**FORMATO PARA EL DESARROLLO DE COMPONENTE FORMATIVO**

|  |  |
| --- | --- |
| Programa de formación | Algoritmo de agrupamiento no supervisado K-means con Python |

|  |  |  |  |
| --- | --- | --- | --- |
| Competencia | 220501114- Sistematizar datos masivos de acuerdo con métodos de analítica y herramientas tecnológicas. | Resultado de aprendizaje | 220501114-1  -Determinar las características y criterios de agrupamiento de datos según el objetivo del análisis. |

|  |  |
| --- | --- |
| Número del componente formativo | 01 |
| Nombre del componente formativo | Introducción al *Machine Learning* |
| Breve descripción | Este componente formativo aborda aspectos generales y claves sobre el *Machine Learning*, la forma en que las máquinas de aprendizaje son un campo de la inteligencia artificial que se enfoca, principalmente, en el uso de datos y algoritmos. Con el estudio responsable de estos contenidos, el aprendiz se afianzará en el manejo de máquinas y uso de algoritmos de aprendizaje supervisado, no supervisado y semisupervisado. |
| Palabras clave | Aprendizaje no supervisado, aprendizaje supervisado, *clustering*, inteligencia artificial, máquina de aprendizaje. |

|  |  |
| --- | --- |
| Área ocupacional | 2 - Ocupaciones en ciencias naturales, aplicadas y relacionadas |
| Idioma | Español |

**A. TABLA DE CONTENIDOS**

**1. *Machine Learning***

1.1. Conjuntos de datos

1.2. *Clustering*

**2. Aplicaciones de *clustering* de datos**

2.1. Modelos de clasificación

2.2. Algoritmos no supervisados

**B. INTRODUCCIÓN**

Le damos la bienvenida al componente formativo denominado “**Introducción al *Machine Learning***”, el cual hace parte del programa de formación complementaria “Algoritmo de agrupamiento no supervisado K-means con Python”.

Para comenzar, preste atención al video que se muestra enseguida:

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

**C. DESARROLLO DE CONTENIDOS**

# ***Machine Learning***

Las máquinas de aprendizaje son un campo de la inteligencia artificial que se enfoca, principalmente, en el uso de datos y algoritmos: con ellos, imitan la forma en la que las personas aprenden y, de manera progresiva, van mejorando su precisión.



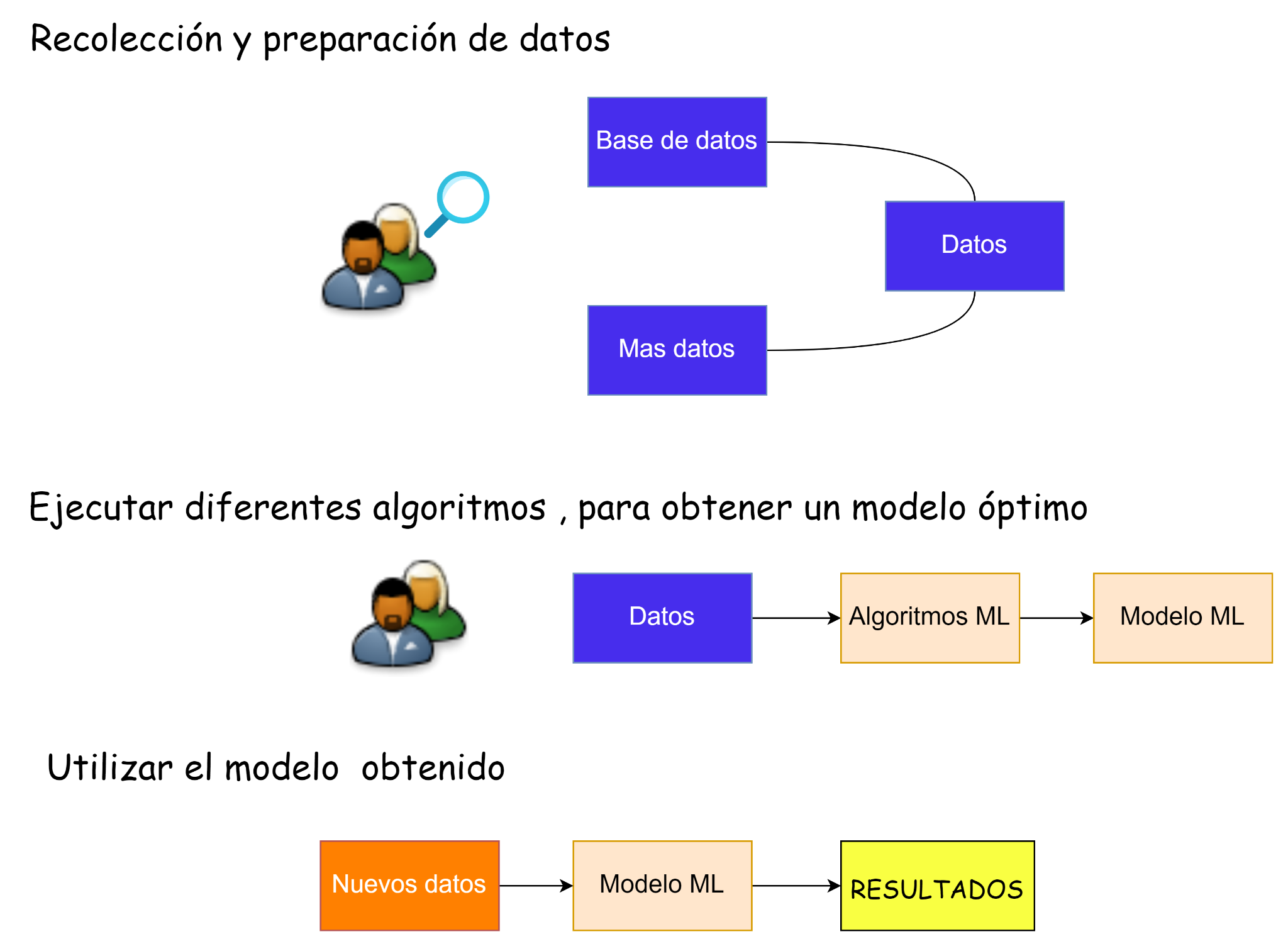
Sobre la implementación de sistemas de aprendizaje automático, es posible afirmar que:

|  |
| --- |
| **DI\_CF01\_1\_LineaTiempo\_SistemasAprendizajeAutomatico** |

El esquema que se muestra a continuación, expone el modelo *Machine Learning*:

**Figura 1**

*Modelo Machine Learning*



Las máquinas de aprendizaje se dividen en tres diferentes tipos de aprendizaje:

* Aprendizaje no supervisado
* Aprendizaje supervisado
* Aprendizaje semisupervisado

Algunas generalidades importantes de tener en cuenta, sobre el *machine learning*, son:

|  |
| --- |
| **DI\_CF01\_1\_Slide\_GeneralidadesDelMachineLearning** |

## 

## **1.1. Conjuntos de datos**

Los datos son el insumo más valioso cuando se habla de ciencia de datos y un término muy usado en la actualidad es el *Big Data* que, como su nombre lo indica, hace referencia a conjuntos de datos muy grandes.



Se han definido tres cualidades para *Big Data*: volumen, velocidad y variedad. Eentonces es posible ver que este implica grandes conjuntos de datos (volumen), de diversos tipos (variedad) y que se generan muy rápidamente (velocidad).

La siguiente infografía muestra la relación entre cada concepto, preste atención:

|  |
| --- |
| **DI\_CF01\_1-1\_Infografia\_RelacionCualidadesDelBigData** |

A continuación, se mencionan eventos sobre volumen, velocidad y variedad:

|  |
| --- |
| **DI\_CF01\_1-1\_Pestañas\_VolumenVelocidadVariedad** |

Habiendo hablado de ciencia de datos, se debe conectar con el concepto de datos en informática, los cuales son representaciones simbólicas (vale decir: numéricas, alfabéticas, algorítmicas, etc.) de un determinado atributo o variable cualitativa o cuantitativa, o sea: la descripción codificada de un hecho empírico, un suceso, una entidad.

En general, existe un conjunto de valores de variables cualitativas o cuantitativas, los cuales se pueden detallar a continuación:

|  |
| --- |
| **DI\_CF01\_1-1\_Acordeon\_VariablesCuantitativasCualitativas** |

En relación con la Big Data, se destacan algunos aspectos como:

|  |
| --- |
| **DI\_CF01\_1-1\_Tarjetas\_LaBigData** |

A través de las siguientes figuras (figuras 2 y 3), conozca un ejemplo de identificación de tipos de datos en el P*ython.* Se trata de un *Dataset* cuyos datos contienen cuatro características (longitud y anchura de sépalos y pétalos) de 50 muestras de tres especies de iris (iris setosa, iris virginica e iris versicolor).

**Figura 2**

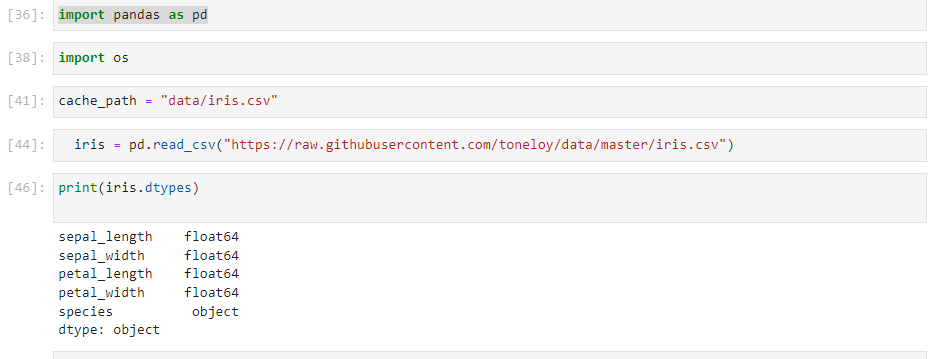
*Extracto de datos de dataset*

|  |  |
| --- | --- |
| ... |  |

En Python*,* usando Jupyter Lab*,* se pueden ver algunos tipos de variables usando el siguiente código:

**Figura 3**

*Variables*

****

Se observan, después de imprimir los tipos de datos de iris, las variables sepal\_length (longitud del sépalo), sepal\_width (ancho de sépalo), petal\_length (longitud de pétalo), petal\_width (ancho de pétalo), que son de tipo float64 y, por tanto, se habla de variables cuantitativas.

La variable *Species* (especie) es de tipo *object*, que puede ser de cualquier tipo, por ejemplo, texto; en este caso, representa una variable cualitativa.

**1.2. *Clustering***

En el **aprendizaje supervisado** el objetivo es mapear una entrada con una salida, con valores correctos suministrados por un supervisor. Por su parte, en **el aprendizaje no supervisado** no existe ese supervisor y solo existen datos de entrada; aquí el objetivo es describir similitudes en los datos de entrada y así observar patrones que aparecen con más frecuencia que otros, entonces la idea es investigar qué es lo que sucede con estos datos y qué no.



En términos estadísticos, lo que se hace es estudiar la densidad de los datos, agrupándolos en clústeres; todos los patrones, asociaciones, relaciones; y los clústeres son extraídos de los mismos datos.

Existen varias técnicas para realizar *clustering* basadas en agrupamiento con *K-Means*. A continuación, se presenta un *clustering* jerárquico:

|  |
| --- |
| **DI\_CF01\_1-2\_LineaTiempo\_ClusteringJerarquico** |

# **Aplicaciones de *clustering* de datos**

La clusterización usando *K-means* tiene un espectro bastante amplio de aplicaciones muy útiles en todas las ramas de la ciencia, tecnología, salud, transporte, astronomía, industria, educación, etc.



**Clasificación de documentos**

Se pueden usar aplicaciones con *K-means* para clasificar documentos en múltiples categorías, basadas en temas del texto o documentos, dependiendo del contenido de las cadenas.

A continuación, se presenta la clasificación:

|  |
| --- |
| **DI\_CF01\_2\_Slide\_ClasificacionDocumentos** |

## **2.1. Modelos de clasificación**

A continuación, una breve introducción a **algoritmos de aprendizaje supervisado**. Por su importancia en la clasificación de datos se puede encontrar que en este tipo de aprendizaje los algoritmos trabajan con observaciones y contienen variables de entrada y variables de salida o etiquetas relacionadas con las variables de entrada.



Las técnicas de aprendizaje supervisado consisten en dos clases principales: **clasificación** y **regresión**, dependiendo del tipo de problema de aprendizaje automático a resolver.

**Figura 4**

*Aprendizaje supervisado*

Diagrama

Descripción generada automáticamente

**El modelo entrenado**: puede ser utilizado para predecir salidas de cualquier conjunto nuevo de datos de entrada. Estas técnicas se definen como supervisados puesto que el modelo aprende de muestras de datos y sus salidas en la fase de entrenamiento.

A continuación, se muestran más generalidades y aspectos destacados sobre las técnicas del aprendizaje supervisado:

|  |
| --- |
| **DI\_CF01\_2-1\_Pestañas\_TecnicasAprendizajeSupervisado** |

## **2.2. Algoritmos no supervisados**

**El aprendizaje no supervisado** tiene la misión de descubrir similitudes, patrones o uniformidades dentro de los datos de entrada; en este caso, no existe un supervisor que etiquete los datos. Los algoritmos de este tipo de aprendizaje forman clústeres en forma autónoma y asignan observaciones a estos clústeres.



Por ejemplo, si los datos son miles de fotos de leones y tigres, en el aprendizaje no supervisado el programa ordena las fotos de los leones en un clúster y la de los tigres en otro. Este algoritmo toma decisiones de ordenamiento de forma independiente, de acuerdo con características similares y características diferentes.

Un algoritmo no supervisado aprende de un modelo de clústeres a partir de los datos de entrenamiento que se pueden utilizar más tarde para asignar nuevos datos a los clústeres.

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

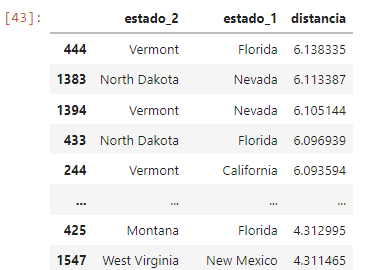
|  |
| --- |
| **DI\_CF01\_2-2\_Pestañas\_ProblemasAsociacionAgrupacion** |

Por medio de un ejemplo en Python y distancia euclídea, se identificarán los estados más similares en Estados Unidos, en cuanto a porcentaje de asaltos, asesinatos y secuestros por cada 100.000 habitantes, para cada uno de los 50 estados, usando el código fuente en Python. Este código fuente se puede ejecutar usando *Jupyter lab*

Los resultados son los siguientes, ordenados en forma descendente y ascendente:

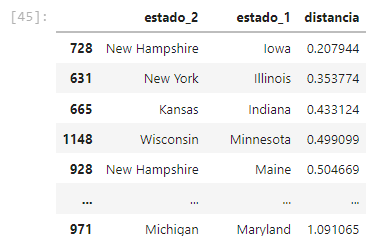
**Figura 5**

*Resultado 1 de ejemplo*



**Figura 6**

*Resultado 2 de ejemplo*

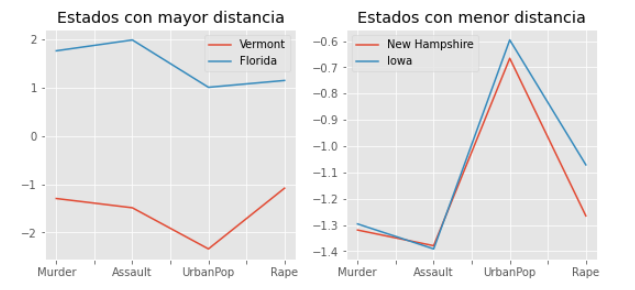


Los estados con mayor distancia y, por tanto, completamente diferentes en porcentaje de asaltos, asesinatos y secuestros son Vermont y Florida, que presentan una distancia euclidiana calculada de 6.14, estos estarán en clústeres separados.

Los estados con menor distancia y, por tanto, muy similares en porcentaje de asaltos, asesinatos y secuestros son New Hampshire y Iowa, que presentan una distancia euclidiana calculada de 0.21; estos podrían estar en el mismo clúster de porcentaje de asaltos, asesinatos y secuestros.

**Figura 8**

*Distancia euclídea*

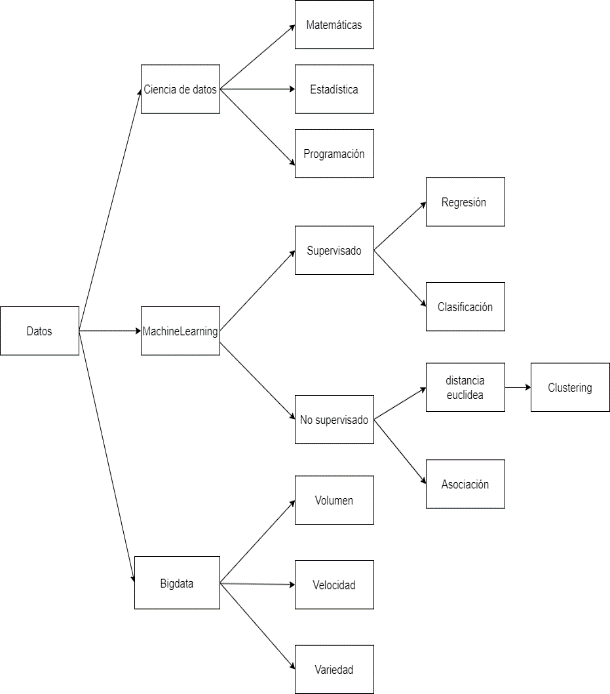


La industria 4.0 va tomando gran importancia en estos últimos años; uno de sus objetivos es que cualquier proceso de producción, sea cual sea, esté completamente automatizado, evitando en lo posible que las personas trabajen en forma manual y dejando que las máquinas o equipos trabajen por sí solos, traduciendo esto en reducción de costos, seguridad, eficiencia y productividad en los resultados.

Pero para llevar a cabo todo esto se requieren datos y estos datos no trabajan solos, existe el aprendizaje automático, que aprende de los datos que recibe constantemente, realiza predicciones, identifica patrones y es capaz de obtener conocimiento y conclusiones.

**D. SÍNTESIS**

Aquí finaliza el estudio de los temas de este componente formativo. En este punto, analice el esquema que se muestra enseguida y haga su propia síntesis de los temas vistos. ¡**Adelante**!



# 

**E. ACTIVIDAD DIDÁCTICA**

|  |  |
| --- | --- |
| Descripción de actividad didáctica | |
| Nombre de la Actividad | Máquinas de aprendizaje |
| Objetivo de la actividad | Reforzar los conceptos y elementos para la gestión de máquinas de aprendizaje y uso de algoritmos, con base en los temas desarrollados en el componente formativo. |
| Tipo de actividad sugerida | Falso-Verdadero |
| Archivo de la actividad  (Anexo donde se describe la actividad propuesta) | Actividad\_Didactica\_1 |

**F. 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 |
| 1.1 Conjunto de Datos | Anaconda. (s. f.). *Installing on Windows*. Anaconda Documentation. | Documento web | <https://docs.anaconda.com/anaconda/install/windows/> |
| 2. *Clustering*. | Singh, S. (2017). *Iris.csv*. Kaggle. | Dataset | <https://www.kaggle.com/datasets/saurabh00007/iriscsv> |

# **G. 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 supervisado: | cuando el algoritmo recibe datos de entrenamiento consistentes en datos etiquetados. |
| Aprendizaje automático no supervisado: | cuando el algoritmo identifica patrones y saca conclusiones de los datos que se le proporcionan. |
| Clúster: | conjunto de objetos o registros que son similares entre sí. |
| *Clustering:* | proceso de dividir un conjunto de objetos o registros en subconjuntos llamados clústeres, que tienen similaridades. |
| Distancia euclídea: | es la longitud del segmento entre dos puntos; en el caso del *clustering*, define las observaciones más cercanas para asignarlas a un clúster. |
| Inteligencia artificial: | sistemas informáticos que pueden aprender como aprende un ser humano. |
| Machine Learning: | aprendizaje automático o máquinas de aprendizaje. |
| *K-means*: | lenguaje de alto nivel usado para construir todo tipo de aplicaciones y muy usado en ciencia de datos. |
| Python: | proceso criptográfico que proporciona comunicaciones seguras a través de las redes, haciendo que la información entre extremos se transporte en forma segura, mediante uso de criptografía. |

# **H. Referencias bibliográficas**

Abdulhamit, S. (2020). *Data analysis using python*. Academic Press.

Anaconda. (s. f.). *Installing on Windows*. Anaconda Documentation. <https://docs.anaconda.com/anaconda/install/windows/>

Akram. (2018). *Mall-customers*. Kaggle. <https://www.kaggle.com/akram24/mall-customers>

Rusell, R. (2018). *Machine Learning. Step-by-Step Guide to implement machine learning algorithms with python*.

Severance, C. (2020). *Python para todos: explorando la información con Python 3*.

Singh, S. (2017). *Iris.csv*. Kaggle. <https://www.kaggle.com/datasets/saurabh00007/iriscsv>

1. **CONTROL DEL DOCUMENTO**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Nombre | Cargo | Dependencia | Fecha |
| Autor(es) | Héctor Henry Jurado Soto | Experto Temático | Regional Cauca – Centro de Teleinformática y Producción Industrial | Abril de 2022 |
| Caterine Bedoya Mejía | Diseñadora Instruccional | Regional Distrito Capital – Centro de Gestión Industrial | Abril de 2022 |
| Carolina Coca Salazar | Metodóloga | Regional Distrito Capital – Centro de Diseño y Metrología | Abril de 2022 |
| Darío González | Corrector de Estilo | Regional Distrito Capital – Centro de Diseño y Metrología | Mayo de 2022 |
| Fabián Leonardo Correa Díaz | Diseñador Instruccional | Regional Santander – Centro Industrial del Diseño y la Manufactura | Julio de 2023 |
| Rafael Neftalí Lizcano Reyes | Responsable Equipo de Desarrollo Curricular | Regional Santander - Centro Industrial del Diseño y la Manufactura | Julio de 2022 |

**J. CONTROL DE CAMBIOS**

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