UAX Universidad Alfonso X el Sabio

Ejercicio de Regresión con PySpark

PROGRAMACIÓN PARALELA Y DISTRIBUIDA

DAVID DE LA ANTONIA

Mediante las técnicas de programación paralela y distribuida que ofrece PySpark, debemos de determinar el índice de Rendimiento Académico (Performance Index) que tienen los estudiantes atendiendo a una serie de variables.

El dataset del que partimos tiene las siguientes variables independientes:

- **Hours Studied**: El número total de horas que ha estudiado cada estudiante.
- **Previous Scores**: Las notas que ha obtenido un estudiante en exámenes anteriores.
- **Extracurricular Activities**: Si el estudiante participa en actividades extraescolares (Yes or No).
- **Sleep Hours**: El número de horas que un estudiante duerme al día.
- **Sample Question Papers Practiced**: El número de hojas de ejercicios que el estudiante realizó.

Se trata de predecir el valor de la variable dependiente:

- **Performance Index**: La medida del rendimiento académico de cada estudiante. Los valores están comprendidos entre 10 y 100, indicando los valores más altos un mayor rendimiento académico.

Se valuará el modelo mediante Root Mean Squared Error.