INICIO GRABACIÓN

MINERIADEDATOS//DATAWAREHOUSE
MACHINE LEARNING VS DEEP LEARNING

SANJOSÉ
FUNDACIÓN DE EDUCACIÓN SUPERIOR

SANJOSÉ
FUNDACIÓN DE EDUCACIÓN SUPERIOR

INDICE

SANJOSÉ

FUNDACIÓN DE EDUCACIÓN SUPERIOR

PRESENTACIÓN

MACHINE LEARNING VS DEEP LEARNING

¿Qué Es Machine Learning?

La capacidad y al abaratamiento de las tecnologías de la información y de los sensores, podemos producir, almacenar y enviar más datos que nunca antes en la historia. De hecho, se calcula que el 90% de los datos disponibles actualmente en el planeta se ha creado en los últimos dos años, produciéndose actualmente en torno a 2,5 quintillones (2.500.000.000.000.000.000) de bytes por día, siguiendo una tendencia fuertemente creciente. Estos datos alimentan los modelos de Machine Learning y son el impulso principal del auge que esta ciencia ha experimentado en los últimos años.

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

MACHINE LEARNING

Machine Learning es uno de los subcampos de la Inteligencia Artificial y puede ser definido como:

“Machine Learning es la ciencia que permite que las computadoras aprendan y actúen como lo hacen los humanos, mejorando su aprendizaje a lo largo del tiempo de una forma autónoma, alimentándolas con datos e información en forma de observaciones e interacciones con el mundo real.” — Dan Fagella

Ofrece una manera eficiente de capturar el conocimiento mediante la información contenida en los datos, para mejorar de forma gradual el rendimiento de modelos predictivos y tomar decisiones basadas en dichos datos. Se ha convertido en una tecnología con una amplia presencia, y actualmente está presente en: filtros anti-spam para correo electrónico, conducción automática de vehículos o reconocimiento de voz e imágenes.

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

MACHINE LEARNING

Ejemplo, el siguiente vídeo muestra una detección de eventos en tiempo real para una aplicación de vídeo-vigilancia basada en Machine Learning.

En Machine Learning generalmente se utilizan matrices y notaciones vectoriales para referirnos a los datos, de la siguiente forma:

- Cada fila de la matriz es una muestra, observación o dato puntual.

- Cada columna es una característica (o atributo), de la observación mencionada en el punto anterior (“feature” en la imagen inferior).

En el caso más general habrá una columna, que llamaremos objetivo, etiqueta o respuesta, y que será el valor que se pretende predecir. (“label” en la imagen inferior.

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

MACHINE LEARNING

← Features →					Label
Position	**Experience**	**Skill**	**Country**	**City**	**Salary ($)**
Developer	0	1	USA	New York	103100
Developer	1	1	USA	New York	104900
Developer	2	1	USA	New York	106800
Developer	3	1	USA	New York	108700
Developer	4	1	USA	New York	110400
Developer	5	1	USA	New York	112300
Developer	6	1	USA	New York	114200
Developer	7	1	USA	New York	116100
Developer	8	1	USA	New York	117800
Developer	9	1	USA	New York	119700
Developer	10	1	USA	New York	121600

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

MACHINE LEARNING

Existen algoritmos específicos cuyo propósito es “entrenar” los modelos de Machine Learning. Dichos algoritmos proporcionan datos de entrenamiento que permiten a los modelos aprender de ellos.

Con respecto a los algoritmos de Machine Learning, normalmente tienen determinados parámetros “internos”. Por ejemplo en los árboles de decisión, hay parámetros como profundidad máxima del arbol, número de nodos, número de hojas, a estos parámetros se les llama “hiperparametros”.

Llamamos “generalización” a la capacidad del modelo para hacer predicciones utilizando nuevos datos.

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

MACHINE LEARNING

Tipos de Machine Learning

Los tipos de Machine Learning que se tratarán son:

- Aprendizaje supervisado
- Aprendizaje no supervisado
- Aprendizaje profundo

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

MACHINE LEARNING

Aprendizaje No Supervisado

En el aprendizaje no supervisado, trataremos con datos sin etiquetar cuya estructura es desconocida. El objetivo será la extracción de información significativa, sin la referencia de variables de salida conocidas, y mediante la exploración de la estructura de dichos datos sin etiquetar.

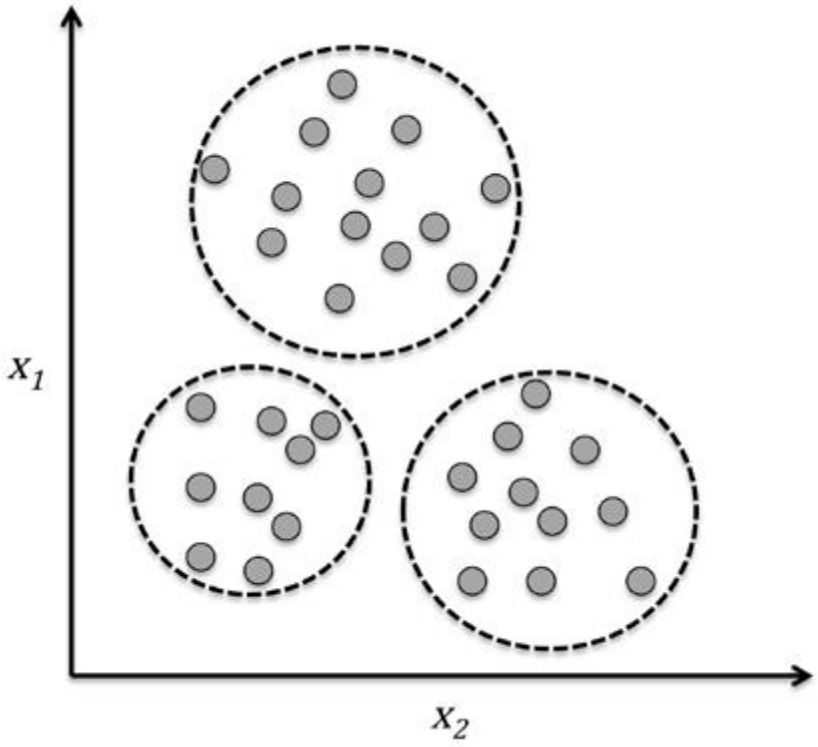
Hay dos categorías principales: agrupamiento y reducción dimensional.

Agrupamiento ó Clustering: El agrupamiento es una técnica exploratoria de análisis de datos, que se usa para organizar información en grupos con significado sin tener conocimiento previo de su estructura. Cada grupo es un conjunto de objetos similares que se diferencia de los objetos de otros grupos. El objetivo es obtener un numero de grupos de características similares.

Un ejemplo de aplicación de este tipo de algoritmos puede ser para establecer tipos de consumidores en función de sus hábitos de compra, para poder realizar técnicas de marketing efectivas y “personalizadas”.

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

MACHINE LEARNING



Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

MACHINE LEARNING

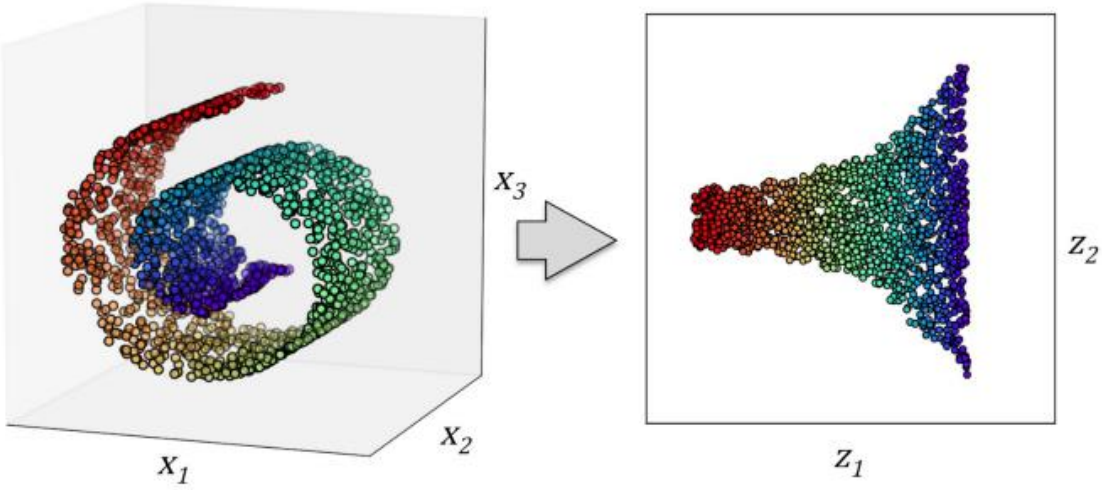
2. Reducción dimensional

Es común trabajar con datos en los que cada observación se presenta con alto número de características, en otras palabras, que tienen alta dimensionalidad. Este hecho es un reto para la capacidad de procesamiento y el rendimiento computacional de los algoritmos de Machine Learning. La reducción dimensional es una de las técnicas usadas para mitigar este efecto.

La reducción dimensional funciona encontrando correlaciones entre las características, lo que implica que existe información redundante, ya que alguna característica puede explicarse parcialmente con otras (por ejemplo, puede existir dependencia lineal). Estas técnicas eliminan “ruido” de los datos (que puede también empeorar el comportamiento del modelo), y comprimen los datos en un sub-espacio más reducido, al tiempo que retienen la mayoría de la información relevante.

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

MACHINE LEARNING



Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

MACHINE LEARNING

Aprendizaje Supervisado

Se refiere a un tipo de modelos de Machine Learning que se entrenan con un conjunto de ejemplos en los que los resultados de salida son conocidos. Los modelos aprenden de esos resultados conocidos y realizan ajustes en sus parámetros interiores para adaptarse a los datos de entrada. Una vez el modelo es entrenado adecuadamente, y los parámetros internos son coherentes con los datos de entrada y los resultados de la batería de datos de entrenamiento, el modelo

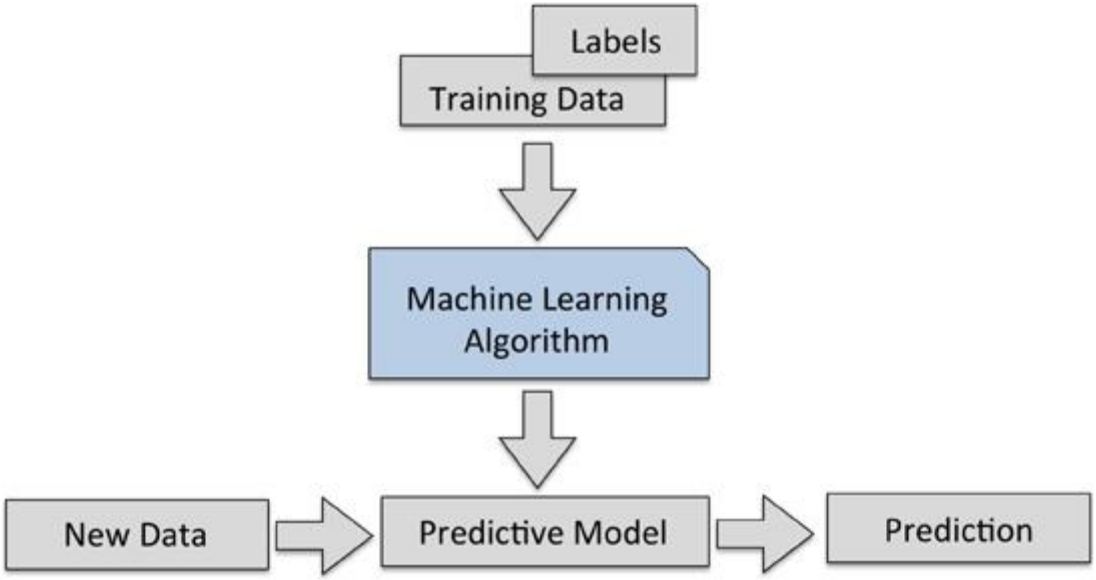
Se podrá realizar predicciones adecuadas ante nuevos

datos no procesados previamente.

De forma gráfica:

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

MACHINE LEARNING



Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

MACHINE LEARNING

Hay dos aplicaciones principales de aprendizaje supervisado: clasificación y regresión:

Clasificación:

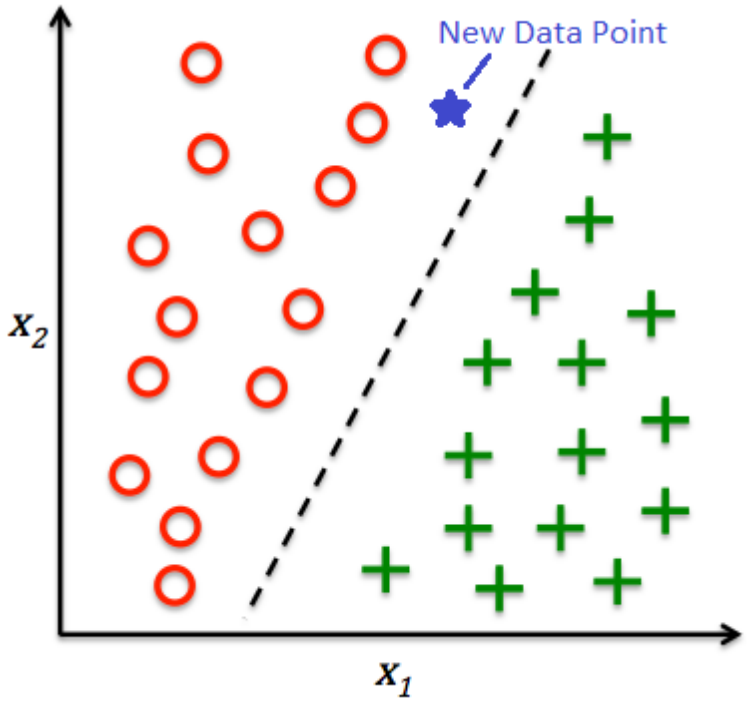
Clasificación es una sub-categoría de aprendizaje supervisado en la que el objetivo es predecir las clases categóricas (valores discretos, no ordenados, pertenencia a grupos). El ejemplo típico es la detección de correo spam, que es una clasificación binaria (un email es spam — valor “1”- o no lo es — valor “0” -).

También hay clasificación multi-clase, como el reconocimiento de caracteres escritos a mano (donde las clases van de 0 a 9).

Un ejemplo de clasificación binaria: hay dos clases de objetos, círculos y cruces, y dos características de los objetos, X1 y X2. El modelo puede encontrar las relaciones entre las características de cada punto de datos y su clase, y establecer la linea divisoria entre ellos. Así, al ser alimentado con nuevos datos, el modelo será capaz de determinar la clase a la que pertenecen, de acuerdo con sus características.

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

MACHINE LEARNING



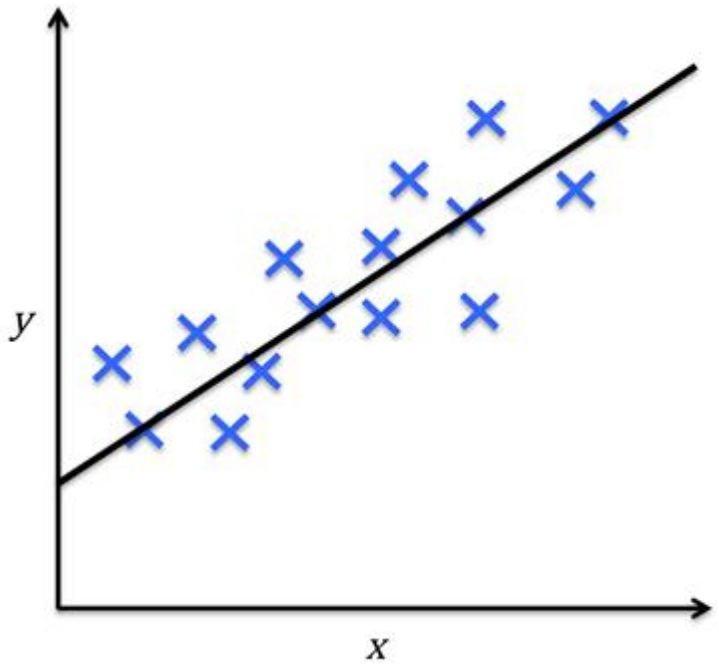
En este caso, el nuevo punto de datos entra en el área correspondiente al subespacio de círculos y por tanto, el modelo predecirá que la clase del objeto es círculo.

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

MACHINE LEARNING

2. Regresión:

La regresión se utiliza para asignar categorías a datos sin etiquetar. En este tipo de aprendizaje tenemos un número de variables predictoras (explicativas) y una variable de respuesta continua (resultado), y se tratará de encontrar una relación entre dichas variables que nos proporciones un resultado continuo.
Un ejemplo de regresión lineal: dados X e Y, establecemos una linea recta que minimice la distancia (con el método de mínimos cuadrados) entre los puntos de muestra y la línea ajustada. Después, utilizaremos las desviaciones obtenidas en la formación de la línea para predecir nuevos datos de salida.



Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

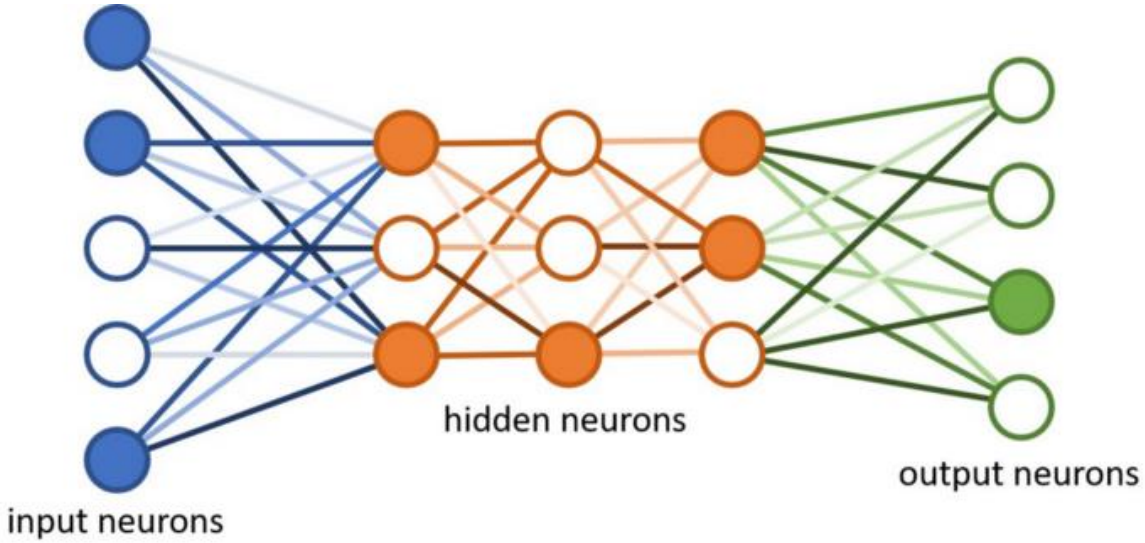
Deep Learning

El aprendizaje profundo ó Deep Learning, es un subcampo de Machine Learning, que usa una estructura jerárquica de redes neuronales artificiales, que se construyen de una forma similar a la estructura neuronal del cerebro humano, con los nodos de neuronas conectadas como una tela de araña. Esta arquitectura permite abordar el análisis de datos de forma no lineal.

La primera capa de la red neuronal toma datos en bruto como entrada, los procesa, extrae información y la transfiere a la siguiente capa como salida. Este proceso se repite en las siguientes capas, cada capa procesa la información proporcionada por la capa anterior, y así sucesivamente hasta que los datos llegan a la capa final, que es donde se obtiene la predicción. Esta predicción se compara con el resultado conocido, y así por análisis inverso el modelo es capaz de aprender los factores que conducen a salidas adecuadas.

Es uno de los principales algoritmos utilizados en la creación de aplicaciones y programas para reconocimiento de imágenes.

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

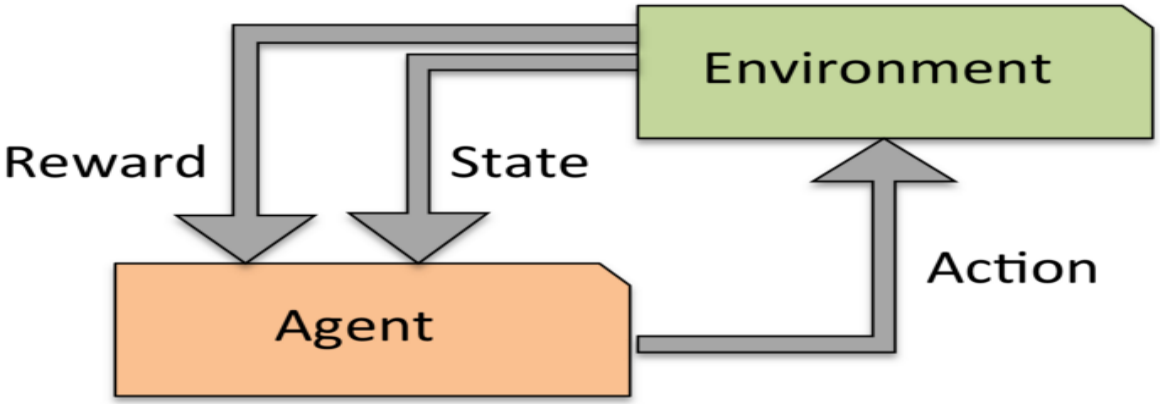


Aprendizaje reforzado

El aprendizaje reforzado es una de las ramas más importantes del aprendizaje profundo. El objetivo es construir un modelo con un agente que mejora su rendimiento, basándose en la recompensa obtenida del entorno con cada interacción que se realiza. La recompensa es una medida de lo correcta que ha sido una acción para obtener un objetivo determinado. El agente utiliza esta recompensa para ajustar su comportamiento futuro, con el objetivo de obtener la recompensa máxima.

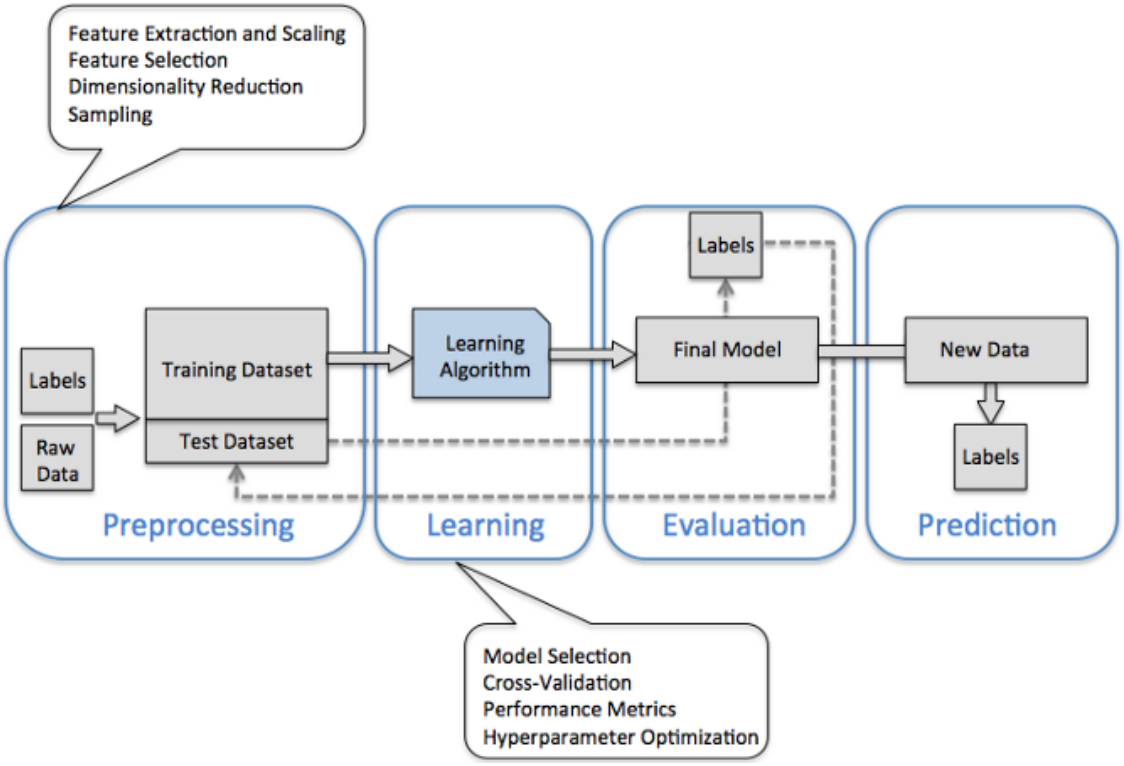
Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Un ejemplo común es una máquina de ajedrez, donde el agente decide entre una serie de posibles acciones, dependiendo de la disposición del tablero (que es el estado del entorno) y la recompensa se recibe según el resultado de la partida.



Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Metodología general para construir modelos de Machine Learning



Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Preprocesamiento:

Este es uno de los pasos más importantes en cualquier aplicación de Machine Learning. Usualmente los datos se presentan en formatos no óptimos (o incluso inadecuados) para ser procesados por el modelo. En estos casos el pre-procesamiento de datos es una tarea que se debe realizar de manera obligatoria.

Muchos algoritmos requieren que las características estén en la misma escala (por ejemplo, en el rango [0,1]) para optimizar su rendimiento, lo que se realiza frecuentemente aplicando técnicas de normalización o estandarización en los datos. Podemos también encontrar en algunos casos que las características seleccionadas están correlacionadas, y por tanto son redundantes para extraer información con significado correcto de ellas. En este caso tendremos que usar técnicas de reducción dimensional para comprimir las características en subespacios con menores dimensiones.

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Entrenando y seleccionando un modelo

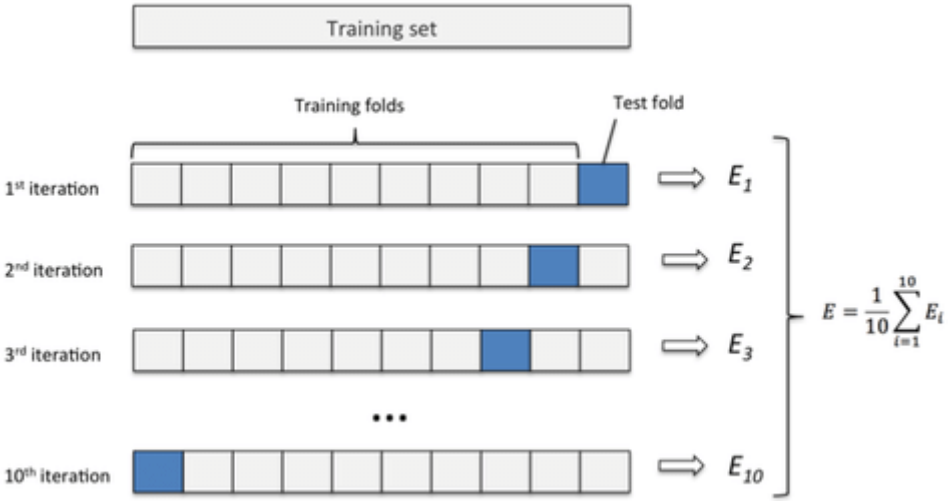
Es esencial comparar los diferentes algoritmos de un grupo para entrenar y seleccionar el de mejor rendimiento. Para realizar esto, es necesario seleccionar una métrica para medir el rendimiento del modelo. Una de ellas comúnmente usada en problemas de clasificación es la precisión de clasificación, que es la proporción de instancias correctamente clasificadas.

En los problemas de regresión, uno de los más comunes es el Error Cuadrático Medio (MSE), que mide la diferencia media cuadrática entre los valores estimados y los reales.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2$$

where N is the number of data points,
f_i the value returned by the model and
y_i the actual value for data point i.

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)



En general, los parámetros por defecto de los algoritmos de Machine Learning proporcionados por las librerías no son los mejores para utilizar con nuestros datos, por lo que usaremos técnicas de optimización de “hiperparámetros” para ayudarnos a realizar el ajuste fino del rendimiento del modelo.

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Evaluando Modelos y Prediciendo con Datos Nuevos

Una vez que hemos seleccionado y ajustado un modelo a nuestro conjunto de datos de entrenamiento, podemos usar los datos de prueba para estimar el rendimiento del modelo en los datos nuevos, por lo que podemos hacer una estimación del error de generalización del modelo, o evaluarlo utilizando alguna otra métrica.

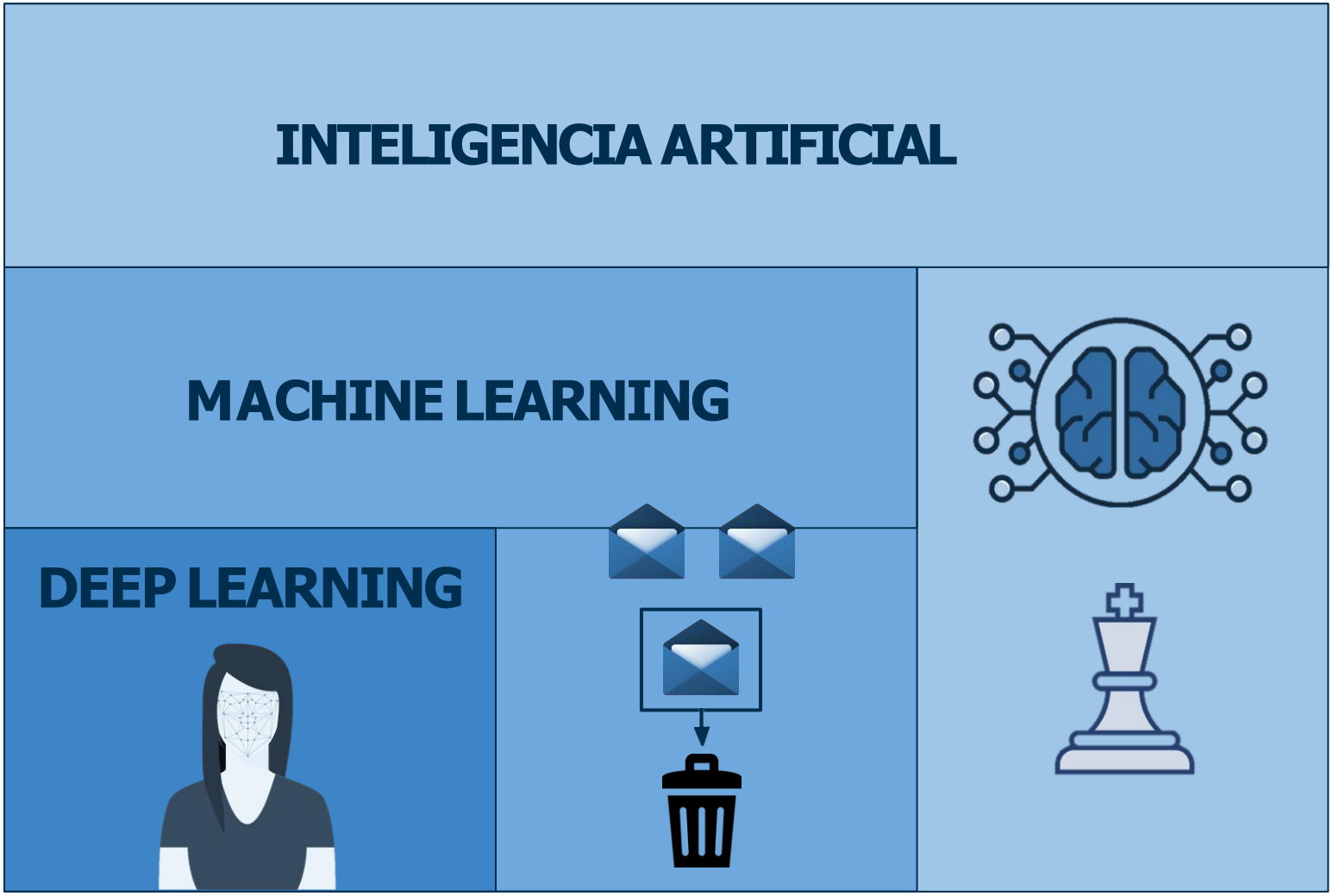
Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Algoritmos de Machine Learning

Existen muchos algoritmos de Machine Learning, desde los más básicos a otros más complejos. Podemos destacar algunos como los siguientes:

- Regresión lineal
- Regresión logística
- Árboles de decisión
- Random Forest
- XGBoost
- Gradient Boosting
- Isolation Forest
- Redes neuronales
- Support Vector Machines
- K-Means

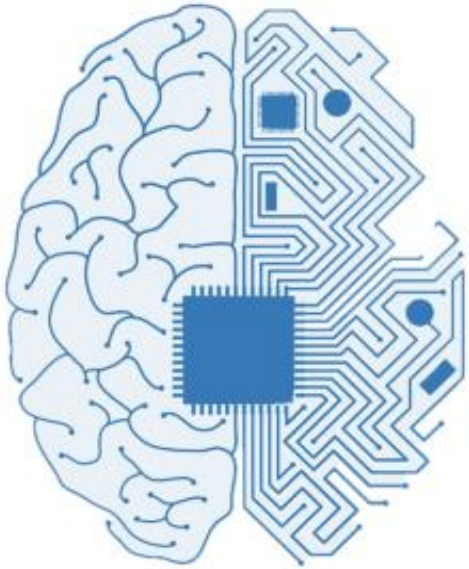
Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)



Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Machine learning

- Detección de fraudes
- Búsqueda web
- Anuncios a tiempo real
- Análisis de textos
- Next best action



Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

¿Qué es deep learning?

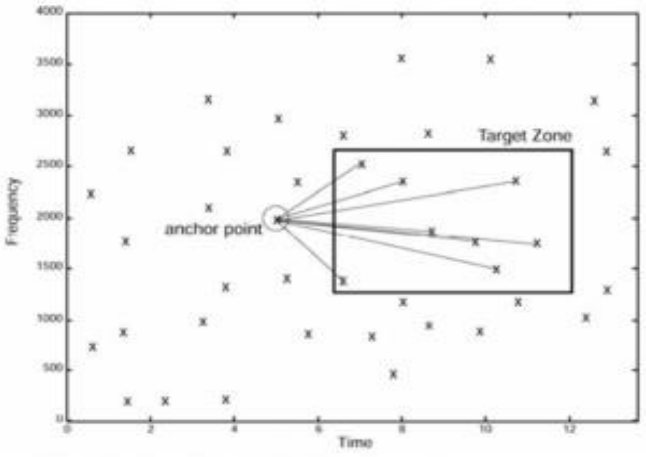
Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Aprender en qué consiste el deep learning y dónde está presente.

Deep learning

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Deep learning



Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Roles en datos

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Herramientas para cada etapa del análisis de datos

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Extracción de información con SQL

- Extracción de información
- Síntesis de la base de datos
- Cuadros de control de la operación

Análisis y visualización con R y Python

- Análisis de datos con enfoque estadístico
- Más complejo al inicio
- Análisis descriptivo y exploratorio
- Packages: ggplot2, dplyr

- Análisis de datos con enfoque en ingeniería
- Parecido a otros lenguajes
- Análisis descriptivo y exploratorio
- Packages: Pandas, Numpy

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Creación de queries en SQL

Queries para extraer información de una tabla.

Ejemplo de SQL

Objetivo

Saber CUÁNTAS bocinas hemos vendido por más de 600 MXN desde 2019

Tabla en Excel

VENTAS_2020

Día	Mes	Año	Producto	id	Valor
1	2	1998	Bocina	24	$528
12	4	2004	Auriculares	31	$240
14	8	2016	Auriculares	14	$315
16	10	2019	Bocina	200	$1,050
21	12	2020	Bocina	304	$680

Comprensión

TABLA: VENTAS_2020

COLUMNAS

- Día (1-31)
- Mes (1-12)
- Año (1990-2020)
- Producto: bocinas y auriculares
- Valor (MXN)

Código SQL

```
SELECT COUNT (DISTINCT id)
FROM VENTAS_2020
WHERE Producto ='Bocina'
 AND Valor >600
 AND Año >=2019
```

Resultado

2

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Distinguir información sensible y crear un criterio ético sobre los usos de los datos.

Ética en el procesamiento de imágenes

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Estructura del problema

PROBLEMA	Algunos clientes contactan a soporte en exceso. No los podemos identificar. No podemos prevenir este comportamiento.
SOLUCIÓN	Script que identifique y clasifique a los Top Offenders. Entender sus motivaciones - clasificarlos. Definir acciones para prevenir esta tendencia.
ALCANCE	LATAM con distinción por ciudades. Clientes. Actualización mensual.

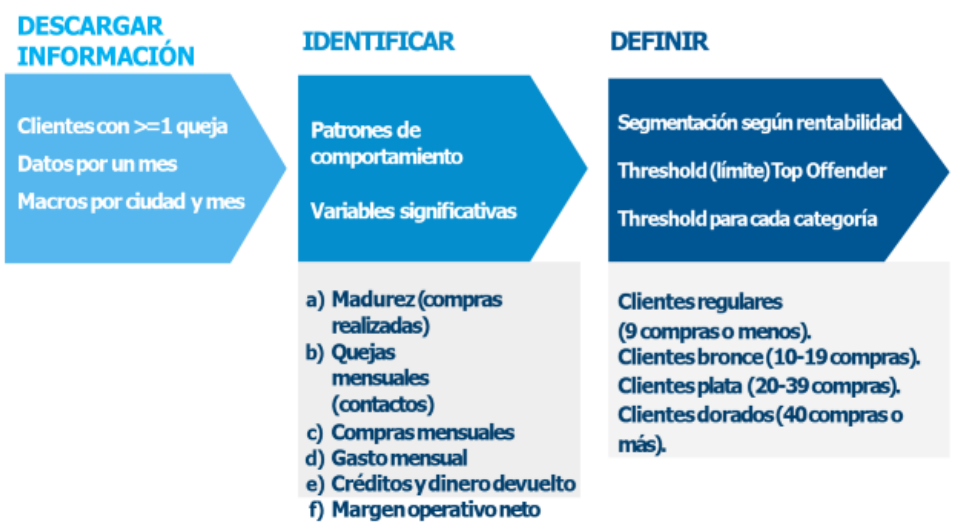
Hipótesis / Storytelling

QUÉ	POR QUÉ	CÓMO
Algunos clientes contactan a soporte en exceso.	a) Motivaciones económicas b) Preguntas c) Problemas tecnológicos d) Política de empresa	1. Análisis cuantitativo. 2. Análisis cualitativo. 3. Matriz cuantitativa - cualitativa. 4. Definir acciones de prevención. 5. Validación.

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Análisis cuantitativo en un caso de negocio

Análisis cuantitativo

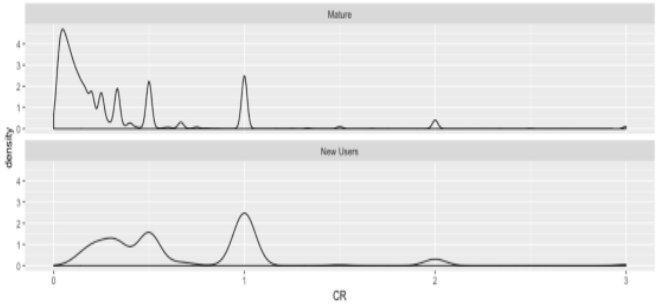


Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Análisis cuantitativo: mapeo

	Regular 280K	Bronce 56K	Plata 17K	Oro 6K	Total Activos 6M
Promedio compras	3	12	28	52	3
Promedio quejas	1.3	1.5	1.7	2	0.1
Ratio de contacto	43%	12%	6%	4%	3%

Hipótesis



	Regular 280K	Bronce 56K	Plata 17K	Oro 6K
Promedio CR (ratio contacto) Promedio ajustado CR	43% 12%*	5%	4%	3%

*Eliminar los clientes que hicieron menos de diez compras en total y 1&1 quejas vs. 1&2 compras (100% CR)

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Análisis cuantitativo: aplicación

TOP OFFENDERS				
	Regular	Bronce	Plata	Oro
Porcentaje límite	20%	5%	1%	1%
Volumen quejas	35%	17%	7%	7%
# Clientes	40K	3K	200	50

20% de las quejas las hacen estos usuarios

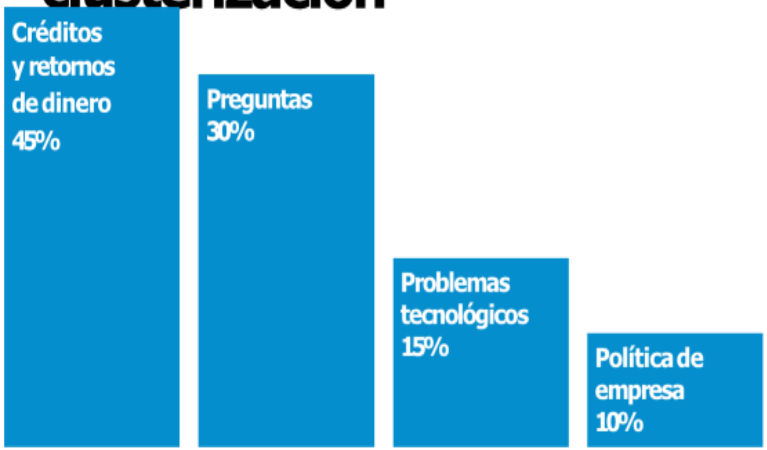
Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Identificar las variables cualitativas que nos ayudarán a resolver el ejercicio.

Análisis cualitativo

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Análisis cualitativo: clusterización



Fusión Cuanti-Cualitativa en un caso de negocio

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Matriz cuantitativa y cualitativa

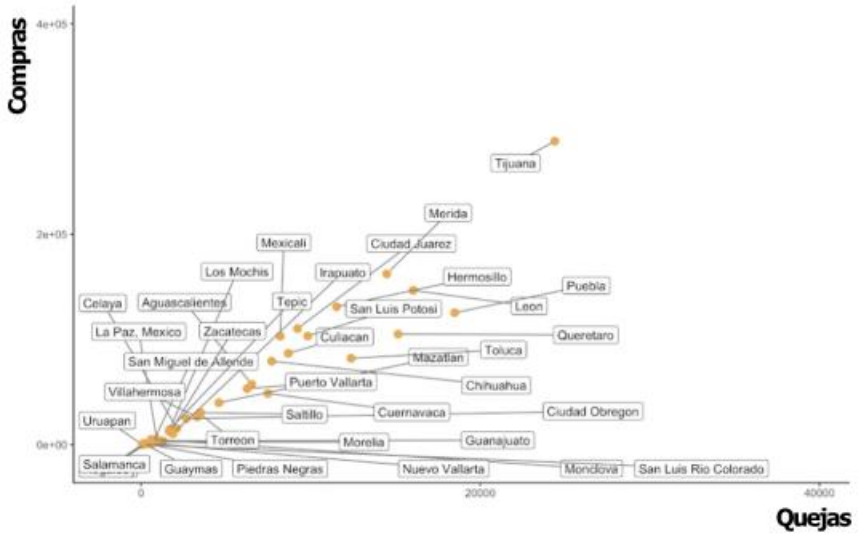
	Créditos y retornos de dinero	Preguntas	Problemas tecnológicos	Política de empresa
Regular	35%	35%	20%	10%
Bronce	30%	25%	25%	20%
Plata	25%	20%	25%	30%
Oro	20%	25%	20%	35%

Motivos de contacto

Regular	Bronce	Plata	Oro
Tarifa de devolución Tasa de envío Cómo embalar para devolución	Tarifa de devolución Tasa de envío Estado del producto	Tarifa de devolución Facturas Estado del producto	Facturas Estado del producto Log in

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

Geolocalización



Acciones derivadas del análisis

Algoritmos usados

a) Minería de datos para clasificación de motivos de contacto.
b) Correlaciones y patrones de comportamiento.
c) Árboles de decisión y teoría de juegos para predecir y tomar decisiones.
d) Validación con bayesianos y MCMC.

Acciones

a) Taggear a los Top Offenders identificados mensualmente.
b) Advertirlos.
c) Llamar usuarios.
d) Bloquear usuarios.
e) Validación con A/B Tests.

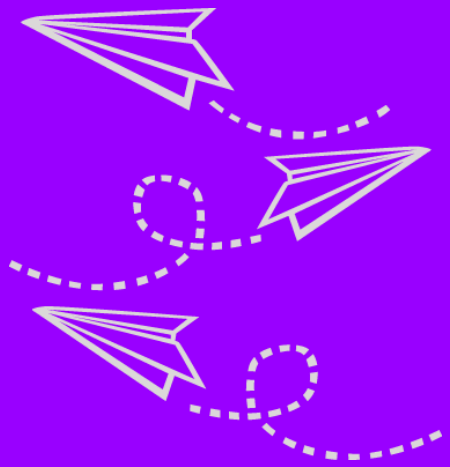
Disminuyeron las quejas en un 30% a nivel de LatAm

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)

CONCLUSIONES

- Conocer cuál es el público objetivo y qué variable se quiere predecir.
- Saber qué datos se tienen disponibles y si estarán disponibles a la hora de ejecutar el modelo.

Bibliografía: • Machine Learning Engineer NanoDegree (Udacity)



FRASE

«"Algunas personas llaman a esto inteligencia artificial, pero la realidad es que esta tecnología nos mejorará. Entonces, en lugar de inteligencia artificial, creo que aumentaremos nuestra inteligencia"».

«"Con mucha diferencia, el mayor peligro de la Inteligencia Artificial es que las personas concluyen demasiado pronto que la entienden"».

INICIO RECESO

FIN DE RECESO

FUNDACIÓN DE EDUCACIÓN SUPERIOR
SANJOSÉ
INSTITUCIÓN TECNOLÓGICA

FIN DE GRABACIÓN