# Curso Actualízate – Machine Learning Gijón (Módulo 5)

Nombre: María Cristina

Apellidos: Calderón Amor

Fecha: 07/07/2023

Responde a las siguientes preguntas. Justifica la respuesta.

1. **¿Qué es un ETL?**

Proceso que permite mover datos desde múltiples fuentes, reformatearlos y limpiarlos y cargarlos en otra base de datos para analizar o en otro sistema operacional para apoyar un proceso de negocio. El proceso es el siguiente: Extract, Transform y Load.

Extract: Obtención, colección e identificación del tipo de dato (estructurado o no).

Transform: Se transforman y unifican para darles estructura única.

Load: Operación de actualización o carga de datos transformados en el sistema de vistualizado.

OK

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

Multiplataforma: Puede ejecutarse en diferentes sistemas operativos como Windows, macOs, Linux y otros.

Multiparadigma: Admite diferentes estilos de programación. La programación orientada a objetos, estructurada y la programación funcional.

Interpretado: Significa que el código fuente se ejecuta directamente por el intérprete de python línea por línea, en lugar de ser compilado en un programa ejecutable antes de su ejecución.

Expresivo y legible: La sintaxis es clara y precisa por lo que la lectura y la escritura del código hace que sea más sencilla. La organización se potencia con la indentación puesto que te obliga a mantener una estructura reglada.

OK

1. **¿Cuáles son los tipos primitivos en Python y qué valores pueden contener cada uno de ellos? Pista: Son 3 tipos.**

Texto o cadenas de texto (string)

Numéricos (enteros)

Booleanos (True/False, 0/1, Si/No).

OK

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.**

Las estructuras de datos más complejas en Python son las listas, las tuplas, los diccionarios y los conjuntos.

OK

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

X = 9

print(X)

OK

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

def area\_poligono (x, y):

return x \* y/2

x = 6

y = 8

print(‘El area de este poligono’, x, ‘es, x \* y/2)

OK

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

variable = 30

print(variable)

OK

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

nombre\_funcion(argumento1, argumento2)

OK

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

La librería Pandas nos permite disponer de funciones de todo tipo y está orientada a la manipulación y análisis de todo tipo de datos.

La librería Numpy nos permite disponer de operaciones matemáticas y sirve para definir arrays y matrices.

OK

***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 si es un dato sin tipo ni valor.

OK

1. **¿Qué es un IDE?**

Es un software que nos permite trabajar y desarrollar proyectos con Python de manera más sencilla. Un ejemplo de IDE es Google Collaboratory.

OK

1. **¿Qué es el CRISP-DM?**

Representa las siglas Cross Industry Standard Process for Data Mining. Es el modelo para el proceso de minería de datos que consta de las siguientes pasos, cabe decir que, el movimiento entre los mismos es continuo y podemos volver atrás si se considera necesario para alcanzar un buen resultado:

Business understanding

Data understanding

Data preparation

Modeling

Evaluation

Deployment

OK

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

Hay cuatro tipos:

Aprendizaje supervisado: Entrenas al algoritmo con un conjunto de datos de los cuales sabemos los datos de entrada y de salida, por lo que, el control sobre los datos es muy alto.

Aprendizaje no supervisado: Busca patrones entre los datos (clustering) y divide en grupos pero no conoce qué datos hay en cada cluster. No tiene ningún tipo de etiqueta.

Aprendizaje semisupervisado: El algoritmo está entrenado con datos hasta cierto punto y después aprende ~~por sí mismo~~ bajo revisión. La mano humana está presente para corregirlo y de ello es un buen ejemplo Chat GPT ya que, se han capado posibles respuestas a preguntas capciosas y se revisan ciertas respuestas en base a quejas de los usuarios).

Aprendizaje reforzado: No utiliza datos para aprender. Marcas las condiciones y el sistema hace pruebas de manera iterativa y la mano humana premia o castiga en función del desempeño del modelo.

***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.***

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

Con el tipo de aprendizaje supervisado por batch learning.

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

Prophet es un algoritmo desarrollado por Facebook (Meta) con el objetivo de realizar mejores predicciones escaladas y muy ligadas a negocios, está enfocada en series de datos temporales.

Scikit Learn, Keras y Tensorflow son librerias.

En el caso de Scikit Learn es una buena librería para trabajar con datos etiquetados dentro del aprendizaje supervisado. Mientras que Keras y Tensorflow son librerías más ajustadas a un tipo de desarrollo concreto del trabajo, Tensoflow sería más utilizada en Machine Learning tradicional y Keras en Deep Learning.

OK

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

La regresión permite la relación entre una variable de entrada (los datos que tenemos de entrada y procesamos) con la variable de salida (qué resultado pretendemos obtener con los datos que contamos en un inicio).

OK

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

La regresión lineal simple.

OK

1. **¿Cuáles son los problemas principales del Machine Learning? Justifica tu respuesta.**

Los principales problemas son la mala calidad de los datos, no se puede trabajar con datos que no nos vayan a proporcionar los resultados que necesitamos e independientemente de lo bueno que sea el modelo que usamos, no saldrá adelante. Se distinguen dos problemas graves: Underfitting y Overfitting.

Underfitting: Los datos son insuficientes. Dada la situación, el margen de error se incrementa y puede intentarse resolver un problema muy grande con un algoritmo muy sencillo por lo tanto ese porcentaje de error crece.

Overfitting: Los datos que tienes para entrenar son buenos pero para testear no lo son tanto. Se suele dar cuando entrenas el modelo con el 100% de los datos y cuando introduces nuevos datos el margen de error crece exponencialmente ya que el modelo está acostumbrado a dar unos resultados.

OK

***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.

1. **¿Qué es el residuo en Machine Learning?**

Es la diferencia entre el valor real observado y el valor predicho por un modelo. Se observa cuando no se obtiene el resultado esperado y la salida del modelo es muy distinta.

OK