Hoja informativa: Tareas de negocio que involucran al machine learning

Práctica

```
# importa la clase classifier
from sklearn.tree import DecisionTreeClassifier
# crea un objeto classifier
model = DecisionTreeClassifier()
# entrena el modelo
model.fit(X, y)
# obtén las predicciones
predictions = model.predict(X)
```

```
# divide el DataFrame en una variable objetivo y características de entrenamiento

y = data['target']
X = data.drop(['target'], axis = 1)
```

```
# Divide el conjunto en datos de entrenamiento y validación
# test_size: la proporción del dataset que se va a dividir para validación
# random_state: el parámetro para reproducir el resultado

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

Teoría

El **aprendizaje automático** es el proceso en el que un modelo busca interrelaciones entre valores basados en varios objetos.

Un **modelo** es un sistema de interrelaciones entre características y una variable objetivo, o entre observaciones que reflejan fielmente la realidad.

Una **variable objetivo** es una variable que puede ser **predicha** o **pronosticada** en función de unas **características** específicas.

Las **observaciones** son ejemplos existentes para los que los valores de característica y variable objetivo ya se conocen.

El **aprendizaje supervisado** consiste en establecer interrelaciones entre características y variables objetivo de acuerdo con los datos etiquetados ya existentes.

La **clasificación** es un caso de aprendizaje supervisado donde la variable objetivo toma un valor de un conjunto limitado de posibles valores.

La **regresión** es un tipo de aprendizaje supervisado donde la variable objetivo es un valor continuo.

El **aprendizaje no supervisado** consiste en establecer interrelaciones entre objetos con variables objetivo desconocidas.

Clustering es un ejemplo de aprendizaje no supervisado donde los objetos se dividen en grupos en función de las interrelaciones entre ellos.

La **reducción dimensional** es un caso de aprendizaje no supervisado donde los vectores de característica cambian a medida que el número de características disminuye.

Los datos de entrenamiento son los datos con los cuales se entrena el modelo.

Los datos de validación se utilizan para ajustar el modelo durante el entrenamiento.

Los datos de prueba se utilizan para probar el modelo después del entrenamiento.