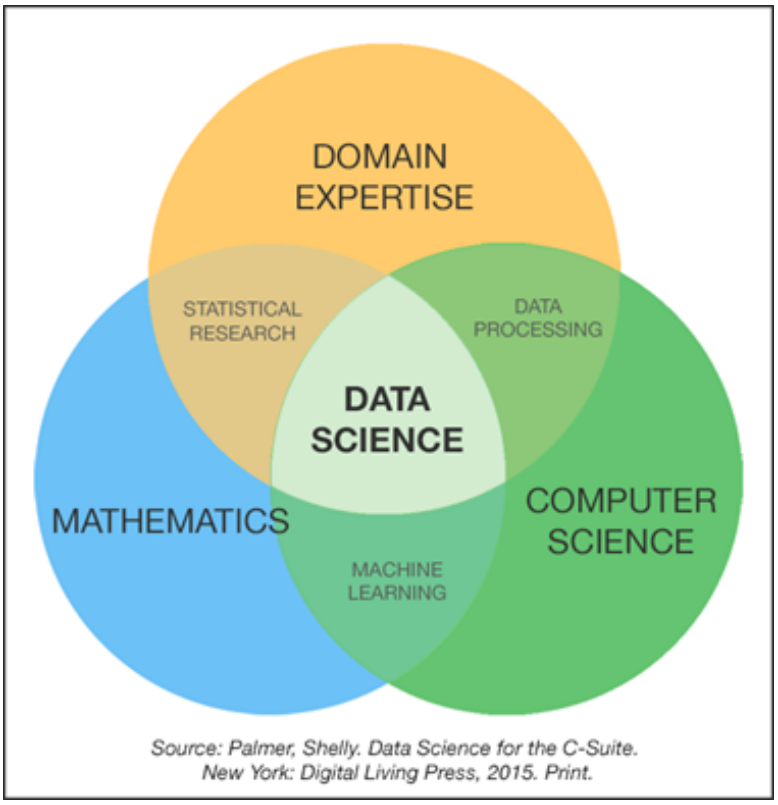
CTIC UNI
CENTRO DE TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIONES
UNIVERSIDAD NACIONAL DE INGENIERÍA

CIENCIA DE DATOS

2022



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

SESION 01

INTRODUCCION AL MACHINE LEARNING

Juan Antonio Chipoco Vidal
jchipoco@gmail.com

CTIC UNI
CENTRO DE TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIONES
UNIVERSIDAD NACIONAL DE INGENIERÍA

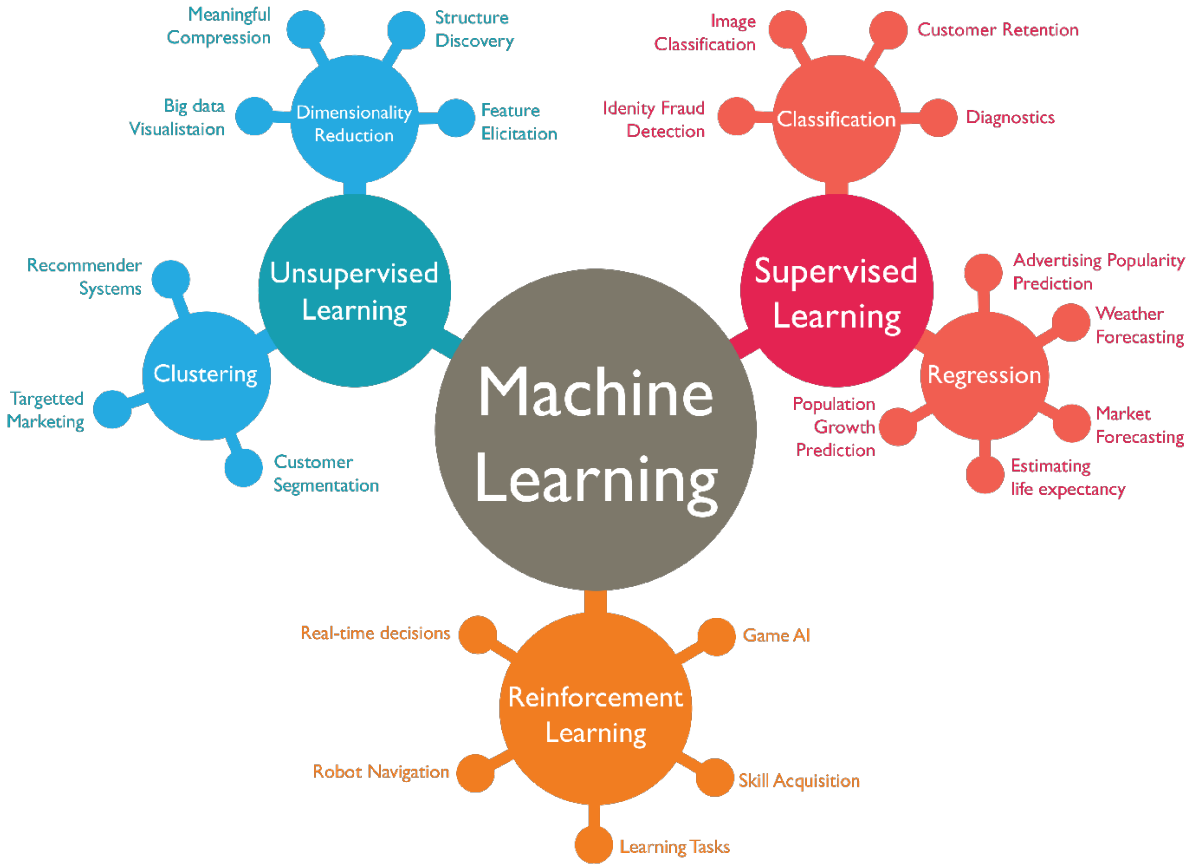
INTRODUCCION AL MACHINE LEARNING

El objetivo de esta sesion es revisar y tener claros los conceptos basicos del aprendizaje automatico (Machine Learning) con el objeto de conocer la terminologia necesaria para las proximas sesiones.

ÍNDICE

OBJETIVO....4
INTRODUCCIÓN....5
¿QUÉ ES CIENCIA DE DATOS?....6
¿CUÁL ES EL CICLO DE VIDA DE LA CIENCIA DE DATOS?....7
¿QUÉ ES EL MACHINE LEARNING?....10
MACHINE LEARNING: CONSTRUCCION DEL MODELO....11
MACHINE LEARNING: APRENDIZAJE SUPERVISADO....12
MACHINE LEARNING: APRENDIZAJE SUPERVISADO....13
MACHINE LEARNING: APRENDIZAJE NO SUPERVISADO....14
MACHINE LEARNING: APRENDIZAJE NO SUPERVISADO....15
MACHINE LEARNING: APRENDIZAJE REFORZADO....16
MACHINE LEARNING: FEATURE ENGINEERING....17
MACHINE LEARNING: FEATURE ENGINEERING....18
MACHINE LEARNING: FEATURE ENGINEERING....19
MACHINE LEARNING: FEATURE ENGINEERING....20
MACHINE LEARNING: TRAIN, VALIDATION Y TEST SETS....21
MACHINE LEARNING: OVERFITTING....22
MACHINE LEARNING: UNDERFITTING....23
MACHINE LEARNING: BIAS-VARIANCE TRADEOFF....24
MACHINE LEARNING: EVALUACION DEL MODELO....25
MACHINE LEARNING: FUNCION DE PERDIDA Y FUNCION DE COSTO....26
MACHINE LEARNING: EVALUACION DEL MODELO....27
MACHINE LEARNING: METRICAS DE REGRESION....28
MACHINE LEARNING: METRICAS DE CLASIFICACION....29
PRACTICA LABORATORIO....30
ANEXO: INSTALACION DE ANACONDA EN WINDOWS....31
ANEXO: INSTALACION DE ANACONDA EN WINDOWS....32
ANEXO: INSTALACION DE ANACONDA EN WINDOWS....33
ANEXO: INSTALACION DE ANACONDA EN WINDOWS....34

Objetivo

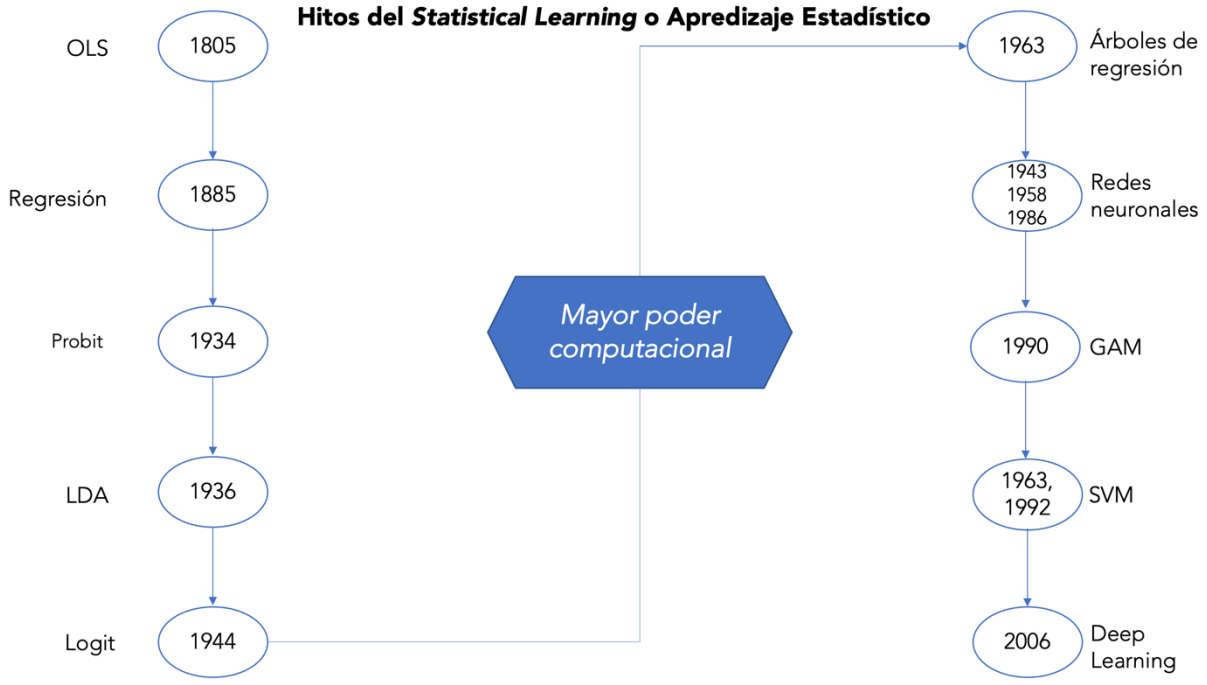


El objetivo de esta semana es conocer la terminologia, principales algoritmos y areas del Machine Learning asi como las herramientas que nos permitiran trabajar en nuestras practicas semanales del curso.

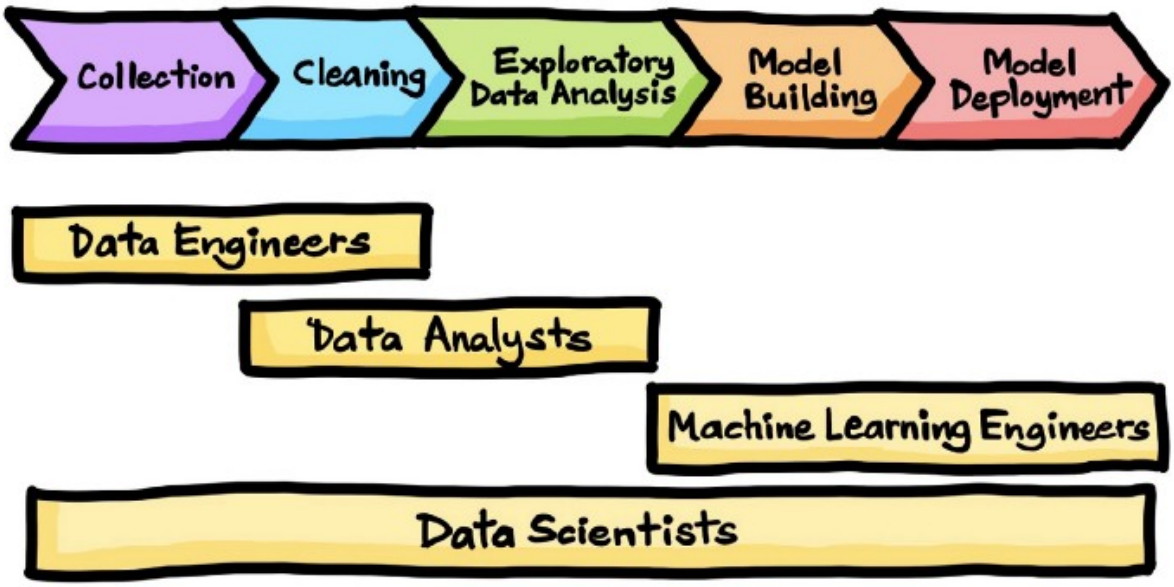
En esta primera sesion la practica de laboratorio consistira en conocer las interfaces graficas que nos facilitaran el procesamiento y visualizacion de nuestros resultados asi como las principales librerias de Python para nuestro curso, como son: numpy, pandas, scikit learn, matplotlib.

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.

¿Cuál es el ciclo de vida de la Ciencia de Datos?

https://towardsdatascience.com/the-data-science-process-a19eb7ebc41b

- 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.
- **Recoleccion 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?
- Analisis 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.

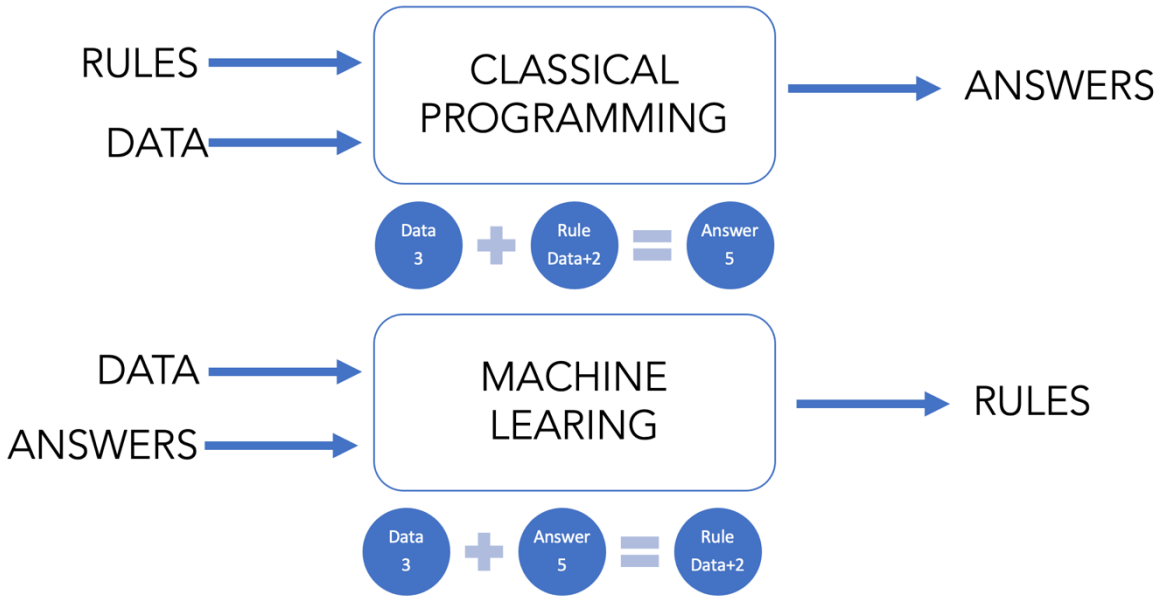
  Recuerde que la calidad de sus entradas decide la calidad de su salida. Por lo tanto, una vez que tenga lista su hipótesis comercial, tiene sentido dedicar mucho tiempo y esfuerzo aquí.
- Modelamiento de datos, modelamiento machine learning, esta etapa parece ser la más interesante para casi todos los 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.

- Evaluacion 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 deben 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.
- Visualizacion 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.

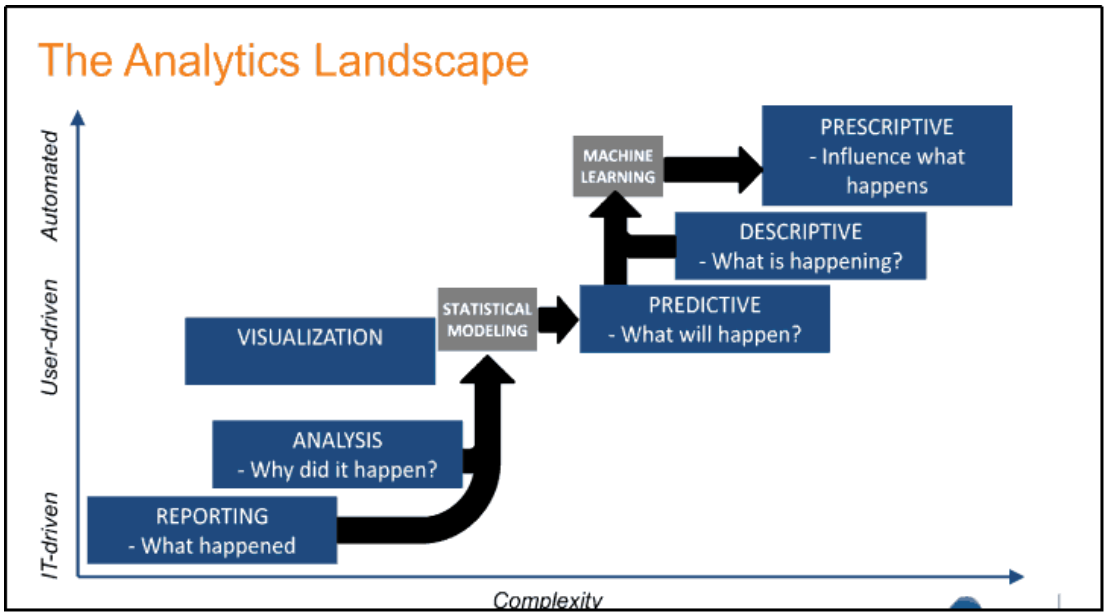
- **Despliegue de modelo,** después de construir modelos, primero se implementa en un entorno de preproducción o prueba antes de implementarlos en producción. Cualquiera que sea la forma en que se implemente su modelo de datos, debe exponerse al mundo real. Una vez que los humanos reales lo usen, seguramente recibirás comentarios. Capturar esta retroalimentación se traduce directamente en la vida o la muerte para cualquier proyecto.

¿Qué es el Machine Learning?



- Es una rama de la inteligencia artificial
- Es la capacidad de las maquinas 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 ultimo es una habilidad indispensable para contruir 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: Construccion del modelo



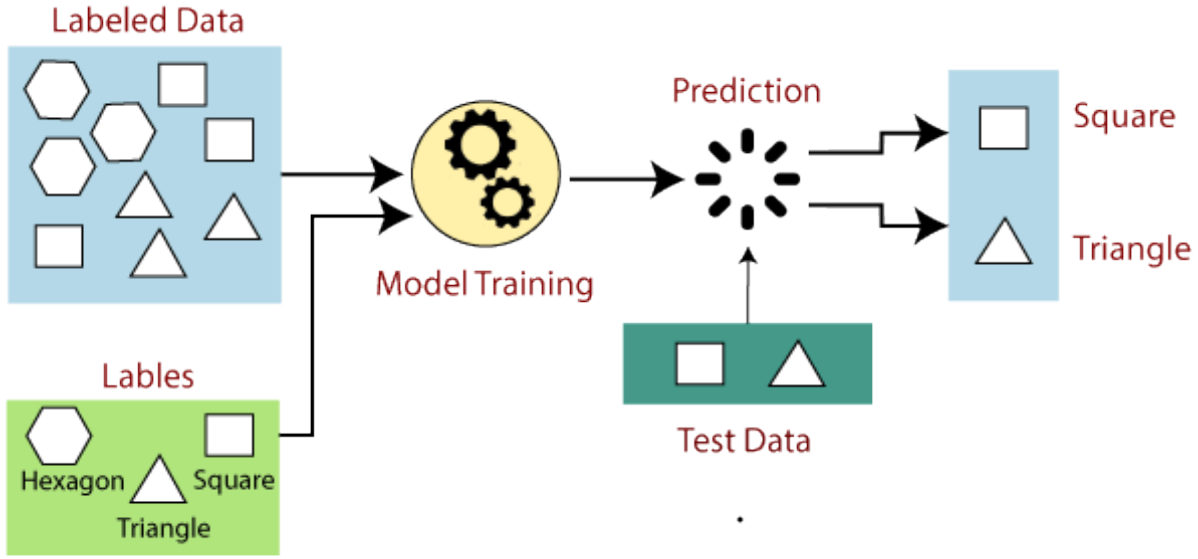
- El modelado se utiliza para encontrar patrones o comportamientos en los datos. Estos patrones nos ayudan de una de dos maneras:

1 **Modelado descriptivo** (aprendizaje no supervisado): sistemas de recomendación que son si a una persona le gustó la película Matrix, también le gustaría la película Inception.

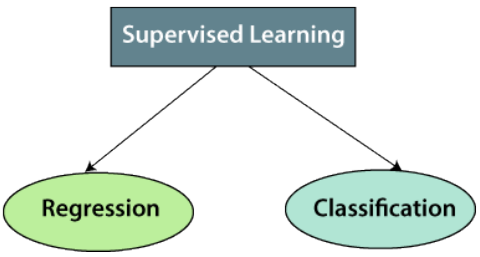
2 **Modelado predictivo** (aprendizaje supervisado): esto implica obtener una predicción sobre tendencias futuras, por ejmplo regresión lineal en la que podríamos querer predecir los valores bursátiles.

	Supervised	Unsupervised
Discrete	Classification or Categorization	Clustering
Continuous	Regression	Dimensionality Reduction

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:



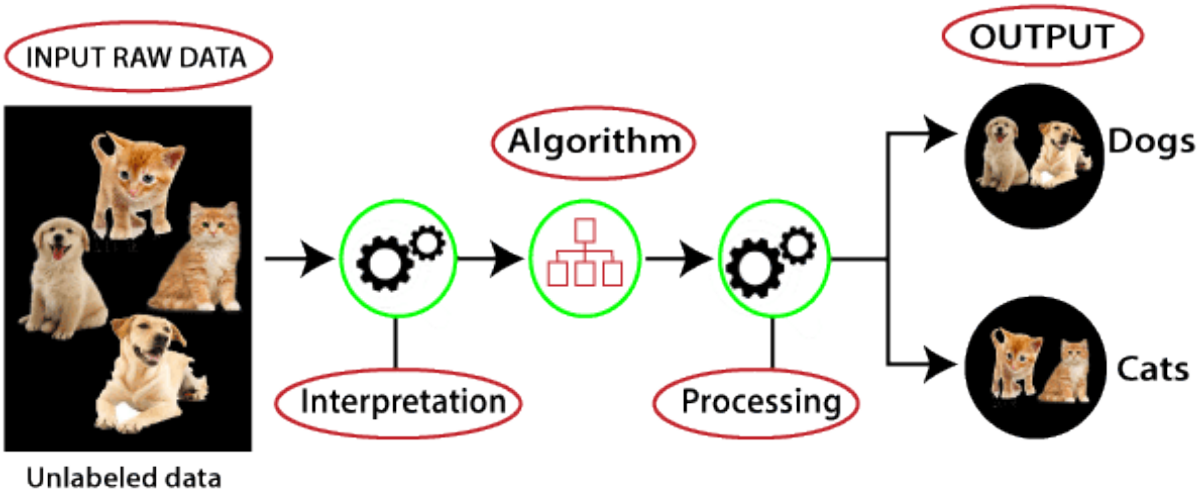
- Algunos algoritmos supervisados:

1. Naive Bayes (Clasificacion, modelo no lineal)
2. Neural Networks

Machine Learning: Aprendizaje supervisado

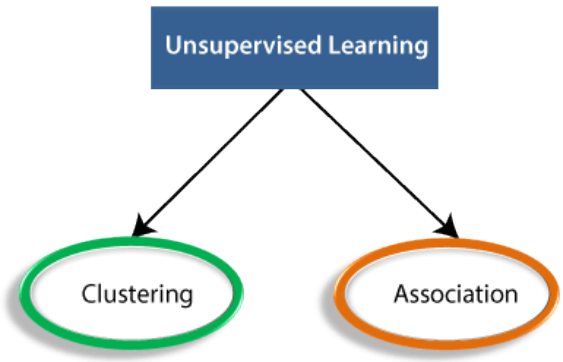
3. k-Nearest Neighbor (kNN) (Clasificacion, modelo no lineal)
4. Linear Regression (Regresion)
5. Logistic Regression (Clasificacion, modelo lineal)
6. Support Vector Machines(SVM) (Clasificacion, modelo lineal)
7. Decision Trees (Clasificacion, modelo no lineal)
8. Random Forest (Clasificacion, modelo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

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:

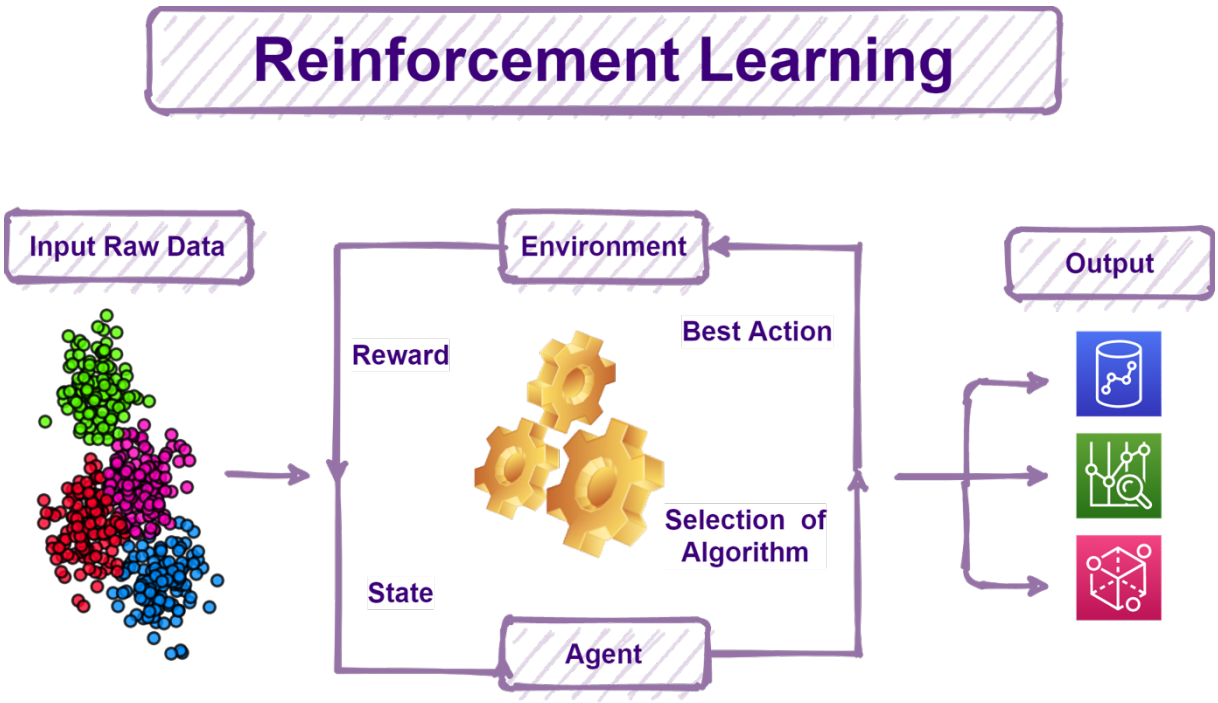


Machine Learning: Aprendizaje no supervisado

- Algunos ejemplos de algoritmos no supervisados:

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

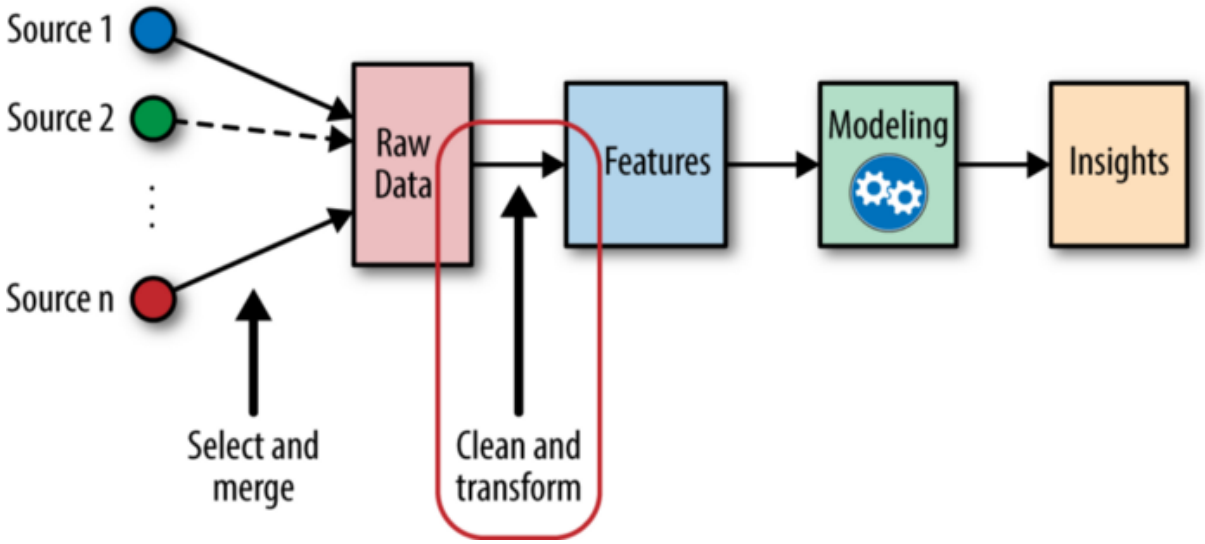
Machine Learning: Aprendizaje reforzado



- 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.

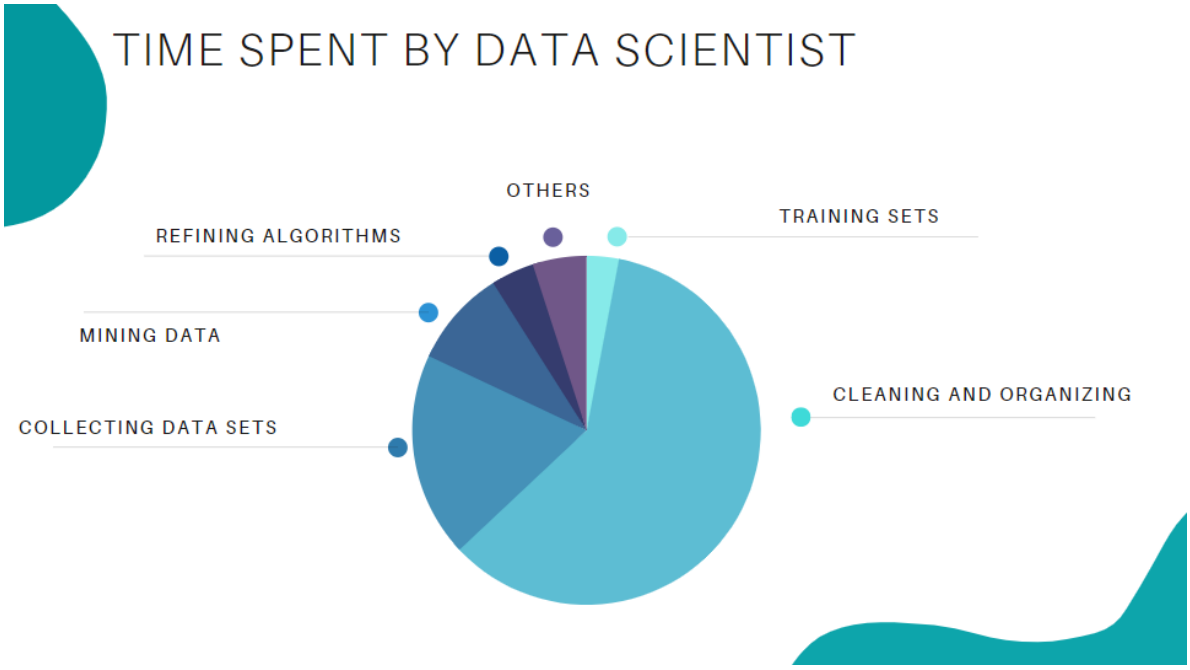
Machine Learning: Feature Engineering



- 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.

Machine Learning: 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.



Machine Learning: 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.

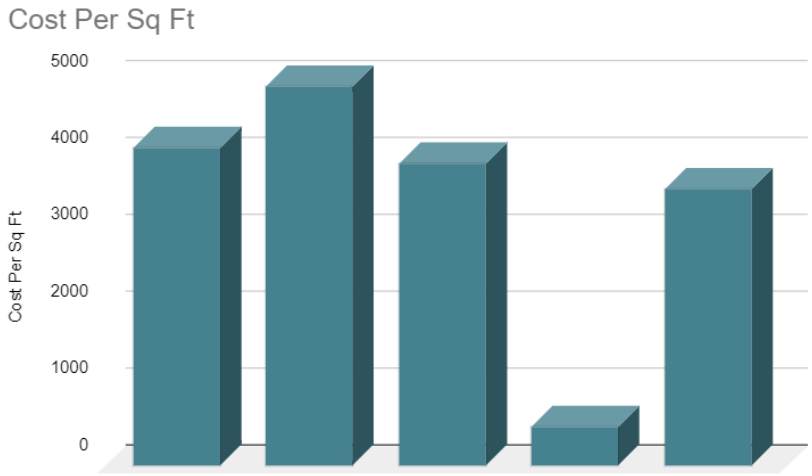
Sq Ft.	Amount
2400	9 Million
3200	15 Million
2500	10 Million
2100	1.5 Million
2500	8.9 Million

- 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.

Machine Learning: 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

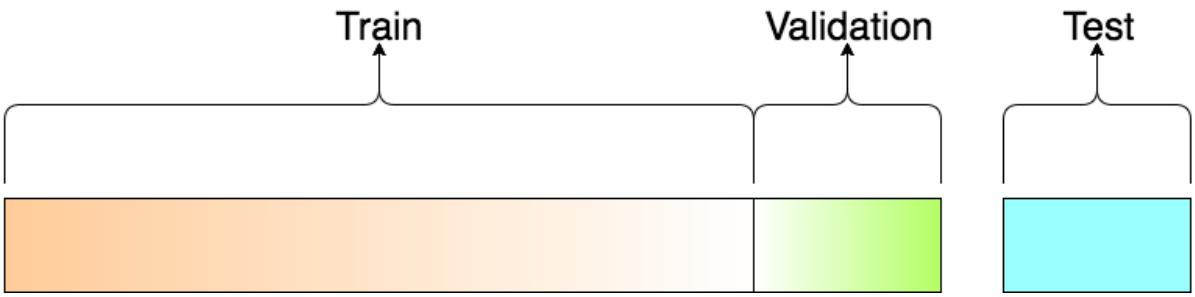
  Transformations

  Feature Extraction

  Exploratory Data Analysis

  Benchmark

Machine Learning: Train, validation y test sets

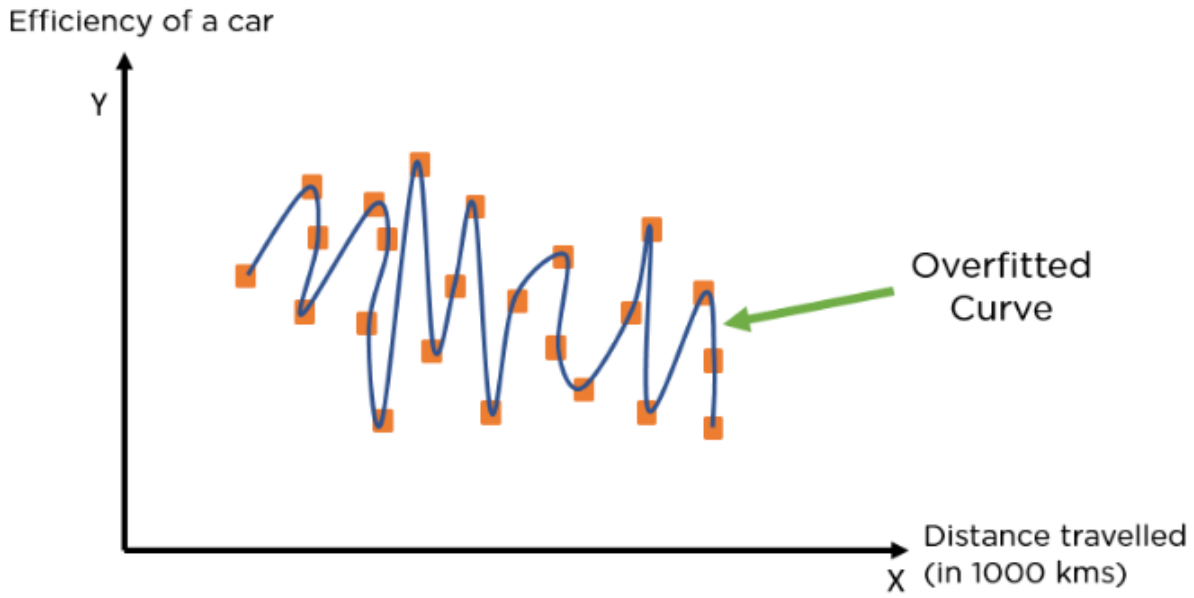


- **Conjunto de entrenamiento:** Conjunto de ejemplos utilizados para el aprendizaje. El conjunto de entrenamiento suele ser el conjunto más grande, en términos de tamaño, que se crea a partir del conjunto de datos original y se utiliza para encontrar el modelo. En otras palabras, los puntos de datos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

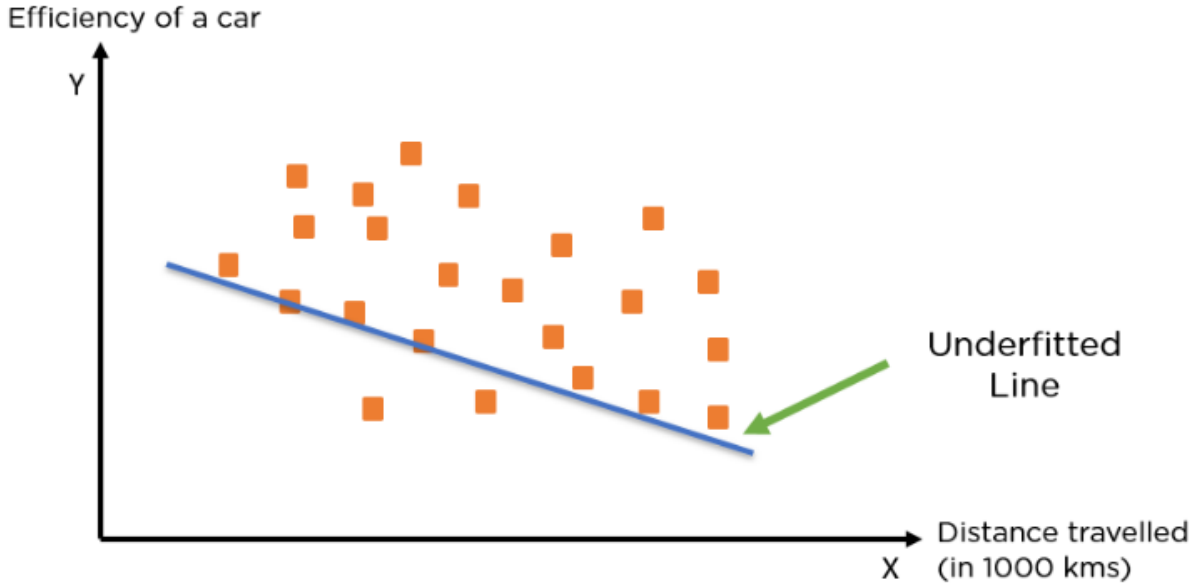
- **Conjunto de prueba:** un conjunto de ejemplos utilizados solo para evaluar el rendimiento de un clasificador completamente especificado.
  El conjunto de prueba se utiliza para evaluar el rendimiento de este modelo y garantizar que pueda generalizarse bien a puntos de datos nuevos e invisibles.

Machine Learning: Overfitting



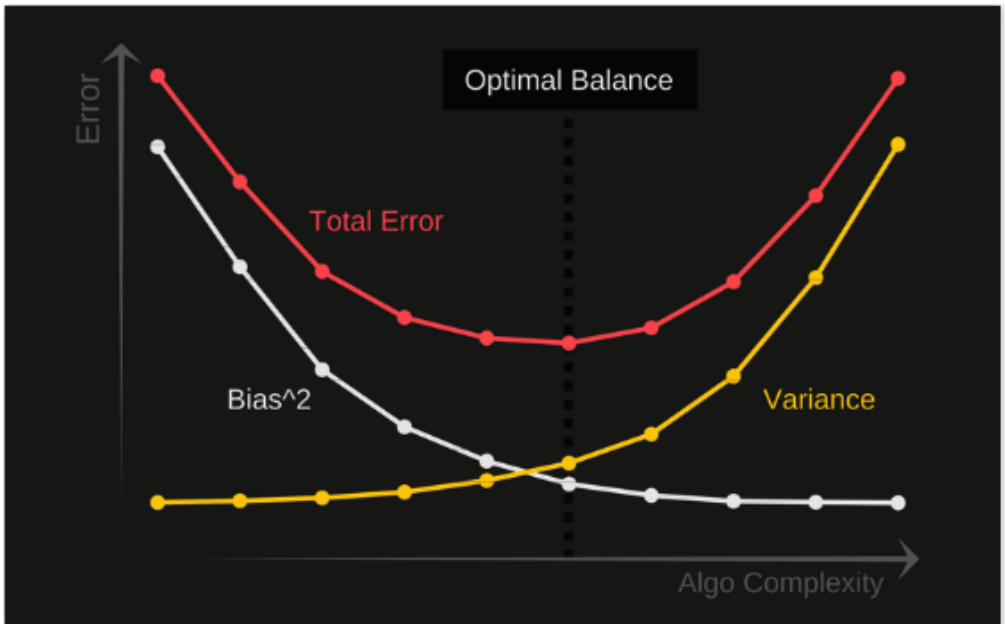
- Cuando un modelo funciona muy bien para los datos de entrenamiento pero tiene un rendimiento deficiente con los datos de prueba (datos nuevos), se conoce como sobreajuste. En este caso, el modelo de aprendizaje automático aprende los detalles y el ruido en los datos de entrenamiento de manera que afecta negativamente el rendimiento del modelo en los datos de prueba. El sobreajuste puede ocurrir debido al bajo sesgo y la alta varianza.

Machine Learning: Underfitting



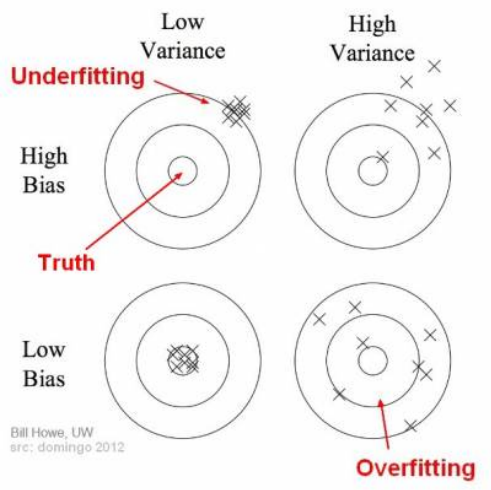
- Cuando un modelo no ha aprendido bien los patrones en los datos de entrenamiento y no puede generalizar bien los nuevos datos, se conoce como ajuste insuficiente. Un modelo inadecuado tiene un rendimiento deficiente en los datos de entrenamiento y dará como resultado predicciones poco confiables. El ajuste insuficiente ocurre debido al alto sesgo y la baja varianza.

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: Evaluacion del Modelo

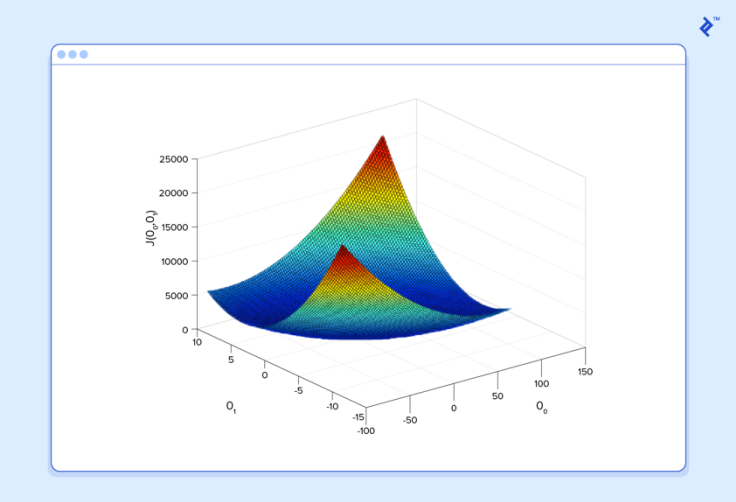


- En el diagrama anterior, el centro del objetivo es un modelo que predice perfectamente los valores correctos. A medida que nos alejamos de la diana, nuestras predicciones empeoran cada vez más. Podemos repetir nuestro proceso de creación de modelos para obtener resultados separados en el objetivo.

- 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.

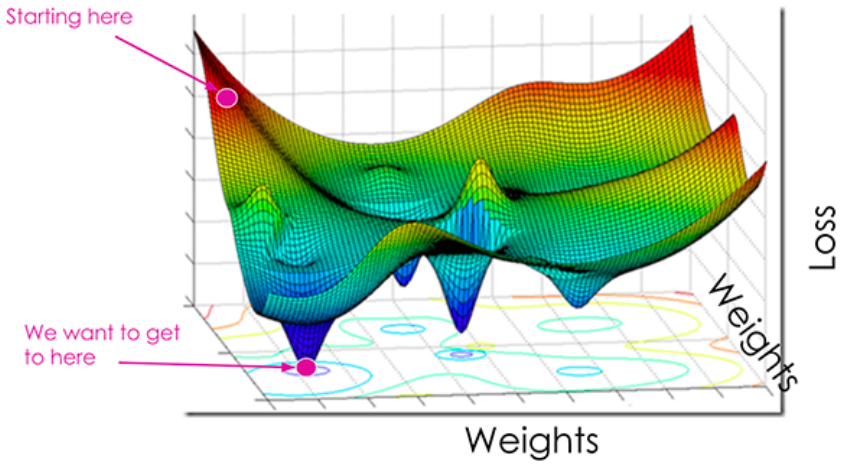
- En el aprendizaje supervisado, el *overfitting* ocurre cuando nuestro modelo captura el ruido junto con el patrón subyacente en los datos. Ocurre cuando entrenamos mucho nuestro modelo sobre un conjunto de datos ruidoso. Estos modelos tienen un sesgo bajo y una varianza alta.

Machine Learning: Funcion de Perdida y Funcion 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 de el 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.
- Funcion de costo es el promedio de todos los errores de la muestra en todo el conjunto de entrenamiento.

Machine Learning: Evaluacion del Modelo

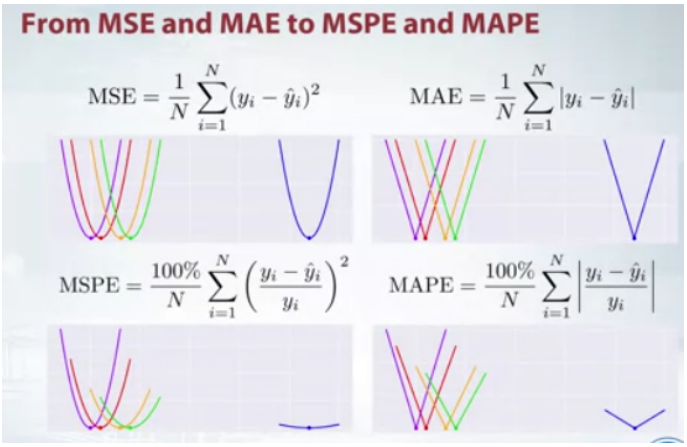


- Imagínese esto: ha entrenado un modelo de aprendizaje automático en un conjunto de datos determinado y está listo para ponerlo frente a su cliente. Pero, ¿cómo puede estar seguro de que este modelo dará el resultado óptimo? ¿Existe alguna métrica o técnica que lo ayude a evaluar rápidamente su modelo 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: Metricas de 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:

  Error absoluto medio (MAE)
  Error cuadrático medio (MSE)
  Error cuadrático medio de la raíz (RMSE)
  MAPE

Machine Learning: Metricas de 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
  Matriz de confusión (no es una métrica pero es fundamental para otros)
  Precisión y recuperación
  Recall
  Puntuaje F1
  AU-ROC

Practica Laboratorio

1. Se realizara en cada sesion una practica de los temas
2. Se entregara en cada clase una practica para realizar en el domicilio durante la semana.

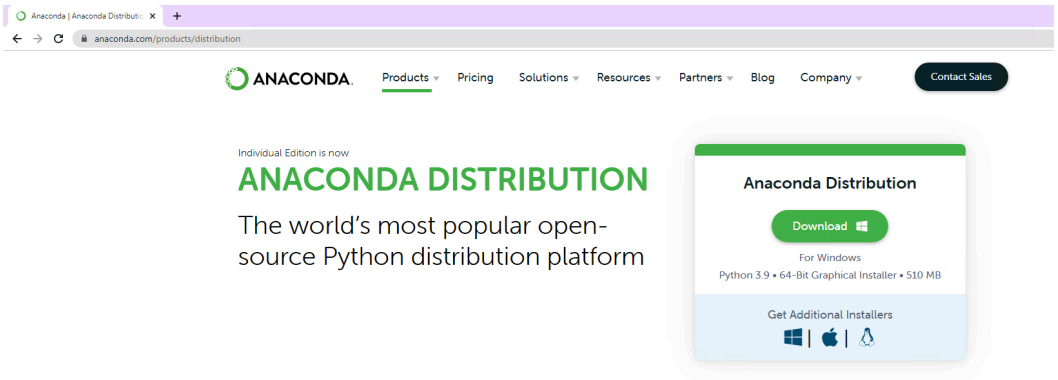
Anexo: Instalacion de Anaconda en Windows

- 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.).

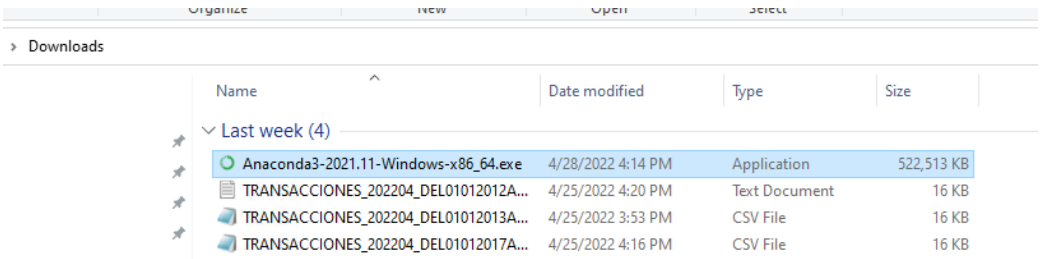
  Tiene como ventaja **simplificar la gestión e implementación de paquetes**. La distribución incluye paquetes de "data science" adecuados para Windows, Linux y macOS.

- Para instalar Anaconda ingresar a la pagina siguiente:

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

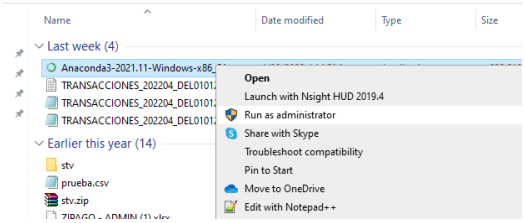


- Descargar el instalador presionando en **Download**

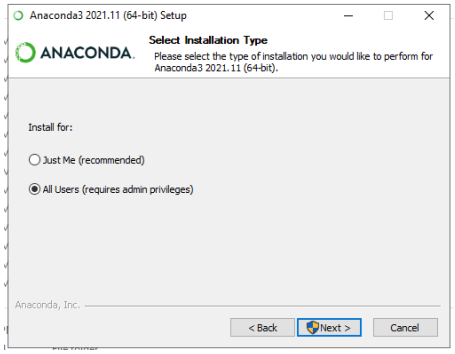


Anexo: Instalacion de Anaconda en Windows

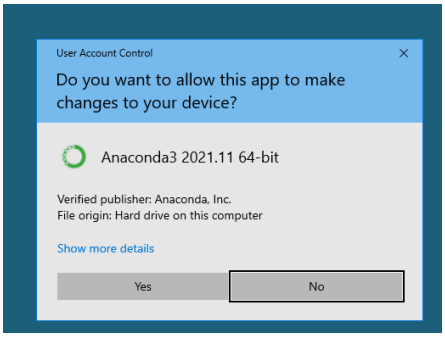
- Una vez descargado ejecutar el instalador **como usuario administrador**:



Presionar Next

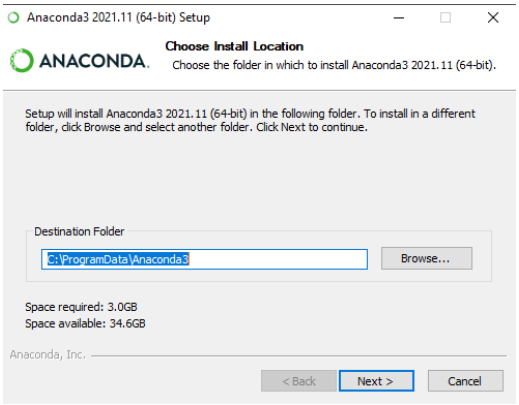


Presionar Next

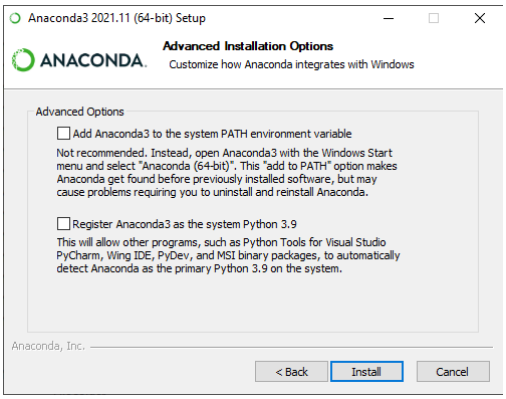


Presionar Yes

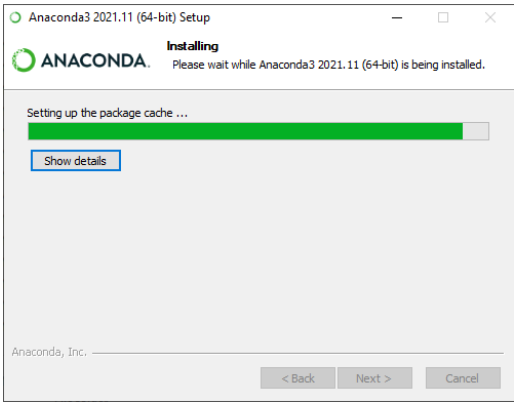
Anexo: Instalacion de Anaconda en Windows



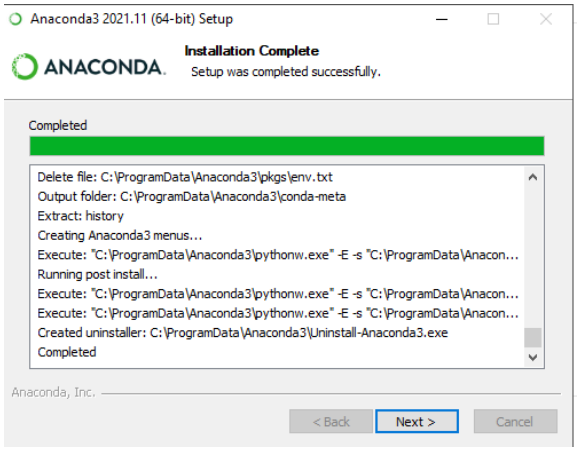
Presionar Next



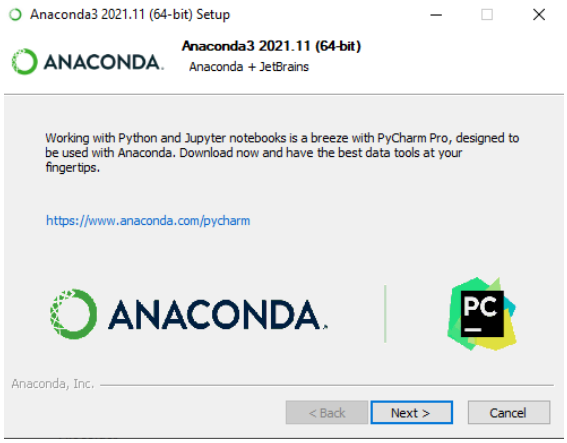
Presionar Install



Anexo: Instalacion de Anaconda en Windows



Presionar Next



Presionar Next