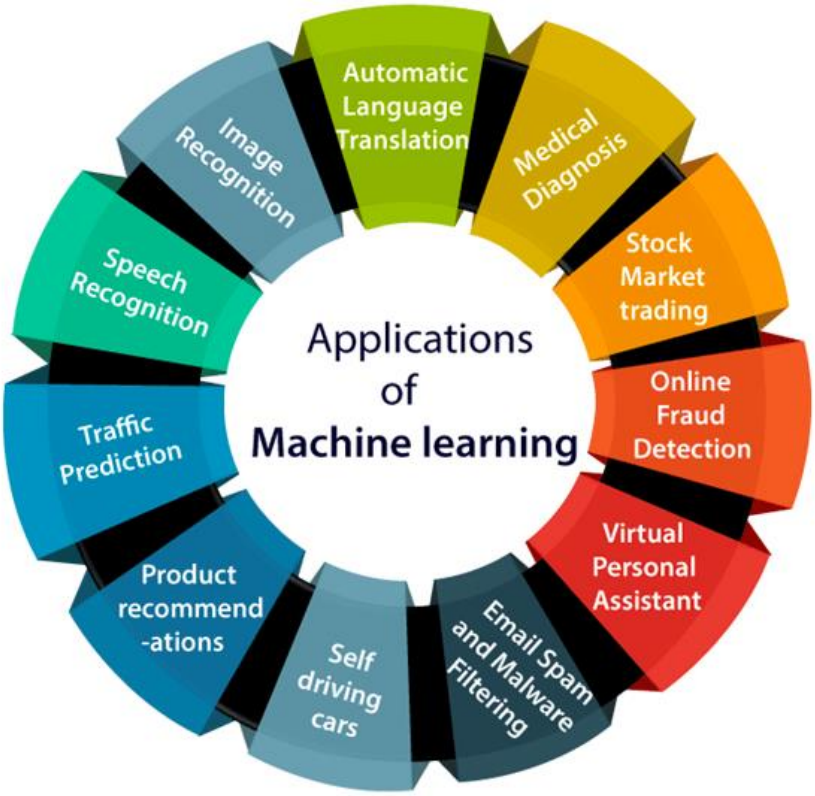
UNIVERSIDAD NACIONAL DE INGENIERÍA

OTI UNI

CIENCIA DE DATOS I



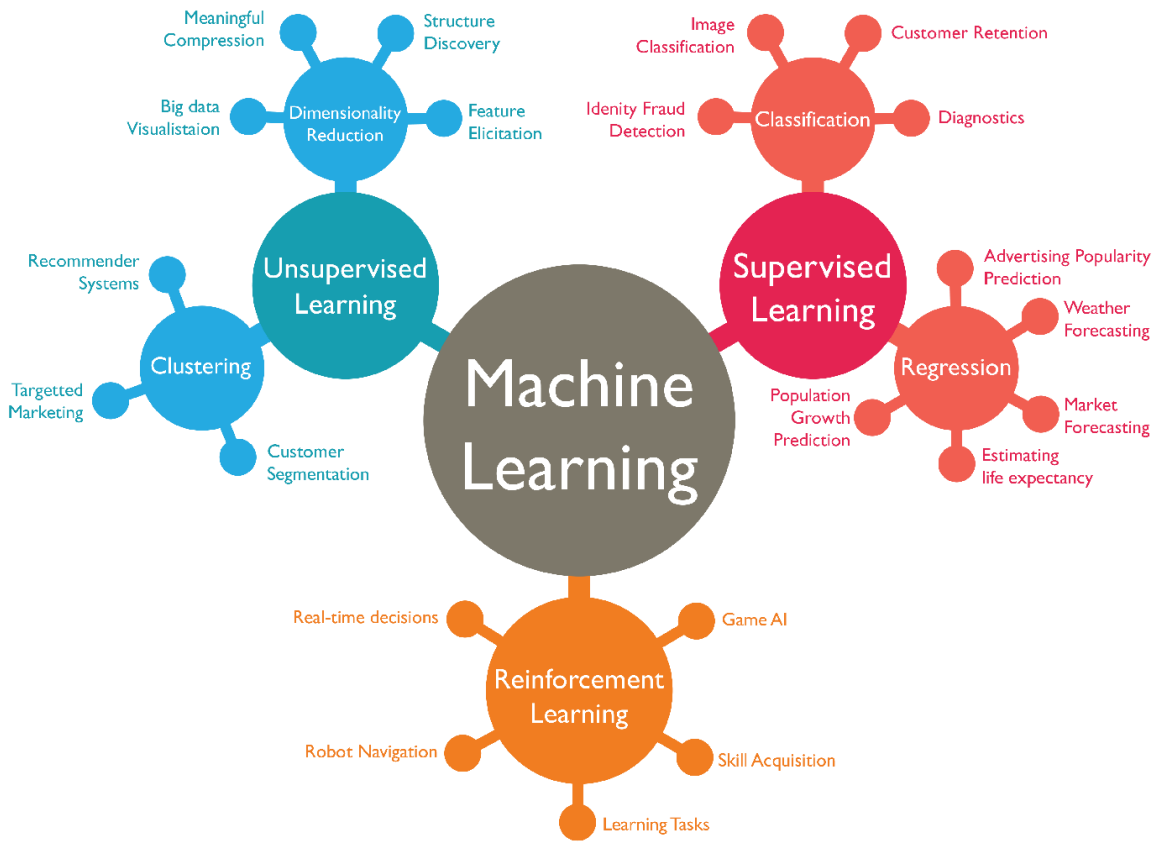
SESION 01

FUNDAMENTOS DE DATA SCIENCE Y MACHINE LEARNING

Todas las imágenes pertenecen a sus autores, se usan solo con fines educativos

ÍNDICE

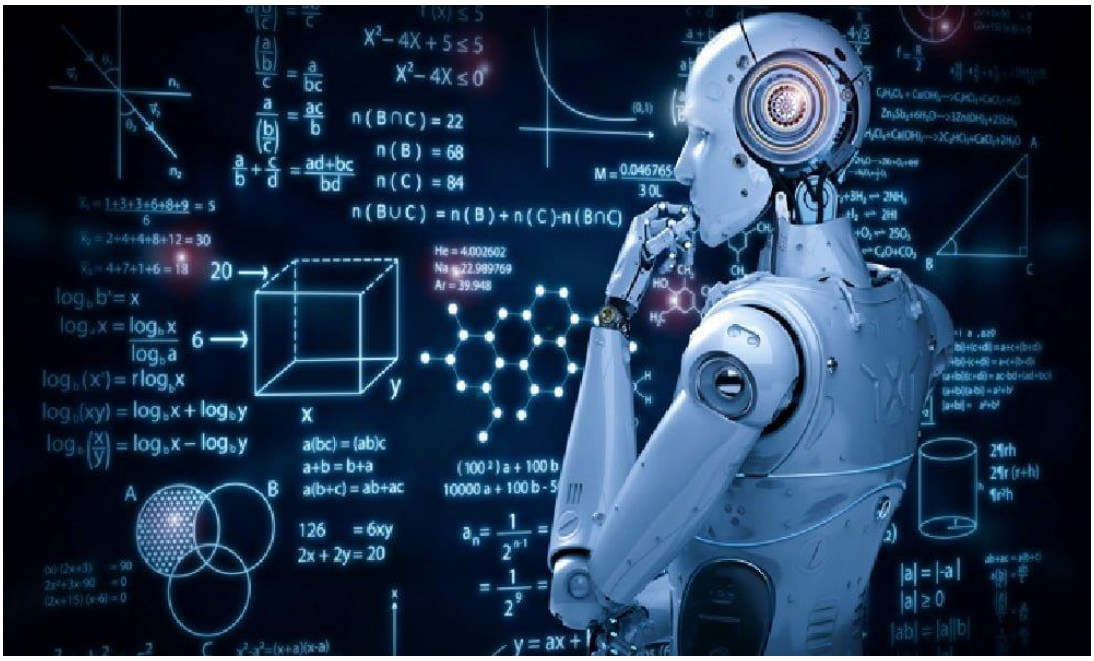
OBJETIVO



El objetivo de esta semana es conocer la terminología, principales algoritmos y áreas del Machine Learning, así como las herramientas que nos permitirán trabajar en nuestras practicas semanales del curso.

En esta primera sesión la práctica de laboratorio revisaremos una introducción a las interfaces graficas que nos facilitaran el procesamiento y visualización de nuestros resultados, así como las principales librerías de Python para nuestro curso, como son: numpy, pandas, scikit learn, matplotlib.

INTRODUCCION



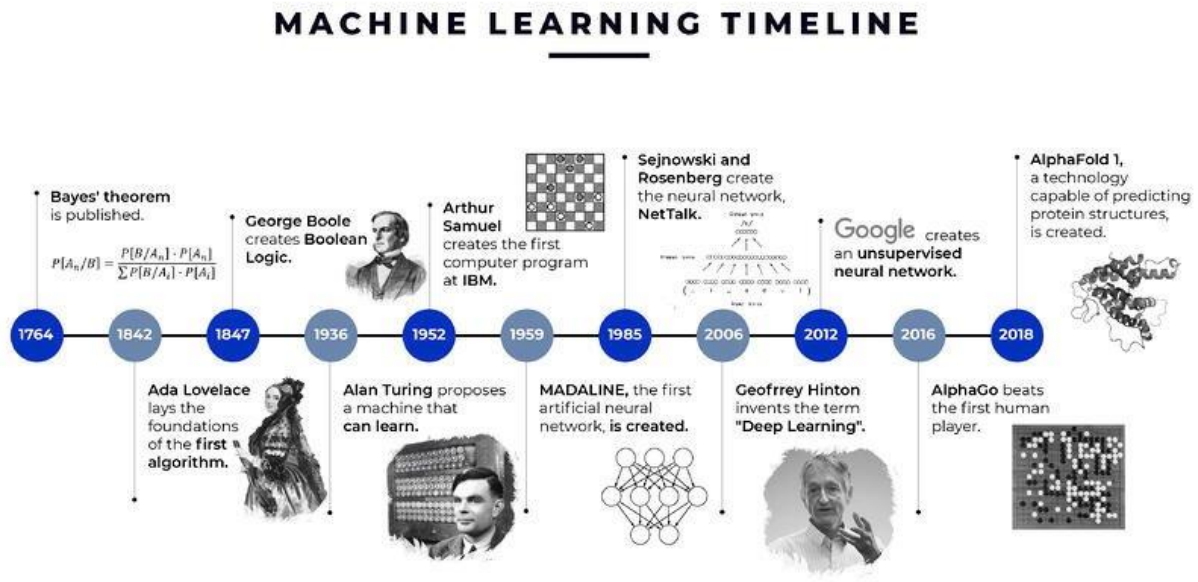
Uno de los vocablos que más se repiten en tecnología en los últimos tiempos es machine learning, o aprendizaje automático, un término que está íntimamente relacionado con la inteligencia artificial.

Brevemente se podría definir machine learning como el aprendizaje automático de los sistemas tecnológicos mediante algoritmos con el objetivo de que puedan llegar a realizar diversas acciones por su cuenta.

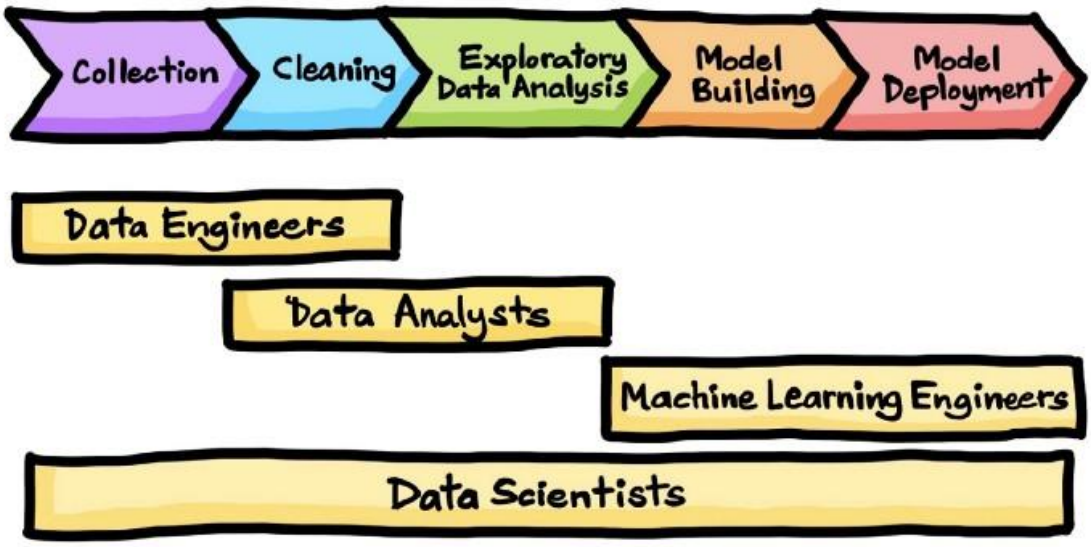
Esto que parece propio de la ciencia ficción o incluso de películas apocalípticas donde las máquinas se rebelan (véase Terminator) ya es una realidad, aunque no tan oscura como esos ejemplos cinematográficos. Es más, machine learning es una excelente noticia para mejorar procesos e impedir que las personas tengan que perder un tiempo muy valioso en realizar ciertas tareas. A fin de cuentas, que los sistemas sean capaces de aprender a partir de los datos que obtienen supervisados o sin supervisar por seres humanos supone una evolución clave para el desarrollo tecnológico durante los próximos años y décadas.

INTRODUCCIÓN

- La ciencia de datos es un campo de estudio que tiene como objetivo utilizar un enfoque científico para extraer significado e información de los datos.
- El aprendizaje automático, por otro lado, se refiere a un grupo de técnicas utilizadas por los científicos de datos que permiten que las computadoras aprendan de los datos.
- La ciencia de datos y el aprendizaje automático son palabras muy populares en la actualidad. Estos dos términos a menudo se juntan, pero no deben confundirse con sinónimos. Aunque la ciencia de datos utiliza el aprendizaje automático, estos son campos amplios con muchas herramientas diferentes.

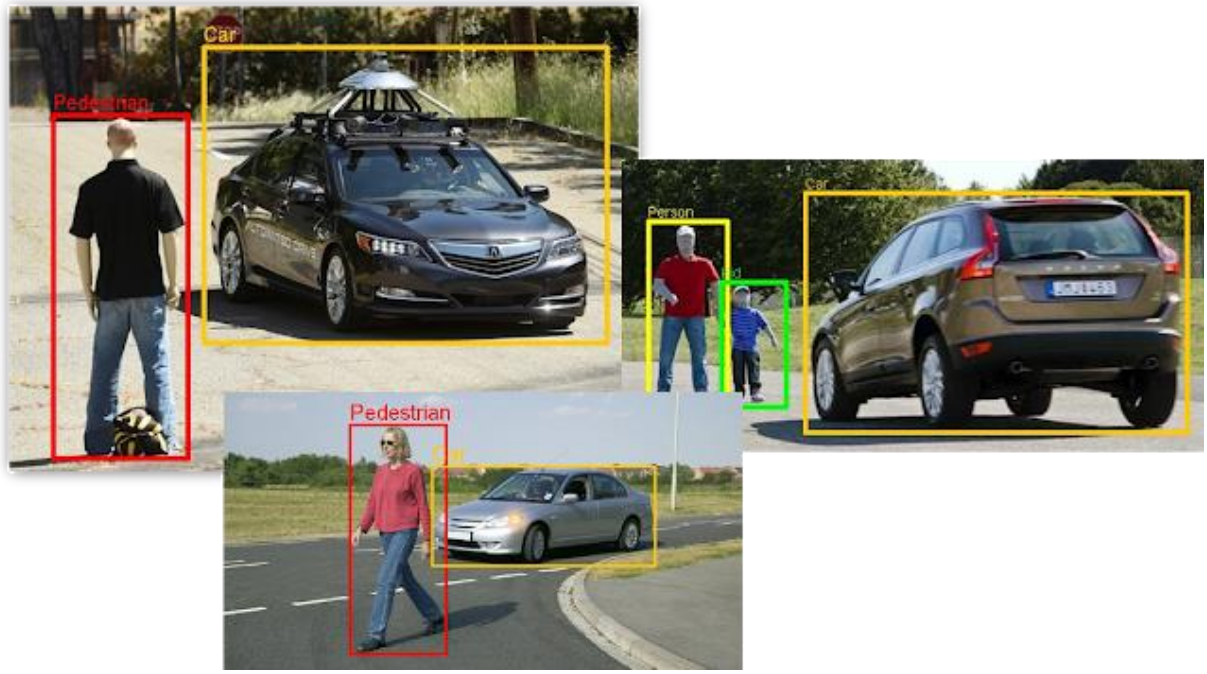


¿QUÉ ES CIENCIA DE DATOS?



- La ciencia de datos es el campo de estudio que combina la experiencia en el dominio, las habilidades de programación y el conocimiento de las matemáticas y las estadísticas para extraer información significativa de los datos.
- Los profesionales de la ciencia de datos aplican algoritmos de aprendizaje automático a números, texto, imágenes, video, audio y más para producir sistemas de inteligencia artificial (AI) con el objetivo de realizar tareas que normalmente requieren inteligencia humana.
- A su vez, estos sistemas generan conocimientos que los analistas y usuarios comerciales pueden traducir en valor comercial tangible.

DATA SCIENCE: data labeling

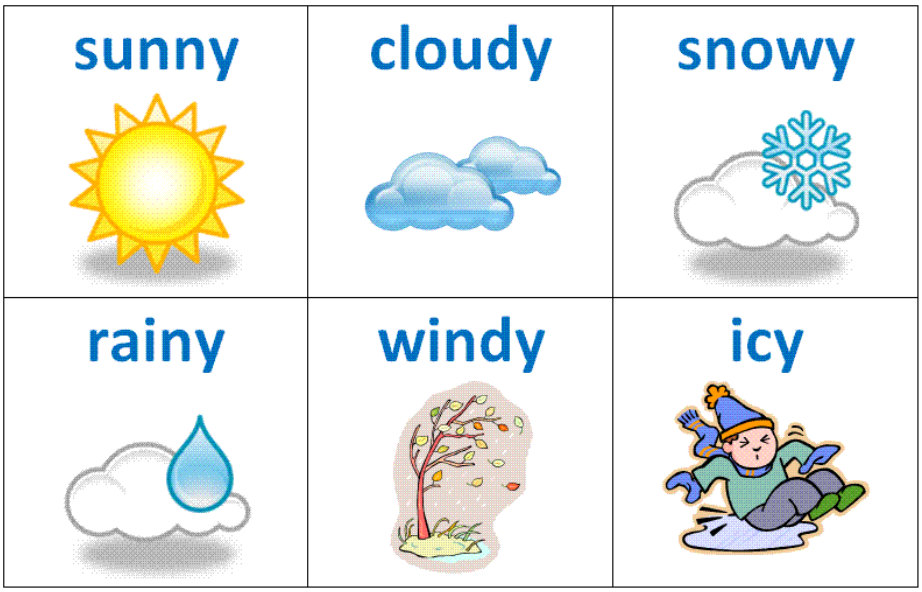


- Como sugiere el nombre, los datos etiquetados (data *labels*) son datos sin procesar (*raw data*) que hemos recopilado a los cuales les hemos agregado descripciones significativas o les hemos asignado una clase. También se les conoce como datos anotados.
- ¿Qué es una *etiqueta* en el aprendizaje automático? Supongamos que estamos construyendo un sistema de reconocimiento de imágenes y ya hemos recopilado varios miles de fotografías. Tal como vemos en la imagen superior las etiquetas le estarían diciendo a la IA que las fotos contienen una 'persona', un 'árbol', un 'automóvil', etc.
- Las funciones y etiquetas de aprendizaje automático son asignadas por expertos humanos, y el nivel de experiencia necesario puede variar. En el ejemplo anterior, no necesita personal altamente especializado para etiquetar las fotos. Sin embargo, si tiene, por ejemplo, un conjunto de radiografías y necesita entrenar la IA para buscar tumores, es probable que necesite médicos para trabajar como anotadores de datos. Naturalmente, debido a los recursos humanos necesarios, la fase del etiquetado manual de datos es mucho más costoso que la fase de recopilación de datos los cuales por lo general se encuentran sin etiquetar.

DATA SCIENCE: datos categóricos

- Cuando recopilamos datos para una investigación, es importante conocer la forma de sus datos para poder interpretarlos y analizarlos de manera efectiva. Existen principalmente dos tipos de datos: datos categóricos y datos numéricos.
- **Datos Categóricos,** Los datos categóricos se refieren a un tipo de datos que se pueden almacenar e identificar en función de los nombres o etiquetas que se les asignan. Se realiza un proceso llamado coincidencia, para extraer las similitudes o relaciones entre los datos y luego se agrupan en consecuencia.

Los datos recopilados en forma categórica también se conocen como datos cualitativos. Cada conjunto de datos se puede agrupar y etiquetar según sus cualidades coincidentes, en una sola categoría. Esto hace que las categorías sean mutuamente excluyentes.



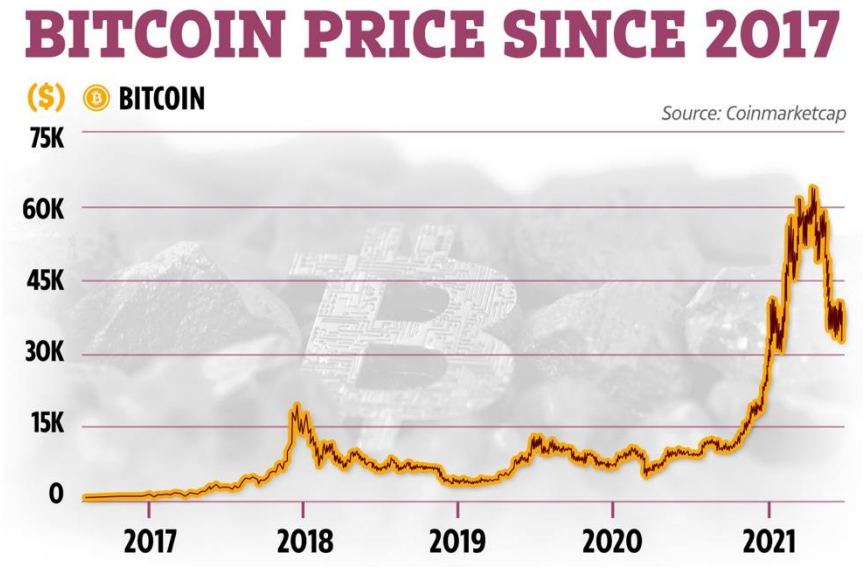
Hay dos subtipos de datos categóricos, a saber: datos nominales y datos ordinales.

Datos nominales: también se denominan datos de nombres. Este es un tipo que nombra o etiqueta los datos y sus características son similares a un sustantivo. Ejemplo: nombre de la persona, género, nombre de la escuela.

Datos ordinales: esto incluye datos o elementos de datos que se clasifican, ordenan o utilizan en una escala de calificación. Puedes contar y ordenar datos ordinales pero no te permite medirlos. Ejemplo: Calificar el resultado de un seminario entre 1 y 5

DATA SCIENCE: datos numéricos

- **Datos Numéricos,** los datos numéricos se refieren a los datos que están en forma de números, y no en ningún idioma o forma descriptiva. A menudo denominados datos cuantitativos, los datos numéricos se recopilan en forma de números y se diferencian de cualquier forma de tipos de datos numéricos debido a su capacidad para calcularse estadística y aritméticamente.



También tiene dos subtipos conocidos como datos discretos y datos continuos.

Datos discretos, los datos discretos se utilizan para representar elementos que se pueden contar. Puede tomar formas tanto numéricas como categóricas y agruparlas en una lista. Esta lista puede ser finita o infinita también.
Los datos discretos básicamente toman números contables como 1, 2, 3, 4, 5, etc. En el caso del infinito, estos números continuarán.

Ejemplo: días de la semana, días de meses, calificaciones de una prueba, talla de zapatos.

Datos continuos, son un tipo de datos cuantitativos que se pueden medir. Los datos numéricos continuos representan medidas y sus intervalos caen en una recta numérica.

Ejemplo: temperatura, humedad, viscosidad.

CICLO DE VIDA DE LA CIENCIA DE DATOS

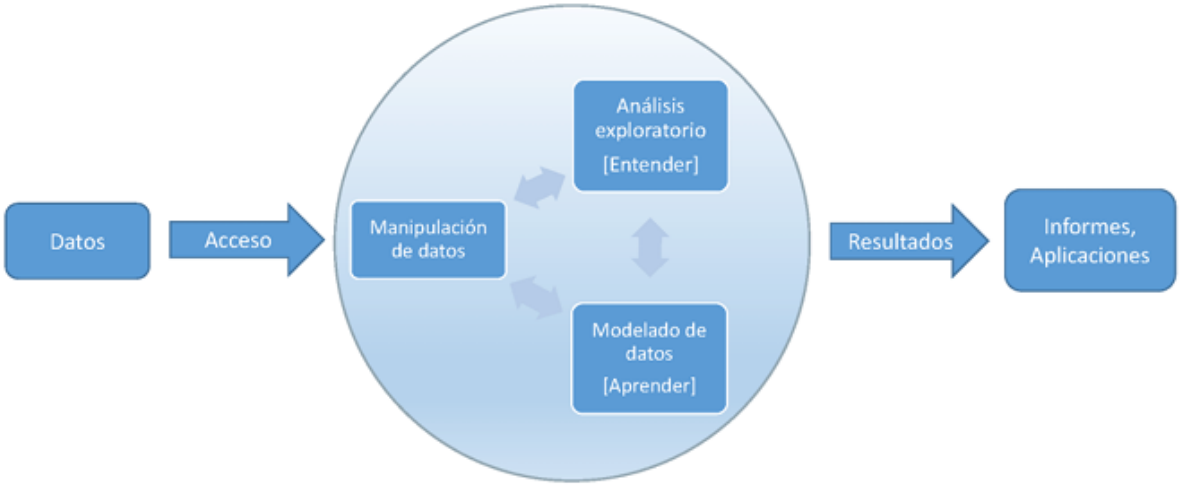
El ciclo de vida de la ciencia de datos se compone esencialmente de:

- **Entender el negocio**, es el punto de partida en el ciclo de vida. Por lo tanto, es importante comprender cuál es la declaración del problema y hacer las preguntas correctas al cliente que nos ayuden a comprender bien los datos y obtener información significativa de los datos.
- **Recolección de datos,** el paso principal en el ciclo de vida de los proyectos de ciencia de datos es identificar primero a la persona o personas que sabe qué datos adquirir y cuándo adquirirlos en función de la pregunta a responder. No es necesario que la persona sea un científico de datos, pero cualquiera que conozca la diferencia real entre los diversos conjuntos de datos disponibles y tome decisiones contundentes sobre la estrategia de inversión de datos de una organización, será la persona adecuada para el trabajo.
- **Limpieza de datos,** en este paso, comprendemos más acerca de los datos y los preparamos para un análisis posterior. La sección de comprensión de datos de la metodología de ciencia de datos responde a la pregunta: ¿Son los datos que recopiló representativos del problema a resolver?
- **Análisis de datos**, el análisis exploratorio a menudo se describe como una filosofía, y no hay reglas fijas sobre cómo abordarlo. No hay atajos para la exploración de datos.

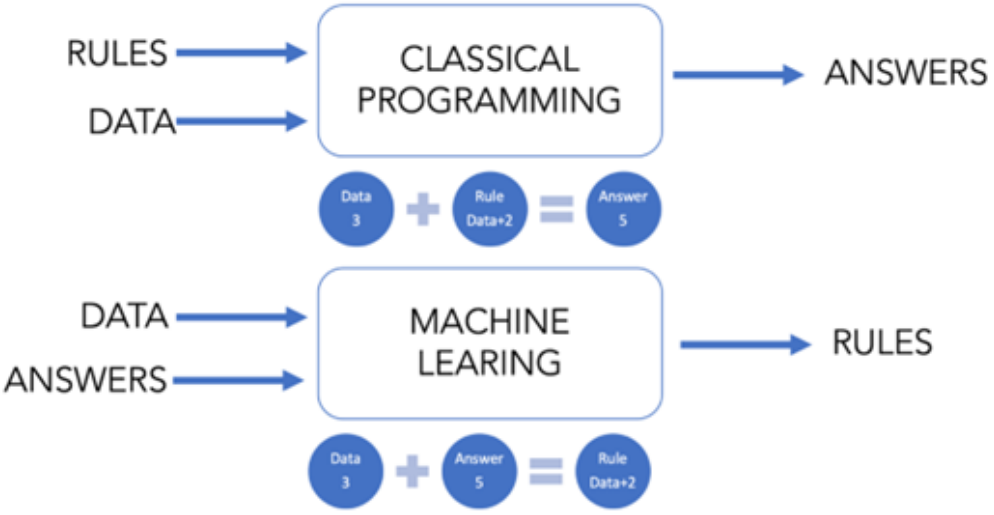
  Recuerde que la calidad de sus entradas decide la calidad de su salida. Por lo tanto, una vez que tenga lista su hipótesis comercial, tiene sentido dedicar mucho tiempo y esfuerzo aquí.

CICLO DE VIDA DE LA CIENCIA DE DATOS

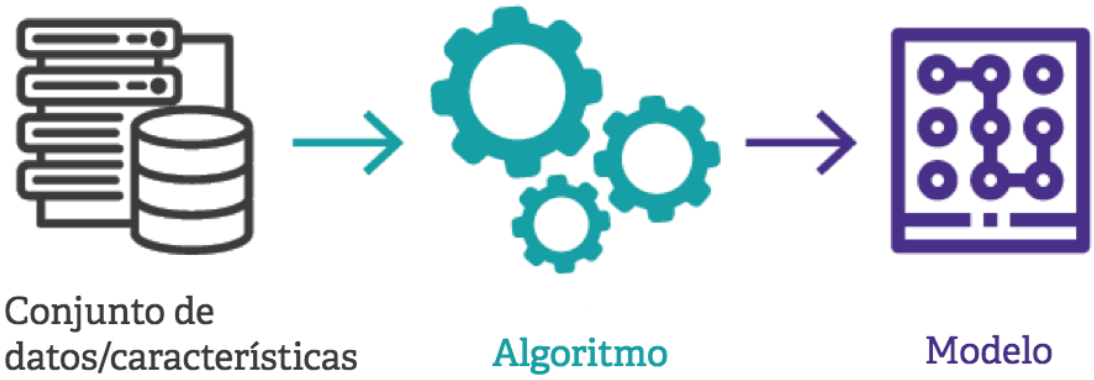
- **Modelamiento de datos, modelamiento machine learning**, esta etapa parece ser la más interesante para casi todos los científicos de datos. Mucha gente lo llama “un escenario donde ocurre la magia”. Pero recordemos que la magia solo puede suceder si tienes los accesorios y la técnica correctos. En términos de ciencia de datos, "Datos" es ese apoyo, y la preparación de datos es esa técnica. Entonces, antes de saltar a este paso, asegúrese de pasar suficiente tiempo en los pasos anteriores. El modelado se utiliza para encontrar patrones o comportamientos en los datos. Acá es donde encaja el Machine Learning.
- **Evaluación del modelo**, una pregunta común que los profesionales suelen tener al evaluar el rendimiento de un modelo de aprendizaje automático en qué conjunto de datos debe usar para medir el rendimiento del modelo de aprendizaje automático. Mirar las métricas de rendimiento en el conjunto de datos entrenado es útil, pero no siempre es correcto porque los números obtenidos pueden ser demasiado optimistas, ya que el modelo ya está adaptado al conjunto de datos de entrenamiento. El rendimiento del modelo de aprendizaje automático debe medirse y compararse mediante conjuntos de validación y prueba para identificar el mejor modelo en función de la precisión y el sobreajuste del modelo.
- **Visualización y reportes**, en este proceso, las habilidades técnicas por sí solas no son suficientes. Una habilidad esencial que necesita es poder contar una historia clara y procesable. Si su presentación no desencadena acciones en su audiencia, significa que su comunicación no fue eficiente. Debe estar en consonancia con las cuestiones comerciales. Debe ser significativo para la organización y las partes interesadas. La presentación a través de la visualización debe ser tal que desencadene la acción en la audiencia. Recuerde que se presentará a una audiencia sin conocimientos técnicos, por lo que la forma en que comunica el mensaje es clave.
- **Despliegue de modelo**, después de construir modelos, primero se implementa en un entorno de preproducción o prueba antes de implementarlos en producción. Cualquiera que sea la forma en que se implemente su modelo de datos, debe exponerse al mundo real. Una vez que los humanos reales lo usen, seguramente recibirás comentarios. Capturar esta retroalimentación se traduce directamente en la vida o la muerte para cualquier proyecto.


Datos
Acceso
Análisis exploratorio
[Entender]
Manipulación de datos
Modelado de datos
[Aprender]
Resultados
Informes, Aplicaciones

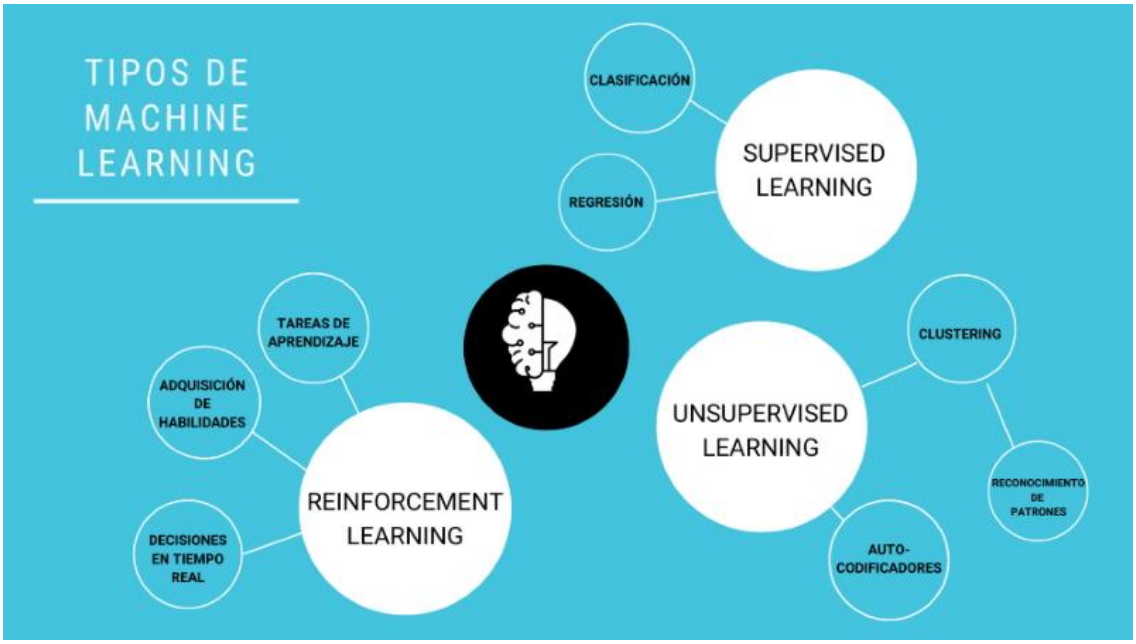
¿QUÉ ES EL MACHINE LEARNING?



- Es una rama de la inteligencia artificial. El termino se usa desde 1959.
- Es la capacidad de las máquinas para aprender a partir de los datos de manera automatizada.
- Al aprender de manera automatizada, esto implica que no necesitan ser programadas para dicha tarea.
- Esto último es una habilidad indispensable para construir sistemas capaces de identificar patrones entre los datos para hacer predicciones de manera eficiente y confiable.
- El aprendizaje automático es excelente para resolver problemas que requieren mucho trabajo para los humanos, mucho procesamiento de datos.

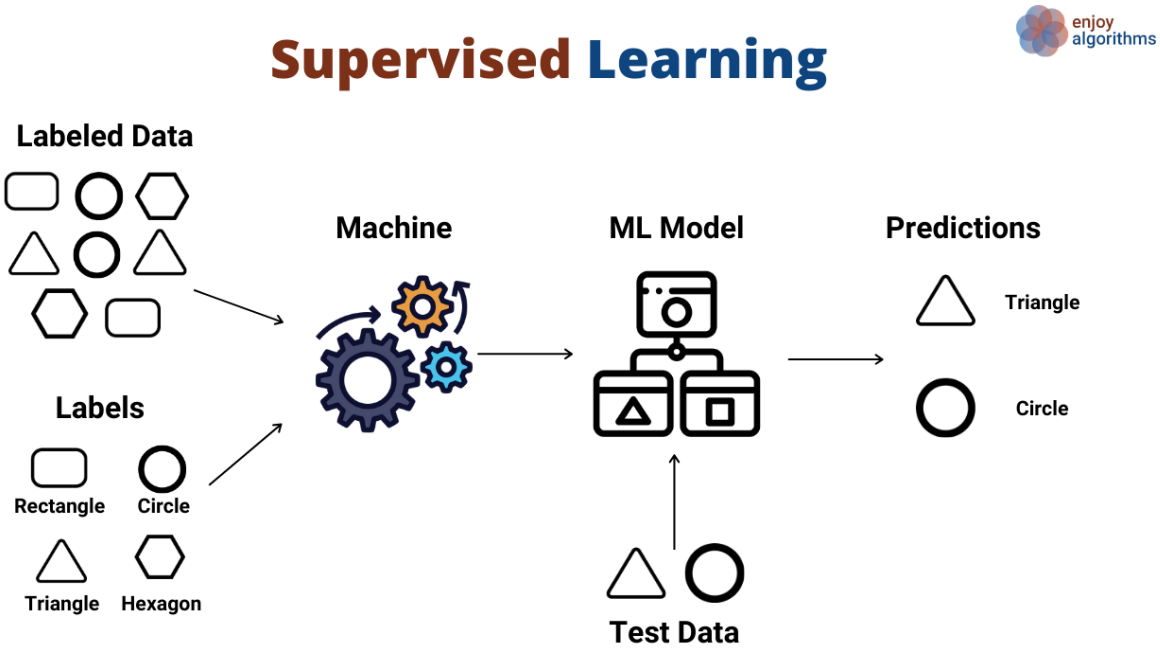


MACHINE LEARNING: construcción del modelo

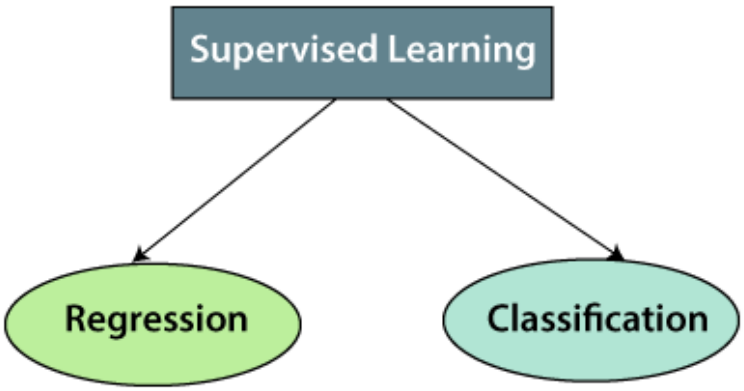


- **¿Qué es un modelo?** Un modelo es simplemente un sistema para mapear entradas a salidas. Por ejemplo, si queremos predecir los precios de las casas, podríamos hacer un modelo que tome los pies cuadrados de una casa y genere un precio. Un modelo representa una teoría sobre un problema: hay alguna conexión entre los pies cuadrados y el precio y hacemos un modelo para aprender esa relación. Los modelos son útiles porque podemos usarlos para predecir los valores de las salidas para nuevos puntos de datos dadas las entradas.
- El modelado se utiliza para encontrar patrones o comportamientos en los datos. Estos patrones pueden ser extraídos dependiendo de las necesidades del problema, el ambiente en el que se van a desenvolver y los factores que afectarán la toma de decisiones, utilizando algunos tipos de algoritmos de aprendizaje, entre los cuales vamos a hablar de 3 de ellos: supervisado, no supervisado y por refuerzo.

MACHINE LEARNING: aprendizaje supervisado



- El aprendizaje supervisado es el tipo de aprendizaje automático en el que las máquinas se entrenan utilizando datos de entrenamiento bien "etiquetados" y, sobre la base de esos datos, las máquinas predicen el resultado. Los datos "etiquetados" significan que algunos datos de entrada ya están marcados con la salida correcta.
- Los algoritmos supervisados pueden dividirse en dos tipos de problemas:



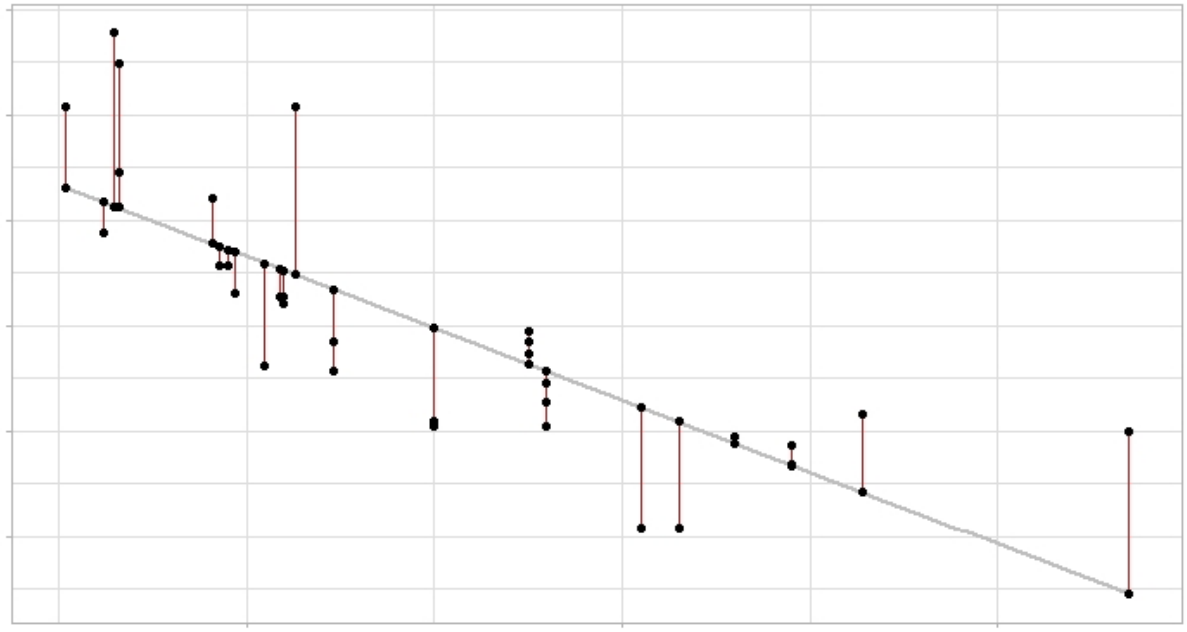
MACHINE LEARNING: aprendizaje supervisado

Algunos algoritmos supervisados

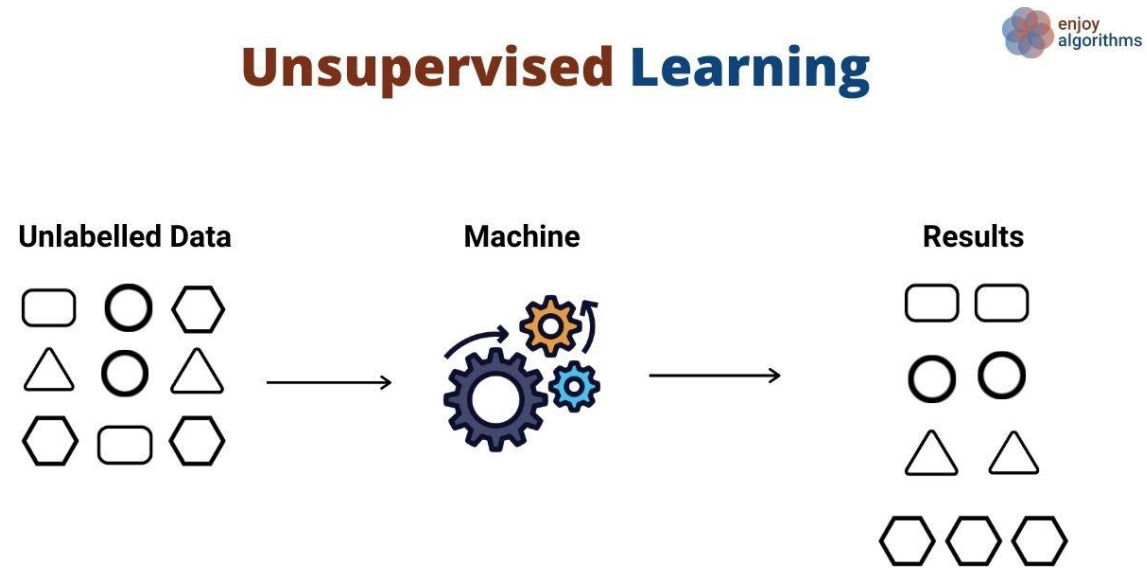
1. Naive Bayes (Clasificacion, modelo no lineal)
2. Neural Networks
3. k-Nearest Neighbor (kNN) (Clasificacion, modelo no lineal)
4. Linear Regression (Regresion)
5. Logistic Regression (Clasificacion, modelo lineal)
6. Support Vector Machines(SVM) (Clasificacion, modelo lineal)
7. Decision Trees (Clasificacion, modelo no lineal)
8. Random Forest (Clasificacion, modelo no lineal)

Ejemplo

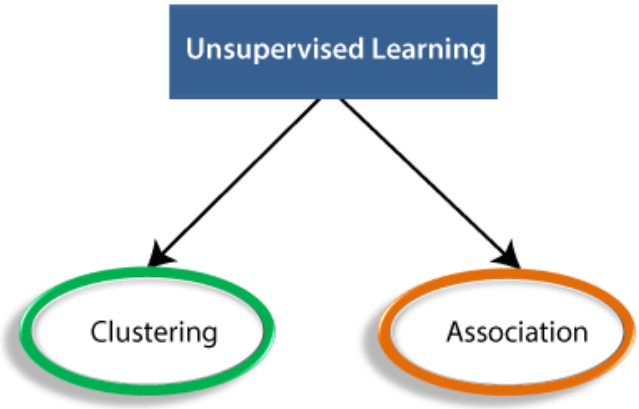
Regresión Lineal



MACHINE LEARNING: aprendizaje no supervisado



- Como sugiere el nombre, el aprendizaje no supervisado es una técnica de aprendizaje automático en la que los modelos no se supervisan mediante un conjunto de datos de entrenamiento. En cambio, los propios modelos encuentran los patrones ocultos y los conocimientos de los datos proporcionados. Se puede comparar con el aprendizaje que tiene lugar en el cerebro humano mientras aprende cosas nuevas.
- Los algoritmos no supervisados pueden dividirse en dos tipos de problemas:



MACHINE LEARNING: aprendizaje no supervisado

- El objetivo del aprendizaje no supervisado puede ser descubrir grupos de ejemplos similares dentro de los datos, lo que se denomina agrupación (*clustering*), o determinar la distribución de datos dentro del espacio de entrada, conocido como *estimación de densidad*. La estimación de densidad y agrupamiento se puede elegir para conocer los patrones en los datos. Proyectar los datos desde un espacio multidimensional hasta dos o tres dimensiones se elegirá con el propósito de poder visualizar los mismos.

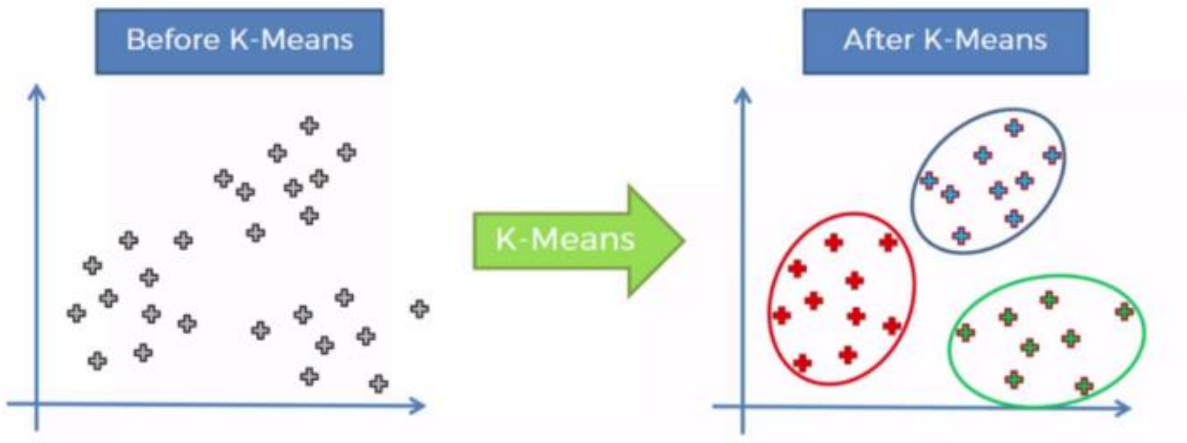
Algunos algoritmos no supervisados

1. Principle Component Analysis (PCA)
2. KMeans/Kmeans++
3. Hierarchical Clustering
4. DBSCAN
5. Market Basket Analysis

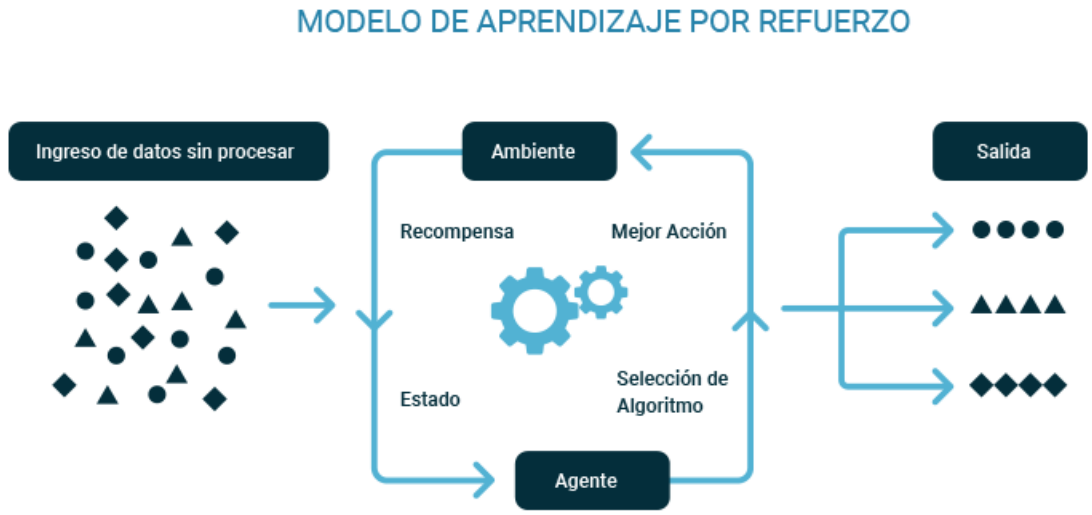
Ejemplo

Ejemplo de clustering con k-means en Python.

http://exponentis.es/ejemplo-de-clustering-con-k-means-en-python



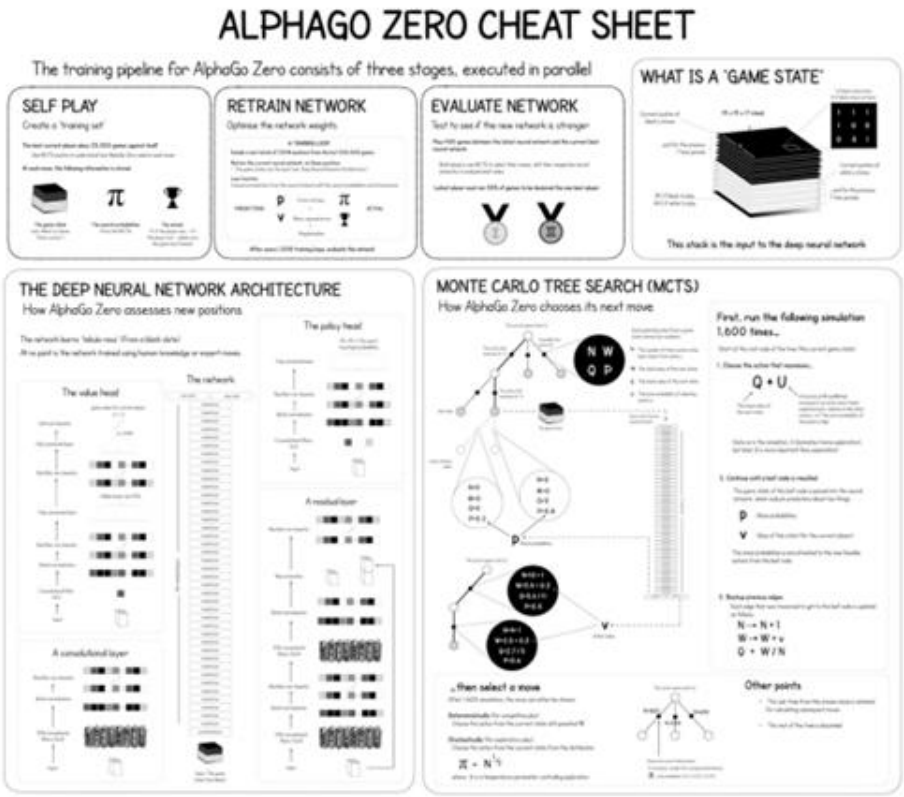
MACHINE LEARNING: aprendizaje reforzado



- RL es una aplicación especializada de técnicas de aprendizaje automático/profundo, diseñadas para resolver problemas de una manera particular. A diferencia del aprendizaje supervisado y no supervisado, el aprendizaje por refuerzo es un tipo de aprendizaje que se basa en la interacción con los entornos. Es decir, los algoritmos aprenden a reaccionar ante un entorno por sí mismos. Por lo tanto, la mayor parte de RL es el proceso de prueba y error.
- Los modelos RL consisten en algoritmos que utilizan los errores estimados como recompensas o penalizaciones. Si el error es grande, entonces la sanción es alta y la recompensa baja. Si el error es pequeño, la penalización es baja y la recompensa alta. La Figura es una ilustración simple de RL. La forma en que el aprendizaje por refuerzo resuelve problemas es permitiendo que una pieza de software llamada "agente" explore, interactúe y aprenda del entorno.

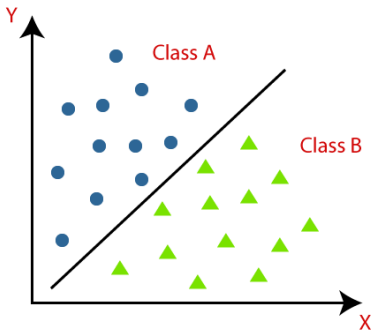
MACHINE LEARNING: aprendizaje reforzado

- El Go es un juego de mesa tradicional chino con más de 2500 años de antigüedad. Se trata de un juego para 2 personas que, por turnos, van colocando piezas blancas y negras en un tablero estándar de 19×19. El objetivo es capturar las piezas del oponente, eliminándolas así del tablero, o rodear espacios vacíos para hacer puntos de territorio.



APRENDIZAJE SUPERVISADO: Algoritmo de Clasificacion

- El algoritmo de clasificación es una técnica de aprendizaje supervisado que se utiliza para identificar la categoría de nuevas observaciones sobre la base de datos de entrenamiento. En Clasificación, un programa aprende del conjunto de datos u observaciones dado y luego clasifica la nueva observación en una serie de clases o grupos. Por ejemplo, Sí o No, 0 o 1, Spam o No Spam, gato o perro, etc. Las clases se pueden denominar como objetivos/etiquetas o categorías.
- A diferencia de la regresión, la variable de salida de Clasificación es una categoría, no un valor, por ejemplo "Verde o Azul", "fruta o animal", etc. Dado que el algoritmo de Clasificación es una técnica de aprendizaje supervisado, toma datos de entrada etiquetados, que significa que contiene entrada con la salida correspondiente.



- El algoritmo que implementa la clasificación en un conjunto de datos se conoce como clasificador. Hay dos tipos de Clasificaciones:

  Clasificador binario: si el problema de clasificación tiene solo dos resultados posibles, se denomina clasificador binario.

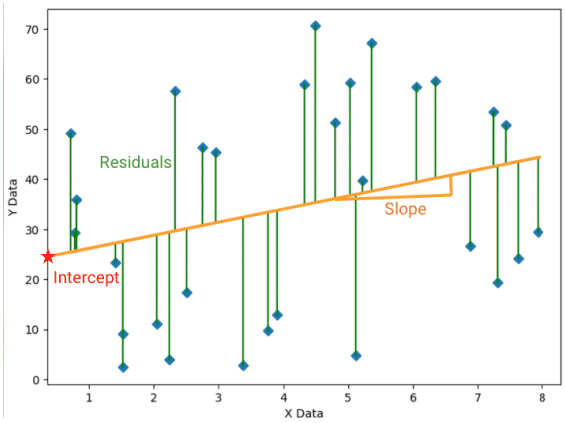
  Ejemplos: SI o NO, MASCULINO o FEMENINO, SPAM o NO SPAM, GATO o PERRO, etc.

  Clasificador multiclase: si un problema de clasificación tiene más de dos resultados, se denomina clasificador multiclase.

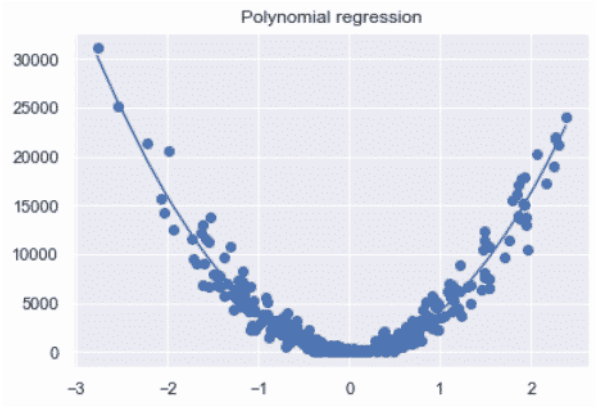
  Ejemplo: Clasificaciones de tipos de cultivos, Clasificación de tipos de música.

APRENDIZAJE SUPERVISADO: Algoritmo de Regresion

- Los algoritmos de regresión intentan estimar la función de mapeo (f) a partir de las variables de entrada (x) a variables de salida numéricas o continuas (y). Ahora, la variable de salida podría ser un valor real, que puede ser un número entero o un valor de coma flotante. Por lo tanto, los problemas de predicción de regresión suelen ser cantidades o tamaños.
- Por ejemplo, si se le proporciona un conjunto de datos sobre casas y se le pide que prediga sus precios, se trata de una tarea de regresión porque el precio será una salida continua.
- Los ejemplos de los algoritmos de regresión comunes incluyen la regresión lineal, la regresión de vectores de soporte (SVR) y los árboles de regresión.
- El *análisis de regresión* trata de ajustar una línea (o curva) en un gráfico de dispersión de dos variables continuas de manera que los puntos de datos se encuentren colectivamente lo más cerca posible de la línea.
- A continuación se muestra un ejemplo de una regresión lineal en la que la intersección y la pendiente de la línea se colocan de manera que se minimiza la suma de los residuos.



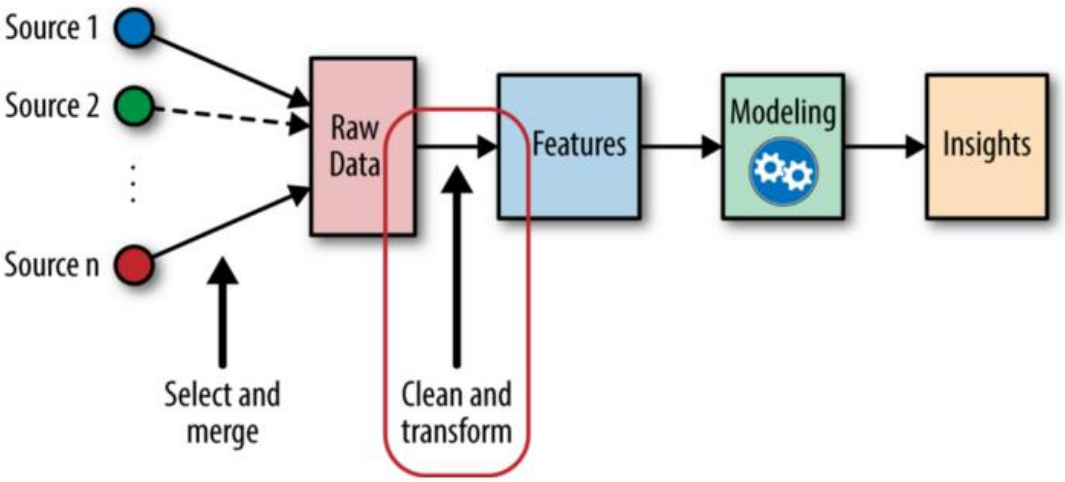
- Ejemplo de regresion polinomial:



DS LIFE CYCLE: análisis exploratorio de datos

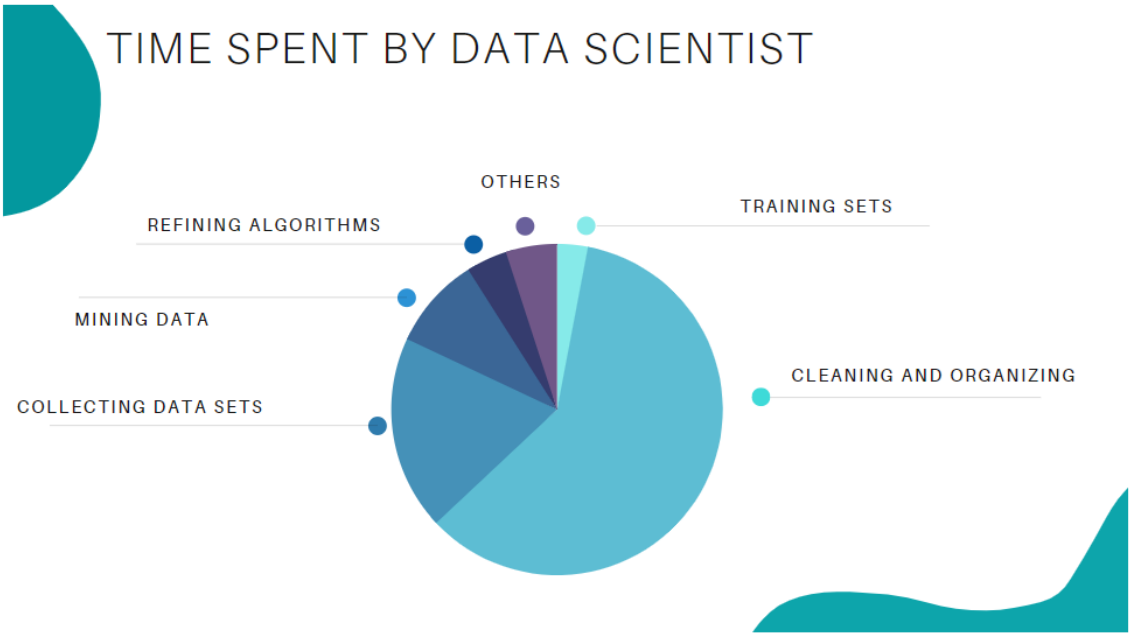
- Antes de realizar análisis de datos, con fines estadísticos o predictivos, usando por ejemplo técnicas de aprendizaje automático, es necesario entender la materia prima (raw data) con la que vamos a trabajar. Es necesario comprender y evaluar la calidad de los datos para, entre otros aspectos, detectar y tratar los datos atípicos (outliers) o incorrectos, evitando posibles errores que puedan repercutir en los resultados del análisis.
- EDA consiste en aplicar un conjunto de técnicas estadísticas destinadas a explorar, describir y resumir la naturaleza de los datos, de forma que podamos entender claramente como están relacionadas nuestras variables de interés.
- Todo esto nos permite identificar posibles errores, revelar la presencia de outliers, comprobar la relación entre variables (correlaciones) y su posible redundancia, y realizar un análisis descriptivo de los datos mediante representaciones gráficas y resúmenes de los aspectos más significativos.

DS LIFE CYCLE: feature engineering



- La ingeniería de variables es el proceso de seleccionar, manipular y transformar datos sin procesar en características que se pueden usar en el aprendizaje supervisado. Para que el aprendizaje automático funcione bien en tareas nuevas, puede ser necesario diseñar y entrenar mejores características. Como sabrá, una "variable" es cualquier entrada medible que se puede usar en un modelo predictivo; podría ser el color de un objeto o el sonido de la voz de alguien. La ingeniería de variables, en términos simples, es el acto de convertir observaciones sin procesar en características deseadas utilizando enfoques estadísticos o de aprendizaje automático.
- La ingeniería de variables es una técnica de aprendizaje automático que aprovecha los datos para crear nuevas variables que no están en el conjunto de entrenamiento. Puede producir nuevas funciones para el aprendizaje supervisado y no supervisado, con el objetivo de simplificar y acelerar las transformaciones de datos y, al mismo tiempo, mejorar la precisión del modelo. Se requiere ingeniería de funciones cuando se trabaja con modelos de aprendizaje automático. Independientemente de los datos o la arquitectura, una característica terrible tendrá un impacto directo en su modelo.

DS LIFE CYCLE: feature engineering



La ingeniería de variables es un paso muy importante en el aprendizaje automático. La ingeniería de variables se refiere al proceso de diseñar características artificiales en un algoritmo. Estas características artificiales son luego utilizadas por ese algoritmo para mejorar su rendimiento o, en otras palabras, obtener mejores resultados. Los científicos de datos pasan la mayor parte de su tiempo con los datos, y se vuelve importante hacer que los modelos sean precisos.

DS LIFE CYCLE: feature engineering

- Ahora, para entenderlo de una manera mucho más fácil, tomemos un ejemplo simple. A continuación, se muestran los precios de las propiedades en una ciudad x. Muestra el área de la casa y el precio total.

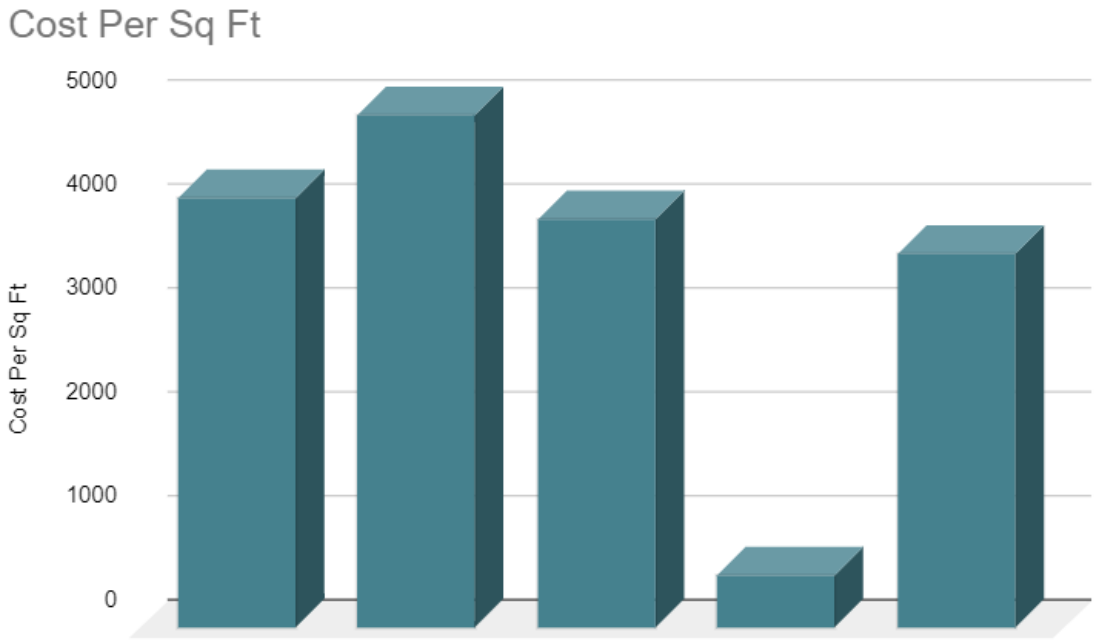
Sq Ft.	Amount
2400	9 Million
3200	15 Million
2500	10 Million
2100	1.5 Million
2500	8.9 Million

- Ahora bien, estos datos pueden tener algunos errores o pueden ser incorrectos, no todas las fuentes en Internet son correctas. Para comenzar, agregaremos una nueva columna para mostrar el costo por pie cuadrado.

Sq Ft.	Amount	Cost Per Sq Ft
2400	9 Million	4150
3200	15 Million	4944
2500	10 Million	3950
2100	1.5 Million	510
2500	8.9 Million	3600

- Esta nueva característica nos ayudará a entender mucho sobre nuestros datos. Entonces, tenemos una nueva columna que muestra el costo por pie cuadrado. Hay tres formas principales de encontrar cualquier error. Puede ponerse en contacto con un asesor inmobiliario o agente de bienes raíces y mostrarle la tarifa por pie cuadrado. Si su abogado afirma que el precio por pie cuadrado no puede ser inferior a 3400, es posible que tenga un problema. Los datos se pueden visualizar.

DS LIFE CYCLE: feature engineering

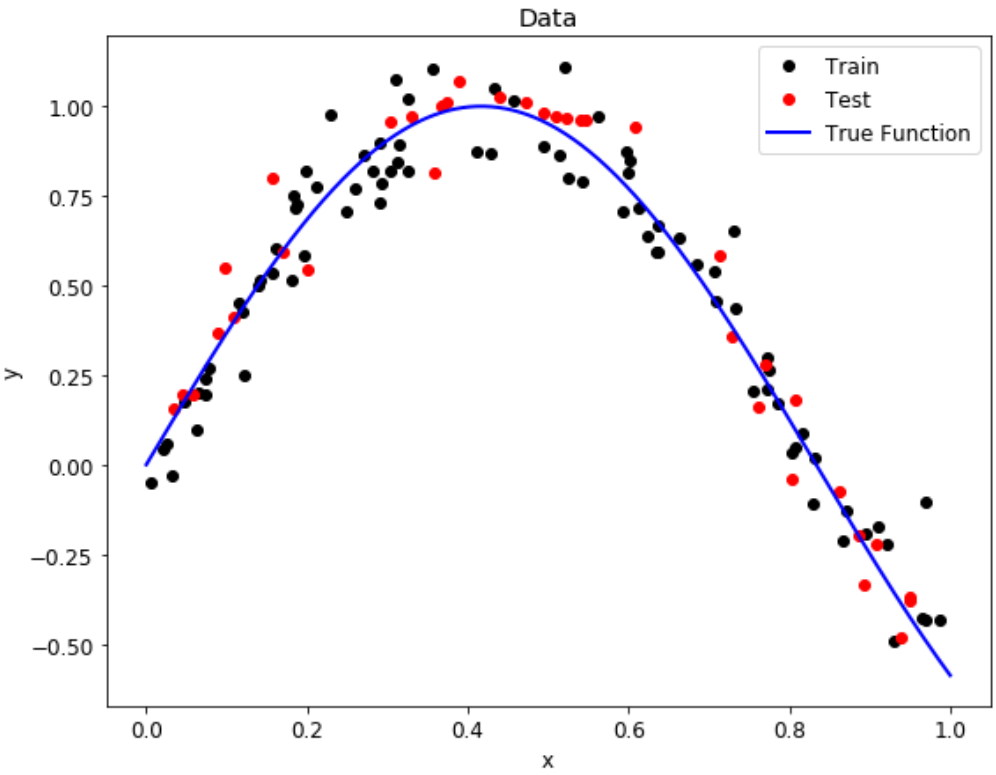


Cuando graficamos los datos, notaremos que un precio es significativamente diferente del resto. En el método de visualización, puede notar fácilmente el problema. La tercera forma es usar Estadísticas para analizar sus datos y encontrar cualquier problema. La ingeniería de características consta de varios procesos:

- Feature Creation
- Transformations
- Feature Extraction
- Exploratory Data Analysis
- Benchmark

DS LIFE CYCLE: model building

- Para hacer un modelo, primero necesitamos datos que tengan una relación subyacente. Para este ejemplo, crearemos nuestro propio conjunto de datos simple con valores x (características) y valores y (etiquetas). Una parte importante de nuestra generación de datos es agregar ruido aleatorio a las etiquetas. En cualquier proceso del mundo real, ya sea natural o artificial, los datos no se ajustan exactamente a una tendencia. Siempre hay ruido u otras variables en la relación que no podemos medir.



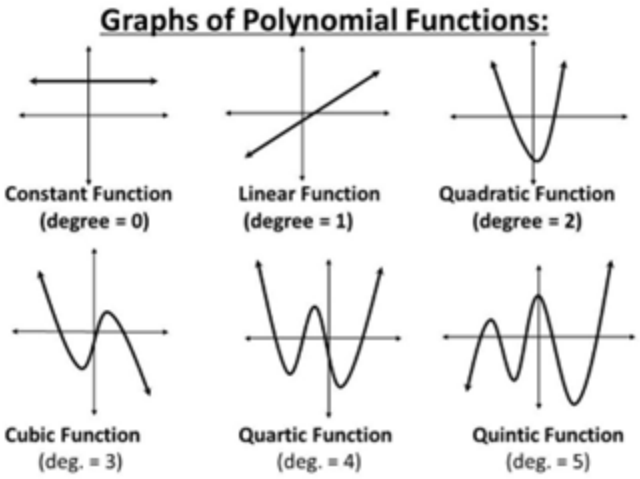
- Podemos ver que nuestros datos se distribuyen con alguna variación alrededor de la función verdadera (una onda sinusoidal parcial) debido al ruido aleatorio que agregamos. Durante el entrenamiento, queremos que nuestro modelo aprenda la verdadera función sin ser "distraído" por el ruido.

DS LIFE CYCLE: model building

Para elegir un modelo una buena regla es comenzar de manera simple y luego avanzar. El modelo más simple es una regresión lineal, donde las salidas son una combinación ponderada linealmente de las entradas. En nuestro modelo, usaremos una extensión de la regresión lineal llamada regresión polinomial para conocer la relación entre x e y. La regresión polinomial, donde las entradas se elevan a diferentes potencias, todavía se considera una forma de regresión "lineal" aunque el gráfico no forma una línea recta. La ecuación general para un polinomio se encuentra a continuación.

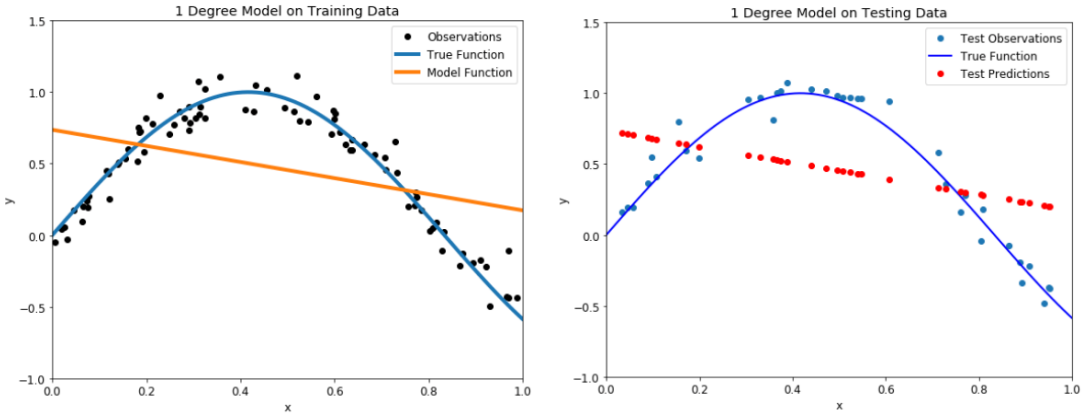
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon.$$

Aquí y representa la etiqueta y x es la característica. Los términos beta son los parámetros del modelo que se aprenderán durante el entrenamiento, y la épsilon es el error presente en cualquier modelo. Una vez que el modelo ha aprendido los valores beta, podemos introducir cualquier valor para x y obtener una predicción correspondiente para y. Un polinomio se define por su orden, que es la potencia más alta de x en la ecuación. Una recta es un polinomio de grado 1 mientras que una parábola tiene 2 grados.



MODEL BUILDING: underfitting

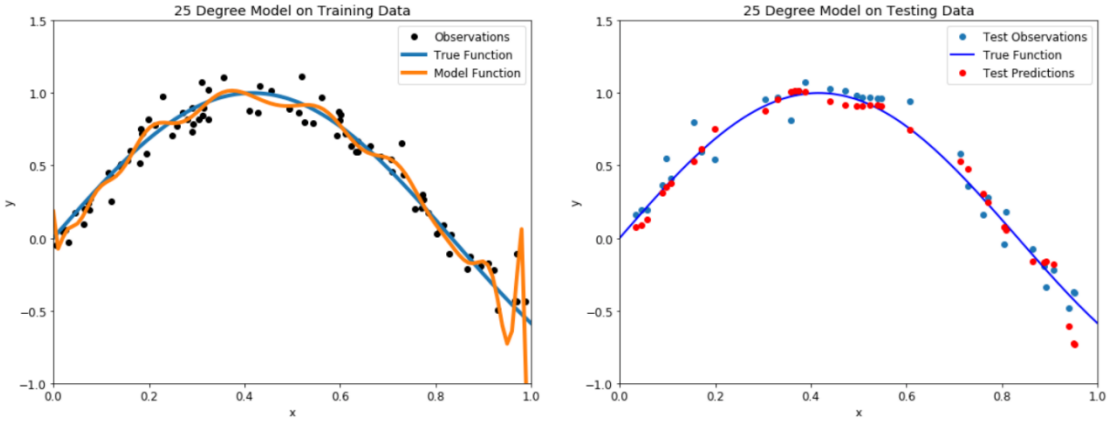
- Cuando un modelo no ha aprendido bien los patrones en los datos de entrenamiento y no puede generalizar bien los nuevos datos, se conoce como ajuste insuficiente. Un modelo inadecuado tiene un rendimiento deficiente en los datos de entrenamiento y dará como resultado predicciones poco confiables. El ajuste insuficiente ocurre debido al alto sesgo y la baja varianza.
- Un modelo de ajuste insuficiente con un ajuste polinomial de 1 grado. En la imagen inferior izquierda, la función del modelo en naranja se muestra encima de la función real y las observaciones de entrenamiento. A la derecha, se muestran las predicciones del modelo para los datos de prueba en comparación con la función real y los puntos de datos de prueba.



Nuestro modelo presenta **underfitting** pues tiene una varianza baja y un sesgo alto. La varianza se refiere a cuánto depende el modelo de los datos de entrenamiento. Para el caso de un polinomio de 1 grado, el modelo depende muy poco de los datos de entrenamiento porque apenas presta atención a los puntos. En cambio, el modelo tiene un alto sesgo, lo que significa que hace una fuerte suposición sobre los datos. Para este ejemplo, la suposición es que los datos son lineales, lo que evidentemente es bastante erróneo. Cuando el modelo hace predicciones de prueba, el sesgo lo lleva a hacer estimaciones inexactas. El modelo no pudo aprender la relación entre x e y debido a este sesgo, un claro ejemplo de ajuste insuficiente.

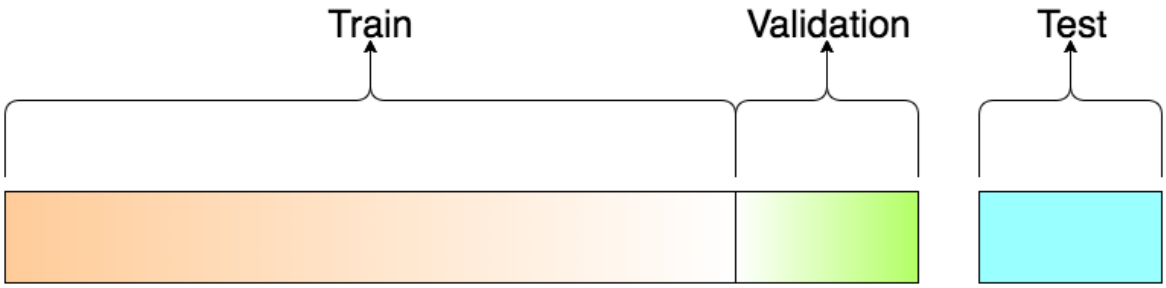
MODEL BUILDING: overfitting

Vimos que un grado bajo conduce al ***underfitting***. Una conclusión natural sería aprender bien los datos de entrenamiento, solo deberíamos aumentar el grado del modelo para capturar cada cambio en los datos.



Esto no es siempre o mejor. Con un grado tan alto de flexibilidad, el modelo hace todo lo posible para tener en cuenta cada punto de entrenamiento. Esto puede parecer una buena idea, ¿no queremos aprender de los datos? Además, el modelo tiene una gran puntuación en los datos de entrenamiento porque se acerca a todos los puntos. Si bien esto sería aceptable si las observaciones de entrenamiento representaran perfectamente la función verdadera, debido a que hay ruido en los datos, nuestro modelo termina ajustando el ruido. Este es un modelo con una varianza alta, porque cambiará significativamente dependiendo de los datos de entrenamiento. Las predicciones en el conjunto de prueba son mejores que el modelo de un grado, pero el modelo de veinticinco grados aún no aprende la relación porque esencialmente memoriza los datos de entrenamiento y el ruido.

DS LIFE CYCLE: train, validation y test sets



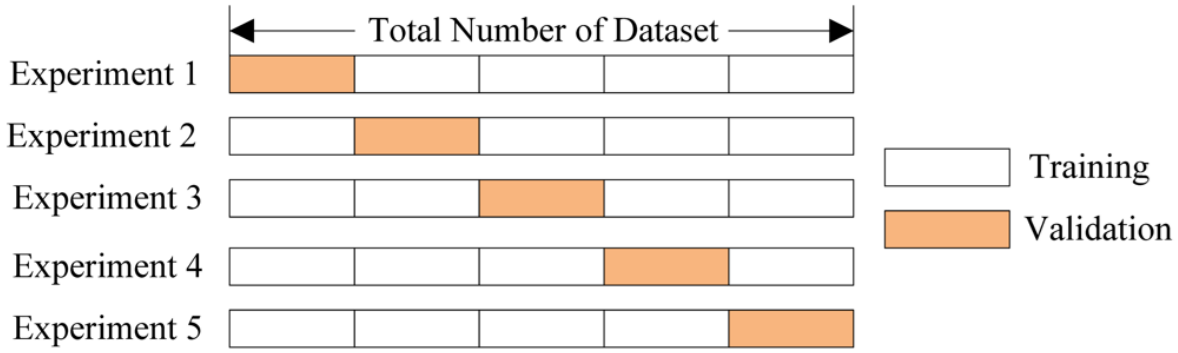
- **Conjunto de entrenamiento:** Conjunto de ejemplos utilizados para el aprendizaje. El conjunto de entrenamiento suele ser el conjunto más grande, en términos de tamaño, que se crea a partir del conjunto de datos original y se utiliza para encontrar el modelo. En otras palabras, los puntos de datos incluidos en el conjunto de entrenamiento se utilizan para aprender los parámetros del modelo de interés.

- **Conjunto de validación:** un conjunto de ejemplos utilizados para ajustar los híper parámetros de un clasificador, por ejemplo, para elegir el número de unidades ocultas en una red neuronal. Luego veremos como introducir la validación como una validación k-fold.

- **Conjunto de prueba:** un conjunto de ejemplos utilizados solo para evaluar el rendimiento de un clasificador completamente especificado. El conjunto de prueba se utiliza para evaluar el rendimiento de este modelo y garantizar que pueda generalizarse bien a puntos de datos nuevos e invisibles.

MODEL BUILDING: enter validation

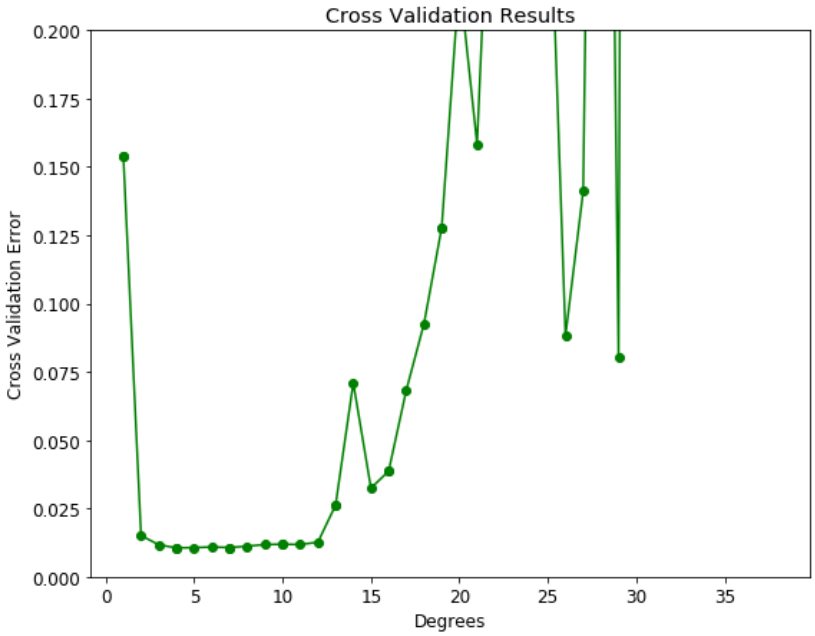
Necesitamos que nuestro modelo no "memorice" los datos de entrenamiento, sino que aprenda la relación real. ¿Cómo podemos encontrar un modelo balanceado con el grado polinomial correcto? Si elegimos el modelo con la mejor puntuación en el conjunto de entrenamiento, simplemente seleccionaremos el modelo de sobreajuste, pero esto no se puede generalizar bien para los datos de prueba. Afortunadamente, existe una técnica de ciencia de datos bien establecida para desarrollar el modelo óptimo: la validación.

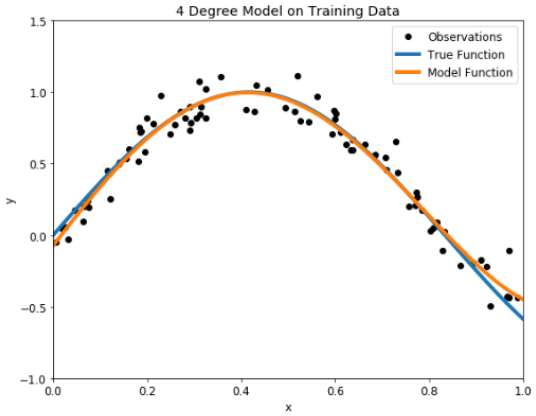


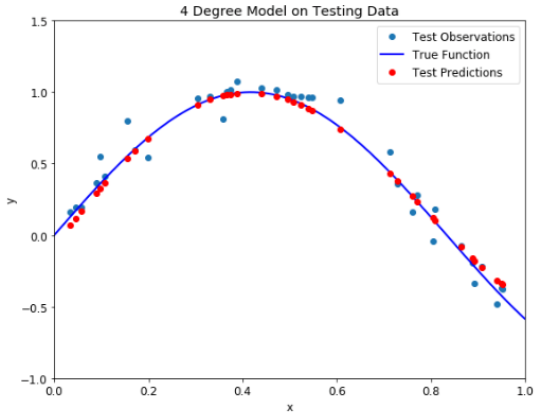
En lugar de usar un conjunto de validación separado, dividimos el conjunto de entrenamiento en varios subconjuntos, llamados folds. Usemos cinco folds como ejemplo. Realizamos una serie de ciclos de entrenamiento y evaluación donde cada vez entrenamos en 4 de los folds y probamos en el quinto, llamado conjunto de espera. Repetimos este ciclo 5 veces, cada vez usando un fold diferente para la evaluación. Al final, promediamos las puntuaciones de cada uno de los folds para determinar el rendimiento general de un modelo determinado. Esto nos permite optimizar el modelo antes de la implementación sin tener que utilizar datos adicionales.

MODEL BUILDING: enter validation

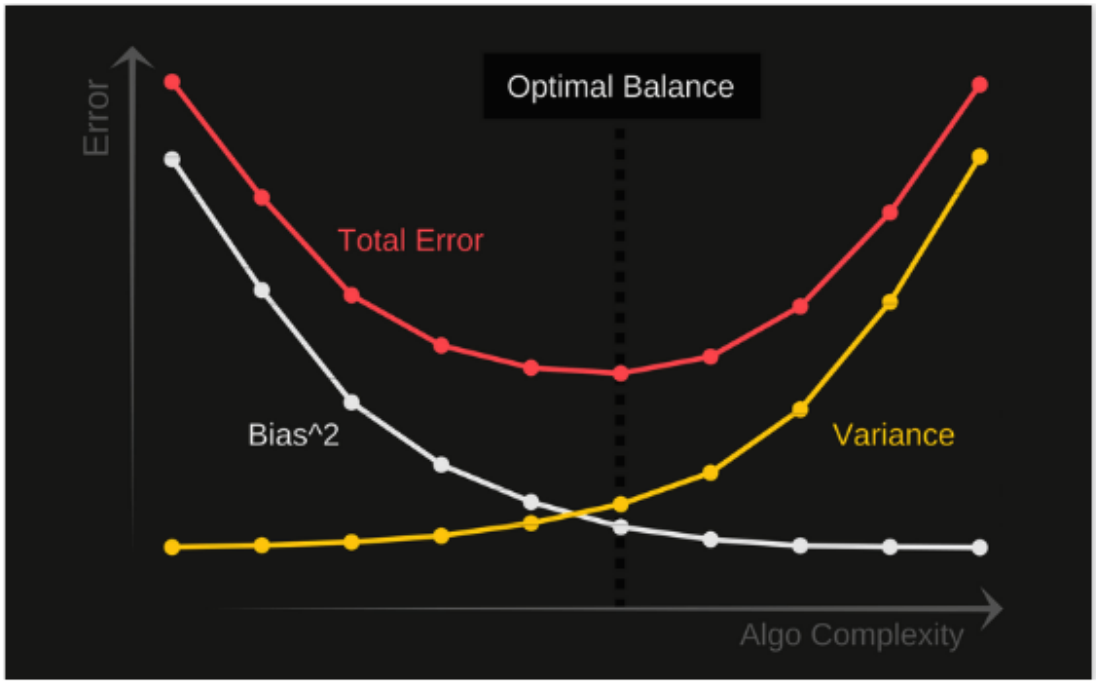
Para nuestro problema, podemos usar la validación cruzada para seleccionar el mejor modelo creando modelos con un rango de diferentes grados y evaluar cada uno usando la validación cruzada de 5 veces. El modelo con la puntuación de validación cruzada más baja se desempeñará mejor en los datos de prueba y logrará un equilibrio entre el ajuste insuficiente y el ajuste excesivo. Se sugiere usar modelos con grados del 1 al 40 para cubrir una amplia gama. Para comparar modelos, calculamos el error cuadrático medio, la distancia promedio entre la predicción y el valor real al cuadrado. La siguiente tabla muestra los resultados de la validación cruzada ordenados por menor error y el gráfico muestra todos los resultados con error en el eje y.





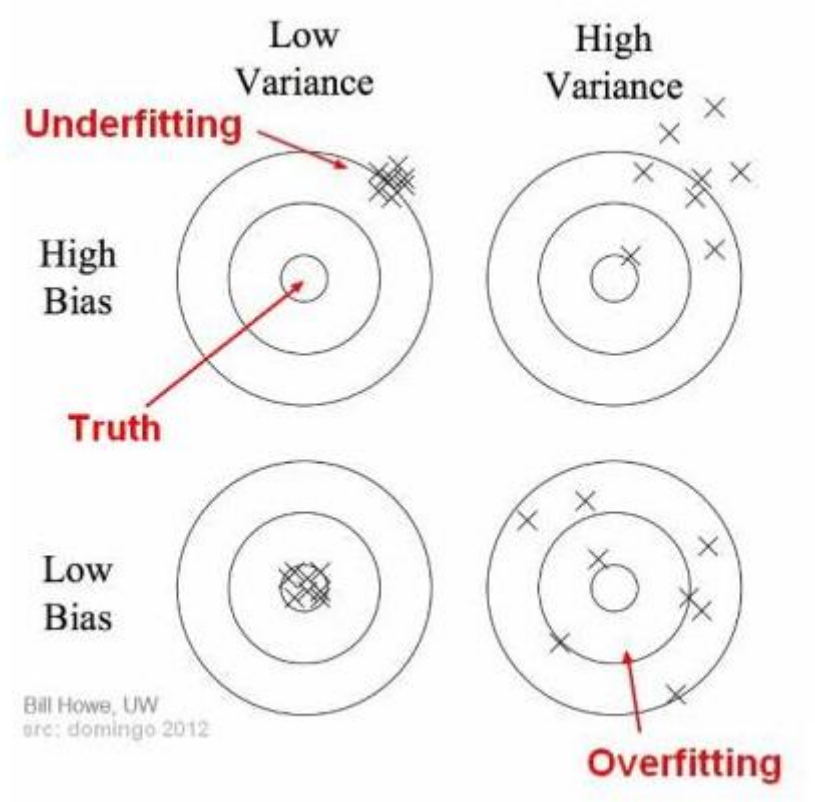


MACHINE LEARNING: bias-variance tradeoff



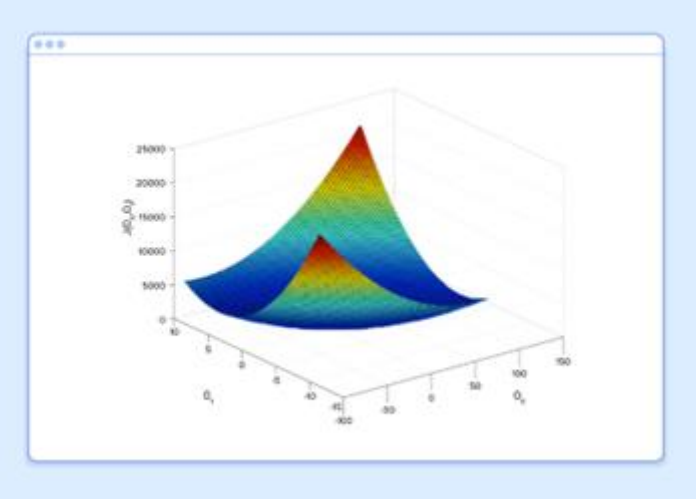
- Cada vez que discutimos la predicción del modelo, es importante comprender los errores de predicción (sesgo y varianza). Existe una compensación entre la capacidad de un modelo para minimizar el sesgo y la varianza. Obtener una comprensión adecuada de estos errores nos ayudara no solo a construir modelos precisos, sino también a evitar el error de sobreajuste y ajuste insuficiente.
- **Bias:** El sesgo es la diferencia entre la predicción promedio de nuestro modelo y el valor correcto que estamos tratando de predecir. El modelo con alto sesgo presta muy poca atención a los datos de entrenamiento y simplifica demasiado el modelo. Siempre conduce a un alto error en los datos de entrenamiento y prueba.
- **Variance:** La varianza es la variabilidad de la predicción del modelo para un punto de datos dado o un valor que nos indica la dispersión de nuestros datos. El modelo con alta varianza presta mucha atención a los datos de entrenamiento y no generaliza sobre los datos que no ha visto antes. Como resultado, estos modelos funcionan muy bien con los datos de entrenamiento, pero tienen altas tasas de error con los datos de prueba.

MACHINE LEARNING: evaluación del modelo



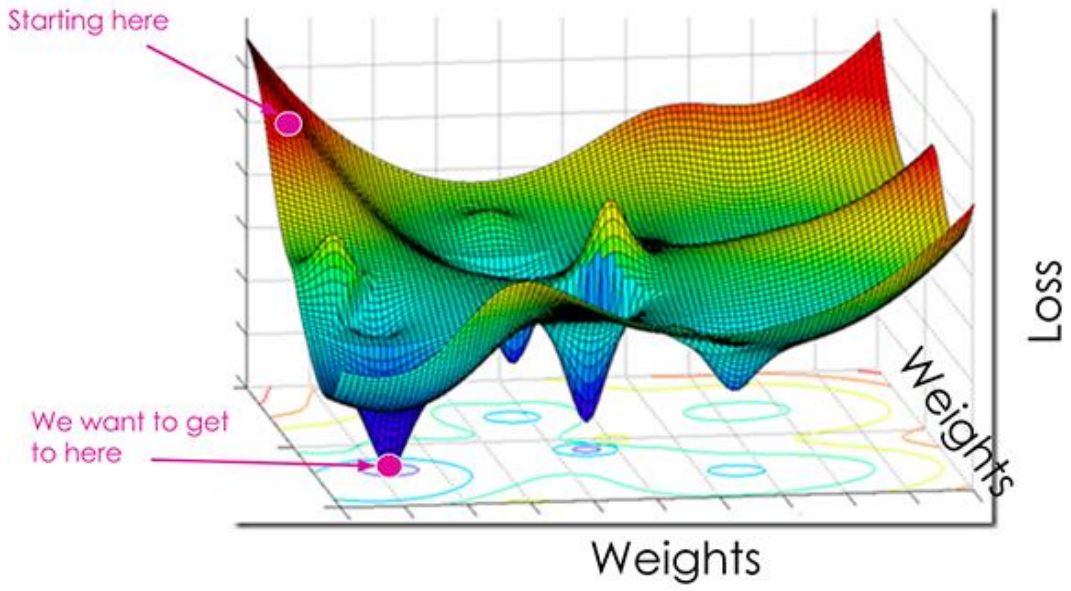
- En el diagrama anterior, el centro del objetivo es un modelo que predice perfectamente los valores correctos. A medida que nos alejamos de la diana, nuestras predicciones empeoran cada vez más. Podemos repetir nuestro proceso de creación de modelos para obtener resultados separados en el objetivo.
- En el aprendizaje supervisado, el ***underfitting*** ocurre cuando un modelo no puede capturar el patrón subyacente de los datos. Estos modelos suelen tener un alto sesgo y una baja varianza. Ocurre cuando tenemos muy poca cantidad de datos para construir un modelo preciso o cuando intentamos construir un modelo lineal con datos no lineales.
- En el aprendizaje supervisado, el ***overfitting*** ocurre cuando nuestro modelo captura el ruido junto con el patrón subyacente en los datos. Ocurre cuando entrenamos mucho nuestro modelo sobre un conjunto de datos ruidoso. Estos modelos tienen un sesgo bajo y una varianza alta.

MACHINE LEARNING:
función de perdida y función de costo



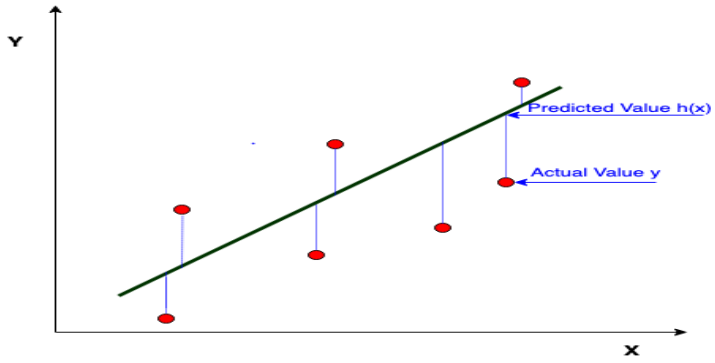
- Una función de pérdida J(x) mide que tan insatisfechos estamos con las predicciones de nuestro modelo con respecto a una respuesta correcta y utilizando ciertos valores de θ. Existen varias funciones de pérdida y la selección de uno de ellos depende de varios factores como el algoritmo seleccionado o el nivel de confianza deseado, pero principalmente depende del objetivo de nuestro modelo.
- La palabra 'Pérdida' establece la penalización por no lograr el resultado esperado. Si la desviación en el valor predicho del valor esperado por nuestro modelo es grande, entonces la función de pérdida da el número más alto como resultado y si la desviación es pequeña y mucho más cercana al valor esperado, genera un número menor.
- Función de costo es el promedio de todos los errores de la muestra en todo el conjunto de entrenamiento.

MACHINE LEARNING: evaluación del modelo



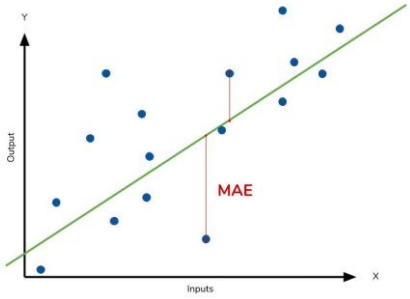
- Imagínese esto: ha entrenado un modelo de aprendizaje automático en un conjunto de datos determinado y está listo para ponerlo frente a su cliente. Pero, ¿cómo puede estar seguro de que este modelo dará el resultado óptimo? ¿Existe alguna métrica o técnica que lo ayude a evaluar rápidamente su modelo en el conjunto de datos?
- La evaluación del modelo es un método para evaluar la corrección de los modelos en los datos de prueba. Los datos de prueba consisten en puntos de datos que el modelo no ha visto antes.
- Los modelos se pueden evaluar utilizando múltiples métricas. Sin embargo, la elección correcta de una métrica de evaluación es crucial y, a menudo, depende del problema que se está resolviendo. Una comprensión clara de una amplia gama de métricas puede ayudar al evaluador a encontrar una coincidencia adecuada entre el enunciado del problema y una métrica.

MACHINE LEARNING: métricas de performance para regresion



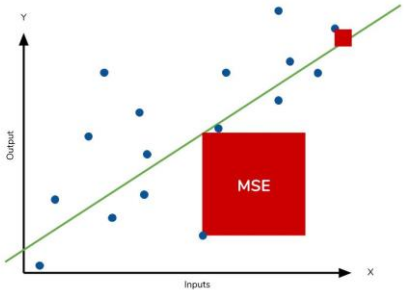
- Los modelos de regresión tienen una salida continua. Por lo tanto, necesitamos una métrica basada en el cálculo de algún tipo de distancia entre la realidad predicha y la real.
- Para evaluar los modelos de regresión, tenemos las siguientes métricas de error:

Error absoluto medio (MAE)



$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - y_j|$$

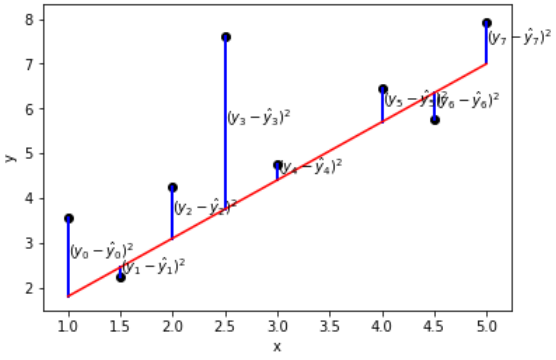
Error cuadrático medio (MSE)



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

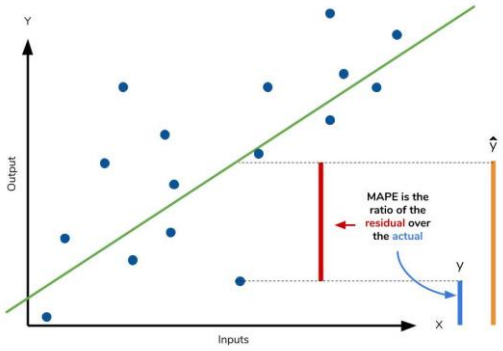
MACHINE LEARNING: métricas de performance para regresion

Raiz del error cuadrático medio (RMSE)



$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Mean absolute percentage error (MAPE)

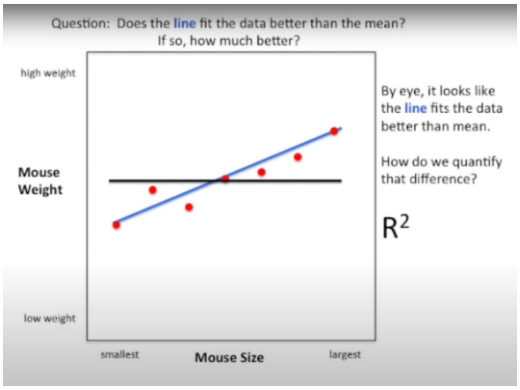


$$MAPE = \frac{100\%}{n}\ \Sigma\ \left|\frac{y - \hat{y}}{y}\right|$$

- Metrica R^2

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$

$$\text{Adjusted } R^2 = 1 - \frac{SS_{residuals}/(n-K)}{SS_{total}/(n-1)}$$



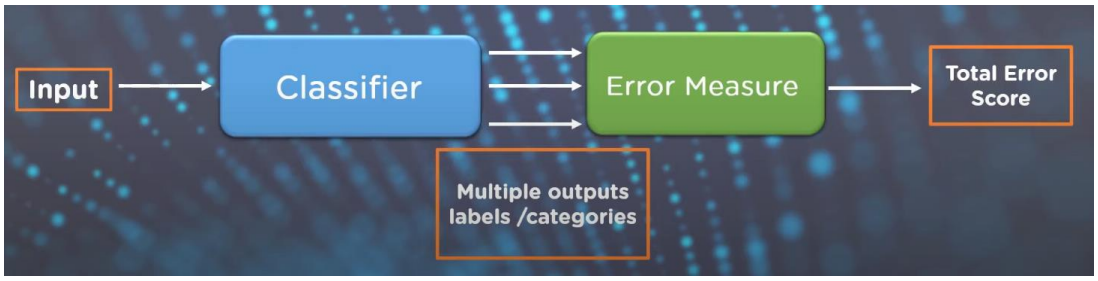
MACHINE LEARNING: métricas de performance para clasificacion

- Los problemas de clasificación son una de las áreas más investigadas del mundo. Los casos de uso están presentes en casi todos los entornos industriales y de producción. Reconocimiento de voz, reconocimiento facial, clasificación de texto: la lista es interminable.
- Los modelos de clasificación tienen una salida discreta, por lo que necesitamos una métrica que compare clases discretas de alguna forma. Las métricas de clasificación evalúan el rendimiento de un modelo y te dicen qué tan buena o mala es la clasificación, pero cada una de ellas lo evalúa de manera diferente.
- Para evaluar los modelos de clasificación, discutiremos estas métricas en detalle:
  - ✓ Exactitud (Accuracy)
  - ✓ Matriz de confusión
  - ✓ Precisión
  - ✓ Recall
  - ✓ Puntuaje F1 (F1-Score)

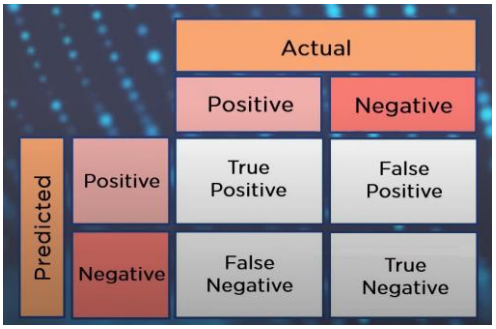
✓ ROC-AUC

ANEXO: Matriz de confusion

- Los modelos de clasificacion tienen multiples categorias de salida. La mayoria de las metricas de error nos indicaran el error total de un modelo, pero a partir de eso no podremos averiguar errores individuales en nuestro modelo.



- Una matriz de confusion es una tabla con las diferentes salidas predichas en un problema de clasificacion que nos ayuda a visualizar los resultados en una manera mas clara.



- Supongamos que deseamos predecir cuántas personas están infectadas con un virus peligroso antes de que muestren los síntomas y en base al eso aislarlos de la población sana.
- Nuestro conjunto de datos es un ejemplo de un conjunto de datos no balanceados (*imbalanced dataset*). Hay 947 puntos de datos para la clase negativa y 3 puntos de datos para la clase positiva.

ANEXO: Matriz de confusion

ID	Actual Sick?	Predicted Sick?	Outcome
1	1	1	TP
2	0	0	TN
3	0	0	TN
4	1	1	TP
5	0	0	TN
6	0	0	TN
7	1	0	FP
8	0	1	FN
9	0	0	TN
10	1	0	FP
:	:	:	:
1000	0	0	FN

- Así es como calcularemos la exactitud:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- De la tabla obtenemos TP = 30, TN = 930, FP = 30, FN = 10, calculamos la exactidud:

$$Accuracy = \frac{30 + 930}{30 + 30 + 930 + 10} = 0.96$$

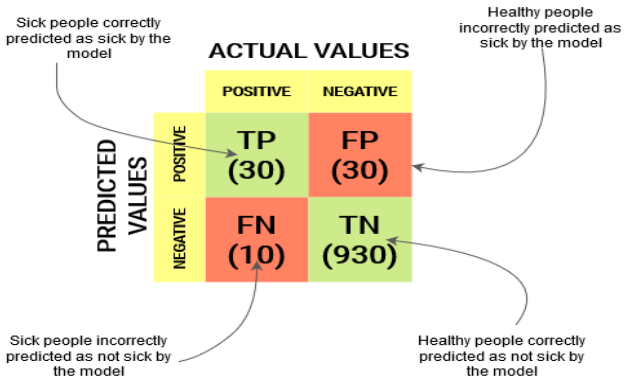
- El 96% a primera vista parece decirnos que porcenaje de la gente se enfermara.
- Si analizamos bien el resultado, el modelo esta prediciendo que porcentaje de la gente no se enfermara.
- Para analizar mejor esto introducimos los conceptos de Precision y Recall
- **Precision**,nos dice cuántos de los casos predichos como positivos resultaron ser positivos realmente.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall**, nos dice cuántos de los casos positivos reales pudimos predecir correctamente con nuestro modelo.

$$Recall = \frac{TP}{TP + FN}$$

ANEXO: Matriz de confusion



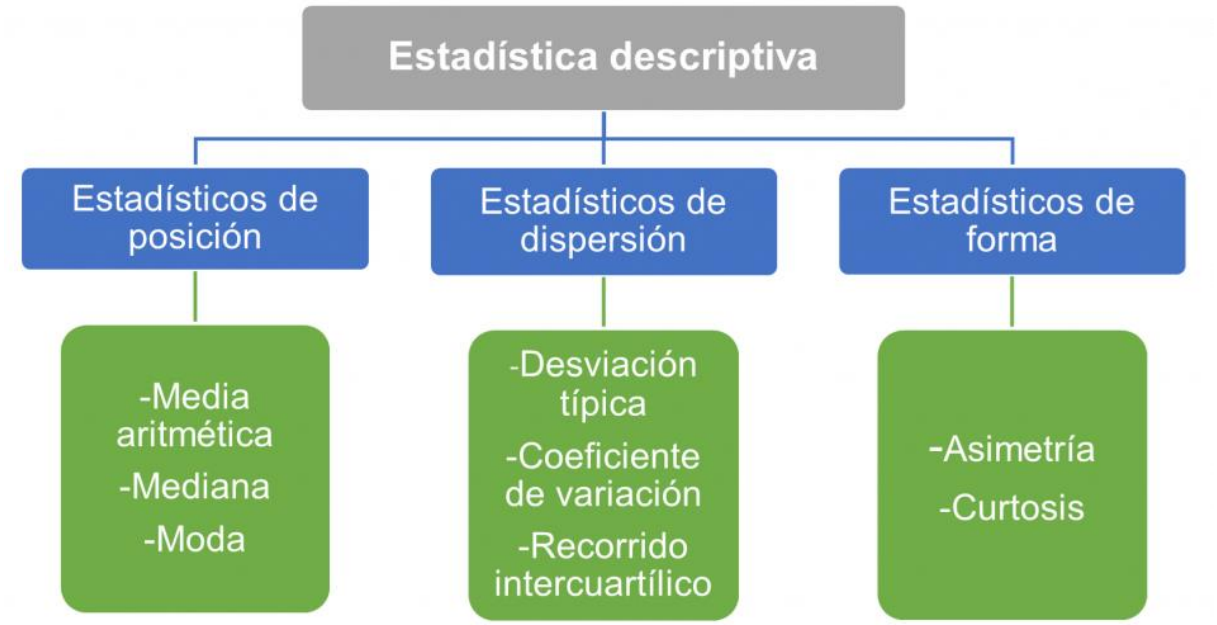
$$Precision = \frac{30}{30+30} = 0.5$$

$$Recall = \frac{30}{30+10} = 0.75$$

- En nuestro ejemplo, *Recall* sería una mejor métrica porque no queremos dar de alta accidentalmente a una persona infectada y dejar que se mezcle con la población sana, propagando así el virus contagioso.
- Recall es importante en los casos médicos en los que no importa si activamos una falsa alarma, ¡pero los casos positivos reales no deben pasar desapercibidos!
- La precisión es importante en los sistemas de recomendación de música o video, sitios web de comercio electrónico, etc. Los resultados incorrectos pueden provocar la pérdida de clientes y ser perjudiciales para el negocio.
- En los casos que no esta claro cual de los dos es mas importante procedemos a combinarlos:

$$F1 - score = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

REVISION DE ESTADISTICA



La estadística descriptiva es el término que se le da al análisis de datos que ayuda a describir, mostrar o resumir datos de una manera significativa de modo que puedan surgir patrones a partir de dichos datos, eso se realiza con la ayuda de gráficos o valores de resumen. Sin embargo, la estadística descriptiva no nos permitirá sacar conclusiones más allá de los datos que hemos analizado ni llegar a conclusiones con respecto a las hipótesis que podríamos haber hecho. Es simplemente una manera de describir nuestros datos.

La estadística descriptiva es muy importante porque si presentamos nuestros datos sin procesar, sería difícil visualizar lo que muestran los datos, especialmente si son muchos. Por lo tanto, la estadística descriptiva nos permite presentar los datos de una manera más significativa, lo que permitirá una interpretación más sencilla de los datos. Por ejemplo, si tenemos los resultados de 100 exámenes de un curso, podríamos estar interesados en el desempeño general de los estudiantes. También estaríamos interesados en la distribución o difusión de las notas. La estadística descriptiva nos permite hacer esto.

MEDIDAS DE TENDENCIA CENTRAL

Las medidas de tendencia central son formas de describir la posición central de una distribución de frecuencia para un grupo de datos. En nuestro ejemplo anterior, la distribución de frecuencias es simplemente la distribución y el patrón de calificaciones obtenidas por los 100 estudiantes desde la más baja hasta la más alta. Podemos describir esta posición central utilizando una serie de estadísticas, incluida la moda, la mediana y la media.
Una medida de tendencia central es un valor único que intenta describir un conjunto de datos mediante la identificación de la posición central dentro de ese conjunto de datos. Como tales, las medidas de tendencia central a veces se denominan medidas de ubicación central. También se clasifican como estadísticas de resumen. La media (a menudo llamada promedio) es probablemente la medida de tendencia central que más conocemos, pero hay otras, como la mediana y la moda.
La media, la mediana y la moda son todas medidas válidas de tendencia central, pero en diferentes condiciones, algunas medidas de tendencia central se vuelven más apropiadas que otras. En las siguientes secciones, veremos la media, la moda y la mediana, y aprenderemos cómo calcularlos y en qué condiciones son más apropiados para su uso.

Media

La media (o promedio) es la medida de tendencia central más popular y conocida. Se puede usar tanto con datos discretos como continuos, aunque su uso es más frecuente con datos continuos. La media es igual a la suma de todos los valores del conjunto de datos dividida por el número de valores del conjunto de datos.

Media de la muestra:

$$\overline{x} = \frac{\sum x}{n}$$

Media de la población:

$$\mu = \frac{\sum x}{n}$$

Una propiedad importante de la media es que incluye todos los valores de su conjunto de datos como parte del cálculo. Además, la media es la única medida de tendencia central donde la suma de las desviaciones de cada valor de la media es siempre cero.
La media tiene una desventaja principal: es particularmente susceptible a la influencia de valores atípicos. Estos son valores que son inusuales en comparación con el resto del conjunto de datos por ser especialmente pequeños o grandes en valor numérico.

Mediana

La mediana es la puntuación media de un conjunto de datos que se ha organizado en orden de magnitud. La mediana se ve menos afectada por los valores atípicos y los datos sesgados.
Para calcular la mediana, supongamos que tenemos los siguientes datos (número impar de datos):

65	55	89	56	35	14	56	55	87	45	92

Primero ordenamos los datos en orden ascendente:

14	35	45	55	55	**56**	56	65	87	89	92

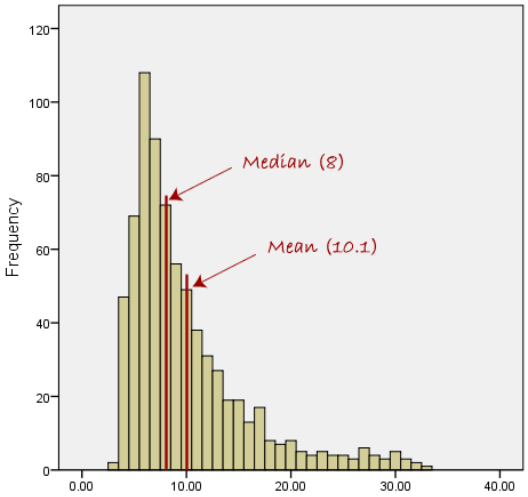
En el siguiente ejemplo tenemos un numero par de datos:

65	55	89	56	35	14	56	55	87	45

Primero ordenamos los datos en orden ascendente:

14	35	45	55	**55**	**56**	56	65	87	89

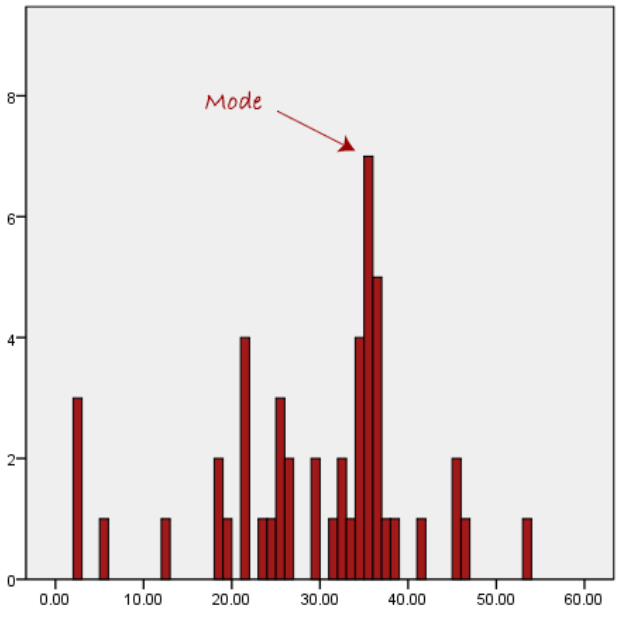
La mediana será el resultado del promedio de las notas 5 y 6 es decir: 55.5
Preferiremos usar la mediana sobre la media (o la moda) cuando nuestros datos están sesgados (es decir, la distribución de frecuencia de nuestros datos está sesgada). Si consideramos la distribución normal, ya que es la más evaluada en estadística, cuando los datos son perfectamente normales, la media, la mediana y la moda son idénticas. Además, todos representan el valor más típico en el conjunto de datos. Sin embargo, a medida que los datos se vuelven sesgados, la media pierde su capacidad de proporcionar la mejor ubicación central para los datos porque los datos sesgados los alejan del valor típico. Sin embargo, la mediana conserva mejor esta posición y no está tan fuertemente influenciada por los valores sesgados.
En el caso de datos sesgados, en la siguiente imagen encontramos que la media está siendo arrastrada en el sentido directo del sesgo. En estas situaciones, generalmente se considera que la mediana es el mejor representante de la ubicación central de los datos. Cuanto más sesgada sea la distribución, mayor será la diferencia entre la mediana y la media, y se debe poner mayor énfasis en usar la mediana en lugar de la media. Un ejemplo clásico de la distribución sesgada hacia la derecha salario, donde los que ganan más brindan una representación falsa del ingreso típico si se expresan como una media y no como una mediana.



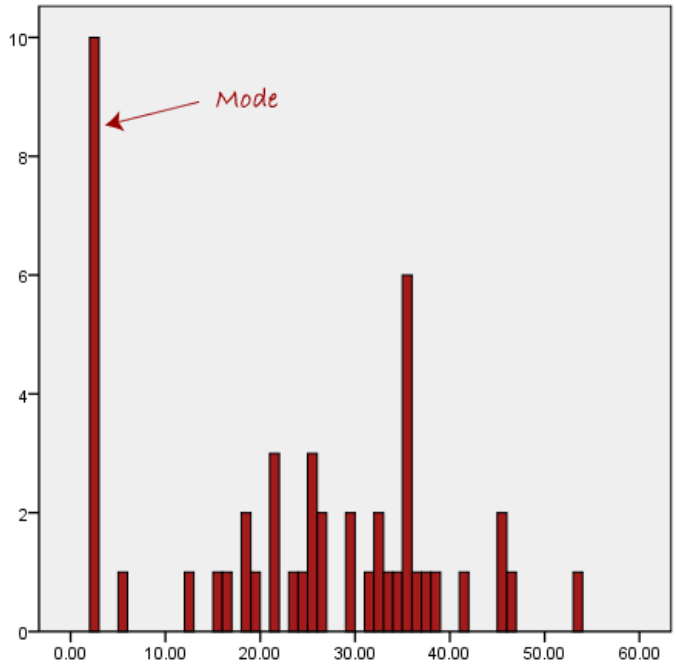
Moda

La moda es la puntuación más frecuente en nuestro conjunto de datos. En un histograma representa la barra más alta en un gráfico de barras o histograma. Por lo tanto, a veces puede considerar la moda como la opción más popular.

A continuación, se presenta un ejemplo de una moda:



Un problema con la moda es que no nos proporcionará una muy buena medida de tendencia central cuando la marca más común está lejos del resto de los datos en el conjunto de datos, como se muestra en el siguiente diagrama:


Mode
10
8
6
4
2
0
0.00
10.00
20.00
30.00
40.00
50.00
60.00

MEDIDAS DE DISPERSIÓN

Introducción

Las medidas de dispersión, se utiliza para describir la variabilidad en una muestra o población. Por lo general, se usa junto con una medida de tendencia central, como la media o la mediana, para proporcionar una descripción general de un conjunto de datos.
Por ejemplo, la puntuación media de los 100 alumnos puede ser de 65 sobre 100. Sin embargo, no todos los alumnos habrán obtenido 65 puntos. Más bien, sus puntajes se distribuirán. Unos serán más bajos y otros más altos. Las medidas de dispersión nos ayudan a resumir cuán dispersas están estas puntuaciones. Para describir este diferencial, tenemos a nuestra disposición una serie de estadísticas, algunas de ellas son los cuartiles, la varianza, la desviación standard y la correlación.

Cuartiles

Los cuartiles nos informan sobre la dispersión de un conjunto de datos dividiéndolo en cuartos, al igual que la mediana lo divide por la mitad. Por ejemplo, considere las calificaciones de los 100 estudiantes, a continuación, que se han ordenado de la calificación más baja a la más alta. En este caso datos no agrupados.

$$Q_1 = x_i + d \cdot (x_{i+1} - x_i)$$

Order	Score	Order	Score	Order	Score	Order	Score	Order	Score
1st	35	21st	42	41st	53	61st	64	81st	74
2nd	37	22nd	42	42nd	53	62nd	64	82nd	74
3rd	37	23rd	44	43rd	54	63rd	65	83rd	74
4th	38	24th	44	44th	55	64th	66	84th	75
5th	39	25th	45	45th	55	65th	67	85th	75
6th	39	26th	45	46th	56	66th	67	86th	76
7th	39	27th	45	47th	57	67th	67	87th	77
8th	39	28th	45	48th	57	68th	67	88th	77
9th	39	29th	47	49th	58	69th	68	89th	79
10th	40	30th	48	50th	58	70th	69	90th	80
11th	40	31st	49	51st	59	71st	69	91st	81
12th	40	32nd	49	52nd	60	72nd	69	92nd	81
13th	40	33rd	49	53rd	61	73rd	70	93rd	81
14th	40	34th	49	54th	62	74th	70	94th	81
15th	40	35th	51	55th	62	75th	71	95th	81
16th	41	36th	51	56th	62	76th	71	96th	81
17th	41	37th	51	57th	63	77th	71	97th	83
18th	42	38th	51	58th	63	78th	72	98th	84
19th	42	39th	52	59th	64	79th	74	99th	84
20th	42	40th	52	60th	64	80th	74	100th	85

N = 100;
Primer Cuartil => (N + 1)/4 = 25.25;
Segundo Cuartil => 2 (N + 1)/4 = 50.50;
Tercer Cuartil => 3(N + 1)/4 = 75.75
Q1 = 55 + 0.25(45 – 45) = 25
Q2 = 58 + 0.50(59 – 58) = 58.5
Q3 = 71 + 0.75(71 – 71) = 71

Varianza

Los cuartiles son útiles, pero también son algo limitados porque no tienen en cuenta todas las notas de nuestro grupo de datos. Para tener una idea más representativa de la dispersión, debemos tener en cuenta los valores reales de cada puntaje en un conjunto de datos. La varianza y la desviación estándar son tales medidas.
La varianza alcanza valores positivos elevando al cuadrado cada una de las desviaciones. La suma de estas desviaciones al cuadrado nos da la suma de los cuadrados, que luego podemos dividir por el número total de notas en nuestro grupo de datos (en otras palabras, 100 porque hay 100 estudiantes) para encontrar la varianza. Por lo tanto, para nuestros 100 estudiantes, la varianza es 211,89, como se muestra a continuación:

$$variance = \frac{\sum(X-\mu)^2}{N}$$

$$= \frac{21188.75}{100}$$

$$= 211.89$$

Where μ = mean, X = score, ∑= the sum of, N = number of scores, ∑X = "add up all the scores"

Como medida de variabilidad, la varianza es útil. Si las notas en nuestro grupo de datos están muy dispersas, la varianza será un número grande. Por el contrario, si las notas se distribuyen muy cerca de la media, la varianza será un número menor. Sin embargo, hay dos problemas potenciales con la varianza. En primer lugar, debido a que las desviaciones de las notas con respecto a la media se elevan al cuadrado, esto da más peso a las puntuaciones extremas. Si nuestros datos contienen valores atípicos (en otras palabras, uno o un pequeño número de puntajes que están particularmente lejos de la media y quizás no representan bien nuestros datos en su conjunto), esto puede deshacer el peso de estas notas. En segundo lugar, la varianza no está en las mismas unidades que las puntuaciones en nuestro conjunto de datos: la varianza se mide en unidades al cuadrado. Esto significa que no podemos ubicarlo en nuestra distribución de frecuencia y no podemos relacionar directamente su valor con los valores de nuestro conjunto de datos. Por lo tanto, la cifra de 211,89, nuestra varianza, parece algo arbitraria. Calcular la desviación estándar en lugar de la varianza corrige este problema. No obstante, el análisis de la varianza es extremadamente importante en algunos análisis estadísticos.

Desviación Standard

La desviación estándar es una medida de la dispersión de puntajes dentro de un conjunto de datos. Por lo general, estamos interesados en la desviación estándar de una población. Sin embargo, como a menudo se nos presentan datos de una muestra solamente, podemos estimar la desviación estándar de la población a partir de una desviación estándar de la muestra. Estas dos desviaciones estándar (desviaciones estándar de la muestra y de la población) se calculan de manera diferente. En estadística, generalmente se nos presenta el tener que calcular las desviaciones estándar de la muestra, aunque también se mostrará la fórmula para una desviación estándar de la población.
La desviación estándar se usa junto con la media para resumir datos continuos, no datos categóricos. Además, la desviación estándar, como la media, normalmente solo es adecuada cuando los datos continuos no están significativamente sesgados o tienen valores atípicos.
Fórmula de la desviación standard para una muestra:

$$s = \sqrt{\frac{\sum(X - \bar{X})^2}{n - 1}}$$

Fórmula de la desviación estándar para una población:

$$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{n}}$$

Covarianza

La desviación típica es un indicador de dispersión de una variable. ¿Qué pasa cuando tienes más de una variable? ¿Existe alguna forma de saber cómo se relaciona una con la otra?
La Covarianza es la media aritmética de los productos de las desviaciones de cada una de las variables respecto a sus medias respectivas.

$$\sigma_{XY} = \sum_{i=1}^{N} \frac{(X_i - \overline{X})(Y_i - \overline{Y})}{N}$$

La covarianza positiva: Cuando uno variable crece la otra variable también. Tienen una relación directa.

La covarianza negativa: Cuando una variable crece la otra variable decrece. Tienen una relación Inversa.

Coeficiente de Correlación

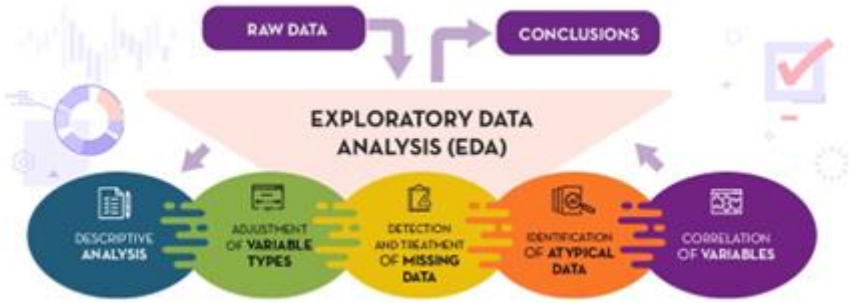
La correlación es un indicador para saber si hay relación (LINEAL) entre dos variables numéricas para esto se utiliza el coeficiente de correlación o correlación de Pearson.
La correlación contesta preguntas como las siguientes:

- ¿La práctica de algún deporte está relacionada con una vida más longeva?
- ¿Existe una relación entre la cantidad de carne ingerida diariamente y el cáncer?
- ¿Mayor estudio implica mejores notas en un examen?

La correlación es un ratio entre la dispersión entre las dos variables conjuntamente (covarianza) y la desviación standard de cada variable.

$$\rho_{xy} = \frac{Cov_{xy}}{\sigma_x \sigma_y}$$

ANÁLISIS EXPLORATORIO DE DATOS - EDA



Antes de realizar análisis de datos, con fines estadísticos o predictivos, usando por ejemplo técnicas de aprendizaje automático, es necesario entender la materia prima (raw data) con la que vamos a trabajar. Es necesario comprender y evaluar la calidad de los datos para, entre otros aspectos, detectar y tratar los datos atípicos (outliers) o incorrectos, evitando posibles errores que puedan repercutir en los resultados del análisis.

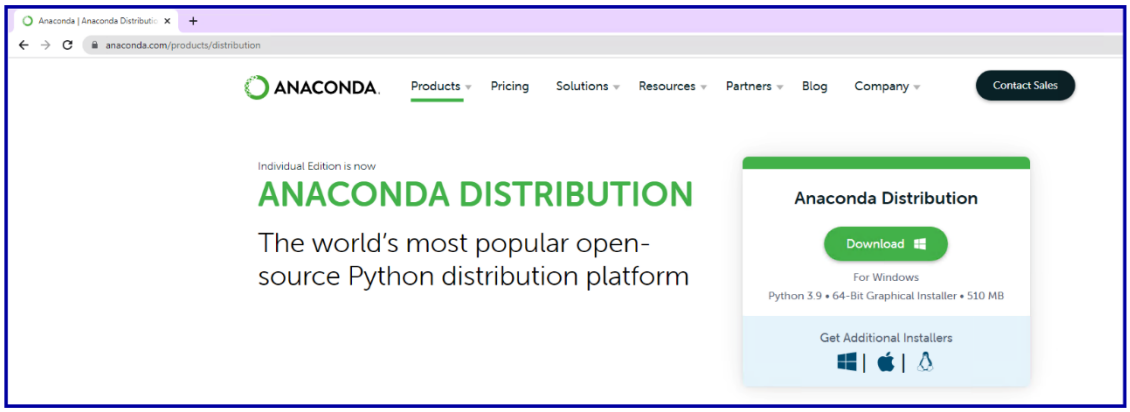
EDA consiste en aplicar un conjunto de técnicas estadísticas destinadas a explorar, describir y resumir la naturaleza de los datos, de forma que podamos entender claramente como están relacionadas nuestras variables de interés.

Todo esto nos permite identificar posibles errores, revelar la presencia de outliers, comprobar la relación entre variables (correlaciones) y su posible redundancia, y realizar un análisis descriptivo de los datos mediante representaciones gráficas y resúmenes de los aspectos más significativos.

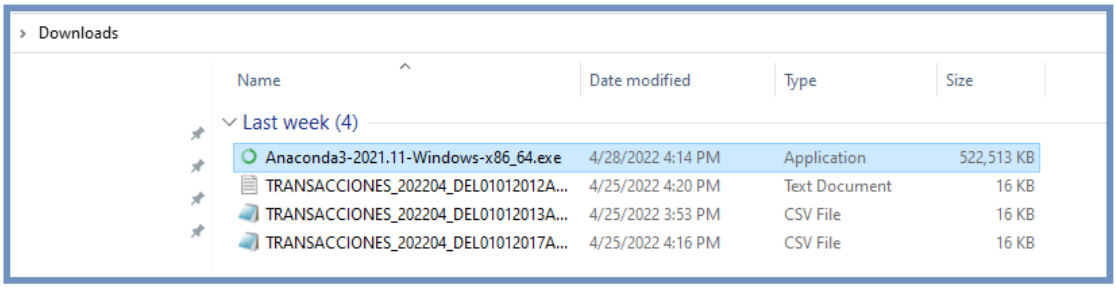
ANEXO: Instalación de Anaconda en Windows

- Anaconda es una distribución de los lenguajes de programación Python y R para computación científica (ciencia de datos, aplicaciones de Machine Learning, procesamiento de datos a gran escala, análisis predictivo, etc.).
- Tiene como ventaja simplificar la gestión e implementación de paquetes. La distribución incluye paquetes de "data science" adecuados para Windows, Linux y macOS.
- Para instalar Anaconda ingresar a la siguiente página:

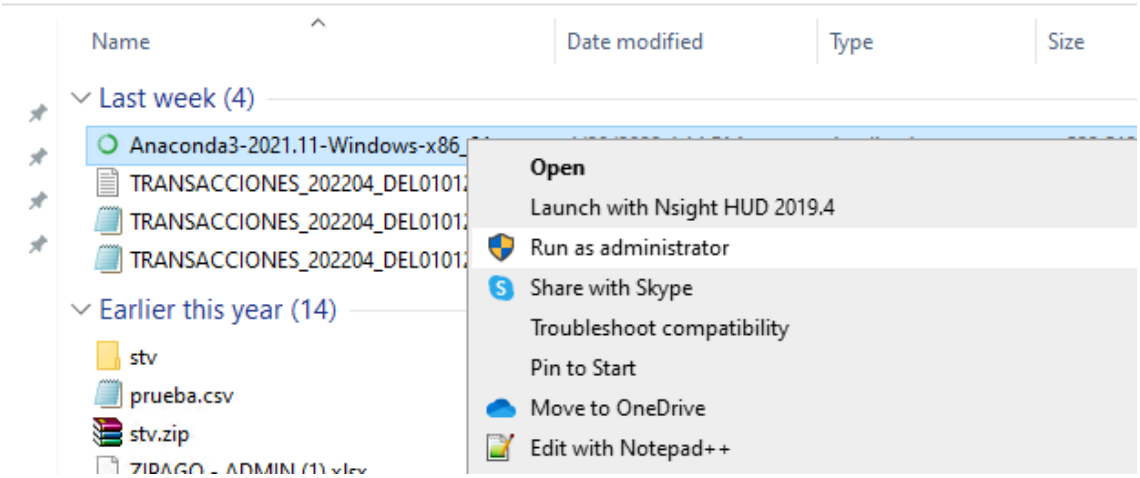
https://www.anaconda.com/products/distribution



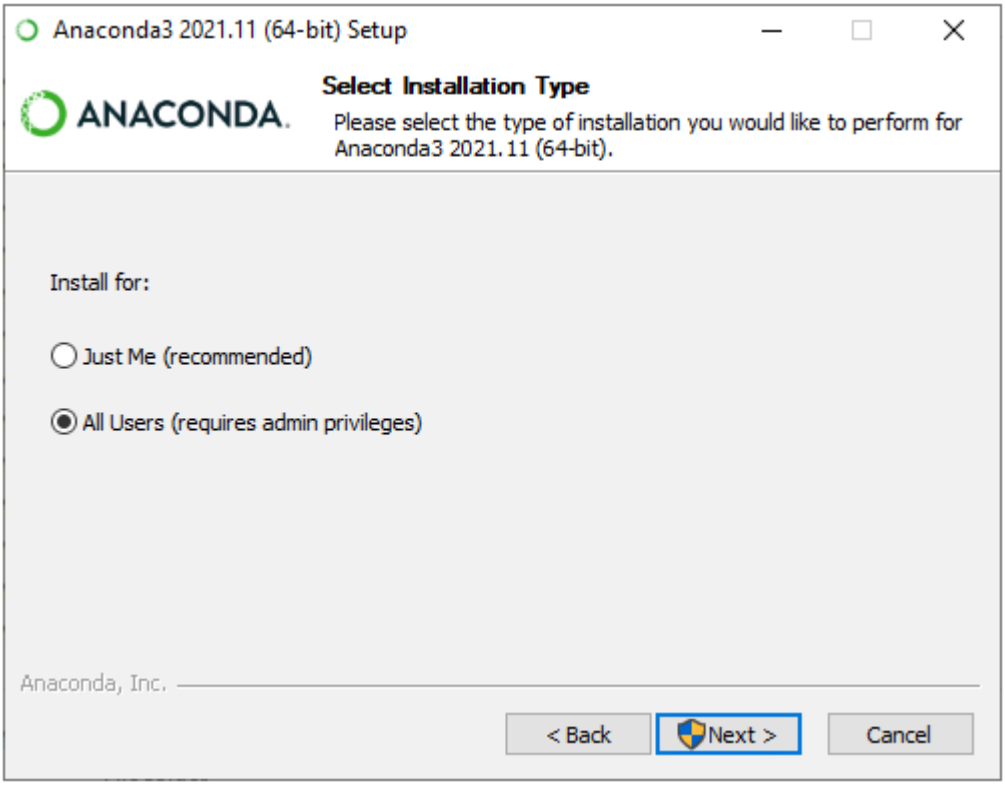
- Descargar el instalador presionando en **Download**:



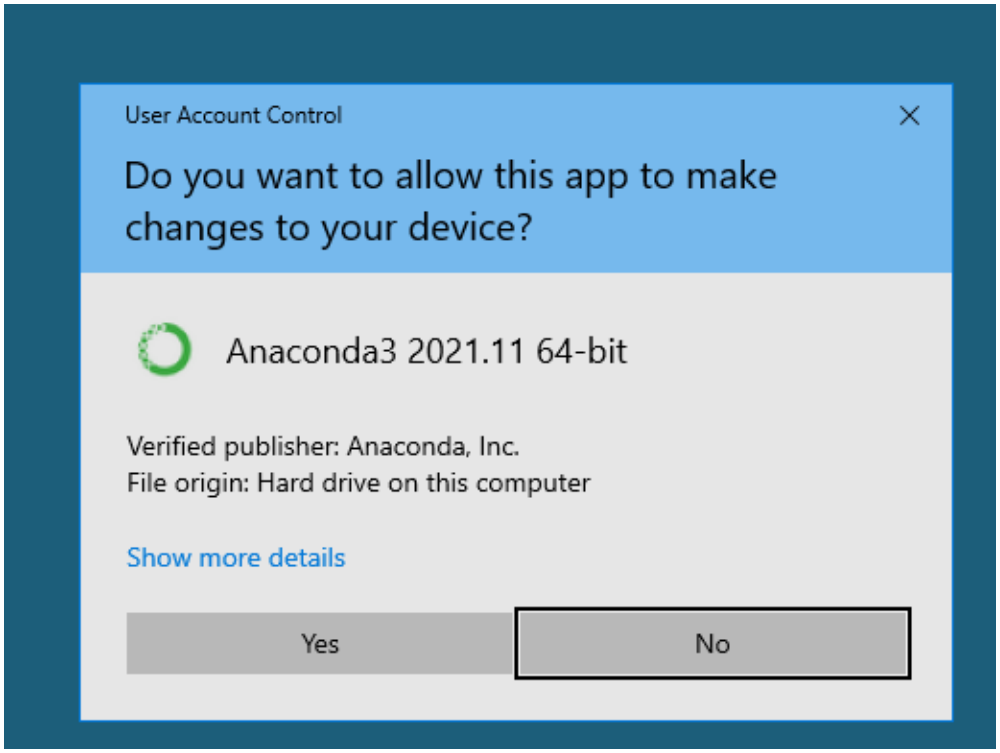
Una vez descargado ejecutar el instalador como usuario administrador:



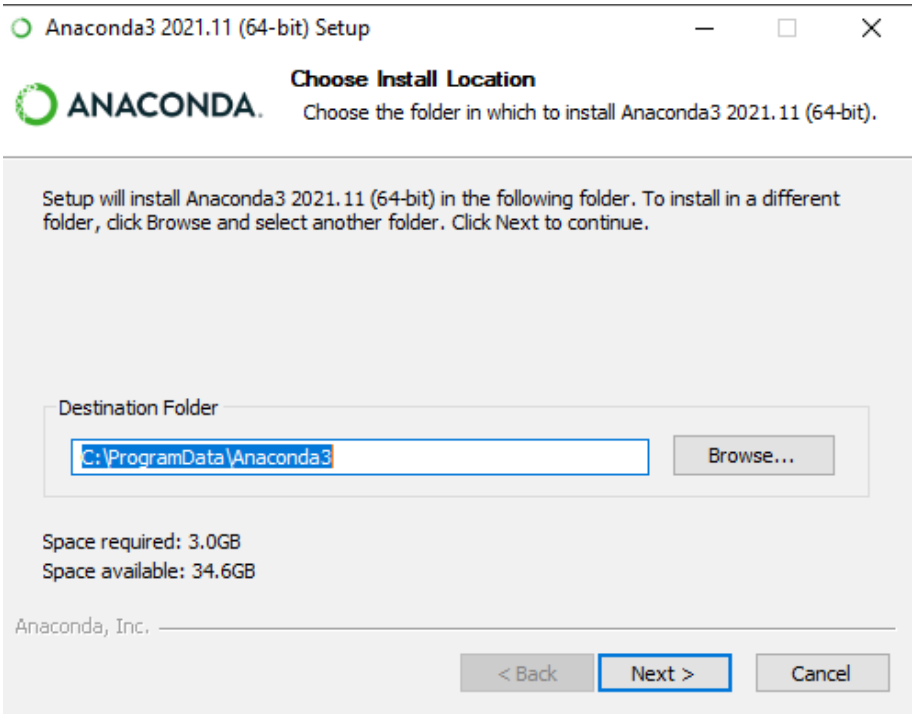
Presionar Next



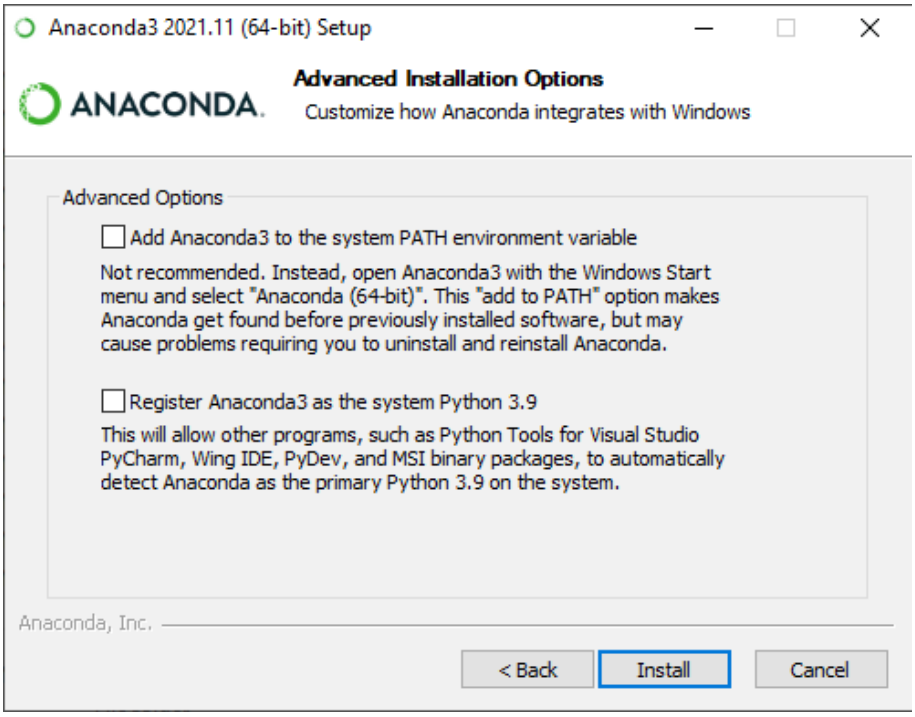
Presionar Next



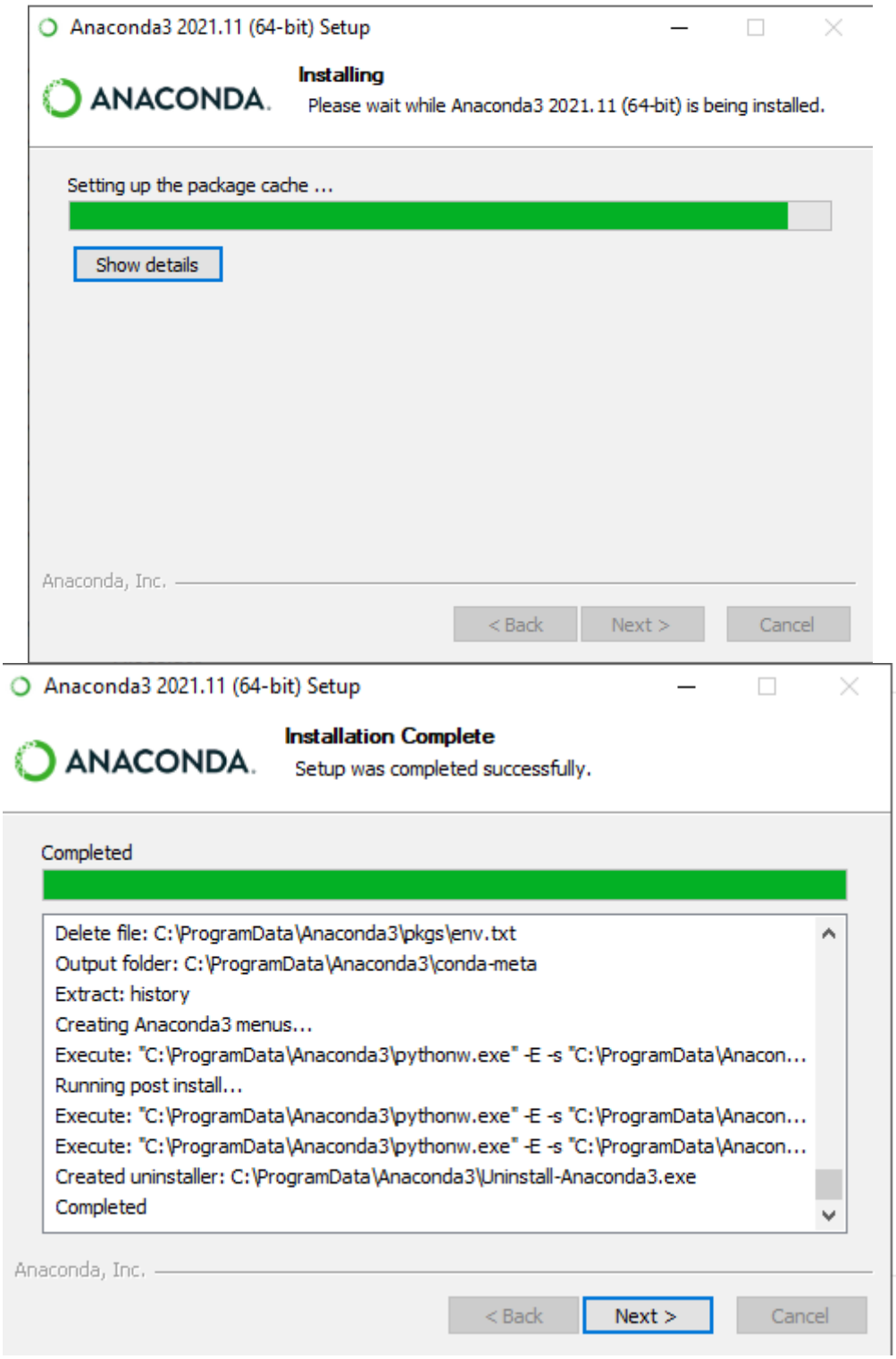
Presionar Yes



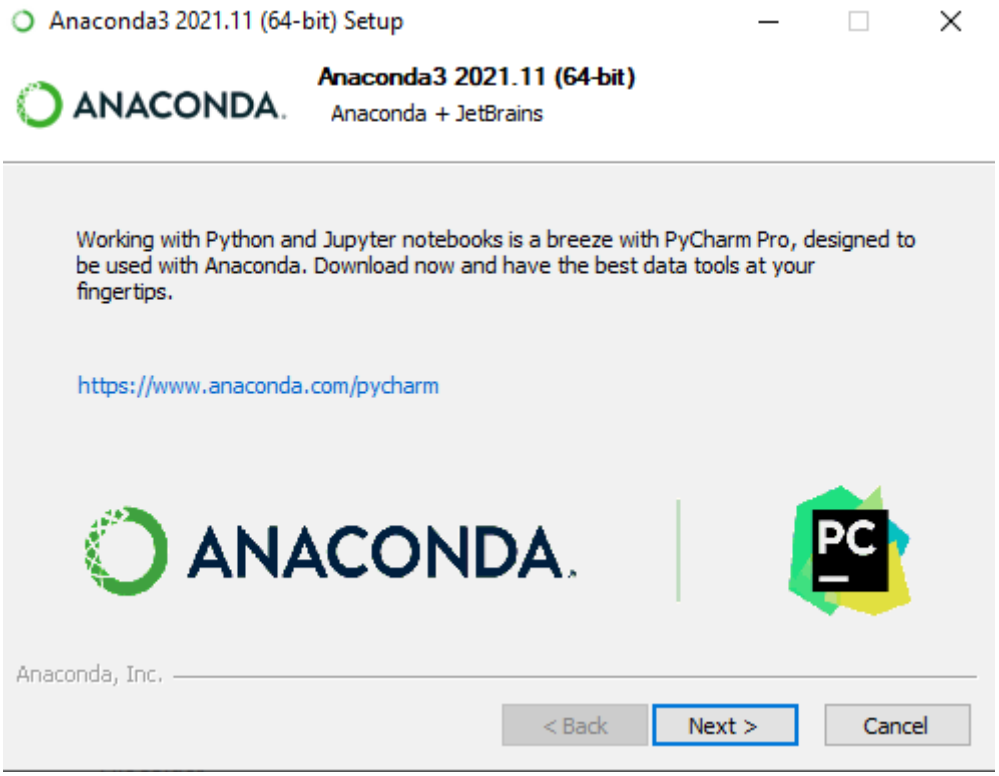
Presionar Next



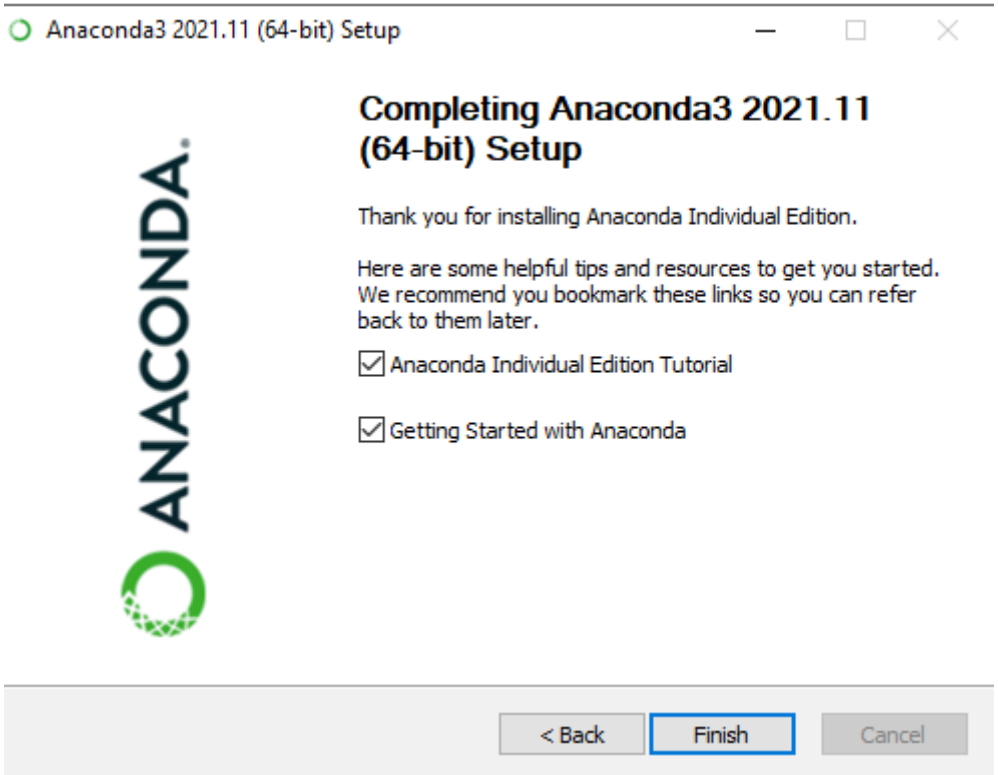
Presionar Install



Presionar Next

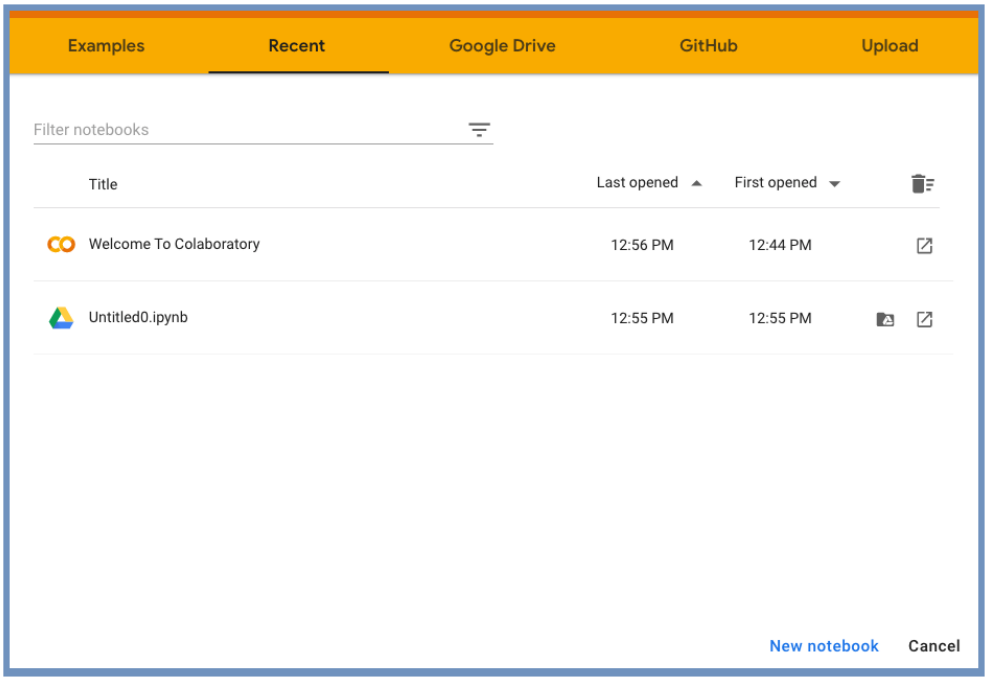


Presionar Next



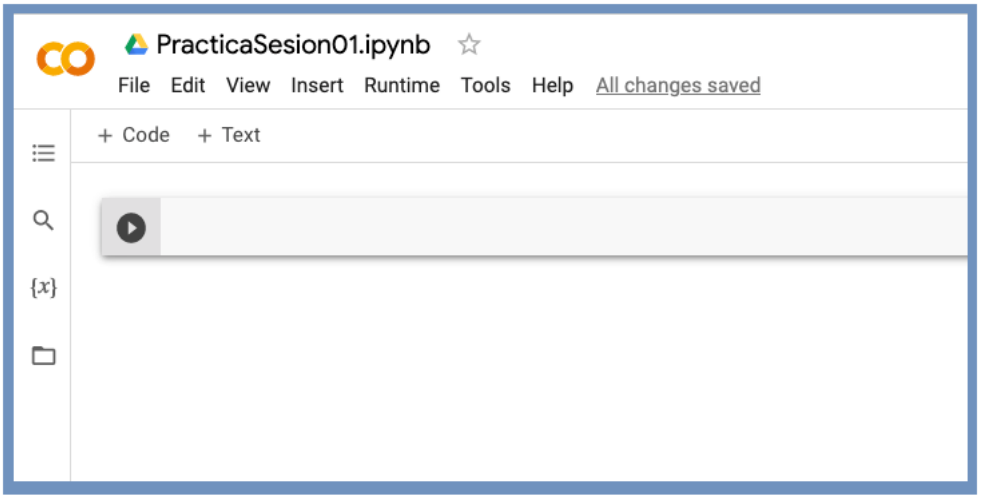
ANEXO: Usando Google Colab

- Para utilizar google colab primero debemos estar logueados a nuestra cuenta de Gmail
- Ingresar a la ruta: https://colab.research.google.com/

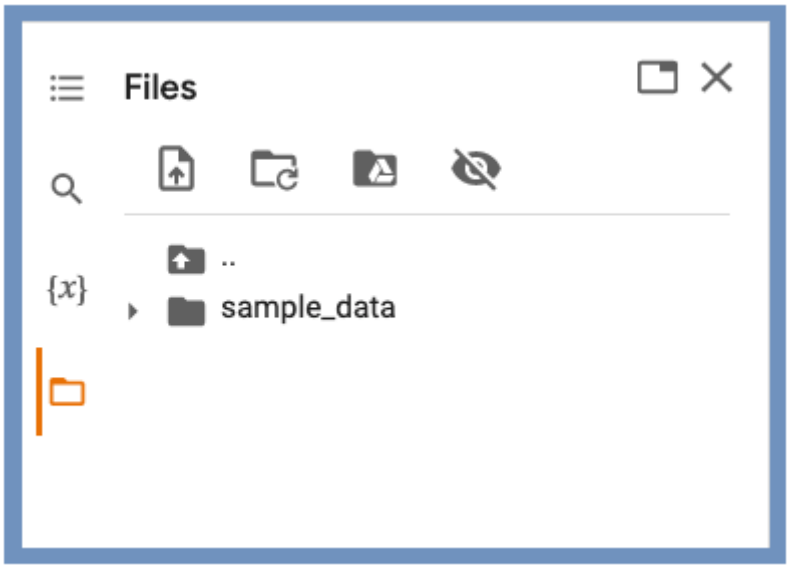


Presionar **New Notebook**

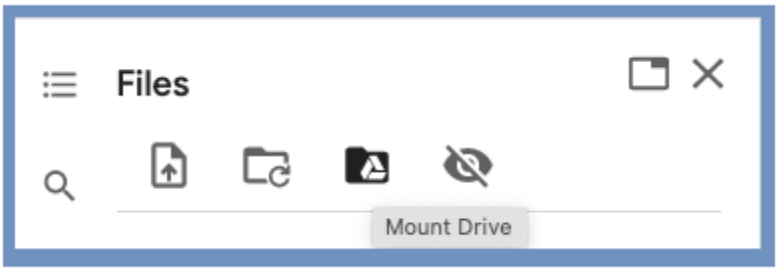
Se abrirá un nuevo IPython Note Book más conocido como Jupyter Notebook, un entorno computacional interactivo, en el que puede combinar ejecución de código, texto enriquecido, matemáticas, gráficos etc. Podemos modificar el nombre a nuestro gusto.

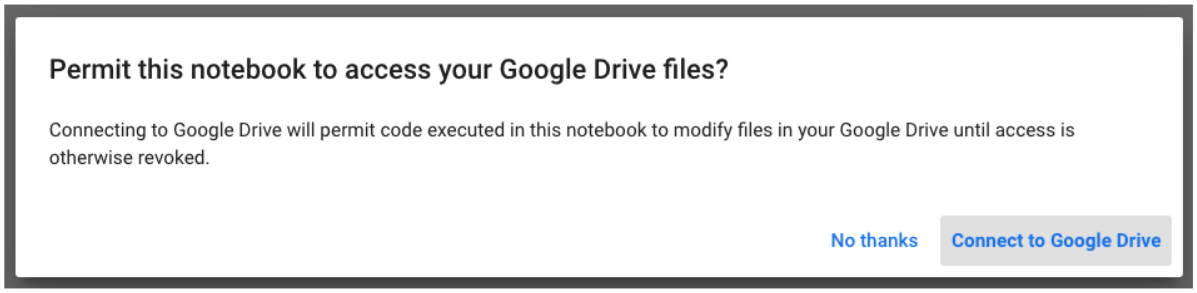


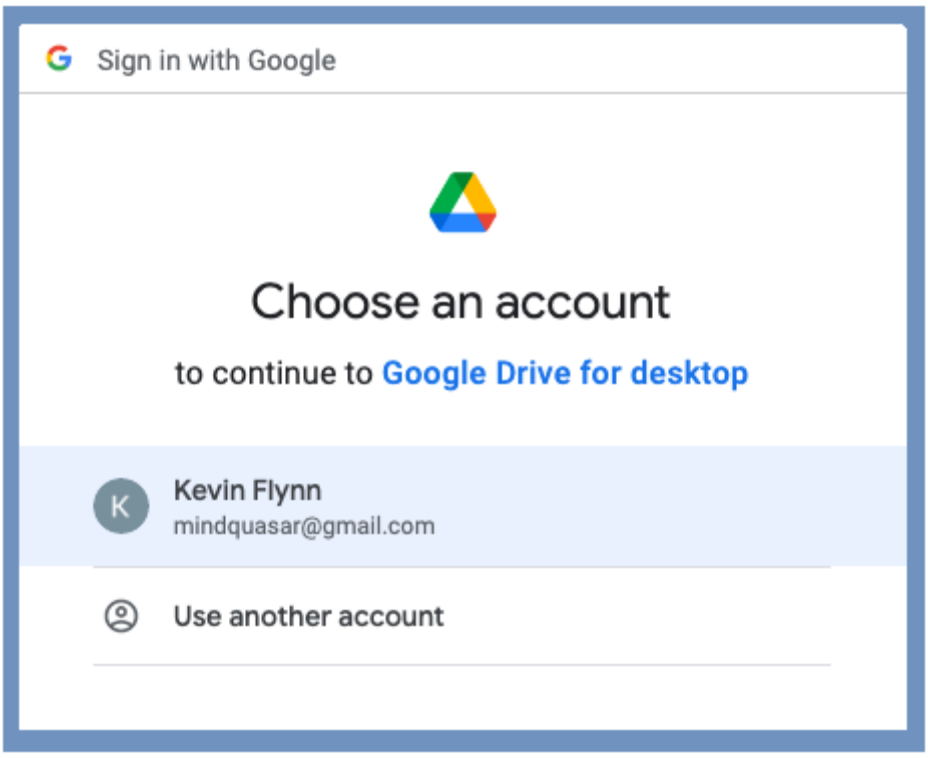
- Seleccionemos el folder en el menú lateral izquierdo y podremos navegar en los directorios de nuestra computadora.



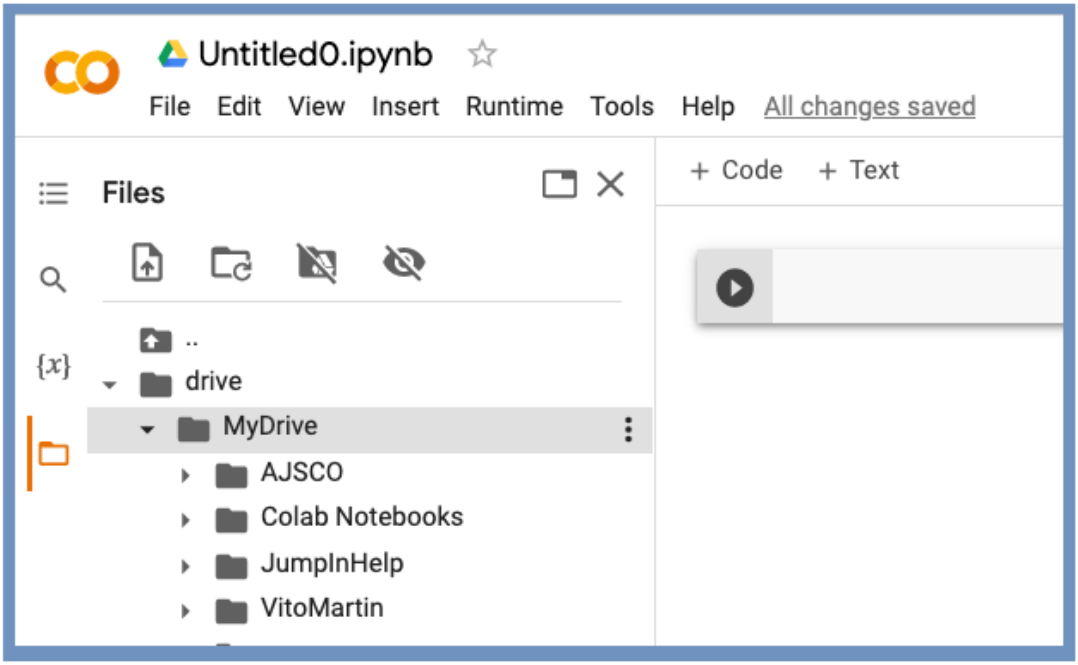
- Seleccionemos el folder con el icono de reciclaje en el menú lateral superior y podremos montar nuestro google drive.





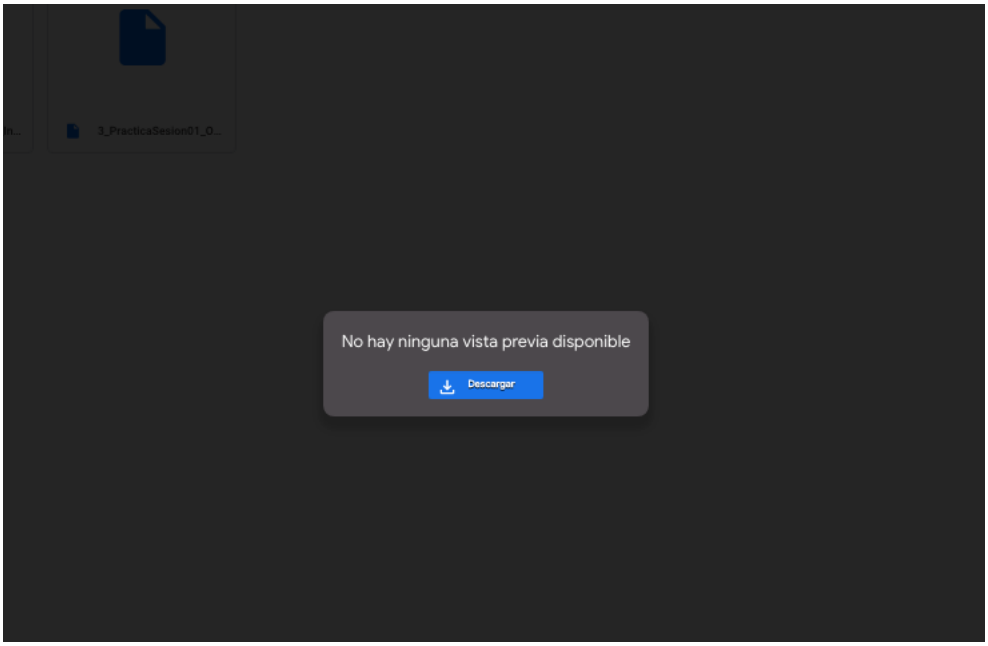


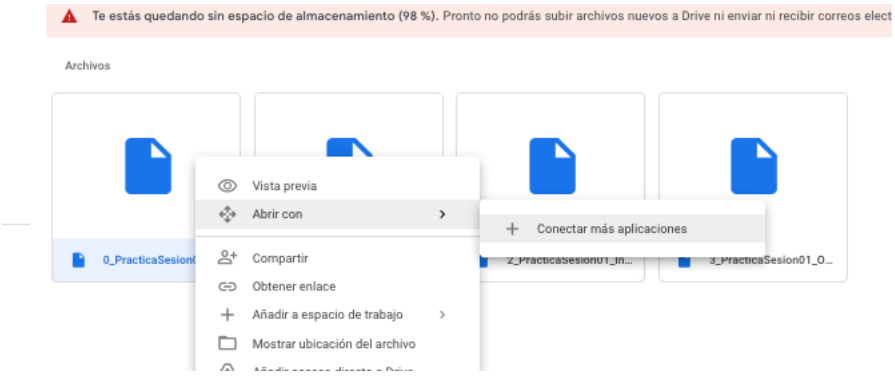
- Una vez montado nuestro google drive podemos tener acceso a nuestros archivos que están en la nube de google.



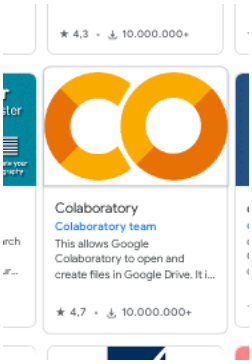
Ahora estamos listos para codificar.

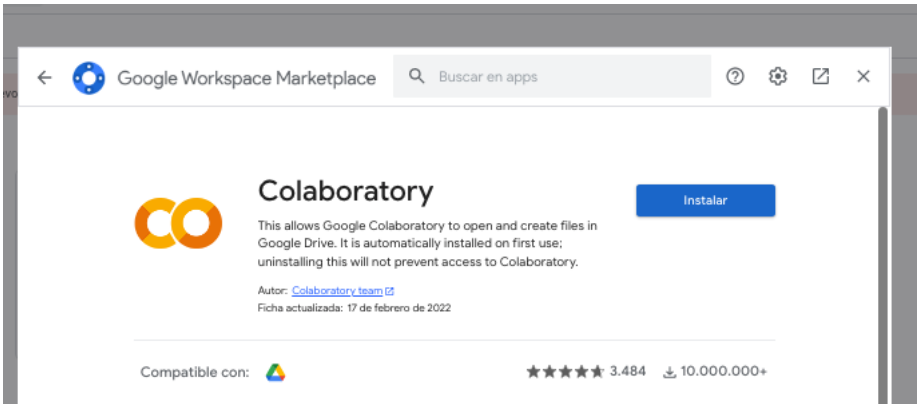
Si les sale el siguiente mensaje:



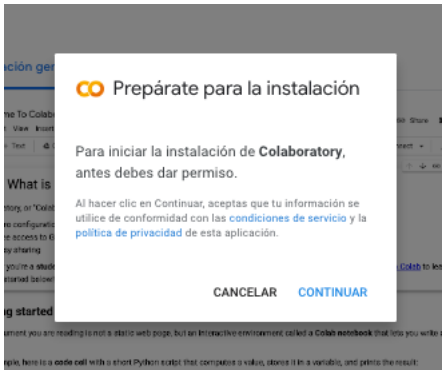


Seleccuibar Colaboratory

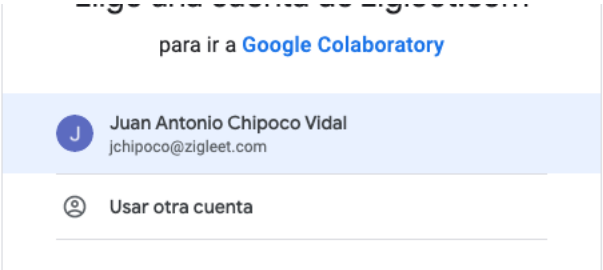




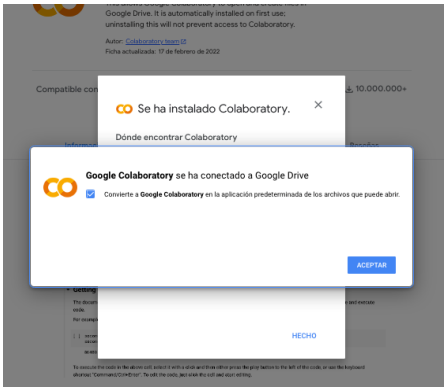
Presionar instalar



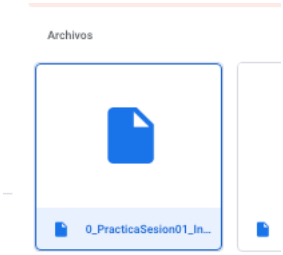
Presionar continuar



Seleccionamos nuestra cuenta gmail

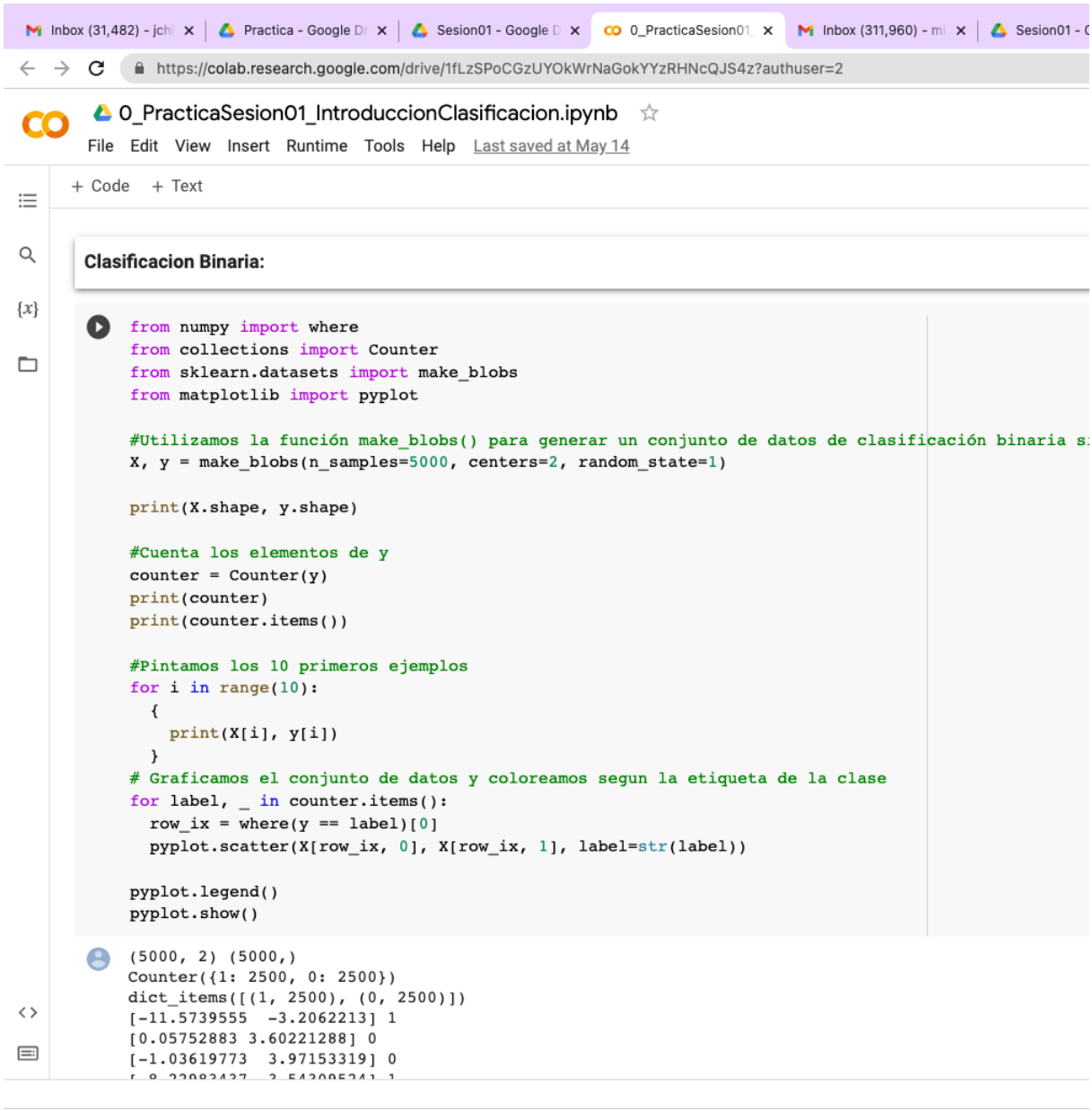


Presionar aceptar



Hacemos doble click sobre nuestro jupyter notebook

Ya podremos ejecutar las fuentes


Inbox (31,482) - jch
Practica - Google Dr
Sesion01 - Google D
0_PracticaSesion01
Inbox (311,960) - mi
Sesion01 - C
https://colab.research.google.com/drive/1fLzSPoCGzUYOkWrNaGokYYzRHNcQJS4z?authuser=2
0_PracticaSesion01_IntroduccionClasificacion.ipynb
File Edit View Insert Runtime Tools Help Last saved at May 14
+ Code + Text
Clasificacion Binaria:
from numpy import where
from collections import Counter
from sklearn.datasets import make_blobs
from matplotlib import pyplot
#Utilizamos la función make_blobs() para generar un conjunto de datos de clasificación binaria s
X, y = make_blobs(n_samples=5000, centers=2, random_state=1)
print(X.shape, y.shape)
#Cuenta los elementos de y
counter = Counter(y)
print(counter)
print(counter.items())
#Pintamos los 10 primeros ejemplos
for i in range(10):
{
print(X[i], y[i])
}
# Graficamos el conjunto de datos y coloreamos segun la etiqueta de la clase
for label, _ in counter.items():
row_ix = where(y == label)[0]
pyplot.scatter(X[row_ix, 0], X[row_ix, 1], label=str(label))
pyplot.legend()
pyplot.show()
(5000, 2) (5000,)
Counter({1: 2500, 0: 2500})
dict_items([(1, 2500), (0, 2500)])
[-11.5739555 -3.2062213] 1
[0.05752883 3.60221288] 0
[-1.03619773 3.97153319] 0