

El algoritmo ideal

**Breve descripción:**

Este componente formativo aborda aspectos generales y claves sobre los principales algoritmos de clasificación que se utilizan en modelos “Machine Learning”, los cuales contribuyen a la solución en diversas problemáticas, a través del desarrollo de sistemas de automatización. Con su estudio responsable, el aprendiz podrá realizar el preprocesamiento de los datos, realizar su depuración, entendimiento y establecer la variable objetivo.

**Noviembre 2023**

Tabla de contenido

[Introducción 1](#_Toc149835301)

[1. Creación de modelos de “Machine Learning” 3](#_Toc149835302)

[1.1. Algoritmos de clasificación 9](#_Toc149835303)

[1.2. K-vecinos cercanos 12](#_Toc149835304)

[1.3. Regresión logística 19](#_Toc149835305)

[1.4. Árboles de decisión 19](#_Toc149835306)

[1.5. Bosques aleatorios 25](#_Toc149835307)

[1.6. “Naive Bayes” 28](#_Toc149835308)

[1.7. Otros algoritmos 29](#_Toc149835309)

[2. Selección y preprocesamiento de datos 34](#_Toc149835310)

[3. Métricas de evaluación 41](#_Toc149835311)

[Síntesis 46](#_Toc149835312)

[Material complementario 48](#_Toc149835313)

[Glosario 49](#_Toc149835314)

[Referencias bibliográficas 50](#_Toc149835315)

[Créditos 51](#_Toc149835316)

Introducción

Se da la bienvenida a este componente formativo denominado “El algoritmo ideal”. Para comenzar el recorrido por el mismo, consulte el video que se propone a continuación:

1. El algoritmo ideal



[**Enlace de reproducción del video**](https://youtu.be/ZgkwSKyGpnY)

|  |
| --- |
| **Síntesis del video: El algoritmo ideal** |
| ¿Sabe por qué la automatización de datos es considerada como uno de los procesos fundamentales para el éxito de una organización?  La organización, selección, jerarquización y análisis de datos son, tal vez, la mejor estrategia que puede implementar una empresa que quiera sobresalir frente a la competencia. Por ello es clave conocer el modelo “Machine Learning”: una de las disciplinas más utilizada para que una empresa pueda tomar decisiones acertadas a partir de la aplicación de algoritmos y el análisis de variables.  Esta disciplina hace parte del campo de la inteligencia artificia y permite identificar patrones masivos y elaborar predicciones, es decir, se realizan tareas específicas de manera autónoma, sin la necesidad de la presencia de un programador.  Está inmersa en la transformación digital y es utilizada para encontrar soluciones en campos de la salud, la economía, la administración, entre otros. |

# Creación de modelos de “Machine Learning”

Un modelo de “Machine Learning” es:

* Un archivo que se entrena con el fin de identificar ciertos tipos de patrones.
* Dicho entrenamiento se realiza a través de una colección de datos y un algoritmo que lo que hace es buscar información de los datos.
* Al entrenar el modelo, este va “aprendiendo” de los datos.
* Por ejemplo, si se lleva al contexto cuando un profesor le muestra el color a un niño, pero, además, le dice el nombre de ese color, el niño aprende, y la próxima vez que vea el color ya lo identificará y podrá reconocerlo.

Cuando el modelo es entrenado, puede usarse para describir datos desconocidos; por ejemplo, si se desea crear una aplicación que permita detectar las emociones del cliente de una empresa, estas emociones se identificarían de acuerdo con sus expresiones faciales. Para realizar el entrenamiento, se le indican a la máquina cientos de imágenes con rostros de diferentes emociones y se le indica qué emoción corresponde a cada imagen; este modelo, creado a partir de este entrenamiento, aprende de los datos y se podría usar en una aplicación que identifique las emociones de cualquier rostro.

Algunas generalidades sobre el “Machine Learning”, las cuales usted debe tener presentes, son:

* **Afianzamiento.** “Machine Learning” es un término que se empieza a escuchar, cada vez, con más frecuencia; es una dependencia de la Inteligencia Artificial (AI).
* **Intencionalidad.** Su principal objetivo es la creación de sistemas que estén en la capacidad de aprender, de manera autónoma, a partir de una colección de datos.
* **Análisis de datos.** Lo anterior, permite tener la percepción y análisis de un gran volumen de datos, lo que facilita muchos procesos del ser humano.
* **Entrenamiento.** Para la construcción de dichos modelos, se requiere, como base fundamental, una colección de datos desde la cual el algoritmo seleccionado empieza su proceso de entrenamiento.
* **Predicciones.** Posteriormente, se realizan las predicciones de acuerdo con una variable objetivo.
* **Algoritmos de clasificación.** En este componente formativo, para la creación de los modelos, se utilizarán específicamente algoritmos que denominamos de clasificación, que están dentro de la categoría de aprendizaje supervisado.

En el aprendizaje supervisado, se le indica a la máquina información histórica para que realice el entrenamiento y aprenda de esa información; una vez ha aprendido, se estaría en la capacidad de realizar el reconocimiento y predecir sin necesidad de saber la respuesta.

Este tipo de aprendizaje toma como base el comportamiento humano, en cual se está bajo la supervisión de un profesor que proporciona a sus estudiantes un número importante de buenos ejemplos para que sean memorizados y, posteriormente, el alumno pueda generar sus propias reglas para estos ejemplos ya aprendidos.

El siguiente, es un ejemplo que permite entender lo anteriormente expuesto; preste suma atención:

* **Entender el modelo.** Se entrenará un modelo teniendo en cuenta una colección de imágenes de animales, en la cual se especifica el nombre de cada uno.
* **Establecer la colección de datos.** Se establece la colección de datos, la cual está formada por una imagen y el nombre asociado a esa imagen; estos datos son procesados por los algoritmos de lenguaje supervisado que le permitirán al modelo aprenderlos.
* **Realizar el proceso de entrenamiento.** Una vez se realice el proceso de entrenamiento, se creará un modelo capaz de identificar una imagen sin que se deba especificar el significado, ya que este ha realizado su aprendizaje con los datos que previamente se le han introducido.

De acuerdo con lo explicado anteriormente, se puede decir que:

* Un modelo de “Machine Learning” es el resultado que se obtiene cuando se realiza el entrenamiento a un algoritmo con datos.
* Después de realizar el entrenamiento, al indicarle un modelo con una entrada, también se le proporciona una salida.
* Por ejemplo, cuando se usa un algoritmo predictivo, es decir, aquel que permite aprender de un conjunto de datos y que, al enfrentarlo a nuevos datos, podría determinar o predecir de qué se tratan, se creará un modelo predictivo.

Es muy importante reafirmar lo dicho hasta ahora. El video que se propone enseguida, destaca los aspectos más relevantes sobre “Machine Learning”; explórelo atentamente:

1. ¿Qué es “Machine Learning”?



[**Enlace de reproducción del video**](https://youtu.be/J9w6KquPKbE)

|  |
| --- |
| **Síntesis del video: ¿Qué es “Machine Learning”?** |
| Es una disciplina de las ciencias informáticas y las neurociencias que parte de la inteligencia artificial y hace referencia a la capacidad que tiene una maquina o un software para aprender de los datos ingresados mediante la adaptación de ciertos algoritmos de su programación, con el fin de reducir la necesidad de que intervengan los seres humanos.  Es importante mencionar que la máquina no aprende por sí misma, sino que por medio del diseño de algoritmos y el desarrollo de modelos de datos sobre activos de su programación, se establecen los parámetros para que ejecuten tareas tanto generales, como específicas, además de predecir escenarios futuros o tomar acciones de manera automática, según ciertas condiciones previamente establecidas.  Actualmente, el “Machine Learning” se utiliza para el razonamiento probabilístico, investigación, recuperación de información, modelos estadísticos y reconocimiento de patrones, además de abordar y resolver problemas prácticos. |

Para empezar a crear los modelos, se requiere definir muy bien la variable objetivo; por ejemplo, saber qué estudiantes pasarán o perderán un curso, qué pacientes de una clínica padecerán cáncer o qué clientes pueden abandonar los servicios que les ofrece una empresa.

Para crear un modelo de “Machine Learning”, hay que tener en cuenta algunas consideraciones, como las que aquí se sugieren:

* **Recolectar datos.** De acuerdo con cada problema, se debe realizar una muy buena investigación y obtención de datos, garantizando que sean de muy buena calidad y cantidad, pues esto determinará su eficiencia.
* Dicha información puede venir de bases de datos privadas o públicas, las cuales proveen información que puede ser descargada en formato CSV o compartida mediante API con estructuras tipo JSON; también puede hacer uso de la técnica “Web Scraping”, que consiste en conseguir información publicada en páginas web mediante un “software” especializado que se va almacenando en tablas, de manera “online” o local.
* **Preparar datos.** Este paso permite hacer visualizaciones de los datos para determinar si existen correlaciones entre las diversas variables; se puede realizar corrección de valores nulos, datos atípicos, conversión de datos categóricos a numéricos, establecer si es necesario la creación de columna “dummy”, normalización de los datos y corrección de errores. Es importante hacer un mezclado y separación de los datos, pues de estos normalmente se utiliza 80 % para realizar el entrenamiento y 20 % se dejan para realizar las pruebas al modelo.
* **Escogiendo el modelo.** Son múltiples los modelos que existen y que se pueden escoger dependiendo de la problemática u objetivo que se quiera resolver.
* **Entrenando la máquina.** Luego de seleccionar el modelo, se realiza el entrenamiento. En este proceso, se deben visualizar mejoras. Es importante inicializar los pesos del modelo aleatoriamente, ya que estos afectan la relación entre la entrada y la salida. Los pesos se ajustan automáticamente a medida que se realizan más entrenamientos. Se deben observar todas las respuestas obtenidas y proceder con las correcciones necesarias.
* **Evaluación.** Lo que se busca con la evaluación es obtener la precisión del modelo entrenado, en la cual se verifica su exactitud; dependiendo del requerimiento del negocio, se aceptarán diferentes porcentajes, lo ideal es siempre buscar un nivel de exactitud superior al 90 %.
* **Configuración de parámetros.** Con los resultados obtenidos, se determina si se realizaron buenas predicciones; en caso de no obtener el resultado esperado, es muy probable que se estén presentando problemas de “overfitting” o de “underfitting”, y se deba retroceder a entrenar nuevamente el modelo.
* Se podría pensar en aumentar el número de iteraciones del modelo o realizar ajustes al valor máximo de error permitido, entre otros. Cada algoritmo tiene sus propias configuraciones, que es necesario conocer para que puedan ser ajustadas a determinadas situaciones.
* **Predicción.** En este punto, ya se han logrado buenas evaluaciones y el modelo debería estar listo para empezar a predecir datos con información completamente nueva.

## Algoritmos de clasificación

Este tipo de algoritmos es utilizado en problemas cuyos resultados estén basados en la obtención de etiquetas discretas, lo que quiere decir que las respuestas a una pregunta o problema están alojadas dentro de una colección de datos finita; por ejemplo, en la detección de correo malicioso, se puede detectar si es “spam” o no, o en la calificación de servicio de un restaurante se puede determinar si es bueno, malo o excelente; en otras palabras, el resultado está enmarcado en un rango de salidas definido.

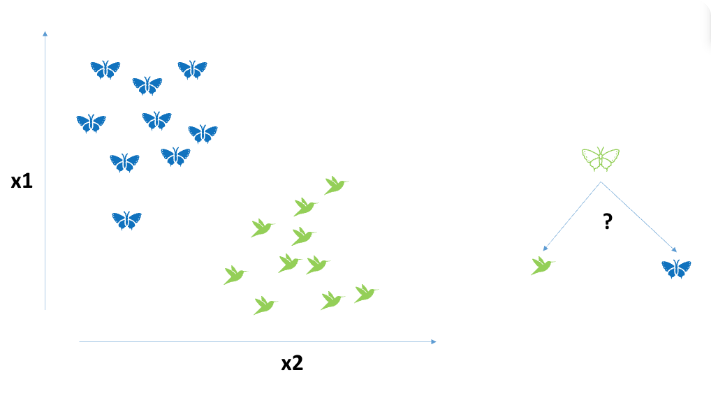
Cuando el modelo es entrenado para obtener resultados del tipo verdadero o falso, se le conoce como clasificación binaria. Por ejemplo, se podrían crear modelos de clasificación con respuestas binarias para predecir si un alumno aprueba un curso o no, o los clientes comprarán un producto nuevo o no; este tipo de etiquetas se representan con un 0 para falso y 1 para verdadero.

Así mismo:

* Estos algoritmos también pueden implementarse cuando se requiera predecir más de dos objetivos.
* En tal caso, se denominan de clasificación multicategoría, y son usados en la segmentación de clientes, en la categorización de imágenes y audios.
* Otro uso que se les puede dar es el de analizar textos que ayuden a establecer cuál es el sentimiento de una persona.
* La forma en cómo trabajan los algoritmos consiste en entregarles una colección de muestras, patrones, ejemplos o prototipos base, para que se tomen como la representación de las clases o variables.
* En esta misma colección, se relacionan con la etiqueta de clase correcta las colecciones, valga la redundancia, que se relacionan correctamente con la etiqueta resultado, a esto se le denomina conjunto de entrenamiento.
* El algoritmo aprende de ese conjunto de datos, ya que se le indica qué variables se asocian al resultado y, con ese conocimiento, se estará en la capacidad de predecir el resultado para nuevas entradas con las que el modelo nunca haya tenido relación.

La siguiente gráfica, muestra cómo se clasifican los datos. Suponga que el modelo se entrenó con mariposas, colibríes y colores, estos son clasificados de acuerdo con sus características. Cuando el modelo se enfrente a nuevos datos con características similares al modelo entrenado, pero que no conoce este, estaría en la capacidad de identificarlos asociándolos a una clase, ya sea por el color o por la figura de la imagen.

1. Clasificación de datos



Los conjuntos mostrados en la figura (mariposas y colibríes), y la conformación de otro conjunto donde aparece un colibrí más una mariposa, sugieren la importancia de entender que cuando se realiza la clasificación, se está realizando el reconocimiento de patrones; los modelos aprenden de los valores y buscan patrones, de tal manera que, cuando se enfrentan a un valor nuevo, se clasifica según sus características más cercanas.

Los principales algoritmos de clasificación son:

* K-vecinos cercanos.
* Regresión logística.
* Árboles de decisión.
* Bosques aleatorios.
* “Naive Bayes”.
* Otros algoritmos.

## K-vecinos cercanos

“K-nearest-neighbor” es un algoritmo que se basa en el aprendizaje supervisado, el cual, se utiliza para clasificar muestras nuevas de valores discretos o realizar predicciones en valores continuos. Es un algoritmo fácil de entender, lo que lo hace ideal para principiantes en el mundo del “Machine Learning”.

Este algoritmo funciona comparando los datos nuevos con los datos de entrenamiento, encontrando, los puntos o características más parecidas a los datos nuevos. De esta forma, se identifica cuáles han sido aprendidas durante el proceso de entrenamiento.

En aprendizaje no supervisado, se encuentra un algoritmo que se denomina “k-means”, el cual se debe diferenciar, ya que, para este, la k tiene que ver con el número de grupos que se desean clasificar, mientras que en el algoritmo Knn (“K-nearest neighbor”), la k tiene que ver con el número de puntos vecinos que se deben tener en cuenta, por su cercanía, para realizar la clasificación en n grupos que sean conocidos previamente, pues este se trata de un algoritmo basado en la supervisión.

La técnica “K-nearest neighbor” intenta predecir realizando una clasificación del dato tomando como base los datos que lo rodean.

En un algoritmo supervisado, la colección de datos con los que se entrena está etiquetada con una clase o resultado esperado.

Por otro lado, un algoritmo basado en instancia no realiza un aprendizaje explícitamente de un modelo, (como lo haría el algoritmo de árbol de decisión o el de regresión lineal); lo que realiza es un entrenamiento, memorizando las instancias que serán la base principal de conocimiento, que se utilizará para predecir nuevos datos.

Comprenda, con mayor propiedad, los aspectos mencionados hasta este punto; observe la ejemplificación que se propone enseguida:

* **Planteamiento.** Para una colección de datos donde cada instancia se refiere a un cliente de un banco, en la cual se conocen para cada uno sus características o atributos, como pueden ser edad, patrimonio, ingresos, gastos, hijos, entre otros; además, se tiene un atributo o clase objetivo que indica si cada uno de ellos puede ser candidato o no para realizarle un préstamo, el atributo o clase objetivo es el que se intentará predecir.
* **Situación.** Para realizar el modelo, se divide la colección de datos en dos partes y se deben seleccionar de manera aleatoria; la parte del entrenamiento con el 80 % de los datos y se deja otro 20 % con el resto de los datos para hacer las pruebas que ayudarán a validar el modelo.
* **Acción.** Se debe establecer un valor para k, preferiblemente que sea pequeño; luego se le pasa al algoritmo una instancia de la colección que se ha separado para pruebas, para que sea procesado. Tras realizar este proceso, el algoritmo se encarga de seleccionar las k instancias de la colección de datos de entrenamiento que más se parezcan o sean más cercanas, dicha cercanía se establece dependiendo de la métrica de similitud que se utilice; la asignación se realiza a la instancia de la clase con mayor frecuencia dentro de las k instancias que han sido seleccionadas como más cercanas.

El algoritmo “K-Nearest-Neighbor” es un algoritmo muy simple, pero con el que se obtienen buenos resultados.

Para tener en cuenta:

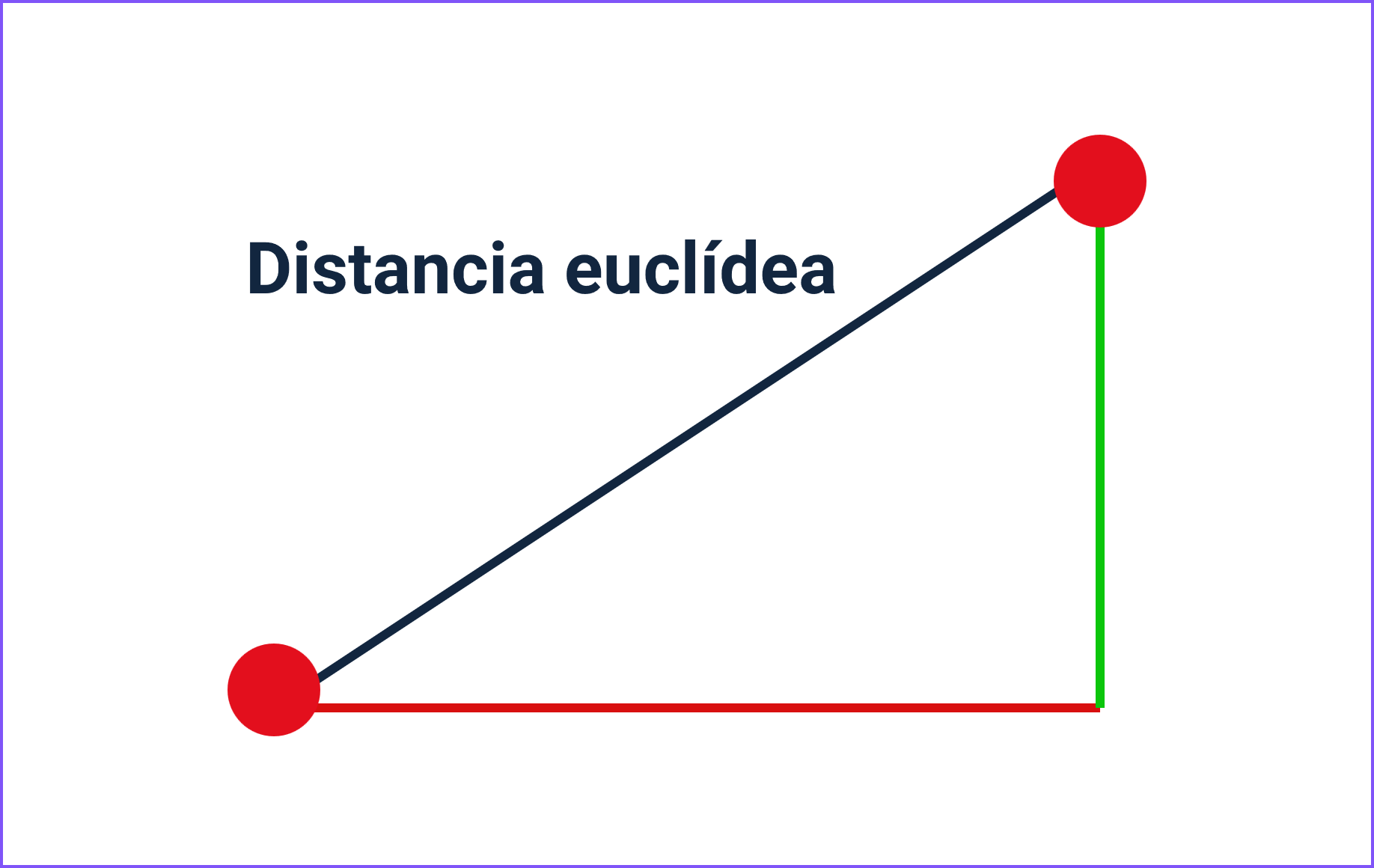
* Este algoritmo es muy sensible a la variable k.
* En la medida en que se cambie este valor, también cambiarán los resultados, para lo cual se recomienda realizar muchas pruebas con distintas instancias para fijarlo.
* También se debe tener en cuenta la métrica de similitud que se vaya a implementar, ya que esto determina fuertemente la relación de cercanía que se irá obteniendo en la construcción del modelo.
* Para establecer los puntos que sean similares y con mayor cercanía, se debe encontrar la distancia entre esos puntos.

Dentro de las técnicas más utilizadas para el cálculo de estas medidas, se pueden encontrar:

**Distancia euclidiana**

Es una de las técnicas más básicas, pero muy utilizada en los proyectos de “Machine Learning”. En principio, está pensada solo para usarla con variables numéricas. En caso de requerirse para trabajar con valores categóricos, se debe, antes, utilizar un mecanismo de conversión a números.

1. Distancia euclidiana

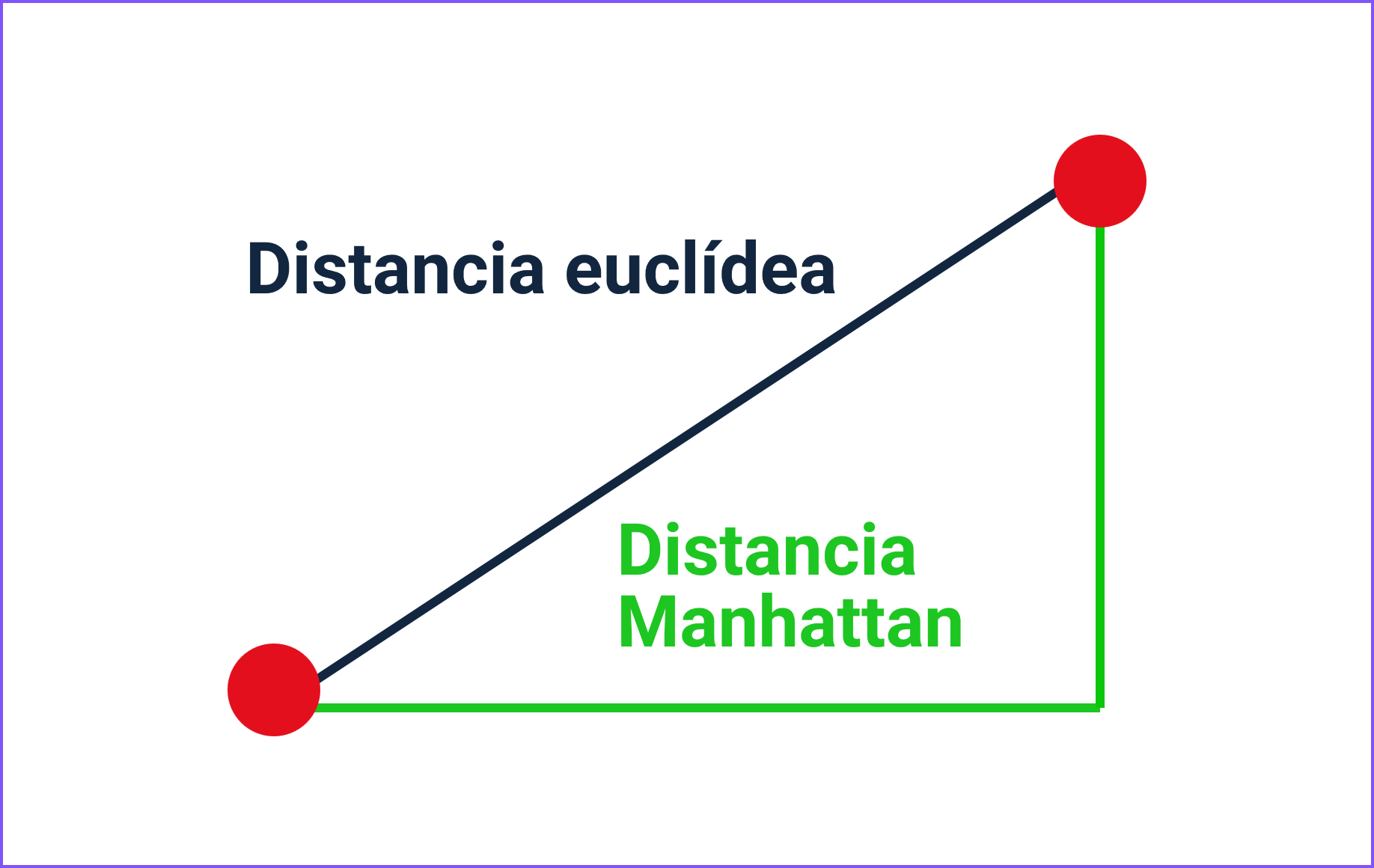


La distancia euclidiana representa la distancia más corta, en línea recta, entre un punto A y un punto B.

**Distancia Manhattan**

También denominada distancia de taxi. En el siguiente gráfico, se puede apreciar que la línea verde no es el camino más cercano de un punto a otro. Ello, porque el recorrido bordea otro punto, antes del punto de llegada.

1. Distancia Manhattan

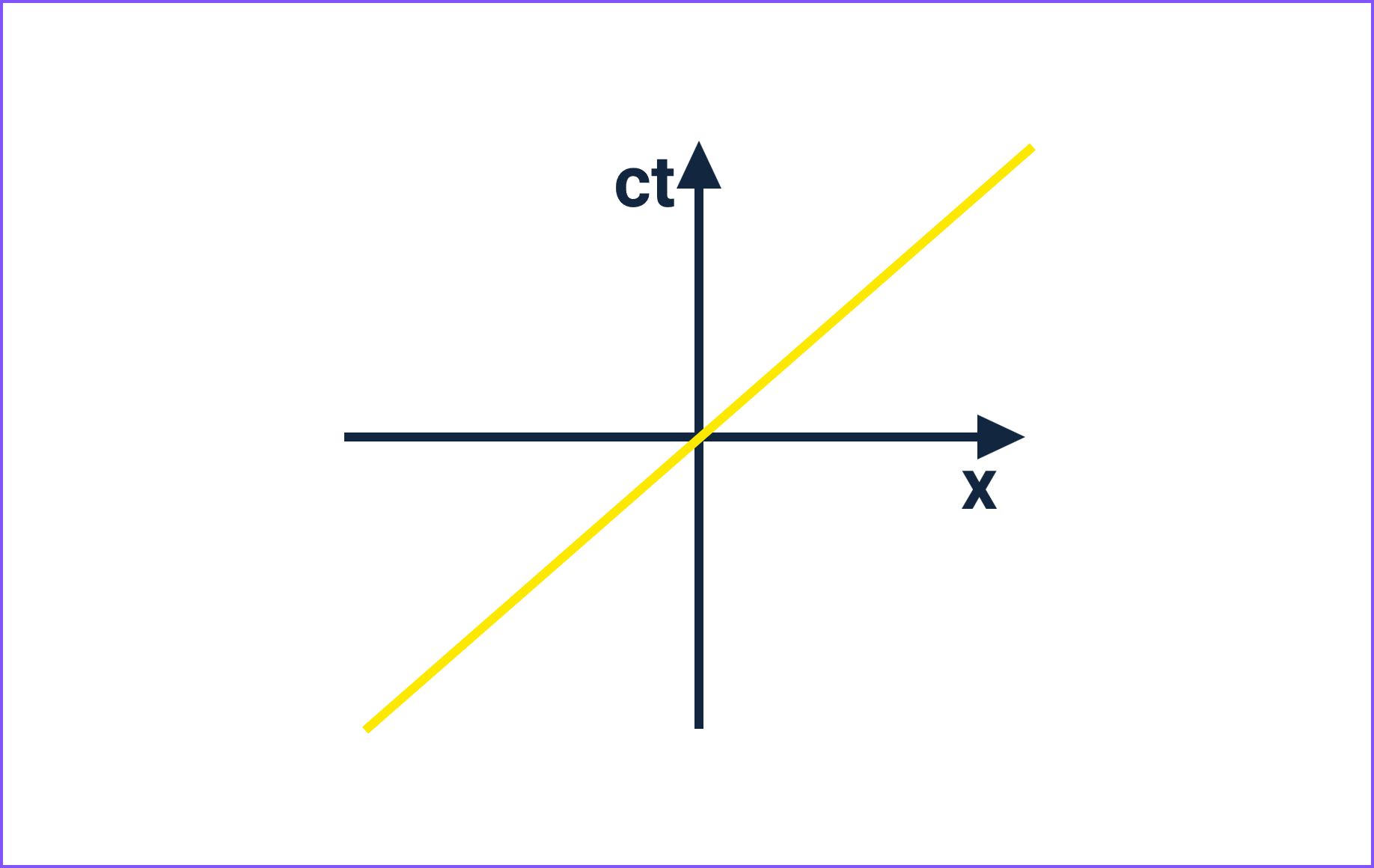


La distancia Manhattan representa la distancia más corta entre dos puntos, habiendo bordeado uno más puntos de manera previa a la llegada.

**Distancia Minkowski**

Es una métrica en un espacio vectorial normalizado. Es considerada como una generalización de la distancia euclidiana y la distancia de taxi.

1. Distancia Minkowski



El diagrama de Minkowski, supone la generalización de la distancia euclidiana y la distancia Manhattan.

En general, se puede decir que en el algoritmo Knn se deben realizar los siguientes pasos:

* Encontrar los vecinos más cercanos.
* Votar por las etiquetas.
* Calcular las distancias.

Hasta este punto, ya se sabe cómo funciona el algoritmo “K-nearest-neighbor”, y con qué parámetro fundamental es la determinación de k. No obstante, tenga presente que:

* **Número de vecinos.** El número de vecinos se debe elegir cuando se está realizando la construcción del modelo; se puede decir que k es una variable o parámetro de control en la predicción.
* **Colecciones de datos y sus exigencias propias.** Pero no es un valor que esté definido fácilmente, dependiendo de alguna problemática específica o una colección de datos puntual; cada colección de datos tendrá sus propias exigencias.
* **Sesgos.** Se han realizado investigaciones que han logrado determinar que las cantidades pequeñas de vecinos son con las que se obtendrían sesgos más bajos, pero también implican tener una alta varianza.
* **Límite de decisión y varianza.** Por otro lado, al seleccionar un gran número de vecinos, se logra obtener un límite de decisión más suave, lo que implica una varianza menor, pero el sesgo sería más alto.
* **Modelos y valores.** La recomendación es elegir números impares para las clases en las que las cantidades sean par y, definitivamente, realizar varios modelos en los que se apliquen diferentes valores para k e ir determinando con cuál se obtienen mejores resultados.
* **Su debilidad.** La debilidad de este algoritmo es su lentitud cuando realiza el proceso de clasificación, debido a que el objetivo principal de este algoritmo no es la búsqueda de un modelo óptimo.
* **Su estrategia.** Su estrategia se basa en que cada instancia de los datos que se seleccionaron para la prueba sea comparada con toda la colección de datos destinados para el entrenamiento y los resultados obtenidos dirán cuáles serán los ajustes que se deban tomar.
* **Validación.** Ello implica validaciones con distintas k, evaluar si las instancias seleccionadas para el entrenamiento fueron las correctas o validar las métricas de similitud utilizadas.

## Regresión logística

Las herramientas que son utilizadas para la clasificación son fundamentales en el mundo del “Machine Learning”. Cerca de un 70 % de los problemas corresponden a clasificación. En esta oportunidad, se abordará el algoritmo de regresión logística, que es muy común y útil cuando se trata de resolver problemas de clasificación binaria, es decir, cuando el resultado es un 0 o un 1.

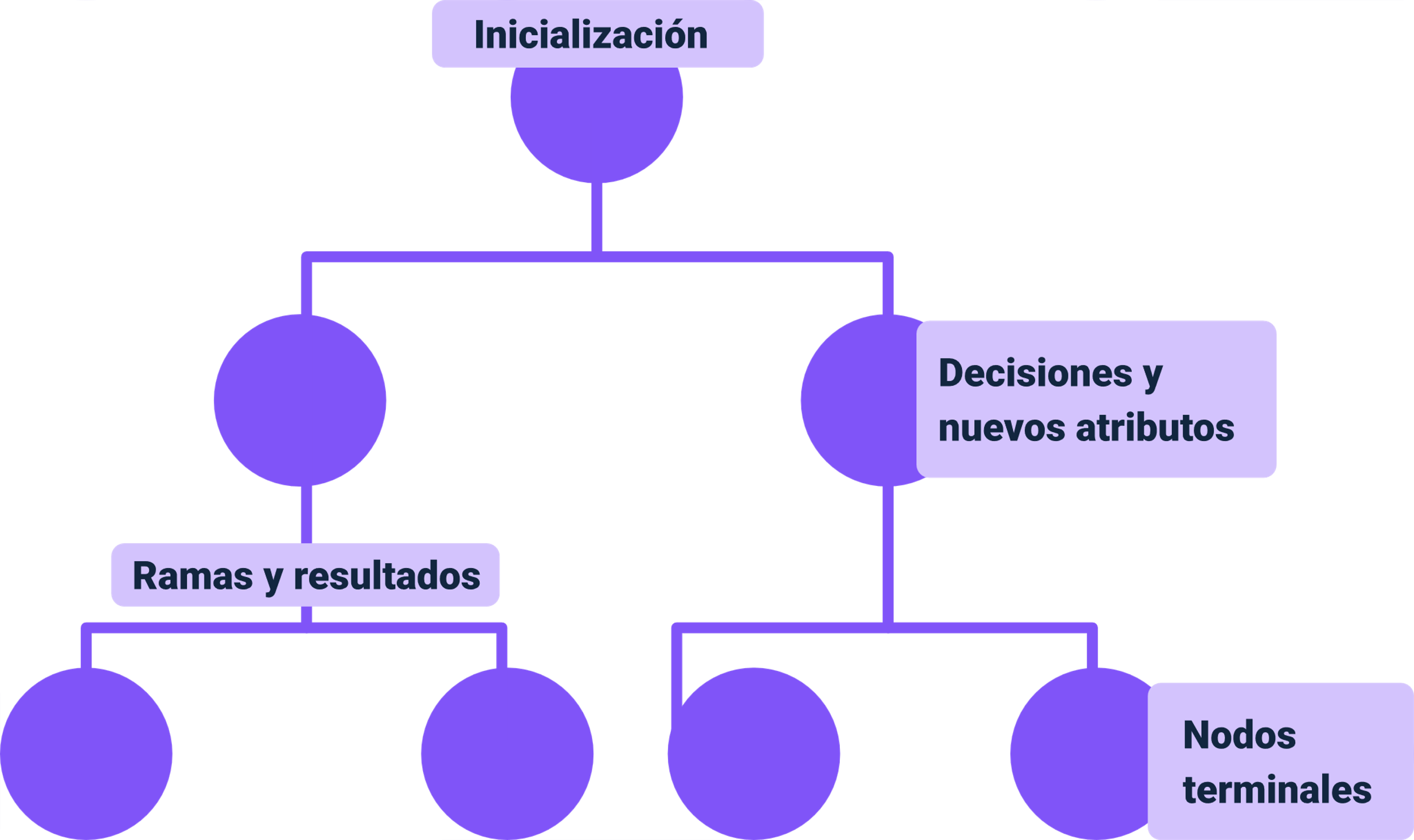
Explore el PDF **Algoritmo de regresión**, el cual se encuentra en la carpeta Anexos y profundice en las generalidades y aspectos clave de la regresión logística. ¡Adelante!

## Árboles de decisión

Esta es una técnica muy utilizada en el mundo del “Machine Learning”, la estrategia que sigue este algoritmo consiste en crear un modelo capaz de predecir el valor o clase de una observación, partiendo de una serie de reglas obtenidas de datos históricos. Son muchos los profesionales que utilizan esta técnica para generar estrategias comerciales de negocio, o en el área de la salud, en modelos que ayudan con los diagnósticos médicos.

A continuación, se muestra la estructura general básica de un árbol de decisión. Detállela con atención:

1. Estructura general básica de un árbol de decisión



* **Inicialización.** El árbol se inicializa con un nodo principal, llamado nodo raíz; en este nodo, se empieza tomando las decisiones teniendo en cuenta los atributos de los datos.
* **Decisiones y nuevos atributos.** Dichos nodos raíz se separan en nodos intermedios, en los cuales se continúa realizando la toma de decisiones a partir de nuevos atributos.
* **Ramas y resultados.** Los nodos intermedios se interconectan con las ramas, en las que se indican cuáles fueron los resultados tomados.
* **Nodos terminales.** Finalmente, se encuentran los nodos terminales, donde se muestra el resultado obtenido, luego de seguir una ruta que conllevó una serie de decisiones.

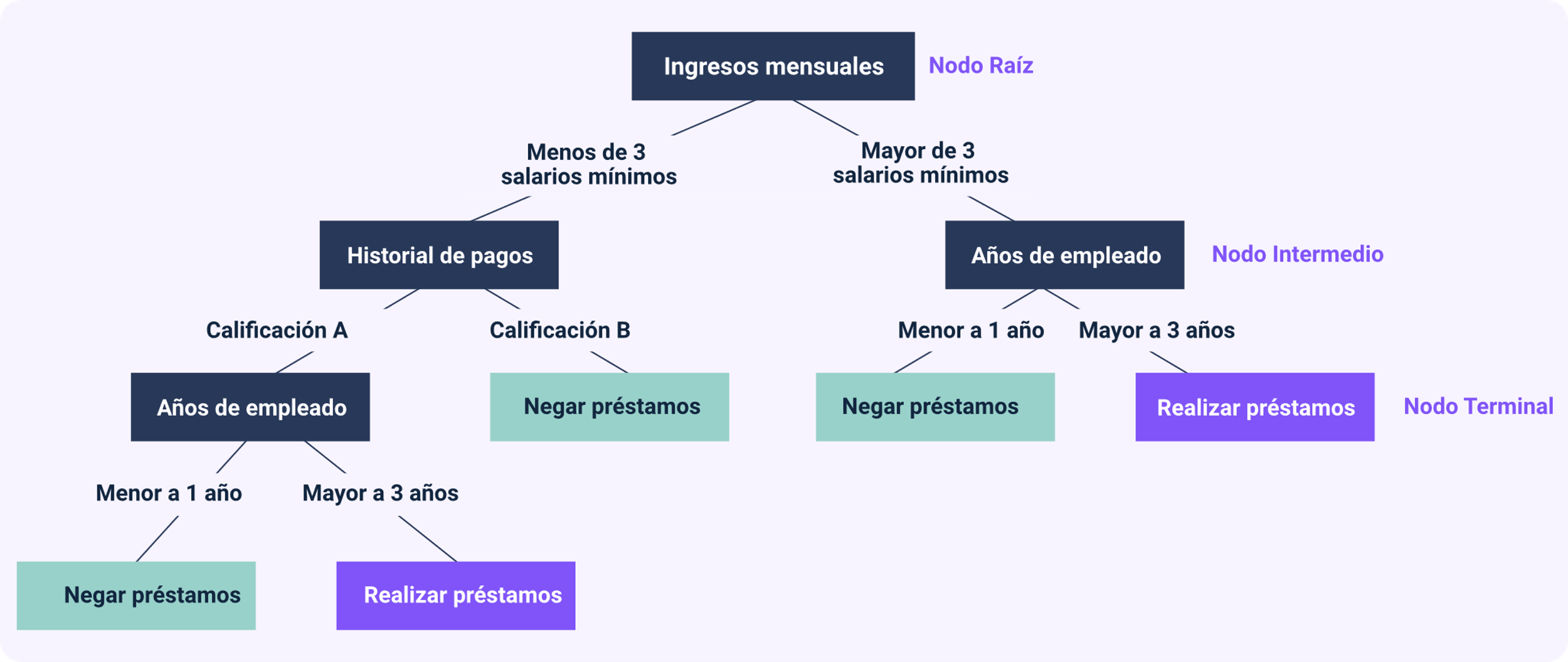
Sobre el árbol de decisión, tenga presente los siguientes aspectos:

* **Uso.** El árbol de decisión puede ser utilizado en la predicción de cualquier tipo de dato, ya sea categórico o numérico.
* **Otras denominaciones.** A los árboles de decisión también se les conoce como árboles de clasificación y árboles de regresión.
* **Método de construcción.** La construcción de los árboles de decisión se basa en un método heurístico denominado partición recursiva.
* **Finalidad.** El objetivo es que, a medida que el árbol vaya tomando decisiones, se armen grupos, cada vez más consistentes, en los que predomine un valor o clase.

Para entender mejor el funcionamiento de este algoritmo, se presenta un ejemplo en el que una entidad bancaria quiere realizar un modelo para tomar decisiones sobre a cuáles clientes se les aprueban préstamos o no.

Analice atentamente la gráfica que se muestra enseguida:

1. Ejemplo árboles de decisión



Los siguientes son los criterios de análisis del ejemplo de la figura:

* Como se muestra en el gráfico, se empieza con una decisión inicial, en el nodo raíz, correspondiente al salario.
* De acuerdo con el resultado en cada rama, se separa en dos nodos intermedios, uno valida la calificación en el historial de pagos y el otro se inclina por el número de años de empleado.
* Por un lado, se obtienen los resultados en los nodos terminales, mientras que el de la izquierda sigue tomando más decisiones hasta llegar más adelante a nodos terminales.
* Los grupos fueron personas a los que se les aprueba el préstamo y a los que no, encontrando así patrones, que son los que se usarían para la predicción frente a nuevos clientes.

¿Cuándo terminar las ramificaciones? A continuación, se dan algunas opciones muy prácticas y de acertado uso:

* **Opción 1.** En cada rama, se terminaría el proceso en la medida en que las variables dejen de generar grupos uniformes.
* **Opción 2.** Cuando al separar por una variable en la iteración n, no se genera un grupo más uniforme que el que se tiene en la iteración n-1.
* **Opción 3.** Si este paso no terminara el proceso, este se acabaría cuando se acaben las comparaciones con todas las variables.
* **Opción 4.** También, como alternativa, se podría definir que el árbol llegue a un límite que esté previamente parametrizado por el usuario.
* **Opción 5.** Normalmente, se establece un máximo para la profundidad del árbol, o también por un número específico de muestras en los nodos internos.

Es muy importante que se limite el crecimiento del árbol, ya que se haría tan complejo que puede llevar a lo que se conoce como:

* **Sobreajuste (“overfitting”).** Que básicamente se refiere a que se lograrían muy buenos resultados con los datos de entrenamiento, pero predicciones deficientes para nuevas observaciones.
* **“Pruning” (poda del árbol).** Consiste en realizar una revisión a los datos y descartar nodos que no aportan información. Realmente, este proceso es importante para la precisión de las predicciones, pues permite un equilibrio en lo simple de las decisiones que deba tomar el árbol.

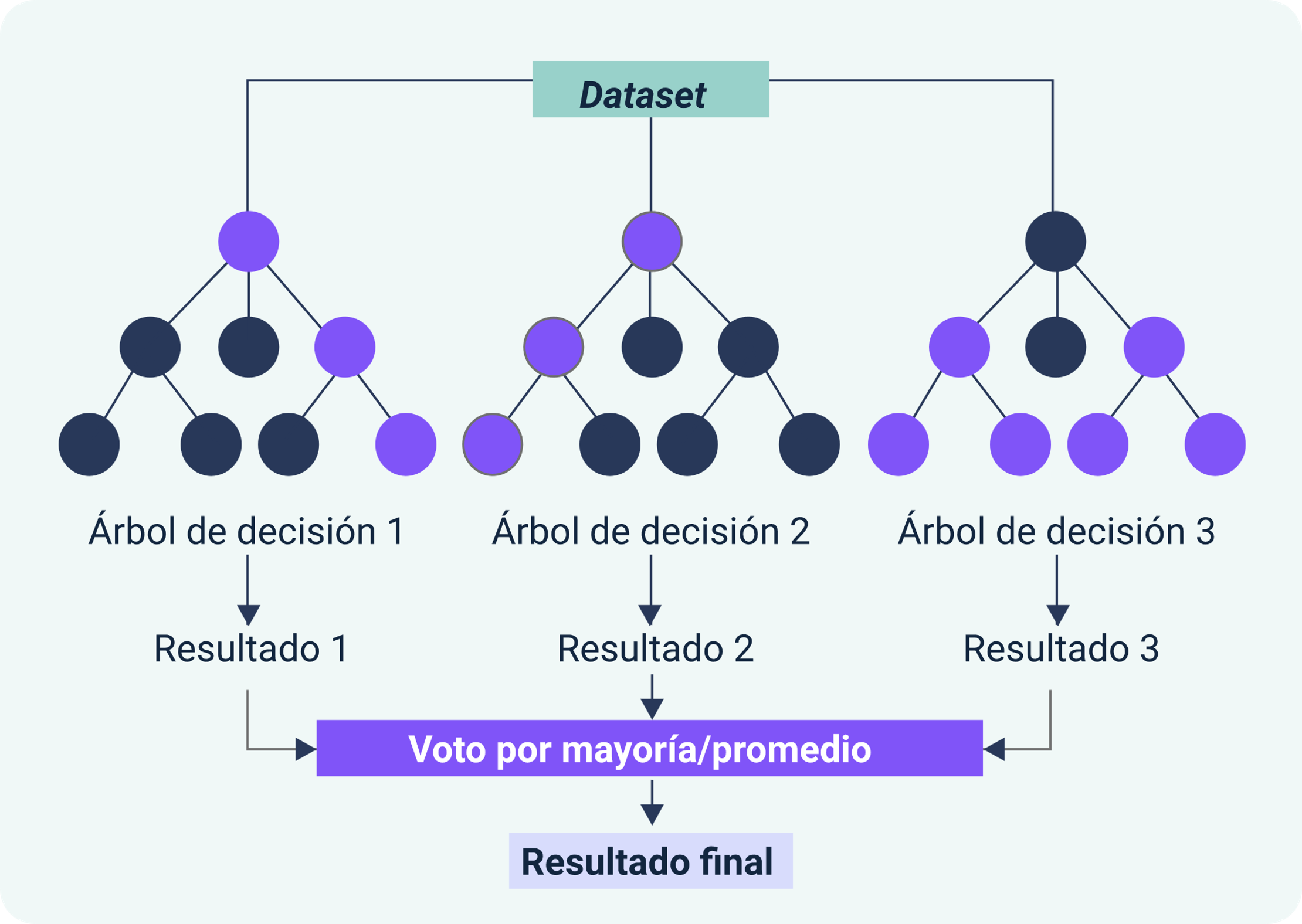
A continuación, se presentan las ventajas y las desventajas de la técnica de árbol de decisión:

* **Fácil de entender.** No son necesarios mayores conocimientos estadísticos para poder entender o interpretar las salidas que generan los árboles de decisión.
* **Exploración de datos.** Es una de las técnicas con las que se pueden identificar, rápidamente, las variables más importantes y la relación entre ellas; es posible la creación de nuevas características con las que se pueda predecir mucho mejor la variable objetivo.
* **Limpieza de datos.** El algoritmo de decisión, comparado con otras técnicas de modelado, es menos exigente en lo que tiene que ver con la limpieza de los datos; es importante recordar que este es uno de los pasos en ciencia de datos que mayor tiempo consume.
* **Tipo de datos.** Se puede trabajar con todo tipo de variables, ya sean numéricas o categóricas.
* **No paramétrico.** Este algoritmo se considera no paramétrico, lo que quiere decir que no se tienen suposiciones sobre la distribución del espacio y la estructura del clasificador.
* **Sobreajuste.** Es uno de los problemas que se presentan en este algoritmo, para solucionarlo, se pueden colocar restricciones en la parametrización del modelo y eliminar ramas en el análisis.
* **Dispersión.** Las técnicas basadas en árboles no fueron pensadas para funcionar con características que estén dispersas, se tendrían que reprocesar estas características en el caso de que las variables sean categóricas y de gran dimensión, usando estadísticas numéricas, o se podría utilizar un modelo que sea lineal, que para estos casos sería el más indicado.

## Bosques aleatorios

Los algoritmos de bosques aleatorios, hacen parte del aprendizaje supervisado. Esta técnica se basa en los algoritmos de árboles de decisión y es mucho más fácil y flexible de usar. Se puede implementar en la solución a problemas tanto de clasificación como de regresión.

1. Ejemplo “Dataset” bosques aleatorios



La imagen muestra un ejemplo de bosques aleatorios, mediante tres árboles de decisión, con sus respectivas ramificaciones, especificando cuáles se determinaron por mayoría o por promedio, llegando a un resultado final.

La estrategia de los bosques aleatorios se basa en la combinación de una gran cantidad de árboles de decisión, en la que se realiza un entrenamiento de cada uno de esos árboles de decisión para obtener una muestra diferente de las observaciones.

Los siguientes, son algunos aspectos importantes sobre los bosques aleatorios, que se deben tener en cuenta:

* Para obtener las predicciones de un algoritmo de bosque aleatorio, se toman todos los resultados de los árboles de decisión y se promedian las predicciones individuales de cada uno de ellos.
* Una de las principales desventajas de los árboles de decisión es el sobreajuste (“overfit”), con los datos, mediante el cual se realiza el entrenamiento.
* El bosque aleatorio resulta siendo de gran ayuda, pues permite la disminución de ese problema, ya que promedia cada resultado obtenido por la predicción de cada árbol.
* Esta técnica ofrece una mejor precisión en la predicción que cuando se trabaja con un solo árbol de decisión.
* También este algoritmo puede ayudar a buscar características que sean importantes en la colección de datos, a través del algoritmo de Boruta.

Esta técnica se ha utilizado en múltiples aplicaciones, de las cuales se pueden destacar las siguientes:

* Cuando vemos diversos productos al visitar una página de comercio electrónico.
* En el área de la salud, se puede implementar para la identificación de las enfermedades de pacientes, tomando como base el análisis del historial médico.
* En las entidades financieras, puede ser utilizada para determinar con facilidad si un cliente es fraudulento o legítimo.

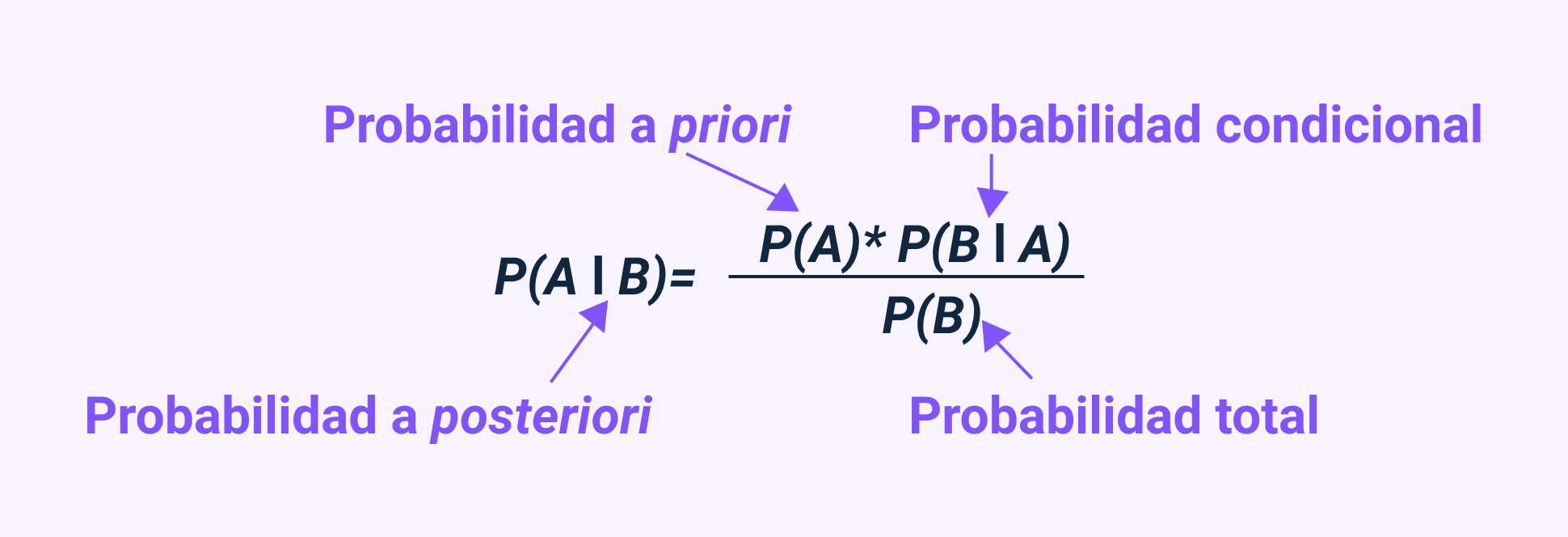
Para implementar este tipo de algoritmo, se deben contemplar algunos pasos, como los que se muestran a continuación:

* **Opción 1.** Lo primero que realiza el algoritmo es seleccionar las muestras de las bases de datos o colección de datos de manera aleatoria.
* **Opción 2.** Con cada muestra que selecciona el algoritmo, se crea un árbol de decisión, para luego obtener el resultado de la predicción de cada uno.
* **Opción 3.** Con cada uno de los resultados previstos, se procede a realizar una votación; en caso de ser un problema de clasificación, se tomará la moda; y para los casos de regresión, se usa la media.
* **Opción 4.** Para finalizar, el algoritmo de bosques aleatorios escogerá el resultado de las predicciones que más votos acumulen como la predicción final.

## “Naive Bayes”

Este algoritmo se basa en el teorema de Bayes, también conocido como el teorema de la probabilidad condicionada, es un clasificador muy útil y ampliamente usado, por su sencilla implementación y rapidez. Consiste en una técnica para clasificación y predicciones supervisadas, mediante la cual se diseña un modelo para predecir la probabilidad de los posibles resultados.

1. Algoritmo de Bayes



Probabilidad a “posteriori” es igual a Probabilidad a “priori” por Probabilidad condicional sobre Probabilidad total.

Para entender mejor el funcionamiento de esta técnica, preste atención al siguiente video:

1. Funcionamiento de “Naive Bayes”



[**Enlace de reproducción del video**](https://youtu.be/BfgKd2sDWto)

|  |
| --- |
| **Síntesis del video: Funcionamiento de “Naive Bayes”** |
| Tutorial que presenta el funcionamiento de “Naive Bayes”. |

## Otros algoritmos

Otros algoritmos de clasificación que se pueden encontrar son los siguientes:

**Máquina de vectores de soporte**

Los SVM, por sus siglas en inglés, generan una muy alta precisión comparándolos con otros algoritmos de clasificación, como lo pueden ser los árboles de decisión o el de regresión logística; su estrategia se basa en su Kernel, que permite manejar espacios de entradas que son no lineales; este algoritmo es utilizado en una gran cantidad de aplicaciones, como, por ejemplo, la detección de intrusos, clasificar el correo electrónico, páginas web, artículos de noticias, la detección de rostros, entre otros.

Este algoritmo separa las diferentes clases, gracias a que construye un hiperplano en un espacio multidimensional óptimo e iterativo que es utilizado para minimizar los errores; el objetivo principal de los SVM es buscar el hiperplano marginal que mejor separe la colección de los datos en clases.

Estos son algunos conceptos clave, relacionados con la máquina de vectores de soporte:

* **Vector de soporte.** Se define como los puntos que están más cerca del hiperplano; estos puntos ayudan a definir mucho mejor la línea que los separa mediante el cálculo de los márgenes; estos puntos toman mayor relevancia cuando se realiza la construcción del clasificador.
* **Hiperplano.** Es una frontera de decisión que se encarga de separar un conjunto de ejemplos que tienen una asociación a la clase diferente.
* **Margen.** Es un espacio entre las dos líneas en los puntos más cercanos de la clase. Se calcula como la distancia perpendicular desde la línea hasta los vectores de soporte o puntos más cercanos. Si el margen es mayor entre las clases, entonces se considera un buen margen; un margen menor es un mal margen.

A continuación, se presentan algunas ventajas y desventajas del uso de la máquina de vectores de soporte:

1. Ventajas:

* Buena precisión.
* Predicciones rápidas con relación al algoritmo de “Naive Bayes”.
* Utiliza menos recursos computacionales.
* Con un buen margen de separación y espacio dimensional elevado, funcionan mucho mejor.

1. Desventajas

* No se recomienda para grandes colecciones de datos.
* Es sensible con el tipo de núcleo que se utilice.
* Con clases que están superpuestas, funciona mal.

**Métodos combinatorios**

También se conocen como métodos de conjunto; sus estrategias se basan en una combinación de múltiples algoritmos de aprendizaje con el objetivo de buscar un mejor rendimiento predictivo, que los algoritmos utilizados solos.

El uso de esta técnica aumenta los recursos computacionales, ya que estos supuestos se analizan simultáneamente y una manera de mitigar estos costos computacionales es usar algoritmos rápidos, como los árboles de decisión.

Los siguientes, son los métodos combinatorios más comunes:

1. **Agregación “Bootstrap”.** La idea es simple, si se tienen muchas muestras de entrenamiento, se puede usar cada una de ellas para entrenar un modelo, que luego se usará para hacer predicciones. De esta forma, se tendrán tantas predicciones como modelos y, por tanto, muestras de entrenamiento. El proceso de promediar todas las predicciones tiene dos ventajas importantes: simplifica la solución y reduce en gran medida la varianza.

* El conjunto de datos de entrenamiento se vuelve a muestrear repetidamente.
* Entrene un modelo con cada conjunto de datos.
* Las predicciones se obtienen promediando las predicciones del modelo (decisión mayoritaria en el caso de clasificación).

1. **“Boosting”.** Es un método general de aprendizaje lento, en el que muchos modelos obtenidos por un método con poco poder predictivo se combinan y potencian para producir un mejor predictor. Los árboles de decisión pequeños (construidos con poca profundidad) son excelentes para esta tarea, son predictores muy pobres (aprendices débiles), son fáciles de componer y generar muy rápidamente.

* Los árboles se cultivan secuencialmente, se trata de mejorar las clasificaciones anteriores.
* A diferencia de los bosques aleatorios y de embolsado, puede haber problemas de sobreajuste (si la cantidad de árboles es grande y la tasa de aprendizaje es alta).
* Las observaciones pueden considerarse ponderadas iterativamente, asignando más peso a aquellas que son más difíciles de clasificar.
* El modelo final es un modelo aditivo (promedio ponderado de árboles).

1. **Subespacios aleatorios.** En este método, cada modelo se entrena con todos los ejemplos, pero solo considera un subconjunto de los atributos. El tamaño de estos subconjuntos es el parámetro del método, y de nuevo, el resultado es el promedio o votación de los resultados individuales de los modelos.

# Selección y preprocesamiento de datos

La selección de los datos y el preprocesamiento de estos es, sin lugar a dudas, una de las tareas que más demanda tiempo y dedicación; los datos son recopilados desde diferentes fuentes de información; actualmente, casi toda esta captura de información se realiza desde aplicativos en línea, que guardan los datos en grandes bases de datos; cada día se recopila mayor volumen de información y se necesitan sistemas actualizados y robustos que lo soporten, por lo que las empresas le están apostando más a la modernización tecnológica.

La información puede ser recopilada por sistemas de información privada que almacenan información propia de cada negocio; por ejemplo, en un sistema de “e-commerce”, la empresa puede almacenar información de sus clientes, compras, productos más vistos, número de “clics”, horas de mayor acceso, lugar desde donde se accede, entre cientos de posibilidades; las empresas quieren tener la mayor cantidad de información de sus clientes, pues esto les permite entenderlos y, de cierta manera, fidelizarlos.

Sobre selección y preprocesamiento de datos, tenga en cuenta:

* **Uso frecuente y común.** Estas fuentes de información cada día se utilizan más debido al incremento de métodos de ciencia de datos para estudiar los datos y generar valor al negocio.
* **Otras fuentes.** También pueden existir fuentes de información públicas, ya sea gratuitas o de pago; cuando se realizan proyectos de ciencia de datos.
* **Dónde seleccionar los datos.** Depende del objetivo, las necesidades y los requerimientos del negocio, por eso es muy importante conocerlo a profundidad y, a partir de allí, determinar si basta con las fuentes privadas o si es necesario recurrir a fuentes externas.
* **Saberes claves.** Durante la búsqueda y selección de la información, se deben tener habilidades en el manejo de archivos, conocimientos de bases de datos, manejo de sistemas de información, ya que estas herramientas son las que normalmente se utilizan para la administración y manipulación de los datos.

Estos son algunos tipos de archivos necesarios para el almacenamiento de la información:

* **Archivos XLS y XLSX.** Este tipo de archivos se genera desde uno de los programas más conocidos en el mundo de la informática para herramientas ofimáticas, como lo es Microsoft Excel. Las extensiones de los archivos se generan con extensión XLS para versiones 2003, y XLSX para versiones desde 2007 en adelante; este tipo de archivos se genera por múltiples sistemas de información, formularios en línea, Office en línea y muchas otras aplicaciones.
* **Archivos CSV.** Es un tipo de archivo que maneja una estructura simple que almacena la información en forma de filas que corresponden a cada uno de los registros, y los campos de esos registros están, a su vez, separados normalmente por comas, aunque los podemos encontrar separados por punto y coma o tabulaciones.

Es un tipo de formato que se puede abrir casi en cualquier programa, como, por ejemplo, un bloc de notas, aunque la compresión de los datos no es tan buena. Para un mejor entendimiento de los datos, se recomienda abrirlo en hojas de cálculo, como Excel, Google Sheet, u otros programas compatibles con esta estructura. Es un formato que no admite gráficos, tablas, colores o fuentes, está diseñado simplemente como un contenedor de información.

* **Archivos JSON.** Más que un tipo de archivo, es una estructura simple que se utiliza para el intercambio de información entre sitios web. Cuando se realiza la exportación de estos, se utilizan archivos de texto simple con extensión TXT; sin embargo, existen aplicaciones, como Google+, que utiliza la extensión JSON para guardar los perfiles de usuario.

Si se trabaja frecuentemente con datos, se deben dominar muy bien estos tipos de archivos; cada uno de ellos ofrece la posibilidad de verlos desde, directamente, su programa de creación. Por ejemplo, los tipos XLSX se podrían ver y analizar desde el programa Microsoft Excel; de igual manera, los de extensión CSV; y para analizar los tipos JSON, se utilizan editores de texto normales o programas especializados que entiendan esta estructura.

**¿Cuándo se puede empezar a trabajar con los datos?**

Si bien ya se realizó la búsqueda y selección de los datos, aún falta mucho para empezar con la creación de los modelos de “Machine Learning”, esto debido a que cada algoritmo de clasificación tiene sus propios requerimientos en cuanto a la colección de datos con los que va a entrenar, es decir, la adaptación de los datos a un modelo de entrenamiento específico; además de que los datos nunca serán ideales, pues se pueden encontrar datos nulos, datos inconsistentes, valores atípicos, con errores, entre otros problemas que normalmente se presentan.

Entonces:

* Es cuando entra en juego una de las tareas más importantes en los proyectos de ciencia de datos, constituyendo aproximadamente el 70 % del tiempo del proyecto, y se trata del preprocesamiento de los datos.
* Es una labor ardua, en la que el objetivo es realizar la depuración de los datos, de tal manera que estos queden lo más consistentes posibles.
* Al realizar el preprocesamiento, se realiza una limpieza, evitando entradas y salidas de características que no se van a utilizar; también, al mejorar los datos, se mejora el rendimiento del modelo, con mejores resultados.

Antes de realizar cualquier acción para preparar los datos, normalmente, se realiza una exploración de los datos, con el objetivo de ir entendiéndolos y los problemas que se puedan presentar en ellos; dentro de esas actividades de exploración, las más frecuentes pueden ser:

* Determinar el número de registros.
* Establecer las características o atributos de los datos.
* Determinar cuáles de los atributos corresponden a valores categóricos, numéricos o nominales.
* Contar y ubicar los valores faltantes.
* Establecer si existen formatos incorrectos.
* Análisis de datos atípicos.

Luego de realizar la exploración de los datos, se continúa con el preprocesamiento de datos, y estas son las actividades que normalmente se realizan:

* **Limpieza de datos.** Rellenar los valores que hacen falta, identificar y eliminar datos atípicos y los ruidosos.
* **Transformación de datos.** Normalización de los datos para reducción de ruido y las dimensiones.
* **Reducción de datos.** Atributos o registros de datos de ejemplo, para un control de datos más sencillo.
* **Discretización de datos.** Realizar la conversión de valores categóricos a numéricos, ya que es requerido por algunas técnicas de aprendizaje automático.

En cada una de las tareas que se realizan en el preprocesamiento, pueden surgir una serie de preguntas claves. Conózcalas a continuación:

1. **¿Qué hacer con los valores faltantes?** Para atacar este problema, lo mejor es investigar cuál es la razón del porqué hay ausencias de valores, lo que permitirá realizar un mejor control de la situación. De acuerdo con esto, las acciones que se pueden realizar son:

* Eliminación definitiva: se quitarían los datos que faltan en la colección de datos.
* Sustituciones con valor fijo: se realiza un reemplazo de los valores que hacen falta por un valor fijo definido de acuerdo con el análisis del negocio; por ejemplo, colocar un 0 para valores numéricos, o la palabra otro para variables categóricas.
* Sustituciones con la media: para atributos faltantes, puede realizar el cálculo de la media para esa columna y reemplazar los valores vacíos con ese resultado.
* Sustitución frecuente: consiste en reemplazar los datos que son categóricos por la etiqueta que más se repita.
* Sustitución de regresión: se implementa un método de regresión con el que se reemplazan los valores faltantes por los valores con regresión.

1. **¿Cómo se normalizan los datos?** Con la normalización, se escalan los datos numéricos en un intervalo determinado. Dentro de las técnicas para realizar la normalización de los valores más conocidos, se puede encontrar:

* Normalización mínima máxima: se convierten los datos linealmente a un intervalo específico, por ejemplo, al llevarlos a un intervalo que esté entre 0 y 1, el valor mínimo de esa columna tomará el 0 y el valor máximo, el 1.
* Normalización de puntuación z: con esta técnica, se escalan los datos teniendo en cuenta la desviación estándar y media: se divide la diferencia entre los valores y la media por la desviación estándar.
* Escalado decimal: escale los datos moviendo la coma decimal del valor del atributo.

1. **¿Qué hacer para discretizar?** Para la discretización de los valores, se puede hacer con la conversión de datos continuos en intervalos o en atributos nominales. Dentro de las estrategias para lograrlo, se puede usar lo siguiente:

* Discretizar con el mismo ancho: se deben tomar todos los valores del atributo y realizar la división en N grupos con el mismo tamaño, luego, se asignan los valores que se encuentran en una ubicación con el número de la ubicación.
* Discretización igual alto: consiste en dividir el rango de todos los valores posibles para un atributo en N grupos, cada uno con el mismo número de instancias, y asignar el valor en una posición al número de posición.

1. **¿Cómo reducir los datos?** Hay varias formas de reducir el tamaño de los datos para facilitar su procesamiento. Dependiendo del tamaño y dominio de los datos, se pueden aplicar los siguientes métodos:

* Muestreo de registros: muestreo de registros de datos y selección de solo un subconjunto representativo de datos.
* Muestreo de atributos: selecciona solo un subconjunto de los atributos más importantes de los datos.
* Agregación: agrupa datos y almacena números para cada grupo. Por ejemplo, los datos de ingresos diarios de una cadena de restaurantes durante los últimos 20 años se pueden agregar a los ingresos mensuales para reducir la cantidad de datos.

# Métricas de evaluación

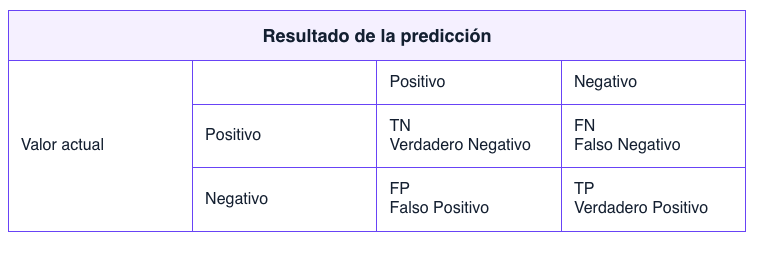
Las métricas de evaluación sirven para medir el rendimiento de un modelo entrenado; lo que se busca es mejorar el poder predictivo del modelo, antes de enviarlo a producción.

Al no realizar las métricas de evaluación, se corre el riesgo de obtener malas predicciones, lo cual se debe a que el modelo no aprende; en estos casos, solo memoriza. Por lo tanto, no puede generalizar a causa de datos no vistos anteriormente.

Una de las métricas de evaluación más usada es la Matriz de confusión. Se trata de una representación de los resultados de las predicciones; estos resultados son representados en forma de matriz, son obtenidos a través de las pruebas binarias que se utilizan para descubrir el rendimiento del modelo de clasificación y comparados con un conjunto de datos de prueba, de los cuales ya se conocen los valores reales.

La matriz de confusión se representa de la siguiente manera:

1. Representación matriz de confusión



Como se muestra en la anterior figura, las predicciones pueden ser uno de 4 resultados posibles; se basa en si coincide, o no, con el valor real:

* Verdadero Positivo: valor predicho es verdadero y el valor es verdadero en realidad.
* Verdadero Negativo: valor predicho es falso y el valor es falso en la realidad.
* Falso Positivo: valor predicho es verdadero y el valor es falso en la realidad.
* Falso Negativo: valor predicho es falso y el valor es verdadero en la realidad.

Para aceptar o rechazar una hipótesis, se debe tener en cuenta: si esta es nula y falsa, debe ser descartada; y si es nula y verdadera, debe ser aceptada.

**Tipos de errores**

Existen dos tipos de errores que pueden ocurrir, se les conoce como errores de TIPO I y errores de TIPO II:

* **Error de TIPO I.** Este error equivale a los falsos positivos (FP), es el rechazo de una hipótesis nula, pero esta es verdadera.
* **Error de TIPO II.** Este error equivale a los falsos negativos (FN), consiste en aceptar una hipótesis falsa nula.

Conozca cómo se pueden evaluar este tipo de errores:

* **Exactitud.** Esta métrica es usada cuando las clases son aproximadamente iguales en tamaño; se encarga de medir el porcentaje de casos en que el modelo ha acertado:

Exactitud = (TP + TN) / (TP + TN + FP + TP)

* **Precisión.** Esta métrica es usada para medir la calidad del modelo en tareas de clasificación. Se usa la siguiente fórmula:

Precisión = TP / (TP + FP)

* **Exhaustividad.** Esta métrica es usada para mostrar la cantidad de verdaderos positivos que el modelo es capaz de identificar. Se usa la siguiente fórmula para calcular la exhaustividad:

Exhaustividad = TP / (TP + FN)

* **Puntuación F1.** Esta métrica combina la precisión y exhaustividad en un solo valor, comprendido entre 0 y 1, donde la mejor puntuación es 1 y la peor es 0. Esta métrica se calcula utilizando la fórmula de la media armónica entre la precisión y la exhaustividad:

F1 = 2 \*((precisión \* exhaustividad) / (precisión + exhaustividad))

Otra forma de desarrollar el proceso de evaluación es por medio del Análisis ROC (“Receiver Operating Characteristic”), que se define como característica operativa del receptor; es una forma útil de evaluar la precisión de las predicciones de modelo al trazar la sensibilidad frente a la especificidad de una prueba de clasificación (ya que el umbral varía en todo un rango de resultados de pruebas de diagnóstico).

El área completa bajo una curva ROC determinada, o AUC, formula una estadística importante que representa la probabilidad de que la predicción esté en el orden correcto cuando se observa una variable de prueba (para un sujeto seleccionado aleatoriamente del grupo de casos, y el otro seleccionado aleatoriamente del grupo de control).

El Análisis ROC admite la inferencia en relación con una sola AUC, curvas de precisión/exhaustividad (PR), y proporciona opciones para comparar dos curvas ROC que se generan de grupos independientes o sujetos emparejados.

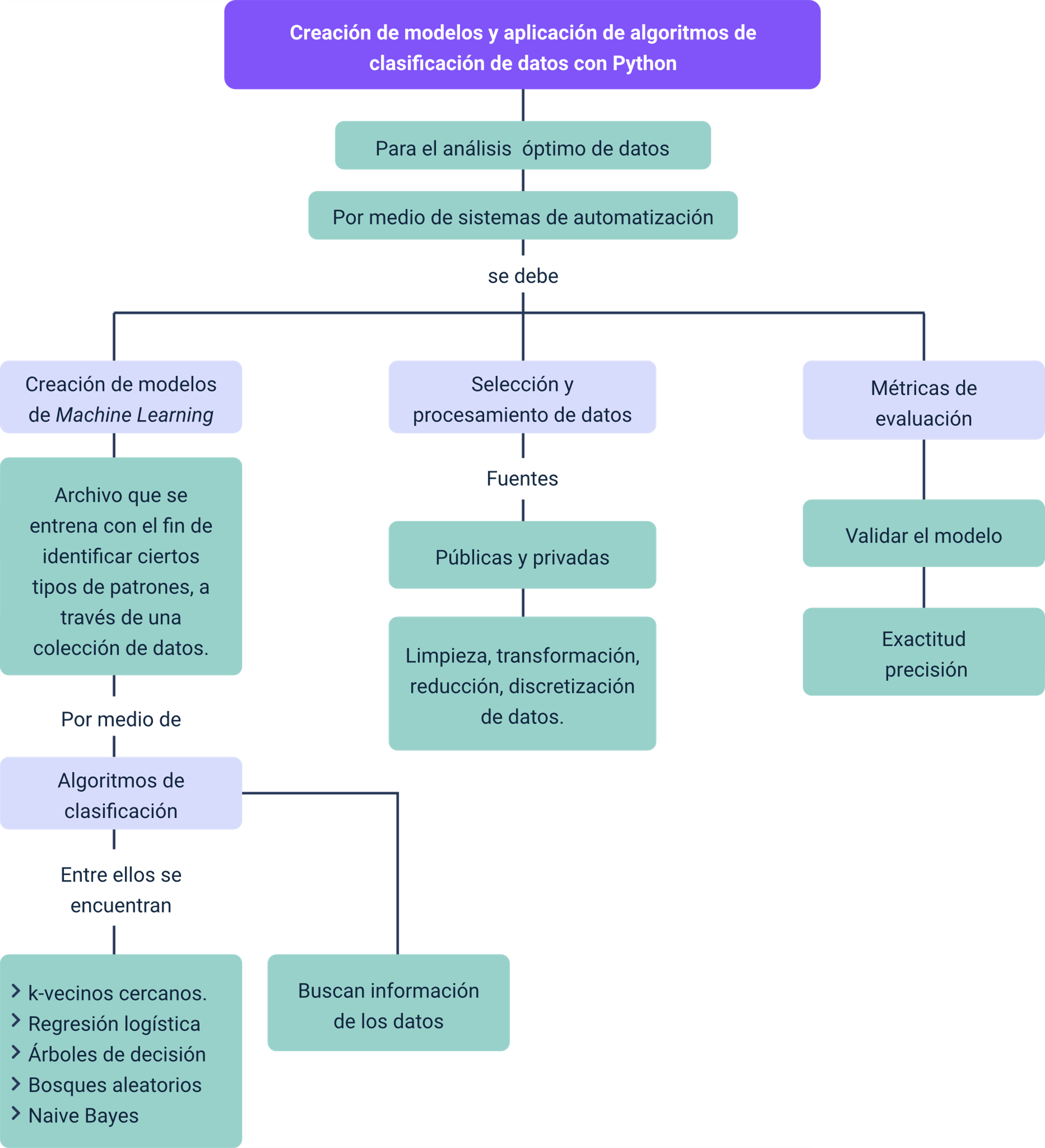
Las curvas PR representan la precisión frente a la exhaustividad, tienden a ser más informativas cuando las muestras de datos observados son muy asimétricas, y proporcionan una alternativa a las curvas ROC para datos con una alta asimetría en la distribución de clase.

Preste atención al siguiente ejemplo de aplicación del método ROC:

* **Conjunto de archivos.** Un banco tiene interés en clasificar a sus clientes dependiendo de si se retrasarán o no en el pago de sus préstamos; por tanto, se desarrollan modelos especiales para tomar estas decisiones. Se puede utilizar el Análisis ROC para evaluar y valorar la precisión de las predicciones del modelo.
* **Estadísticos.** AUC, grupo negativo, valores perdidos, clasificación positiva, valor de corte, grado de convicción, intervalo de confianza asintótica bilateral, distribución, error estándar, diseño de grupo independiente, diseño de muestra emparejada, supuesto no paramétrico, supuesto de distribución exponencial binegativo, punto medio, punto de corte, curva PR, interpolación por pasos, significación asintótica (bilateral), Sensibilidad y (1 - Especificidad), Precisión y Exhaustividad.
* **Métodos.** Se comparan las áreas bajo dos curvas ROC que se generan a partir de grupos independientes o sujetos emparejados. Comparar dos curvas ROC puede proporcionar más información en cuanto a la precisión que resulta de dos acercamientos de diagnósticos comparativos.
* **Datos.** Consideraciones de datos sobre el Análisis ROC. Las curvas PR representan la precisión frente a la exhaustividad y tienden a ser más informativas cuando las muestras de datos observados son muy asimétricas. Una interpolación lineal simple puede producir erróneamente una estimación excesivamente optimista de una curva PR.
* **Supuestos.** La predicción estará en el orden correcto cuando se observa una variable de prueba para un sujeto seleccionado aleatoriamente del grupo de casos, y el otro seleccionado aleatoriamente del grupo de control. Cada grupo definido contendrá al menos una observación válida. Solo se utiliza una variable de agrupación única para un procedimiento único.

Síntesis

Aquí finaliza el estudio de los temas de este componente formativo “El algoritmo ideal”. En este punto, analice el esquema que se muestra a continuación y haga su propia síntesis de los conceptos y aspectos estudiados. ¡Adelante!



Este esquema general de contenidos y conceptos, explica cómo el componente formativo desarrolló aspectos generales y claves sobre los principales algoritmos de clasificación que se utilizan en modelos “Machine Learning”, los cuales contribuyen a la solución en diversas problemáticas, a través del desarrollo de sistemas de automatización.

Material complementario

|  |  |  |  |
| --- | --- | --- | --- |
| Tema | Referencia | Tipo de material | Enlace del recurso |
| 1. Creación de modelos de “Machine Learning” | SDC LEARNING. (2022). Webinar Gratuito Mi Pri-mer Modelo de Machine Learning en Python | Video | <https://www.youtube.com/watch?v=9HKfqinJJAo> |
| 1.1. Algoritmos de clasificación | Parra, F. (2019). Estadística y Machine Learning con R. Book-down. | Página web | <https://bookdown.org/content/2274/metodos-de-clasificacion.html> |
| 3. Métricas de evaluación | González, L. (2020). Evaluando el error en los modelos de clasificación. Aprende IA. | Página web | <https://aprendeia.com/evaluando-el-error-en-los-modelos-de-clasificacion-machine-learning/> |

Glosario

**Algoritmos de clasificación:** se utilizan cuando el resultado deseado es una etiqueta discreta.

**Entrenamiento:** es una forma de la Inteligencia Artificial que permite a un sistema aprender de los datos, en lugar de aprender mediante la programación explícita.

**Exactitud:** esta métrica es usada cuando las clases son aproximadamente iguales en tamaño, se encarga de medir el porcentaje de casos en que el modelo ha acertado.

**Exhaustividad:** esta métrica es usada para mostrar la cantidad de verdaderos positivos que el modelo es capaz de identificar.

**Hipótesis:** suposición hecha a partir de unos datos, que sirve de base para iniciar una investigación o una argumentación.

**Media armónica:** es una medida estadística recíproca a la media aritmética, que es la suma de un conjunto de valores entre el número de observaciones.

**Precisión:** esta métrica es usada para evaluar la calidad del modelo en tareas de clasificación.

**Predicción:** los sistemas de predicción son técnicas que construyen y estudian nuevas previsiones a través de una rama de la Inteligencia Artificial denominada aprendizaje automático (“Machine Learning”).

Referencias bibliográficas

González, L. (2020). ¿Cuál es la diferencia entre Regresión Lineal y Regresión Logística? Aprende IA. <https://aprendeia.com/diferencia-entre-regresion-lineal-y-regresion-logistica-machine-learning/>

Miller, V. (2018). Explorando Algoritmos de Aprendizaje Automático Supervisado. Developers. <https://www.toptal.com/machine-learning/explorando-algoritmos-de-aprendizaje-automatico-supervisado>

Roman, V. (2019a). Aprendizaje Supervisado: Introducción a la Clasificación y Principales Algoritmos. Ciencia & Datos. <https://medium.com/datos-y-ciencia/aprendizaje-supervisado-introducción-a-la-clasificación-y-principales-algoritmos-dadee99c9407>

Roman, V. (2019b). Machine Learning: Cómo Desarrollar un Modelo desde Cero. Ciencia & Datos. <https://medium.com/datos-y-ciencia/machine-learning-cómo-desarrollar-un-modelo-desde-cero-cc17654f0d48>

Sotaquirá, M. (2021). ¿Se requiere SQL para trabajar en Machine Learning? Codificando Bits. <https://www.codificandobits.com/blog/sql-machine-learning/>

Créditos

| Nombre | Cargo | Centro de Formación y Regional |
| --- | --- | --- |
| Claudia Patricia Aristizábal | Responsable del Ecosistema | Dirección General |
| Rafael Neftalí Lizcano Reyes | Responsable de Línea de Producción | Regional Santander - Centro Industrial del Diseño y la Manufactura |
| Rónald Alexánder Vacca Ascanio | Experto Temático | Regional Distrito Capital – Centro de Diseño y Metrología |
| Fabián Leonardo Correa Díaz | Diseñador Instruccional | Regional Santander – Centro Industrial del Diseño y la Manufactura |
| Blanca Flor Tinoco Torres | Diseñador web | Regional Santander – Centro Industrial del Diseño y la Manufactura |
| Francisco José Lizcano Reyes | Desarrollador Fullstack | Regional Santander – Centro Industrial del Diseño y la Manufactura |
| Wilson Andrés Arenales Cáceres | Storyboard e Ilustración | Regional Santander – Centro Industrial del Diseño y la Manufactura |
| Carlos Eduardo Garavito Parada | Animador y Productor Multimedia | Regional Santander – Centro Industrial del Diseño y la Manufactura |
| Daniela Muñoz Bedoya | Locución | Regional Santander – Centro Industrial del Diseño y la Manufactura |
| Emilsen Alfonso Bautista | Actividad didáctica | Regional Santander – Centro Industrial del Diseño y la Manufactura |
| Zuleidy María Ruiz Torres | Validador de Recursos Educativos Digitales | Regional Santander - Centro Industrial del Diseño y la Manufactura |
| Luis Gabriel Urueta Álvarez | Validador de Recursos Educativos Digitales | Regional Santander - Centro Industrial del Diseño y la Manufactura |
| Daniel Ricardo Mutis Gómez | Evaluador para contenidos inclusivos y accesibles | Regional Santander - Centro Industrial del Diseño y la Manufactura |