Curso Actualízate – Machine Learning (Módulo 5)

**Nombre:** SOLUCIÓN

**Apellidos:**

**Fecha:**

Responde a las siguientes preguntas. Justifica la respuesta.

1. **¿Qué es un ETL?**

ETL es un proceso, que proviene del anglicismo y acrónimo ***“extract, transform and load (extracción, transformación y carga)”***, utilizado en el campo de la informática con la idea de integrar y consolidar datos para un único sistema de destino.

En este proceso, se extraen datos de una o varias fuentes, las cuales pueden estar en diferentes formatos y almacenados en distintos tipos de repositorios. El objetivo principal del proceso es asegurar que ***los datos sean precisos, completos y coherentes entre sí antes de ser utilizados*** y tratados en análisis o replicados para el uso de aplicaciones.

***Una buena transformación de los datos***durante el proceso ETL ***debe incluir la conversión de los formatos, la eliminación de duplicados, la normalización, la aplicación de reglas y/o la agregación de nuevas variables.***

1. **Enumera las características de Python explicando cada una de ellas (menciona 4 al menos).**

Python es un***lenguaje de programación de alto nivel,*** que se ha convertido en una de las opciones más populares para desarrolladores de software en todo el mundo. Sus principales características son:

1. Es un ***lenguaje simple y claro***, esto hace que Python sea más fácil de entender y de usar para programadores a cualquier nivel.
2. Es un ***lenguaje expresivo y legible***, los desarrolladores pueden escribir código que se asemeja mucho al lenguaje natural, lo que hace que sea más fácil para otros programadores que no trabajan en este sistema entender lo que otros están haciendo.
3. Tiene una ***sintaxis elegante y concisa*,** lo que hace que su código sea fácil de leer y de escribir. Python ha sido diseñado para ser intuitivo y fácil de usar.
4. Tiene un ***tipado dinámico y fuerte***, lo que significa que los tipos de datos se comprueban en tiempo real de ejecución así como posibles errores a la hora de generar código.
5. Es ***multiparadigma***,soporta varios estilos de programación. Los programadores pueden usar Python para escribir código orientado a objetos, funcional o procedimental, lo que lo hace ideal para diferentes tipos de proyectos.
6. Es ***interpretado***, lo que significa que no es necesario compilar su código antes de ejecutarlo. Esto hace que el proceso de desarrollo sea más rápido, ya que los desarrolladores ven los resultados del código inmediatamente.
7. Es ***multiplataforma***, es compatible con diferentes sistemas operativos y plataformas, haciéndolo ideal para desarrolladores que trabajan en diferentes entornos sin necesidad de modificar el código.
8. **¿Cuáles son los tipos primitivos en Python y qué valores pueden contener cada uno de ellos? Pista: Son 3 tipos.**

En Python, existen tres tipos primitivos de datos:

* ***Numéricos,*** este tipo de dato representa números enteros sin decimales. Pueden ser positivos o negativos y no tienen límites de tamaño en Python***.*** Dentro de ellos contamos con,
  + ***enteros (int)sin decimales (1, 10, -15).*** Este tipo de dato representa números enteros sin decimales. Pueden ser positivos o negativos y no tienen límites de tamaño en python,
  + ***y flotantes (float) con decimales, (3.14, 2.718, -0.5).*** Este tipo de dato representa números con decimales. Los flotantes en python se escriben con un punto decimal y pueden ser positivos o negativos.
* ***Booleanos (bool),*** se representan con los valores, ***verdadero (True) y falso (False).*** Los booleanos son útiles en la programación para tomar decisiones y controlar el flujo del programa.
* ***String (str),*** este tipo de dato representa cadenas de texto. Las cadenas de texto en Python se escriben entre comillas simples o dobles, y pueden contener letras, números, signos de puntuación y espacios. Ejemplos de cadenas de texto son ***"hola", "Python me mola", "123"***.

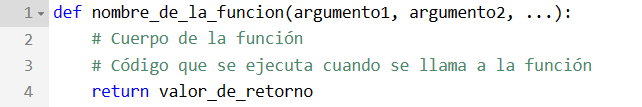
1. **Menciona alguna estructura de datos más compleja que los tipos primitivos que conozcas.**

Una de las estructuras de datos más compleja en Python es el ***diccionario (dictionary), que se utiliza para almacenar una colección de pares clave-valor***. Un diccionario en Python es una estructura de datos no ordenada y mutable que se compone de una serie de elementos, cada uno de los cuales tiene una clave única y un valor asociado.

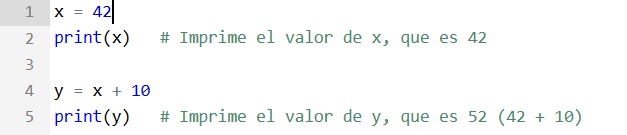
1. **Escribe la sintaxis para crear variables.**



1. **Escribe la sintaxis para crear funciones.**



1. **Escribe la sintaxis para llamar a variables.**



1. **Escribe la sintaxis para llamar a funciones.**



1. **Explica con tus palabras para qué sirven las librerías: Pandas y Numpy.**

***Numpy es una librería que proporciona soporte para cálculos científicos y matemáticos en Python***. Es muy utilizada para la manipulación de matrices y arreglos multidimensionales. Con Numpy, se pueden realizar operaciones numéricas eficientes y complejas, como la multiplicación de matrices, operaciones estadísticas y algebraicas, y otras operaciones matemáticas.

Por otro lado, ***Pandas es una librería de análisis de datos que proporciona estructuras de datos de alta eficiencia, como dataframes y series*** ***par*a manipular y analizar datos.** Pandas es muy útil para explorar, analizar y transformar datos, y se puede utilizar para tareas como la limpieza de datos, el filtrado, la transformación y la agregación de datos. Además, Pandas tiene una serie de herramientas para la visualización de datos y para el manejo de datos faltantes. Pandas también permite integrar los datos de múltiples fuentes de datos, como bases de datos, archivos CSV, archivos Excel, etc.

Resumiendo, ***Numpy es una librería para manipulación de matrices y cálculos numéricos***, mientras que ***Pandas es una librería para análisis de datos y manipulación de datos tabulares***.

1. **¿Cómo representamos el valor vacío en Python?**

El valor vacío se representa mediante la palabra clave None. None se utiliza para indicar que una variable o expresión no tiene un valor válido o que una función no devuelve ningún valor. Ej, ***x = None.***

1. **¿Qué es un IDE?**

Un IDE en Python ***es un programa que proporciona herramientas para facilitar el desarrollo de software***. En particular, el IDE de Python ofrece características como la edición de código, la depuración, la ejecución y la gestión de proyectos, todo en una sola interfaz gráfica de usuario.

1. **¿Qué es el CRISP-DM?**

Es una ***metodología para la minería de datos***, es decir, para el proceso de descubrir patrones y relaciones en grandes conjuntos de datos.

Es un modelo de proceso estándar para el desarrollo de proyectos de minería de datos y análisis de datos.

1. **Explica cada uno de los tipos de Machine Learning.**

***Aprendizaje supervisado:*** ***Conjunto de datos etiquetados***, es decir, que ***se tiene información de salida para cada ejemplo de entrada***. El objetivo es entrenar un modelo que pueda predecir correctamente las etiquetas para nuevos datos de entrada. Algunos ***ejemplos*** de aplicaciones de aprendizaje supervisado incluyen la ***clasificación de imágenes o el reconocimiento de voz***.

***Aprendizaje no supervisado:*** En este tipo de ML no se cuenta con datos etiquetados, el objetivo es ***encontrar patrones y relaciones en los datos sin información previa sobre las etiquetas.*** Algunos ejemplos de aplicaciones de aprendizaje no supervisado incluyen ***la segmentación de clientes, la agrupación de datos y la reducción de la dimensionalidad.***

***Aprendizaje reforzado:*** En este tipo de ML no se necesita datos, analiza comportamientos. Algunos ***ejemplos*** de aplicaciones de aprendizaje por refuerzo ***incluyen la robótica, los videojuegos y la optimización de sistemas de control***.

***Aprendizaje semi-supervisado:*** *Es un enfoque híbrido que combina elementos del aprendizaje supervisado y no supervisado. En este tipo de aprendizaje, el modelo se entrena con un conjunto de datos que contiene tanto ejemplos etiquetados como no etiquetados.*

*La idea es utilizar la información disponible en los datos no etiquetados para mejorar el rendimiento del modelo. Algunos* ***ejemplos serían el reconocimiento facial o la clasificación de documentos.***

**14. ¿Con qué tipo de Machine Learning hemos estado trabajando nosotros?.**

***Batch Learning***, aprendizaje supervisado.

**15. ¿Qué es el Prophet? ¿Y Scikit Learn, Keras y Tensorflow?.**

Prophet es una ***biblioteca de Python****,* que se utiliza para ***realizar análisis de series temporales*** y hacer pronósticos en el futuro.

Scikit-Learn es una ***biblioteca de aprendizaje que*** ***proporciona una amplia variedad de algoritmos de aprendizaje automático supervisado y no supervisado***, herramientas de preprocesamiento de datos y utilidades para evaluar modelos de aprendizaje automático.

Keras es una ***biblioteca de alto nivel*** ***que proporciona una interfaz para construir y entrenar modelos de aprendizaje profundo***. Keras se centra en la facilidad de uso, la modularidad y la extensibilidad, lo que la hace muy popular entre la comunidad de aprendizaje profundo. ***Tanto Prophet, Scikit-Learn como Keras son de código abierto***.

**16. ¿Qué significa en ML regresión? Responde brevemente.**

*Es un tipo de* ***tarea de aprendizaje supervisado******en la que se busca predecir un valor numérico continuo*** *a partir de una serie de variables de entrada.*

**17. ¿Cuál es el tipo de predicción más sencilla, pero a la vez la más usada?**

***Regresión lineal simple,*** busca modelar la relación entre una variable de entrada y una variable de salida numérica.

**18. ¿Cuáles son los problemas principales del machine Learning? Justifica tu respuesta.**

1. ***Datos de mala calidad:*** Si los datos de entrada contienen errores o inconsistencias, puede ser difícil para un modelo de aprendizaje automático aprender patrones significativos.
2. ***Variables irrelevantes:*** Si los datos de entrada contienen características irrelevantes o redundantes, puede ser difícil para un modelo de aprendizaje automático aprender patrones significativos y puede llevar a una disminución en el rendimiento del modelo.
3. ***Underfitting:*** Si el modelo tiene una capacidad limitada o hay muy pocos datos de entrenamiento, puede ser difícil para el modelo aprender patrones significativos y hacer predicciones precisas.
4. ***Overfitting:*** Si el modelo tiene demasiada capacidad o se entrena con demasiados datos, puede ser difícil para el modelo generalizar y hacer predicciones precisas en nuevos datos.

19. ¿Qué es el residuo en Machine Learning?

***El residuo se refiere a la diferencia entre el valor observado y el valor predecido por un modelo****.* Los residuos se utilizan comúnmente para evaluar el rendimiento de un modelo y ajustar sus parámetros para mejorar la precisión de las predicciones.