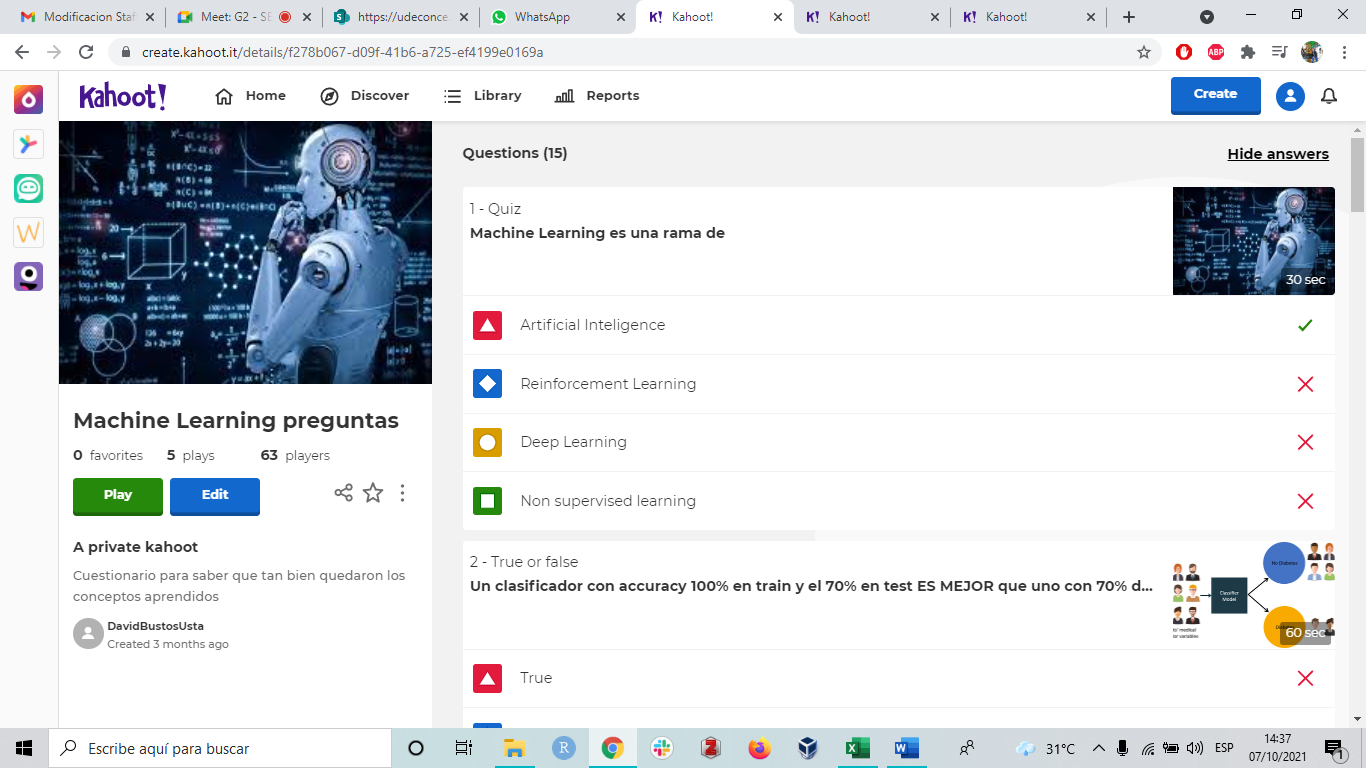
EVALUACION MACHINE LEARNING



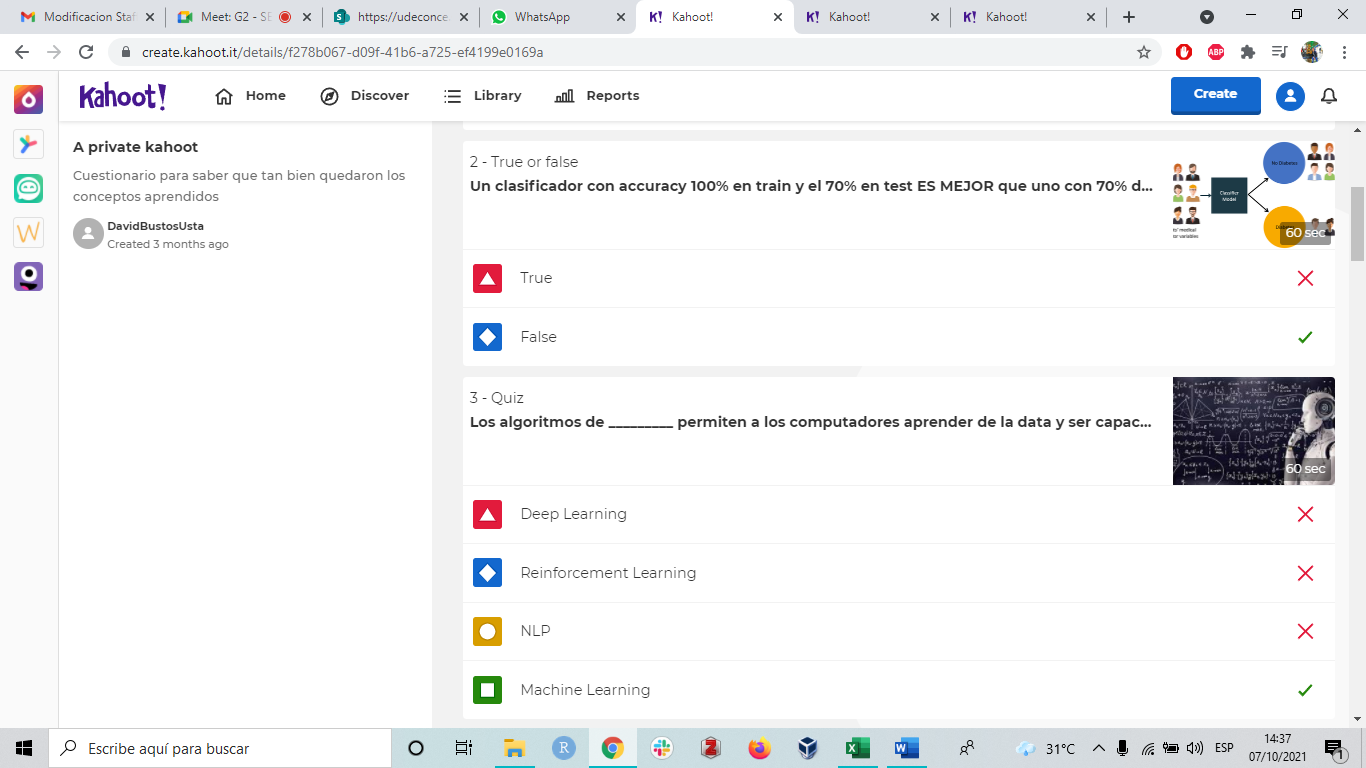
El '**machine learning**' –aprendizaje automático– es una **rama** de la inteligencia artificial que permite que las máquinas aprendan sin ser expresamente programadas para ello. Una habilidad indispensable para hacer sistemas capaces de identificar patrones entre los datos para hacer predicciones.

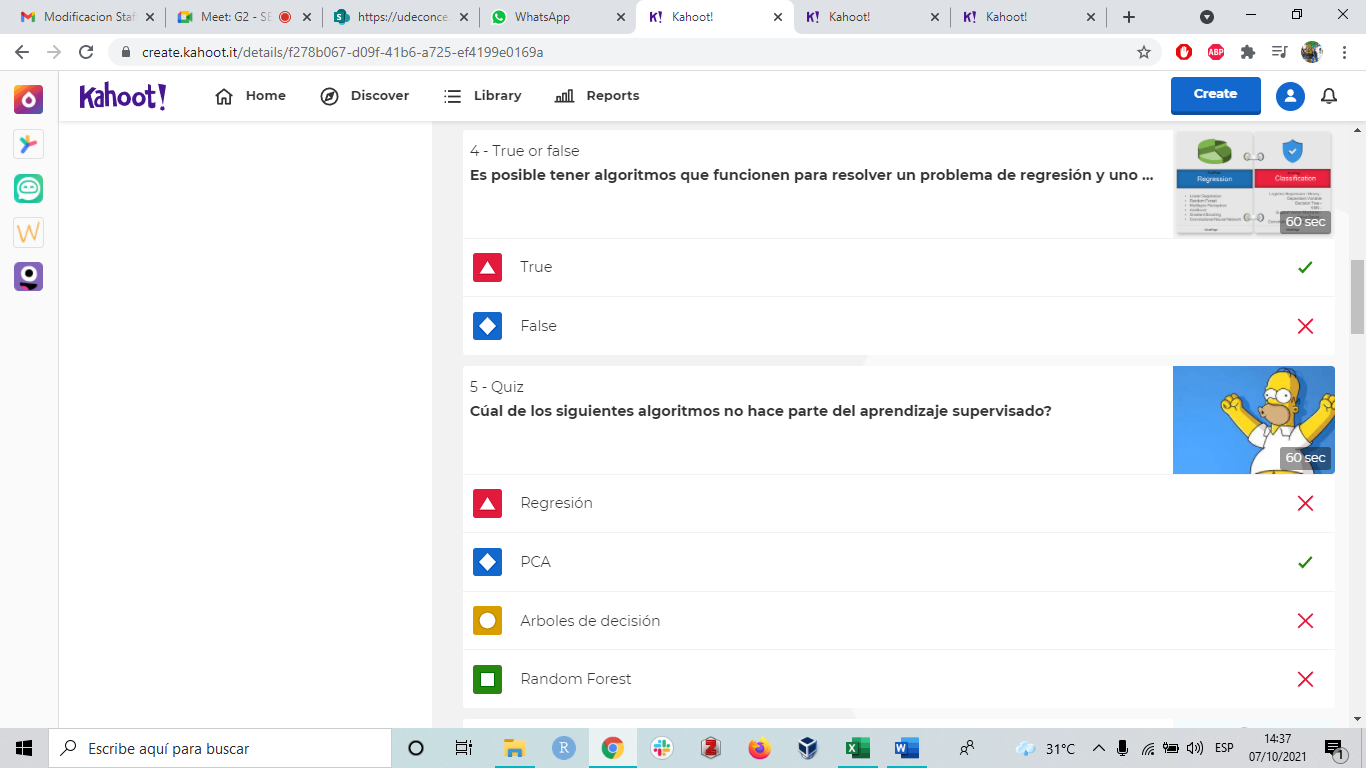
Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Generalmente en el mundo del Machine Learningse organizan los datos disponibles de manera que una parte sirve para entrenar y otra para hacer una prueba final (Test dataset) usándolos una sola vez como evaluación del modelo final. Pero de los datos que hemos destinado para entrenar reservamos una parte como datos de validación (validation dataset).

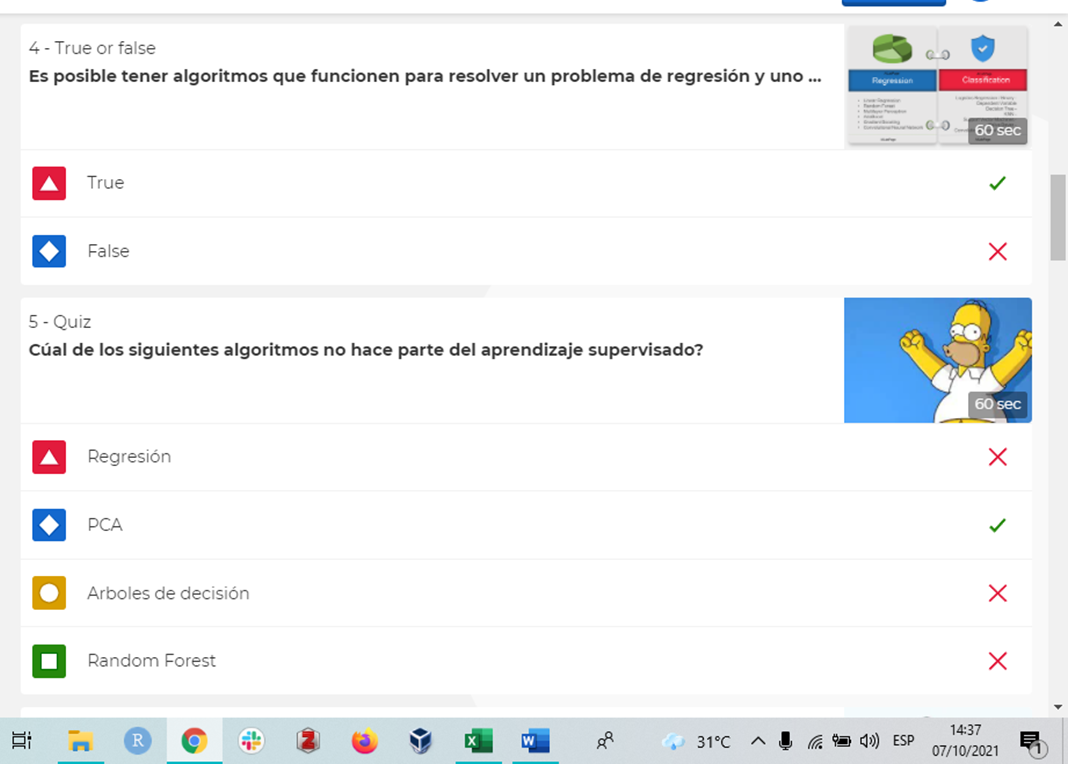
El conjunto de datos que nos quedan para el entrenamiento (Training dataset) son datos que se utilizan para entrenar diciéndole al modelo de red neuronal que “así es como se ve un gato” y “así es como se ve un perro”. En cambio el conjunto de datos de validación son imágenes de gatos y perros que la red neuronal no verá como parte del entrenamiento pero serán usadas para comprobar qué tan bien o qué tan mal se ha entrenado el modelo en cada epoch. Con las métricas, como la Accuracy, que se pueden obtener de este conjunto de datos de validación, nos guiamos para decidir como sintonizar los hiperparámetros del algoritmo antes de repetir el proceso de entrenamiento para otra epoch.





una de las cosas indispensables a la hora de aprender MACHINE LEARNING, es de entender la diferencia entre los algoritmos de clasificación y los de regresión.

La regresión tiene el objetivo de predecir valores continuos (Números pues, como el 1, 2.3, 3.1416 etc…), Y la clasificación tiene la tarea de asignar una clase, es decir predecir a que clase pertenece un conjunto de datos, aquí es muy importante entender que en los problemas de clasificación los valores son discretos .



El análisis de componentes principales (Principal Component Analysis PCA) es un método de reducción de dimensionalidad que permite simplificar la complejidad de espacios con múltiples dimensiones a la vez que conserva su información

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

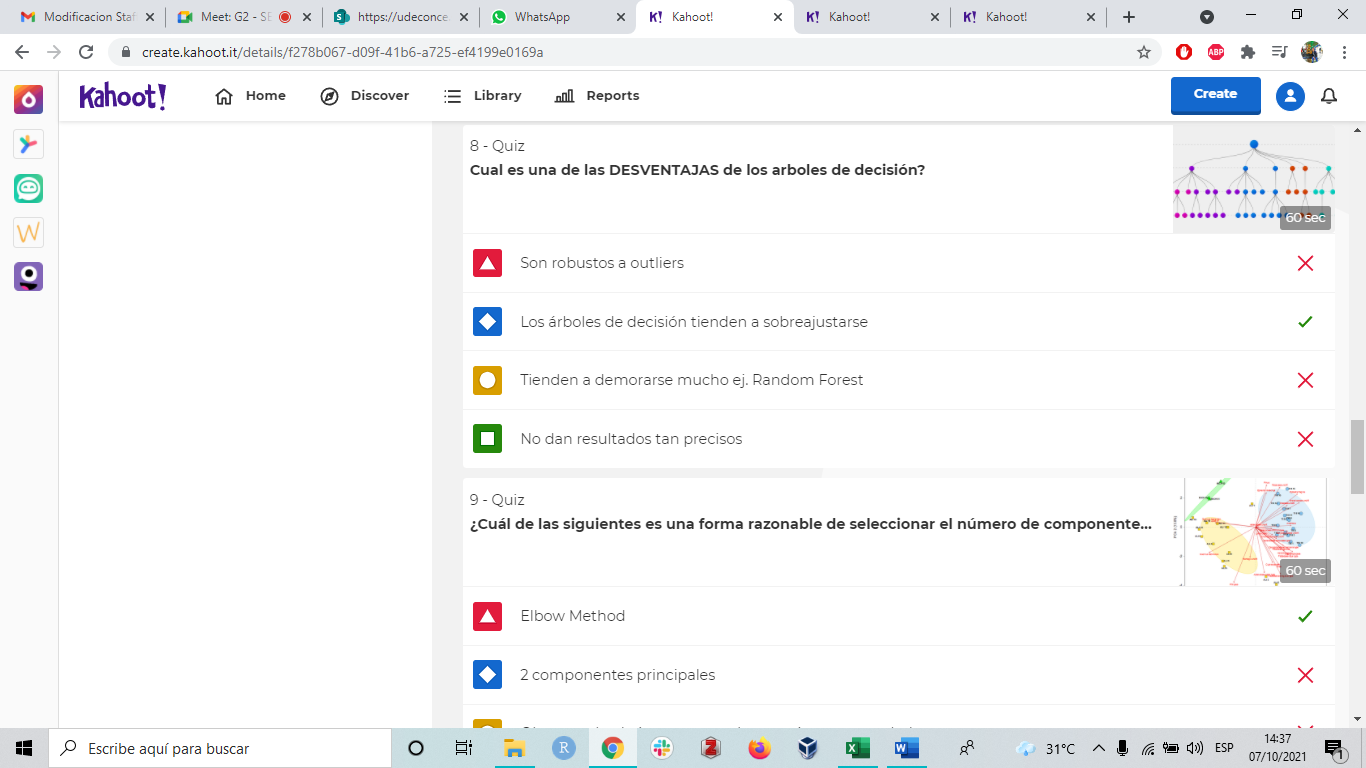
El **Clustering** o la clusterización es un proceso importante dentro del **Machine learning**. Este proceso desarrolla una acción fundamental que le permite a los algoritmos de **aprendizaje automatizado**entrenar y conocer de forma adecuada los datos con los que desarrollan sus actividades.

Este proceso ayuda a las máquinas a generar capacidades de análisis de forma rápida, en grandes volúmenes y con la menor cantidad de errores posibles. A continuación, conoceremos un poco más sobre este importante proceso de estudio y agrupamiento de datos para aprovecharlo al máximo en nuestros proyectos.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Primero debe normalizar los datos siempre. De lo contrario, PCA u otras técnicas que se utilizan para reducir las dimensiones darán resultados diferentes.



Un árbol de decisión es un diagrama en forma de árbol que muestra la probabilidad estadística o determina un curso de acción. Muestra a los analistas y, a los que toman las decisiones, qué pasos deben tomar y cómo las diferentes elecciones podrían afectar todo el proceso. Todo ello soportado en datos.

Por otro lado, hay una serie de desventajas, las cuales quizás no sean tan evidentes,

Son inestables: cualquier pequeño cambio en los datos de entrada puede suponer un árbol de decisión completamente diferente.

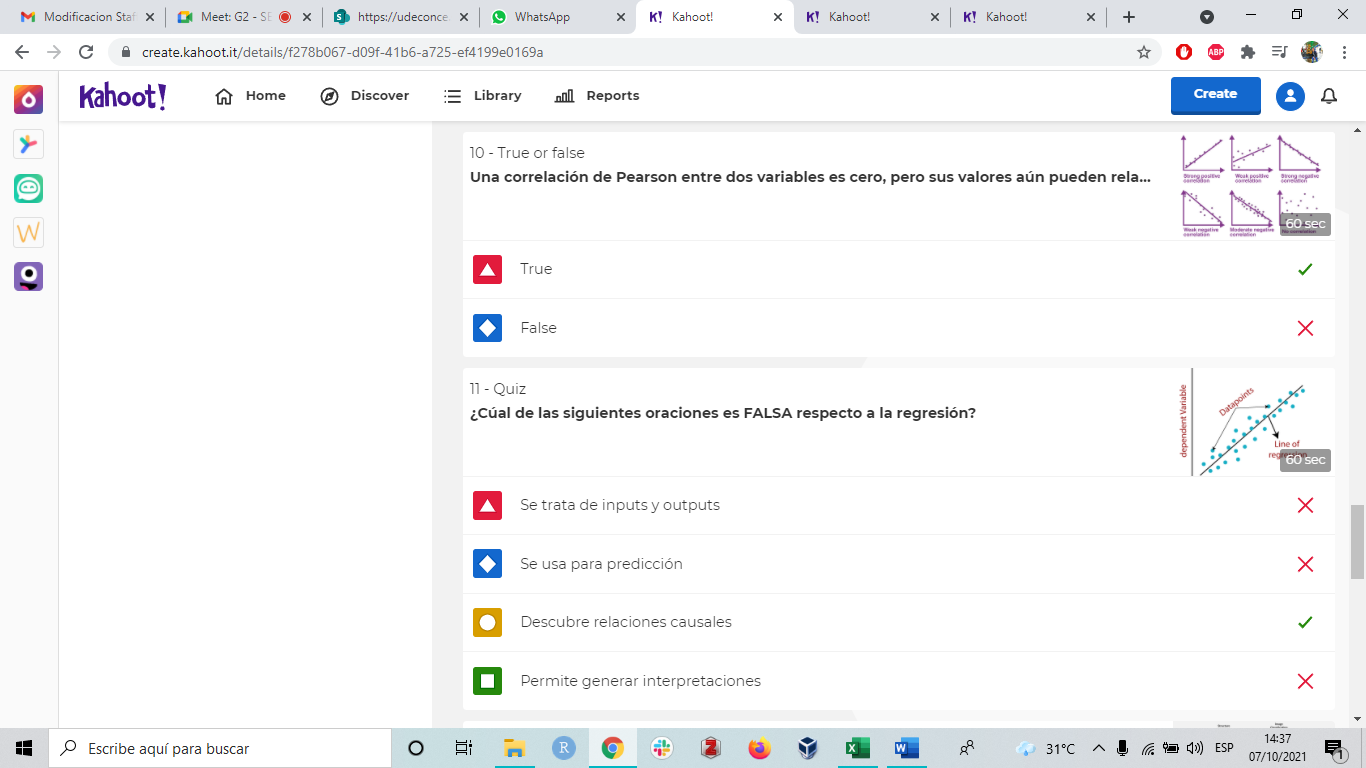
No se puede garantizar que el árbol generado sea el óptimo.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación

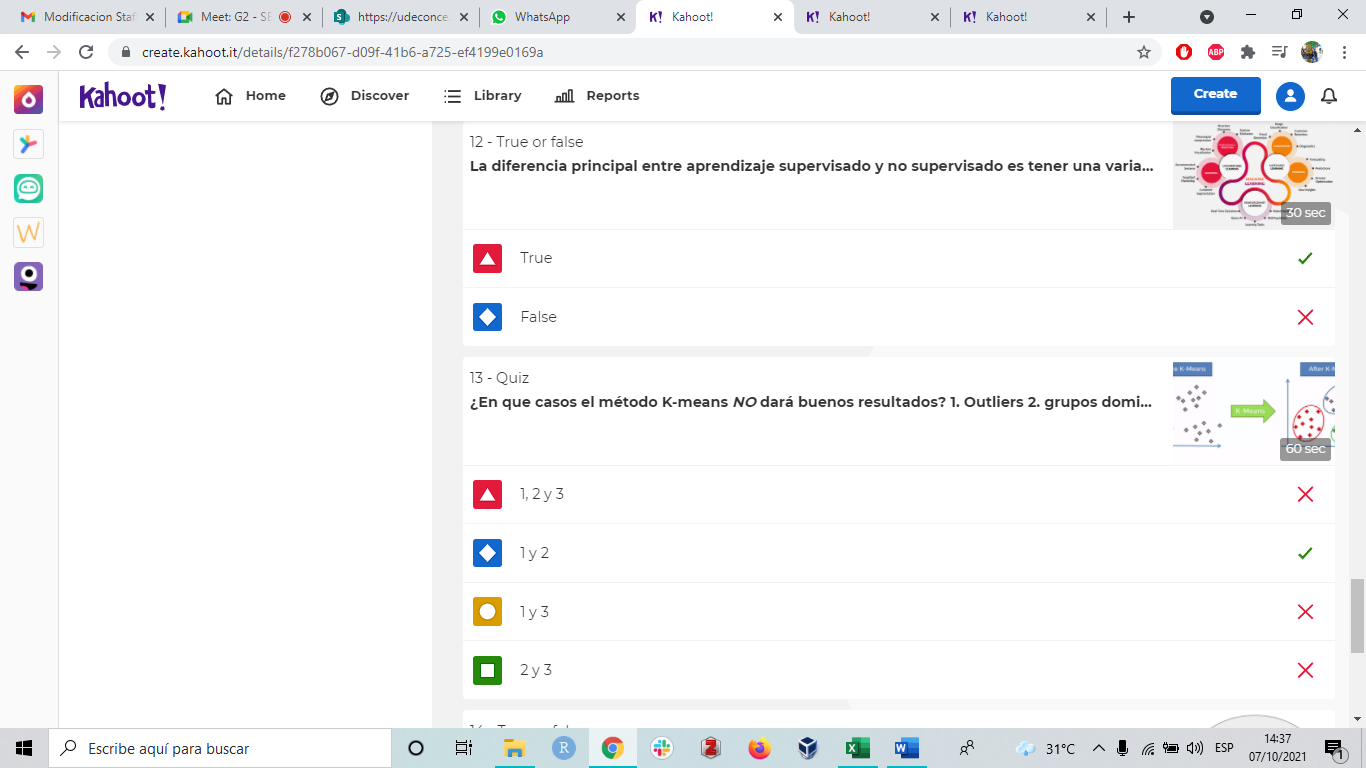
Descripción generada automáticamente

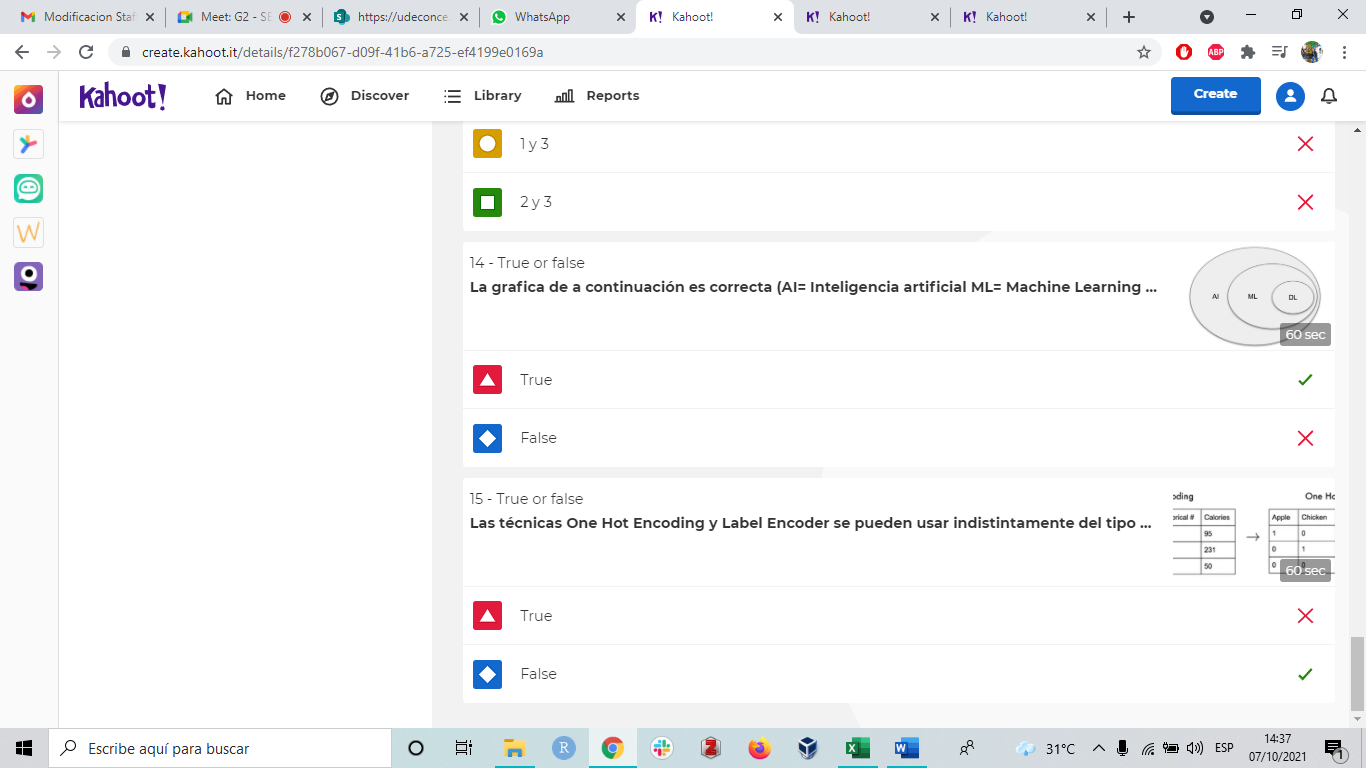


Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

el **aprendizaje supervisado** utiliza datos de entrada y salida etiquetados, mientras que un algoritmo de **aprendizaje** no **supervisado** no lo hace.





Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

El método de Label Encoding visto tiene la ventaja de ser sencillo de implementar. Sin embargo tiene un cierto problema: los valores numéricos pueden ser malinterpretados por algunos algoritmos: si hemos codificado varias ciudades con los valores 0, 1, 2 y 3 ¿significa que la ciudad correspondiente al valor 3 es el triple que la que ha recibido el valor 1 (según algún criterio)? La respuesta es no, por supuesto.

Una alternativa al Label Encoding es el método llamado One Hot Encoding. La estrategia que implementa es crear una columna para cada valor distinto que exista en la característica que estamos codificando y, para cada registro, marcar con un 1 la columna a la que pertenezca dicho registro y dejar las demás con 0.