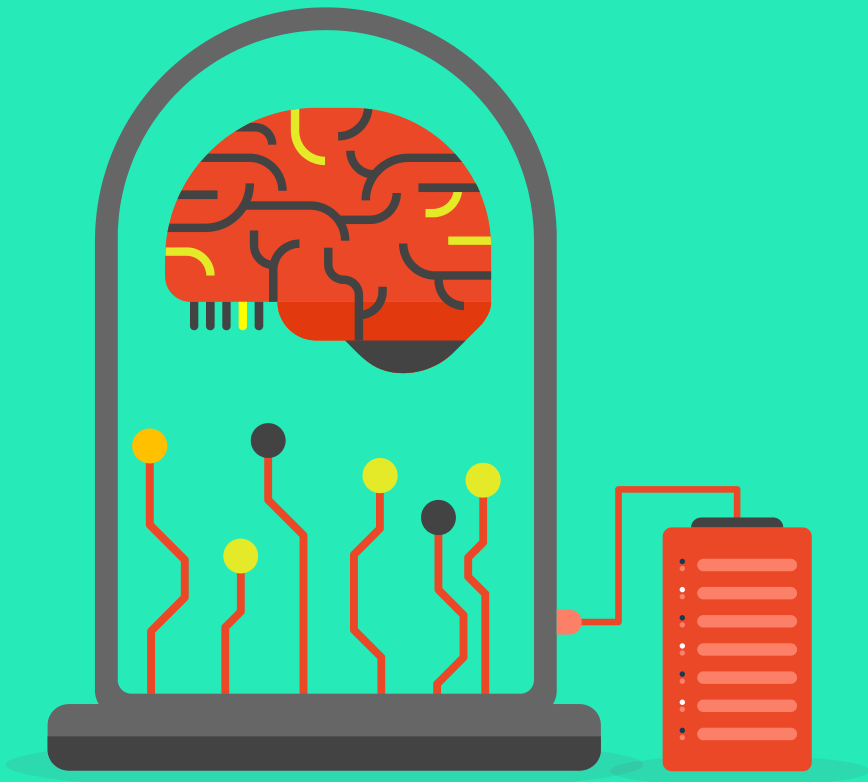


9. Regresión con Bosques Aleatorios



Métodos de Conjuntos (Ensemble Learning)

Los Métodos de Conjuntos o *Ensemble Learning* en Machine Learning son los algoritmos basados en tomar un grupo de predicciones provenientes de diferentes modelos, y agregar los resultados de cada uno para generar una predicción definitiva.

Por ejemplo, para un problema de regresión dado, podríamos construir una serie de árboles de decisión diferentes, cada uno entrenado con un subconjunto al azar de los datos de entrenamiento. Luego, para obtener el valor final de la regresión de un conjunto de datos nuevo, podríamos promediar las predicciones de cada uno de los árboles de decisión entrenados. A este “conjunto” de árboles de decisión se le llama **Bosque Aleatorio**.

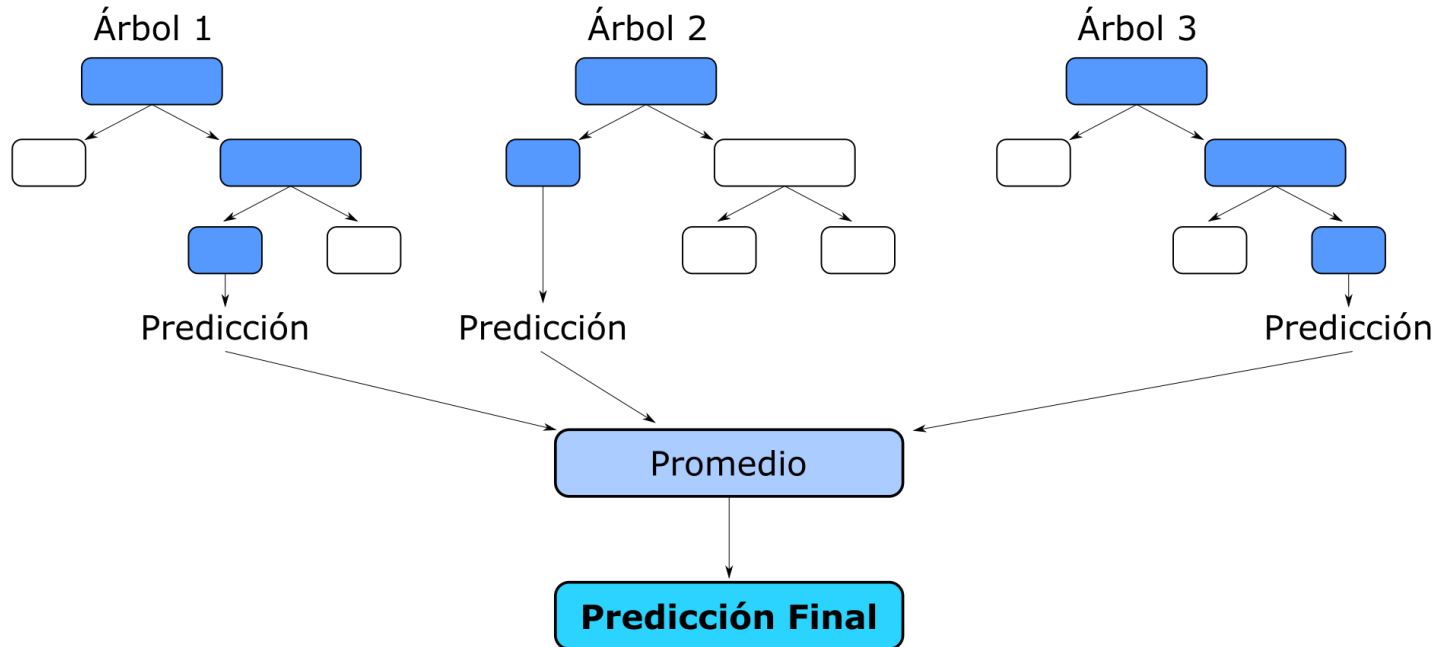
Bagging

Sin embargo, un ensemble podría estar compuesto por varios modelos combinados (por ejemplo, regresión lineal + SVR + árbol de decisión) y, de manera similar, obtener la predicción final como el promedio de las salidas de cada modelo.

En el caso de los **Bosques Aleatorios**, al método de *ensemble* empleado en el que se construyen muchos modelos similares entrenados sobre subconjuntos aleatorios (y con reemplazo, es decir, que el mismo dato puede seleccionarse varias veces) de los datos se le conoce como **Bagging** (por *bootstrap aggregating*). En el caso en el que no se produzca reemplazo de los datos, se le llama **Pasting**.

Bosques Aleatorios

De esta manera, se consigue que el bosque aleatorio, como conjunto, es capaz de ofrecer mejores resultados de predicción que incluso el mejor de los árboles de decisión que lo componen.



Bosques Aleatorios

Los hiperparámetros de cada árbol de decisión seguirán siendo los mismos vistos anteriormente, y para el bosque aleatorio como tal, el hiperparámetro fundamental será la cantidad de árboles que conformará el bosque.

Sin embargo, durante el entrenamiento del bosque aleatorio, como cada subconjunto de entrenamiento para cada árbol es distinto, se introduce al modelo una aleatoriedad tal que se garantiza que cada árbol que compone el bosque sea lo suficientemente diferente como para, en conjunto, garantizar un balance entre los sesgos y las varianzas individuales de cada uno.

En general, los bosques aleatorios suelen ofrecer resultados tan buenos (en especial para datos estructurados) que representan uno de los modelos más poderosos y utilizados dentro del Machine Learning.

Décimo Notebook Práctico

Veamos en nuestro próximo notebook práctico el cómo se implementa un bosque aleatorio para conjuntos de datos conocidos como los que hemos trabajado hasta este momento.

¡Adelante!

Décimo Notebook Práctico