En la mayoría de los modelos de Machine Learning, es recomendable partir de datos estandarizados o normalizados para obtener un mejor desempeño del modelo a entrenar. Por esta razón, resulta interesante verificar el efecto de estas transformaciones previas en el resultado final del modelo.

Para evaluar empíricamente este efecto, se propone utilizar la herramienta Rapidminer y generar un flujo que valide la precisión de un modelo en particular. Para esto, el flujo deberá verificar la métrica, a partir de un modelo entrenado con datos estandarizados (o normalizar) y otro con datos sin estandarizar (o normalizar).

El dataset a utilizar será Wine UCI y el modelo a entrenar será del tipo clasificador Naive Bayes.

A los efectos de entender mejor la problemática, resulta conveniente explicitar la diferencia entre los conceptos “normalización” y “estandarización”. Si bien ambas técnicas implican llevar los predictores a una escala común, su implementación es distinta. La normalización se centra en el intervalo objetivo de la transformación, escalando los valores según su proporción original (p.e.: El intervalo 0.0 a 1.1). Por otra parte, la estandarización implica llevar los valores de los atributos a un conjunto con media cero y desviación estándar uno.

Para comenzar se crea un nuevo flujo en Rapidminer y se agregan los operadores de Naive Bayes. Cada uno de los operadores se nutren con el modelo Wine UCI, con el detalle de que uno de ellos recibe los datos normalizados utilizando “Transformación Z” en la configuración del operador.

Diagrama

Descripción generada automáticamente

El modelo por evaluar no debería ser afectado por las escalas de los atributos, ya que la base del algoritmo es la probabilidad condicional y no implica un cálculo de distancia.  
  
  
  
  
Al ejecutar el flujo de Rapidminer se obtienen como salidas las siguientes matrices de confusión:

Modelo sin normalización

Interfaz de usuario gráfica

Descripción generada automáticamente

Modelo normalizado

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Como conclusión, se verificó para un caso particular, que la estandarización o normalización de los datos no impacta significativamente en la precisión de un modelo Naive Bayes. Esto condice con la teoría mencionada anteriormente sobre la base del modelo. Sin embargo, existen diversos modelos de Machine Learning de clasificación que requieren la normalización para funcionar correctamente, por lo que este experimento no es suficiente para descartar la técnica de normalización en la preparación de los datos.