

Introducción a la analítica

Profesores César Augusto Gómez, Mauricio Alejandro Mazo y
Juan Carlos Salazar

jcsalaza@unal.edu.co



What is analytics? Analytics is an encompassing and multidimensional field that uses mathematics, statistics, predictive modeling and machine learning techniques to find meaningful patterns and knowledge in recorded data¹.

Today, we add powerful computers to the mix for storing increasing amounts of data and running sophisticated software algorithms – producing the fast insights needed to make fact-based decisions. By putting the science of numbers, data and analytical discovery to work, we can find out if what we think or believe is really true. And produce answers to questions we never thought to ask. That's the power of analytics.

¹https://www.sas.com/en_us/insights/analytics/what-is-analytics.html

Introducción y motivación

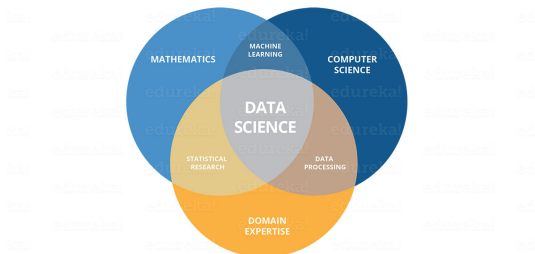


Figura 1: Habilidades en ciencia de datos: ML

Fuente: https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.rebelliouspixels.com%2Fblog%2Fhow-to-become-a-data-scientist-an-experts-guide%2F&psig=AOvVaw1oIIX7AiHwaj6B-eY_c4A1&ust=1596312685898000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCPDS36am-OoCFQAAAAAdAAAAABAJ

Introducción y motivación

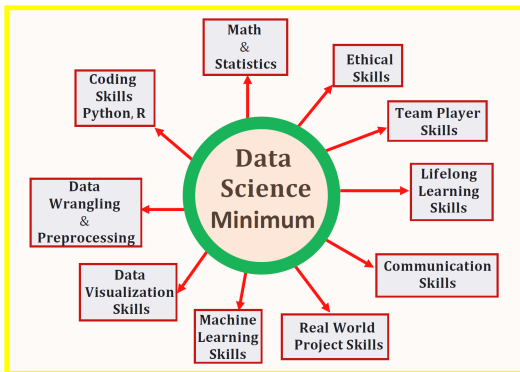


Figura 2: Habilidades en ciencia de datos: ML

Fuente: <https://www.google.com/url?sa=i&url=https%3A%2F%2Ftowardsdatascience.com%2Fdata-science-minimum-10-essential-skills-you-need-to-know-to-start-doing-data-science-e5a5a9be5991&psig=AOvVaw2hqqOMu5XLUDUn5ig6lv1&ust=1596311206280000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCLCqsOGg-OoCFQAAAAAdAAAAABAD>

Introducción y motivación

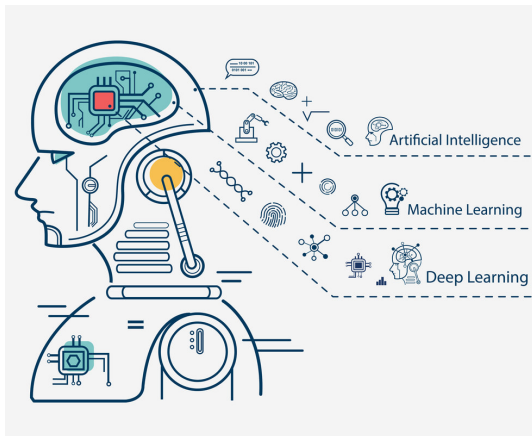


Figura 3: Importancia del ML en la estadística moderna

Fuente: <https://www.google.com/url?sa=i&url=https%3A%2F%2Fmachinelearningasaservice.weebly.com%2Fblog%2Fhow-to-use-machine-learning-in-data-quality-testing&psig=AOvVaw2o-XOkSSYs1eFKywbN63t&ust=1596300994332000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCLCXwuD69-oCFQAAAAAdAAAAABAK>


 Data Science vs Data Analytics		
	Data Science	Data Analytics
SKILLSET	<ul style="list-style-type: none">• Data Modelling• Predictive Analytics• Advanced Statistics• Engineering/Programming	<ul style="list-style-type: none">• BI Tools• Intermediate Statistics• Solid Programming Skills• Regular Expression (SQL)
SCOPE	Macro	Micro
EXPLORATION	<ul style="list-style-type: none">• Search Engine Exploration• Machine Learning• Artificial Intelligence• Big data - Often Unstructured	<ul style="list-style-type: none">• Data Visualization Techniques• Designing Principles• Big Data - Mostly Structured
GOALS	Discover New Questions to Drive Innovation	Use Existing Information to Uncover Actionable Data

Figura 4: Ciencia de datos vs Analítica

Fuente: <https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.pinterest.com%2Fpin%2F787355947333257839%2F&psig=AOvVaw0t-cVZqvSXsmBKq8S7Sckg&ust=1596312390277000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCNDQu5al-OoCFQAAAAAdAAAAABAI>

Why is analytics important? From the first known population data collection project by the Swedish government in 1749, to Florence Nightingale recording and analyzing mortality data in the 1850s, to British scholar Richard Doll's tobacco and lung cancer study in the 1950s, the analysis of data has fueled knowledge discovery for hundreds of years.

Each of the above scenarios required an answer to a heretofore unanswerable question. In the 1700s, the Swedes wanted to know the geographical distribution of their population to learn the best way to sustain an appropriate military force. Nightingale wanted to know the role that hygiene and nursing care played in mortality rates. Doll wanted to know if people who smoked were more likely to suffer from lung cancer. Each of these pioneers knew that instinct wasn't good enough. **Analysis of data** can uncover correlations and patterns.

En este curso se cubrirá mayoritariamente una rama de la estadística moderna y de la analítica (Analytics en inglés) conocida con **Aprendizaje Estadístico (AE)** y que hace parte de la era de la computación (Efron & Hastie, 2016)² que ha venido desarrollandose con mucha fuerza desde los años 50 hasta llegar a lo que hoy en día se conoce como la cuarta revolución industrial. Sin lugar a dudas, un estadístico moderno debe estar muy familiarizado y sentirse cómodo con este tipo de técnicas para garantizar un buen desempeño en el mundo académico e industrial y poder enfrentar los retos de la era moderna.

²Efron, B. and Hastie, T. (2016) *Computer Age Statistical Inference, Algorithms, Evidence, and Data Science*. Cambridge University Press

El aprendizaje estadístico (AE) está pensado y diseñado con el propósito de extraer y generar conocimiento a partir de datos a fin de orientar y facilitar la toma de decisiones en situaciones de alta variabilidad e incertidumbre. Es un campo de investigación en la intersección de estadística, inteligencia artificial y ciencias de la computación y también se conoce como análisis predictivo. La aplicación de los métodos de aprendizaje estadístico en los últimos años se ha vuelto omnipresente en la vida cotidiana.

Desde las recomendaciones automáticas de qué películas ver, hasta qué alimentos pedir o qué productos comprar, hasta la radio en línea personalizada y reconocer a sus amigos en sus fotos, muchos sitios web y dispositivos modernos tienen algoritmos de aprendizaje estadístico en su núcleo. Cuando se observa un sitio web complejo como Facebook, Amazon o Netflix, es muy probable que cada parte del sitio contenga varios modelos de aprendizaje estadístico³.

³Guido, S. and Muller, A.C. (2017). *Introduction to Machine Learning with Python*. O'relly

Fuera de las aplicaciones comerciales, el aprendizaje estadístico ha tenido una gran influencia en la forma en que se realiza la investigación basada en datos en la actualidad, como la comprensión de las estrellas, la búsqueda de planetas distantes, el descubrimiento de nuevas partículas, el análisis de secuencias de ADN y el suministro de tratamientos personalizados para el cáncer. Virtualmente, se puede encontrar en casi todas las actividades de la vida moderna. Todos estos argumentos deberían ser suficientes para motivar el estudio a profundidad del AE.

Algunos problemas que se pueden enfrentar con AE

- Identificar factores de riesgo para el cáncer de próstata.
- Clasificar archivos de audio con base en periodogramas.
- Predecir si alguien tendrá un ataque al corazón teniendo en cuenta mediciones demográficas, dietéticas y clínicas.
- Personaliza un sistema de detección de correo no deseado (SPAM).

Algunos problemas que se pueden enfrentar con AE

- Identificar los números en un código postal escrito a mano.
- Clasificar una muestra de tejido en una de varias clases de cáncer, basados en expresión genética.
- Establecer la relación entre salario y variables demográficas provenientes de encuestas poblacionales.
- Clasificar los píxeles en una imagen de satélite sobre uso de terrenos (LANDSAT).

¿Qué es AE? De acuerdo a James et al. (2014)⁴, AE se refiere a un conjunto extenso de herramientas para entender los datos (usualmente complejos), para generar conocimiento de calidad a partir de datos. Estas herramientas se clasifican en **supervisadas** y **no supervisadas** (También hay unas nuevas conocidas como **aprendizaje por refuerzo** (aprendizaje a través de errores) y hacen parte de inteligencia artificial (IA) y el **aprendizaje semi-supervisado**, que no se cubren en este curso, pero que se definen más adelante).

⁴James, G., Witten, D., Hastie, T. and Tibshirani, R. *An Introduction to Statistical Learning with Applications in R*. Springer

Aprendizaje estadístico supervisado

Por *aprendizaje estadístico supervisado* se entiende el proceso de construir un modelo estadístico para predecir o estimar un output o salida (en estadística usualmente es la variable respuesta y , también llamada variable dependiente, respuesta, objetivo) con base en un conjunto de inputs (en estadística, usualmente son las variables x , También llamadas, Regresores, covariables, características, variables independientes). Las respuestas pueden ser de tipo cuantitativo (continuas) y cualitativo (categorías)

Aprendizaje estadístico no supervisado

Por *aprendizaje estadístico no-supervisado* se entiende el proceso de encontrar patrones o agrupaciones en datos sin etiqueta; es decir sin variable y , solo se cuenta con los inputs o features x , y el objetivo es inferir propiedades de la distribución de probabilidad $P(x)$ sin la ayuda de un 'supervisor' y .

El **aprendizaje por refuerzo (RL)** es un área de aprendizaje automático que se ocupa de cómo el software debe tomar medidas en un entorno para maximizar la noción de recompensa acumulada. El aprendizaje por refuerzo es uno de los tres paradigmas básicos de aprendizaje automático, junto con el aprendizaje supervisado y el aprendizaje no supervisado. Se diferencia del aprendizaje supervisado en que los pares de entrada/salida etiquetados no necesitan presentarse, y las acciones subóptimas no necesitan ser explícitamente correctas. En cambio, el enfoque es encontrar un equilibrio entre la exploración (de un territorio desconocido) y la explotación (del conocimiento actual)⁵

⁵https://en.wikipedia.org/wiki/Reinforcement_learning

El aprendizaje semi-supervisado⁶ es una clase de tareas y técnicas de aprendizaje automático que también utilizan datos no etiquetados para el entrenamiento, generalmente una pequeña cantidad de datos etiquetados con una gran cantidad de datos no etiquetados (sin la variable respuesta y). El aprendizaje semi-supervisado⁷ se ubica entre el aprendizaje no supervisado (sin ningún dato de entrenamiento etiquetado, sin la variable respuesta y) y el aprendizaje supervisado (con datos de entrenamiento completamente etiquetados, con y)

⁶https://en.wikipedia.org/wiki/Semi-supervised_learning

⁷<https://stackoverflow.com/questions/19170603/what-is-the-difference-between-labeled-and-unlabeled-data/19172720#19172720>

AE es un área reciente en estadística y se mezcla con desarrollos paralelos en ciencias de la computación, en particular con aprendizaje de máquina (Machine learning, ML). AE encuentra aplicaciones en marketing, finanzas, ciencias médicas, biología, genética, ciencias de la tierra, ciencias sociales, astronomía, economía, entre otras.

Con la llegada de lo que se conoce como Big Data (*Big Data is a term that describes the large volume of data (both structured and unstructured) that inundates a business on a day-to-day basis. Big Data can be analyzed for insights that lead to better decisions and strategic business moves*)⁸, AE se ha convertido en un tópico de mucha actividad en diversas áreas científicas.

⁸www.sas.com/insights/big-data/what-is-big-data.html

¿Qué son datos estructurados? Los datos estructurados suelen clasificarse como datos cuantitativos, y es el tipo de datos con el que la mayoría de nosotros estamos acostumbrados a trabajar. Piense en datos que se ajusten perfectamente a campos y columnas fijos en bases de datos relacionales y hojas de cálculo. Ejemplos de datos estructurados incluyen nombres, fechas, direcciones, números de tarjetas de crédito, información sobre acciones, geolocalización y más.

¿Qué son los datos no estructurados? Los datos no estructurados suelen clasificarse como datos cualitativos, y no pueden procesarse ni analizarse utilizando herramientas y métodos convencionales. Ejemplos de datos no estructurados incluyen texto, video, audio, actividad móvil, actividad en redes sociales, imágenes satelitales, imágenes de vigilancia: la lista sigue y sigue.

El problema de aprendizaje supervisado Punto de partida:

- Medida de resultado Y .
 - Vector de p mediciones del predictor X (también llamadas inputs, Regresores, covariables, características (features), variables independientes).
- En el **problema de regresión**, Y es **cuantitativo** (por ejemplo, precio, presión sanguínea).
- En el **problema de clasificación**, Y toma valores en un conjunto finito no ordenado (sobrevive / muere, un dígito del 0-9, tipo de cáncer).
- Se tienen datos de entrenamiento $(x_1, y_1) \dots (x_N, y_N)$. Estos son Observaciones de estas medidas.

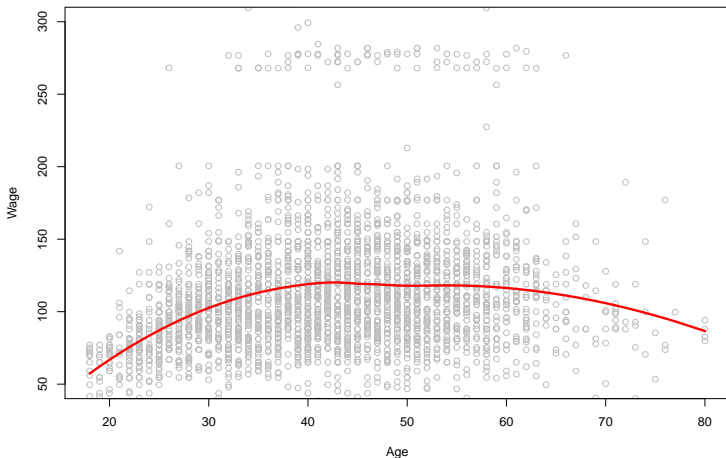
Sobre la base de los datos de entrenamiento se quiere:

- Predecir con precisión las observaciones nuevas (Test data o datos de prueba).
- Entender qué inputs afectan el resultado y cómo.
- Evaluar la calidad de nuestras predicciones e inferencias.

- Es importante entender las ideas detrás de las diversas técnicas, para saber cómo y cuándo usarlas.
- Se deben entender los métodos más simples primero, a fin de comprender los más sofisticados.
- **Es importante evaluar con precisión el rendimiento de un método, para saber qué tan bien o qué mal está funcionando.** (¡Los métodos más simples a menudo funcionan tan bien como los más sofisticados!)
- AE es un área de investigación emocionante, que tiene importantes aplicaciones en ciencia, industria y finanzas.
- El aprendizaje estadístico es un ingrediente fundamental en la formación de un científico moderno de datos.

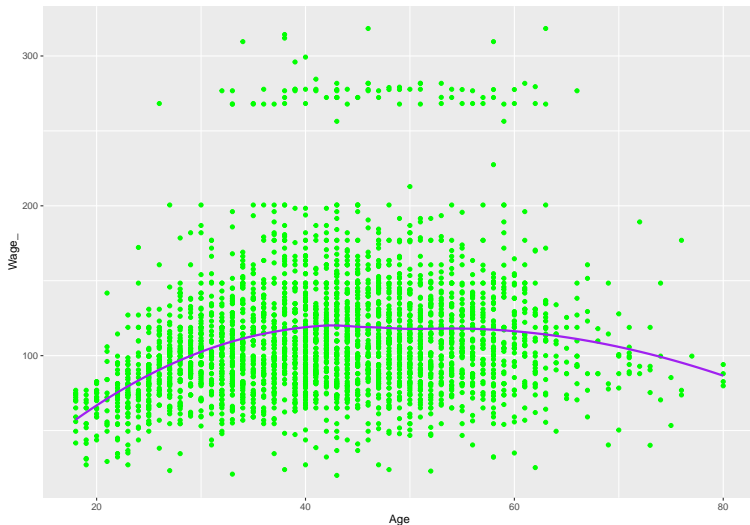
Introducción y motivación

- Datos sobre salarios (Wage Data, James et al. 2014). **Objetivo:** Establecer la relación entre salario y variables demográficas provenientes de encuestas poblacionales.



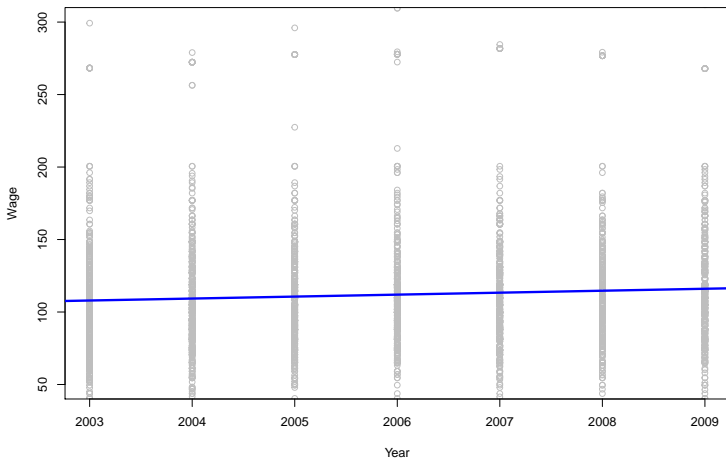
Introducción y motivación

- Establecer la relación entre salario y variables demográficas provenientes de encuestas poblacionales.



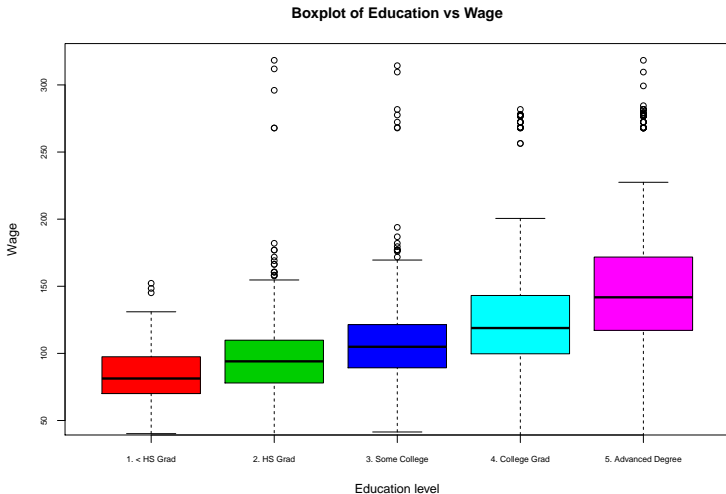
Introducción y motivación

- Establecer la relación entre salario y variables demográficas provenientes de encuestas poblacionales.



Introducción y motivación

- Establecer la relación entre salario y variables demográficas provenientes de encuestas poblacionales.



Específicamente, se quiere entender la relación entre el salario y la edad de un empleado, el año y el nivel de educación. De acuerdo al gráfico anterior, hay evidencia de que el salario se incrementa con la edad pero luego decrece aproximadamente después de los 60 años (relación no lineal, entre una variable continua o cuantitativa y otra cuantitativa). La línea ajustada, la cual es una estimación del salario promedio para una edad dada, captura esta tendencia. dada la edad de un empleado, se puede usar esta curva para predecir su salario. Sin embargo, se observa una alta variabilidad y por lo tanto la edad por si sola es improbable que proporcione una estimación precisa del salario particular de una persona.

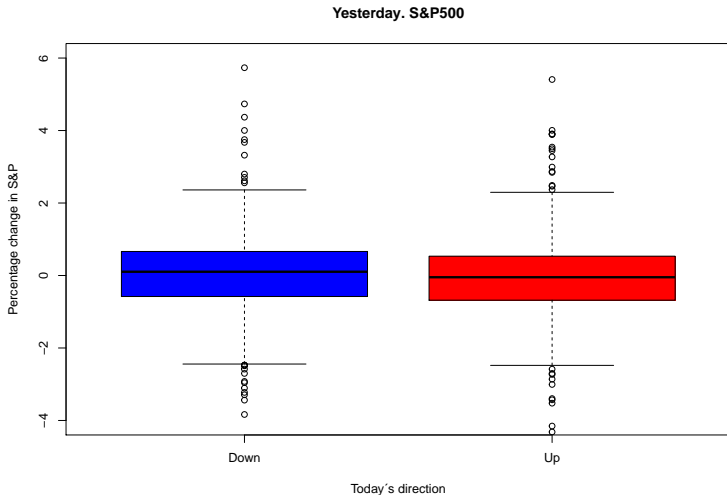
De hecho, si se observan los gráficos de años versus salario y de educación versus salario indican que estos factores están asociados con el salario. Salario se incrementa en aproximadamente 10000 US, casi linealmente, entre el 2003 y el 2009. Salario también es típicamente mayor para personas con un mayor nivel de educación. Tal vez, combinando edad, años y educación podría generar una mejor predicción del salario que usar solo la edad:

$$\text{Salario} \approx f(\text{Edad}, \text{Años}, \text{Educación})$$

Aquí Edad, años y educación son features, inputs, o predictores y salario es la respuesta o target la cual es continua. Este tipo de problemas se conocen como *problemas de regresión*.

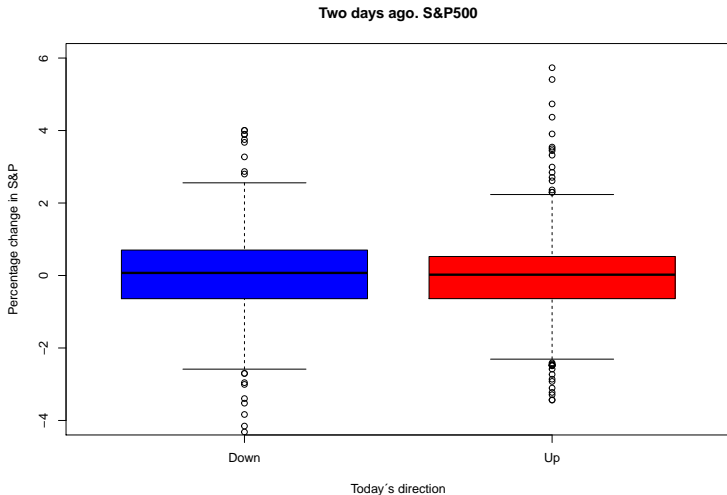
Introducción y motivación

- Datos del mercado de acciones (Stock market data, James et al. 2014)



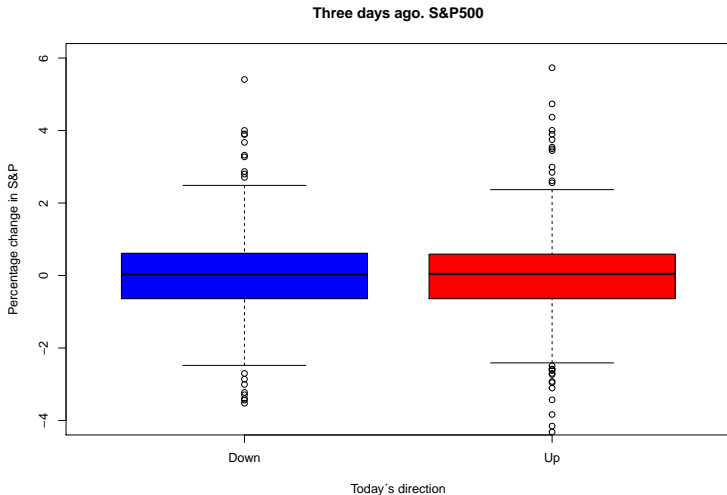
Introducción y motivación

- Datos del mercado de acciones (Stock market data, James et al. 2014)



Introducción y motivación

- Datos del mercado de acciones (Stock market data, James et al. 2014)

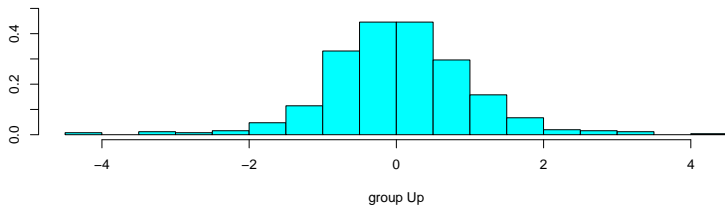
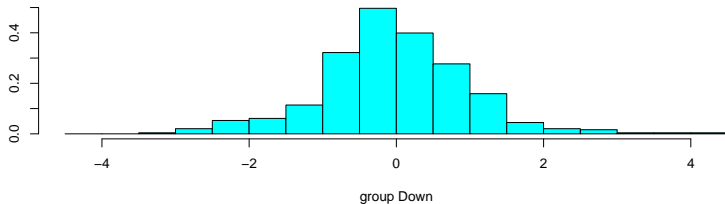


Con estos datos (período de 5 años), se quiere entender la relación entre un valor no numérico (un output o respuesta categórica) y unas features específicas, en este caso movimientos diarios previos. El objetivo es predecir si el índice crecerá o decrecerá en un día dado, usando los porcentajes de cambio en el índice de los últimos 5 días. Este es un *problema de clasificación*. Con base en los gráficos y del hecho de que son muy similares, se concluye que no hay una estrategia simple de usar los movimientos del día anterior para predecir los retornos del día de hoy. Esta falta de tendencia es de esperarse: en presencia de correlaciones fuertes entre retornos de días sucesivos, uno puede adoptar una estrategia de intercambio simple para generar ganancias de ese mercado.

Más adelante, se verá que hay algunas tendencias leves en estos datos que sugieren que, al menos para este período de 5 años, es posible predecir correctamente la dirección del movimiento en el mercado aproximadamente el 60 % de las veces.

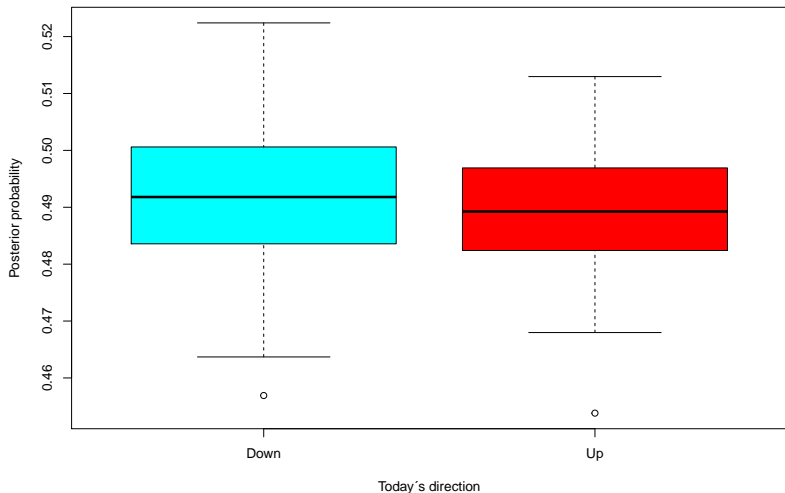
##		Direction.2005	
##		Down	Up
##	Down	35	35
##	Up	76	106

Introducción y motivación



```
##          Direction.2005
##          Down   Up
##   Down    30   20
##   Up      81 121
```

Introducción y motivación



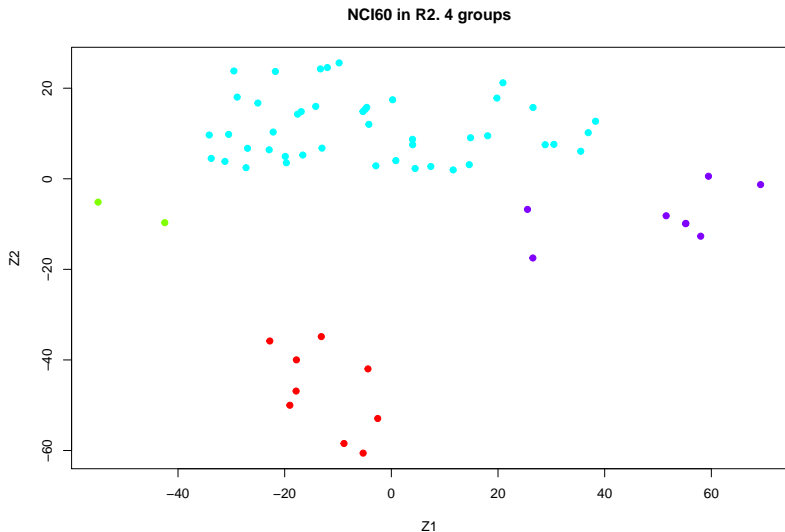
- Datos de expresión genética (Gene expression data, James et al. 2014)

Los dos ejemplos anteriores ilustran conjuntos de datos ambos con variables input y output. Sin embargo, otra clase importante de problemas se relaciona con situaciones en las cuales solo se observan variables input (solo las X) sin su correspondiente output (sin la Y). En estas situaciones, usualmente, no se busca predecir un output sino más bien agrupar individuos de acuerdo a sus características observadas (*clustering problem*).

- Datos de expresión genética (Gene expression data, National Cancer Institute (NCI) James et al. 2014)

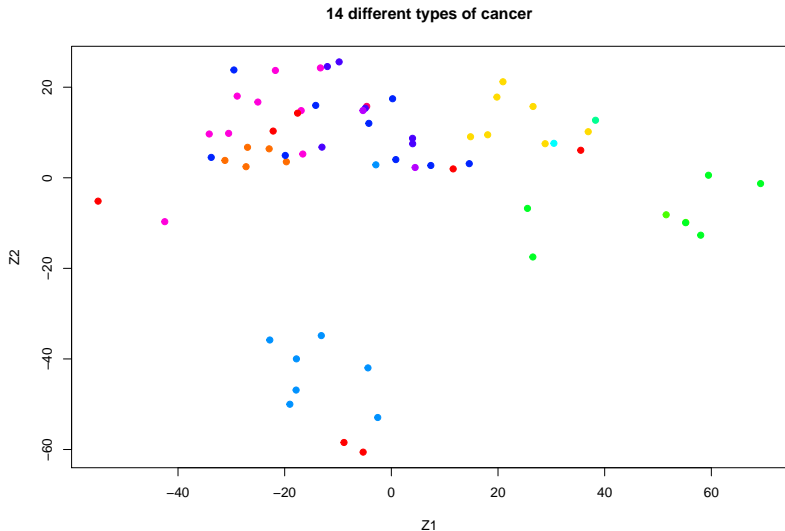
Considere el dataset NCI60, que consiste de 6830 mediciones de expresiones de genes para cada uno de 64 líneas de cáncer. En vez de predecir un output, interesa determinar si hay grupos, o clusters, entre las líneas de células con base en sus medidas de expresión genética. Esta es una pregunta difícil de responder ya que hay miles de expresiones genéticas por línea, lo cual dificulta visualizar los datos.

En la siguiente figura se ilustra como se puede abordar este problema representando cada una de las 64 líneas de células usando solo dos números, Z_1 y Z_2 conocidos como las dos primeras componentes principales de los datos, las cuales resumen las 6830 expresiones genéticas a solo dos dimensiones.



Aunque es muy probable que esta reducción de dimensionalidad haya producido alguna pérdida importante de información, es ahora posible examinar visualmente los datos en búsqueda de evidencia de clustering. Decidir el número correcto de grupos o clusters es frecuentemente un problema difícil, pero el gráfico de arriba sugiere al menos 4 grupos de líneas de células. De esta manera, el problema se reduce a examinar las líneas de células dentro de cada cluster para identificar similitudes en sus tipos de cáncer a fin de entender mejor la relación entre niveles de expresión genética y cáncer.

En este conjunto de datos, pasó que las líneas de células corresponden a 14 tipos distintos de cáncer. Hay evidencia clara de que las líneas de células con el mismo tipo de cáncer tienden a localizarse cerca entre ellas en esta representación bidimensional.



De acuerdo a James et al. 2014, ISLR se basa en cuatro premisas:

- Muchos métodos de AE son relevantes y útiles en una amplia gama de disciplinas académicas y no académicas, más allá de solo las ciencias estadísticas.
- AE no debe ser visto como una serie de cajas negras
- Mientras es importante saber qué trabajo ejecuta cada engranaje, no es necesario tener la habilidad para construir toda la máquina dentro de la caja
- Se asume que las personas están interesadas en aplicar AE a problemas de la vida real.

Aprendizaje estadístico versus aprendizaje automático o de máquina

- El aprendizaje de máquina surgió como un sub campo de la inteligencia artificial.
- El aprendizaje estadístico surgió como un sub campo de la estadística.

Hay mucho parecido entre ellos: ambos campos se centran en **problemas supervisados y no supervisados**

- ML tiene un mayor énfasis en aplicaciones de gran escala y precisión de predicción.
- AE enfatiza más en los modelos y su interpretabilidad, precisión e incertidumbre.

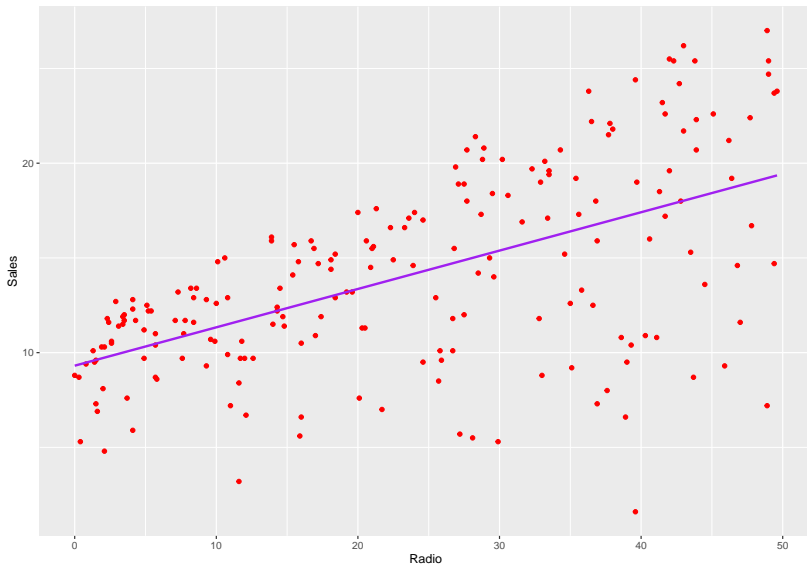
Pero la distinción se ha vuelto cada vez más borrosa y ambas se ayudan entre si. ¡El aprendizaje automático tiene la ventaja en marketing!

¿Qué es AE? A fin de motivar el estudio de AE vamos a usar un ejemplo. Suponga que usted es un consultor estadístico contratado por un cliente para proporcionar consejos de cómo mejorar las ventas de un producto particular. El conjunto de datos de publicidad (Advertising dataset, James et al. 2014) consiste de ventas (y) de ese producto en 200 mercados distintos, junto con presupuestos para publicidad de ese producto en cada uno de los mercados a través de tres medios: Tv, Radio y periódico (las X). Los gráficos de los datos:

Aprendizaje estadístico



Aprendizaje estadístico



Aprendizaje estadístico

