Introducción a la analítica predictiva

Breve descripción:

Este componente formativo está enfocado en el reconocimiento de los algoritmos de aprendizaje supervisado, no supervisado y semisupervisado, que serán aplicados al análisis automático de datos. Ello permite, a cualquier organización, tomar decisiones asertivas e implementar soluciones a diferentes problemas identificados.

Noviembre 2023

Tabla de contenido

Introducción

Tenga una especial bienvenida a este componente formativo denominado **Introducción a la analítica predictiva**; para comenzar el recorrido por el mismo, explore la información que se ofrece en el siguiente video:

Video 1. Introducción a la analítica predictiva

Enlace de reproducción del video

Síntesis del video: Introducción a la analítica predictiva
¿Sabía usted que, en la actualidad, los datos son los insumos más importantes sobre los cuales una empresa toma sus decisiones? En efecto, los datos han sido un recurso altamente valioso desde el comienzo de la historia de las organizaciones. Hoy día, lo es mucho más, en virtud del avance de las tecnologías computacionales. Con ella, se han consolidado infinidad de proveedores de

aplicaciones en la nube y grandes bases de datos o "big data"; las mismas empresas del mundo generan millones y millones de datos todos los días.

Estos datos pueden ser usados para responder grandes preguntas y tomar decisiones de todo tipo; así mismo, han nacido conceptos como la ciencia de datos que involucra elementos estadísticos para realizar la exploración de los datos, aprendizaje automático para identificar patrones o realizar predicciones, matemáticas, ciencias de la computación, limpieza, formateo de datos y visualización de los mismos.

Pero, ¿a qué se refiere el aprendizaje automático?

Este concepto no es nuevo, en 1952 Arthur Lee Samuel, pionero en inteligencia artificial y videojuegos, usó este término y creó el primer juego "Checkers" (Damas), basado en el aprendizaje automático, Frank Rosenblatt en 1957 desarrolló el "perceptron mark one", máquina capaz de aprender mediante un sistema de red nerviosa que simula los procesos del cerebro humano.

En 1967 se escribió el algoritmo "Nearest Neighbor", con el cual nacieron los algoritmos de reconocimiento de patrones. En 1981 Gerald De Jong usa el término aprendizaje basado en experiencia, en el que el programa analiza información de entrenamiento y crea reglas para descartar las observaciones o datos menos importantes.

En 1990 nace el "data driven" que es básicamente aprovechar al máximo los datos existentes para tomar decisiones, por ejemplo, sobre actuales clientes y sobre clientes potenciales. Empresas como Netflix usan este concepto.

Hoy por hoy, empresas como IBM, Watson, Facebook, Amazon, Microsoft, Google, entre otras, presentan sus plataformas de aprendizaje automático; Amazon, por ejemplo, patentó "anticipatory shipping", el cual puede predecir la demanda de productos por parte de los usuarios en un área geográfica determinada y tenerlos listos para ser entregados a los usuarios.

Como es imaginable, siempre ha existido el interés por apoyarse en máquinas para aprovechar el poder de los datos al máximo, de aquí la importancia de mantener datos organizados, confiables y seguros que, respaldados con algoritmos adecuados, creen modelos predictivos potentes que colaboren no solo en empresas comerciales, sino en muchas áreas como la ciencia, la salud, la educación, la investigación, la tecnología y el transporte, para tomar decisiones adecuadas con la mayor probabilidad de ser las correctas.

1. Conjunto de datos

En la actualidad, es muy común escuchar el término ciencia de datos, el cual ha surgido a partir de la necesidad de identificar estrategias para manejar grandes volúmenes de información dentro de determinadas organizaciones.

La ciencia de datos permite hacer uso de herramientas y técnicas para encontrar patrones en la información, realizar predicciones y clasificar información usando algoritmos de aprendizaje.

A partir de ello:

- Han surgido conceptos como "big data".
- "Big data" hace referencia a extensos conjuntos de datos.
- Tales conjuntos de datos pueden ser estructurados, semiestructurados y no estructurados.
- Esto porque son demasiado grandes y difíciles de procesar con las bases de datos y el "software" tradicional.

Para reconocer cada uno de dichos conjuntos de datos revise con atención la siguiente información:

a. **Datos estructurados**. Se refieren a la información organizada. Estos datos se pueden procesar, almacenar y recuperar con un formato establecido desde una base de datos. Entre ellos se encuentran detalles de empleados como salario, cargo, edad, entre otros; detalles de clientes como edad, sexo, capacidad de compra, entre otros; datos de pacientes como edad, sexo, diagnóstico, entre otros; tipos de flores como longitud y ancho del pétalo, longitud y ancho del sépalo.

b. **Datos no estructurados**. Estos datos carecen de cualquier forma o estructura específica, son complejos de procesar y analizar. Por ejemplo, los videos y el audio o los correos electrónicos, documentos en PDF o Word, etc.
c. **Datos semiestructurados**. Contienen los dos tipos de datos anteriores tanto estructurados como no estructurados, estos datos se pueden almacenar en alguna base de datos; pero pueden tener etiquetas que hacen referencia a otros datos de la información como HTML, XML o JSON.

Otro aspecto a tener en cuenta es que el "big data" se ha caracterizado por poseer cuatro grandes cualidades: volumen, velocidad, variedad y veracidad; por tanto, dicho concepto se puede puntualizar como:

Grandes conjuntos de datos (volumen), de diversos tipos (variedad) que son creados, almacenados y procesados a gran rapidez en tiempo real (velocidad), garantizando la fiabilidad de la información recibida (veracidad).

Por otro lado, para entender mejor a qué se hace referencia cuando se habla de un conjunto de datos es importante tener claridad frente a conceptos como población, muestra, unidad de análisis, variables y datos.

A continuación, se enuncian y explican algunos de ellos:

a. **Población**. Conjunto de todos los casos que concuerdan con determinadas características y especificaciones, la población puede estar conformada por personas, animales, muestras de laboratorios, registros, etc.
b. **Muestra**. Puede considerarse como ese subconjunto de la población que hace parte de la investigación.

c. **Unidad de análisis**. Es la entidad principal que un investigador analiza en su estudio. Esta entidad puede ser persona, grupo de personas, negocios, objetos inanimados, transacciones, etc.
d. **Variable**. Es cualquier característica de una unidad de análisis de una población, por ejemplo, los clientes pueden ser clasificados en hombres o mujeres, esta variable se llama sexo.
e. **Dato**. Es el valor que toma una variable de una unidad de análisis.

Además, es importante puntualizar en los tipos de variables que se pueden medir para realizar cualquier estudio de información; entre ellas se encuentran la variables categóricas o cualitativas y las variables numéricas o cuantitativas:

a. **Variables categóricas o cualitativas**. Estas se definen por categorías, clases o dimensiones, por ejemplo, estado civil (puede tomar valores como casado, soltero, separado, viudo, etc.), tipo de ocupación (puede tomar valores como empleado, independiente, empleado público, empleado privado, etc.), nivel educativo (primaria, secundaria, profesional, especialización, maestría, etc.). A su vez, las variables categóricas pueden ser dicotómicas o binarias de acuerdo con la presencia o ausencia de una determinada característica como, por ejemplo, intención de compra de un producto, el cual puede tomar valor sí o no.
b. **Variables numéricas o cuantitativas**. Estas variables se expresan por un valor numérico, dentro de esta clasificación existen las variables discretas y las variables continuas:
    - Las variables discretas pueden tomar valores como un simple conteo o por asignación de ciertos números a categorías cualitativas.

- Las variables continuas se generan midiendo una variable sobre unidades de análisis tales como duración de una llamada en minutos, nivel de ingresos, altura, edad, peso, temperatura, salario, etc.

Pero, ¿qué grados de medición tienen las variables cualitativas o cuantitativas?

a. **Escala nominal**. Clasifica la unidad de análisis de acuerdo con una característica cualitativa, por ejemplo, el estado civil de un cliente puede tener las categorías: soltero, casado, viudo, separado y a cada una de estas se puede agregar un valor numérico: soltero puede tomar el valor 1, casado el 2, viudo el 3 y separado el 4.
b. **Escala ordinal**. Establece una relación de orden entre las distintas variables cualitativas como, por ejemplo, se puede considerar el grado de estudios que tiene una persona 1: puede ser primaria, 2: bachiller, 3: estudios universitarios, 4: si tiene estudios de especialización.
c. **Escala de intervalo**. A veces es importante medir la diferencia entre dos valores de una variable como, por ejemplo, se pueden tomar los grupos de edades de un paciente, rango de temperaturas en la que un paciente se encuentra, etc.
d. **Escala de razón**. También denominada de proporción, se usa para medir variables físicas y naturales como velocidad, longitud, aceleración, peso. Se puede realizar conversión entre estas unidades como, por ejemplo, se encuentra el peso en **Kg**, estatura en **cm** y bilirrubina en suero **mg** por litro.

2. Procesamiento de datos

Cuando se trata de algoritmos de predicción o clasificación en aprendizaje supervisado, existen datos o características independientes y un valor dependiente. Si existen demasiadas variables predictoras, es posible que muchas de ellas sean irrelevantes y es conveniente eliminarlas.

Si una de las características no es informativa es conveniente eliminarla; por ejemplo, de acuerdo con determinada investigación realizada, puede ser más significativa la edad de una persona que el número de su casa.

Existen varias formas para preprocesar la información; entre las más destacadas están:

a. **Binarización de características**. El principal objetivo de esta transformación es mejorar una observación o entidad, eliminando, cambiando o añadiendo información. Una transformación sencilla puede ser transformar un tipo de "data set" en otro, por ejemplo, transformar un atributo de tipo entero a flotante. Una transformación muy común es la binarización que no es más que transformar un atributo de tipo categórico a un atributo booleano.
b. **Discretización**. Consiste en convertir un valor numérico continuo en intervalos, esto es debido a que algunos algoritmos de clasificación solo aceptan atributos categóricos, esto reduce el tamaño de los datos para su mejor entendimiento.

   Existen varias formas de discretización, pero una de las más conocidas es "binning" que consiste en agrupar valores en contenedores "bins"; por ejemplo, se puede agrupar las edades que son valores numéricos continuos

en rangos o en grupos de edades como adultos, adulto mayor, bebé, niño, joven, etc.

c. **Normalización**. Se trata de uniformar los valores de las características no uniformes en rangos diferentes y, de esta forma, hacer que los rangos sean consistentes entre las variables y permitan comparación imparcial entre ellas, de esta manera se facilitará algunos análisis estadísticos posteriores.

d. **Formas de normalizar**.

   - **Escalado simple**: se divide cada valor de una característica por el valor máximo de esa característica. Los valores, por tanto, están entre 0 y 1.
   - **Escalado min-max**: a cada valor de una característica se le resta el valor mínimo de todos los datos de la característica y se divide la diferencia entre el valor mínimo y el máximo, los valores estarán entre 0 y 1.
   - **"Z-Score" o puntuación estándar**: a cada valor de una característica se le resta la media de la característica y este resultado se divide por la desviación estándar, los resultados tomarán valores negativos y positivos.

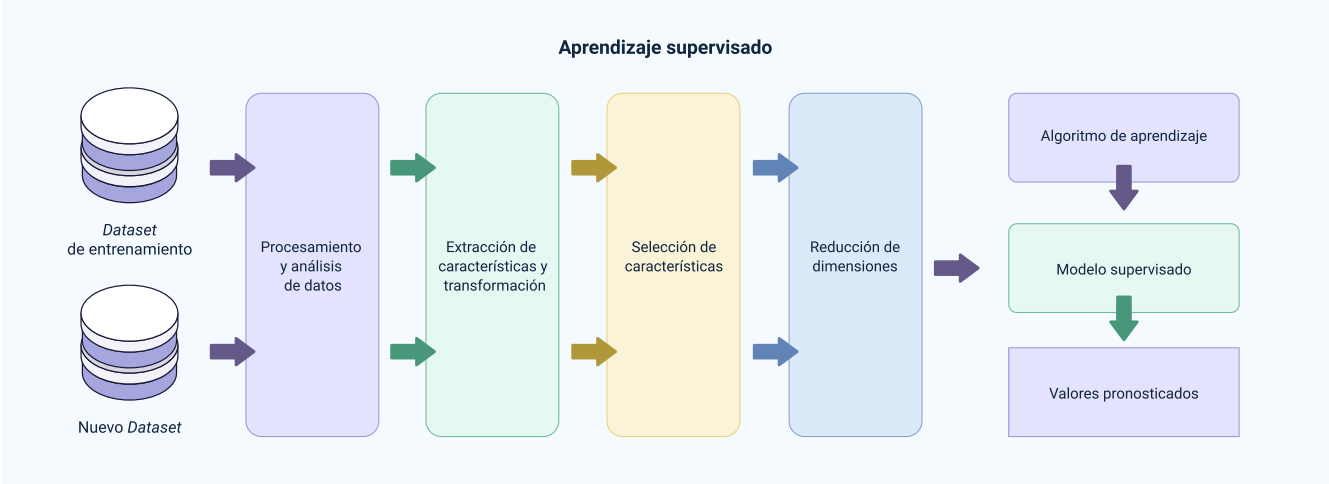
e. **Imputación de valores faltantes**. Es la técnica de rellenar los valores que hacen falta de una variable o característica, se pueden usar medias, medianas, modas para completar los datos. Otra forma es encontrar esos valores usando modelos predictivos, una solución a los valores faltantes puede ser borrar la variable, aunque esto es una medida drástica, otra forma menos drástica puede ser borrar el registro que contiene el dato perdido, pero se puede eliminar información valiosa.

3. Algoritmos de aprendizaje supervisado

Los algoritmos de aprendizaje supervisado son importantes en la predicción de la información, específicamente usando algoritmos de regresión y algoritmos de clasificación de datos. En este tipo de aprendizaje los algoritmos trabajan con observaciones, los cuales contienen variables de entrada y variables de salida o etiquetas relacionadas con las variables de entrada. Lo principal es comprender la relación entre las variables de salida y los datos de entrada.

Estos son los pasos básicos de una máquina de aprendizaje supervisado:

Figura 1. Pasos básicos de una máquina de aprendizaje no supervisado



Una máquina de aprendizaje supervisado, tienen como pasos básicos:

- Procesamiento y análisis de datos
- Extracción de características y transformación
- Selección de características
- Reducción de dimensiones
- Algoritmo de aprendizaje

- Modelo supervisado
- Valores pronosticados

Las dos etapas básicas son: el procesamiento y el análisis de los datos y la construcción del modelo supervisado. Pero muchas veces es necesario realizar la extracción y la transformación de las características, selección de las características relevantes y la reducción de las dimensiones, que se usan tanto para los datos de entrenamiento, como para los nuevos datos (para los cuales el modelo va a predecir los resultados).

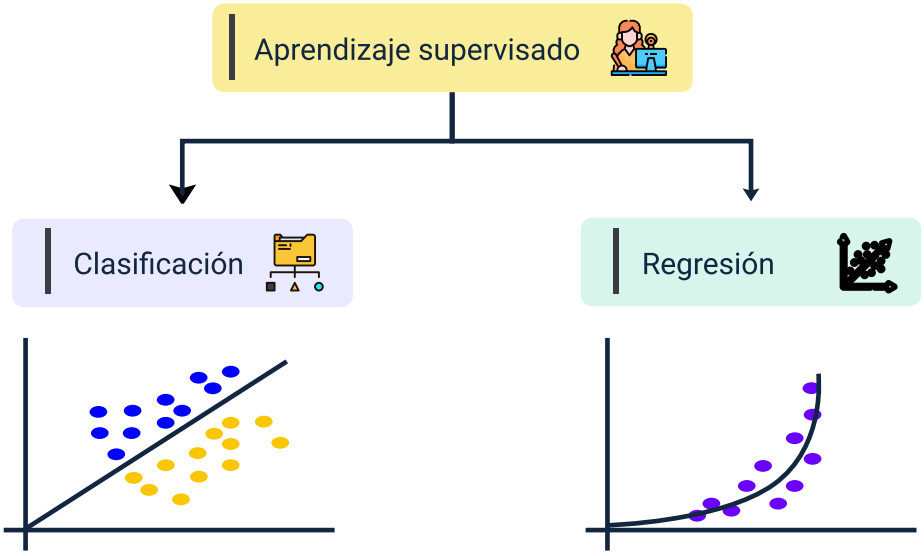
Los datos de entrenamiento y los datos de prueba son otros dos conceptos muy importantes en el aprendizaje automático:

a. **Datos de entrenamiento**. Representan la experiencia que el algoritmo usa para aprender. Cada observación consiste de una variable observada y una o más variables de entrada observadas.
b. **Datos de prueba**. Conjunto de observaciones usadas para evaluar la precisión del modelo usando alguna métrica. Es importante que las observaciones del set de prueba no contengan datos del set de entrenamiento porque, de lo contrario, será difícil saber si el algoritmo aprendió o simplemente memorizó la información.

El modelo entrenado, por tanto, puede ser utilizado, posteriormente, para predecir las salidas de cualquier conjunto nuevo de datos de entrada. Estas técnicas se definen como supervisadas, puesto que el modelo aprende de la muestra de los datos y de sus salidas en la fase de entrenamiento.

Las técnicas de aprendizaje supervisado contemplan dos clases principales: la clasificación y la regresión, dependiendo del tipo de problema de aprendizaje automático por resolver.

Figura 2. Aprendizaje supervisado



Los resultados por predecir en un algoritmo de clasificación son datos categóricos, por tanto, cada salida corresponde a una clase o categoría de tipo discreto; por otro lado, los resultados a predecir, usando un algoritmo de regresión, son valores numéricos continuos.

3.1. Algoritmos de regresión

Uno de los principales objetivos del aprendizaje automático es estimar un valor y esto se puede realizar mediante tareas de regresión.

Sobre los algoritmos de regresión, tenga presente:

- Los datos usados en los modelos de regresión usan atributos (características o variables de entrada).
- Estas características o variables de entrada se conocen como variables independientes, explicativas o predictoras.
- Los respectivos valores de salida numéricos continuos (de estas variables) se conocen como variables de respuesta, dependientes o de resultado.
- Los algoritmos de predicción basados en la regresión hacen uso de esta información y aprenden a relacionar las entradas con sus respectivas salidas.
- Con este conocimiento, entonces, ya se puede predecir las respuestas de los registros nuevos.

Por su parte, la regresión lineal simple es usada para estimar los valores del mundo real, tales como el precio de las casas, el número de llamadas, las ventas totales, etc., basada en las variables continuas.

El objetivo de la regresión es obtener la mejor ecuación lineal que será el modelo para representar una relación entre variables predictoras y variables dependientes. La línea de mejor ajuste es conocida como línea de regresión.

Una empresa, por ejemplo, podría usar la regresión lineal para conocer si se están disminuyendo las ventas o está creciendo el número de clientes, de manera que para las ventas futuras estas predicciones son útiles porque permiten tomar decisiones en cuanto a las ventas y el crecimiento.

Estos son algunos modelos de regresión y generalidades de los mismos:

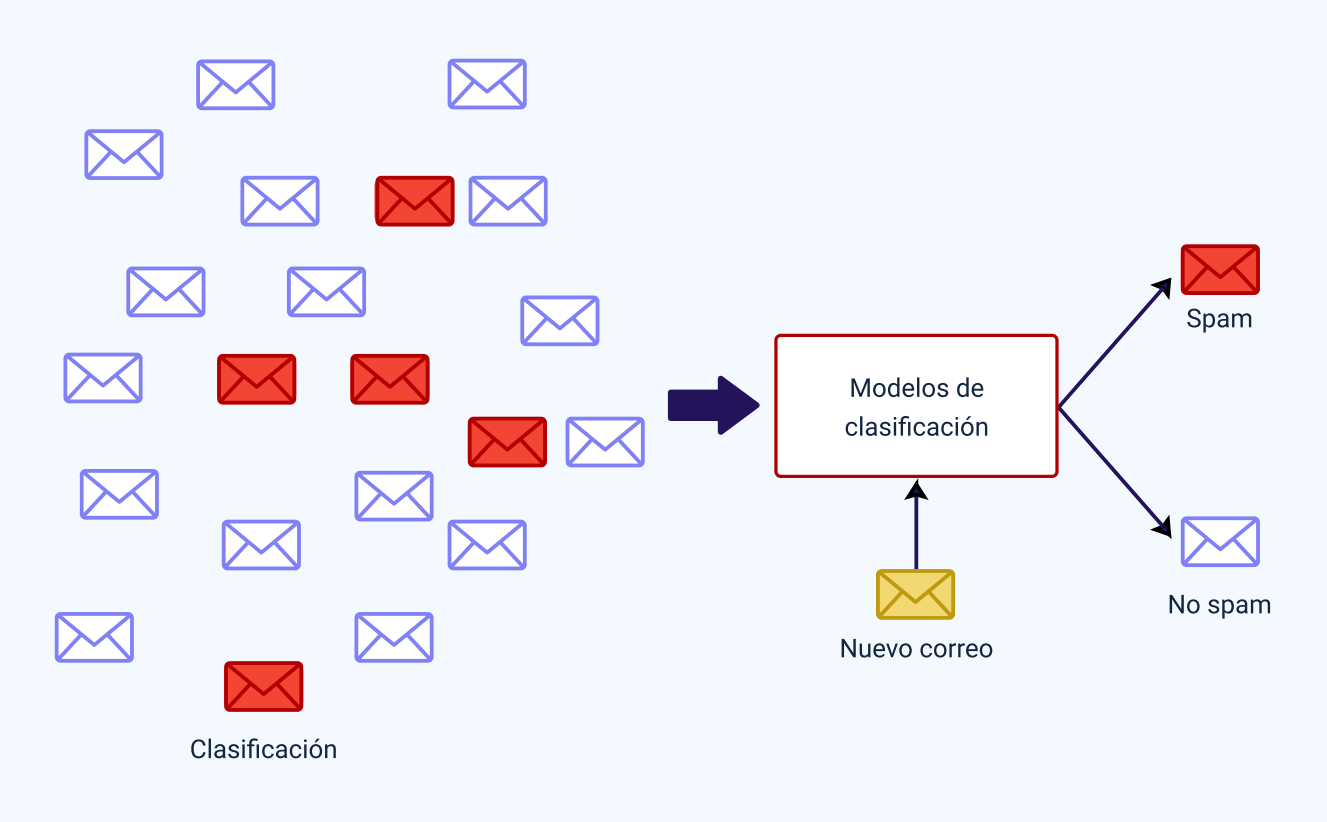
a. **Modelo de regresión lineal simple**. Se caracteriza por establecer la relación entre dos variables mediante una línea recta y una línea recta en el plano cartesiano. Se puede construir con una pendiente y una intersección.
   Un término importante es la propiedad coeficiente de correlación de una regresión lineal que mide el grado de relación entre las variables predictoras y dependientes, las cuales deben ser cuantitativas y continuas.
b. **Modelo de regresión lineal múltiple**. Cuando una variable dependiente no se puede explicar con una sola variable dependiente, se puede usar la regresión lineal múltiple: puede ser que se requiera usar varias variables dependientes para explicar la variable dependiente.
c. **Uso de los modelos de regresión**. Los modelos de regresión, por ejemplo, lineal múltiple, se pueden usar en muchas aplicaciones, pero hay que tener en cuenta algunos posibles problemas que pueden ocasionar que la regresión no sea un modelo óptimo.
d. **Multicolinealidad o dependencia casi lineal**. Se presenta cuando hay una relación suficientemente grande entre las variables predictoras **xi**.
e. **Heterocedasticidad**. Al aumentar la variable explicativa se observa mayor variación en la nube de puntos de la variable dependiente.

3.2. Algoritmos de clasificación

Los algoritmos de clasificación son un campo del aprendizaje supervisado, en el cual el objetivo es predecir etiquetas de salida o variables de naturaleza categórica, relacionadas con lo que el modelo ha aprendido en el entrenamiento. Cada respuesta de salida pertenece a una categoría o clase de tipo discreto.

El siguiente, es un ejemplo de este tipo de algoritmo:

Figura 3. Ejemplo algoritmos de clasificación



La imagen que ejemplifica el algoritmo de clasificación muestra que, si una variable de entrada al algoritmo de clasificación es un correo electrónico, el resultado puede ser que el correo es “spam” o no “spam”. Previamente, el algoritmo ha sido entrenado con correos etiquetados como “spam” y correos etiquetados con no “spam”.

En el caso de las compras de un cliente potencial el resultado es: sí compra o no compra; el resultado de un tipo de tumor puede ser maligno o benigno; un comportamiento de una transacción puede ser una anómala, o no, y un cliente puede pagar un crédito, o no.

Otros ejemplos de aplicaciones de algoritmos de clasificación son:

- La detección de fraudes.
- La detección de “spam”.
- La clasificación de enfermedades.
- Diagnósticos basados en la edad, el sexo y el azúcar en la sangre.
- La clasificación de imágenes.
- La identificación de caracteres y la clasificación de posibles clientes.

Regresión logística

Es un modelo de clasificación que usa un método de clasificación binaria o multinomial. Se trata de un método estadístico muy simple y eficiente para resolver problemas de clasificación lineal y es muy usado en la industria.

La respuesta de la regresión logística puede ser una variable dicotómica, es decir, tiene dos respuestas posibles. Por ejemplo, sí o no, “ok” o no “ok”, positivo o negativo, etc., cuando la regresión logística es binaria; pero también existe la regresión logística multinomial en la que los valores de salida pueden presentar tres o más categorías de tipo nominal:

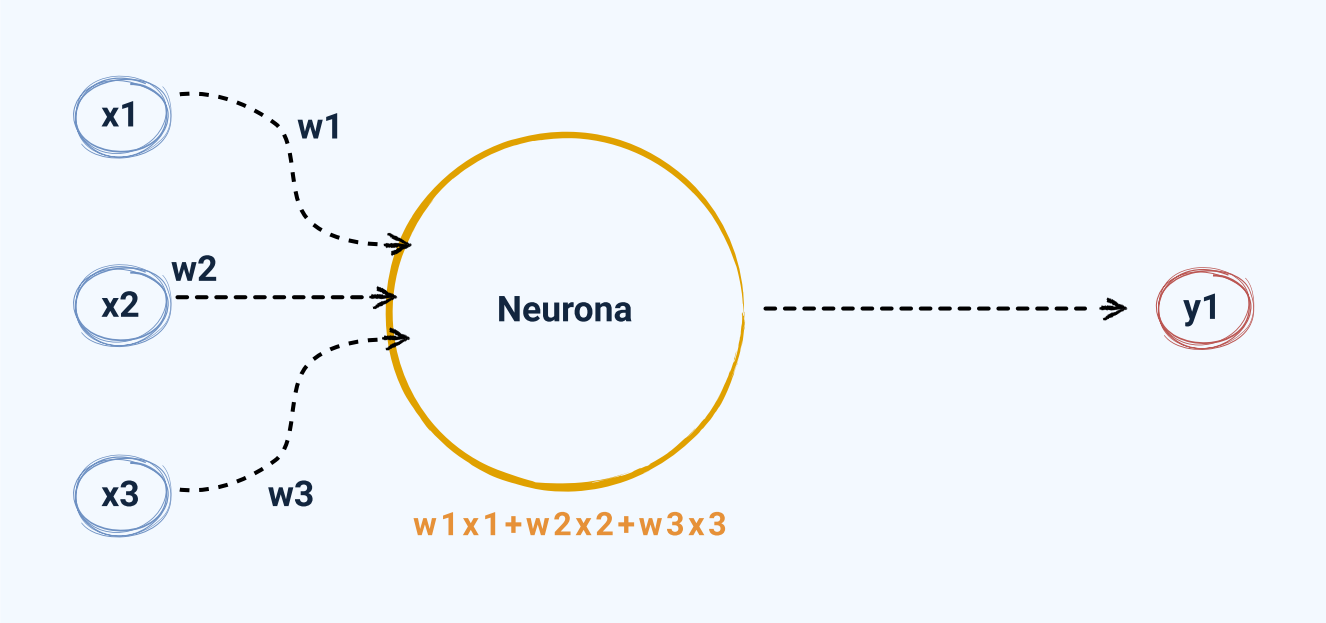
$$logit\ (p) = ln \frac{p}{1 - p} = B_0 + B_1 x_1 + B_2 x_2 + \ldots + B_k x_k$$

La ecuación busca relacionar la probabilidad (p) de que ocurra cierto evento, por ejemplo, la probabilidad de tener una enfermedad de corazón, o no tenerla, depende de las variables independientes, las cuales pueden ser las medidas del nivel de azúcar, el nivel de triglicéridos o alguna lectura necesaria para predecir el resultado.

Redes neuronales artificiales

Una neurona es la unidad básica de procesamiento que se va a encontrar dentro de una red neuronal; es similar a una neurona en el cerebro: recibe estímulos a través de conexiones de entrada, con estos valores la neurona realiza un cálculo interno y genera un valor de salida.

Figura 4. Neurona

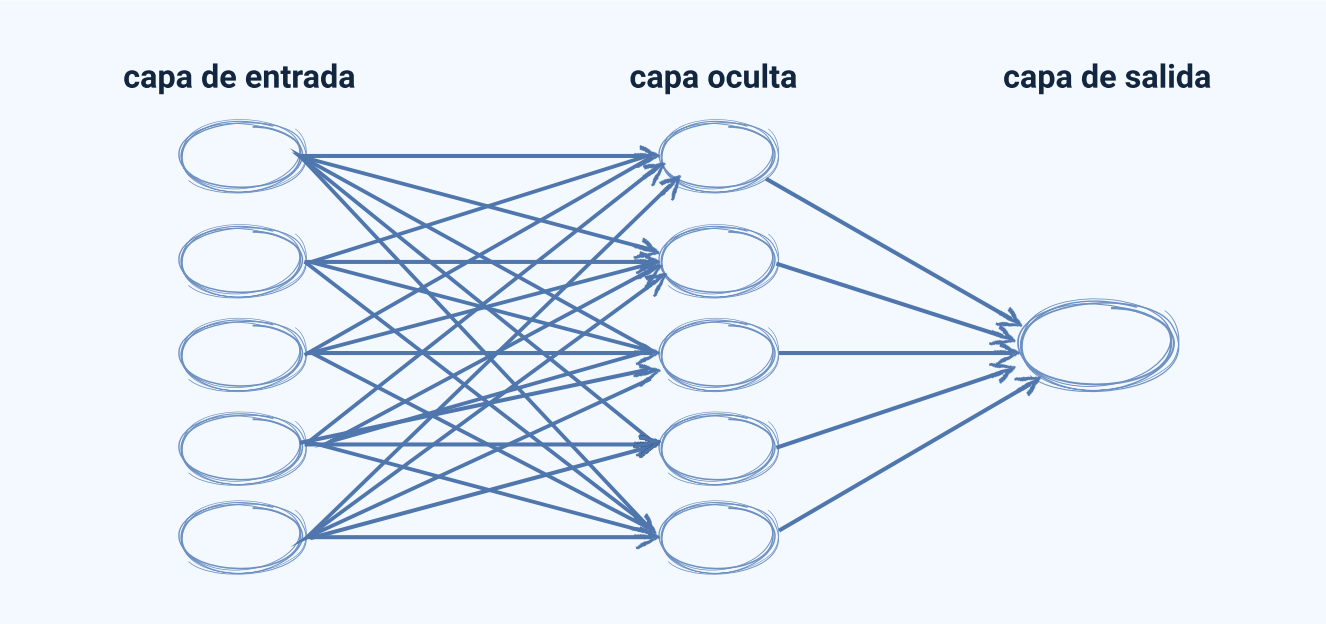


Según el esquema gráfico de la neurona (unidad básica de procesamiento), internamente una neurona utiliza todos los valores de entrada para realizar una suma ponderada de esos valores, la ponderación de cada una de las entradas viene dada por el peso asignado a cada conexión. La ilustración muestra que la función interna de la neurona se asemeja a una regresión lineal.

Como una sola neurona no puede resolver muchos problemas, entonces, se construyen redes neuronales, una red neuronal se construye en capas y cada capa tiene neuronas.

Existen capas intermedias llamadas capas ocultas. Los valores de entrada se fijan en la primera capa o capa de entrada y los valores se transportan por todas las capas hasta la capa de salida. Cada unidad de conexión tiene un peso o una ponderación.

Figura 5. Red neuronal



Una red neuronal tiene una capa de entrada que es donde se reciben los datos de entrada y tiene, al menos, una capa de salida; allí se tienen los resultados calculados.

Al principio todas las ponderaciones son aleatorias y los resultados pueden ser ilógicos y erróneos, pero la red va aprendiendo a través del entrenamiento, a medida que se entrena la red con datos conocidos la predicción de la red mejora sustancialmente.

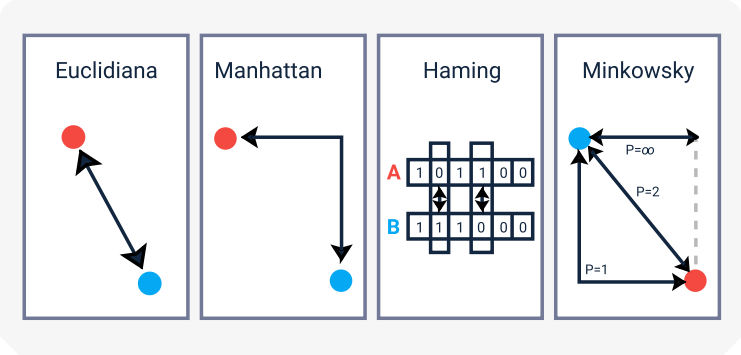
K-vecinos más cercanos

Es un método de clasificación que estima la probabilidad de que un dato sea parte de un grupo u otro, basado en que el grupo de este dato es más cercano. Es un

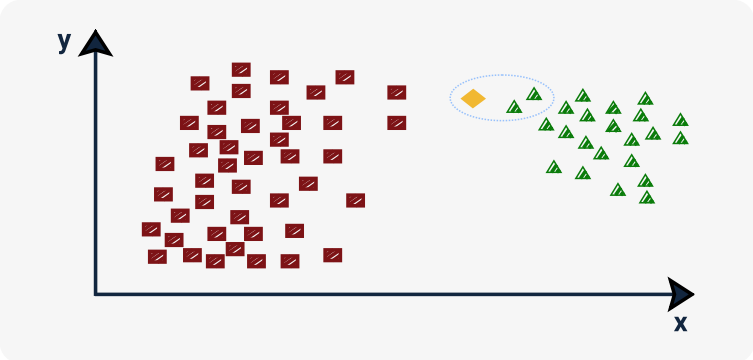
método no paramétrico, no tiene en cuenta las distribuciones de los datos, el principal objetivo es determinar a qué grupo de datos pertenece un punto.

Para entender mejor este tipo de algoritmo revise los siguientes pasos en los que se resaltan algunos ejemplos:

- **Ejemplo 1**. Para clasificar una instancia desconocida el algoritmo calcula la distancia entre el punto y los puntos en los datos de entrenamiento. Esta distancia se calcula usando la distancia euclidiana, manhattan, haming, minkowsky. La más usada es la distancia euclidiana.

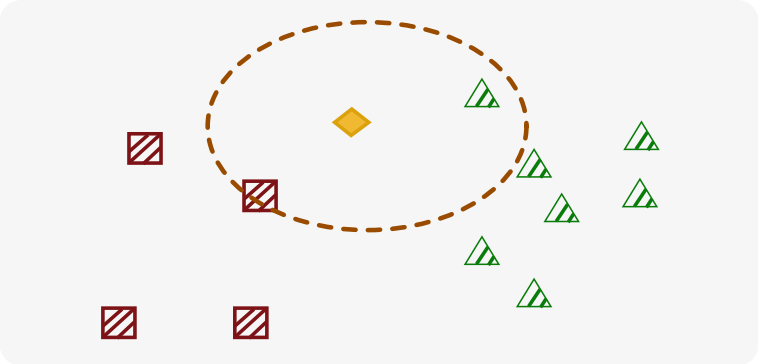


- Ejemplo 2. Si se desea clasificar el punto amarillo y se escoge k = 2, los dos puntos más cercanos son los puntos verdes, en este caso el punto amarillo se clasifica como verde.



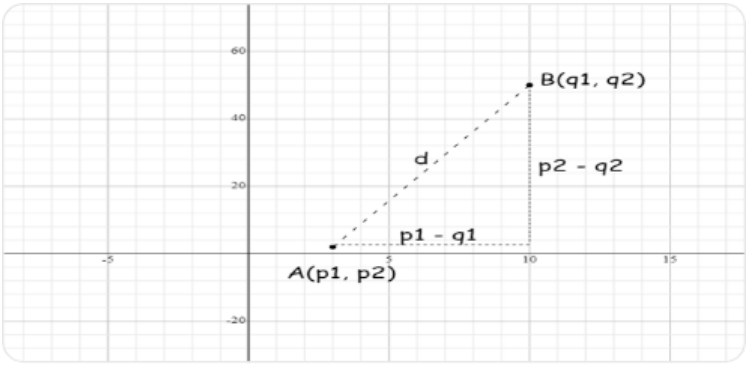
- **Ejemplo 3**. Si se desea clasificar el punto amarillo y se escoge k=3, se observa que los dos puntos más cercanos son verdes y el tercer punto más

cercano es rojo; se tienen dos verdes y uno rojo, por tanto, por mayoría, se clasifica el punto desconocido como parte de los verdes.



- **Ejemplo 4**. La distancia euclideana para calcular distancias entre puntos es una de las más utilizadas en “Machine Learning”.

- **Ejemplo 5**. Si las variables están en rangos muy diferentes, entonces, no tiene mucho sentido calcular la distancia y se deben normalizar las variables.



En esta imagen, las variables se encuentran en plano muy distantes y se hace necesario normalizarlas.

Árboles de clasificación

Son una técnica de "Machine Learning" de aprendizaje supervisado que predice las respuestas mediante las reglas de decisión, así el algoritmo permite segmentar y clasificar los distintos objetos.

Para entender su estructura revise, detenidamente, la siguiente información:

a. **Su estructura**. Es similar a un árbol, se inicia por la raíz que representa la base de datos, luego las ramas representan las subdivisiones que son los criterios que permiten dividir la base de datos, luego las hojas del árbol representan las decisiones o los resultados finales.
b. **"Data"**. En muchos casos los datos no están estrictamente separados, por lo que se deben tomar, entonces, muchas decisiones sobre la data.
c. **Modelo**. Si se tiene un punto cualquiera se inicia la evaluación como en el modelo, revisando la ordenada y abscisa del punto y descartando opciones. Se presume que un árbol de clasificación es muy fácil de entender e interpretar, se identifican fácilmente variables significativas. Además, se requiere menos limpieza de los datos y no es sensible a valores atípicos o "outliers".

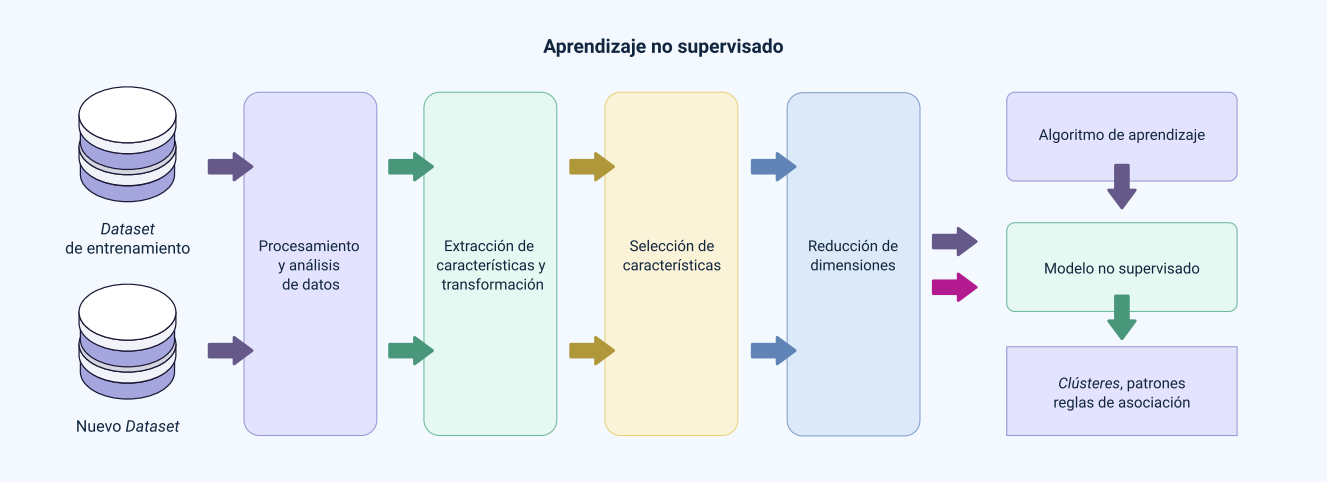
4. Algoritmos de aprendizaje no supervisado

Tienen la misión de descubrir las similitudes, los patrones o uniformidades dentro de los datos de entrada, en este caso no se cuenta con un supervisor que etiquete los datos.

Los algoritmos de este tipo de aprendizaje forman clústeres de manera autónoma y asignar observaciones a estos clústeres.

Estos son los pasos básicos de una máquina de aprendizaje no supervisado:

Figura 6. Pasos básicos de una máquina de aprendizaje no supervisado



La figura muestra las dos etapas básicas que son el procesamiento y análisis de los datos y la construcción del modelo no supervisado; pero muchas veces es necesario realizar la extracción y la transformación de las características, seleccionar las características relevantes y la reducción de las dimensiones, el resultado final son clústeres y reglas de asociación.

Los algoritmos no supervisados se agrupan en problemas de asociación y agrupación, así:

a. **Asociación**. Trabaja en función de reglas de asociación, que permiten establecer asociaciones entre los datos dentro de bases de datos grandes. Por ejemplo, si un usuario compra un carro nuevo, tiene probabilidades de comprar un seguro contra accidentes.

   Así los algoritmos combinan los datos basándose en los atributos que se comparten, aquí la idea no es encontrar semejanzas entre ellos, si no encontrar relaciones entre los datos. Un concepto en este tipo de aprendizaje son las reglas de asociación que se usan para extraer conocimiento, estas reglas se obtienen de los datos históricos, identificando relaciones entre los datos.

b. **Agrupamiento**. Se trata de identificar un patrón en datos no categorizados y agrupados en clústeres o grupos. Se supone que los datos tienen similitudes identificadas por métricas de distancia, tales como la distancia euclídea.

   Así, dos registros que tengan una distancia mínima, en comparación con otras distancias, podrían pertenecer al mismo clúster.

4.1. Reducción de dimensiones

Una de las técnicas para reducir dimensiones es el análisis de componentes principales o PCA, el cual es un procedimiento estadístico que permite resumir el número de variables o reducir el número de dimensiones; este método es parte del aprendizaje no supervisado.

Sobre la reducción de dimensiones, tenga en cuenta:

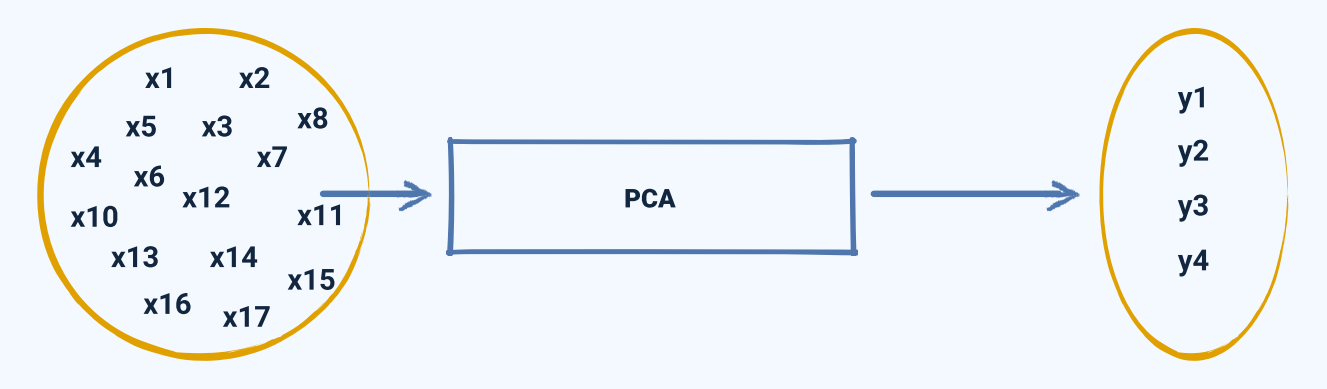
a. Cuando se tienen un conjunto de clientes, usuarios, pacientes etc., lo más común es obtener la media de una de sus variables como la edad o el peso; con un solo número se resumen las observaciones de las variables.

b. De ese conjunto de personas es posible tener muchas variables diferentes, como datos de información personal y transacciones realizadas.

c. Se pueden tener muchas variables obtenidas de las redes sociales y tener esta cantidad abrumadora de variables puede ser un problema a la hora de analizar la información.

La figura que se muestra enseguida, facilita la comprensión del análisis de componentes principales:

Figura 7. Análisis de componentes principales – PCA



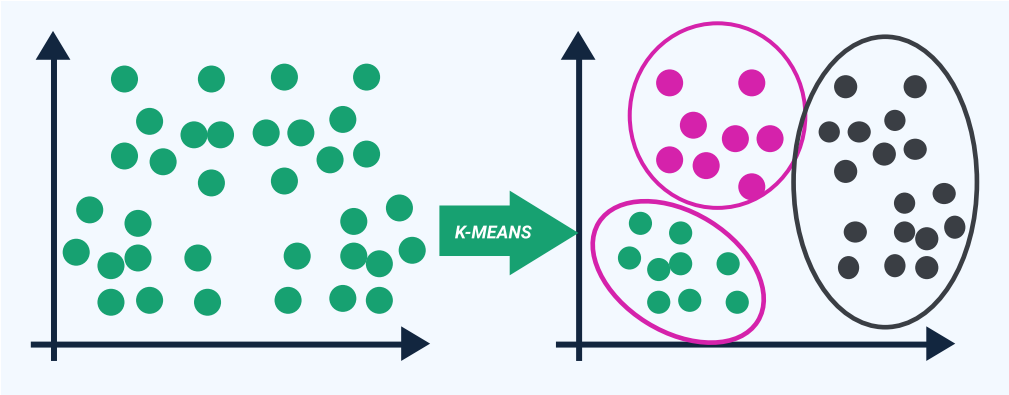
El análisis de los componentes principales o PCA permite resumir, en unas pocas variables, grandes cantidades de variables de información en un número pequeño de variables, conocidas como componentes principales; esta técnica es muy útil en la compresión de imágenes o para reducir variables y usarlas en otros métodos de aprendizaje, como la regresión lineal.

4.2. "Clustering"

Uno de los algoritmos más comunes de aprendizaje no supervisado utilizado en ciencia de datos es el "k-medias" o "k-means"; el objetivo es agrupar datos con

características similares e identificar patrones que muchas veces no se pueden detectar fácilmente o a simple vista.

Figura 8. "Clustering"



Como se muestra en la figura, el algoritmo elige aleatoriamente una cantidad k de centroides iniciales que marcan el centro de cada clúster; cada punto se ubica con su centroide más cercano, usando cualquier medida de distancia tal como la distancia euclidiana.

Una vez que el algoritmo elige una cantidad de centroides para marcar el centro de cada clúster, se actualizan los centroides cambiando su posición al centro de las observaciones asignadas y, nuevamente, cada punto se ubica con su centroide más cercano y así sucesivamente hasta que las asignaciones de clústeres no cambien o se alcance un número de iteraciones determinado.

Ventajas de "K-means"

Las ventajas de "K-means" son:

- Es un algoritmo veloz y eficiente en términos de costo computacional para segmentar los datos.

- Es sencillo de implementar y de aplicar.
- Produce clústeres más definidos que el “clustering” jerárquico.
- Puede manejar grandes datos.

5. Métricas de evaluación

Es fundamental medir el rendimiento del modelo entrenado, el modelo generaliza sobre los datos no vistos, es lo que define a los modelos de aprendizaje automático adaptables frente a los no adaptables.

Al hacer uso de las diferentes métricas para la evaluación del rendimiento del modelo se tiene la posibilidad de mejorar la predicción del mismo antes de ponerlo en marcha en la producción de los datos no vistos con anterioridad.

En relación con las métricas de evaluación, tenga presente:

- Si no se realiza una evaluación adecuada del modelo aprendizaje automático utilizando diferentes métricas y se usa solo la precisión, puede darse un problema cuando el modelo respectivo se despliega sobre datos no vistos y puede dar lugar a malas predicciones.
- Esto sucede porque los modelos no aprenden, sino que memorizan; por lo tanto, no pueden generalizar bien sobre datos no vistos.

Métricas de evaluación del modelo

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:

Figura 9. Representación matriz de confusión

		Resultado de la predicción	
		Positivo	Negativo
Valor actual	Positivo	TN Verdadero Negativo	FP Falso Positivo
	Negativo	FN Falso Negativo	TP Verdadero Positivo

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.

a. **Error de TIPO I**. Este error equivale a los falsos positivos (FP), es el rechazo de una hipótesis nula, pero esta es verdadera.
b. **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:

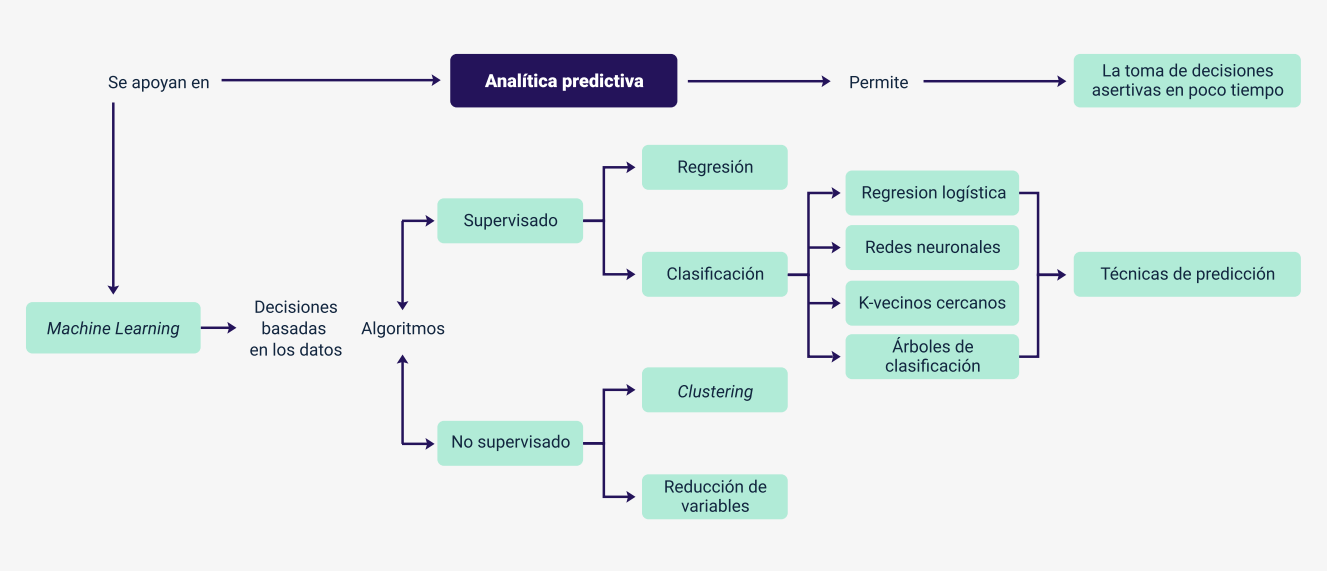
- **Exactitud**. Esta métrica es usada cuando las clases son aproximadamente iguales en tamaño; se encarga de medir el porcentaje de casos en que el modelo ha acertado:
  Exactitud = (TP + TN) / (TP + TN + FP + TP)
- **Precisión**. Esta métrica es usada para medir la calidad del modelo en tareas de clasificación. 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))

Síntesis

Aquí finaliza el estudio de las temáticas de este componente formativo. En este punto, analice el esquema que se presenta a continuación y haga su propia síntesis de los contenidos. ¡**Adelante**!



La estructura de contenidos de este componente formativo abordó generalidades y aspectos clave de la introducción a la analítica de datos. El éxito de las organizaciones depende, en gran medida, de la forma cómo manejan la información, organizan sus datos y de las herramientas que implementan para realizar un análisis adecuado de los mismos, que resulte confiable y que no implique mayor tiempo para desarrollarlo. Desde esta perspectiva conocer la forma de implementar modelos que les permita clasificar y hacer uso de los datos de manera adecuada, alcanzando así las metas organizacionales planteadas.

Material complementario

Tema	Referencia	Tipo de material	Enlace del recurso
3. Algoritmos de aprendizaje supervisado	SDC LEARNING. (2022). Webinar gratuito Mi primer modelo de Machine Learning en Python [video]. YouTube.	Video	https://www.youtube.com/watch?v=9HKfqinJJAo
3. Algoritmos de aprendizaje supervisado	AprendeIA con Ligdi González. (2019). Ventajas y desventajas algoritmos de regresión [video]. YouTube.	Video	https://www.youtube.com/watch?v=TkHO2HHbJDs
3. Algoritmos de aprendizaje supervisado	Parra, F. (2022). Métodos de clasificación. bookdown.org.	Página web	https://bookdown.org/content/2274/metodos-de-clasificacion.html
5. Métricas de evaluación	González, L. (2019). Errores modelos clasificación.	Página web	https://aprendeia.com/evaluando-el-error-en-los-modelos-de-clasificacion-machine-learning/

Glosario

Aprendizaje automático: rama de la inteligencia artificial, cuyo objetivo es implementar técnicas que permitan a los computadores aprender mediante un proceso de inducción del conocimiento.

Aprendizaje automático no supervisado: hace referencia al proceso en el cual el algoritmo identifica patrones y saca conclusiones de los datos que se le proporcionan.

Aprendizaje automático supervisado: hace referencia al proceso en el cual el algoritmo recibe los datos de entrenamiento consistente en los datos etiquetados.

Entrenamiento: proceso que se realiza para que los modelos aprendan de los datos.

Evaluación: análisis de eficiencia con el que el modelo predice los datos, generalmente se contrasta con una colección de pruebas separadas previamente.

Inteligencia artificial: sistemas informáticos que pueden aprender como aprende un ser humano.

"**k-means**": lenguaje de alto nivel, usado para construir todo tipo de aplicaciones y muy usado en la ciencia de datos.

Matriz de confusión: es una métrica para establecer el nivel de error, precisión y otras medidas en los modelos de "Machine Learning".

Predicciones: capacidad del modelo para clasificar entradas nuevas, de acuerdo con un entrenamiento previo.

Preprocesamiento: manipulación que se realiza a los datos con el objetivo de entregarlos al modelo como este lo requiera.

Python: proceso criptográfico que proporciona comunicaciones seguras a través de las redes, haciendo que la información entre extremos se transporte de forma segura mediante el uso de la criptografía.

Referencias bibliográficas

González, L. (2019). Preguntas frecuentes. Regresión lineal y regresión logística. Aprende IA. https://aprendeia.com/diferencia-entre-regresion-lineal-y-regresion-logistica-machine-learning/

Kaggle. (2016). SMS Spam Collection Dataset. Kaggle. https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset

Miller, V. (2018). Explorando algoritmos de aprendizaje automático supervisado. Toptal Engineering Blog. https://www.toptal.com/machine-learning/explorando-algoritmos-de-aprendizaje-automatico-supervisado

Roman, V. (2019). Machine Learning: cómo desarrollar un modelo desde cero. Medium. https://medium.com/datos-y-ciencia/machine-learning-c%C3%B3mo-desarrollar-un-modelo-desde-cero-cc17654f0d48

Sotaquirá, M. (2021). ¿Se requiere SQL para trabajar en Machine Learning? Codificandobits. 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	Centro Industrial del Diseño y la Manufactura - Regional Santander
Héctor Henry Jurado Soto	Experto temático	Centro de Teleinformática y Producción Industrial - Regional Cauca
Fabián Leonardo Correa Díaz	Diseñador instruccional	Centro Industrial del Diseño y la Manufactura - Regional Santander
Carlos Eduardo Garavito Parada	Animador y Productor Multimedia	Centro Industrial del Diseño y la Manufactura - Regional Santander
Wilson Andrés Arenales Cáceres	"Storyboard" e ilustración	Centro Industrial del Diseño y la Manufactura - Regional Santander
Camilo Andrés Bolaño Rey	Locución	Centro Industrial del Diseño y la Manufactura - Regional Santander
Yerson Fabián Zarate Saavedra	Diseñador de Contenidos Digitales	Centro Industrial del Diseño y la Manufactura - Regional Santander
Andrea Paola Botello De la Rosa	Desarrollador "Fullstack"	Centro Industrial del Diseño y la Manufactura - Regional Santander
Emilsen Alfonso Bautista	Actividad didáctica	Centro Industrial del Diseño y la Manufactura - Regional Santander
Daniel Ricardo Mutis Gómez	Evaluador para Contenidos Inclusivos y Accesibles	Centro Industrial del Diseño y la Manufactura - Regional Santander
Zuleidy María Ruíz Torres	Validador de Recursos Educativos Digitales	Centro Industrial del Diseño y la Manufactura - Regional Santander
Luis Gabriel Urueta Álvarez	Validador de Recursos Educativos Digitales	Centro Industrial del Diseño y la Manufactura - Regional Santander