# Curso Actualízate – Machine Learning Gijón (Módulo 5)

Nombre: Vanesa

Apellidos: González Rodríguez

Fecha: 07/07/2023

Responde a las siguientes preguntas. Justifica la respuesta.

1. ¿Qué es un ETL?

Extract, Transform and Load. Es el proceso por el cual se extraen datos de múltiples fuentes, se formatean, se limpian y se cargan en una base de datos más grande (como si fuera un almacén de datos)

OK

1. Enumera las características de Python explicando cada una de ellas (menciona 4 al menos).
   * Es un lenguaje de muy alto nivel (es más fácil de leer y escribir ya que está más próximo al lenguaje natural que al código fuente). Muy expresivo y muy legible
   * Se ejecuta paso a paso con lo que es fácil atajar los errores y corregirlos
   * Es multiplataforma. Se puede utilizar para programación web, móvil, app además de para el análisis de datos
   * Rico juego de estructuras de datos: se pueden almacenar en listas, tuplas, diccionarios, conjuntos…

OK

1. ¿Cuáles son los tipos primitivos en Python y qué valores pueden contener cada uno de ellos? Pista: Son 3 tipos.
   * Cadena (string): almacena texto, números o símbolos. Entre comillas. Aunque almacenase números, al ser string, no podría operar con ellos.
   * Integer (int): almacena sólo números. Se puede operar con ellos. Dentro de los numéricos puede haber float (con decimales).
   * Booleano (bool): Devuelve verdadero o falso, si o no

OK, faltan los floats

● ***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.
   * Diccionario: Estructura de datos compleja. Es una colección no ordenada de datos estructurada en pares clave-valor donde puedes llegar al valor solicitándolo a través de la clave que es única para cada caso. Estas claves pueden ser de los tipos que admite python

OK

1. Escribe la sintaxis para crear variables.
   * nombre\_variable = valor\_de\_la\_variable
     + tienes que darle un nombre coherente e inicializarla (darle un valor inicial) para poder comenzar a trabajar con ella

OK

1. Escribe la sintaxis para crear funciones.
   * def nombre\_funcion (parametros):

cuerpo de la función

Donde el nombre de la función ha de ser coherente con lo que hace dicha función, los parámetros pueden ser uno o más o incluso ninguno (es una variable utilizada para los valores de entrada a la función) y el cuerpo de la función es lo que hace la función en si (tiene que tener un return al final para que nos de un valor de salida)

OK

1. Escribe la sintaxis para llamar a variables.
   * nombre\_variable = valor\_variable

por ejemplo el nombre de la variable puede ser “fruta” y el valor de dicha variable puede ser “pera”. Se recomienda que el nombre de la variable sea coherente con el valor.

OK

1. Escribe la sintaxis para llamar a funciones.
   * nombre\_funcion(parámetro)

por ejemplo en una función que te haga el cuadrado de un número, el parámetro sería el número del que queremos dicho cuadrado

OK

1. Explica con tus palabras para qué sirven las librerías: Pandas y Numpy.
   * Pandas: Librería de Python para la manipulación y el análisis de datos
   * Numpy: Librería que permite definir arrays y matrices de dos o más dimensiones

***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.

1. ¿Cómo representamos el valor vacío en Python?
   * None

OK

1. ¿Qué es un IDE?
   * Es una app que ayuda a desarrollar software de forma sencilla

OK

1. ¿Qué es el CRISP-DM?
   * Estándar que sirve como modelo en el proceso de ciencias de datos

OK

1. Explica cada uno de los tipos de Machine Learning.
   * Supervisado: Le pasas en input y el output al algoritmo
   * No supervisado: No le pasas todo, sólo los datos de entrada y el modelo busca patrones y hace grupos
   * Semisupervisado: Primero aplica el no supervisado y luego lo revisa una persona. Puede usar incluso una votación
   * Aprendizaje reforzado: No usa datos. No hace predicciones. Se le establecen condiciones y reglas y empiezan a probar de forma iterativa. Un humano no lo supervisa y se suele utilizar el sistema premio-castigo para el aprendizaje (devolverle 1 o 0)

OK

***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.*

1. ¿Con qué tipo de Machine Learning hemos estado trabajando nosotros?
   * Con el tipo supervisado. Un 80% de los datos para entrenar el algoritmo y un 20% para testearlo. Y por batch learning.
2. ¿Qué es el Prophet? ¿Y Scikit Learn, Keras y Tensorflow?
   * Prophet: Biblioteca (de código abierto) desarrollada por Facebook para el pronóstico de series temporales
   * Scikit Learn: Biblioteca de aprendizaje automático para Python
   * Keras: Biblioteca de aprendizaje profundo que se ejecuta sobre Tensorflow (marco de aprendizaje automático desarrollado por Google)

OK

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***.

1. ¿Qué significa en ML regresión? Responde brevemente.
   * Es una herramienta que nos ayuda a hacer predicciones sobre valores numéricos continuos en función de otras variables. Nos permite encontrar patrones en los datos y utilizarlos para estimar valores futuros.

OK

1. ¿Cuál es el tipo de predicción más sencilla, pero a la vez la más usada?
   * La más utilizada es la regresión lineal simple. Se utiliza para predecir valores numéricos continuos en función de una variable o características independientes. Se busca encontrar una línea recta que se ajuste lo mejor posible a los datos disponibles.

OK

1. ¿Cuáles son los problemas principales del Machine Learning? Justifica tu respuesta.
   * Falta de datos correctamente clasificados con sus respuestas correctas, abundancia de datos nulos o no completados…
   * El desequilibrio en las clases. Puede haber sobremuestreo (aumento de los datos que tienes poca cantidad) o submuestreo (eliminación de datos de los que hay mucha cantidad)
   * Selección de características útiles para el modelo
   * Sobreajuste y subajuste. Cuando para entrenar el modelo se usa la totalidad de los datos o no se usan los suficientes, luego puede tener problemas a la hora de usar nuevos datos y tiene un índice de fallo bastante alto
   * Interpretabilidad: Hay modelos muy complejos del tipo de las redes neuronales y su manera de analizar, los procesos y las decisiones, son difíciles de interpretar y entender
   * Sesgo y discriminación: Si los datos de entrenamiento están sesgados, puede conducir a que los resultados no sean justos y equitativos

OK

1. ¿Qué es el residuo en Machine Learning?
   * Es la diferencia entre el valor real y el valor predicho por el modelo. Es una medida de cuánto se desvían las predicciones del modelo de los valores reales.

OK