Predict Concrete Strenght con Knime

Marcos Folguera Rivera

Asignatura: Adquisición y Computación del Big Data

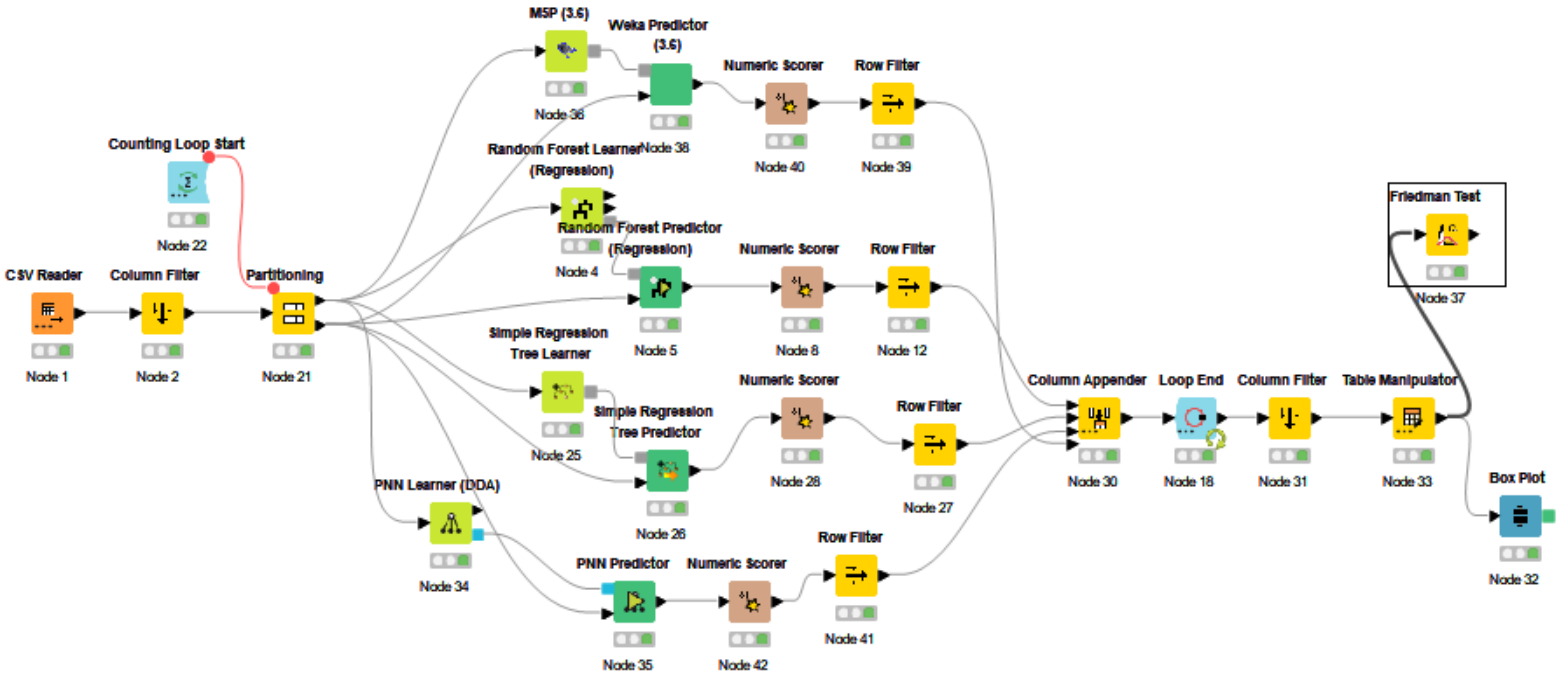
Profesor: Francisco Chávez de la O

Curso 2022/2023

Se nos pide predecir la variable Strenght del conjunto de datos ConcreteStrengthData.

Usando 4 predictores implementados en knime , comparar los predictores mediante análisis estadísticos y realizar la predicción usando un proceso una validación cruzada 5-fold cross valitation (en mi caso usando un partitioning de 80%).Para dividir el conjunto de datos entre entrenamiento y prueba y mandárselo a los Learners y Predictors respectivamente .

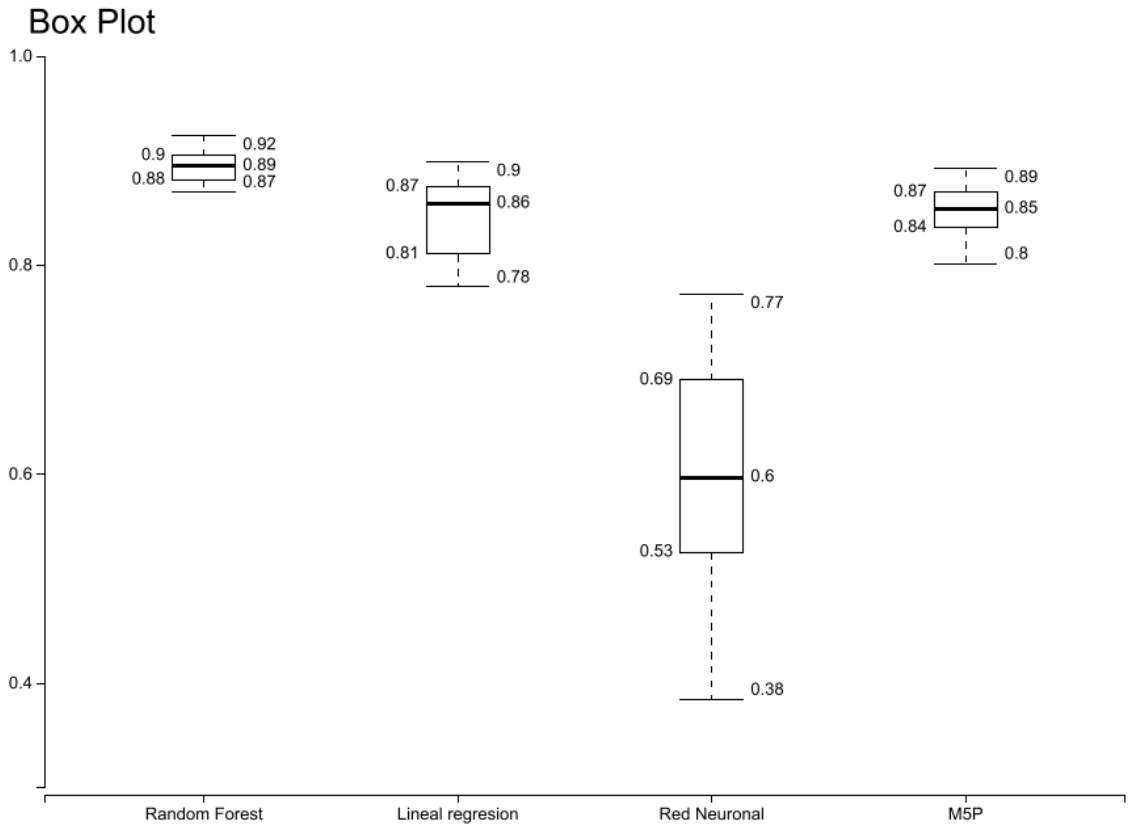
Creamos el siguiente modelo de knime.



Realice un estudio lanzando 20 veces cada tipo de regresión almacenando las R cuadrados.

Rando...	Lineal r...	Red Ne...	M5P
0.904	0.899	0.691	0.878
0.872	0.825	0.697	0.837
0.924	0.873	0.688	0.871
0.87	0.779	0.521	0.802
0.887	0.811	0.597	0.835
0.875	0.789	0.457	0.826
0.882	0.804	0.505	0.871
0.877	0.806	0.558	0.803
0.898	0.877	0.539	0.864
0.907	0.894	0.638	0.863
0.891	0.836	0.622	0.865
0.905	0.859	0.748	0.843
0.9	0.886	0.757	0.893
0.921	0.859	0.486	0.868
0.909	0.86	0.772	0.886
0.902	0.868	0.53	0.829
0.905	0.89	0.597	0.863
0.891	0.842	0.562	0.845
0.892	0.81	0.658	0.837
0.881	0.87	0.385	0.839

Véase en el siguiente diagrama de cajas y patillas que compara los 4 tipos distintos de algoritmos en función de las R^2 de las 20 iteraciones.



Se puede apreciar diferencias entre los modelos pero para saber si de verdad son diferencias significativas hay que aplicar el siguiente test.

Friedman test: es una técnica de prueba no paramétrica utilizada para comparar múltiples muestras relacionadas o medidas repetidas en una variable dependiente continua.

Row ID	B Reject H0	D Q	D Critical ...	D p-Value
Friedman (Random Forest, Lineal regresion, Red Neuronal, M5P)	true	54.06	7.815	1.0894840585251586E-11

Ya que el test de Friedman nos da un P-value es inferior a 0,05 rechazamos la hipótesis nula en favor de la hipótesis alternativa lo que significa que hay diferencias significativas entre los algoritmos. Y elegimos el Random Forest ya que no hay tanta diferencia entre sus valores (caja pequeña) no tiene valores atípicos y es el que mas se acerca al 1.

Conclusión

Elegimos el algoritmo Random Forest para predecir la variable Strenght del conjunto de datos ConcreteStrengthData ya que es con el que mejor resultados obtenemos.