya ven la tecnología potencial, y aquellos que implementan estrategias proactivas de IA anticipan beneficios aún mayores. Ac-

⁸Estimación de los gastos de 2016 de las empresas para desarrollar y desplegar productos basados en IA. Calculado para las principales 35 empresas de alta tecnología y fabricación avanzada.

tualmente las categorías de adaptación de la inteligencia artificial tienden a concentrarse en el mismo tipo de industrias. Generalmente las empresas que tienen una alta escalabilidad en IA tienden a ser high Tech como servicio de telecomunicaciones, empresas de ensamble de automóviles y proveedores de servicios financieros, como el claro ejemplo de Delloite y Mckinsey. En la de nivel medio se encuentran las empresas de retail, entretenimiento y medios y bienes de consumo envasado, aunque para el caso de esta última las inversiones suelen ser más en robótica que Big Data. Para el último caso de las empresas que recién adquieren IA o lo hacen de forma más leve se encuentran los servicios de educación, de salud, agencias de viaje y turismo.

A pesar de todos estos avances en la inversión de inteligencia artificial, es limitada la cantidad de productos que actualmente están en el mercado, pero es probable que lleguen allí pronto para impulsar una adopción inmediata y generalizada, al menos así lo prevén los analistas más optimistas con respecto a esta ciencia. Como resultado, los analistas siguen divididos en cuanto al potencial de la inteligencia artificial: algunos han formado un consenso eufórico sobre el potencial de IA, mientras que otros siguen siendo cautelosos acerca de su verdadero beneficio económico. Esta falta de acuerdo es visible en la gran variación de las previsiones actuales del mercado, que van desde \$ 644 millones hasta \$ 126 mil millones para el año 2025 ⁹. Esto tan sólo es otra evidencia del fenómeno de boom-and-bust que se relaciona con las nuevas tecnologías. Aún así en el peor de los casos la suma es abrumadora, lo que es otra razón del énfasis que se debe adoptar por estudiar esta ciencia. Debido al gran partido que aún se le puede sacar.

5.3.2. Impacto de inteligencia artificial en la industria del Retail

Las empresas de retailers ya han iniciado sus inversiones e investigación de la aplicación de la inteligencia artificial en su cadena de valor. Siendo en algunas el núcleo principal para ofrecer un servicio dedicado a sus clientes. Una de las razones más importante es que este avance permite mejorar las actividades hechas hoy por humanos, tales como **promociones, cantidad de surtidos y reposiciones**. Estas tres áreas permiten disfrutar una mejora en el corto y medio plazo. La predicción de demanda, como ya se ha mencionado antes, es una de las actividades más importante, sobre todo para las empresas de retail digitales pues optimiza el almacenamiento y logística en el área de operaciones. Tal es el progreso de esta tecnología, que algunas empresas apuestan por el envío de la mercancía sin tener aún la confirmación de compra, anticipando totalmente la orden del cliente[4].

Una buena predicción de demanda también se traduce en una mejora en el servicio de compra desde la perspectiva del cliente porque permite tomar decisiones más inteligentes. Por parte de la empresa se refleja en el perfeccionamiento de la administración de provisiones y el cálculo de los precios de los productos comparándose con el mercado de forma más seguida. En segundo punto, toda el área de robótica basada en

⁹<https://www.tractica.com/research/artificial-intelligence-market-forecasts/>

Deep Learning podrá incrementar tanto el número de clientes como el monto promedio que estos gastan, porque crearán experiencias de compra personal y convenientes acordes a la necesidad.

La forma en que las empresas puedan predecir el comportamiento de los clientes en tiempo real se debe a las actividades de identificar y aprender de patrones en grandes volúmenes de datos, abarcando muchas fuentes que no estén necesariamente relacionadas entre sí. e.g.: transacciones previas, pronósticos del clima, tendencias en redes sociales, patrones de compras, historial de visitas en línea, análisis de expresiones faciales, patrones de compras estacionales y un sinnúmero de opciones. Este impacto ya se ha probado en la vida real, como es el caso de una empresa de retail europea donde pudo mejorar sus ganancias entre un 1 % y 2 % antes de los intereses e impuestos (EBIT) a través de un algoritmo de Machine Learning para poder anticipar las ventas de frutas y vegetales. Los beneficios técnicos se vieron aplicados la maximización de rotación y la minimización de desperdicios. Otro caso similar es de una empresa de retail online alemana llamada Otto que ha logrado rebajar el sobre stock existente en un 20 % y la reducción de los retornos de sus productos por partes de sus clientes en una suma de dos millones de artículos al año. Esto lo logró con la utilización de Deep Learning para poder analizar las miles de millones de transacciones y predecir lo que los clientes comprarían antes de un pedido. Acorde a los científicos de datos de esta empresa se consiguió una precisión sobre el 90 % al pronosticar lo que podría vender la empresa en sus siguientes 30 días, así Otto logra ordenar más de 200.000 ítems a sus proveedores sin ninguna intervención humana [8].

Todos estos avances pueden ser una barrera de entrada muy competitiva para el comercio minorista, sobre todo en los países desarrollados donde los gigantes del retail pueden hacer inversiones billonarias en inteligencia artificial. El éxito de los minoristas para mantenerse vigentes dependerá en que tanto puedan participar y asegurar el acceso a los datos estratégicos mientras reinventan la experiencia de compra, quizá la formación de alianzas estratégica sea una buena opción para poder competir, o al menos suplir la demanda de los gigantes del retail.

5.3.2.1. Mejoras en el área de operaciones

Las actividades de almacenamiento y logística presentan un amplio conjunto de oportunidades de optimización por la IA. Para algunos minoristas no digitales, especialmente los supermercados, la automatización de las operaciones podría marcar una diferencia radical. Muchos supermercados ofrecen ventas en línea y entrega a domicilio para competir con las tiendas en línea, pero aún así tienen el costo asociado por ocupar las tiendas físicas, así que los costos de servicio en línea borran muchas veces el margen de ganancias del 2 por ciento de la industria. En el Reino Unido 6.25 % se destinan en recoger un artículo del estante, y un 10 % del producto se va en el reparto de este.

La inteligencia artificial también se hace presente en la robótica de las empresas, tal es el caso de Swisslog que ha reducido el tiempo de almacenamiento en un 30 % desde que usa vehículos autónomos en

sus depósitos y galerías. Esto también se ha implementado en el servicio de encomiendas DHL, que utiliza carros automatizados que siguen a los trabajadores de recolección, reduciendo el esfuerzo físico y agilizando la movilidad de sus artículos.

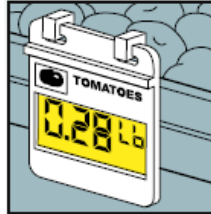
En un área más técnica de geo-localización, minoristas han logrado usar un algoritmo de Machine Learning para mejorar la eficiencia de ubicación de sus surtidos en un 50 %, gracias a esto la empresa de retailer logró aumentar sus ventas entre un 4 % y 6 % determinando el atractivo en el micromercado y minimizando el back order en sus existencias.

5.3.2.2. Proyección del retail a futuro

La siguiente es una lista que proporciona la consulta McKinsey [8], donde se relatan los aspectos que parecen futuristas, pero ya son aplicables a la realidad. Esta nueva red de tecnologías de inteligencia artificial se basa en la rama del aprendizaje de máquinas, y manejo de Big Data.



El software de reconocimiento facial, el aprendizaje automático y el lenguaje natural permiten a los agentes virtuales saludarlo personalmente, anticipar pedidos y brindar instrucciones



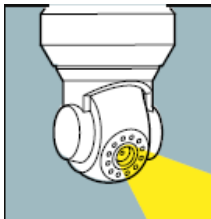
Las tiendas actualizan y optimizan los precios en tiempo real, con el aprendizaje automático aprovechando los datos sobre los precios de los competidores, el clima y los niveles de inventario para maximizar los ingresos



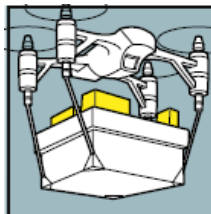
A través de Machine Learning se personalizan las promociones para que coincidan con los perfiles de los compradores; los beacons en la tienda envían ofertas a sus teléfonos inteligentes mientras navegan por la tienda



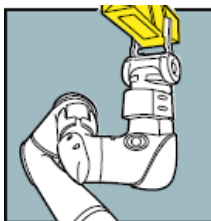
Las pantallas interactivas y los tableros de mesa habilitados con visión por computador y aprendizaje profundo pueden identificar artículos y recomendar productos y usos complementarios que se ajusten al perfil de estilo de vida de los compradores.



La visión por computadora con aprendizaje profundo identifica los artículos en bolsas de los compradores; al agregar datos de los sensores, AI permite el pago sin control y el pago automático



Los drones autónomos que utilizan la tecnología de aprendizaje profundo completan la entrega en el último tramo y son capaces de manejar los obstáculos o los destinatarios ausentes.



Los robots mejorados con IA realizan un seguimiento continuo del inventario, reconocen los estantes vacíos y los reponen; otros robots llenan bolsas en el almacén.

5.3.3. Big Data en Finanzas

La finanza cuantitativa es un área en la cual los datos son la información vital que se procesa en todos los aspectos. Las principales empresas y firmas financieras están adoptando tecnologías Big Data avanzadas para obtener información manejable de datos masivos de mercado, estandarizando datos financieros de una variedad de fuentes, reduciendo el tiempo de respuesta a flujos de datos en tiempo real, mejorando la escalabilidad de algoritmos y gran cantidad de software de tecnología de punta. Hoy en día, estas importantes ganancias están impulsando a los pioneros de los profesionales financieros a desarrollar y desplegar soluciones de Big Data en productos financieros, que van desde la negociación algorítmica de front-office hasta la administración y análisis de datos de back-office[17]. El análisis de Big Data puede ayudar a revelar más oportunidades de mercado ocultas mediante el análisis de datos estructurados de gran volumen y noticias sociales, en contraste con los de bajo desempeño que son incapaces de adoptar técnicas novedosas.

El concepto de “Big” en el contexto de la industria financiera es diferente de lo que es en contextos científicos o minoristas. En las empresas minoristas, por ejemplo, el análisis del perfil de los clientes implica principalmente el análisis de datos no estructurados de las fuentes de las redes sociales, tal como se menciona en la sección de herramientas. Sin embargo, los mercados financieros se ocupan principalmente de datos estructurados recopilados de un conjunto limitado de fuentes, como intercambios y proveedores de datos. Los desafíos de los grandes datos en el contexto financiero generalmente se refieren a proyectos que involucran múltiples factores, como grandes volúmenes de datos complejos que deben tener referencias cruzadas en un marco de tiempo específico[30]. Si bien esto no siempre es necesario los desafíos contemplado también incluyen de identificación de clase heterogénea de activos e información de riesgo, implementando complejas agregaciones de datos para informes normativos ad hoc, análisis de crédito, generación de señales de negociación o gestión de riesgos para instancias, mientras se reducen las latencias de agregación y aumento de datos.

De un estudio actualizado de *guide to big data for finance*, se resume una lista de las actividades en la que el manejo de Big Data y control de finanzas están relacionados:

5.3.3.1. Puntuación de crédito

Hasta que los servicios de puntaje de Big Data no estaban disponibles, este enfoque ha tenido poca innovación al convertir el puntaje en una mercancía. Con las tecnologías de big data, por ejemplo, algoritmos de aprendizaje automático, las decisiones de préstamos y créditos se determinan en segundos mediante procesos automatizados. En algunos casos, la tecnología puede usar puntos de datos a escala millonaria para evaluar los puntajes crediticios de los clientes en tiempo real. Estas tecnologías ofrecen una visión integral de 360 grados de cualquier posible cliente, basándose en sus familiares, colegas e incluso en su navegación web.

Esto finalmente ayuda a expandir la disponibilidad de crédito a aquellos que luchan por obtener préstamos justos. La investigación ha demostrado que todo, desde la inclinación política de los usuarios hasta la orientación sexual, ahora se puede predecir con precisión mediante el análisis de la información disponible públicamente en redes sociales como Facebook y Twitter.

5.3.3.2. Trading algorítmico

A principios de la década de 1990, los intercambios más grandes adoptaron el "motor de emparejamiento" electrónico para reunir compradores y vendedores, lo que se denomina el origen del comercio cuantitativo. Para el año 2009, más del 70 % del volumen de negociación de acciones de EE. UU. Fue contribuido por el comercio cuantitativo. La negociación algorítmica utiliza principalmente datos históricos enormes para calcular la proporción de éxito de los algoritmos que se escriben. Los algoritmos evalúan miles de valores con complejas herramientas matemáticas, mucho más allá de la capacidad humana. También combinan y analizan datos para revelar ideas que no son evidentes para el ojo humano. Aquí es donde ocurre la verdadera innovación: hoy en día, tenemos una cantidad interminable de datos disponibles, y con las herramientas adecuadas, el modelado financiero se ve limitado solo por el poder del cerebro y la imaginación del quant en el trabajo. Las técnicas de Big Data son las herramientas correctas que pueden facilitar el ciclo de generación alfa del desarrollo-gestión del despliegue, que es el proceso que cada inversionista serio ha estado tomando. En base a las funcionalidades, el proceso se puede categorizar como Administración de datos, Desarrollo de estrategias y Despliegue de productos, donde se podrían implementar varias técnicas de Big Data.

5.3.3.3. Gestión de riesgos

A medida que la institución se concentra más, los mercados se vuelven más interconectados y la información atraviesa las conexiones más rápidamente, la complejidad ha crecido en todos los aspectos de la industria. Mientras tanto, los riesgos aumentan con la complejidad. Las demandas para mejorar el monitoreo del riesgo, la cobertura de riesgos y el poder predictivo de los modelos de riesgo nunca han sido tan altos. Las tecnologías Big Data, junto con miles de variables de riesgo, pueden permitir a los bancos, administradores de activos e instituciones de seguros detectar de forma proactiva los riesgos potenciales, reaccionar de manera más eficiente y más efectiva, y tomar decisiones sólidas. Big Data puede estar dirigido a las necesidades particulares de una organización y aplicarse para mejorar diferentes dominios de riesgo, tales como: Riesgo de crédito, riesgo de liquidez, Riesgo de crédito por contraparte y riesgo operacional.



6 | Caso de estudio aplicado

Para el caso de estudio aplicado a Big Data se aborda una propuesta hecha por INFORMS, The Institute for Operations Research and the Management Sciences. Esta es una asociación profesional dedicada a promover buenas prácticas y avances de investigación en operaciones y análisis para mejorar los procesos operativos. La propuesta consta de un desafío en el área de mantenimiento y prevención de fallos para un sistema ferroviario utilizando base de datos masivas las cuales no pueden ser operadas por programas comunes como Excel o MiniTab. Se ofrece una base de datos de entrenamiento y una de testeo las cuales deben cumplir ciertas expectativas del desafío. Cabe destacar que estos datos son reales y provienen de una empresa analista encargada de monitorear el estado y mantenimiento de las ruedas en los trenes en Holanda.

6.1. Introducción al desafío

Dentro de todas las características otorgada por la base de datos se tiene como medición principal los "Peak kips" de cada rueda, estos denotan la fuerza vertical total que una rueda impone sobre el riel. Un "Kip" es una unidad de fuerza que equivale a 1,000 [libras-fuerza], y se puede medir con detectores en tierra ubicados en los rieles. Cuanto mayor sea el valor máximo de los kips, mayor será el daño impuesto tanto a las ruedas como a las pistas. Los ferrocarriles usan esta medida para detectar si hay defectos en las vías o las ruedas. Actualmente se activa una alarma si el valor máximo de los kips es superior o igual a 90, y los carros involucrados y los móviles identificados se deben someter a una inspección y posterior reparación en caso de ser necesario, ya sea inmediatamente o en el próximo lugar de reparación disponible. Aunque no se requiere una acción inmediata para las ruedas con valores máximos de kips cercanos a 90, el daño a las vías y ruedas puede aumentar progresivamente (y los Peak Kips de una rueda defectuosa pueden aumentar aún más) si las ruedas defectuosas no se estabilizan a tiempo.

Sin embargo, instalar un móvil cuando está cargado puede causar muchas pérdidas a la compañía ferroviaria. El envío podría retrasarse y el tráfico de la red podría sufrir interrupciones innecesarias, lo que se traduce en un castigo económico para la empresa. Por lo tanto, la compañía ferroviaria tiene un incentivo para identificar proactivamente ruedas y vehículos problemáticos y excluirlos de ser utilizados para envíos

futuros.

Este proyecto tiene como objetivo construir un modelo predictivo que se pueda usar para proyectar los peak kips de un carro actualmente vacío pero que esté designado a estar cargado en el estado más próximo. Esto implica que la proyección se calculará cuando el automóvil esté vacío. Este modelo ayudará a la compañía ferroviaria a excluir las ruedas y/o autos problemáticos de un viaje futuro, y evitar las pérdidas por negligencias y defectos del ferrocarril.

6.2. Descripción del problema

Cada Peak Kip está compuesto de dos componentes:

$$Peak\ Kip = Kips\ promedio + Kips\ dinamicos \quad (6.1)$$

Donde el kip promedio corresponde a la parte estática de la fuerza del kip asociado al peso del carro. Para el caso de que el carro esté vacío, el kip promedio debería ser bajo, y cuando este esté cargado, el kip promedio debería ser considerablemente mayor. En cuanto al kip dinámico, este se explica como la parte de el kip de fuerza causado por los daños y defectos de las ruedas.

Dentro de las mediciones se encuentra el ratio dinámico, y este se define como los peak kips dividido por los kips promedios. Esto representa la porción de la fuerza total que se debe por los impactos sufridos, a diferencia de la proporción del peso. La Figura 6.1 ilustra la relación entre los peak kips, kips promedios y kips dinámicos

De la ecuación Ecuación 6.1, se infiere que los peak kips de un vehículo vacío son siempre menores a uno estado cargado. En adhesión, una rueda con defectos causará mayo impacto que una con menor defecto bajo el mismo estado de “cargado”. Esto provee una razón de porque los peak kips de algunas ruedas súbitamente se disparan desde una cantidad menor a 50, a una sobre 100 cuando un carro se carga con materia prima. Este comportamiento se puede apreciar en la Figura 6.2 , la cual muestra como el peak kip de una rueda puede cambiar de manera abrupta cuando pasa por la transición de estar en un estado vacío a uno cargado.

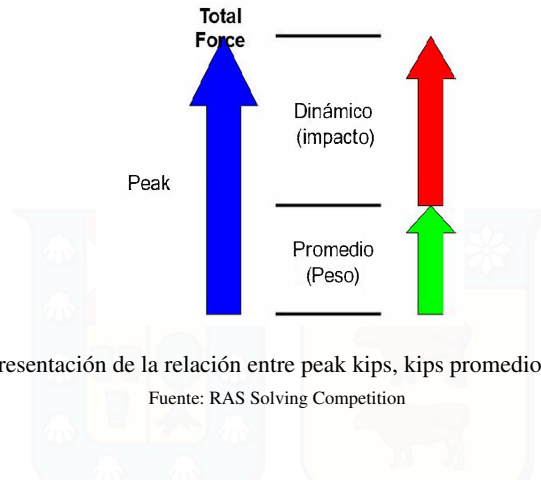


Figura 6.1: Representación de la relación entre peak kips, kips promedios y kips dinámicos

Fuente: RAS Solving Competition

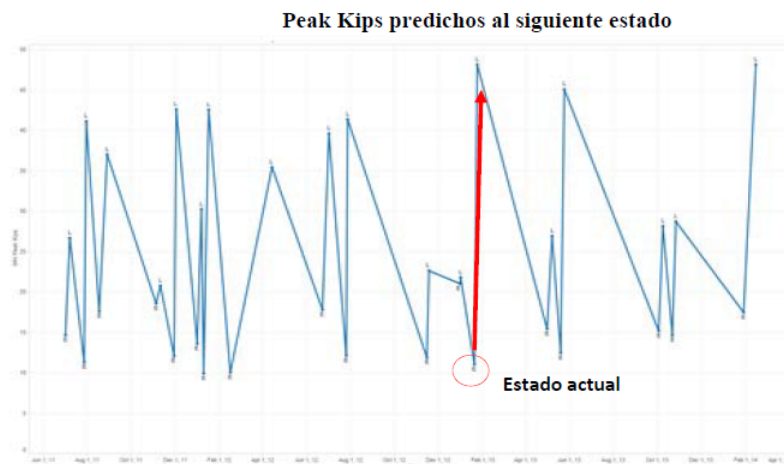


Figura 6.2: Peak kips de una rueda transicionando entre una estado *vacío* y *lleno*

Fuente: RAS Solving Competition

El **objetivo del proyecto** es predecir los valores de los peak kips de un carro vacío en el siguiente estado cargado. Esto apunta a obtener una medida preventiva antes de asignar un tonelaje a los vagones. Para esto se provee de un conjunto de datos empíricos para una base de datos de entrenamiento y una de testeo.

El objetivo se despliega en 2 partes :

1. Predecir el valor del peak kip en el siguiente estado cargado. La solución deseada debería tener una diferencia entre el valor predicho y el verdadero de un rango entre $[-2,2]$ para un 80 % de las observaciones en el conjunto de datos.
2. Predecir cuando una alarma será desencadenada en el siguiente estado cargado. Se considera un evento positivo (Alarma provocada) cuando los peak kips alcancen una lectura sobre 90. El criterio de evaluación corresponde a obtener la mayor cantidad de positivos verdaderos mientras se minimizan los falsos positivos en el conjunto de datos de testeo.

De estos objetivos, se generan dos propuestas distintas de como se planteará el desafío:

Propuesta 1	Propuesta 2
-Para realizar una predicción el tiempo t , se usarán los inputs de tiempos anteriores. ($t-1, t-2, \dots$).	-Para realizar una predicción en el tiempo t , se usarán los inputs del tiempo actual.
-Se podrán usar las 22 características para entrenar la base de datos.	-Se descartarán variables que están relacionadas a la presión (Presión dinámica, promedio, Ratio dinámico) para evitar una correlación.

Tabla 6.1: Propuestas para abordar el desafío

6.3. Resolución del desafío

Para la resolución del problema se decide ocupar la metodología de DataMining aplicando como herramienta principal el aprendizaje de máquinas (Ver Subsección 4.4.4), ya que es una de las ciencias más utilizadas para tratar Big Data. Esto se explica por muchos motivos, entre ellos destacan la gran cantidad de modelos que permite abordar problemas desde varias perspectivas, tanto lineales, como no lineales, estadísticos y otros no basados en la estadística si no la pura optimización de sus parámetros. La basta cantidad de documentación y su fácil acceso tanto en internet como en libros es otro de los motivos del porque es una de las herramientas preferidas por los científicos de datos. Por último la versatilidad para manipular información, ya que puede manejar lenguaje supervisado tanto como no supervisado lo hace una herramienta capaz de abordar todo tipo de problemas.

En cuanto a la herramienta de hardware a ser ocupada, se utilizará un computador Laptop de oficina. Este es un Dell Latitude 3450 con un procesador Intel®Core™ i5-5200U CPU @ 2.20 GHz (Correspondiente a 4 CPUs), con 8192 MB de RAM y sistema operativo KALI Linux, debido a la memoria que consume Windows por defecto es mayor a la de Linux, lo que se manifiesta en una compilación más rápida y efectiva, pues se evita la falla de memoria por sobre procesamiento. Cabe destacar que la cantidad de CPUs y la memoria RAM son fundamentales para el proceso de tratamiento de datos, sobre todo si es masivo porque generalmente estas herramientas trabajan bajo la paralelización del procesador, es decir una tarea se divide en la cantidad de CPU que tenga el sistema en forma paralela, sólo si la memoria RAM lo permite.

Para la implementación de machine learning se decide utilizar el lenguaje Python 2.7.19. Esto porque el alumno ya contaba con nociones básicas de programación y además existe una gran documentación sobre como usar algoritmos de aprendizaje en Python. La variedad de módulos creados para este ámbito permite que la manipulación de la base de datos y algoritmos sea más cómoda y configurable de usar. Entre los

módulos utilizados, se destacan los siguientes:

- NumPy: Proporciona una manipulación de alto rendimiento en objeto de matriz multidimensional, además de herramientas para trabajar con estas matrices
- Pandas: Se usa para estructuras de datos de alto rendimiento, siendo estas fáciles de usar. Con herramientas de análisis de datos.
- Dask: Simil a pandas, pero permite la paralelización del procesador para una lectura más rápida. Su desventaja es que proporciona menos herramientas de manipulación de datos.
- Scikit-Learn: Contiene algoritmos y modelos de machine learning listos para ser usados, como complemento proporciona herramientas de evaluación de rendimiento.
- Keras: Al igual que Scikit-Learn, proporciona algoritmos de machine learning, pero incluye redes neuronales profundas, además de distintos optimizadores de algoritmos.
- TensorFlow: Se usa como complemento a los módulos de Keras y Scikit-Learn. Este posibilita el procesamiento por paralelización de procesador por lo que acelera el proceso de predicción y evaluación.

6.3.1. Canalización de datos

Para explicar el detalle del proceso, se usará la metodología KDD mencionada en la Subsección 4.4.3, dividiendo el trabajo en 5 procesos fundamentales: **Preparación de datos, Creación de modelo, Evaluación, Optimización y por último Predicción de datos**, la cual usará la base de datos de testeo. En la Figura 4.4 se puede visualizar la forma en que se reparten estas actividades.

6.3.1.1. Preparación de datos

Esta etapa concierne desde la ingesta y recopilación de datos, hasta el filtrado y limpieza para ser entregado al modelo. Es una de las actividades más extenuantes ya que la calidad en que la información es entregada suele estar desordenada y dañada, de tal forma que los modelos no puedan interpretarlo directamente. Se necesita de una copiosa manipulación de estructura de datos para trabajar con datos reales.

Para el caso del desafío no es necesario recopilar datos, ya que vienen entregados por la INFORMS. Los archivos a tratar consisten en 2 base datos, "*Training.data*" y "*Test.data*". Estas tienen aproximadamente 7.5 millones y 2.5 millones de muestras, con un peso de 1.1 Gb y 0.4 Gb respectivamente. Ambas bases de datos constan de 22 columnas. Para la propuesta 1, se considerará inicialmente 22 features y un output, en cambio para la propuesta 2 serán 19 features y 1 output. Cada muestra viene separada por un salto de línea, y separadas entre ellas por tabulación, esto último sólo es importante en términos de programación.

La siguiente lista recopila todas las columnas en ambas bases de datos, indicando si estas son de clase continua, binaria o categórica:

- **unq_whl_id** : Identificador único de cada rueda, que consta de la inicial y número de carro, lado y posición de eje, fecha de inicio y término de cada control. Este imput se usará como indicador para diferenciar cada rueda a evaluar. **(Categoría)**
- **load_empty**: Indica si el carro viene en estado cargado o vacío , con una *L* y *E* respectivamente. **(Binaria)**
- **trn_id**: Identificador de tren, donde su primer carácter determina el tipo de tren. **(Categórica)**
- **vnder_edr_ts_cst**: Fecha de la lectura de kips. **(Continua)**
- **vndr_edr_nme** : Identificador de los detectores. **(Categoría)**
- **edr_vndr**: Proveedor del sensor, este puede ser *PRT* o *SALIENT*. **(Binaria)**
- **age**: El día en que fue hecha la medición. **(Continua)**
- **axle_side** : Indica el lado del equipo al que está unida la rueda, con una *L* y *R* para especificar si es de la izquierda o derecha respectivamente. **(Binario)**
- **eqp_axle_nbr** : La posición del eje donde se encuentra la rueda. **(Categoría)**
- **eqp_seq_nbr**: La posición del vagón en el tren de la rueda que se está controlando. **Categoría**
- **edr_eqp_spd** : La velocidad de la rueda en el momento de la medición. **Continua**
- **eqp_grs_tons** : El tonelaje del equipo del que forma parte la rueda en el momento de la medición. **(Continua)**
- **tare**: El peso del vagón cuando está vacío medido en toneladas. **(Continua)**
- **aar_ct_c**: Designación del tipo de carro. **(Categoría)**
- **eqp_grp** : Agrupación de diferentes carros según su tipo. **(Categoría)**
- **grs_rail_wgt**: El tipo de rueda que representa el tonelaje que puede soportar el tipo de rueda en particular, junto con las otras ruedas unidas como parte del mismo automóvil. Un automóvil generalmente tiene varias ruedas del mismo tipo conectadas a él. **(Categoría)**
- **Kipdays**: representa la tensión acumulativa a la que está sometida una rueda a lo largo del tiempo debido a las cargas de impacto. Kip-Days se empiezan a acumular cuando: $\text{Peak} \geq 65 \text{ Kips}$ O $\text{Dynamic} \geq 30 \text{ Kips}$ O $\text{Ratio} \geq 3.0$; Si ninguna de las condiciones de Kipdays se cumple, se otorga -999 de valor. **(Continua)**
- **whl_avg_kips** : Peak Kips promedio. **(Continua)**

- **whl_peak_kips** : Medición de Peak Kip. (**Continua**)
- **whl_dyn_kips** : Kips dinámicos. (**Continua**)
- **whl_dyn_ratio** : Ratio de Kips dinámicos. (**Continua**)

Supuestos

1. Se asume que el siguiente estado cargado a predecir debe estar dentro de los 10 días siguientes desde último estado vacío.
2. Sólo se estará interesado en predecir el siguiente estado cargado cuando el estado inmediato anterior es vacío.
3. Se consideran mediciones de kips fiables sólo cuando los detectores leen una velocidad mayor a los 20[MPH].
4. Puesto que la idea es predecir la presión ejercida en el futuro, se asumirá que no se tiene ninguna medición con respecto a las otras características del futuro (Propuesta 1).
5. Si la medición de una rueda en un vagón vacío sobre pasa los 90 [kips] no se considerará válida.
6. Cualquier medición de WHL_PEAK_KIPS sobre 90[Kips] se considerará como detonante de *Alarma*.

Desarrollo Las bases *Training* y *Data* viene por defecto desordenadas en sus fechas, La Figura 6.3 muestra un ejemplo didáctico para entender como vienen configuradas estas bases de datos. Hay que considerar que se han cambiado el nombre a algunas variables (Características) para facilitar la comprensión en las transformaciones posteriores.

ID_Rueda	Presión	Estado	...	Fecha
ID_B	90	Cargado	...	17-jun
ID_A	18	Vacío	...	07-may
ID_B	47	Cargado	..	03-jun
ID_A	46	Cargado	...	12-may
ID_A	17	Vacío	...	01-may
ID_B	22	Vacío	...	14-jun
ID_A	45	Cargado	...	09-may

Figura 6.3: Ejemplo simplificado de la base de datos

Fuente: Elaboración propia

La variable *vnder_edr_ts_cst* (**Fecha** en ejemplo) representa la fecha y hora de medición, teniendo en cuenta esto, se decide organizar la base de datos según el ID de la rueda, y luego por su fecha. Gracias a la manipulación de datos del módulo pandas esto se realiza en un código compacto A.1.0.0.1.

Ya teniendo la base de datos ordenada, se procede a aplicar los supuestos 1,2 y 4. Para la propuesta 1 se creará una columna nueva, la cual será el nuevo output. Este consistirá en la medición futura de presión en

el siguiente estado cargado. Esta nueva medición en términos de programación y de resolución del problema deberá contemplar 3 condiciones a través del código A.1.0.0.2:

- La presión futura deberá pertenecer al mismo ID de la rueda
- El siguiente estado deberá ser cargado
- La diferencia entre los días debe ser menor o igual a 10

La Figura 6.4 muestra como luce la nueva base de datos, donde se tiene ordenado primeramente por ID de la rueda y luego por las fechas correspondiente a su medición. Parte de la limpieza de los datos será remover las mediciones las cuales no estén interesados, para esto se deshacerán todas las muestras donde haya presencia de valores *Null* (Vacío en Python).

PROPUESTA 1					
ID_Rueda	Presión	Estado	Fecha	Presion_futura	Explicación
ID_A	17	Vacío	01-may	Null	Siguiente estado es vacío
ID_A	18	Vacío	07-may	45	Cumple las condiciones
ID_A	45	Cargado	09-may	46	Cumple las condiciones
ID_A	46	Cargado	12-may	Null	Última medición de rueda
ID_B	47	Cargado	03-jun	Null	Está sobre 10 días
ID_B	22	Vacío	14-jun	90	Cumple las condiciones
ID_B	90	Cargado	17-jun	Null	Última medición de rueda

Figura 6.4: Ejemplo simplificado de base de datos con presión en tiempo $t + 1$

Fuente: Elaboración propia

Para la **Propuesta 2**, no es necesario la adquisición de una nueva columna ya que los datos de entrada serán del periodo actual a la presión a predecir. Por lo tanto sólo será necesaria la remoción de las variables que tengan una correlación de ante mano sabida con la variable de salida, es decir, la variables de presión, como por ejemplo la presión dinámica, ratio dinámico y la presión promedio. La siguiente tabla ilustra el ejemplo de remoción de una de estas variables.

PROPUESTA 2						
ID_Rueda	Presión	Estado	Presión	Dinámica	Carga	Fecha
ID_A	17	Vacío	57	1 ton	01-may	
ID_A	18	Vacío	60	2 ton	07-may	
ID_A	45	Cargado	13,0	3.5 ton	09-may	
ID_A	46	Cargado	13,3	3.4 ton	12-may	
ID_B	47	Cargado	13,7	3.6 ton	03-jun	
ID_B	22	Vacío	73	1.7 ton	14-jun	
ID_B	90	Cargado	30,0	3.9 ton	17-jun	

Figura 6.5: Ejemplo simplificado de base de datos con presión en tiempo $t + 1$

Fuente: Elaboración propia

Para tratar el problema como clasificación se ha de transformar el output (y) como una variable binaria. Esto se hará según el criterio del supuesto 6, es decir, cualquier medición $y > 90$ Kips, se asignará como 1, para el caso que no se cumpla, 0.

De manera similar se deben transformar las variables binarias de las características para un formato que sea legible para los modelos de machine learning. Estas características en su origen, vienen con un formato

string. Aplicando el código de anexo A.1.0.0.3 se logra transformar en formato binario, es decir, adquieren valores de 0 y 1. Este mismo código se aplica a las variables: *load_empty*, *edr_vndr*, *axle_side*

Generalmente en la fase de preparación de datos, uno de los pasos a seguir es borrar las muestras donde se presenten datos outliers, esto se aplica a el desafío de regresión en el anexo A.1.0.0.6 . Es importante destacar que para el caso específico de clasificación esto no se hace, pues se está tratando con una base de datos **NO BALANCEADA**. Es decir la proporción entre *Alarmas* y *No Alarmas* es desequilibrada. Por ejemplo, para la propuesta 1, a base de datos nueva, con todas las transformaciones que se han sucedido queda con un total de 16.466 Alarmas y 3.587.136 de no alarmas, dándose un ratio de 0.9945 entre eventos negativos y positivos. Por lo que se considera que los datos outliers pueden ser significativos para poder identificar outputs la clase minoritaria.

Ya teniendo las variables cuantitativas y discretas, es necesario procesar las de condición categórica. Estas son aquellas que pueden un valor de una lista limitada, generalmente fija, de posibles valores. Estas son de características cualitativas y pueden venir tanto en formato numérico como string. Las variables que cumplen con esta condición son las siguientes: *TRN_ID*, *VNDR_EDR_NME*, *EQP_AXLE_NBR*, *EQP_SEQ_NBR*, *AAR_CT_C*, *EQP_GRP*, *GRS_RAIL_WGT*. En el caso de *TRN_ID*, fue necesario dividir el string, y adquirir sólo la primera letra, pues determina el tipo de carga del tren. A priori esto se considera que puede ser importante para una evaluación futura. El código se aprecia en el anexo A.1.0.0.5. Este paso es requerido tanto para el desafío de clasificación y regresión en ambas propuestas.

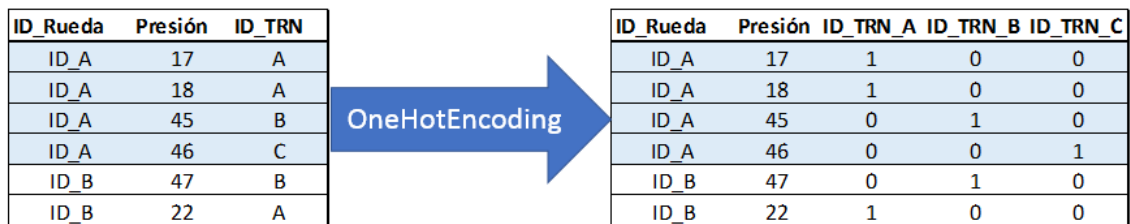


Figura 6.6: Ejemplo ilustrativo de transformación de la base antes y después de OneHotEncoding

Fuente: Elaboración propia

El proceso de convertir las variables categóricas en una entrada legible el modelo no basta con cambiar el formato string a uno numérico, ya que el modelo puede considerar que se está tratando con variables cuantitativas y no cualitativas, por ello se hace el proceso a través de **ONE HOT ENCODING**. Este lo que hace es convertir una entrada categórica en varias sub-entradas binarias, donde cada categoría pasa a ser 1 en el caso que se cumpla, o 0 si no. La Figura 6.6 relata un ejemplo aplicado a la variable *TRN_ID*. Donde de sus categorías A, B y C se han generado 3 categorías nuevas y eliminado la original *TRN_ID*.

Al final de este tratamiento lo que se obtiene es una base de datos legible, sólo queda aplicar los supuestos 5 y 6. Con esto se cambiarán los valores que no cumplan las condiciones: $edr_eqp_spd \leq 20[MPH]$

y $whl_peak_kips \geq 90[Kips]$ si $load_empty = 0$. La remoción se hace de la misma forma que la primera limpieza, asignando *Null* a cada dato que no cumpla la restricción. Esta acción se emplea en el anexo A.1.0.0.5.

En la siguiente tabla se adjunta la lista de inputs que recibirá cada modelo según su propuesta:

Datos de entrada (X_i)					
Propuesta 1			Propuesta 2		
EQP_AXLE_NBR_7	EQP_AXLE_NBR_6	EQP_AXLE_NBR_5	AAR_CT_C_A	AAR_CT_C_B	AAR_CT_C_C
EQP_AXLE_NBR_4	EQP_AXLE_NBR_3	EQP_AXLE_NBR_2	AAR_CT_C_E	AAR_CT_C_F	AAR_CT_C_G
EQP_AXLE_NBR_1	AXLE_SIDE	AGE	AAR_CT_C_H	AAR_CT_C_J	AAR_CT_C_K
trn_id_V	WHL_AVG_KIPS	EQP_AXLE_NBR_8	AAR_CT_C_L	AAR_CT_C_M	AAR_CT_C_P
EQP_GRP_IFLT	WHL_DYN_RATIO	VNDR_EDR_NME_CHEL	AAR_CT_C_Q	AAR_CT_C_R	AAR_CT_C_S
VNDR_EDR_NME_PEST	VNDR_EDR_NME_BAGD	WHL_PEAK_KIPS	AAR_CT_C_T	AAR_CT_C_V	AGE
VNDR_EDR_NME_DEAN	VNDR_EDR_NME_BMRK	VNDR_EDR_NME_PRAI	AXLE_SIDE	EDR_EQP_SPD	EDR_VNDR
EQP_GRP_BOX_C	LOAD_EMPTY	VNDR_EDR_NME_AURO	EQP_AXLE_NBR_1	EQP_AXLE_NBR_2	EQP_AXLE_NBR_3
VNDR_EDR_NME_COCH	VNDR_EDR_NME_ROGN	EQP_GRP_MFLT	EQP_AXLE_NBR_4	EQP_AXLE_NBR_5	EQP_AXLE_NBR_6
EQP_GRP_TANK	trn_id_C	VNDR_EDR_NME_MALT	EQP_AXLE_NBR_7	EQP_AXLE_NBR_8	EQP_AXLE_NBR_9
trn_id_E	KIPDAYS	VNDR_EDR_NME_KYRO	EQP_GRP_BOX_C	EQP_GRP_FLAT	EQP_GRP_GOND
EQP_GRP_FLAT	trn_id_H	trn_id_J	EQP_GRP_HOPP	EQP_GRP_IFLT	EQP_GRP_MFLT
VNDR_EDR_NME_GAGE	EDR_EQP_SPD	EQP_GRS_TONS	EQP_GRP_MISC	EQP_GRP_TANK	EQP_GRP_VFLT
VNDR_EDR_NME_CHRS	EQP_GRP_HOPP	VNDR_EDR_NME_LEED	EQP_GRS_TONS	GRS_RAIL_WGT	KIPDAYS
EQP_AXLE_NBR_9	VNDR_EDR_NME_SEYM	trn_id_X	LOAD_EMPTY	TARE	VNDR_EDR_NME_AMAZ
EDR_VNDR	EQP_GRP_GOND	VNDR_EDR_NME_AMAZ	VNDR_EDR_NME_AURO	VNDR_EDR_NME_BAGD	VNDR_EDR_NME_BMRK
VNDR_EDR_NME_WINT	VNDR_EDR_NME_KIRK	VNDR_EDR_NME_VELO	VNDR_EDR_NME_CHEL	VNDR_EDR_NME_CHRS	VNDR_EDR_NME_COCH
AAR_CT_C_T	AAR_CT_C_S	AAR_CT_C_R	VNDR_EDR_NME_DEAN	VNDR_EDR_NME ESTR	VNDR_EDR_NME_GAGE
AAR_CT_C_Q	AAR_CT_C_P	GRS_RAIL_WGT	VNDR_EDR_NME_JOPP	VNDR_EDR_NME_KIRK	VNDR_EDR_NME_KYRO
VNDR_EDR_NME_JOPP	AAR_CT_C_V	EQP_GRP_VFLT	VNDR_EDR_NME_LEED	VNDR_EDR_NME_MALT	VNDR_EDR_NME_PEST
TARE	VNDR_EDR_NME ESTR	AAR_CT_C_G	VNDR_EDR_NME_PRAI	VNDR_EDR_NME_ROGN	VNDR_EDR_NME_SEYM
AAR_CT_C_F	AAR_CT_C_E	EQP_GRP_MISC	VNDR_EDR_NME_VELO	VNDR_EDR_NME_WINT	WHL_PEAK_KIPS
AAR_CT_C_C	AAR_CT_C_B	AAR_CT_C_A	trn_id_C	trn_id_E	trn_id_H
AR_CT_C_M	AAR_CT_C_L	AAR_CT_C_K	trn_id_J	trn_id_V	trn_id_X
AAR_CT_C_J	AAR_CT_C_H	WHL_PEAK_NEXT			

Tabla 6.2: Lista de características (inputs o features) que entran al modelo

Si bien se parte con 22 datos de entrada para la propuesta y 18 para la propuesta 2, estas se ven aumentadas a 76 y 73 respectivamente. Este gran incremento es principalmente por el proceso de *One Hot Encoding*. Puesto que para cada variable categórica se crean sub-variables dependiendo de sus clases, por lo que es de esperarse este aumento. Este incremento puede afectar a la velocidad de ejecución y compilación

6.3.1.2. Creación de modelos

Ya obtenidas unas bases de datos legible, se procede al siguiente paso de la canalización de datos. En esta se hará un análisis incluyendo los modelos visto en la Subsubsección 4.4.4.3, en adhesión a estos algoritmos se incluirán modelos típicos usados por la mayoría de los científicos de datos. La lista queda conformada con los siguientes tipos de aprendizajes:

- **Regresión multilíneal:** Asume una relación lineal entre las variables de entrada (x_i) y la variable de salida única (y).
- **LDA discriminant:** Asume que la distribución de los datos es de forma gaussiana, luego usa funciones de probabilidad para hacer predicciones.
- **SVM:** De un modelo con N inputs de entrada, crea un hiperplano de $N - 1$ para poder separar y clasificar.
- **TreeClassifier:** Crea un árbol de decisiones, y va seleccionando el output según la probabilidad generada en la base de entrenamiento
- **RandomForest:** Genera una serie de árboles de decisiones aleatorios, y calcula el output a través de una función
- **Neural Network:** Mapea los inputs a través de una función en una cantidad determinada de parámetros, luego calcula el peso de cada parámetro para minimizar la función de pérdida.

Debido a la gran lista de modelos a probar, se iniciará la prueba con sus hiper-parámetros por defecto. Para probar todos los modelos de forma automatizada, se crea una lista de objetos y se llamarán a través del módulo Skicit-Learn. El anexo A.1.0.0.7 muestra la programación de esta acción.

Recordar que cada modelo en cada propuesta recibirá los mismos datos, y cantidad de parámetros vistos en la Tabla 6.2.

6.3.1.3. Evaluación

Se debe considerar que en el caso de la clasificación el desafío contempla la resolución de un modelo óptimo para una base de datos No Balanceada, por lo que si bien la exactitud por sí sola no será un buen parámetro para reflejar el rendimiento del test, refleja la probabilidad que tiene de acierto para evaluar correctamente un evento. Por lo que sólo en caso de tener varios modelos con exactitud alta, se hará un diferenciamiento a través de un análisis de sensibilidad, o sea, la métrica *Recall*, que representa la cantidad de eventos positivos determinados correctamente sobre la cantidad de eventos positivos totales. A continuación la siguiente tabla refleja los resultados obtenidos para ambas propuestas en ambos desafíos.

Modelo	Propuesta1		Propuesta 2	
	Clasificación	Regresión	Clasificación	Regresión
	Exactitud[%]	[%]	Exactitud[%]	[%]
Random Forest	95.7	84 %	96.8	87 %
Support Vector Machine	>99 %	N.A	>99 %	N.A
Árbol de Decisión	92.3	N.A	95.3	N.A
Red Neuronal	90.1	44 %	92	54 %
Regresión Lineal	81.1	22.1	83	26 %
LDA Discrimant	80.9	N.A	80	N.A

Tabla 6.3: Tabla comparativa de resultados obtenido en la base de entrenamiento con validación cruzada K=10

Si bien se aprecia que la exactitud de Support Vector Machine es en extremo superiora los demás algoritmos, al momento de la evaluación se encontró que no fue capaz de predecir ni un resultado positivo. Esto se explica por el sobreajuste del modelo. Esto es una muestra de lo necesario que son otras medidas para corroborar el rendimiento del modelo. Este modelo con el objetivo de obtener una exactitud cercana al 1, predice todos los eventos como negativo (*No alarma*) y en una base donde el 99.66 % de los datos son negativos se asegura una exactitud favorable, pero inservible al fin y al cabo, ya que el costo asociado por no predecir alguna alarma y sufrir un desperfecto en pleno viaje significa una gran pérdida monetaria como se explica al inicio del desafío.

Pasando a analizar los otros resultados, el que tiene una mejor exactitud y sensibilidad es el Random Forest, por el que pasará a la siguiente etapa de afinación. Cabe destacar que si bien la exactitud obtenida por el árbol de clasificación es similar, se privilegiará el hecho de tener una mayor sensibilidad pues el costo de no detectar una alarma puede ser crucial en las operaciones según la descripción del desafío. El hecho de que sea Random Forest el algoritmo con mejor rendimiento es práctico, puesto que la visualización del peso de los parámetros es explicativa y fácil de visualizar, no así en una red neuronal. En estas no se podría explicar el porque un parámetro tiene más peso que otro por lo complejo de su proceso de optimización.

6.3.1.4. Afinación de modelo

Teniendo en cuenta la misma consideración sobre el punto anterior, siendo esta una base de datos no balanceada se decide primeramente abordar este problema para la optimización. Una vez que se logre balancear los datos, la lectura de las bases será más rápida lo que se traduce en eficiencia para el analista de datos. Para realizar esto, se ejecuta una técnica de Resample (Re-muestro). Existen dos tipos de resamples. Para este caso ejecutaremos las dos.

Randon Under-Sampling: Se reducen las muestras de eventos negativos, en este caso, no alarmas. Esto se hace por un medio aleatorio. Ya que el porcentaje de no negativos supera el 99,9 % de las muestras. El numero original de eventos negativos para la propuesta 1 y 2 son 3.587.136 e 3.605.679 respectivamente. Luego de aplicado este proceso, las muestras negativas de la propuesta 1 y 2 quedan reducidas ambas a 150.000.

Tomek Links Under-Sampling: Los enlaces Tomek eliminan la superposición no deseada entre las clases donde se eliminan los enlaces de clase mayoritaria hasta que todos los pares vecinos más cercanos mínimamente distanciados son de la misma clase. Esta por su forma calculada de reducir la base de datos toma más tiempo de ejecución, pero aplicada luego del Random Under Sampling el tiempo empleado es despreciable. Aplicada esta transformación, las muestras de *No Alarmas* para ambas propuestas quedan contabilizadas en 40.000 y 27.000 respectivamente.

SMOTE Over-Sampling: A diferencia de las técnicas anteriores, el over-sampling se usa para aumentar la clase minoritaria pero con la misma finalidad de establecer una proporción más equitativa entre los datos positivos y negativos. En el caso específico del modo SMOTE, se realiza una creación sintética de muestras positivas. Este funciona tomando el vector de inputs para una muestra de *Alarma*, luego toma un conjunto de datos y sus k-vecinos más cercanos, teniendo la distancia entre el k-vecino y el dato actual, multiplica el vector por un número aleatorio entre 0 y 1. Se agrega esta a la base de datos actual, creando una nueva entrada sintética. Para la propuesta 1 y 2, se aumentan las muestras positivas de 16466 y 8292 a 20.000 y 11.000 respectivamente.

Se remarca el hecho de que si bien todos estos cambios son abruptos, pues la mayoría de los datos son eliminados, estos sólo se aplican en la base de datos de entrenamiento (*Training.data*). En cambio la base de testeo, que también sufre de un no balanceo, se dejará intacta para su uso.

La **remoción de parámetros** se puede emplear una vez realizado el resample. Este proceso se va adaptando para ver si este obtiene mejoras, y así descartar parámetros que sólo atrasen el procesamiento del algoritmo. La evaluación de los parámetros de entrada se hace a través del índice de impureza de Gini. Esta medida indica la probabilidad que tuvo cada característica para acertar en el resultado final. El proceso total de esta optimización parte con una iteración, donde inicialmente se alimenta el modelo con todos los parámetros (n parámetros), para este caso n corresponde a 76 y 72 parámetros para la propuesta 1 y 2 respectivamente. Una vez las nuevas bases de datos son leídas, se calculan las distintas métricas para evaluar si el modelo fue aceptable (Exactitud, sensibilidad). Al final de cada iteración se calcula la impureza de Gini para cada característica, finalmente se remueve la que tenga menor pureza dentro de este vector. Esta nueva lista, con $n - 1$ elemento es reemplazada por la anterior y así se van sucediendo una iteración tras otra hasta obtener un vector de tan sólo un elemento. Este proceso se repite varias veces, para evaluar su coherencia y poder evaluar con que lista de características se tuvieron mejores resultados. A continuación, se adjunta la Tabla 6.4 que muestra los parámetros que finalmente resistieron la remoción sin afectar el rendimiento del modelo, o inclusive ayudaron a mejorarlo. En esta también se indica la pureza, que es un reflejo de la importancia que

tiene cada variable dentro del modelo. La suma total de la columna debe ser igual 1.

Propuesta 1		Propuesta 2	
Variable	Pureza	Variable	Pureza
WHL_DYN_KIPS	0.3377	KIPDAYS	0.408
KIPDAYS	0.2177	EQP_GRS_TONS	0.1407
WHL_DYN_RATIO	0.1427	AGE	0.1301
AGE	0.115	LOAD_EMPTY	0.0922
WHL_AVG_KIPS	0.0504	EDR_EQP_SPD	0.0762
EQP_GRS_TONS	0.0458	TARE	0.0723
EDR_EQP_SPD	0.044	VNDR_EDR_NME_DEAN	0.0442
TARE	0.0297	EDR_VNDR	0.0362
GRS_RAIL_WGT	0.017	N.A	N.A

Tabla 6.4: Tabla que indica la significancia de cada parámetro dentro del modelo.

6.3.1.5. Predicción y Resultados

Esta corresponde a la etapa final. Donde se pone a prueba el modelo definitivo para ver los resultados obtenidos y evaluar el rendimiento que este ha tenido. A continuación se prepara una tabla con los resultados para cada propuesta. Se adjuntan también las matrices de confusión correspondientes a cada una de estas.

6.3.1.5.1 Propuesta 1

Matriz de confusión

Matriz de Confusión		Real	
		Alarma	No Alarma
Predicción	Alarma	6.822	163.000
	No Alarma	1.663	1.055.340

Tabla 6.5: Matriz de confusión para propuesta 1

De la matriz anterior, se desprende los siguientes resultados resumidos en la tabla Tabla 6.6

Métrica	Resultado	Interpretación
Regresión	76,4 %	La cantidad de presiones predecida con una diferencia de rango [-2,2].
Exactitud	0,865781	La probabilidad de predecir correctamente un evento.
Sensibilidad	0,804007	La razón entre la cantidad de eventos positivos predichos correctamente y los eventos positivos totales. El modelo mejora mientras más se acerque a 1.
Especificidad	0,866211	La razón entre la cantidad de eventos negativos predichos correctamente y los eventos negativos totales. El modelo mejora mientras más se acerque a 1.
G-Mean	0,83453	Se considera un buen rendimiento de clasificación entre la clase mayoritaria y la minoritaria.
Razón de verosimilitud (L):	6,010	Entre un rango de [5,10] se considera un modelo aceptable.
Discriminante de poder	0,785	Un valor bajo 1, indica un modelo pobre en cuanto a su clasificador para distinguir entre los casos positivos y negativos.
Youden	0,670	El modelo es aceptable, pero no fiable para evadir clasificaciones erróneas.
Kappa	0,064	Existe un bajo acuerdo entre la clase <i>“actual”</i> y la <i>“clasificada”</i>
AuRoc	0.76	Está dentro de lo que se considera un Test bueno. Esta métrica indica que existe un 76 % de probabilidad de que un análisis a un evento con alarma sea más correcto que un evento de no alarma elegido al azar.

Tabla 6.6: Resultados e interpretaciones para la regresión y clasificación de la propuesta 1

6.3.1.5.2 Propuesta 2

Matriz de confusión

Matriz de Confusión		Real	
		Alarma	No Alarma
Predicción	Alarma	3.301	185.043
	No Alarma	572	1.131.487

Tabla 6.7: Matriz de confusión para propuesta 2

De la matriz anterior, se desprende los siguientes resultados resumidos en la tabla Tabla 6.8

Métrica	Resultado	Interpretación
Regresión	82 %	La cantidad de presiones predecida con una diferencia de rango [-2,2].
Exactitud	0,859425	La probabilidad de predecir correctamente un evento.
Sensibilidad	0,852311	La razón entre la cantidad de eventos positivos predichos correctamente y los eventos positivos totales. El modelo mejora mientras más se acerque a 1.
especificidad	0,859446	La razón entre la cantidad de eventos negativos predichos correctamente y los eventos negativos totales. El modelo mejora mientras más se acerque a 1.
G-Mean	0,855871	Se considera un buen rendimiento de clasificación entre la clase mayoritaria y la minoritaria.
Razón de verosimilitud (L):	6,063957	Entre un rango de [5,10] se considera un modelo aceptable.
Discriminante de poder	0,853251	Un valor bajo 1, indica un modelo pobre en cuanto a su clasificador para distinguir entre los casos positivos y negativos.
youden	0,711757	El modelo es aceptable, pero no fiable para evadir clasificaciones erróneas.
Kappa	0,028764	Existe un acuerdo casi nulo entre la clase <i>actual</i> y la <i>clasificada</i>
AuRoc	0.83	Está dentro de lo que se considera un Test Bueno. Esta métrica indica que existe un 83 % de probabilidad de que un análisis a un evento con alarma sea más correcto que un evento de no alarma elegido al azar.

Tabla 6.8: Resultados e interpretaciones para la regresión y clasificación de la propuesta 2

7 | Conclusiones y perspectivas

La medida de evaluación del modelo es un factor clave en el éxito del proceso de minería de datos; en el contexto de datos no balanceados, este tema se vuelve más complejo ya que las métricas de evaluación más frecuentemente adoptadas comparten inconvenientes, que no cumplen con una evaluación fidedigna.

Para el caso entregado por INFORMS, ambas propuestas tuvieron resultados similares en cuanto a su rendimiento para predecir eventos. Por una parte, resultados como Coeficiente de verosimilitud, G-Mean, AuRoc y Sensibilidad entregaron resultados favorables, que indican que el modelo funciona de forma aceptable o buena. Por otro lado indicadores como el coeficiente de Kappa e Índice de Youden indican que el modelo no es del todo fiable para poder predecir eventos. De esto nace un tema importante de discusión: la métrica ideal usada para datos no balanceados. Si bien estos indicadores mencionados anteriormente nos permiten tener una visión global del rendimiento, su interpretación va a depender de la distribución de los datos y lo importante que se consideren para el evaluador. Para este caso se manipuló la información de tal forma que se pudiera privilegiar la minimización de *Falsos negativos*. Esto por que en el desarrollo de la problemática se explica que el gasto por tener una falla y no haberla prevista en un carro con carga asignada es en demasía significativa para la empresa ferroviaria. Si bien en este caso se quiso tener un ratio más grande entre los *Verdaderos positivos* y *Falsos negativos*, se genera un problema de Trade off entre la métrica de Sensibilidad, que mide la integridad de los eventos positivos, y la métrica de especificidad, que mide la integridad de los negativos. Teniendo en cuenta que los datos de alarmas correspondían a menos del 0.45 % del total de los datos resulta difícil tener una buena medida en ambas métricas, y aunque existía la posibilidad de privilegiar los resultados de las métricas para estar en un rango de excelencia de rendimiento (pudiéndose haber obtenido 94 % de Especificidad y 50 % de Sensibilidad), se ha preferido mantenerse dentro de un modelo aceptable, pero con un buen nivel de predicción de eventos positivos, sobre pasando el 85 % de estos incluso. Es decir, este rendimiento se vuelve favorable siempre cuando el costo ahorrado por haber predicho correctamente una alarma, supera con creces el costo por haber acometido una inspección de seguridad por error.

El caso de la evaluación del proceso resulta interesante analizar también el porque ciertos modelos se mostraron tan ineficiente cuando son reconocidamente efectivos en estos casos, como es el ejemplo de Support

Vector Machine o la Red Neuronal. Existe una amplia gama de opciones para modificar la configuración de estos algoritmos a través de los hiper-parámetros. Como la cantidad de neuronas por capas, el número de capas o la constante de margen en el caso de Support Vector Machine. Para este caso una de las razones en que se haya fallado, puede ser a un sobreajuste al no tener una consideración especial sobre la clase minoritaria, por lo que la ejecución de under y over sampling no fueron suficientes para que estos modelos optimizaran la predicción de eventos positivos.

En cuanto al análisis del panorama del Big Data en la ingeniería industrial, se aprecia que el interés por esta área va en un aumento exponencial, donde se refleja por la cantidad de personas que se están dedicando a estudiar esta ciencia, triplicándose por año, como por la cantidad de empresas que están invirtiendo, siendo las más reconocidas mundialmente las mayores inversionistas, superando incluso los 30 billones de dólares en tan sólo un año. La indiferencia existente por esta área actualmente se debe a que se asocia a carreras más técnicas como ingeniería informática o eléctrica, pero lo cierto es que la versatilidad que ofrece machine learning permite ser usado en cualquier campo de las ciencias y empresas. Incluso para un ingeniero industrial un conocimiento técnico de esta área puede significar una característica competitiva que permita sobresalir en el mercado. Y si bien este campo es vasto y tiene tema para ser profundizado, no es necesario ser un experto en estadística o informática para poder implementar un algoritmo funcional en problemas que traten datos reales, tal como se vio en este caso, lo que debería ser un fomento adicional para ser una habilidad adquirida por esta rama de la ingeniería.

Y si bien esta tesis no ha resuelto el tema completo sobre como este tema sustancial se une o justifica la mejora en el área de la ingeniería industrial, es de esperar que haya proporcionado algunas discusiones útiles y un marco para los investigadores a futuro. En cuanto a esto último temas que pueden ser interesantes para seguir investigando puede ser:

- La relación existente entre la adquisición de manejo de Big Data y minería de datos, con la mejora en Supply Chain Management.
- Cómo afectaría una afinación más personalizada en los hiper parámetros a algoritmos como Support Vector Machine o una Red Neuronal para tratar bases de datos con clases altamente desequilibradas.
- Un review sobre el estado del arte de los nuevos tipos de compañía ligada a la ciencia de datos, como Data driven, data analytics y su escalabilidad en Chile.
- El uso de herramientas alternativas a machine learning para la resolución del desafío, tales como Visualización de datos, Cartas de control o herramientas de probabilidad

Bibliografía

- [1] Akosa, Josephine S (2017). Predictive accuracy : A misleading performance measure for highly imbalanced data. *International Knowledge Sharing Platform*, x(1), 1–12. 4.4.4.4
- [2] Andrea McAffe, Erik Brynjolfsson (2012). *Big Data: The Management Revolution*. Estados Unidos: Harvard Business School. 5.1, 5.2
- [3] Bekkar, Mohamed; Kheliouane Djemaa, Dr.Hassiba; y Akrouf Alitouche, Dr.Taklit (2013). Evaluation measures for models assessment over imbalanced data sets. *International Knowledge Sharing Platform*, 3(10), 2224–5782. 4.4.4.4
- [4] Grewal, Dhruv; Motyka, Scott; ; y Levy, Michael (2018). The evolution and future of retailing and retailing education. *Journal of Marketing Education*, (pp. 1–9). 5.3.2
- [5] Hazen, Benjamin T; Skipper, Joseph B; Ezzel, Jeremy D; y Boone, Cristopher A (2016). Bid data and predictive analytics for supply chain sustainability: A theory-driven research agenda. *Computers and Industrial Engineering*, 101(2), 592–598. 5.1, 5.1
- [6] Henrik Brink, Joseph W Richards (2010). *Real World- Machine Learning*. Estados Unidos: Manning Publications Co., 1ra. edición. 4.2, 4.4.4, 4.4.4.3.2
- [7] Institute, McKinsey Global (2011). *Big data: The next frontier for innovation, competition, and productivity*. Estados Unidos: Mckinsey Domain. 4.1
- [8] Jacques Bughin, Eric Hazan, Sree Ramaswamy (2017). *ARTIFICIAL INTELLIGENCE THE NEXT DIGITAL FRONTIER?* Estados Unidos: Mckinsey Domain. 5.3.1, 5.3.2, 5.3.2.2
- [9] James Manyika, Michael Chui (2010). “*Clouds, big data, and smart assets: Ten tech-enabled trends to watch*”. Estados Unidos: Mckinsey Domain. 4.1, 4.2
- [10] Kotsiantis, SB (2007). Supervised machine learning: A review of classification techniques. *S. B. Kotsiantis Department of Computer Science and Technology*, 31(10), 249–268. 4.4.4.3.3
- [11] Lucas, J y Sacucci, M (1990). Exponentially weighted moving average control schemes: Properties and enhancements. *Technometrics*, 32(1), 1–12. 4.4.2
- [12] Matthew A Waller, Stanley E Fawcett (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics Strategic Supply Chain Research*, 34(2), 77–84. 5.2, 5.2.1
- [13] McAlone, Nathan (2016). Why netflix thinks its personalized recommendation engine is worth 1 billion dollar per year. *Business Insider*, 3(7), 1–5. 5.3.1
- [14] Mell, Peter y Grance, Timothy (2011). The nist definition of cloud computing. *National Institute of Standards and Technology*, (pp. 1–11). 4.1
- [15] Mosavi, Amir; Lopez, Alvaro; R, Annamaria; y VarkonyiKoczy (2018). Industrial application of big data: State of the art survey. *Institute of Automation: Obuda University*, 3(2), 9–29. 5.2.1

- [16] of Technology, Massachusetts Institute (2018). Digital economy. <http://ide.mit.edu/>. [Acceso Marzo de 2018]. 5.2
- [17] Peat, Maurice (2013). *Big Data In Finance*. Estados Unidos: InFinance: The Magazine for Finsia Members. 5.3.3
- [18] Poole, David Lynton; Mackworth, Alan K; y Goebel, Randy (1998). *Computational Intelligence: A Logical Approach*. Estados Unidos: Oxford University Press. 4.4.1
- [19] Qiu, Peihua (2017). *Statistical Process Control Charts as a Tool for Analyzing Big Data*, (pp. 123–138). Springer International Publishing: Cham. 4.4.2, 4.3
- [20] Russel, Stuart y Norvig, Peter (2010). *Artificial Intelligence: A modern approach*. Estados Unidos: Pearson Education, 3ra. edición. 4.4.1, 4.4.4.1, 4.4.4.3.1, 4.4.4.3.3
- [21] Sarle, Warren (1994). Neural networks and statistical models. *SAS Institute Inc., Cary*, (pp. 1–11). 4.4.4.3.3
- [22] Saul IGass, Michael C Fu (2013). *Encyclopedia of Operations Research and Management Science*. Estados Unidos: Springer International Publishing AG. 4.4.2
- [23] Seref Sagiroglu, Duygu Sinanc (2013). Big data: A review. *Department of Computer Engineering Faculty of Engineering*, 3(7), 42–46. 4.2, 5.2
- [24] Stross, Randall (2009). *Planet Google: One Company's Audacious Plan to Organize Everything We Know*. Estados Unidos: Simon and Schuster., 1ra. edición. 5.2
- [25] Thomas H Davenport, D Patil (2012). *Data Scientist: The Sexiest job of the 21st Century*. Estados Unidos: Harvard Business School, 1ra. edición. 1.1, 5.2.1
- [26] Triss Ashton, Nicholas Evangelopoulos, Victor R. Prybuto (2014). Quantitative quality control from qualitative data: control charts with latent semantic analysis. *The University of Texas*. 4.4.2
- [27] Usama Fayyad; Piatetsky-Shapiro, Gregory; y Padhraic Smyth (1996). From data mining to knowledge discovery in databases. *AI MAGAZINE*, 5(10), 37–54. 4.4.3
- [28] Vidhya, Analytics (2018). Learn everything about analytics. <http://www.analyticsvidhya.com>. [Acceso Marzo de 2018]. 4.4.4.3.3
- [29] White, Tom (2012). *Hadoop: The Definitive Guide*. Estados Unidos: O'Reilly Publications., 3ra. edición. 4.4.5
- [30] Yu Shui, Guo Song (2016). *Big Data Concepts, Theories, and Applications*. Estados Unidos: Springer International Publishing AG. 5.3.3
- [31] Zikopoulos, Paul y Eaton, Chris (2011). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. Estados Unidos: McGraw-Hill Osborne Media, 1ra. edición. 4.2

A | Anexo

A.1. Código en python

A.1.0.0.1 Orden por fecha

```
datos['vnder_edr_ts_cst'] = datos['vnder_edr_ts_cst'].map(lambda x:
datetime.datetime.strptime(str(x)[0:9], "%d %b %X").strftime("%X-%m-%d"))
datos= datos.sort_values(['unq_whl_id', 'vnder_edr_ts_cst'], ascending=[True, True])
```

A.1.0.0.2 Preprocesamiento de datos

```
aux=[]
auxL_E=[]
for i in range(len(datos['WHL_PEAK_KIPS'])-1):
    if datos['unq_whl_id'][i+1]==datos['unq_whl_id'][i]:
        a= datetime.datetime.strptime((datos['vnder_edr_ts_cst'][i]), "%X-%m-%d")
        b= datetime.datetime.strptime((datos['vnder_edr_ts_cst'][i+1]), "%X-%m-%d")
        delta= b-a
        if datos['LOAD_EMPTY'][i+1]=='L':
            if delta.days <= 10:
                aux.append(datos['WHL_PEAK_KIPS'][i + 1])
            else:
                aux.append(np.nan)
        else:
            aux.append(np.nan)
        else:
            aux.append(np.nan)
```

A.1.0.0.3 Función de mapeo - Transformación binaria

```
datos['WHL_PEAK_NEXT']= datos['WHL_PEAK_NEXT'].map(lambda x: "A" if x >=90 else "N")
datos['LOAD_EMPTY']=datos['LOAD_EMPTY'].map({ 'L':1, 'E':0})
```

A.1.0.0.4 Parcelación de TRN_ID

```
data['TRN_ID']=data['TRN_ID'].map(lambda x: str(x)[1])
```

A.1.0.0.5 Supuesto 5 y 6

```
datos['WHL_PEAK_NEXT']= datos['WHL_PEAK_NEXT'].map(lambda x: "A" if x >=90 else "N")
datos['EDR_EQP_SPD']=datos['EDR_EQP_SPD'].map(lambda x: np.nan if x <= 20 else float(x))
```

A.1.0.0.6 Remoción de OutLiers

```
data= data[(numpy.abs(stats.zscore(data)) < 3).all(axis=1)]
data= data.reset_index(drop=True)
```

A.1.0.0.7 Lista de modelos

```
models = []
models.append(('RF', RandomForestClassifier()))
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
#models.append(('SVM', SVC()))
seed = 7
scoring = 'accuracy'
print "se procede a evaluar cada modelo"
results = []
names = []
for name, model in models:
    kfold = model_selection.KFold(n_splits=10, random_state=seed)
    cv_results = model_selection.cross_val_score(model, X_train, Y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
```

```
msg = "%s: %f %f" % (name, cv_results.mean(), cv_results.std())
print(msg)
```

A.1.0.0.8 Código de regresión

```
import numpy
from sklearn.ensemble import RandomForestRegressor
import pandas as pd
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
import numpy as np
import imblearn.under_sampling as us
from sklearn.ensemble import RandomForestClassifier
from imblearn import over_sampling as ov
from math import sqrt
from sklearn.svm import SVC

f = open('resultados_futuro.txt','w')

f.write('Resultados'+'\n')

data = pd.read_csv("training_ordenado.data", low_memory=False, sep='\t',nrows=10000)
data2= pd.read_csv("testing_ordenado.data", low_memory=False, sep='\t',nrows=50000)
print "lectura de datos"
interseccion= list( set(list(data)) & set(list(data2)) )
lista_remove= ['vnder_edr_ts_cst','EQP_SEQ_NBR','unq_whl_id','RPR_WHY_CD']
for i in lista_remove:
interseccion.remove(i)
interseccion.remove('WHL_PEAK_NEXT')
interseccion.append('WHL_PEAK_NEXT')
data= data[interseccion]
data2= data2[interseccion]

#####Particularidades#####
#####Los valores se reducen a solo el primer caracter, ya que es el identifica
```

```

particularidades = ['trn_id', 'AAR_CT_C', 'EQP_AXLE_NBR']
for i in particularidades:
    data[str(i)] = data[str(i)].map(lambda x: str(x)[0])
    data2[str(i)] = data2[str(i)].map(lambda x: str(x)[0])
#####
#####
for i in list(data):
    clases = data[str(i)].unique()
    clases2 = data2[str(i)].unique()

    if 2 < len(clases) <= 30 :
        data = pd.concat([data, pd.get_dummies(data[str(i)], prefix=str(i))], axis=1)
        data.drop([str(i)], axis=1, inplace=True)
        data2 = pd.concat([data2, pd.get_dummies(data2[str(i)], prefix=str(i))], axis=1)
        data2.drop([str(i)], axis=1, inplace=True)
        interseccion2 = list(set(list(data)) & set(list(data2)))
        interseccion2.remove('WHL_PEAK_NEXT')
        interseccion2.append('WHL_PEAK_NEXT')
        data = data[interseccion2]
        data2 = data2[interseccion2]

    elif len(clases) == 2:
        data[str(i)] = data[str(i)].map({clases[0]: 0, clases[1]: 1})
        data2[str(i)] = data2[str(i)].map({clases[0]: 0, clases[1]: 1})

    else:
        data[str(i)] = data[str(i)].map(lambda x: np.nan if x < 0 else x)
        data2[str(i)] = data2[str(i)].map(lambda x: np.nan if x < 0 else x)

data = data.dropna().reset_index(drop=True)
data2 = data2.dropna().reset_index(drop=True)
#####

columnas = interseccion2

```

```
print len(columnas)
print columnas
```

```
#Remocion de Outliers
```

```
data= data[(numpy.abs(stats.zscore(data)) < 3).all(axis=1)]
data2= data2[(numpy.abs(stats.zscore(data2)) < 3).all(axis=1)]
data= data.reset_index(drop=True)
data2= data2.reset_index(drop=True)
```

```
Feature_a_removal=[]
f.write("Feature_a_removal: \n")
while len(columnas) > 2:
print("Cantidad de Features: " + str(len(columnas)))
f.write("Cantidad de Features: " + str(len(columnas)) + '\n')
```

```
if len(Feature_a_removal) != 0:
columnas.remove(Feature_a_removal)
f.write("Feature_a_removal: " + str(Feature_a_removal) + '\n')
data = data[columnas]
data2 = data2[columnas]
```

```
valores = data.values
valores2= data2.values
X_train = valores[:,0:(len(columnas)-1)] # valores[:,[1,6,10,13,12,14,16,20,21]]
Y_train = valores[:,(len(columnas)-1)]
X_validation= valores2[:,0:(len(columnas)-1)] #valores2[:, [1,6,10,13,12,14,16,20,21]]
Y_validation= valores2[:,(len(columnas)-1)]
```

```
contador = 0
```

```
RFR= RandomForestRegressor(verbose=True,n_jobs=-1)
RFR.fit(X_validation,Y_validation)
```

```

predicciones= RFR.predict(X_validation)
for i in range(len(predicciones)):
    if abs(predicciones[i] - Y_validation[i]) <= 3:
        contador = contador +1
    print("Total_de_Y: "+ str(len(Y_validation)) + '\n')
    f.write("Total_de_Y: "+ str(len(Y_validation)) + '\n')
    print("Contador: "+ str(contador) + '\n')
    f.write("Contador: "+ str(contador) + '\n')
    print("Porcentaje: " + str(float(contador)/(len(Y_validation))) + '\n')
    f.write("Porcentaje: " + str(float(contador)/(len(Y_validation))) + '\n')
Indice_impuridad= sorted(zip(map(lambda x: round(x, 4), RFR.feature_importances_),columns))
print("Indice_de_impuridad: " + '\n' + str(Indice_impuridad) + '\n' + '\n' + '\n' + '\n')
f.write("Indice_de_impuridad: " + '\n' + str(Indice_impuridad) + '\n' + '\n' + '\n' + '\n')
Feature_a_remover = str(Indice_impuridad[len(Indice_impuridad) - 1][1])

```