

UCLA
Department of Statistics Papers

Title

Cómo convertir recursos de alfabetización estadística en recursos de ciencia de datos

Permalink

<https://escholarship.org/uc/item/3jf6c838>

Author

Sanchez, Juana

Publication Date

2026-10-13

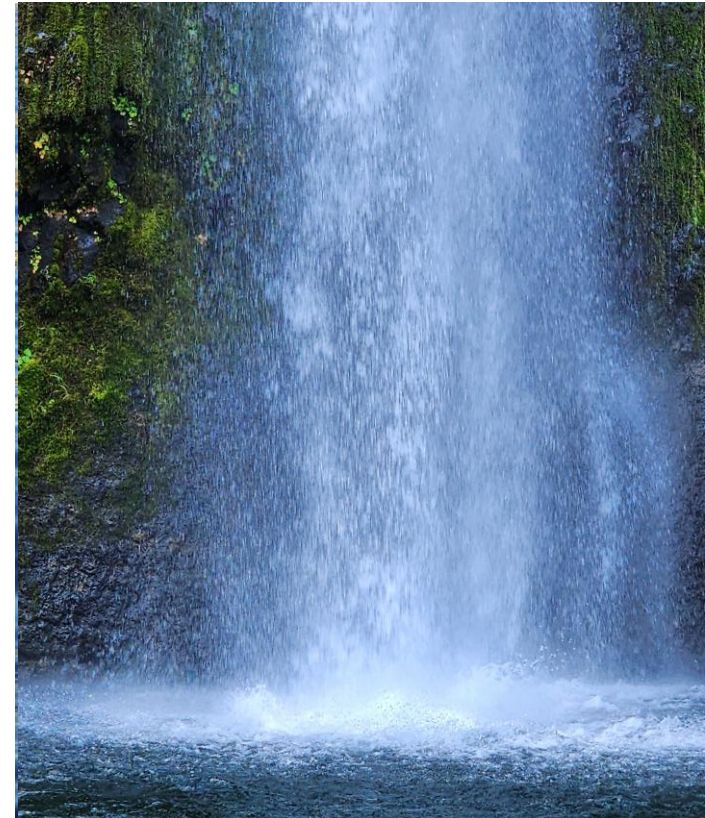
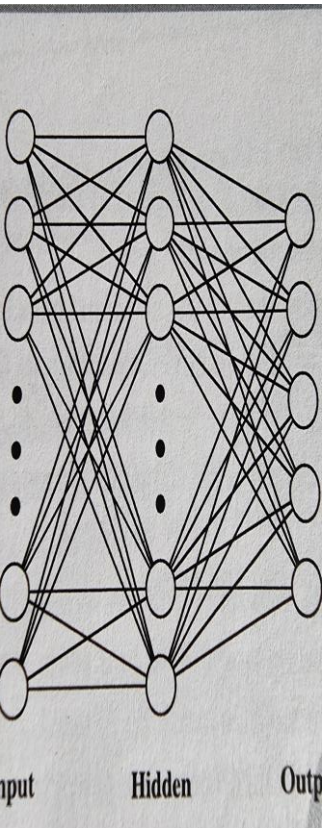
Cómo convertir recursos de alfabetización estadística en recursos de ciencia de datos

Juana Sanchez, Ph.D.

UCLA Dept of Statistics and Data Science

Editor-in-Chief, Journal of Statistics and Data Science Education

XI Jornada de Educación Estadística «Martha Aliaga», 13 de Octubre, 2026



Gracias por invitarme a participar en la Conferencia Inaugural de la XI Jornada de Educación Estadística «Martha Aliaga»



Aquí vemos a Martha en acción, participando en la *IASE Roundtable Conference : Challenges for Teaching and Teacher Education*, Monterrey Mexico, 2008. Martha era una firme creyente en el aprendizaje activo y una ferviente promotoradora de la educación estadística.

Gracias en particular a Adriana D'Amelio por todo lo que hace para la promoción de la educación estadística desde 2007 hasta hoy. Sus iniciativas sostenidas, equitativas y de gran impacto elevan la presencia internacional de los estudiantes argentinos en el mundo. Su liderazgo no tiene igual.



Aquí vemos Adriana D'Amelio durante la *IASE Roundtable Conference : Challenges for Teaching and Teacher Education*, Monterrey Mexico, 2008 (izquierda), y durante la 57.^a Sesión (conferencia bienal) del Instituto Internacional de Estadística (ISI), 2009, en Durban, Sudáfrica, con miembros del ISLP (en medio y derecha)

Esquema de la presentación

1. Introducción
2. Contexto educativo en el que vivo. Marco, directrices
3. Conectando la ciencia de datos (la base de la IA) y la estadística.
4. Ejemplo 1. ¿Son justos los algoritmos de ciencia de datos (y, por ende, de la IA)? En el curso: Introducción a la probabilidad
5. Ejemplo 2. Tema del ejemplo: Introduciendo diagramas de cajas comparativos a través de series temporales para predecir la demanda de electricidad en el sur d California. En el curso: Introducción a la estadística.
6. Ejemplo 3. Micromovilidad a pequeña escala. ¿Dónde hay mayor demanda u oferta de patinetes eléctricos? En el curso: Introducción a la probabilidad
7. Ejemplo 4.
8. Conclusiones

Alerta

El inglés utiliza un punto como separador decimal y una coma para los millares, mientras que el español tradicional emplea una coma como separador decimal y un punto (o un espacio) para los millares. Los números de esta presentación están escritos siguiendo el formato inglés.

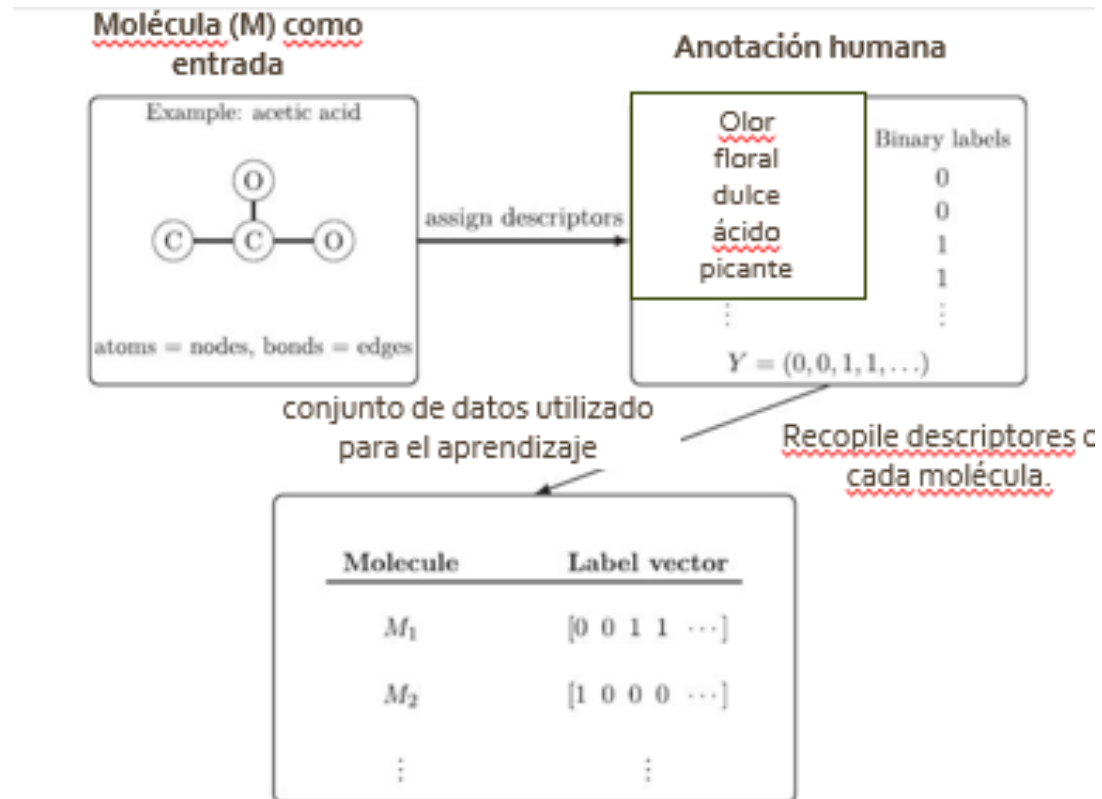
1. Introducción

Esta charla aborda estrategias y ejemplos sobre cómo ayudo a estudiantes de grado —en cursos introductorios de estadística o probabilidad en mi departamento de estadística— a comprender que los conceptos que están aprendiendo constituyen un componente fundamental del emergente campo de la Ciencia de Datos (CD) en ámbitos muy diversos . Eso conlleva:

- **Enfatizar el diverso origen e importancia del manejo de datos** porque eso ayuda a crear datos conjuntos formateados como los que se utilizan en las clases, aunque mucho mas voluminosos. La representación gráfica, los conceptos de probabilidad, los resúmenes aprendidos en clase se pueden utilizar de muchas maneras. La calidad y el origen de los datos determina si los algoritmos con ellos entrenados dan conclusiones sesgadas. De ahí su importancia.
- **Abordar la importancia de la predicción** en el panorama actual de la ciencia de datos y cómo esta impulsa el desarrollo de la inteligencia artificial.
- **Conectar lo conocido con lo desconocido a un nivel adecuado para la clase.** Sin sacrificar el plan de estudios de estadística, se toca la CD lo suficiente para hacer que los alumnos se sientan un poco parte del nuevo paisaje.

Idea principal: Datos surgen de preguntas y ayudan a predecir. Por ejemplo: ¿Qué es este olor?

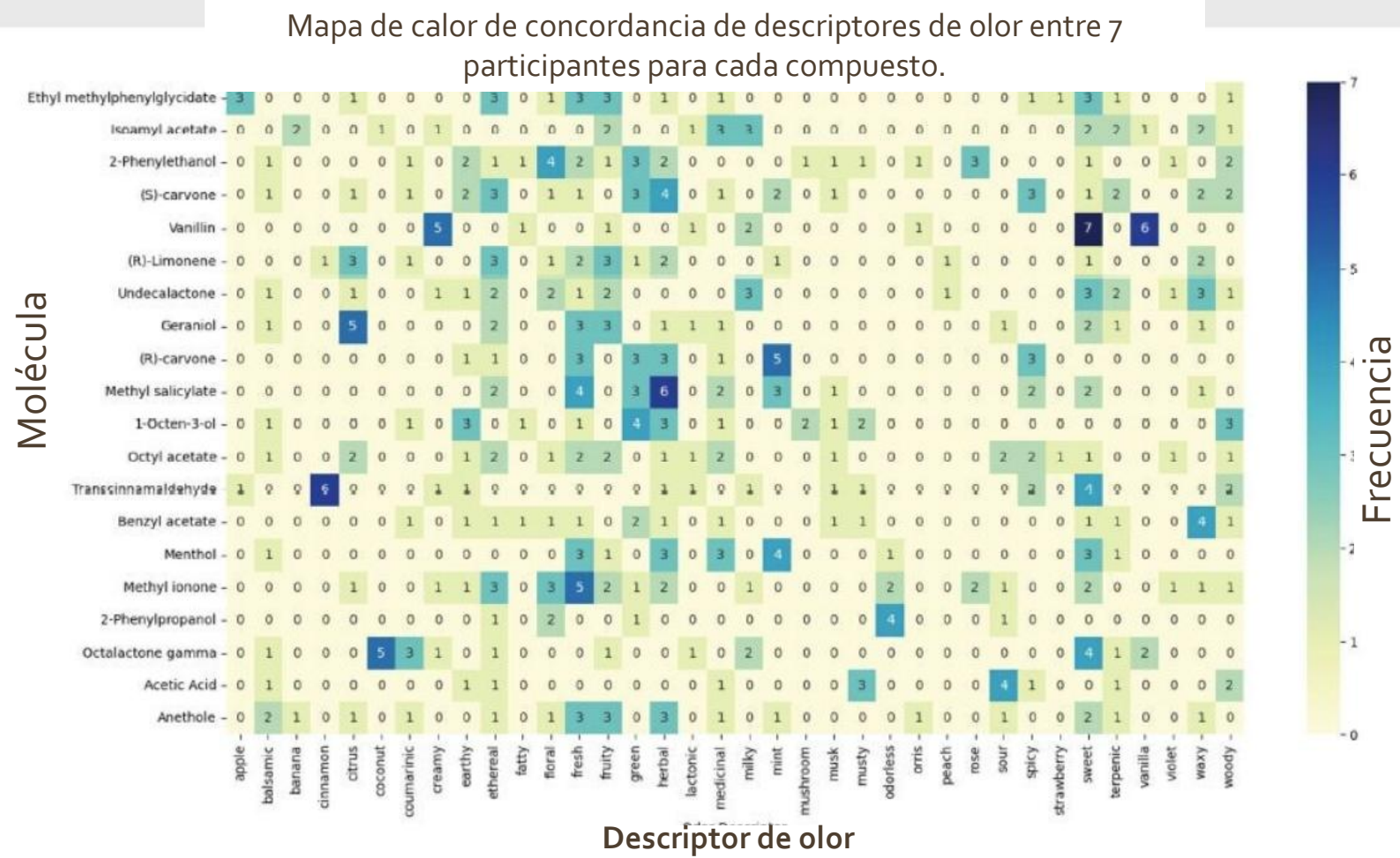
Figure 2. Representación esquemática del sistema de olfacción artificial. Una molécula se representa como un grafo en el que los átomos son nodos y los enlaces químicos son aristas. Mediante anotación humana se asignan descriptores de olor a la molécula, generando un vector de etiquetas binarias que indica qué descriptores son aplicables. Un conjunto de datos para aprendizaje automático se compone de numerosos pares de molécula y etiqueta de este tipo. **Fuente: Han et al. (2026)**



- Podríamos usar un olfato electrónico, una máquina que ha sido entrenada con datos de olores previamente y que al dar un olor nuevo predecirá lo que el nuevo olor es.
- Que son datos de olores? Datos obtenidos de moléculas odoríferas
- Para ver como producir datos para entrenar la máquina, se puede hacer un ejercicio de olfato (detección y procesamiento de moléculas químicas presentes en el aire).
- Al acabar el ejercicio, se les dice a los alumnos que usando Redes neuronales de grafos (GNN) la máquina predice lo que el nuevo olor es.

Figura 3. Frecuencia de selección de descriptores entre siete participantes para cada compuesto. Cada celda [r, c] indica cuántos estudiantes seleccionaron el descriptor c para el compuesto r (0-7). La figura pone de relieve tanto el consenso (celdas de alta frecuencia) como la variabilidad (selecciones dispersas de baja frecuencia) en los descriptores de olor señalados por los estudiantes para los distintos compuestos.

Fuente: Han et al. (2026)



La fuente del ejemplo sobre el olfato y de los gráficos es un artículo publicado por el *Journal of Statistics and Data Science Education* recientemente

Han, Y., Kydd, M. K., Ward, J., & Tewari, A. (2026). Teaching Machine Olfaction in an Undergraduate Deep Learning Course: A Pilot Integrating Chemistry, Machine Learning, and Sensory Evaluation. *Journal of Statistics and Data Science Education*, 34(3), 372–380.

<https://doi.org/10.1080/26939169.2026.2665103>

Los materiales facilitados a los estudiantes para crear el conjunto de datos se encuentran en un repositorio creado por los autores de dicho artículo:

<https://osf.io/jqn4y/>

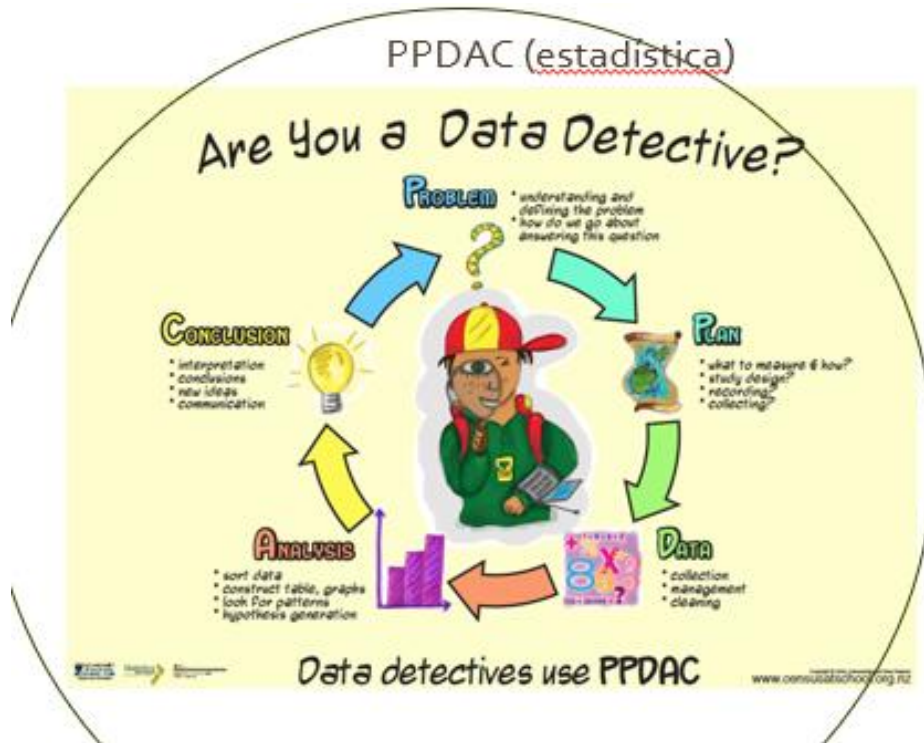
Este será el tema del seminario web de JSDSE/CAUSE del 27 de octubre de 2026, de 16:00 a 16:45 h ET (17:00 h en Argentina). La actividad es gratuita, pero requiere inscripción antes del 27 de octubre en

<https://www.causeweb.org/cause/webinar/jsdse/2026-10>

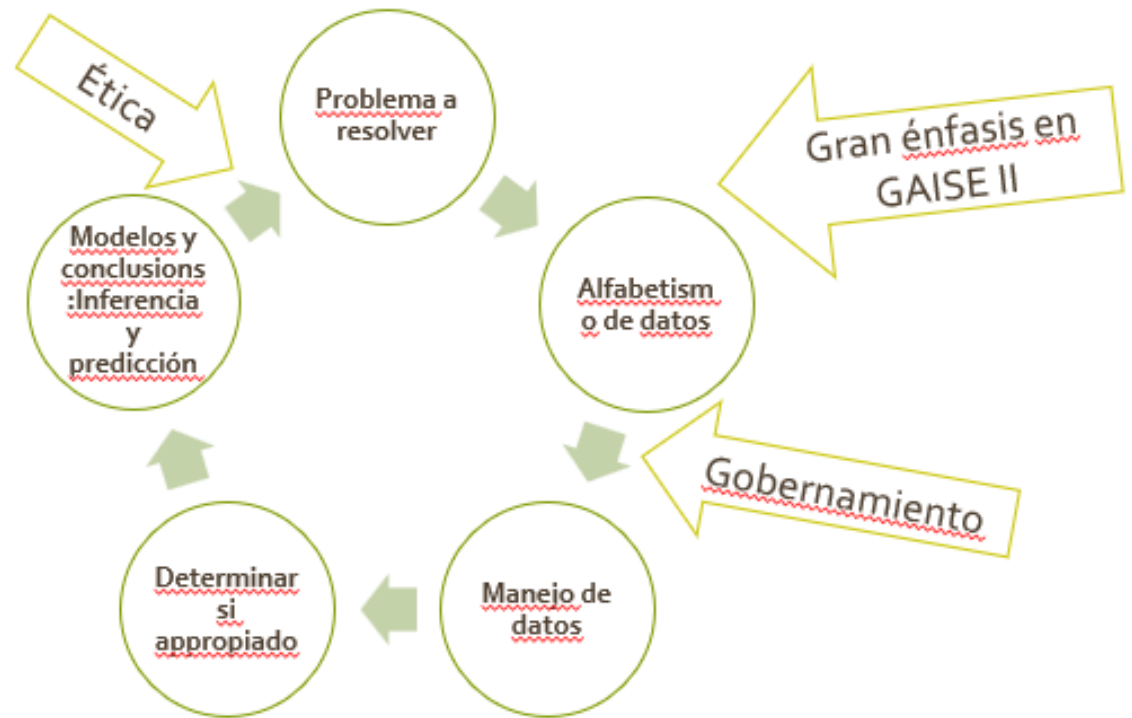
2. El contexto educativo en el que se implementaron las actividades que serán mencionadas en esta charla.

- Desde 1999, en mi departamento y en GAISE la estadística es la ciencia de los datos (prácticas de análisis de datos en cursos introductorios, manuales de laboratorio y libros de texto que hacían hincapié en el análisis de datos se llevan usando desde entonces).
- La Universidad de California en Los Ángeles se considera una universidad de investigación. Tanto la docencia de grado como la investigación son altamente valoradas por profesores y estudiantes, a pesar de su fama como universidad de investigación de postgrado.
- A principios de la década de 2020 los alumnos de los cursos introductorios se empezaron a percibir del aprendizaje electrónico (ML), la inteligencia artificial, redes neuronales, alfabetización en datos y tuve la necesidad de satisfacer la curiosidad de los alumnos en cursos introductorios de estadística y probabilidad con respecto a la “nueva” ciencia de datos. Se requería información accesible que les hiciera ver la utilidad de lo aprendido en la clase para la nueva ciencia de datos. Hacían falta ejercicios y debates adaptados al nivel del curso que destacaran los aspectos clave—procedencia de los datos, predicción, ética—, partiendo de conceptos que ya resultaran familiares a los estudiantes de estas asignaturas introductorias.

3. Conectando. Los que aplican ciencias de datos utilizan conocimientos estadísticos, pero con mayor volumen, velocidad y tipos de datos, con más potencia informática que en el pasado y con otros énfasis.



GAISE I, GAISE II, Census@School, ISLP, OECD, y en numerosos escenarios de todo el mundo, así como en libros de introducción a la estadística desde hace va muchos años.



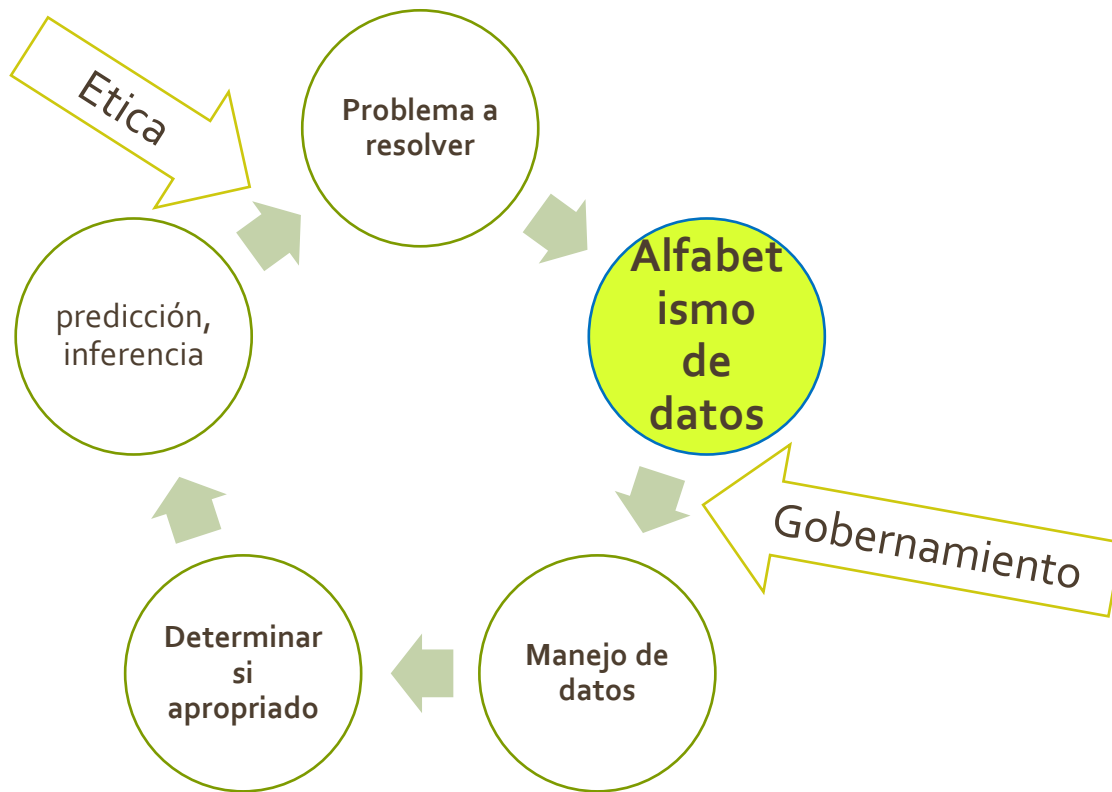
Keller, S.A, et al. (2020): Doing **Data Science**: A Framework and Case Study.

Importante comparar los dos lenguajes

- Los profesionales de la ciencia de datos provienen de diferentes campos (informática, estadística, ingeniería, humanidades, economía, contadores, empleados comerciales, etc.)
- Hay diferentes conjuntos de habilidades en cada profesional de la ciencia de datos.
- Como resultado, los términos que utilizamos en estadística han sido renombrados de diferentes formas. Es necesario hacer conscientes a los alumnos."

Acción	Nombre en las estadísticas	Nombre en CD, ML
Órdenes impartidas a las funciones de algoritmos de software	Argumentos de funciones (por ejemplo, en R, <code>modelo1=lm(sueldo~educación)</code>)	Hiperparámetros
Nombres asignados a los datos recopilados	Variables	Características
Transformaciones o combinaciones de variables	Preparación o gestión de datos (limpieza, preparación, vinculación, exploración)	Ingeniería de características
Determinar el modelo de población	Estimar el modelo	Entrenamiento del modelo
Datos sobre los datos (metadatos, procedencia)	Quién, qué, cuándo, cómo, dónde	alfabetización en datos
Generación de conocimiento a partir de datos	proceso de investigación	canalización de datos
¿Qué nos permite generar números aleatorios multivariantes?	distribución de probabilidad conjunta	modelo generativo

La profundidad y el alcance de la conexión que establezco entre la estadística clásica y el entorno del profesional de la ciencia de datos dependen de las competencias de los alumnos.



Conjunto mínimo de competencias: «ser capaz de comprender información extraída de datos y resumida en estadísticas simples, realizar cálculos adicionales utilizando dichas estadísticas y emplear estas últimas para tomar decisiones». Bonikowska et al. (2019) —lo que va más allá de esto se adquiere en la educación superior.

Un conjunto de habilidades más amplio: «la capacidad de plantear y responder a una pregunta del mundo real a partir de conjuntos de datos grandes y pequeños mediante un proceso de indagación, teniendo en cuenta el uso ético de los datos» (Wolff et al., 2016). Suena a todo el ciclo PPDAC, con distintos niveles de competencias informáticas de por medio.

Definición restringida: capacidad para elaborar un inventario de datos y utilizar todo tipo de datos disponibles en la mayor variedad de formatos posible. Keller, S. A. et al. (2020)



Ejemplo 1. Contexto: Introducción a la probabilidad

¿Son justos los algoritmos de inteligencia artificial?

Algoritmos utilizados para extraer conocimiento a partir de datos. Se dice que son desconocidos para el o demasiado complejos; sin embargo, podemos medir su equidad utilizando datos sobre sus resultados y conceptos básicos de estadística o probabilidad. Modelos generativos.

Introducción a la probabilidad: probabilidad condicional, probabilidades conjuntas, probabilidades marginales, construcción de tablas de contingencia a partir de datos.

The New York Times

Biased Algorithms Are Easier to Fix Than Biased People

Racial discrimination by algorithms or by people is harmful — but that's where the similarities end.

Give this article



Tim Cook

By Sendhil Mullainathan
Dec. 6, 2019

In [one study](#) published 15 years ago, two people applied for a job. Their résumés were about as similar as two résumés can be. One person was named Jamal, the other Brendan.

In [a study](#) published this year, two patients sought medical care. Both were grappling with diabetes and high blood pressure. One

<https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html>

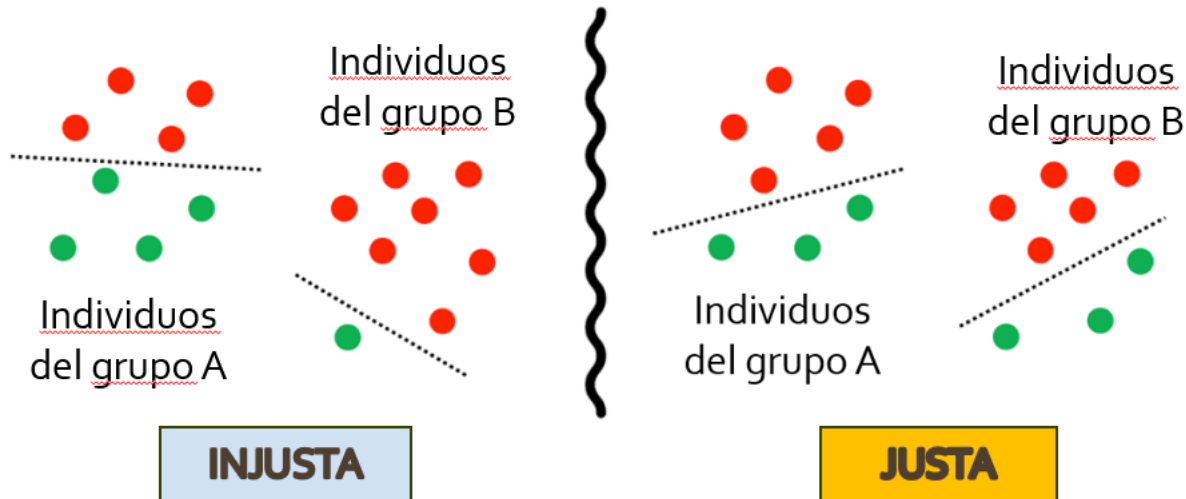
Equidad algorítmica con respecto a un grupo demográfico protegido de naturaleza binaria

Datos utilizados para entrenar el algoritmo



¿Dados esos datos que entrenaron el algoritmo, qué predice el algoritmo de nuestro banco sobre la probabilidad de que este nuevo cliente pague el préstamo solicitado?

Tipo de predicciones



Equidad algorítmica



LoanID	G	T	D
201	1	1	1
210	0	1	0
214	1	0	1
290	1	1	0
310	1	1	1
340	1	1	1
...	...	...	...
...	...	...	...

$D = 0$

$D = 1$

	$G = 0$	$G = 1$
$T = 0$	0.21	0.32
$T = 1$	0.07	0.28

	$G = 0$	$G = 1$
$T = 0$	0.01	0.01
$T = 1$	0.02	0.08

adolfoeliazat.com

Las tablas podían contabilizarse como recuentos antes de esto.

Se utilizará un algoritmo de inteligencia artificial para realizar una predicción binaria sobre si una persona devolverá un préstamo. Surge la siguiente cuestión: ¿es el algoritmo «justo» con respecto a un grupo demográfico protegido de naturaleza binaria? Notación: $G=1$ (predicción de que la persona pagará el préstamo); D = grupo demográfico; $T=1$ (la persona paga el préstamo).

Fuente de este ejemplo

:<https://chrispiech.github.io/probabilityForComputerScientists/en/examples/fairness/>

G=1 (predicción de que paga);
D = grupo demográfico;
T=1 (paga)

	D = 0			D = 1	
	G = 0	G = 1		G = 0	G = 1
T = 0	0.21	0.32	T = 0	0.01	0.01
T = 1	0.07	0.28	T = 1	0.02	0.08

$$\begin{aligned}
 P(G = 1|D = 1) &= \frac{P(G = 1, D = 1)}{P(D = 1)} \\
 &= \frac{P(G = 1, D = 1, T = 0) + P(G = 1, D = 1, T = 1)}{P(D = 1)} \\
 &= \frac{0.01 + 0.08}{0.12} = 0.75 \\
 \\
 P(G = 1|D = 0) &= \frac{P(G = 1, D = 0)}{P(D = 0)} \\
 &= \frac{P(G = 1, D = 0, T = 0) + P(G = 1, D = 0, T = 1)}{P(D = 0)} \\
 &= \frac{0.32 + 0.28}{0.88} \approx 0.68
 \end{aligned}$$

Concepto de equidad
algorítmica 1: paridad
demográfica:
P(G=1|D=1)=P(G=1|D=0)

Fuente: <https://chrispiech.github.io/probabilityForComputerScientists/en/examples/fairness/>

Tras recibir esta formación orientada a la evaluación formativa, los alumnos realizan una encuesta entre los estudiantes de la UCLA para responder a preguntas de interés, elaborar tablas similares y demostrar el teorema de Bayes. Se observa la similitud entre la equidad algorítmica y otros problemas.

Para profundizar en el tema, los estudiantes analizan y debaten cómo los modelos de IA generativa utilizan probabilidades conjuntas para crear datos nuevos (sintéticos) y cómo los modelos de IA discriminativa emplean probabilidades condicionales y datos existentes para clasificarlos. Asimismo, buscan bibliografía de su propia disciplina que aplique estos enfoques.

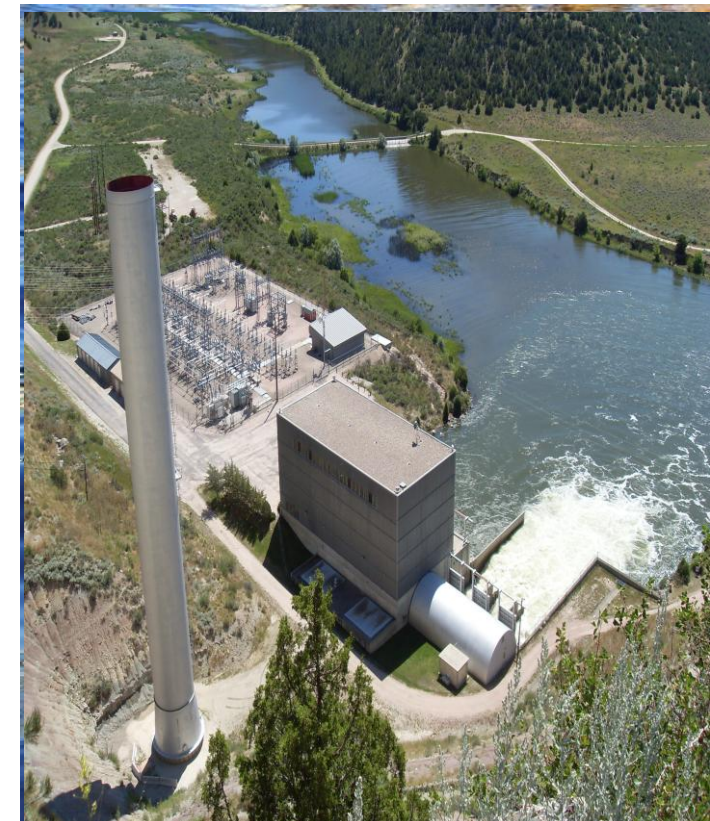
Todo esto puede abordarse durante las dos primeras semanas de un curso introductorio de probabilidad. En esas dos primeras semanas se adquieren algunos fundamentos de la inteligencia artificial.

SCE's service area includes portions of 15 counties and hundreds of cities and communities in a 50,000-square-mile service area within Central, Coastal and Southern California.

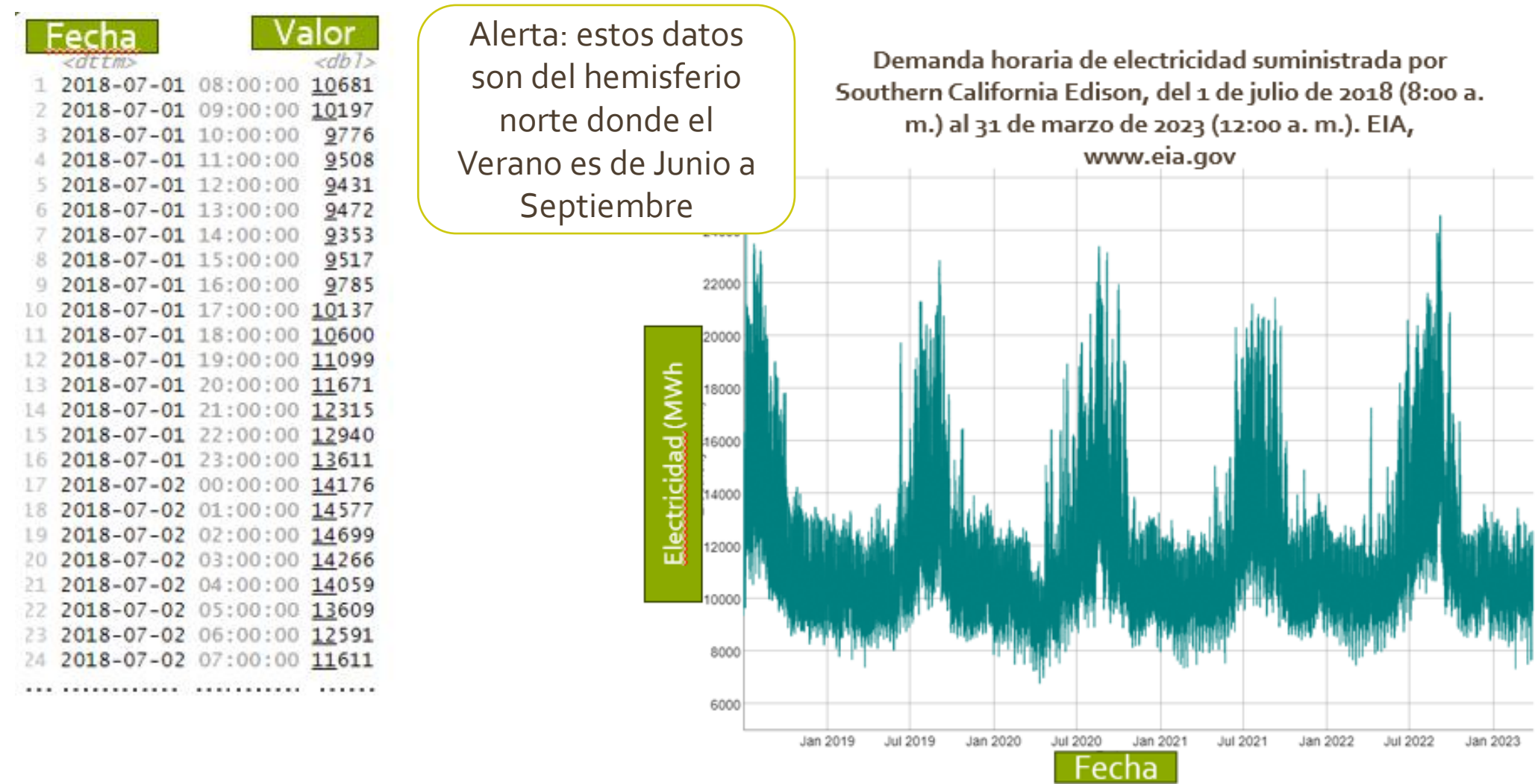


Contexto para el profesional de ciencia de datos: ingeniería de características, regresión múltiple y regresión mediante aprendizaje automático (ML). Aprendizaje automático supervisado.

Contexto en estadística: preparación de datos (data wrangling), datos multivariantes, introducción a la estadística descriptiva, regresión e inferencia, incluyendo variables que reflejan la naturaleza temporal de los datos (mes, día, hora...).



La mayor parte de los datos que se recopilan hoy en día llevan marca temporal (sobretudo por internet). Es importante hacer a los alumnos conscientes de esto.



Fuente: Sanchez, J. (2023) , Estudio de caso para el capítulo 10, timeserietime.org

Preparar los datos para modelos de aprendizaje automático (RF, GB, NN) y para regresión múltiple convencional (así como para análisis multivariante de datos propio de la estadística introductoria)

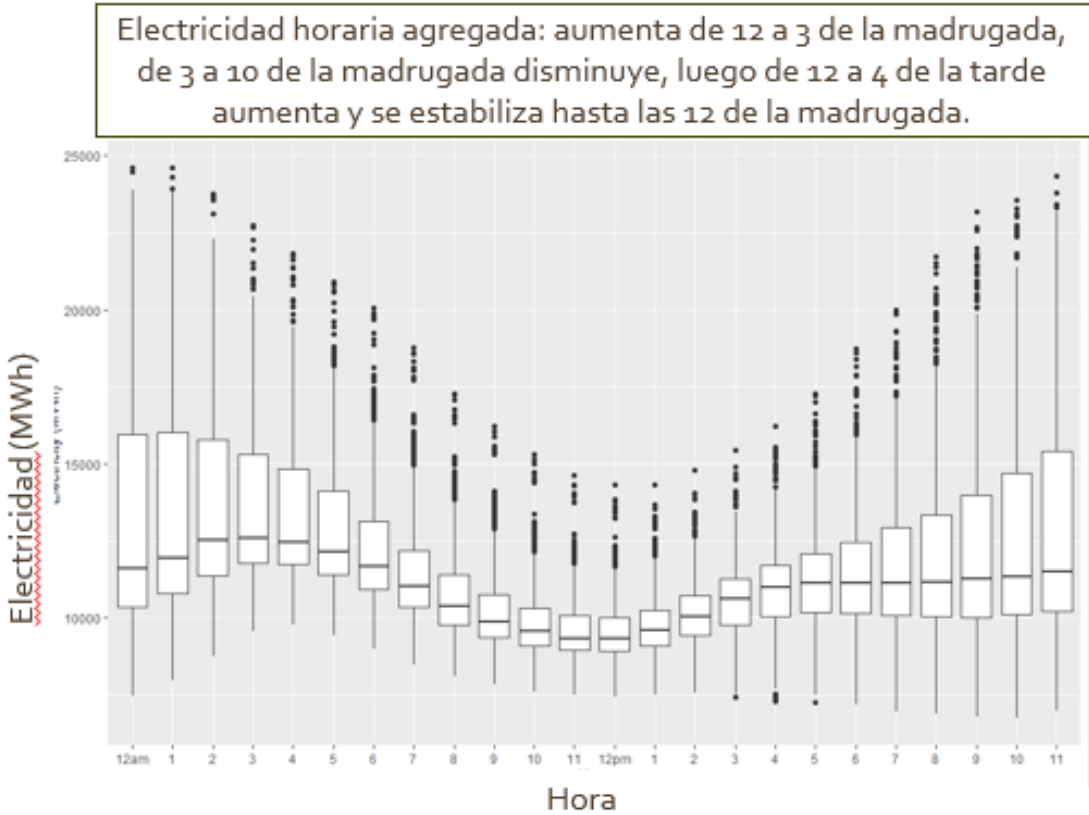
Los datos de la izquierda se transforman en un formato de conjunto de datos multivariante familiar para los estudiantes de estadística básica (para análisis multivariante elemental) o para estudiantes más avanzados (para el entrenamiento de modelos de aprendizaje automático como redes neuronales, bosques aleatorios o bien para realizar una regresión múltiple, que suele servir como modelo de referencia). La variable de fecha ya no es necesaria.

	Fecha	Valor
	<dtm>	<dbl>
1	2018-07-01 08:00:00	10681
2	2018-07-01 09:00:00	10197
3	2018-07-01 10:00:00	9776
4	2018-07-01 11:00:00	9508
5	2018-07-01 12:00:00	9431
6	2018-07-01 13:00:00	9472
7	2018-07-01 14:00:00	9353
8	2018-07-01 15:00:00	9517
9	2018-07-01 16:00:00	9785
10	2018-07-01 17:00:00	10137
11	2018-07-01 18:00:00	10600
12	2018-07-01 19:00:00	11099
13	2018-07-01 20:00:00	11671
14	2018-07-01 21:00:00	12315
15	2018-07-01 22:00:00	12940
16	2018-07-01 23:00:00	13611
17	2018-07-02 00:00:00	14176
18	2018-07-02 01:00:00	14577
19	2018-07-02 02:00:00	14699
20	2018-07-02 03:00:00	14266
21	2018-07-02 04:00:00	14059
22	2018-07-02 05:00:00	13609
23	2018-07-02 06:00:00	12591
24	2018-07-02 07:00:00	11611
***	*****	*****

Ingeniería de características (preparación de datos)

# A tibble: 32,801 x 22										
Fecha	valor (y)	hora	día de semana	més	año	covid	desfase1	desfase2	desfase3	desfase4
<date>	<dbl>	<int>	<ord>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1 2019-07-02	9869	11	Tue	7	2019	0	10149	10646	11244	12161
2 2019-07-02	9982	12	Tue	7	2019	0	9869	10149	10646	11244
3 2019-07-02	10412	13	Tue	7	2019	0	9982	9869	10149	10646
4 2019-07-02	10864	14	Tue	7	2019	0	10412	9982	9869	10149
5 2019-07-02	11351	15	Tue	7	2019	0	10864	10412	9982	9869
6 2019-07-02	11745	16	Tue	7	2019	0	11351	10864	10412	9982
7 2019-07-02	12207	17	Tue	7	2019	0	11745	11351	10864	10412
8 2019-07-02	12643	18	Tue	7	2019	0	12207	11745	11351	10864
9 2019-07-02	13189	19	Tue	7	2019	0	12643	12207	11745	11351
10 2019-07-02	13716	20	Tue	7	2019	0	13189	12643	12207	11745
11 2019-07-02	14398	21	Tue	7	2019	0	13716	13189	12643	12207
12 2019-07-02	15073	22	Tue	7	2019	0	14398	13716	13189	12643
13 2019-07-02	15594	23	Tue	7	2019	0	15073	14398	13716	13189
14 2019-07-03	15931	0	wed	7	2019	0	15594	15073	14398	13716
15 2019-07-03	16037	1	wed	7	2019	0	15931	15594	15073	14398
16 2019-07-03	15878	2	wed	7	2019	0	16037	15931	15594	15073
17 2019-07-03	15363	3	wed	7	2019	0	15878	16037	15931	15594
18 2019-07-03	15010	4	wed	7	2019	0	15363	15878	16037	15931
19 2019-07-03	14466	5	wed	7	2019	0	15010	15363	15878	16037

Sorprendentemente, los datos multivariantes preparados para el aprendizaje automático (ML), obtenidos a partir de una serie temporal, nos permiten completar el ciclo PPDAC. Existen muchas preguntas posibles por las que empezar.



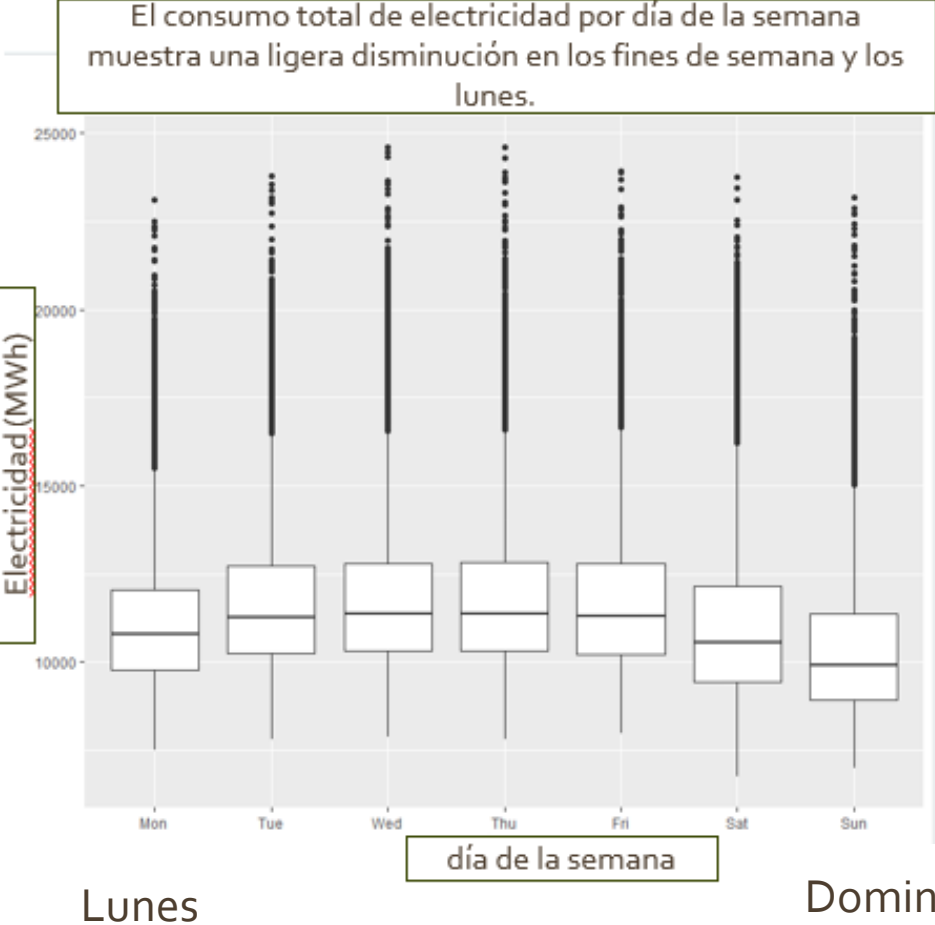
Características (variables)

A tibble: 32,801 × 22

Fecha	valor (y)	hora	día de semana	mes	año	covid	desfase1	desfase2	desfase3	desfase4
<date>	<dbl>	<int>	<ord>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1 2019-07-02	9869	11	Tue	7	2019	0	10149	10646	11244	12161
2 2019-07-02	9982	12	Tue	7	2019	0	9869	10149	10646	11244
3 2019-07-02	10412	13	Tue	7	2019	0	9982	9869	10149	10646
4 2019-07-02	10864	14	Tue	7	2019	0	10412	9982	9869	10149
5 2019-07-02	11351	15	Tue	7	2019	0	10864	10412	9982	9869
6 2019-07-02	11745	16	Tue	7	2019	0	11351	10864	10412	9982
7 2019-07-02	12207	17	Tue	7	2019	0	11745	11351	10864	10412
8 2019-07-02	12643	18	Tue	7	2019	0	12207	11745	11351	10864
9 2019-07-02	13189	19	Tue	7	2019	0	12643	12207	11745	11351
10 2019-07-02	13716	20	Tue	7	2019	0	13189	12643	12207	11745
11 2019-07-02	14398	21	Tue	7	2019	0	13716	13189	12643	12207
12 2019-07-02	15073	22	Tue	7	2019	0	14398	13716	13189	12643
13 2019-07-02	15594	23	Tue	7	2019	0	15073	14398	13716	13189
14 2019-07-03	15931	0	Wed	7	2019	0	15594	15073	14398	13716
15 2019-07-03	16037	1	Wed	7	2019	0	15931	15594	15073	14398
16 2019-07-03	15878	2	Wed	7	2019	0	16037	15931	15594	15073
17 2019-07-03	15363	3	Wed	7	2019	0	15878	16037	15931	15594
18 2019-07-03	15010	4	Wed	7	2019	0	15363	15878	16037	15931
19 2019-07-03	14466	5	Wed	7	2019	0	15010	15363	15878	16037

Pregunta: ¿Influye la hora del día en la demanda de electricidad? Puede realizar este diagrama de caja estacional con estudiantes de cursos introductorios de estadística utilizando el conjunto de datos destacado y empleando las variables valor (y) y hora.

Cuestionamiento a través del análisis. ¿Es importante el día de la semana?



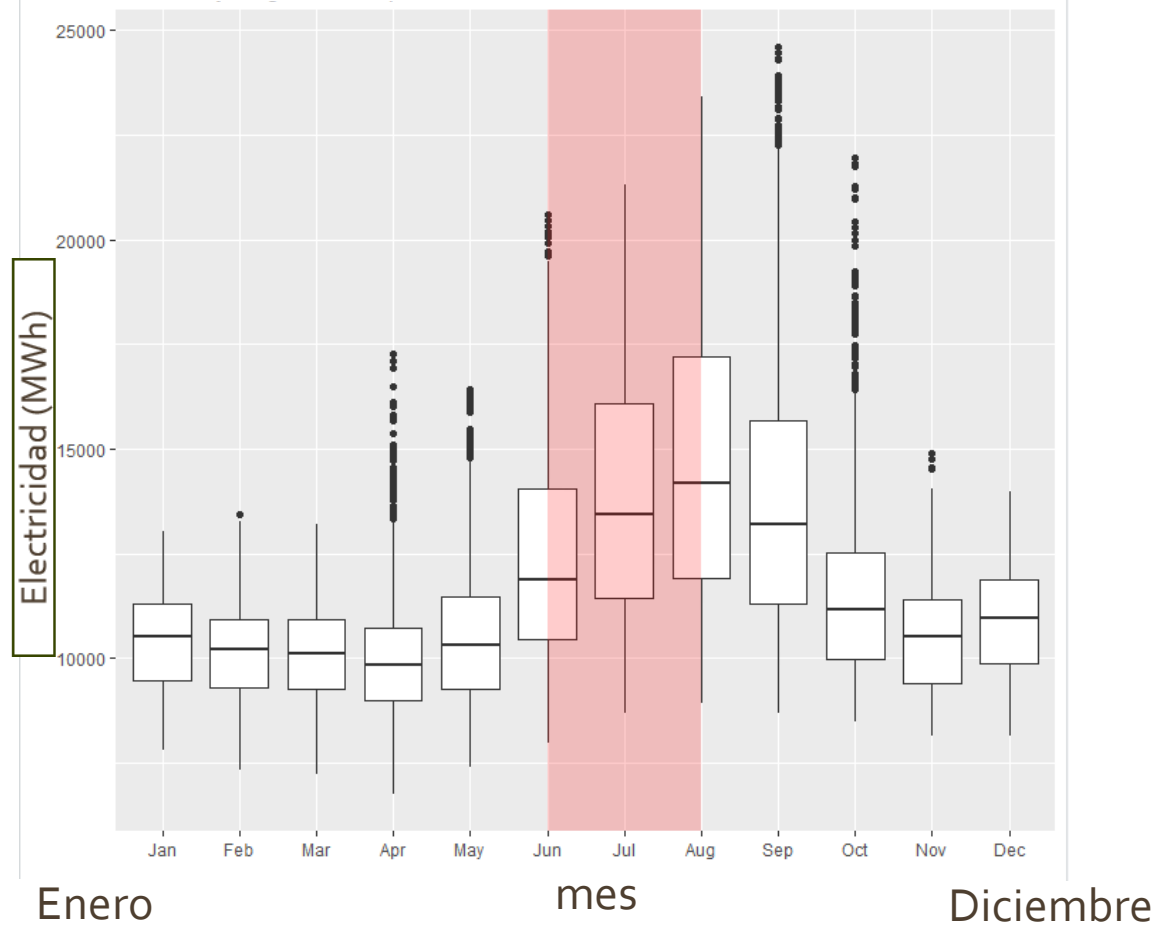
Características (variables)

A tibble: 32,801 x 22

	Fecha	valor (y)	hora	día de semana	mes	año	covid	desfase1	desfase2	desfase3	desfase4
	<date>	<dbl>	<int>	<ord>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	2019-07-02	9869	11	Tue	7	2019	0	10149	10646	11244	12161
2	2019-07-02	9982	12	Tue	7	2019	0	9869	10149	10646	11244
3	2019-07-02	10412	13	Tue	7	2019	0	2982	9869	10149	10646
4	2019-07-02	10864	14	Tue	7	2019	0	10412	9982	9869	10149
5	2019-07-02	11351	15	Tue	7	2019	0	10864	10412	9982	9869
6	2019-07-02	11745	16	Tue	7	2019	0	11351	10864	10412	9982
7	2019-07-02	12207	17	Tue	7	2019	0	11745	11351	10864	10412
8	2019-07-02	12643	18	Tue	7	2019	0	12207	11745	11351	10864
9	2019-07-02	13189	19	Tue	7	2019	0	12643	12207	11745	11351
10	2019-07-02	13716	20	Tue	7	2019	0	13189	12643	12207	11745
11	2019-07-02	14398	21	Tue	7	2019	0	13716	13189	12643	12207
12	2019-07-02	15073	22	Tue	7	2019	0	14398	13716	13189	12643
13	2019-07-02	15594	23	Tue	7	2019	0	15073	14398	13716	13189
14	2019-07-03	15931	0	wed	7	2019	0	15594	15073	14398	13716
15	2019-07-03	16037	1	wed	7	2019	0	15931	15594	15073	14398
16	2019-07-03	15878	2	wed	7	2019	0	16037	15931	15594	15073
17	2019-07-03	15363	3	wed	7	2019	0	15878	16037	15931	15594
18	2019-07-03	15010	4	wed	7	2019	0	15363	15878	16037	15931
19	2019-07-03	14466	5	wed	7	2019	0	15010	15363	15878	16037

¿Tienen algunos meses más demanda que otros?

Electricidad mensual agregada: constante de enero a abril, aumenta de mayo a agosto y luego disminuye hasta diciembre.



- Demanda horaria de electricidad suministrada por Southern California Edison: del 1 de julio de 2018 (8:00 a. m.) al 31 de marzo de 2023 (12:00 a. m.)

Características (variables)

```
# A tibble: 32,801 x 22
  Fecha      valor (y) hora día de semana mes año covid desfase1 desfase2* desfase3* desfase4
  <date>      <dbl> <int> <ord>      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 2019-07-02  9869    11 Tue       7 2019    0 10149 10646 11244 12161
2 2019-07-02  9982    12 Tue       7 2019    0 9869 10149 10646 11244
3 2019-07-02 10412    13 Tue       7 2019    0 9982 9869 10149 10646
4 2019-07-02 10864    14 Tue       7 2019    0 10412 9982 9869 10149
5 2019-07-02 11351    15 Tue       7 2019    0 10864 10412 9982 9869
6 2019-07-02 11745    16 Tue       7 2019    0 11351 10864 10412 9982
7 2019-07-02 12207    17 Tue       7 2019    0 11745 11351 10864 10412
8 2019-07-02 12643    18 Tue       7 2019    0 12207 11745 11351 10864
9 2019-07-02 13189    19 Tue       7 2019    0 12643 12207 11745 11351
10 2019-07-02 13716    20 Tue       7 2019    0 13189 12643 12207 11745
11 2019-07-02 14398    21 Tue       7 2019    0 13716 13189 12643 12207
12 2019-07-02 15073    22 Tue       7 2019    0 14398 13716 13189 12643
13 2019-07-02 15594    23 Tue       7 2019    0 15073 14398 13716 13189
14 2019-07-03 15931     0 wed       7 2019    0 15594 15073 14398 13716
15 2019-07-03 16037     1 wed       7 2019    0 15931 15594 15073 14398
16 2019-07-03 16037     2 wed       7 2019    0 16037 15931 15594 15073
17 2019-07-03 15363     3 wed       7 2019    0 15878 16037 15931 15594
18 2019-07-03 15010     4 wed       7 2019    0 15363 15878 16037 15931
19 2019-07-03 14466     5 wed       7 2019    0 15010 15363 15878 16037
```

Noten que el aire acondicionado funciona con electricidad en California. ¿Cómo sería este diagrama de caja comparativo para Argentina o el hemisferio sur en general?

Otras preguntas posibles para practicar

- (1) ¿Se ve afectada la demanda en la hora t por la demanda en la hora $t-1$ (desfase1) o $t-2$ (desfase2)?
- (2) Si ajustamos una regresión de la demanda de electricidad en función de las otras variables, ¿cuáles serían significativas? (Pregunta de estadística). En el ML, CD, se pregunta ¿Qué variable nos ayuda a predecir mejor?
- (3) ¿Qué variable es la más importante para predecir la demanda horaria de electricidad? (Pregunta de ML). Es difícil de responder con una regresión múltiple, pero más fácil con un árbol de regresión. Una buena excusa para hablar sobre árboles de regresión. Observe el énfasis en la predicción que es el objetivo de ML, CD.



Fuente: Uber movement (<https://movement.uber.com>)

year	month	day	hour	osm_way_id	osm_start_node_id	osm_end_node_id	speed_mph_mean	speed_mph_stddev
2020	1	1	1	40722998	62385707	4927951349	26.636	4.483
2020	1	31	21	40722998	62385707	4927951349	25.513	4.276
2020	1	1	0	40722998	62385707	4927951349	27.521	5.105
2020	1	1	0	40722998	5780849015	4927951349	26.05	3.803
2020	1	1	1	40722998	5780849015	4927951349	25.459	3.585
2020	1	30	8	417094233	4714793573	1014244233	27.761	3.679
2020	1	7	15	416137931	239464357	4318478540	25.721	1.649
2020	1	30	18	416137931	239464357	4318478540	25.222	7.128
2020	1	4	11	416137931	239464357	4318478540	23.629	3.669
2020	1	17	17	416137931	239464357	4318478540	22.642	3.554
2020	1	22	17	416137931	239464357	4318478540	23.842	4.381
2020	1	9	17	416137931	239464357	4318478540	29.338	14.674
2020	1	29	10	416137931	239464357	4318478540	23.056	3.197
2020	1	17	15	416137931	239464357	4318478540	27.031	5.015
2020	1	5	18	416137931	239464357	4318478540	23.461	3.422
2020	1	30	19	416137931	239464357	4318478540	23.45	1.53
2020	1	25	14	416137931	239464357	4318478540	26.481	2.493
2020	1	27	14	416137931	239464357	4318478540	26.054	3.478
2020	1	27	17	416137931	239464357	4318478540	32.316	18.225
2020	1	14	17	416137931	239464357	4318478540	28.82	7.825

Para una mayor evaluación formativa, seguir la discusión y practicar después del entrenamiento descrito antes, los alumnos pueden usar datos de Uber anonimizados para ayudar a la planificación urbana en ejercicios o proyectos.

Para profundizar en el tema: ¿cómo se construiría un árbol de regresión si utilizáramos únicamente la regresión como algoritmo en cada paso? ¿Cómo lo describirías utilizando papel y lápiz?

Uber ya publica sus datos en un formato de ciencia de datos actual, listo para su uso en modelos de aprendizaje automático (ML). Los alumnos no necesitan hacer lo que hicimos con la serie temporal de la electricidad.

O bien, en lugar de Uber, Kaggle o repositorios, participar en proyectos de ciencia ciudadana (citizen science), utilizar Kaggle o recurrir a los numerosos repositorios de grandes volúmenes de datos.

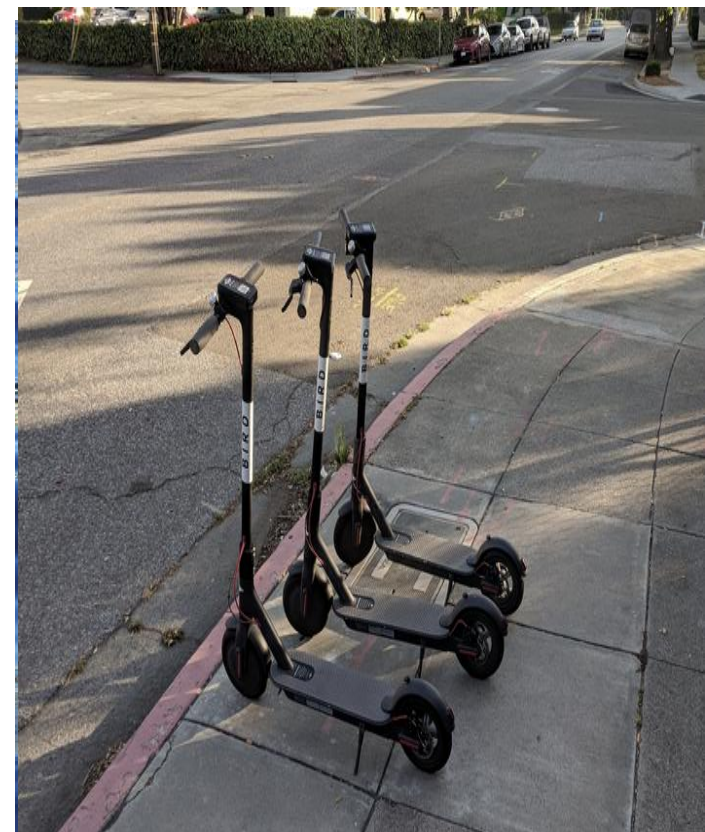


Ejemplo 3. Contexto: Introducción a la probabilidad-Modelo Poisson

Micromovilidad a pequeña escala. ¿Dónde hay mayor demanda u oferta de patinetes eléctricos?

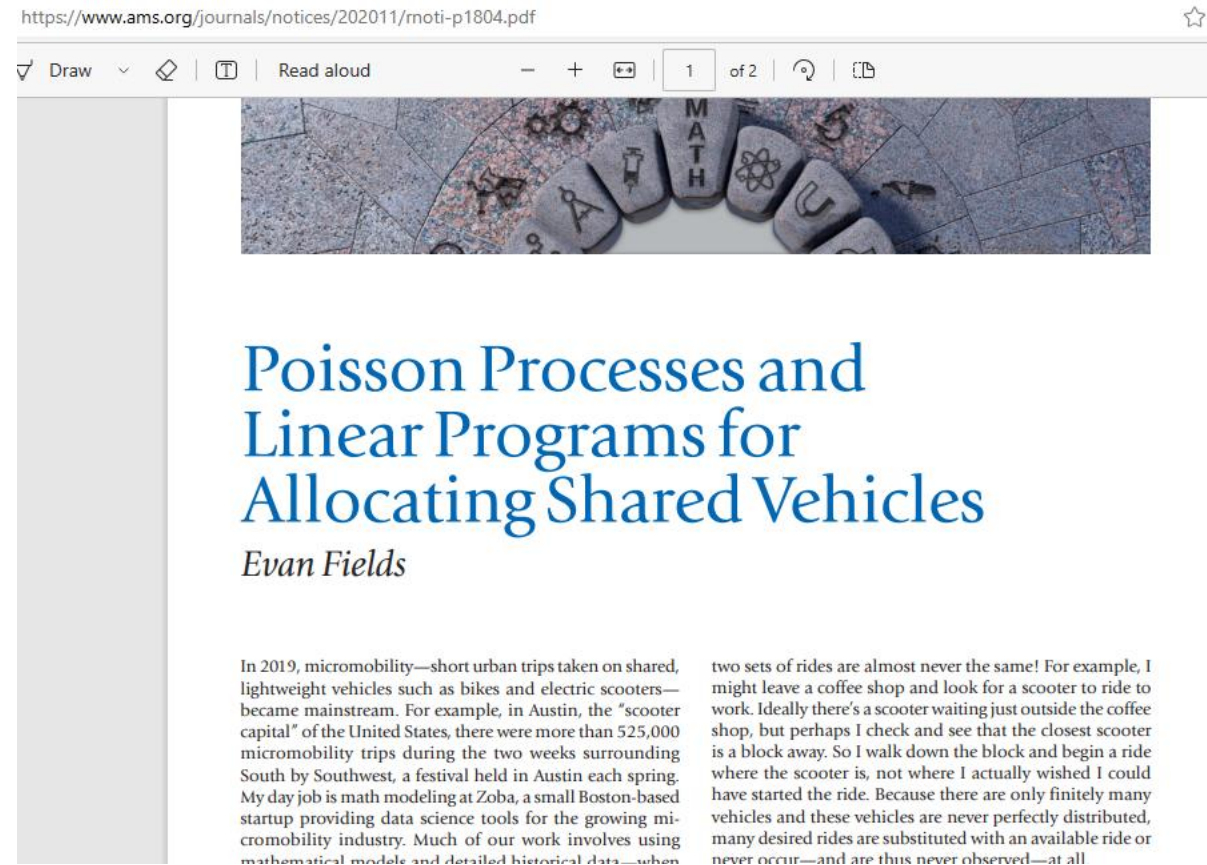
Contexto para el profesional de la ciencia de datos: ciudades inteligentes. Modelado predictivo que incorpora la incertidumbre.

Contexto estadístico: el ciclo PPDAC completo. Preparación de datos (*data wrangling*). Ajuste de modelos.



Los estudiantes leen sobre los patinetes y la micromovilidad.

Leí por casualidad un artículo en una revista de matemáticas (Notices of the AMS) sobre «procesos de Poisson y programación lineal para la asignación de vehículos compartidos» y pensé: ¿cómo podría convertir esto en un proyecto para mi clase de probabilidad? ¿Puede servir para cuando estudiemos la distribución de Poisson?



Entrenamiento de los alumnos: reciben datos temporales sobre nacimientos durante 24 horas en un hospital y aprenden el uso del modelo Poisson

Cronología de los nacimientos

1951 2 3402 1191
.....
2010 1 3500 1210
2037 2 3736 1237
2051 2 3370 1251
.....
2104 2 2121 1264
2123 2 3150 1283
.....
2217 1 3866 1337
.....
2327 1 3542 1407
2355 1 3278 1435
.....
etc.

4 nacimientos en la vigésima hora

3 nacimientos en la hora 21

2 nacimientos en la hora 22

1 nacimientos en la hora 23

2 nacimientos en la hora 24

¿En cuántas horas del día nacieron tantos niños por hora?

Nacimientos por hora

Probabilidad empírica (frecuencia relativa observada)

Probabilidad teórica (utilizando el modelo de Poisson con parámetro $\lambda = 44/24 = 1,83$)

Number of Births per hour	Tally (in how many of the hours did we observe the number of births in column 1) (Observed)	Empirical Probability (this is the observed relative frequency)	Theoretical Probability (with Poisson model with $\lambda=44/24=1.83$ births per hour)
0	3	$3/24 = 0.125$	$\frac{1.83^0 e^{-1.83}}{0!} = 0.160$
1	8	$8/24 = 0.333$	$\frac{1.83^1 e^{-1.83}}{1!} = 0.293$
2	6	0.250	0.269
3	4	0.167	0.164
4	3	0.125	0.075
5+	0	0.000	0.039
Total	24 hours	1	1

Source: Sanchez, J. 2020.

Nacimientos por hora

¿En cuántas horas del día nacieron tantos niños por hora?

Probabilidad empírica (frecuencia relativa observada)

Probabilidad teórica (utilizando el modelo de Poisson con parámetro $\lambda = 44/24 = 1.83$)

Number of Births per hour	Tally (in how many of the hours did we observe the number of births in column 1) (Observed)	Empirical Probability (this is the observed relative frequency)	Theoretical Probability (with Poisson model with $\lambda = 44/24 = 1.83$ births per hour)
0	3	$3/24 = 0.125$	$\frac{1.83^0 e^{-1.83}}{0!} = 0.160$
1	8	$8/24 = 0.333$	$\frac{1.83^1 e^{-1.83}}{1!} = 0.293$
2	6	0.250	0.269
3	4	0.167	0.164
4	3	0.125	0.075
5+	0	0.000	0.039
Total	24 hours	1	1

$$\text{Sum of } \frac{(O - E)^2}{E} = 0.18375 + \dots + 0.9360 = 2.08625$$

The Chi-square statistic equals 2.08625.

Looking at the app,

$P(\text{"Chi-square with 5 degrees of freedom"} > 2.08625) = 0.83709$

El estadístico de la prueba de chi-cuadrado de bondad de ajuste nos indica que el modelo de Poisson con un valor esperado de 1.83 es un buen modelo para estos datos de nacimientos.

Con este ejemplo, los alumnos no solo usan el modelo de Poisson para ajuste de modelos, sino que también nota que la inferencia usa modelos de probabilidad.

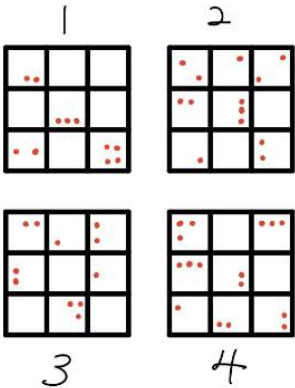
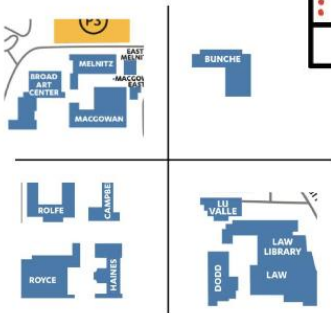
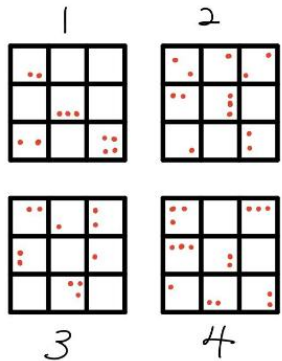
Source: Sanchez, J. 2020.

Con esos conocimientos previos, los estudiantes hacen un proyecto y salen al terreno (el campus de la UCLA), crean una cuadrícula del campus y cuentan y describen la cantidad de patinetes eléctricos observados en cada celda de dicha cuadrícula. Cada grupo hace su análisis. Luego se junta los datos de todos los grupos para mejor “fit”

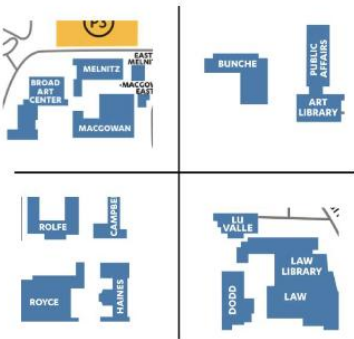
Ejemplo: datos de un grupo



Campus map



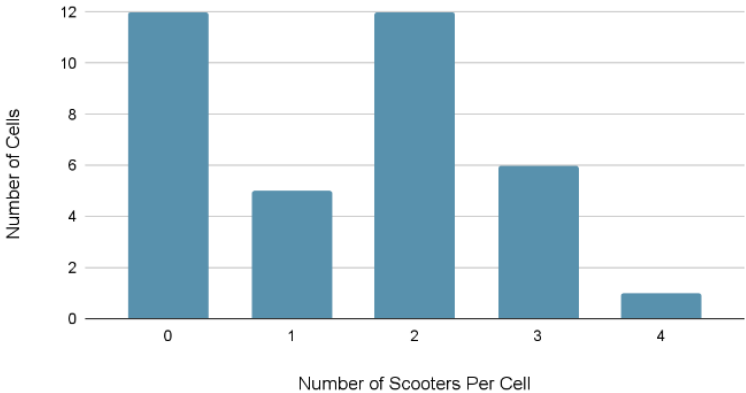
Campus map



Recuento y resumen del mismo grupo

Number of Scooters Per Cell	Number of Cells With That Number of Scooters
0	12
1	5
2	12
3	6
4	1

Number of Scooters Per Cell vs. Number of Cells



Los estudiantes en ese grupo ajustan un modelo de probabilidad estimada (algunos a mano, otros con computadoras).

Calculate what is needed

Estimate of $\lambda = \frac{0 \times 12 + 1 \times 5 + 2 \times 12 + 3 \times 6 + 4 \times 1}{36} = 1.42$

$P(X=x) = \frac{1.42^x e^{-1.42}}{x!}$

Theoretical Probabilities:

$$\frac{1.42^0 e^{-1.42}}{0!} = 0.24 \quad \frac{1.42^2 e^{-1.42}}{2!} = 0.24 \quad \frac{1.42^4 e^{-1.42}}{4!} = 0.04$$

$$\frac{1.42^1 e^{-1.42}}{1!} = 0.34 \quad \frac{1.42^3 e^{-1.42}}{3!} = 0.115$$

Predicted # Per Cell:

$$0.24 \times 36 = 8.64 \quad 0.24 \times 36 = 8.64 \quad 0.04 \times 36 = 1.44$$

$$0.34 \times 36 = 12.24 \quad 0.115 \times 36 = 4.14$$

Realize that probability is also used to draw inferences

# of scooters per cell (X)	Observed (O)	P(X=x)	# of scooters Predicted (E)	$\frac{(E-O)^2}{E}$
0	12	0.24	8.64	1.3
1	5	0.34	12.24	4.28
2	12	0.24	8.64	1.3
3	6	0.115	4.14	0.83
4	1	0.04	1.44	0.13
		≈ 1	≈ 36	7.84

$\chi^2 = \text{Chi square statistic} = 7.84$
 $5 - 1 = 4 \text{ degrees of freedom}$
 $P(\chi^2_4 > 7.84) = 0.097$

Poisson Distribution Formula

$$P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where
 $x = 0, 1, 2, 3, \dots$
 λ = mean number of occurrences in the interval
 e = Euler's constant ≈ 2.71828

Because the p-square statistic is larger than 0.05, we can conclude the Poisson model with $\lambda = 1.42$ is a good fit to the data.

Source: students' paper.

Los estudiantes critican el enfoque y sugieren:

Más variables ayudarían a predecir mejor.

La recopilación de datos no se realizó el mismo día ni a la misma hora. Esto podría sesgar las conclusiones.

Se requieren más datos y una mejor cobertura de las áreas del campus en el muestreo.

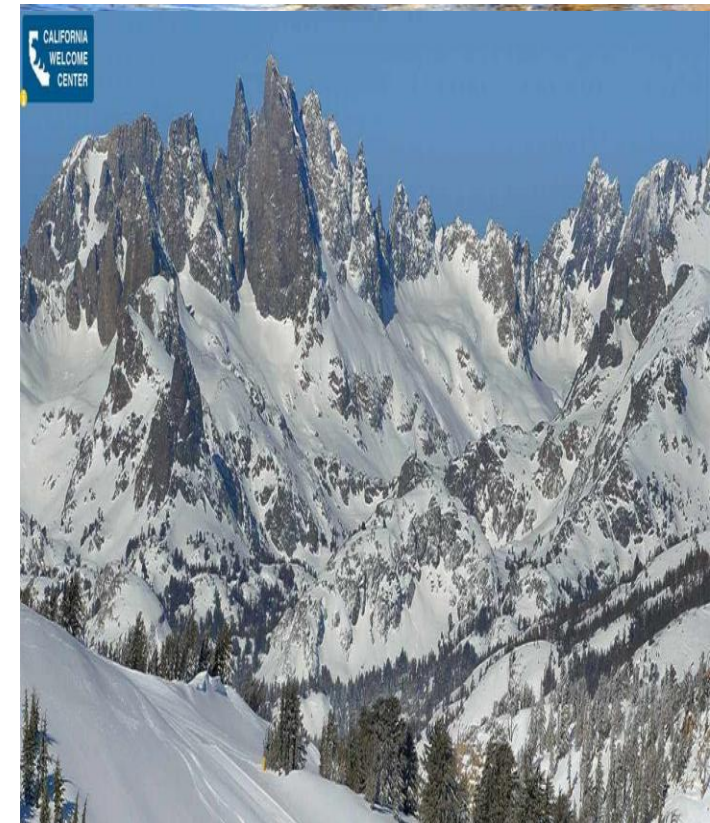


Example 4 – Introducción a la estadística. Caracterización de series temporales de caudal fluvial mediante estadísticos resumen.

¿Qué tienen en común los ríos de California?

Contexto para el profesional de la ciencia de datos: Ingeniería de características, aprendizaje automático no supervisado. Modelos discriminantes.

Contexto estadístico: Preparación de datos, el ciclo completo de PPADC.



Datos de series temporales convertidos en datos de características resumidas —características simples— para aprendizaje automático no supervisado.

Datos anuales brutos sobre el caudal fluvial medio anual.

Basin	ID	1930	1931	1932	1933	1934	1935	1936	1937	1938	1939	1940	1941
Trinity	TNL	319.6	201.72	425.5	578.9	264.22	559.31	535.51	800.01	1087.4	256.21	620.28	1400.3
Sac/Delta	SDT	170.1	69.2	170.9	245.1	123.4	308.7	219.2	398.5	499.3	101.7	272.4	636.1
McCloud	MSS	284.1	193.9	285.4	342.8	242.4	459.98	318.68	470.82	744.07	279.12	481.56	748.03
Pit	PSH	703.1	480.1	805.1	759.7	535.3	1374.3	766.4	969	1615.6	592.4	968.2	1202.6
Shasta	SIS	1143	764	1312	1374	926	2275	1343	1969	3060	996	1849	2791
Bend	SBB	1644	943	1689	1770	1186	3336	1854	2600	4062	1267	2660	4235
Feather	FTO	1426	502.1	1742	1142	594.2	2892	1648	1940	4321	748.5	1833	2569
Yuba	YRS	752.8	279.8	1226.1	775.1	310.5	1547.2	1240.6	1220.2	2075.2	450.21	1056.32	1434.55
American	AMF	829.2	363.9	1579.8	977.2	361.7	1915	1663.8	1476.8	2475.1	572.7	1378.5	1531
Cosumnes	CSN	61.97	12.3	114.23	79.33	17.85	257.16	149.88	177.51	276.75	35.25	130.68	156.39
Mokelumne	MKM	463.32	219.65	101.59	1038.00	3	13						

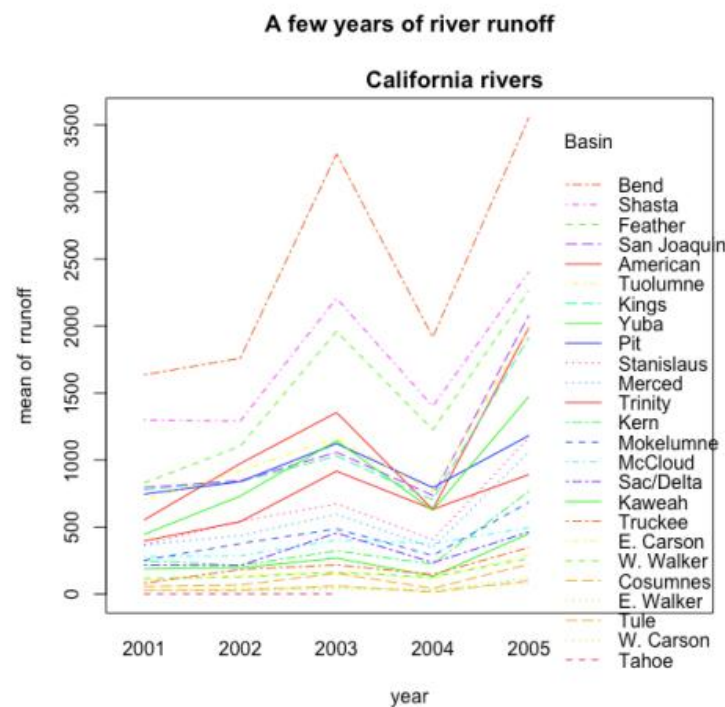
Ingeniería de
características
sencilla
(preparación de
datos)

Las características resumen conforman un conjunto de datos multivariante de características resumen. Son adecuadas para el aprendizaje no supervisado con ML.

Cuenca	Media	de	Min	Max	Cuantos +2de	+1de
Trinity	641.12	294.28.	116.62.	1593.35	4	10
Sac/Delta	292.66	141.48	63.42	711.20	4	8
McCloud	397.87	132.07	184.67	748.03	3	9
Pit	1037.95	332.54	480.10	2097.72	3	8
Shasta	1795.64	660.68	764.00	3525.31	3	9
Bend	2453.62	989.65	943.00	5075.46	2	7
Feather	1822.91	962.31	391.85	4676.00	4	11
Yuba	1036.31	501.31	199.88	2424.09	4	12
American	1275.89	658.01	228.96	2912.26	3	13
Cosumnes	127.56	93.13	7.96	362.84	6	10
Mokelumne	463.32	219.65	101.59	1038.00	3	13
Stanislaus	710.88	351.49	115.51	1636.18	3	12
Tuolumne	1210.10	559.48	301.02	2645.28	3	13
Merced	623.45	332.30	123.29	1587.46	3	11
San Joaquin	1233.34	641.00	261.91	2898.00	4	10
Kings	1203.78	633.85	274.49	3112.61	2	13
Kaweah	283.91	170.07	61.72	799.70	3	10
Tule	63.26	56.19	2.36	259.14	3	3
Kern	452.69	328.28	84.39	1657.07	2	6
Truckee	261.81	147.88	52.42	712.73	3	12
Tahoe	1.39	0.82	0.17	3.57	4	15
W. Carson	54.12	26.40	12.06	135.21	3	12
E. Carson	184.80	88.67	42.57	406.72	5	12
W. Walker	149.87	63.97	34.79	303.33	3	13
E. Walker	62.13	45.11	6.66	209.04	3	7

Sanchez, J. (2023), Chapter 1.

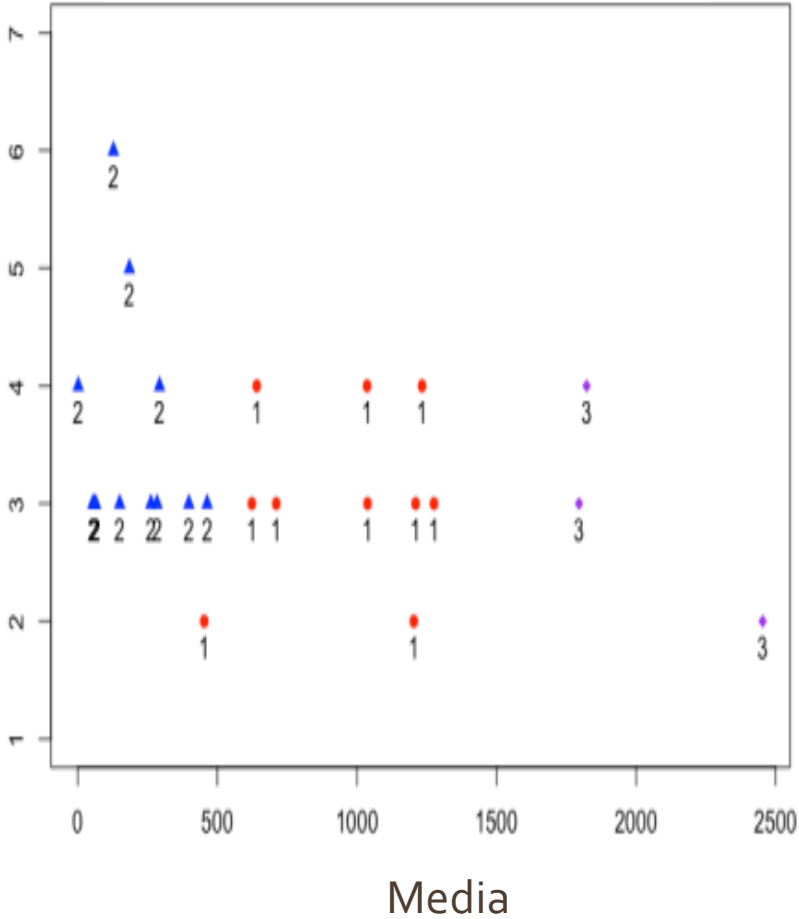
Aprendizaje automático no supervisado.



Ingeniería de características sencilla (preparación de datos)

Agrupamiento de ríos mediante el algoritmo K-means utilizando un resumen de los datos basado en ingeniería de características.

¿Cuántas veces el caudal del río superó en dos desviaciones estándar el caudal medio?



Los estudiantes reflexionan sobre características más significativas para incluir en los datos y examinan software de generación automatizada (en R, por ejemplo, tsfeatures, feasts, timetk) de características (en R, por ejemplo, tsfeatures, feasts, timetk) .Y especulan:

¿Quizás el número de puntos de inflexión debería ser una característica

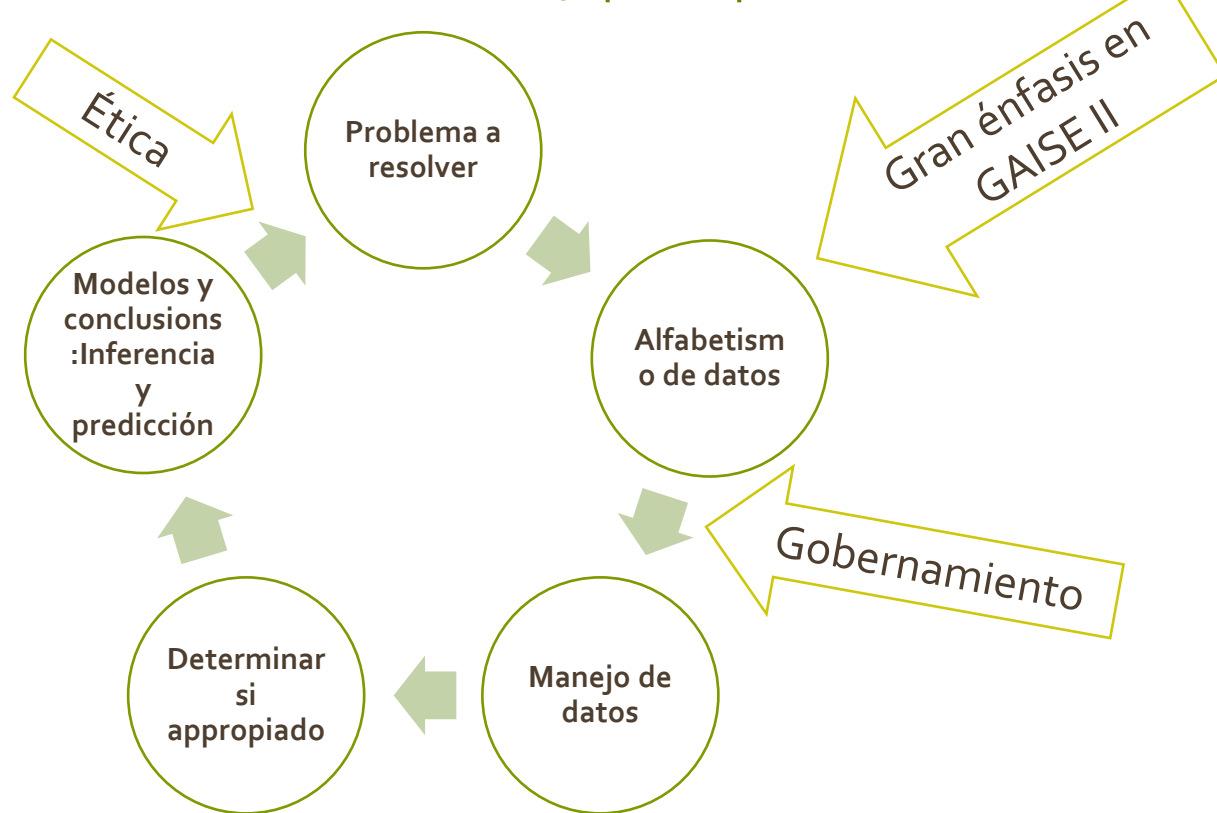
¿Quizás se deberían incluir las precipitaciones y la temperatura media de la región del río?

¿Son aplicables a los datos fluviales todas las características generadas por los programas de generación automatizada de características? No deberíamos utilizar características aplicables a datos financieros para datos de caudal fluvial, ¿verdad? Analice esta cuestión.

Conclusiones

- En todos los ejemplos mencionados se han aplicado una o varias etapas del ciclo de ciencia de datos (o, de forma equivalente, del ciclo PPDAC), adaptadas al nivel de competencia y al momento del aprendizaje de los estudiantes.
- Los ejemplos abarcan una variedad de conjuntos de datos, incluidos algunos de gran tamaño. En algunos casos, presentamos los mismos datos de formas muy distintas según los objetivos planteados. La mayoría de las aplicaciones de aprendizaje automático (ML) consisten en transformar diversos tipos de datos al formato habitual de tabla rectangular de observaciones y variables —proceso conocido como ingeniería de características (*feature engineering*)— para prepararlos para su uso en redes neuronales y modelos de ML. Se pone especial énfasis en la alfabetización en datos, porque el origen de y como se tratan los datos afectan las decisiones que los algoritmos toman. La estadística tradicional no tiene en cuenta la incertidumbre generada por este proceso.
- No obstante, todas las actividades incorporan conceptos estadísticos introductorios propios de los planes de estudio tradicionales de estadística, probabilidad o series temporales. Los estudiantes abordan simultáneamente dichos contenidos curriculares y los de aprendizaje automático; por ello, resulta fundamental hacer hincapié en la terminología para que sean conscientes de esta relación.
- Al mismo tiempo que se hacen las actividades, los alumnos notan que la historia que se desprende de los datos será tan buena como lo sean el origen y el tratamiento de dichos datos. Los cursos en los que se trabajan estos ejemplos generan relatos a través de foros de debate y discusiones en clase; pero también suscitan críticas cuando estas son pertinentes (no solo respecto a los métodos estadísticos, sino también a los mecanismos de generación de datos, algo que no suele abordarse habitualmente en nuestras clases).

Esta charla ha presentado ejemplos y actividades derivadas que he utilizado en las clases de introducción a la probabilidad y a la estadística para ir más allá del modelo PPDAC —que empleo habitualmente- para resaltar ante los estudiantes el papel de los datos y la predicción en el DSLC (Ciclo de Vida de la Ciencia de Datos) y consecuentemente en la IA, que depende del DSLC. Y sobre todo para resaltar similitudes entre los dos.



Keller, S.A, et al. (2020): Doing **Data Science**: A Framework and Case Study.



GAISE I, GAISE II, Census@School, ISLP, OECD, y en numerosos escenarios de todo el mundo, así como en libros de introducción a la estadística desde hace ya muchos años.

Bibliografía

1. Bryce, G. R. (2002). Undergraduate Statistics Education: An Introduction and Review of Selected Literature. *Journal of Statistics Education*, 10(2). <https://doi.org/10.1080/10691898.2002.11910664>
2. Bonikowska, A. et al. (2019). Data Literacy: What it is and How to Measure it in the Public Service. https://publications.gc.ca/collections/collection_2019/statcan/11-633-x/11-633-x2019003-eng.pdf
3. Fields, E. (2020). *Poisson Processes and Linear Programs for Allocating Shared Vehicles*. Notices of the American Mathematical Association. Vol 67, No. 11, page 1804-1805.
4. Han, Y., Kydd, M. K., Ward, J., & Tewari, A. (2026). Teaching Machine Olfaction in an Undergraduate Deep Learning Course: A Pilot Integrating Chemistry, Machine Learning, and Sensory Evaluation. *Journal of Statistics and Data Science Education*, 34(3), 372–380. <https://doi.org/10.1080/26939169.2026.2665103>
5. Keller, S.A., Shipp, S.S., Schroeder, A.D. and Korkmaz, G. (2020). *Doing Data Science: A Framework and Case Study*. HDSR, Issue 2.1. Winter 2020. <https://doi.org/10.1162/99608f92.2d83f7f5>
6. Piech, C. *Course Reader for CS 109*. Stanford University. Ongoing. <https://chrispiech.github.io/probabilityForComputerScientists/en/examples/fairness/>
7. Sanchez, J. (2023) *Time Series for Data Scientists*. Cambridge University Press.
8. Sanchez, J. (2023) timeseriestime.org A companion to Sanchez, J. Time Series Time for Data Scientists. <https://timeseriestime.org/>
9. Sanchez, J (2020) *Probability for Data Scientists*. Cognella.
10. Wolff, A., D. Gooch, J.J. Caverio Montaner, U. Rashid, and G. Kortuem. 2016. “Creating an understanding of data literacy for a data-driven society.” *The Journal of Community Informatics* 12 (3): 9–26

Gracias