

BIG DATA & DATA SCIENCE

VÍCTOR LEIVA

www.victorleiva.cl

ESCUELA DE INGENIERÍA INDUSTRIAL

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

AGENDA

BUSINESS INTELLIGENCE (BI)

BIG DATA

DATA MINING

DATA SCIENCE

EXTRACT, TRANSFORM, LOAD (ETL)

PRESENTACIÓN DE CASOS

SOFTWARE R

ANÁLISIS DE CASOS EN R

A semi-transparent image of Larry Page, co-founder of Google, is on the left side of the slide. He has grey hair and is wearing a grey sweater over a black shirt. He is holding a silver Shure microphone and speaking into it. The background behind him is a warm orange and red gradient.

BIG DATA

«ELEGIMOS ESTO PORQUE
TRABAJAMOS CON GRANDES
CANTIDADES DE DATOS Y
ADEMÁS SUENA COOL.»

— LARRY PAGE, GOOGLE

BIG DATA EN EL MUNDO

BIG DATA UNIVERSITY



BIG DATA EN CHILE



DataChile
beta

El esfuerzo más completo para integrar y visualizar los datos
públicos de Chile

Busque lugares, industrias, productos...

EXPLORAR PERFILES >

¿QUÉ ES BIG DATA?

BIG DATA: VISIÓN CLÁSICA

SEGÚN GARTNER:

VOLUMEN

VELOCIDAD

VARIEDAD

VALOR

OBJETIVO: PROPORCIONAR ALTA TECNOLOGÍA DE **HARDWARE Y SOFTWARE** PARA ALMACENAR, PROCESAR Y ANALIZAR GRANDES (BIG) CANTIDADES DE DATOS (GB, TB, PB, EB, ZB O YB) PARA GENERAR **VALOR** EN UNA ORGANIZACIÓN.

BIG DATA: OTRAS VISIONES

Organización	Descripción
Oracle	BBDD relacionales aumentadas con nuevas fuentes de datos no estructurados: enfatiza tecnologías como NoSQL , Hadoop , R y Python .
Intel	Datos generados a un ritmo de 500TB por semana. Se centra en velocidad = (volumen / tiempo).
Microsoft	El proceso de aplicar seriamente poder computacional , aprendizaje de máquinas e inteligencia artificial en conjuntos de datos masivos y, con frecuencia, altamente complejos: explicita el poder computacional y se relaciona a machine learning e inteligencia artificial .
MIKE2.0	<i>Method for an Integrated Knowledge Environment:</i> Múltiples fuentes de datos independientes con potencial de interactuar (<i>Big Data</i> puede ser 'pequeño').

TIPOS DE BIG DATA

ESTRUCTURADOS

NO ESTRUCTURADOS

SEMI ESTRUCTURADOS

TIPOS DE BIG DATA

(EL ORIGEN DE LOS DATOS)

BIG DATA TRANSACCIONAL

M2M

WEB & SOCIAL MEDIA

DATOS BIOMÉTRICOS

GENERACIÓN ‘HUMANA’



A photograph showing a group of miners in a mine shaft. They are wearing various colored shirts and hard hats. In the foreground, a large yellow metal support beam is visible. In the background, a white flag with the Chilean coat of arms and the word "CHILE" is hanging from a structure.

**33 MINEROS
AGOSTO DE 2010
COPIAPÓ, CHILE**

ANTONIO BANDERAS RODRIGO SANTORO JULIETTE BRIE JAMES LOU DIAMOND MARC ADRIANA KATE COTE BOB GABRIEL
BRODIE BROLIN PHILLIPS CASAS BARRAZA DEL CASTILLO DE PABLO WITH GUNTON AND BYRNE

69 DAYS. 2300 FT UNDERGROUND. 1 BILLION PEOPLE WATCHING.

DISCOVER THE MIRACULOUS TRUE STORY



DOS PELÍCULAS

MACHAMA'S

33
33
33
33

THE STORY OF HOW 33 MINERS BRAVELY SURVIVED, AS ONE NATION ROSE TO THE CHALLENGE.

AMERICA FILM FESTIVAL PRESENTS MACHAMA'S 33 AND PANTERA'S 33 FROM DYNAMIC FACTORY PRODUCTION, STARRING ALBERTO A., ALBERTO DE LA TORRE, LUIS MIGUEL, PRODUCTION DESIGN ALICIA ALERO,
DIRECTOR OF PHOTOGRAPHY ALBERTO FUENTES, MUSIC ALBERTO DE LA TORRE, EDITORS ANDRES SANCHEZ, FELIPE BONILLA, EXECUTIVE PRODUCER JACINTO DE LA TORRE, EXECUTIVE PRODUCER ANTONIO BANDERAS

NOVEMBER 13

AMERICA
TELEVISION

RECOLECTANDO BIG DATA



Menú

NEWS | MUNDO

Noticias | América Latina | Internacional | Economía | Tecnología | Ciencia | Salud | Cultura

La chaqueta inteligente que mide en terreno la salud de los mineros

Constanza Hola Chamy
BBC Mundo

⌚ 15 octubre 2013

f 📢 🌐 📧 Compartir

A simple vista, una chaqueta reflectante. Pero esconde un salvavidas, literalmente.

En realidad, no es una chaqueta, sino una Estación de Monitoreo Móvil (EMM), equipada con un conjunto de sensores portátiles capaces de medir datos medioambientales y médicos.



Estos son transmitidos en tiempo real tanto al teléfono inteligente del trabajador como a la nube, para ser revisados y analizados.



A wide-angle photograph of a busy stock exchange trading floor, likely the New York Stock Exchange (NYSE). Numerous traders in dark suits are positioned behind long, curved trading desks, each equipped with multiple computer monitors displaying financial data and charts. The room is filled with the glow of screen lights and the hum of activity. In the background, a large digital sign prominently displays the letters "NYSE".

**DATA
ARE
EVERWHERE**

Y SI QUIERO HACER BIG DATA A QUIÉN RECURRO

BIG DATA & AI LANDSCAPE 2018



A man with grey hair, wearing a dark suit, white shirt, and brown tie, is gesturing with his right hand while speaking. He is positioned on the left side of the frame against a solid blue background.

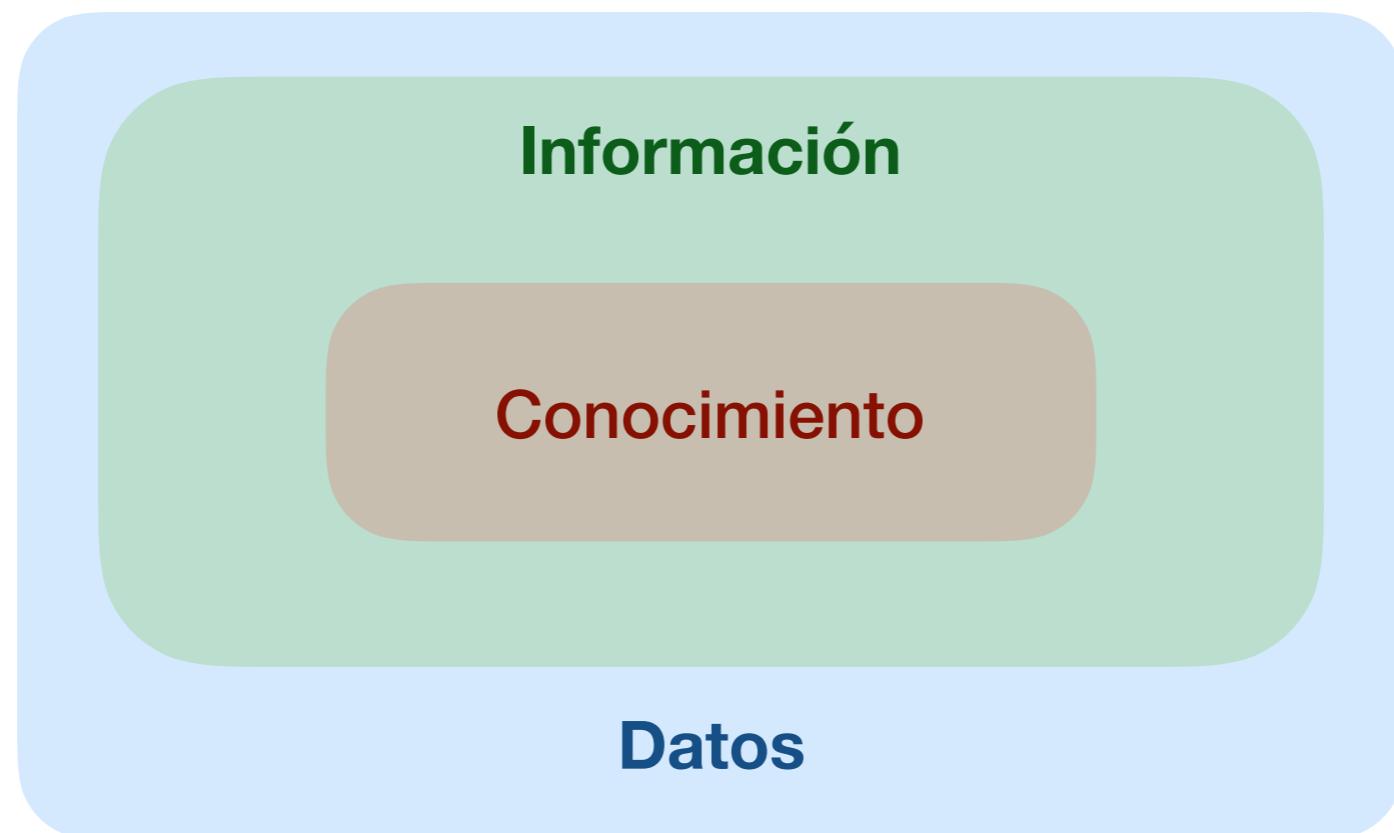
«LOS DATOS SON EL PETRÓLEO DEL
SIGLO XXI, Y ANALYTICS ES SU
MOTOR DE COMBUSTIÓN.»

(PETER SONDERGAARD)

DATA
ANALYTICS

¿QUÉ ES BUSINESS INTELLIGENCE?

BUSINESS INTELLIGENCE ES UN CONJUNTO DE HERRAMIENTAS Y MÉTODOS INFORMÁTICOS, ESTADÍSTICOS, MATEMÁTICOS, DE OPTIMIZACIÓN, COMPUTACIONALES Y DE NEGOCIOS QUE PERMITEN **TRANSFORMAR DATOS EN INFORMACIÓN**, Y LUEGO **TRANSFORMAR LA INFORMACIÓN EN CONOCIMIENTO** PARA OPTIMIZAR LA TOMA DE DECISIONES EN CONTEXTOS DE NEGOCIO.



¿QUÉ ES DATA SCIENCE?

DATA SCIENCE ES UN CAMPO INTERDISCIPLINARIO SOBRE PROCESOS PARA EXTRAER CONOCIMIENTO DESDE LOS DATOS QUE SE MANIFIESTAN DE DISTINTAS FORMAS. **DATA SCIENCE** ESTÁ RELACIONADO A LAS ÁREAS DE ANÁLISIS DE DATOS: **ESTADÍSTICA**, **DATA MINING** Y **ANÁLISIS PREDICTIVO**, SIMILARES AL DESCUBRIMIENTO DE CONOCIMIENTO EN DATOS (KNOWLEDGE DISCOVERY IN DATA, KDD).

EL **DATA SCIENTIST** HA SIDO DECLARADO COMO EL PERFIL PROFESIONAL MÁS ATRACTIVO DEL SIGLO XXI POR THE NEW YORK TIME Y LINKEDIN. LOS **DATA SCIENTISTS** SON ALTAMENTE DEMANDADOS POR GRANDES COMPAÑÍAS DEL MUNDO Y CORRESPONDEN A UNA MEZCLA DE DIVERSAS HABILIDADES.

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ★ Machine learning
- ★ Statistical modeling
- ★ Experiment design
- ★ Bayesian inference
- ★ Supervised learning: decision trees, random forests, logistic regression
- ★ Unsupervised learning: clustering, dimensionality reduction
- ★ Optimization: gradient descent and variants

DOMAIN KNOWLEDGE & SOFT SKILLS

- ★ Passionate about the business
- ★ Curious about data
- ★ Influence without authority
- ★ Hacker mindset
- ★ Problem solver
- ★ Strategic, proactive, creative, innovative and collaborative

PROGRAMMING & DATABASE

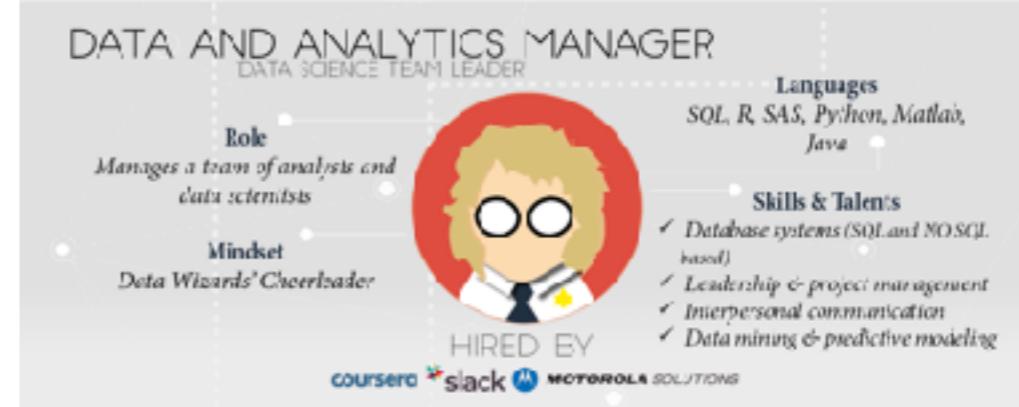
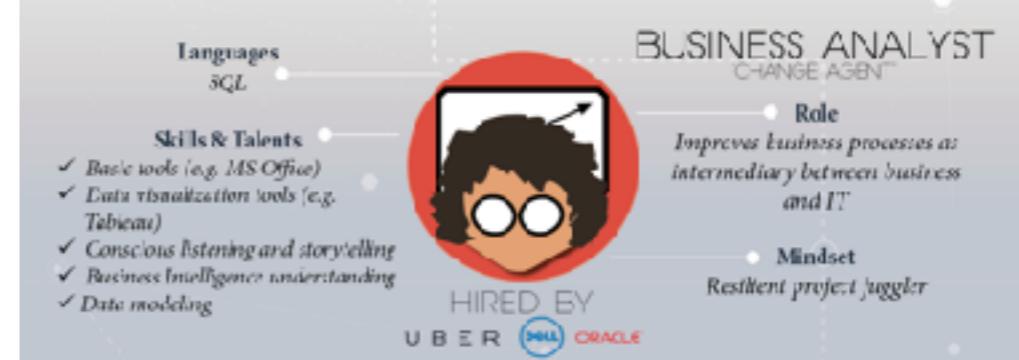
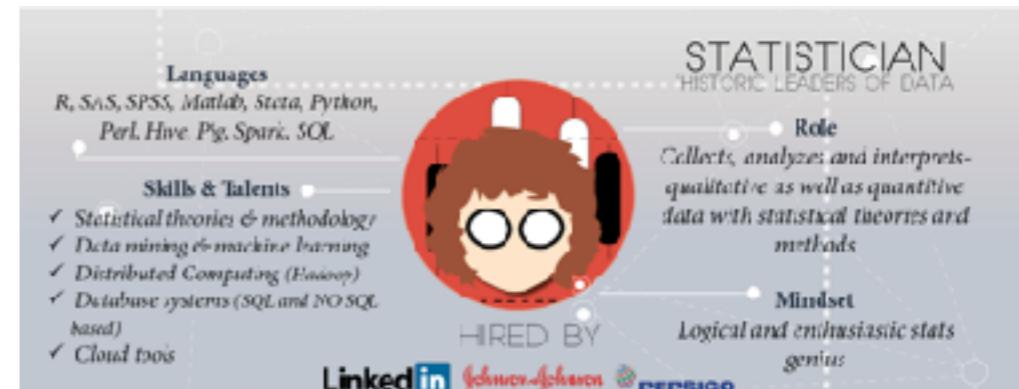
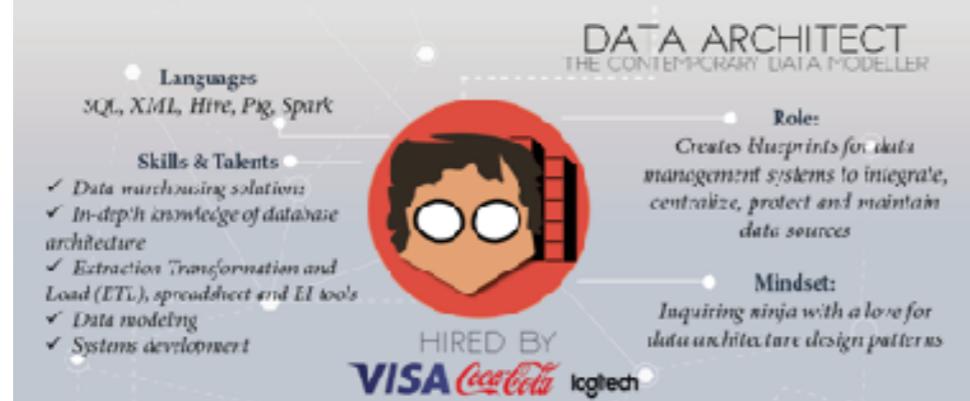
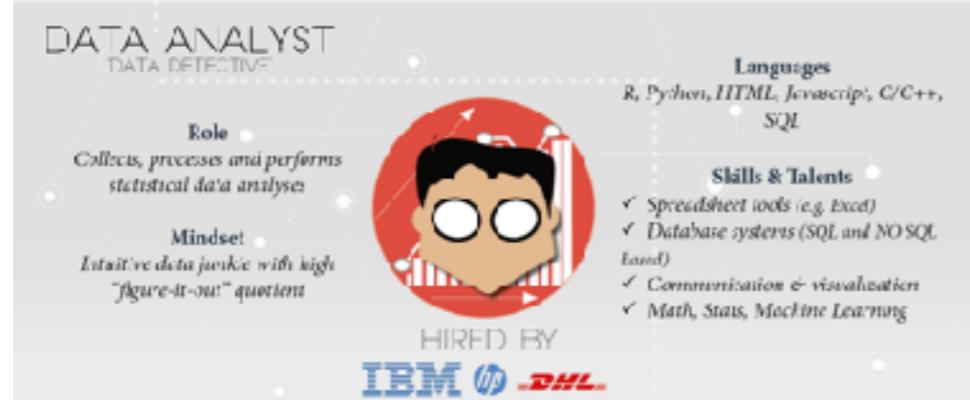
- ★ Computer science fundamentals
- ★ Scripting language e.g. Python
- ★ Statistical computing package e.g. R
- ★ Databases SQL and NoSQL
- ★ Relational algebra
- ★ Parallel databases and parallel query processing
- ★ MapReduce concepts
- ★ Hadoop and Hive/Pig
- ★ Custom reducers
- ★ Experience withaaS like AWS

COMMUNICATION & VISUALIZATION

- ★ Able to engage with senior management
- ★ Story telling skills
- ★ Translate data-driven insights into decisions and actions
- ★ Visual art design
- ★ R packages like ggplot or lattice
- ★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau



EXISTEN OTROS PERFILES



¿QUÉ ES DATA MINING?

DATA MINING ES UN PROCESO AUTOMÁTICO Y NO TRIVIAL DE RECONOCIMIENTO DE PATRONES QUE ESTÁN OCULTOS EN LOS DATOS. ESTE PROCESO DEBE SER COMPRENSIBLE, POTENCIALMENTE ÚTIL, VÁLIDO E INNOVADOR.

DATA MINING UTILIZA MÉTODOS DE **INTELIGENCIA ARTIFICIAL**, DE **MACHINE LEARNING**, DE **ESTADÍSTICA** Y DE **BASES DE DATOS**.

INTELIGENCIA
ARTIFICIAL

MACHINE LEARNING

DEEP LEARNING

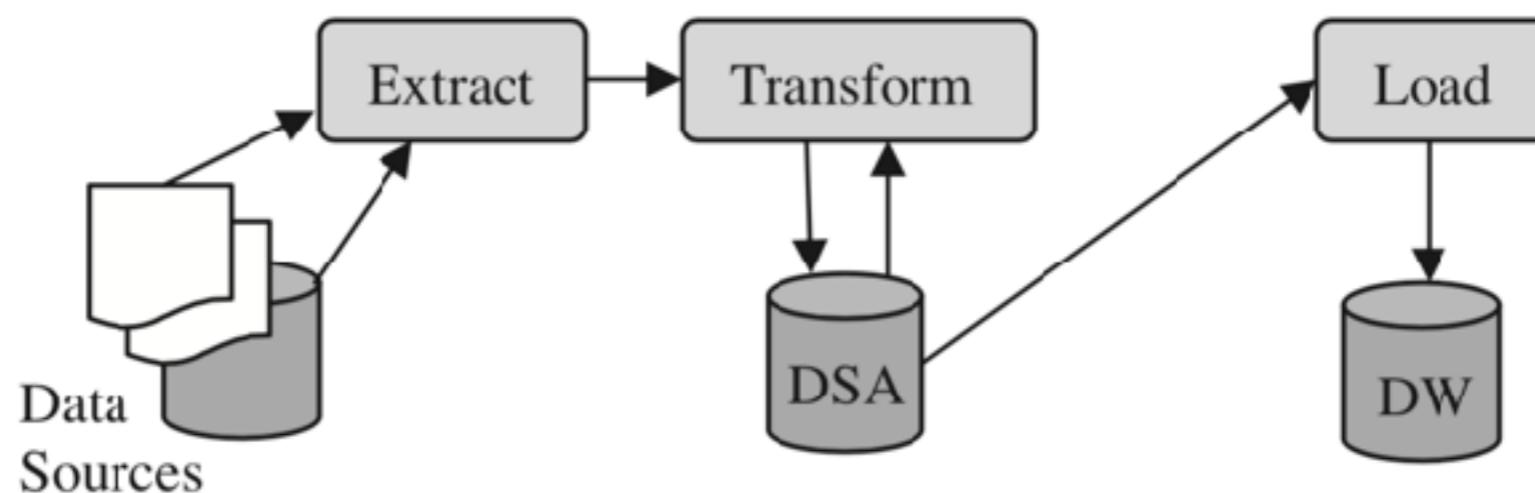
CIENCIAS APLICADAS

EL **DEEP LEARNING** AYUDA A DAR SENTIDO A GRANDES CANTIDADES DE DATOS.

MEDIANTE EL USO DE **REDES NEURONALES** LOS COMPUTADORES VEN, APRENDEN Y REACCIONAN ANTE SITUACIONES COMPLEJAS.

¿QUÉ ES ETL?

EXTRAER, TRANSFORMAR Y CARGAR (ETL) ES EL CONJUNTO DE HERRAMIENTAS Y TÉCNICAS QUE COMPONEN EL PROCESO MEDIANTE EL CUAL LAS ORGANIZACIONES MUEVEN DATOS DESDE MÚLTIPLES ORÍGENES. ESTO INCLUYE: REFORMATEAR, LIMPIAR Y CARGAR EN OTRO PUNTO DE ALMACENAMIENTO, TÍPICAMENTE UNA BASE DE DATOS, EN UN **DATA WAREHOUSE**. ETL ES EL PRIMER PASO ANTES DE REALIZAR ANALYTICS, PARA POSTERIORMENTE TOMAR DECISIONES QUE APOYEN UN PROCESO DE NEGOCIOS.



ENCAPSULACIÓN DE DATOS

SERVIDOR

BASE DE DATOS

TABLAS

CAMPOS

4 ETAPAS DEL ANÁLISIS

FILTRADO DE DATOS

SELECCIÓN DE VARIABLES

GENERACIÓN DE CONOCIMIENTO

INTERPRETACIÓN Y EVALUACIÓN

ALGUNOS MÉTODOS DE DATA ANALYTICS

CLASIFICACIÓN Y AGRUPACIÓN

MODELADO

REDUCCIÓN DE DIMENSIÓN

DATA ANALYTICS

SUPERVISADO

CLASIFICACIÓN

REGRESIÓN

NO SUPERVISADO

CLUSTERING

ANÁLISIS CORRELACIONAL

REGLAS DE ASOCIACIÓN

ANÁLISIS DE SIMILITUD

PREDICCIÓN DE ENLACES

REDUCCIÓN DE DATOS

**TODOS LOS MODELOS ESTÁN
EQUIVOCADOS, PERO
ALGUNOS SON ÚTILES.**

—GEORGE E. P. BOX (1919-2013)



BUSINESS INTELLIGENCE

ALGUNAS PREGUNTAS QUE PODEMOS RESPONDER

INDUSTRIA FINANCIERA

- ¿CUÁLES CLIENTES CERRARÁN SUS PRODUCTOS EN LOS PRÓXIMOS MESES?
- ¿DE QUÉ SEGMENTO DE CLIENTES SE COMPONE MI CARTERA?
- ¿DEBE EL BANCO DAR UN CRÉDITO A CIERTO CLIENTE?



BUSINESS INTELLIGENCE

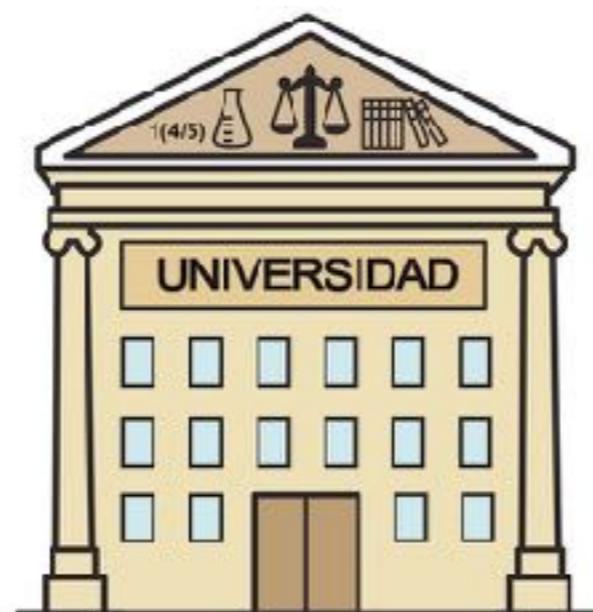
ALGUNAS PREGUNTAS QUE PODEMOS RESPONDER

INSTITUCIÓN EDUCACIONAL

¿POR QUÉ DESERTAN MIS ESTUDIANTES?

¿QUÉ PERFIL TIENEN MIS ESTUDIANTES?

¿ES MEJOR UNA UNIVERSIDAD QUE OTRO CENTRO ACADÉMICO?



BUSINESS INTELLIGENCE

ALGUNAS PREGUNTAS QUE PODEMOS RESPONDER

EMPRESA MINERA

¿CUÁL ES LA LEY DEL MINERAL?

¿PODEMOS REDUCIR COSTOS PLANIFICANDO MEJOR EL CONSUMO DE ENERGÍA?

¿CUÁNTOS AÑOS PUEDO EXPLOTAR ESTA MINA?



© Can Stock Photo - csp6162543

BUSINESS INTELLIGENCE

ALGUNAS PREGUNTAS QUE PODEMOS RESPONDER

CENTRO DE SALUD

¿QUÉ TIPO DE PACIENTES SE ATIENDE CON NOSOTROS?

¿CUÁL SERÁ LA DEMANDA DE LOS PACIENTES EL PRÓXIMO AÑO?

¿MEJORA LA SALUD AL TENER UN SEGURO MÉDICO?



BUSINESS INTELLIGENCE

ALGUNAS PREGUNTAS QUE PODEMOS RESPONDER

CANAL DE TV

¿QUÉ TELEVIDENTES VERÁN UN PROGRAMA A CIERTA HORA?

¿CUÁLES PRODUCTOS SE DEBEN PUBLICITAR A PARTIR DE LAS 22:00?

¿DEBEMOS CREER EN LOS RATINGS?



BUSINESS INTELLIGENCE

ALGUNAS PREGUNTAS QUE PODEMOS RESPONDER

INSTITUCIÓN ELECTORAL

¿QUÉ CANDIDATOS ESTÁN SIENDO ELECTOS HASTA AHORA?

¿CUÁLES PARTIDOS POLÍTICOS ESTÁN MEJOR EVALUADOS ANTE LA CORRUPCIÓN?

¿LAS NUEVAS LEYES MEJORAN LA CALIDAD DE VIDA DE LOS CIUDADANOS?



BUSINESS INTELLIGENCE

ALGUNAS PREGUNTAS QUE PODEMOS RESPONDER

COMPAÑÍA DE SEGUROS

¿CUÁL ES EL RIESGO DE MI CARTERA?

¿QUÉ CLIENTES SON MÁS RENTABLES?

¿ES RENTABLE TENER UN SEGURO?



BUSINESS INTELLIGENCE

ALGUNAS PREGUNTAS QUE PODEMOS RESPONDER

TIENDAS POR DEPARTAMENTO

¿QUÉ CLIENTES COMPRAN POR INTERNET?

¿CUÁLES CLIENTES ABRIRÁN TARJETA DE CRÉDITO CON LA TIENDA?

¿QUÉ PRODUCTOS LE OFREZCO A MIS CLIENTES?



BUSINESS INTELLIGENCE

ALGUNAS PREGUNTAS QUE PODEMOS RESPONDER

SERVICIO DE HEADHUNTER

¿CUÁL CANDIDATO DEBE SER CONTRATADO?

¿QUÉ HABILIDADES DEBE TENER EL CANDIDATO PARA ESA POSICIÓN?

¿QUÉ SALARIO SE DEBE ASIGNAR AL CANDIDATO SELECCIONADO?

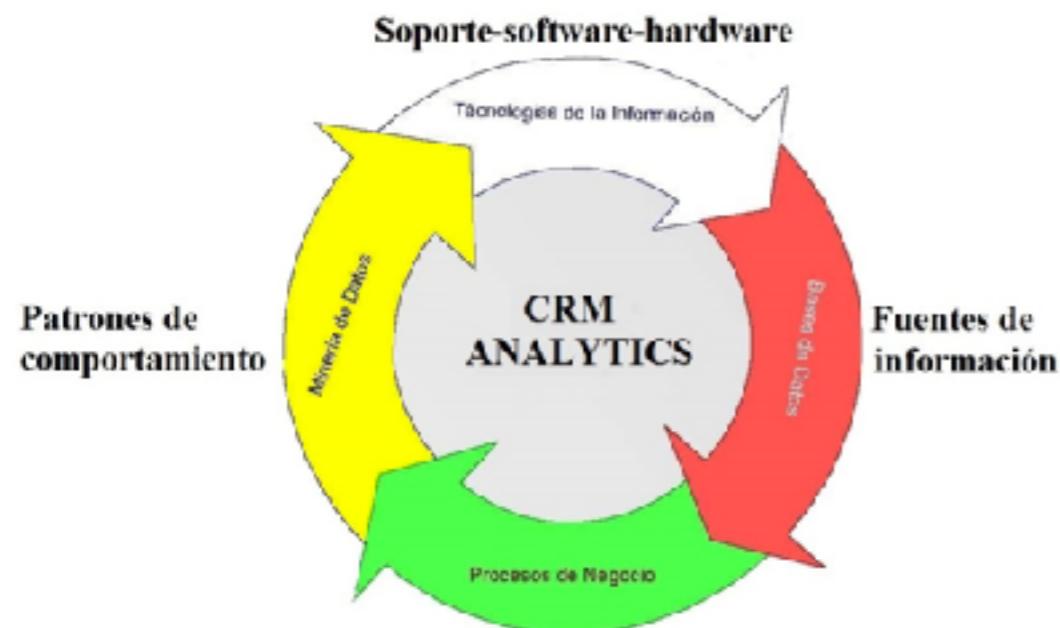


**¿CUÁL ES LA DIFERENCIA
ENTRE DATA SCIENCE Y
BUSINESS INTELLIGENCE?**

BUSINESS INTELLIGENCE



COSTUMER RELATIONSHIP MANAGEMENT



EL **CRM (CUSTOMER RELATIONSHIP MANAGEMENT) ANALYTICS** COMBINA PROCESOS DE NEGOCIO Y TECNOLOGÍAS PARA ANALIZAR DE DATOS DE CLIENTES, DESCUBRIENDO PATRONES DE CONDUCTA O NUEVAS TENDENCIAS DE MERCADO. EL **CRM ANALYTICS** IDENTIFICA PROBLEMAS EN LA CALIDAD DEL SERVICIO AL CLIENTE, AUMENTANDO LOS NIVELES DE SATISFACCIÓN Y RETENCIÓN.

BUSINESS INTELLIGENCE EN ANÁLISIS DE CASOS

- DECIDIENDO SI DAR O NO CRÉDITO
- RESUMIENDO UN DISCURSO PRESIDENCIAL
 - DETERMINANDO UN PERFIL
- SELECCIONANDO UN POSTULANTE
 - ESCOGIENDO UNA MARCA
 - PREDICIENDO LA DEMANDA
- ESTABLECIENDO UNA LÍNEA DE POBREZA

ASIGNACIÓN DE CRÉDITO

LOS BANCOS MANTIENEN DATOS HISTÓRICOS DE SUS CLIENTES A LOS QUE LES HAN PRESTADO DINERO. ESOS DATOS INCLUYEN SI HAN O NO PAGADO EL CRÉDITO. UN BANCO COMERCIAL CONSIDERA UNA MUESTRA DE $N = 2.464$ CLIENTES, SIENDO LAS SIGUIENTES LAS VARIABLES MEDIDAS:

Variable	Característica	Valores
Valoración:	Valoración del cliente	0=Malo y 1=Bueno
Edad:	Edad medida en años	$[0, \infty[$
Ingresos:	Ingresos categorizados	1=Bajo, 2=Medio y 3=Alto
Tarjetas:	Número de tarjetas de crédito	1=Ménos de cinco y 2 = Cinco o más
Educación:	Nivel de educación	1=Enseñanza media y 2=Universitaria
Crédito:	Número de créditos solicitados	1=Uno o ninguno y 2=Dos o más

OBJETIVO: ENCONTRAR UN MODELO QUE PERMITA CLASIFICAR A FUTUROS CLIENTES COMO BUENOS O MALOS PAGADORES.

UN CLIENTE TIENE 36 AÑOS DE EDAD, ESTUDIOS UNIVERSITARIOS, MÁS DE 5 TARJETAS DE CRÉDITO, INGRESOS BAJOS Y ANTERIORMENTE SOLICITÓ UN (1) CRÉDITO.

¿EL BANCO DEBE ASIGNARLE EL CRÉDITO?

¿EL CLIENTE PAGARÁ EL CRÉDITO AL BANCO?

DISCURSO DE UN PRESIDENTE

Señoras y señores parlamentarios,
chilenas y chilenos,

Nuestra Misión

Hace un año concurrí a este Congreso Pleno para convocar a las chilenas y chilenos a asumir una misión y abrazar una causa noble y factible: hacer de nuestra patria, antes que termine esta década, un país desarrollado y sin pobreza. Este fue el sueño que nuestros padres y abuelos siempre acariciaron, pero nunca lograron. Y esta es la misión y deber de nuestra generación, la generación del Bicentenario.

Los convoqué a que juntos construyéramos una sociedad de seguridades, oportunidades y valores para todos.

Una sociedad de seguridades significa una vida digna para todos los chilenos y chilenas y atenuar los temores que nos oprimen. Para ello necesitamos una red de protección fuerte y siempre disponible para acogerlos. Pero no una telaraña que los atrape en una



PROBLEMA: LA OPOSICIÓN DESEA ANALIZAR EL DISCURSO EN UN BREVE PLAZO PARA EMITIR UN COMUNICADO.
OBJETIVO: IDENTIFICAR LOS ELEMENTOS MÁS DESTACADOS DEL DISCURSO.

MINERÍA DE TEXTO

SE PUEDE UTILIZAR PARA ANALIZAR REDES SOCIALES.
PUEDE SER EL FUTURO DE LAS ENCUESTAS.



PERFIL DE TRABAJADORES

PROBLEMA: UNA EMPRESA CONSIDERA UNA MUESTRA DE 474 DE SUS TRABAJADORES.
LAS VARIABLES MEDIDAS A CADA TRABAJADOR SON:

ID: CÓDIGO DEL TRABAJADOR

SEXO: GÉNERO (MASCULINO, FEMENINO)

FECHNAC: FECHA DE NACIMIENTO

EDUC: AÑOS DE ESCOLARIDAD

CATLAB: CATEGORÍA LABORAL (ADMINISTRATIVO, SEGURIDAD, DIRECTIVO)

SALARIO: SALARIO ACTUAL

SALINI: SALARIO INICIAL

TIEMPEMP: MESES DESDE EL CONTRATO

EXPPREV: EXPERIENCIA PREVIA (MESES)

MINORIA: CLASIFICACIÓN DE MINORÍAS (PERTENECE, NO PERTENECE)

OBJETIVO: OBTENER UN PERFIL DE LOS TRABAJADORES DE LA EMPRESA.

PROBLEMA: UNA EMPRESA CONTRATA A 25 CANDIDATOS A UN PUESTO DE TRABAJO. CADA UNO RESPONDE 4 TESTS DE APTITUDES. TODOS FUERON CONTRATADOS POR UN PERÍODO DE PRUEBA. UNA VEZ QUE ESTE PERÍODO CONCLUYÓ, A CADA UNO SE LE MIDE LA COMPETENCIA LABORAL (JOB-PROF).

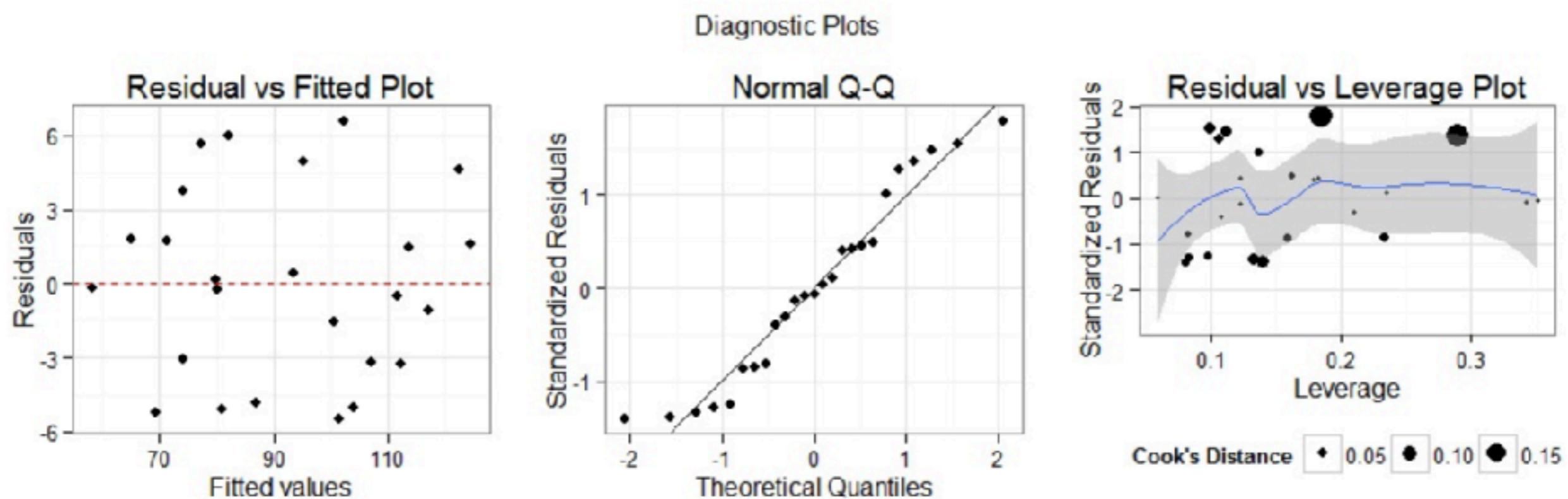
OBJETIVO: IDENTIFICAR LAS VARIABLES QUE MEJOR PREDICEN LA COMPETENCIA LABORAL.



SOLUCIÓN: SE UTILIZÓ UN MODELO DE REGRESIÓN LINEAL MÚLTIPLE Y EL MODELO FINAL PREDICTIVO, QUE NO CONSIDERA EL TEST2, ES:

$$\widehat{\text{Job_prof}} = -124.200 + 1,357 \cdot \text{Test3} + 0,296 \cdot \text{Test1} + 0,517 \cdot \text{Test4}$$

ADEMÁS, SE REALIZÓ EL SIGUIENTE ANÁLISIS DE DIAGNÓSTICO:



ESCOGIENDO UNA MARCA

PROBLEMA: SE CONSIDERA UNA MUESTRA DE 22 VEHÍCULOS ELEGIDOS AL AZAR DE ENTRE LOS OFRECIDOS POR LOS RESPECTIVOS FABRICANTES. LAS VARIABLES MEDIDAS EN CADA VEHÍCULO SON:

PRICE: PRECIO DEL VEHÍCULO.

ACCELER: ACCELERACIÓN DEL VEHÍCULO (DE 0 A 60MPH, EN SEGUNDOS).

BRAKING: RENDIMIENTO DE FRENADO (DISTANCIA DE FRENADO DE 80MPH PARA COMPLETAR LA PARADA).

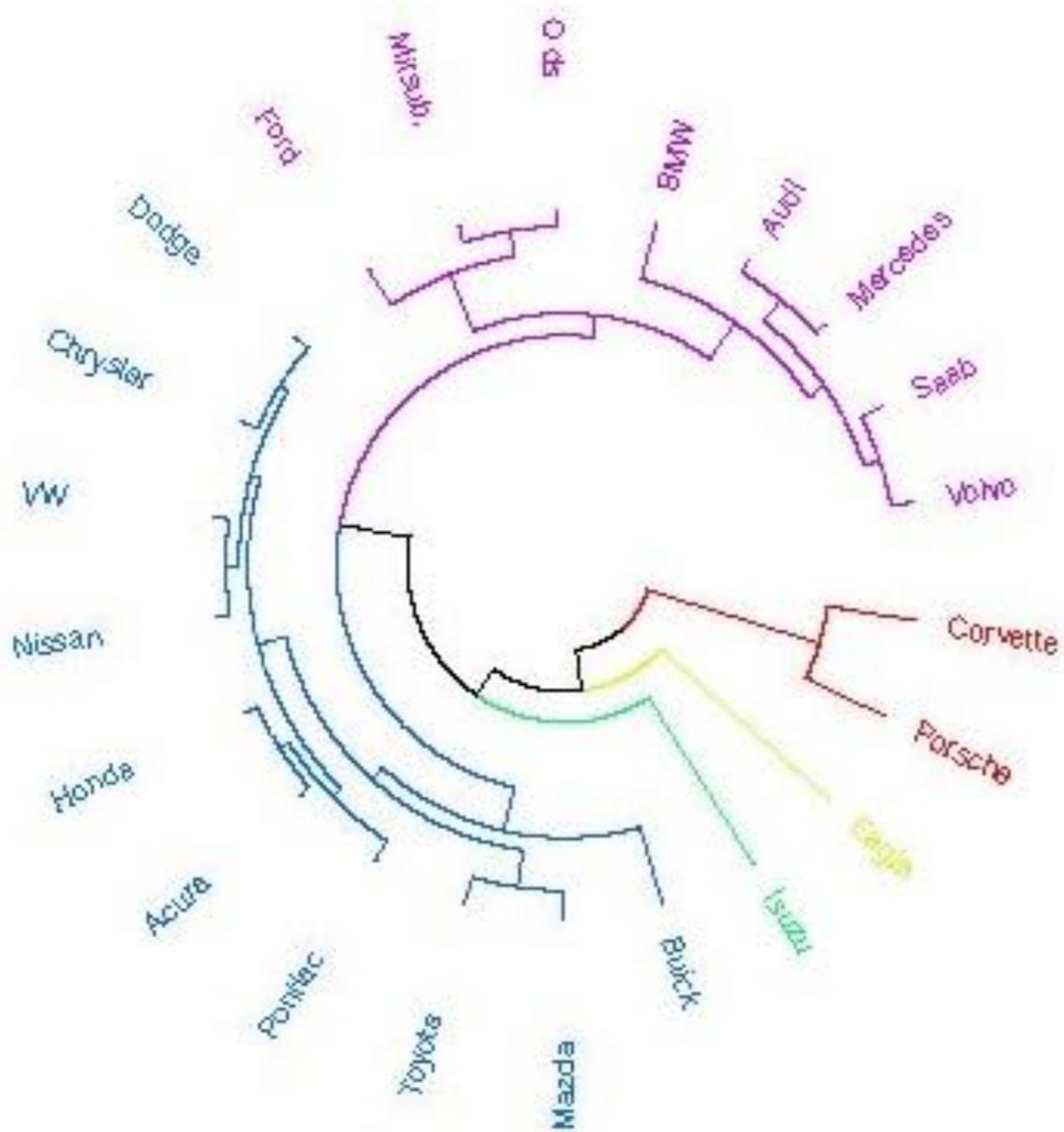
HANDLING: ÍNDICE DE LA CAPACIDAD DE AGARRE EN CARRETERA .

MILEAGE: MILLAS POR GALÓN.

OBJETIVO: DETERMINAR SI EXISTEN GRUPOS DE VEHÍCULOS CON CARACTERÍSTICAS SIMILARES.



SOLUCIÓN: SE UTILIZÓ UN ANÁLISIS DE CLUSTER Y LOS RESULTADOS FUERON:



SE OBSERVA QUE AUDI, MERCEDES Y BMW
COMPARTEN CARACTERÍSTICAS SIMILARES.

EL DIAGRAMA MUESTRA DOS GRANDES GRUPOS Y
LUEGO OTRAS MARCAS QUE QUEDARON AISLADAS.

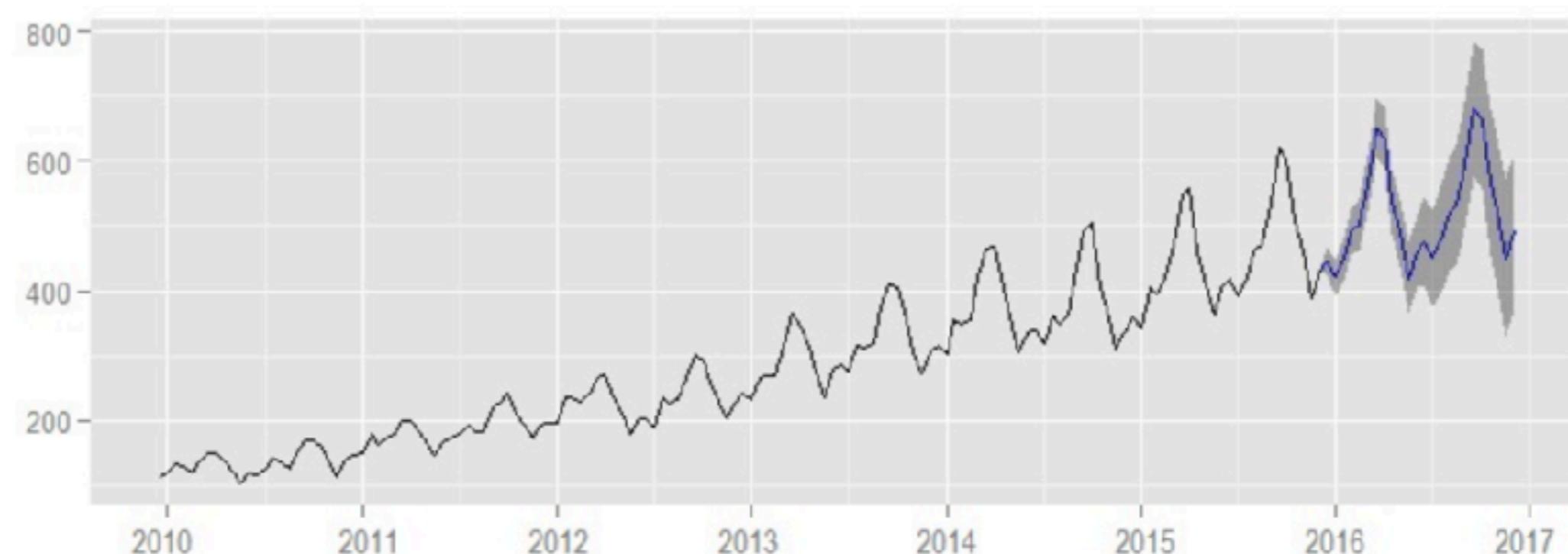
PREDICIENDO LA DEMANDA



PROBLEMA: SE CONSIDERA LA CANTIDAD DE PASAJEROS MENSUALES (EN MILES) EN EL TRANSPORTE AÉREO INTERNACIONAL DURANTE DOCE AÑOS CONSECUTIVOS.

OBJETIVO: PREDECIR LA DEMANDA DE PASAJEROS PARA EL PRÓXIMO AÑO.

SOLUCIÓN: SE USA UN MODELO DE SERIES DE TIEMPO ARIMA PARA PREDECIR LA DEMANDA DE LOS 12 MESES DEL PRÓXIMO AÑO.



LÍNEA DE LA POBREZA

PROBLEMA: SE SELECCIONAN ALEATORIAMENTE 30 ZONAS GEOGRÁFICAS Y SE CONSIDERAN SUS ESTADÍSTICAS DE DOS CENSO CONSECUTIVOS (CADA DIEZ AÑOS). LAS VARIABLES MEDIDAS SON:

POP CHNG: CAMBIO DE LA POBLACIÓN.

N EMPLOD: CANTIDAD DE EMPLEADOS QUE TRABAJAN EN AGRICULTURA.

PT POOR: PORCENTAJE DE FAMILIAS POR DEBAJO DEL NIVEL DE POBREZA.

TAX RATE: TASA DE IMPUESTO A PROPIEDADES RESIDENCIALES Y AGRÍCOLAS.

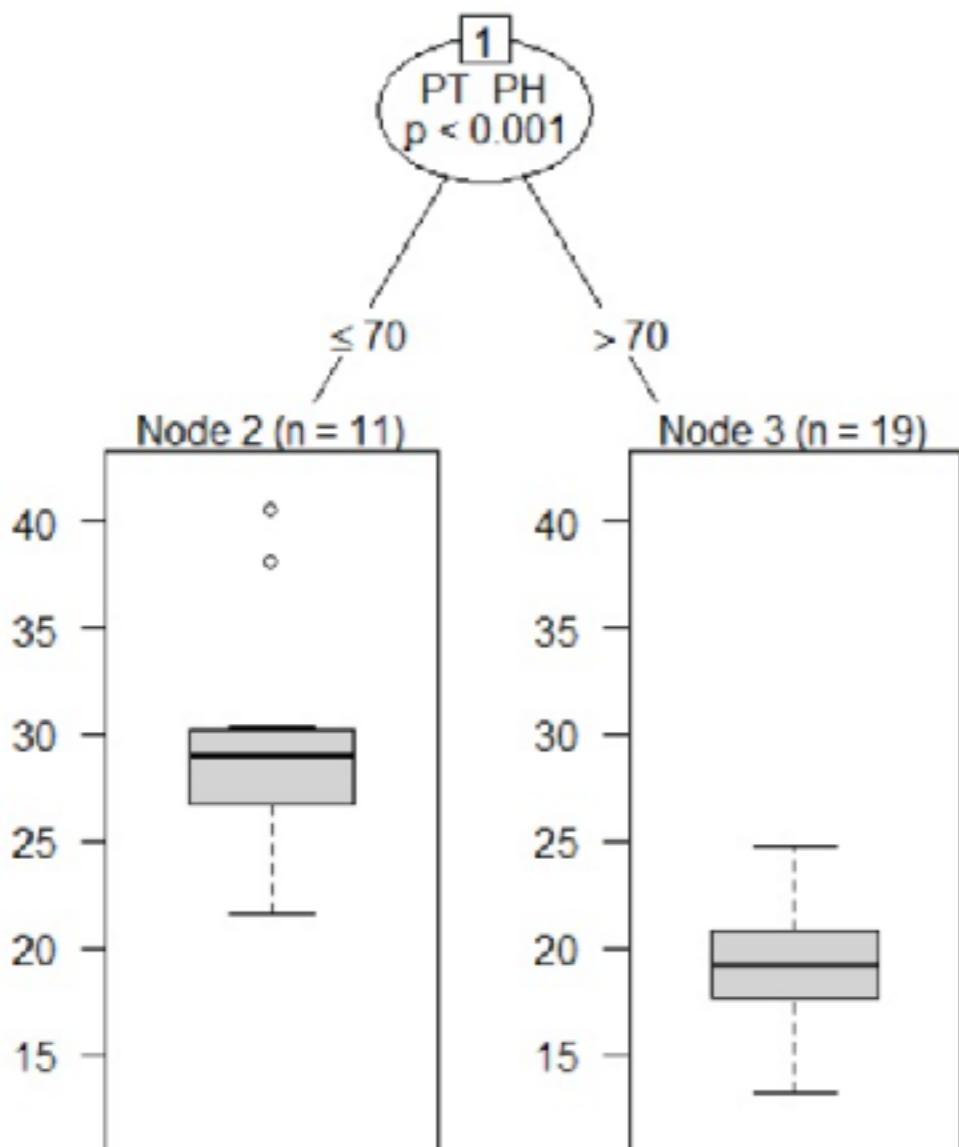
PT PHONE: PORCENTAJE DE RESIDENCIAS CON TELÉFONO

PT RURAL: PORCENTAJE DE POBLACIÓN RURAL

AGE: EDAD PROMEDIO

OBJETIVO: IDENTIFICAR LAS VARIABLES QUE MEJOR EXPLICAN QUE UNA FAMILIA ESTÉ POR DEBAJO DE LA LÍNEA DE POBREZA.

SOLUCIÓN: SE UTILIZA EL MÉTODO DE ÁRBOLES DE CLASIFICACIÓN PARA ENCONTRAR LAS VARIABLES QUE MEJOR EXPLICAN QUE UNA FAMILIA ESTÉ POR DEBAJO DE LA LÍNEA DE POBREZA.



EL PORCENTAJE DE POBREZA PROMEDIO ES DEL 23,0%.

ESTE PORCENTAJE AUMENTA A UN 29,6% SI EL PORCENTAJE DE RESIDENCIAS CON TELÉFONO ES INFERIOR AL 72%.

ESTE PORCENTAJE DISMINUYE A UN 19,2% SI EL PORCENTAJE DE RESIDENCIAS CON TELÉFONO ES SUPERIOR AL 72%.

EL PORCENTAJE DE RESIDENCIAS CON TELÉFONOS ES LA VARIABLE QUE MEJOR EXPLICA LA LÍNEA DE POBREZA DE UNA FAMILIA.

BIG DATA & DATA SCIENCE

VÍCTOR LEIVA

www.victorleiva.cl

ESCUELA DE INGENIERÍA INDUSTRIAL

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO