

Introducción a Machine Learning

Por Sebastian Waisbrot

Antes de ML: Buscar en una lista ordenada

Dada una lista ordenada encontrar si un item si está presente.

Antes de ML: Linear Search

Águila	Cisne	Ganso	Loro	Petirrojo
Ardilla	Codorniz	Gato	Oveja	Pez
Avestruz	Conejo	Gaviota	Pájaro carpintero	Pingüino
Búho	Cuervo	Golondrina	Paloma	Rata
Buitre	Emú	Halcón	Pato	Ratón
Burro	Faisán	Hámster	Pavo	Serpiente
Caballo	Flamenco	Hurón	Pavo real	Tarántula
Canario	Gallina	Iguana	Pelícano	Toro
Cerdo	Gallo	Kiwi	Perico	Tortuga

Antes de ML: Linear Search

¿Esa lista incluye perro?

Antes de ML: Linear Search

Águila	Cisne	Ganso	Loro	Petirrojo
Ardilla	Codorniz	Gato	Oveja	Pez
Avestruz	Conejo	Gaviota	Pájaro carpintero	Pingüino
Búho	Cuervo	Golondrina	Paloma	Rata
Buitre	Emú	Halcón	Pato	Ratón
Burro	Faisán	Hámster	Pavo	Serpiente
Caballo	Flamenco	Hurón	Pavo real	Tarántula
Canario	Gallina	Iguana	Pelícano	Toro
Cerdo	Gallo	Kiwi	Perico	Tortuga

Antes de ML: Linear Search

Águila	Cisne	Ganso	Loro	Petirrojo
Ardilla	Codorniz	Gato	Oveja	Pez
Avestruz	Conejo	Gaviota	Pájaro carpintero	Pingüino
Búho	Cuervo	Golondrina	Paloma	Rata
Buitre	Emú	Halcón	Pato	Ratón
Burro	Faisán	Hámster	Pavo	Serpiente
Caballo	Flamenco	Hurón	Pavo real	Tarántula
Canario	Gallina	Iguana	Pelícano	Toro
Cerdo	Gallo	Kiwi	Perico	Tortuga

Antes de ML: Linear Search

(43 slides después)

Antes de ML: Linear Search

Águila	Cisne	Gancho	Loro	Petirrojo
Ardilla	Codorniz	Gato	Oveja	Pez
Avestruz	Conejo	Gaviota	Pájaro carpintero	Pingüino
Búho	Cuervo	Golondrina	Paloma	Rata
Buitre	Emú	Halcón	Pato	Ratón
Burro	Faisán	Hámster	Pavo	Serpiente
Caballo	Flamenco	Hurón	Pavo real	Tarántula
Canario	Gallina	Iguana	Pelícano	Toro
Cerdo	Gallo	Kiwi	Perico	Tortuga

Antes de ML: Linear Search

Águila	Cisne	Gancho	Loro	Petirrojo
Ardilla	Codorniz	Gato	Oveja	Pez
Avestruz	Conejo	Gaviota	Pájaro carpintero	Pingüino
Búho	Cuervo	Golondrina	Paloma	Rata
Buitre	Emú	Halcón	Pato	Ratón
Burro	Faisán	Hámster	Pavo	Serpiente
Caballo	Flamenco	Hurón	Pavo real	Tarántula
Canario	Gallina	Iguana	Pelícano	Toro
Cerdo	Gallo	Kiwi	Perico	Tortuga

Antes de ML: Binary Search



Antes de ML: Binary Search

Águila	Cisne	Ganso	Loro	Petirrojo
Ardilla	Codorniz	Gato	Oveja	Pez
Avestruz	Conejo	Gaviota	Pájaro carpintero	Pingüino
Búho	Cuervo	Golondrina	Paloma	Rata
Buitre	Emú	Halcón	Pato	Ratón
Burro	Faisán	Hámster	Pavo	Serpiente
Caballo	Flamenco	Hurón	Pavo real	Tarántula
Canario	Gallina	Iguana	Pelícano	Toro
Cerdo	Gallo	Kiwi	Perico	Tortuga

Antes de ML: Binary Search

Águila	Cisne	Gancho	Loro	Petirrojo
Ardilla	Codorniz	Gato	Oveja	Pez
Avestruz	Conejo	Gaviota	Pájaro carpintero	Pingüino
Búho	Cuervo	Golondrina	Paloma	Rata
Buitre	Emú	Halcón	Pato	Ratón
Burro	Faisán	Hámster	Pavo	Serpiente
Caballo	Flamenco	Hurón	Pavo real	Tarántula
Canario	Gallina	Iguana	Pelícano	Toro
Cerdo	Gallo	Kiwi	Perico	Tortuga

Antes de ML: Binary Search

Águila	Cisne	Gancho	Loro	Petirrojo
Ardilla	Codorniz	Gato	Oveja	Pez
Avestruz	Conejo	Gaviota	Pájaro carpintero	Pingüino
Búho	Cuervo	Golondrina	Paloma	Rata
Buitre	Emú	Halcón	Pato	Ratón
Burro	Faisán	Hámster	Pavo	Serpiente
Caballo	Flamenco	Hurón	Pavo real	Tarántula
Canario	Gallina	Iguana	Pelícano	Toro
Cerdo	Gallo	Kiwi	Perico	Tortuga

Antes de ML: Binary Search

Águila	Cisne	Gancho	Loro	Petirrojo
Ardilla	Codorniz	Gato	Oveja	Pez
Avestruz	Conejo	Gaviota	Pájaro carpintero	Pingüino
Búho	Cuervo	Golondrina	Paloma	Rata
Buitre	Emú	Halcón	Pato	Ratón
Burro	Faisán	Hámster	Pavo	Serpiente
Caballo	Flamenco	Hurón	Pavo real	Tarántula
Canario	Gallina	Iguana	Pelícano	Toro
Cerdo	Gallo	Kiwi	Perico	Tortuga

Antes de ML: Binary Search

(4 slides después)

























Antes de ML: Binary Search

Águila	Cisne	Gancho	Loro	Petirrojo
Ardilla	Codorniz	Gato	Oveja	Pez
Avestruz	Conejo	Gaviota	Pájaro carpintero	Pingüino
Búho	Cuervo	Golondrina	Paloma	Rata
Buitre	Emú	Halcón	Pato	Ratón
Burro	Faisán	Hámster	Pavo	Serpiente
Caballo	Flamenco	Hurón	Pavo real	Tarántula
Canario	Gallina	Iguana	Pelícano	Toro
Cerdo	Gallo	Kiwi	Perico	Tortuga

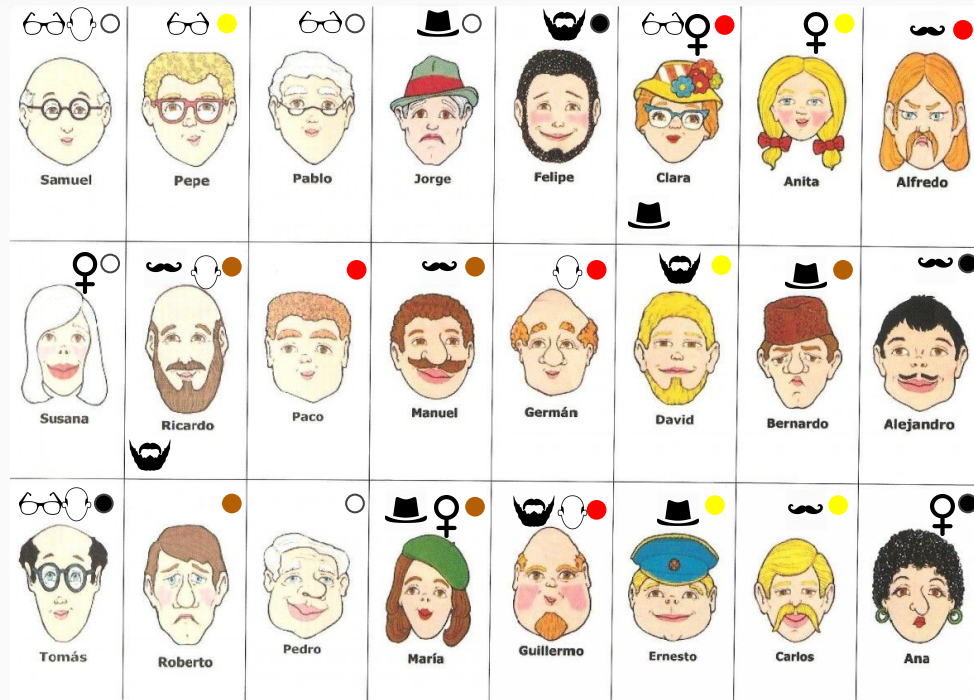
¿Quién es quién?



¿Quién es quién?

 Samuel	 Pepe	 Pablo	 Jorge	 Felipe	 Clara	 Anita	 Alfredo
 Susana	 Ricardo	 Paco	 Manuel	 Germán	 David	 Bernardo	 Alejandro
 Tomás	 Roberto	 Pedro	 María	 Guillermo	 Ernesto	 Carlos	 Ana

¿Quién es quién?



¿Quién es quién?

Mujeres (5)

Anteojos (5)

Pelados (5)

Rubios (5)

Canosos (5)

Pelirrojos (5)

Castaños (5)

Pelo negro (4)

Bigote (5)

Barba (4)

Sombrero (5)

¿Quién es quién?

Mujeres (5)

Anteojos (5)

Pelados (5)

Rubios (5)

Canosos (5)

Pelirrojos (5)

Castaños (5)

Pelo negro (4)

Bigote (5)

Barba (4)

Sombrero (5)

¿Quién es quién?

Mujeres (5) →



Anteojos (1)

Peladas (0)

Rubias (1)

Canosas (1)

Pelirrojas (1)

Castañas (1)

Pelo negro (1)

Bigote (0)

Barba (0)

Sombrero (2)

¿Quién es quién?

Mujeres (5) —————> Sombrero (2) —————>



Anteojos (1)

Pelirrojas (1)

Castañas (1)

¿Quién es quién?

Mujeres (5) → ~Sombrero (3) →



Rubias (1)

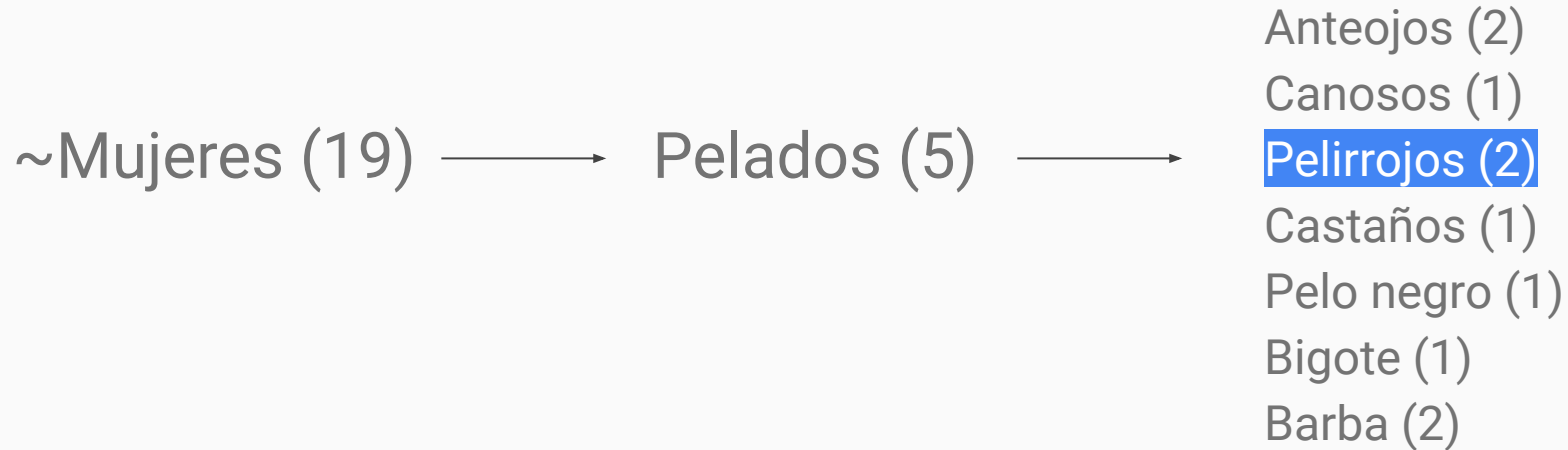
Canosos (1)

Pelo negro (1)

¿Quién es quién?

~Mujeres (19) →	Anteojos (4)	Castaños (4)
	Pelados (5)	Pelo negro (3)
	Rubios (4)	Bigote (5)
	Canosos (3)	Barba (4)
	Pelirrojos (4)	Sombrero (4)

¿Quién es quién?



¿Quién es quién?

~Mujeres (19) → ~Pelados (14) →

Anteojos (2)

Rubios (4)

Canosos (2)

Pelirrojos (2)

Castaños (3)

Pelo negro (2)

Bigote (4)

Barba (2)

Sombrero (4)

Árbol de decisión



¿Quién es quién? - Árbol de decisión



¿Quién es quién? - Árbol de decisión



¿Quién es quién? - Árbol de decisión



¿Quién es quién? - Árbol de decisión

Anteojos 



¿Quién es quién? - Árbol de decisión



¿Quién es quién? - Árbol de decisión

```
if (woman) {  
  if (hat) {  
    if (glasses) {  
      return "Clara";  
    } else {  
      return "María";  
    }  
  } else {  
    if (blond) {  
      return "Anita";  
    }  
  }  
}
```

¿Quién es quién? - Árbol de decisión



I Am Devloper

@iamdevloper

You say: "We added AI to our product"
I hear: "We added a bunch more IF
statements to our codebase"

7:07 AM - 10 Feb 2017

3,315 Retweets 5,622 Likes



44



3.3K



5.6K



Bosques de árboles de decisión

Ya están implementados: Random Forest, XGBoost, otros.

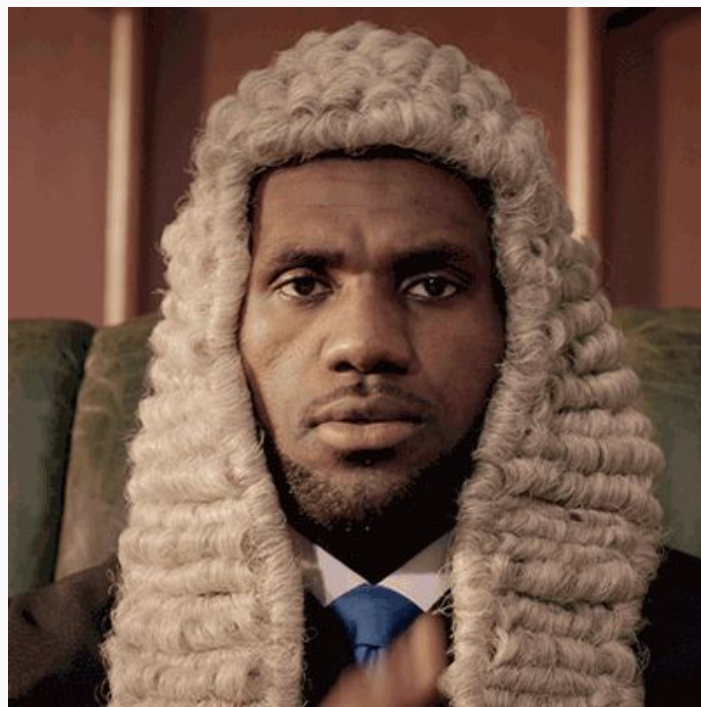
Sólo hay que pasarle una matriz de datos

	A	B	C	D	E	F	G	H	I	J	K	L
1	name	woman	has_glasses	bald	yellow_haired	white_haired	red_haired	brown_haired	black_haired	has_mustache	has_beard	has_hat
2	Samuel	0	1	1	0	1	0	0	0	0	0	0
3	Pepe	0	1	0	1	0	0	0	0	0	0	0
4	Pablo	0	1	0	0	1	0	0	0	0	0	0
5	Jorge	0	0	0	0	1	0	0	0	0	0	1
6	Felipe	0	0	0	0	0	0	0	1	0	1	0
7	Clara	1	1	0	0	0	1	0	0	0	0	0
8	Anita	1	0	0	1	0	0	0	0	0	0	0
9	Alfredo	0	0	0	0	0	1	0	0	1	0	0
10	Susana	1	0	0	0	1	0	0	0	0	0	0
11	Ricardo	0	0	1	0	0	0	1	0	1	1	0
12	Paco	0	0	0	0	0	1	0	0	0	0	0
13	Manuel	0	0	0	0	0	0	1	0	1	0	

Machine Learning

Pero si ya está implementado, ¿cuál es tu trabajo?

Machine Learning



Machine Learning

Recopilar los datos

Prepararlos para los modelos

Tener ideas de cosas que podrían servir y probar si suman

Machine Learning

Por ejemplo en vez de hacer una columna por cada color de pelo hacer una que sea tonalidad de pelo: 0 para blanco, 1 rubio, 2 pelirrojo, 3 castaño, 4 negro.

De esa forma el modelo en vez de preguntar si el pelo es blanco puede preguntar si es pelirrojo o más claro y hacer grupos más convenientes.

¿Quién es quién?

Mujeres (5)

Anteojos (5)

Pelados (5)

Canosos o más claro (5)

Rubios o más claro (10)

Pelirrojos o más claro (15)

Castaños o más claro (20)

Pelo negro o más claro (24)

Bigote (5)

Barba (4)

Sombrero (5)

¿Quién es quién?

Mujeres (5)

Anteojos (5)

Pelados (5)

Canosos o más claro (5)

Rubios o más claro (10)

Pelirrojos o más claro (15)

Castaños o más claro (20)

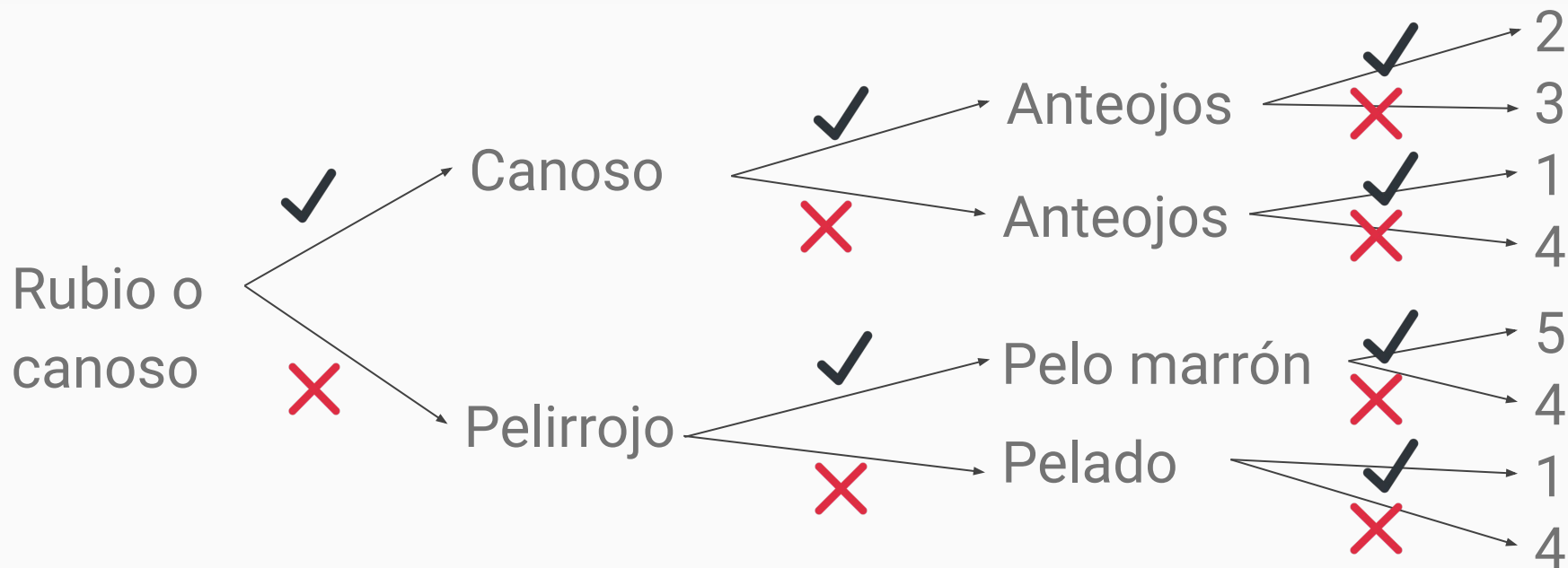
Pelo negro o más claro (24)

Bigote (5)

Barba (4)

Sombrero (5)

¿Quién es quién? - Árbol de decisión agrupando color de pelo



¿Quién es quién? - Árbol de decisión sin agrupar color de pelo



Machine Learning - Para qué



Machine Learning - Para qué

Estimar una propiedad desconocida basado en propiedades conocidas.

- Detección de fraude ¿La compra se parece más a una compra fraudulenta o legítima?
- Predicción de compra ¿El usuario se parece más a un usuario que compra o a uno que no?
- Elegir aviso a mostrar ¿Qué tipo de aviso parece más posible que este usuario clickee?

Modelo para películas

Machine Learning - Para qué

¡Estimar si me va a gustar una película en base a otras películas que vi!

Pueden probarlo

<https://seppo0010.github.io/movies-recommendation/>

Los datos vienen de

<https://www.kaggle.com/tmdb/tmdb-movie-metadata>

Machine Learning - Flujo aproximado

Paso 1: Elegir features. El dataset incluye título, género, palabras clave, fecha de lanzamiento, puntaje promedio en tmdb.

Paso 2: Etiquetar resultado de algunas películas.

Paso 3: Ver qué onda.

Paso 4: Volver a paso 1 o que sean las 6 de la tarde y te vayas del trabajo.

Elegir features

Empecé por features sencillos: los géneros de las películas.

Acción, aventura, animación, fantasía, comedia, drama, romance, guerra.

Advertencia

El siguiente slide contiene mi opinión sobre películas.

La misma no será compartida por el público.

Elegir películas



Accion
Aventura
Fantasía



Accion
Aventura
Fantasía



Animacion



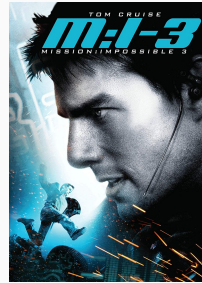
Animación
Aventura
Comedia



Comedia
Fantasía



Accion
Aventura
Fantasía

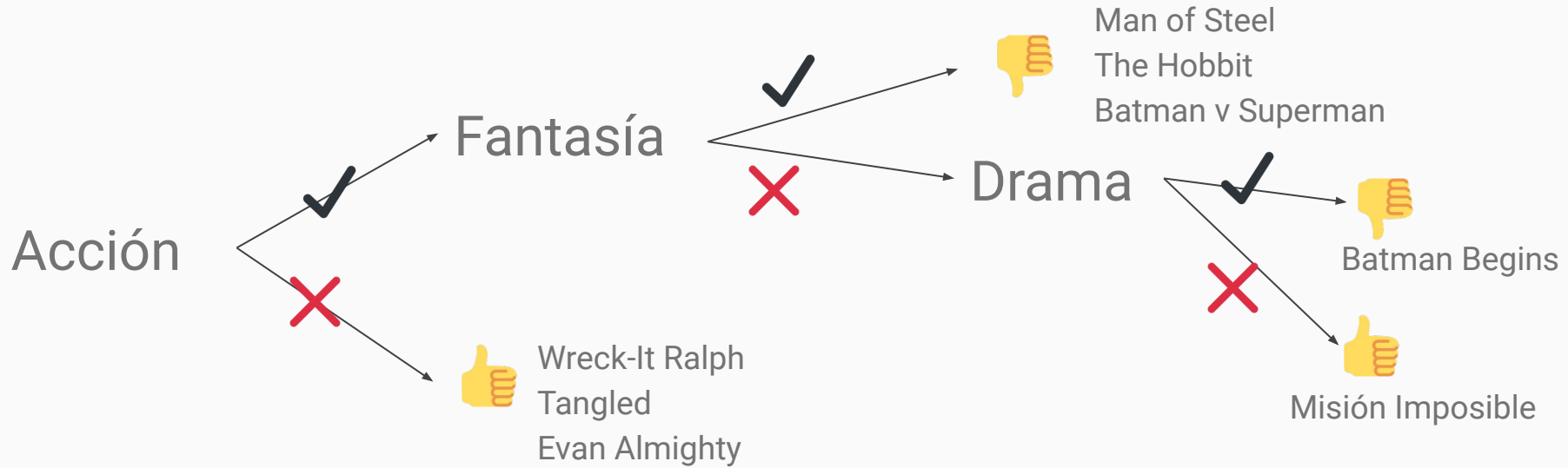


Accion
Aventura

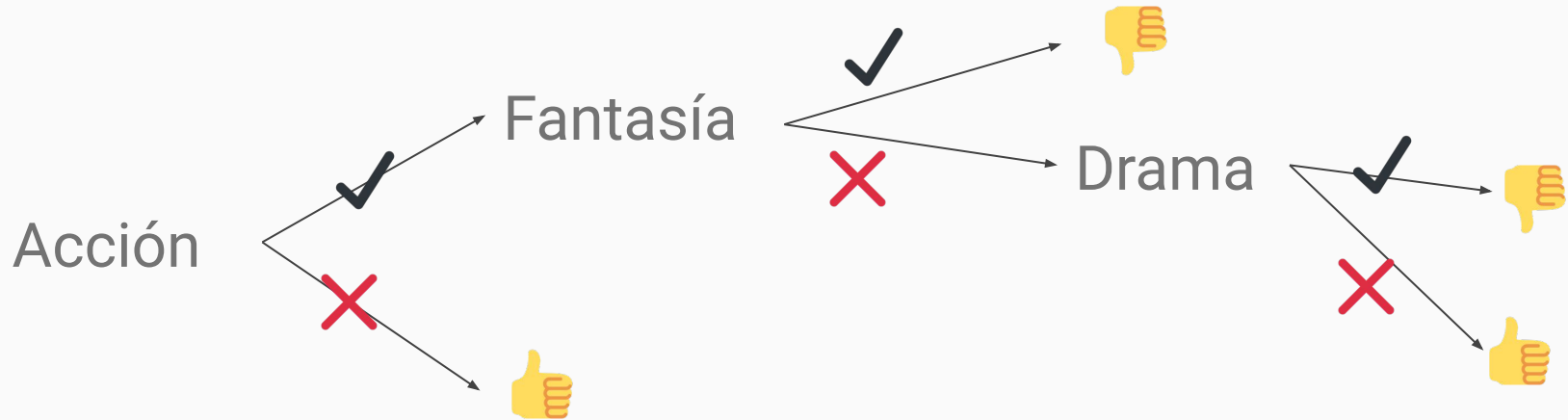


Accion
Drama

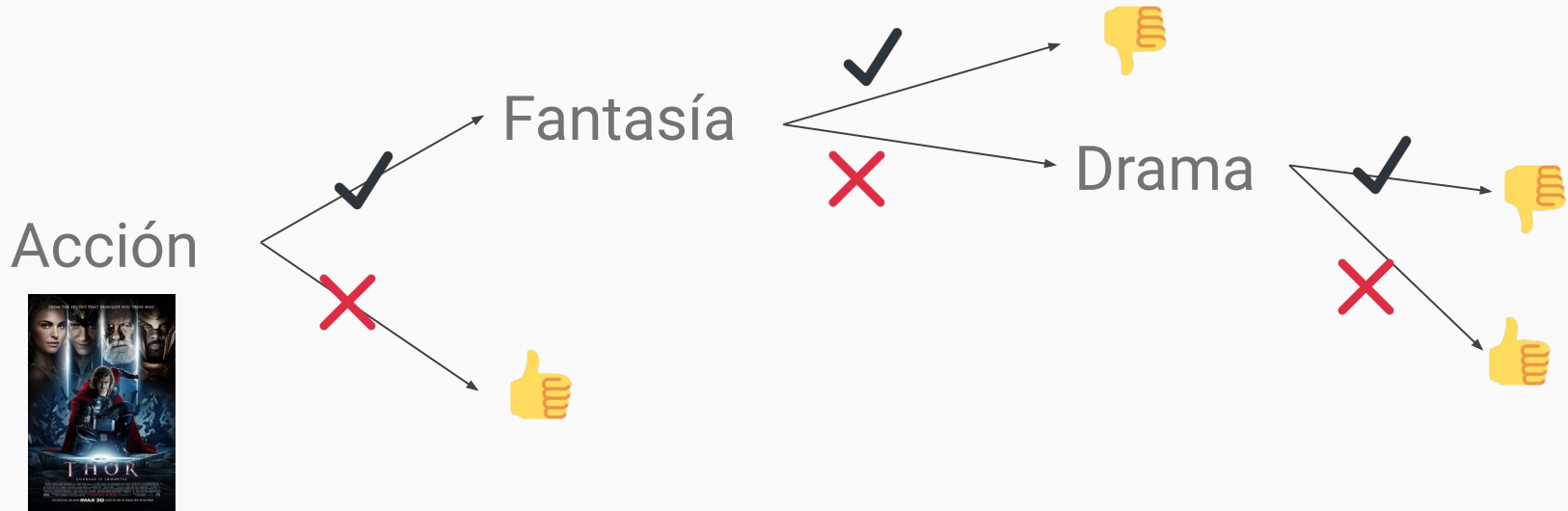
Árbol de decisión



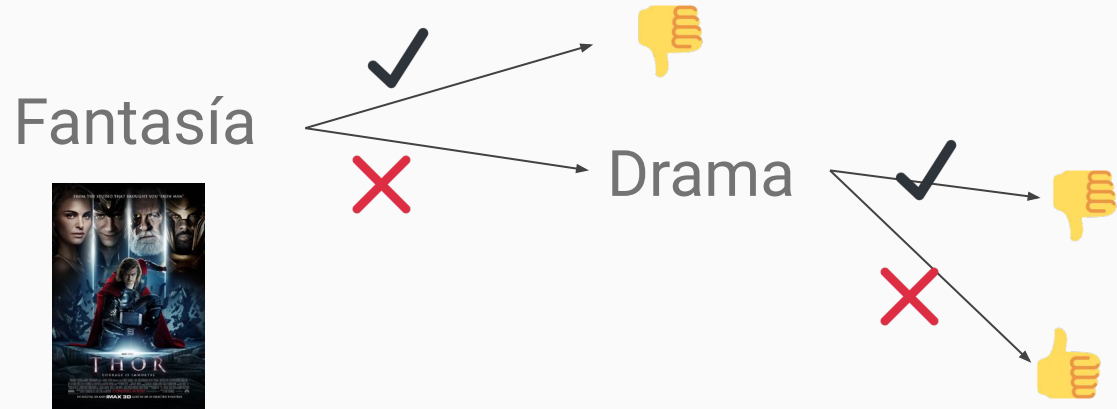
Árbol de decisión - Nuestro modelo de ML



Recorrer el árbol de decisión



Recorrer el árbol de decisión



Recorrer el árbol de decisión

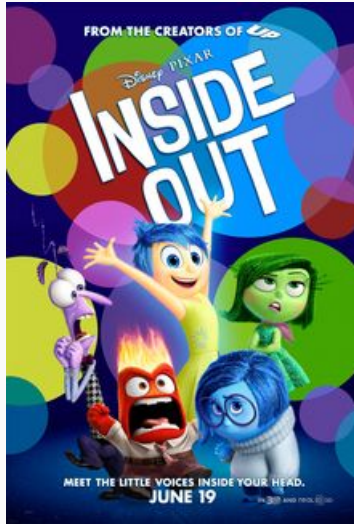


Ahora con más datos

Agregué más películas: 41 buenas y 36 malas en total.



¿Qué tal funciona?



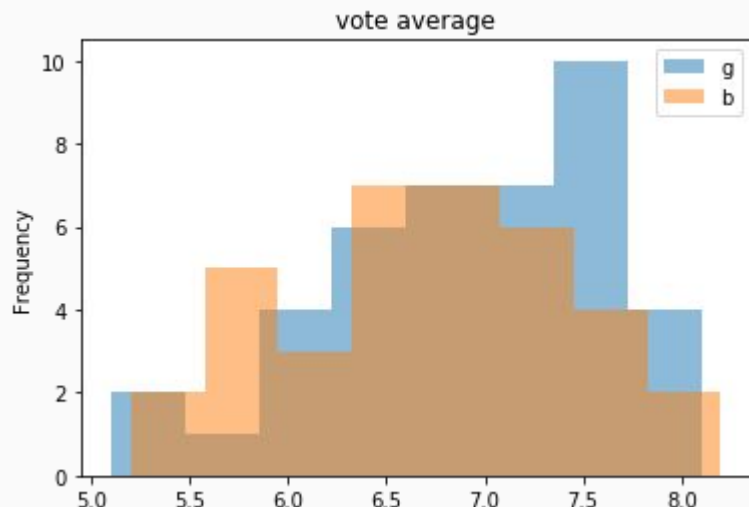
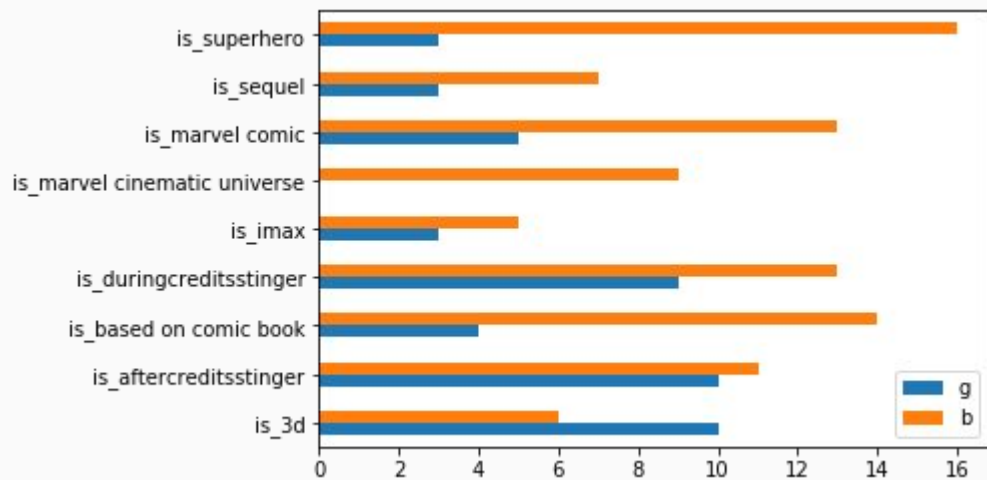
100%



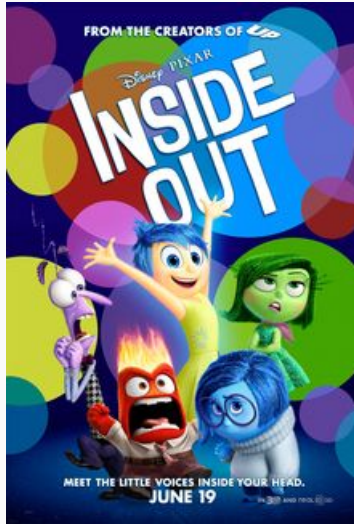
8%

Repetir paso 1: agregar más features

Agregamos promedio de voto y keywords



¿Qué tal funciona?



100%



0%

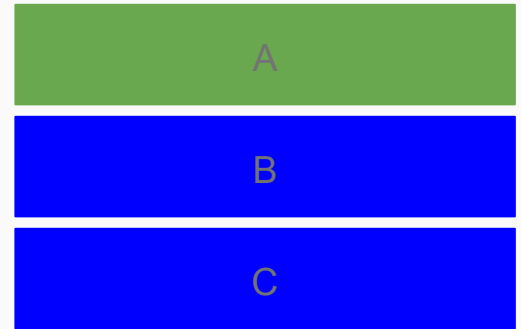
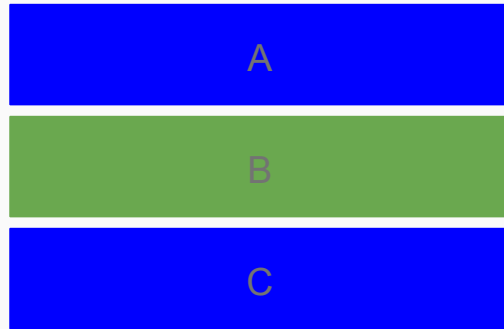
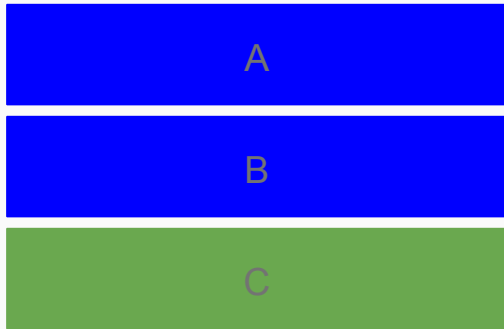
¿Qué tal funciona?

Contra las 77 películas que tengo etiquetadas, acierta 76, un 99% de eficacia.

Demasiado bueno para ser verdad, acá debe haber trampa.

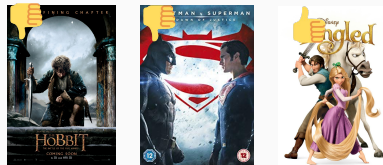
¿Qué tal funciona? Cross Validation

Train Test

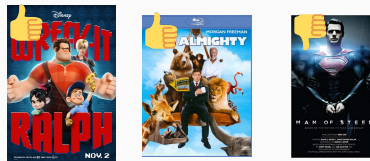


¿Qué tal funciona? Cross Validation

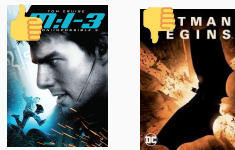
A



B



C



¿Qué tal funciona? Cross Validation

A



B

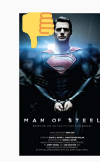


¿Qué tal funciona? Cross Validation

A



B

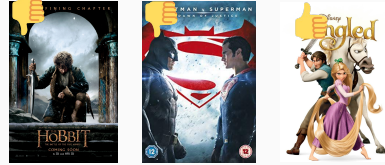


C



¿Qué tal funciona? Cross Validation

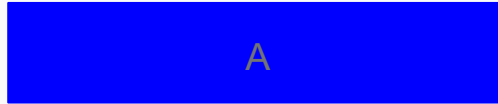
A



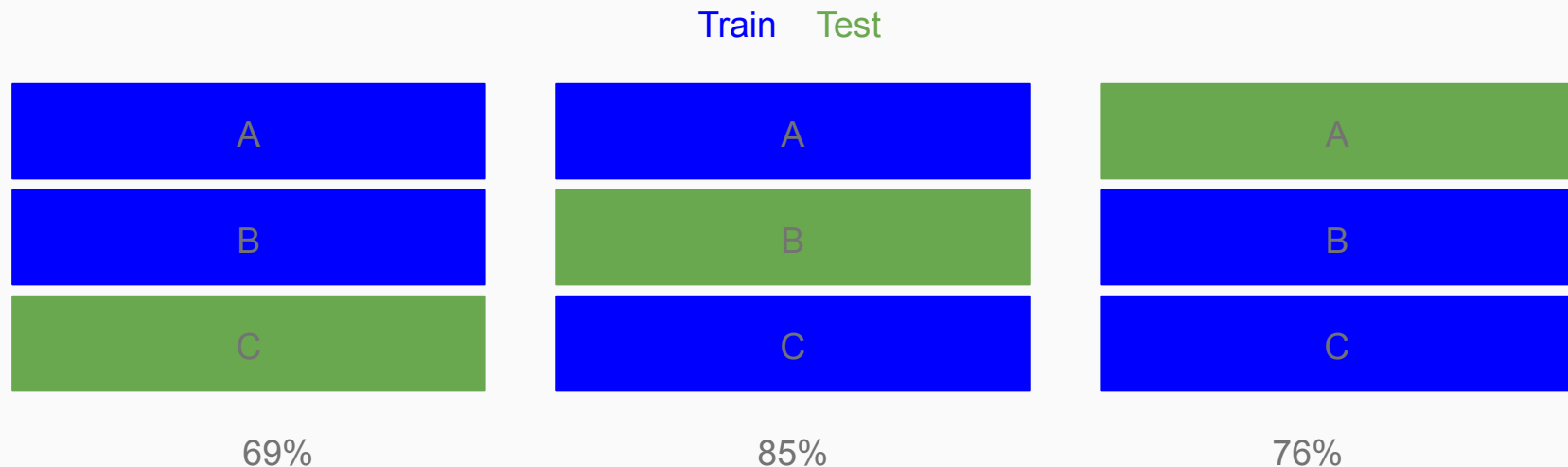
C



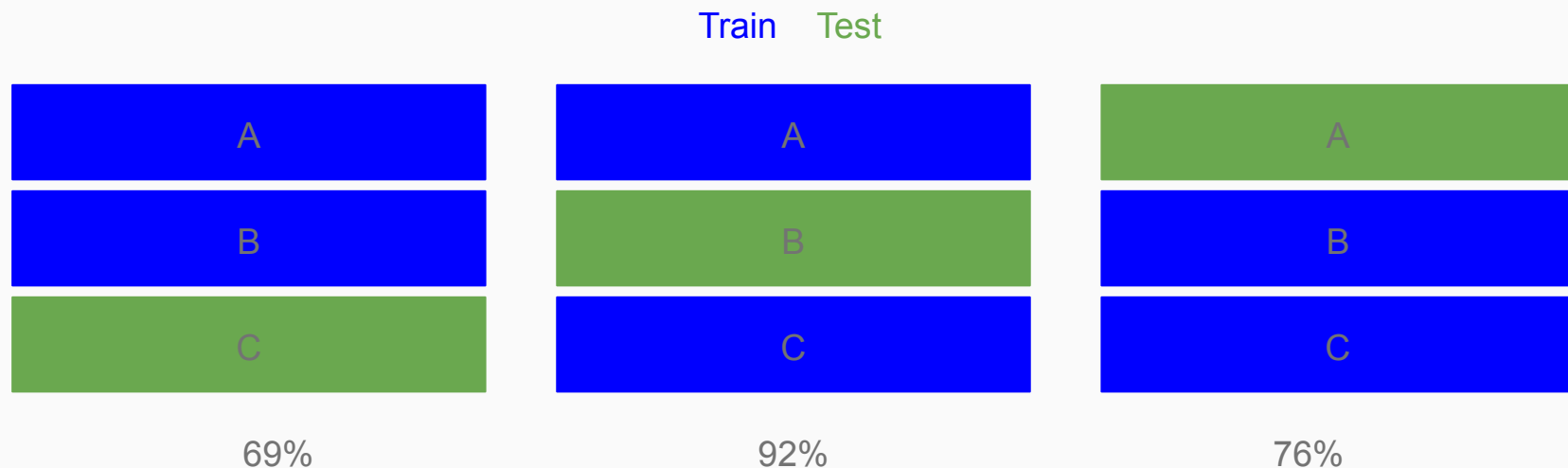
¿Qué tal funciona? Cross Validation



¿Qué tal funciona el modelo 1 (sólo géneros)?



¿Qué tal funciona el modelo 2 (con votos y keywords)?



Siguiente paso

Probar cosas:

- Agregar features
- Sacar features
(¡a veces sacar ayuda!)
- Buscar más datos

Generar un nuevo modelo. Medirlo. Seguir probando.

Siguiente paso

Ideas para features

- ¿Es secuela? Si lo es, ¿la anterior (o la primera) es buena?
- ¿Qué porcentaje de personajes son mujeres?
- ¿Conocemos a otras personas con gustos parecidos? ¿Y a cuántas les gustó la película?

Ejemplos de uso

Ejemplos de uso

Cada seis meses sysarmy hace una encuesta de sueldos de tecnología y publica las respuestas. En enero de 2019 unas 4700 personas respondieron.

Pregunta sexo, edad, educación, experiencia, (...) y sueldos

Me identifico	Tengo	Dónde estás trabajando	Años de experiencia	Años en la empresa actual	Años en el puesto actual	¿Gente a cargo?	Nivel de estudios alcanzados	Estado	Carrera	Universidad
Hombre		22 Catamarca	2	1	1		0 Universitario	En curso	ING Informática	Unca
Hombre		23 Catamarca	2	1	1		0 Secundario	Completado	Ing informática	UNCa
Mujer		35 Chaco	8	7	2		2 Posgrado	En curso	MTI	Universidad d
Hombre		41 Chaco	20	2	10		0 Universitario	Completado	Lic en sistemas	Uba / CAECE
Hombre		35 Chaco	5	3	2		0 Universitario	Completado	Analisis en Sistemas	UTN
Hombre		38 Chaco	16	16	16		0 Terciario	Incompleto	Analista de Sistemas	
Hombre		30 Chaco	9	5	5		3 Terciario	Completado	T S P	Utn fre
Hombre		32 Chaco	10	2	2		0 Universitario	Incompleto	Ing. en Sistemas	UTN FRRe (R

Ejemplos de uso



Ejemplos de uso

Hice un formulario para que cada uno pueda responder y obtener una estimación de cuánto debería ganar

<https://seppo0010.github.io/sysarmy-sueldos/>

Resultados



Santiago Castellanos
@santocaste

Replying to @seppo0011 and @sysarmy

Es bruto ese monto ? Me dió un 25 % abajo. Pero mis tecnologías estrellas no estaban: nano y bash.



Juan Soto
@juansoto86

Replying to @seppo0011 and @sysarmy

Muy bien me dio!



Santiago Basulto
@santiagobasulto

Replying to @seppo0011 and @sysarmy

Se ve que debo ser un gran cebador de mates, pq me tiró un número por el piso 🤔



Cocot ❤️
@arito21

Replying to @seppo0011 and @sysarmy

Muy bueno! Me dijo el sueldo anterior! Y esta bien porque creo que cuando lo hice estaba en el trabajo anterior! Buenísima la idea!



co ❤️
@co_constanza

Replying to @seppo0011 @ramblinn_ and @sysarmy
¡Genial! Muy, muy aproximado 🎯



Black mr Meeseeks
@k_bordon

Replying to @seppo0011 and @sysarmy

Excelente, muy similar a lo que estoy cobrando



Sebastian Cipolat
@seba_cipolat

Replying to @seppo0011 and @sysarmy

Le error un toque pero como idea sirve



SeedVicious
@SeedVicious

Replying to @seppo0011 and @sysarmy
200 \$ de diferencia



Bren 🇪🇺
@bren_sk8

Replying to @seppo0011 and @sysarmy

Muy buena idea ! En mi caso probé para mi pareja y yo . A mí me dio 30% más abajo de lo que gano realmente y a mi novio le dio 70% más arriba de lo que él está ganando actualmente . Me siento bien por mí (? Pero muy mal por el 🤔)



Chilita
@chilitaaaa

Replying to @seppo0011 and @sysarmy

Es extremadamente preciso !! 🎯🎯🎯



Tana
@PeiblTapia

Replying to @seppo0011 and @sysarmy

Veo que ya esta funcionando! Excelente Laburo, me dio solo con 300 pesos de diferencia, una locura



Tamar Anush
@tamar_moz

Replying to @seppo0011 and @sysarmy

Muy bueno! Pero a mí me dio que debería tener un sueldo que es más del doble del actual 😞😞



Pink Whale ❤️
@PinkyWhale

Replying to @seppo0011 and @sysarmy

Le pifió por 7k



Facundo
@FacundSua

Replying to @seppo0011 and @sysarmy

Bueno no le vamos a pasar ésto a mí empleador porque, en papel, me estoy choreando 10 lucas casi



agustin
@Agusttyny

Replying to @seppo0011 and @sysarmy

Le pegó bastante bien, esto también te sirve como par orientarte hacia donde ir \$:p

Resultados



Malena Rey ❤️

@malerey_

Replying to @malerey_ @seppo0011 and @sysarmy

Lo depresivo es que muestra muy claro que el mismo perfil cobra menos solo cambiando la variable hombre/mujer



Gabriel Patiño ❤️💛

@gepatino

Replying to @seppo0011 and @sysarmy

Te baja el sueldo casi un 10% por cambiar de hombre a mujer. Sigue pasando en nuestra industria? Mal ahí...

Riesgos de Machine Learning



Sebastian Ariel Waisbrot @seppo0011 · Mar 12
Usé las respuestas de la encuesta de @sysarmy para armar un modelo de sueldos. Pueden probarlo y ver cómo lo armé acá seppo0010.github.io/sysarmy-sueldo...

52 70 231

Sebastian Ariel Waisbrot
@seppo0011

Le pedí feedback a un amigo

Amigo: hay que tener cuidado por si alguien lo quiere usar para poner sueldos en su empresa
Yo: nadie puede ser tan idiota
A: a la gente de rrhh le interesaría
Y: oh, no, qué he hecho

<https://twitter.com/seppo0011/status/1106941824160223232>

Riesgos de Machine Learning

Predicción de Embarazo Adolescente con Machine Learning (MS+Gob Salta)

En colaboración con el Ministerio de Primera Infancia del Gobierno Provincial de Salta, definimos como objetivo utilizar inteligencia artificial para identificar aquellas adolescentes con mayor riesgo de quedar embarazada.

Afortunadamente, tuvimos acceso a un amplio espectro de datos. Utilizamos un dataset de más de 200.000 residentes de la ciudad de Salta con más de 12.000 mujeres de entre 10 y 19 años , Argentina. Estos datos no contienen información personal identificable sobre las personas, tal como reconoce habeas data.

A través de las herramientas Azure Machine Learning y SQL Server 2016, logramos crear diferentes modelos predictivos que permiten detectar hasta el 90% de los casos en riesgo de embarazo adolescente.

<https://github.com/facundod/case-studies/blob/master/Prediccion%20de%20Embarazo%20Adolescente%20con%20Machine%20Learning.md>

Riesgos de Machine Learning

Sobre la predicción automática de embarazos adolescentes (LIAA UBA)

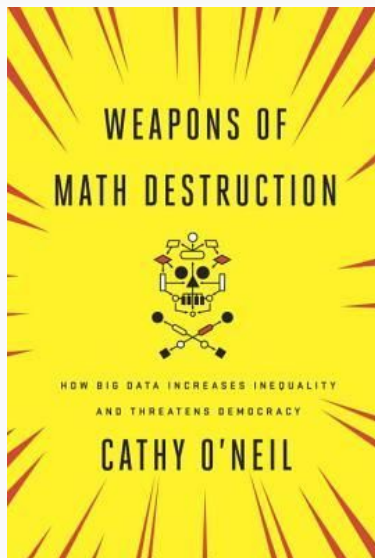
Los datos de evaluación incluyen réplicas casi idénticas de muchos datos de entrenamiento

Los datos de embarazos adolescentes tienen una tendencia a estar sesgados o incompletos, debido a que son un tema sensible y confidencial, de difícil acceso.

Los datos utilizados fueron extraídos de una encuesta a adolescentes residentes en la provincia de Salta conteniendo información personal, sobre su entorno y sobre si había cursado o estaba cursando, al momento de la encuesta, un embarazo. Estos datos no son adecuados para responder a la pregunta planteada: si una adolescente tendrá un embarazo en el futuro.

<https://www.dropbox.com/s/r7w4hln3p5xum3v/%5BLIAA%5D%20Sobre%20la%20predicci%C3%B3n%20autom%C3%A1tica%20de%20embarazos%20adolescentes.pdf?dl=0>

Riesgos de Machine Learning

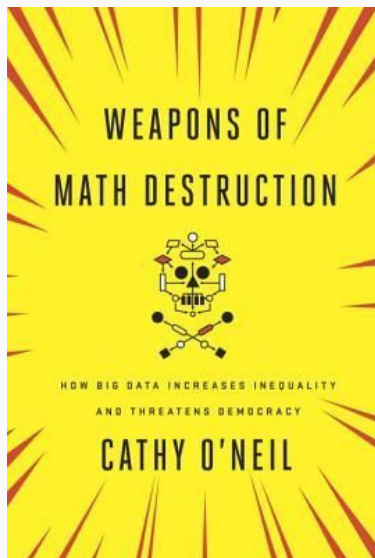


Washington D.C. usó algoritmos de machine learning para evaluar profesores. Los peores 2% fueron despedidos. No se podía apelar porque nadie entendía la evaluación.

Este es un mal caso de uso para machine learning: muchas cosas afectan el desempeño de los alumnos, no se recibe feedback intermedio, no se puede corregir.

<https://www.goodreads.com/book/show/28186015-weapons-of-math-destruction>

Riesgos de Machine Learning

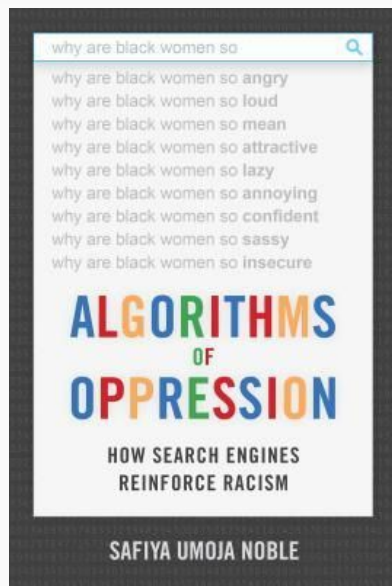


Se consideran datos de una persona para contratar o promover en una empresa, ayuda a jueces a establecer una sentencia, a bancos a dar préstamos, y dan la apariencia de ser neutrales.

El modelo sólo puede ser tan bueno como los datos que recibe. Va a mantener los sesgos preexistentes y justificarlos.

<https://www.goodreads.com/book/show/28186015-weapons-of-math-destruction>

Riesgos de Machine Learning



algunas de las personas que desarrollan algoritmos de búsqueda y su arquitectura están dispuestas a promover actitudes sexistas y racistas abiertamente en el trabajo y más allá, mientras se supone que deberíamos creer que estos mismos empleados están desarrollando herramientas de toma de decisiones “neutrales” u “objetivas”

<https://www.goodreads.com/book/show/34762552-algorithms-of-oppression>

Riesgos de Machine Learning

The problem of bias in word embeddings

Man:Woman as King:Queen

Man:Computer_Programmer as Woman:Homemaker ✗

Father:Doctor as Mother:Nurse ✗

Word embeddings can reflect gender, ethnicity, age, sexual orientation, and other biases of the text used to train the model.

Links de interés

Slides y código

<https://github.com/seppo0010/movies-recommendation>

<https://github.com/seppo0010/movies-recommendation/blob/master/Explore%20data.ipynb>

Cursos (más o menos) gratis

<https://www.coursera.org/specializations/data-science-python>

<https://www.coursera.org/learn/machine-learning>

<https://developers.google.com/machine-learning/recommendation/index>

Datasets y desafíos

<https://www.kaggle.com/>

<https://toolbox.google.com/datasetsearch>

Gracias