**FORMATO PARA EL DESARROLLO DE COMPONENTE FORMATIVO**

|  |  |
| --- | --- |
| Programa de formación | Creación de modelos aplicación de algoritmos de predicción con Python |

|  |  |  |  |
| --- | --- | --- | --- |
| Competencia | 220501114. Sistematizar datos masivos de acuerdo con métodos de analítica y herramientas tecnológicas. | Resultado de aprendizaje | 220501114-01. Seleccionar el algoritmo de predicción según los requerimientos establecidos. |

|  |  |
| --- | --- |
| Número del componente formativo | 01 |
| Nombre del componente formativo | 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. |
| Palabras clave | Análisis de datos, aprendizaje supervisado, aprendizaje no supervisado, inteligencia artificial, métricas de evaluación. |

|  |  |
| --- | --- |
| Área ocupacional | 6 - Ventas y Servicios |
| Idioma | Español |

1. **TABLA DE CONTENIDOS**

**Introducción**

1. **Conjunto de datos**
2. **Preprocesamiento de datos**
3. **Algoritmos de aprendizaje supervisado**

3.1 Algoritmos de regresión

3.2 Algoritmos de clasificación

**4. Algoritmos de aprendizaje no supervisado**

4.1 Reducción de dimensiones

4.2 *Clustering*

**5. Métricas de evaluación**

1. **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:

|  |
| --- |
| DI\_CF01\_0\_Video\_Introduccion |

1. **DESARROLLO DE CONTENIDOS**
   * + 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 *softwar*e tradicional. |

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

|  |
| --- |
| DI\_CF01\_1\_Slider\_ConjuntosDeDatos |

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:

|  |
| --- |
| DI\_CF01\_1\_InfografiaInteractiva\_ConceptosConjuntosDeDatos |

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**:

|  |
| --- |
| DI\_CF01\_1\_Pestañas\_VariablesCualitativasYCuantitativas |

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

|  |
| --- |
| DI\_CF01\_1\_Slide\_GradosDeMedicion |

* + - 1. **Preprocesamiento 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:

|  |
| --- |
| DI\_CF01\_2\_Acordeon\_FormasDeProcesarInformacion |

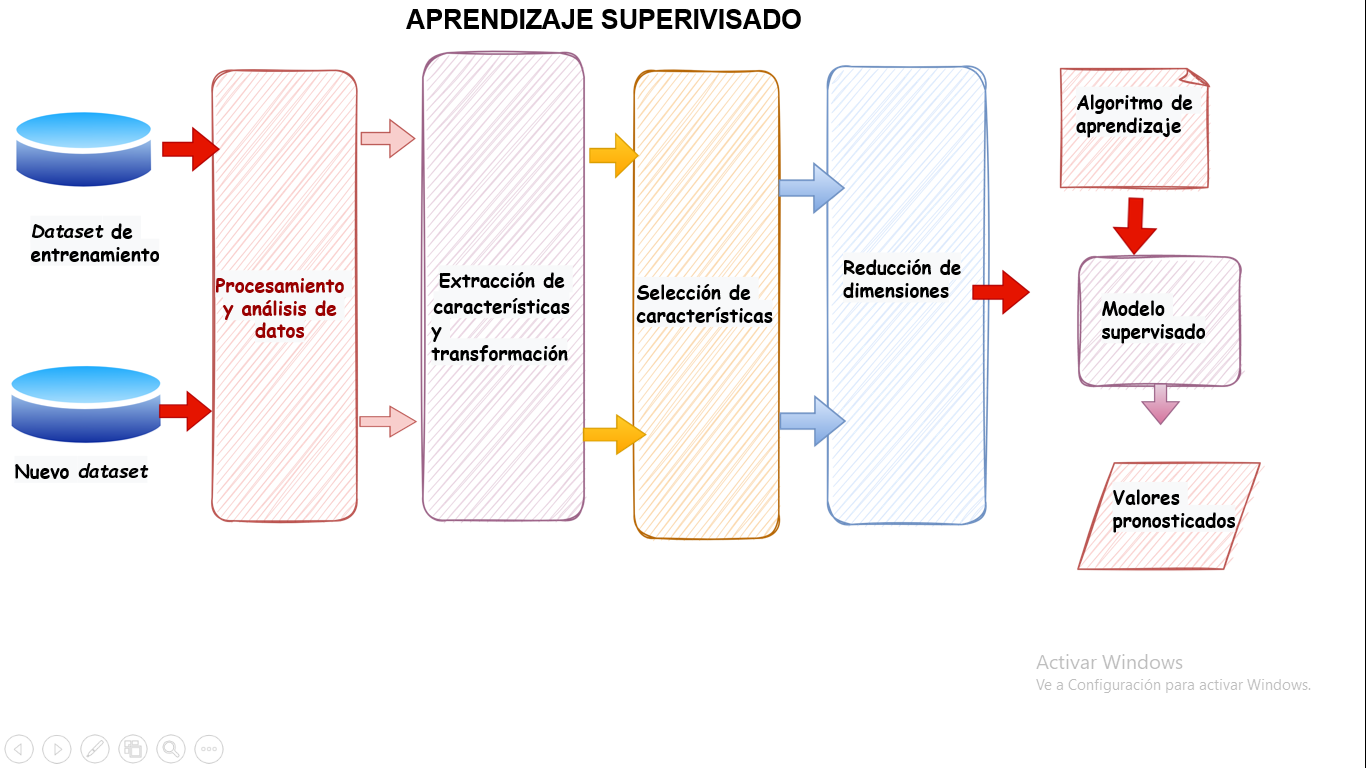
* + - 1. **Algoritmos de aprendizaje supervisado**

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 supervisado*



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:

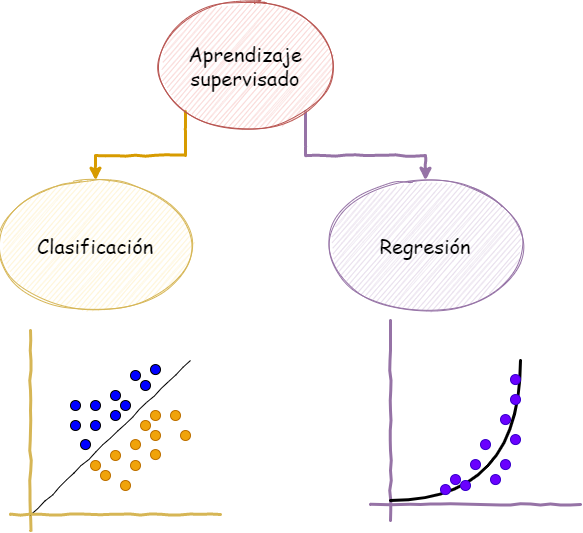
|  |
| --- |
| DI\_CF01\_3\_Slide\_ConceptosAprendizajeAutomatico |

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*



**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:

|  |
| --- |
| DI\_CF01\_3-1\_InfografiaInteractiva\_ModelosDeRegresion |

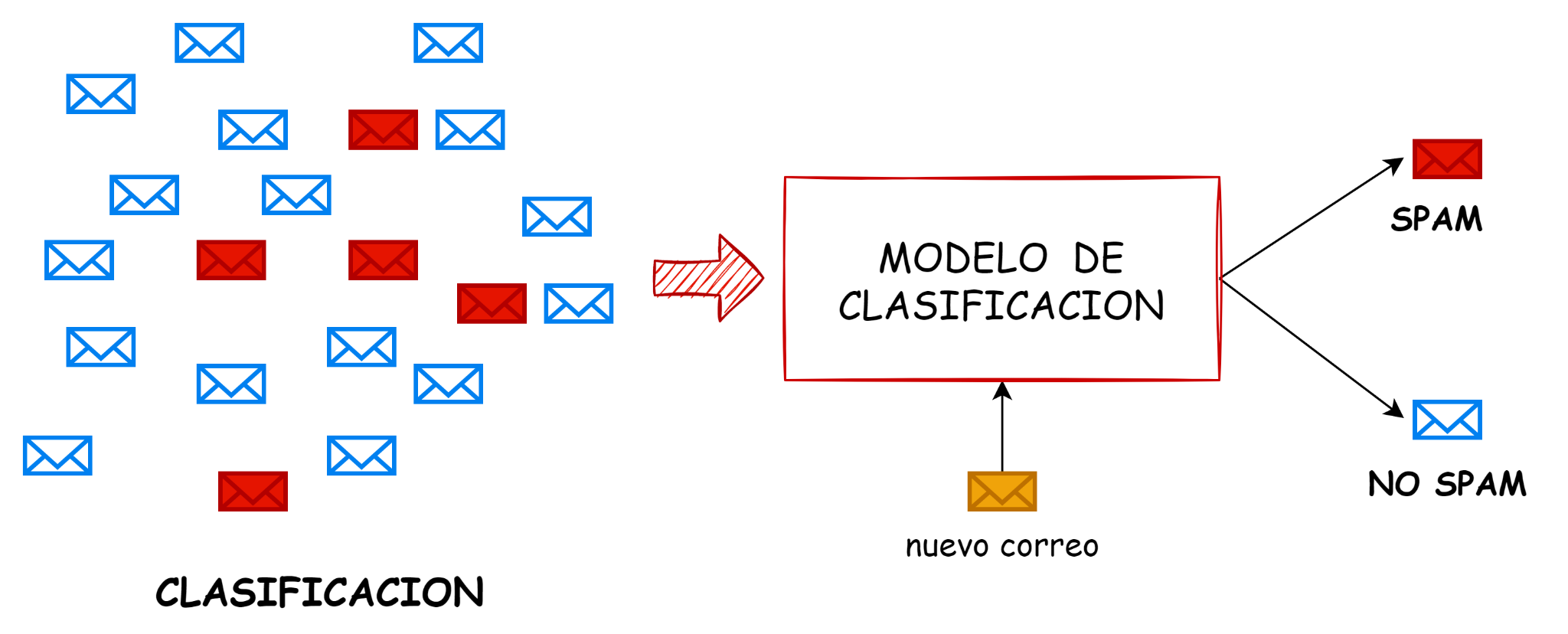
**3.2. Algoritmos de clasificación**

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*



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.

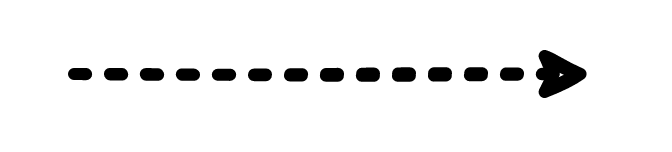
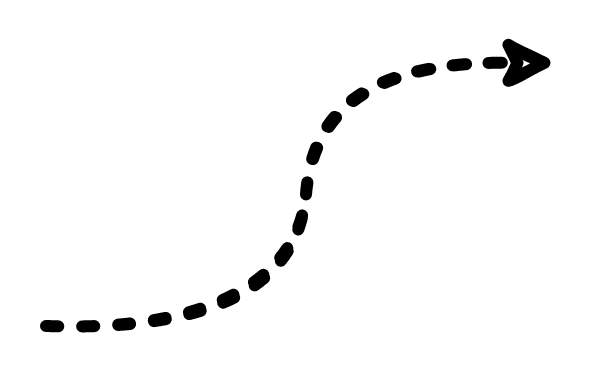
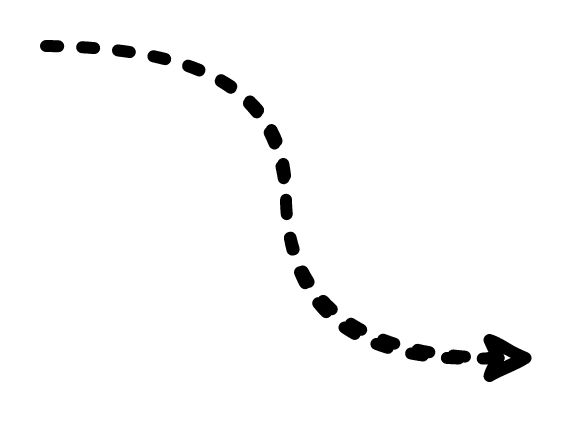
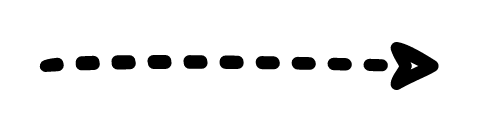
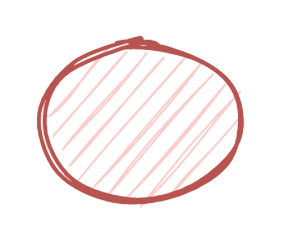
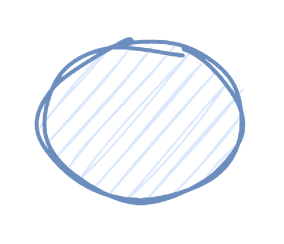
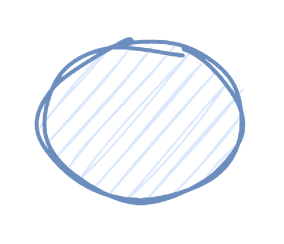
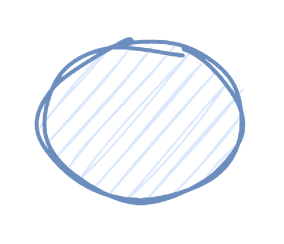
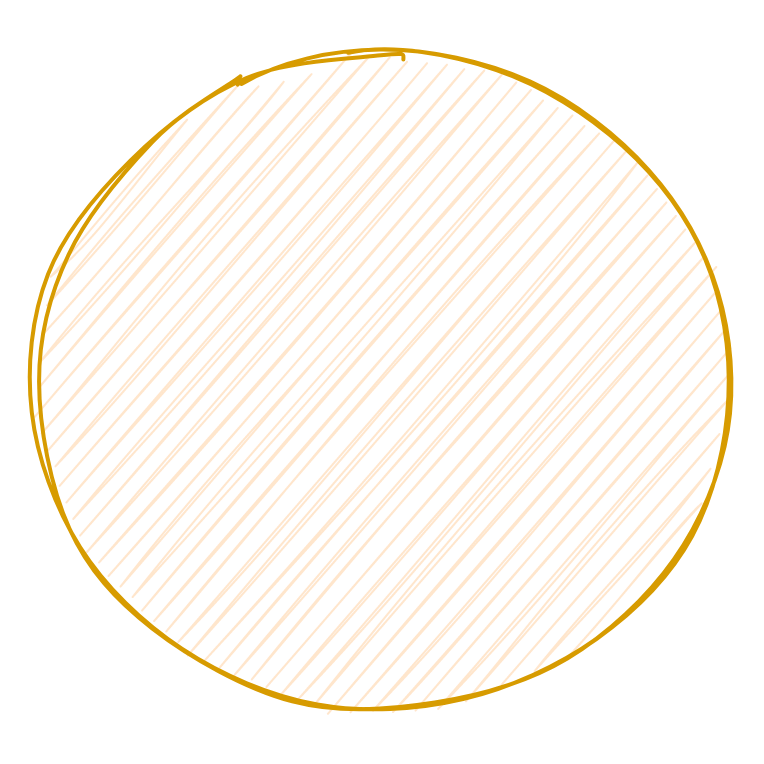
La respuesta de la regresión logística puede ser una variable dicotómica, es decir, tiene dos respuestas posibles. Por ejemplo, sí o no, *ok* o no *ok*, positivo o negativo, etc., cuando la regresión logística es binaria; pero también existe la **regresión logística multinomial** en la que los valores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

**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*



x1

x2

x3

y1

w1

w2

w3

w1x1+w2x2+w3x3

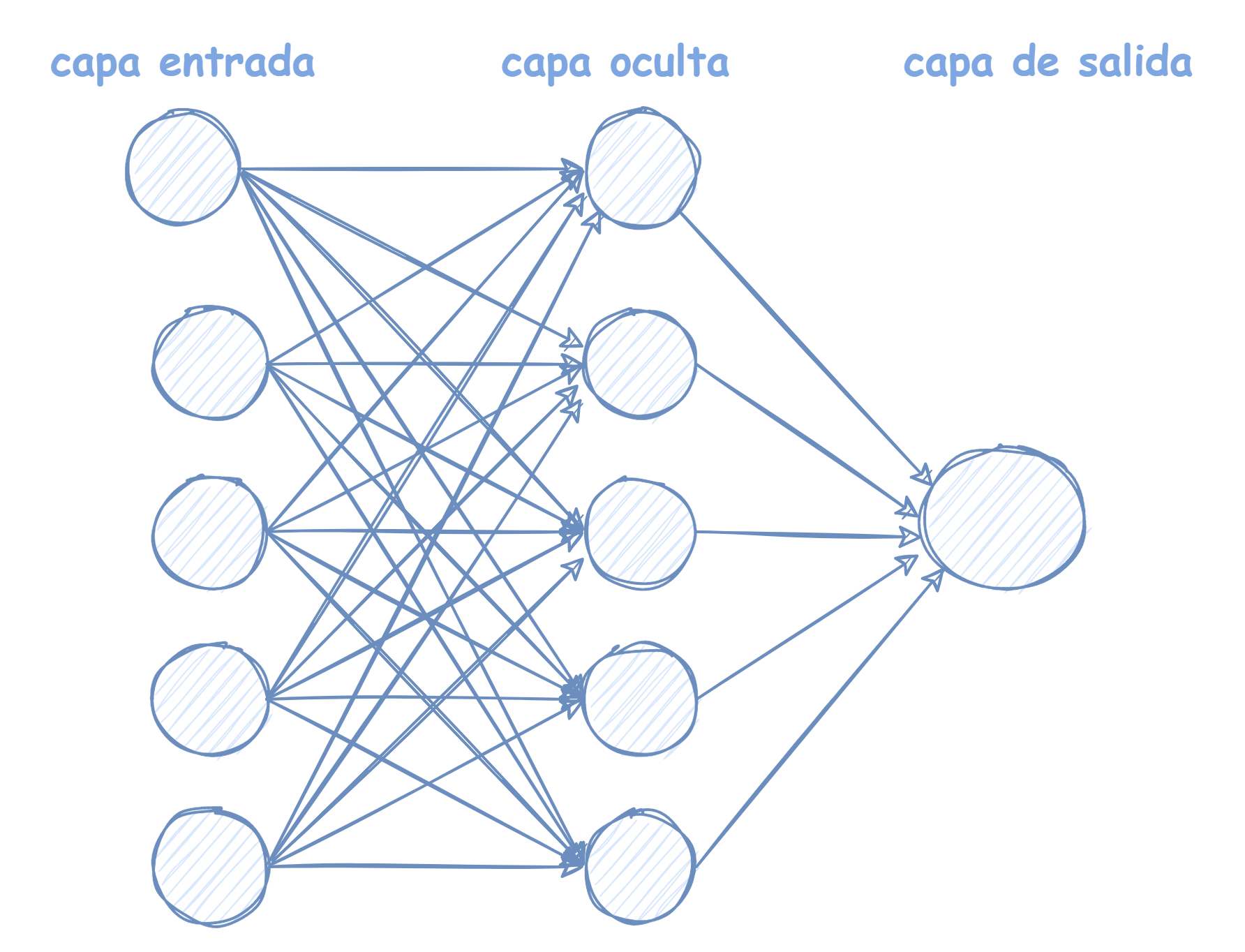
**NEURONA**

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*



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:

|  |
| --- |
| DI\_CF01\_3-2\_Pasos\_EjemploK-vecinosmascercanos |

**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:

|  |
| --- |
| DI\_CF01\_3-2\_Slide\_ArbolesDeClasificacion |

1. **Algoritmos de aprendizaje no supervisado**

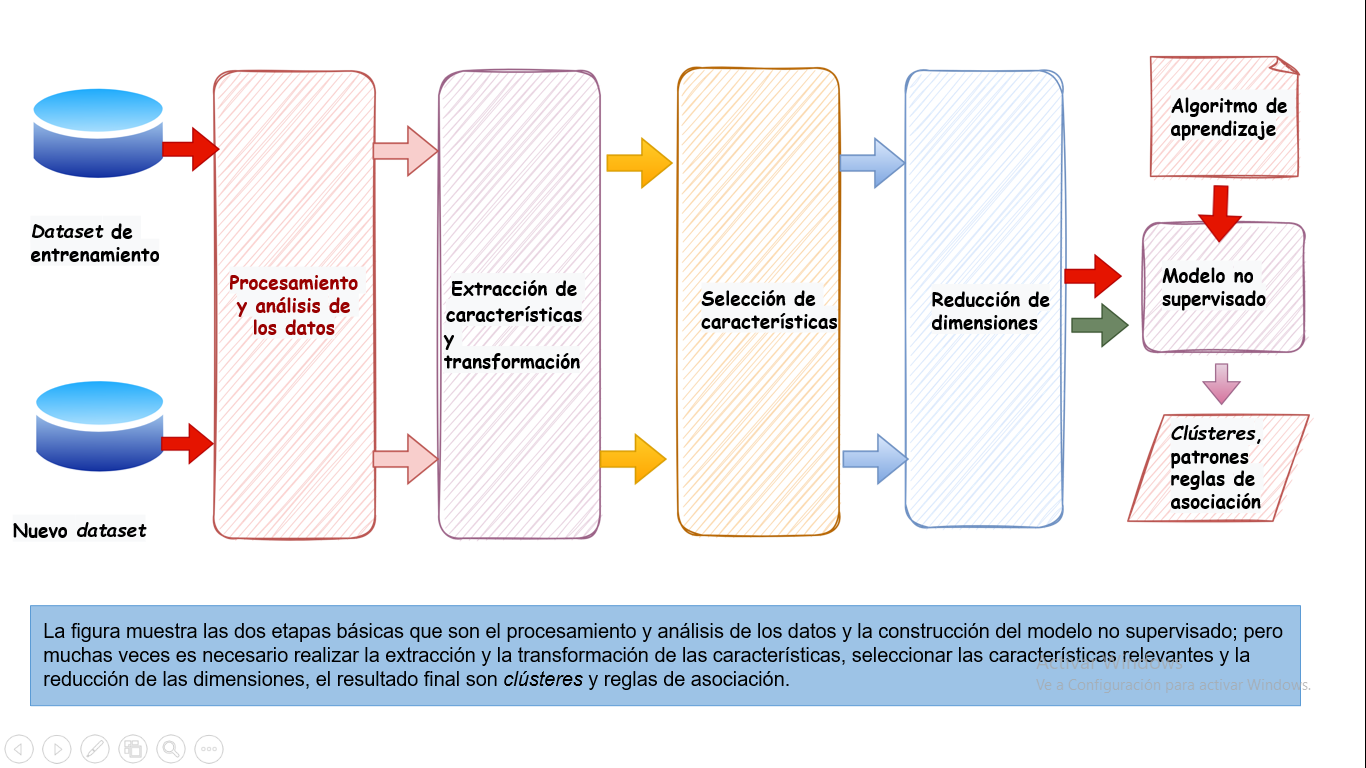
Tienen la misión de descubrir las similitudes, los patrones o uniformidades dentro de los datos de entrada, en este caso no se cuenta con un supervisor que etiquete los datos.

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*



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

|  |
| --- |
| DI\_CF01\_4\_Slide\_ClasificacionAlgoritmosNoSupervisado |

**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.

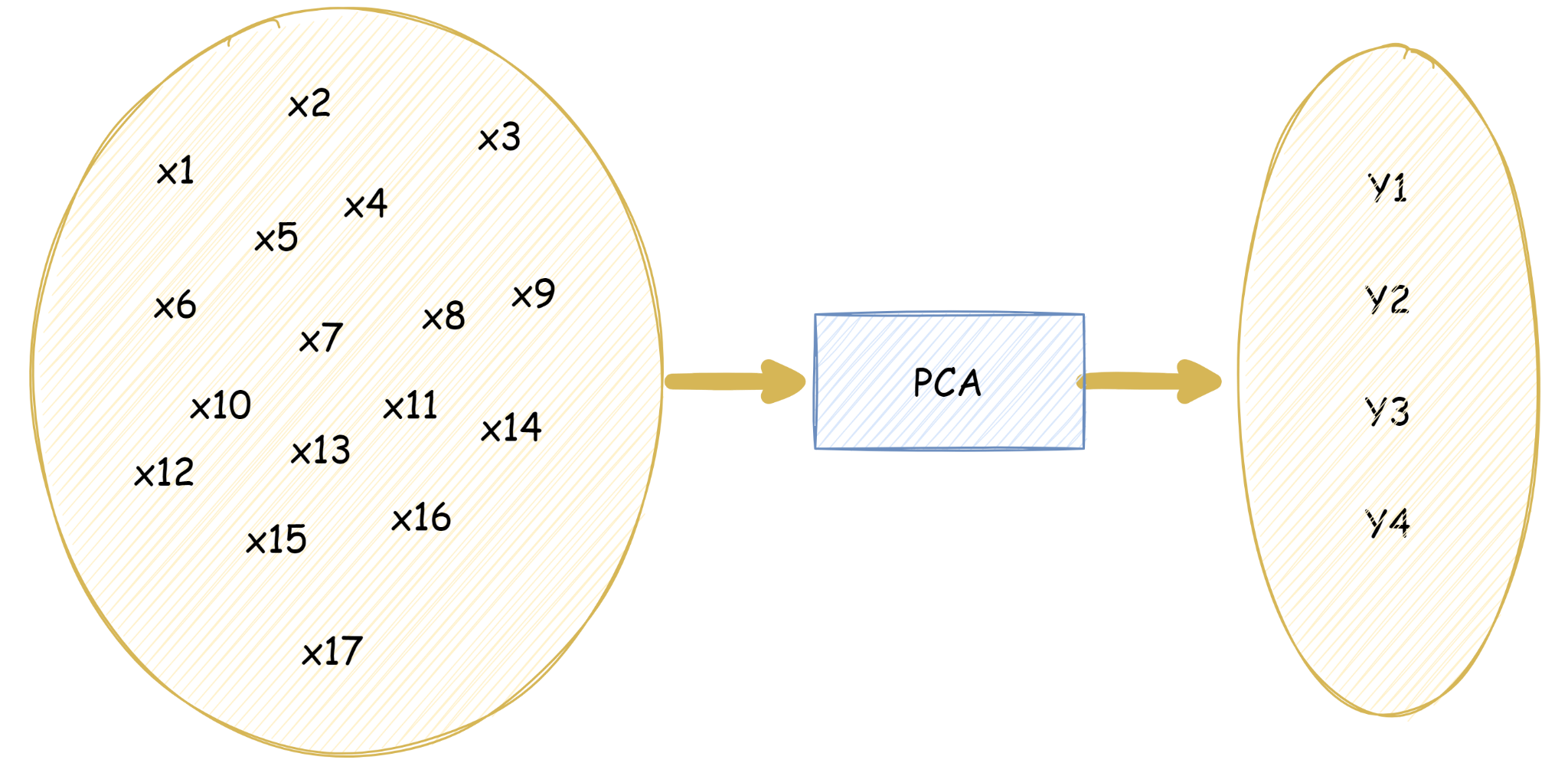
Sobre la reducción de dimensiones, tenga en cuenta:

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

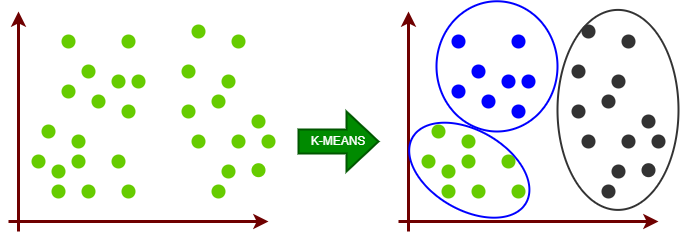


***4.2. Clustering***

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*



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.

1. **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:

**Tabla 1**

*Representación matriz de confusión*

|  |  |  |  |
| --- | --- | --- | --- |
| **Resultado de la predicción** | | | |
| **Valor actual** |  | Positivo | Negativo |
| 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.**

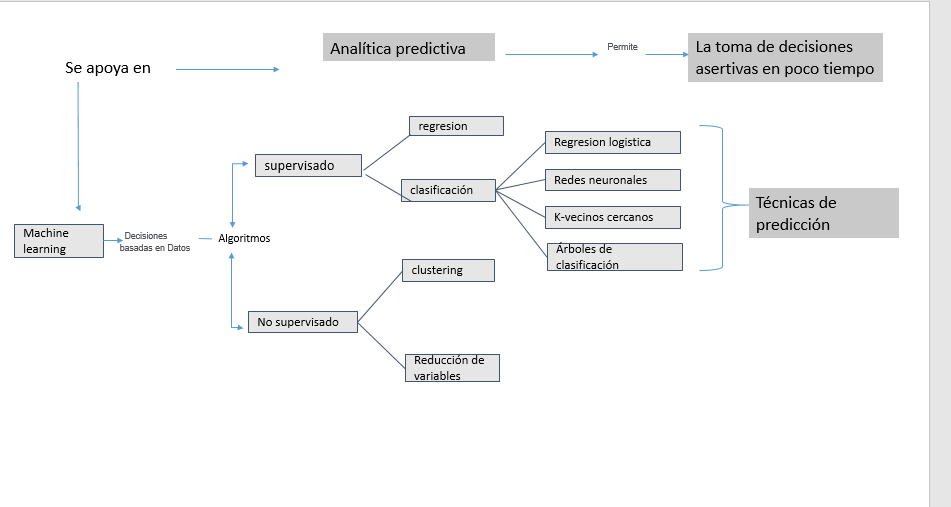
|  |  |
| --- | --- |
| **Error de TIPO I**  Este error equivale a los falsos positivos (FP), es el rechazo de una hipótesis nula, pero esta es verdadera. | **Error de TIPO II**   * Este error equivale a los falsos negativos (FN), consiste en aceptar una hipótesis falsa nula. |

Conozca cómo se pueden evaluar este tipo de errores:

|  |
| --- |
| **DI\_CF01\_5\_LineaDeTiempo\_ComoEvaluarTiposDeErrores** |

1. **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**!



# 

1. **ACTIVIDADES DIDÁCTICAS**

|  |  |
| --- | --- |
| DESCRIPCIÓN DE LA ACTIVIDAD DIDÁCTICA | |
| Nombre de la actividad | Introducción a la analítica predictiva. |
| Objetivo de la actividad | Reforzar conceptos, acciones y elementos clave para la aplicación del modelo de aprendizaje supervisado y no supervisado, en la elaboración de predicciones, a partir de necesidades organizacionales. |
| Tipo de actividad sugerida | Relación de términos. |
| Archivo de la actividad  (Anexo donde se describe la actividad propuesta) | En carpeta Formatos\_DI:  Actividad\_Didactica\_1 |

1. **MATERIAL COMPLEMENTARIO**

|  |  |  |  |
| --- | --- | --- | --- |
| Tema | Referencia APA del material | Tipo de material  (Video, capítulo de libro, artículo, otro) | Enlace del recurso o  archivo del documento o material |
| 3. Algoritmos de aprendizaje supervisado | SDC LEARNING. (2022). *Webinar gratuito Mi primer modelo de Machine Learning en Python* [video]. YouTube. <https://www.youtube.com/watch?v=9HKfqinJJAo> | 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. <https://www.youtube.com/watch?v=TkHO2HHbJDs> | Video | <https://www.youtube.com/watch?v=TkHO2HHbJDs> |
| 3. Algoritmos de aprendizaje supervisado | Parra, F. (2022). *Métodos de clasificación.* bookdown.org. <https://bookdown.org/content/2274/metodos-de-clasificacion.html> | Página web | <https://bookdown.org/content/2274/metodos-de-clasificacion.html> |
| 5. Métricas de evaluación | González, L. (2019). *Errores modelos clasificación*. Aprende IA. <https://aprendeia.com/evaluando-el-error-en-los-modelos-de-clasificacion-machine-learning/> | Página web | <https://aprendeia.com/evaluando-el-error-en-los-modelos-de-clasificacion-machine-learning/> |

1. **GLOSARIO**

|  |  |
| --- | --- |
| TÉRMINO | SIGNIFICADO |
| 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. |

1. **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/>

1. **CONTROL DEL DOCUMENTO**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Nombre | Cargo | Dependencia | Fecha |
| Autor (es) | Héctor Henry Jurado Soto | Experto temático | Regional Cauca – Centro de Teleinformática y Producción Industrial | Junio de 2022 |
| Jeimy Lorena Romero Perilla | Diseñadora instruccional | Regional Norte de Santander - Centro de la Industria, la Empresa y los Servicios | Junio de 2022 |
| Andrés Felipe Velandia Espitia | Asesor metodológico | Regional Distrito Capital - Centro de Diseño y Metrología | Junio de 2022 |
|  | Julia Isabel Roberto | Correctora de estilo | Regional Distrito Capital - Centro de Diseño y Metrología | Septiembre de 2022 |
|  | Fabián Leonardo Correa Díaz | Diseñador Instruccional | Regional Santander - Centro Industrial del Diseño y la Manufactura | Agosto 2023 |
|  | Rafael Neftalí Lizcano Reyes | Responsable Desarrollo Curricular, Ecosistema RED Santander | Regional Santander - Centro Industrial del Diseño y la Manufactura | Agosto 2023 |

1. **CONTROL DE CAMBIOS**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Nombre | Cargo | Dependencia | Fecha | Razón del Cambio |
| Autor (es) |  |  |  |  |  |