Curso

6

MACHINE LEARNING & BIG DATA

Supervised Learning: Random Forest

Algoritmos Supervisados

JOSÉ NELSON ZEPEDA DOÑO

Cluster de Estudio: Advanced Analytics

Este material es el resumen de muchos autores que por medio de sus libros y documentos nos ofrecen fuentes riquísimas de conocimiento sobre los temas de Big Data y Machine Learning.
Algunas citas, figuras y tablas pueden ser encontradas de forma textual tal como lo indica el autor en su material original.

Nelson Zepeda
MIP • V 1.0
San Salvador El Salvador
Phone 503 79074137 • @nelsonzepeda733

Tabla de Contenido

Capítulo 1

Random Forests

¿Y si en lugar de pedir una sola opinión, le pedimos la opinión a un grupo de personas?

Un modelo Random Forests, tal como su nombre lo indica, se compone de un grupo de árboles de decisión ya sea de clasificación o regresión, los cuales son vistos como un bosque.

Estos árboles son construidos mediante un algoritmo que trata de reducir la correlación entre ellos. Una vez el bosque está construido, las predicciones son generadas tomando en cuenta las predicciones individuales de cada árbol.

El autor detrás de Random Forests es Leo Breiman, quien junto con su colaboradora Adele Cutler trabajaron y publicaron diferentes documentos que permitieron establecer el vínculo correcto entre la estadística y las ciencias de la computación.

Historia y Concepto

Veamos otros conceptos de RF[1]:

- Random forest es un método que combina una cantidad grande de árboles de decisión independientes probados sobre conjuntos de datos aleatorios con igual distribución.

- Random Forest son un tipo de método de particionamiento recursivo especialmente adecuado para pequeñas n y grandes p (pocos datos y muchas variables).

- Random forest (o random forests) también conocidos en castellano como "'Bosques Aleatorios'" es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos. Es una modificación sustancial de bagging que construye una larga colección de árboles no correlacionados y luego los promedia.

- Random Forest es como si un grupo de personas buscasen un buen restaurante en una misma zona y entre todos votasen cuál es el mejor.

---

[1] https://quantdare.com/random-forest-vs-simple-tree/ https://blog.bigml.com/2013/04/04/the-three-cardinal-virtues-of-ensemble-learning/

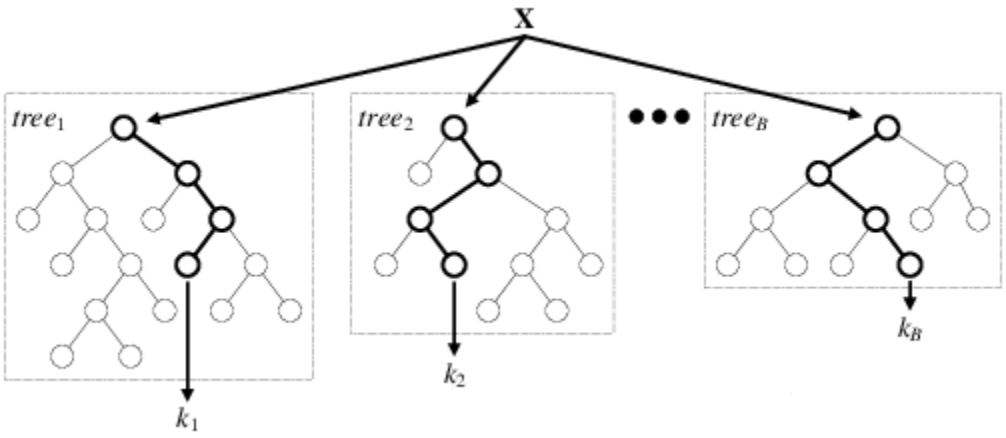

Figura 1-1 Ejemplificación de un modelo RF

El término aparece de la primera propuesta de Random decision forests, hecha por Tin Kam Ho de Bell Labs en 1995 y no esta demás mencionar que el método combina la idea de bagging de Breiman y la selección aleatoria de atributos, introducida independientemente por Ho.

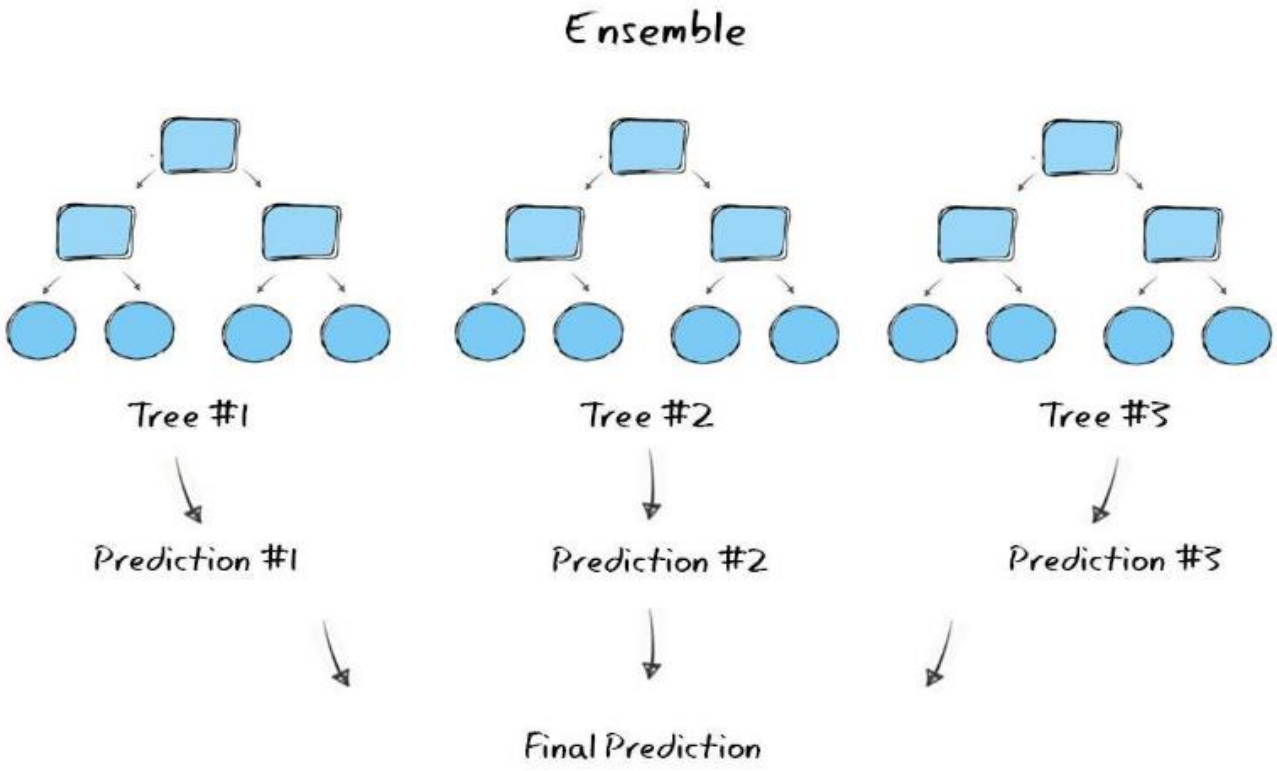

Figura 1-2 Obteniendo una predicción.

La aleatoriedad se introduce en el modelo con el objetivo de reducir la varianza mediante la reducción de la correlación entre los árboles.

La formación de cada árbol con una muestra diferente de los datos iniciales y la partición en cada nodo con distintas variables genera arboles con estructuras diferentes reduciendo la correlación entre los mismos.

Fortalezas y Debilidades

Fortalezas:

- RF es más robusto en términos de exactitud que un algoritmo de árbol de decisión, esto es debido a que reduce el sobre entrenamiento.
- RF presenta más estabilidad que un solo árbol de decisión por lo cual es más consistente en las predicciones.

Debilidades:

- Son difíciles de comprender e interpretar, es un verdadero desafío comprender que es lo que miles de árboles están haciendo de forma aleatoria.
- Requieren de un gran poder computacional.

Proceso general de construcción

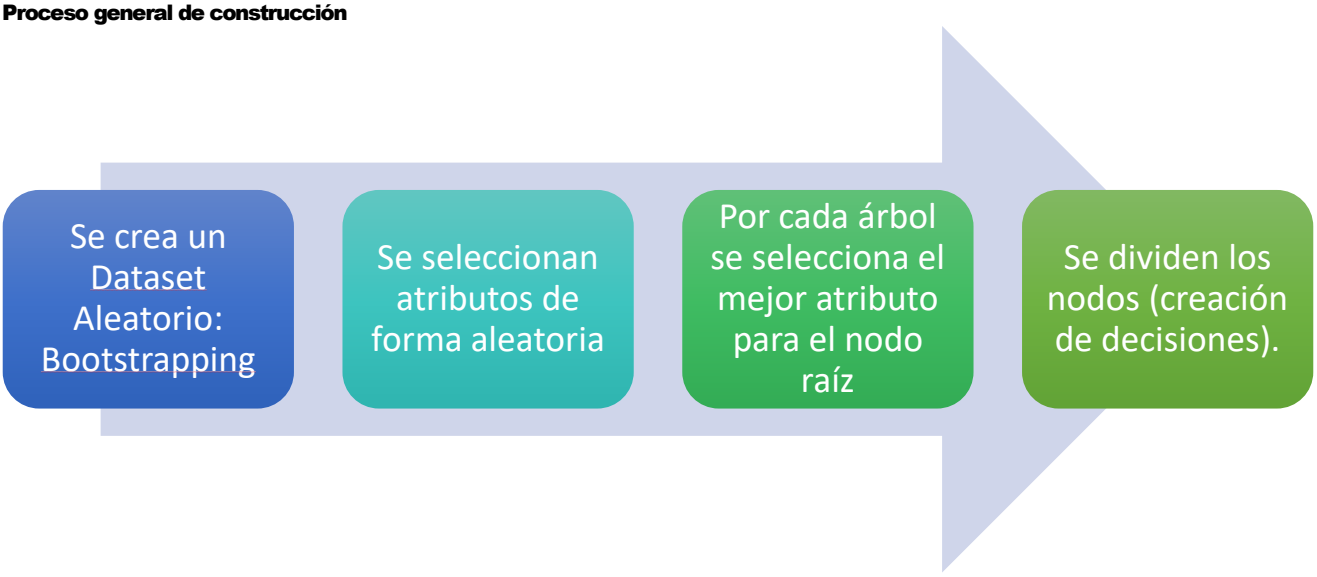

Figura 1-3 Proceso General de construcción de un RF.

El algoritmo de RF puede describirse de la siguiente manera:

- Para cada uno de los árboles, dada la base de entrenamiento, se eligen de forma aleatoria cierta cantidad de datos con reemplazamiento. Esto último se conoce como bootstrapping. El hecho de que cada árbol

se construye con una fuente diferente que ha sido seleccionada aleatoriamente constituye el primer pilar de este algoritmo y el modelo resultante.

- En cada nodo de cada árbol, se eligen de forma aleatoria variables/atributos candidatos para la partición. El número de atributos seleccionados será constante durante todo el proceso de formación del modelo, esta selección aleatoria de atributos constituye el segundo pilar de este algoritmo.

- Se deja crecer el árbol.

El algoritmo es igual tanto para arboles de regresión como de clasificación, una vez construido el bosque, cada predicción se realiza tomando en cuenta las predicciones individuales de cada uno de los árboles.

Por tanto, Random Forests tiene 2 parámetros fundamentales de diseño:

- NTREE: parámetro que indica el número de árboles que constituyen el bosque.

- Mtry: parámetro que indica cuantos atributos se seleccionaran para construir cada árbol.

Variaciones en ambos parámetros conducen a resultados ligeramente diferentes, al reducir la cantidad de atributos, se reduce la correlación entre los árboles debido a que en cada nodo se tienen menos posibilidades de variables entre las que elegir con el objetivo de reducir la impureza, sin embargo, reducir la cantidad de atributos también puede reducir la precisión de cada árbol individual.

Por otro lado, el número de árboles tiene un efecto directo en la precisión de la predicción. En forma lógica, cuantos más arboles individuales se construyan mejor serán las predicciones del modelo, sin embargo, existe un cierto valor de la cantidad de árboles en el cual se estabiliza el error de predicción, es decir que existe un numero óptimo de árboles, una vez alcanzado ese número optimo, cada árbol adicional contribuirá muy poco a la reducción del error en la predicción y por otro lado aumentara en vano los costos computacionales del modelo.

Bootstrapping

Según Wikipedia, el bootstrapping puede definirse como un método de remuestreo. Se extrae una muestra a partir de la muestra original, lo que hace especial al bootstrapping, es que dicha muestra debe extraerse utilizando un muestreo con reposición, de tal forma que algunos elementos no serán seleccionados y otros lo podrán ser más de una vez en cada muestreo.

Aunque pueda parecer una práctica muy compleja a priori, el procedimiento en que se basa el bootstrapping es simplemente la creación de un gran número de muestras reposicionando los datos tomando como referencia una muestra poblacional inicial. Esta técnica resulta especialmente útil en aquellas situaciones en las que las muestras con las que se cuenta son pequeñas o, como se dijo antes, si la distribución es muy sesgada. En ese sentido, ayudan a la resolución de multitud de problemas de probabilidad y estadística aplicada[2].

[2] https://economipedia.com/definiciones/bootstrap.html

De cara a Random Forests, podemos acotar lo siguiente en cuanto al bootstrapping:

- Cada árbol se construye utilizando un dataset de entrenamiento.
- Cada dataset es 100% único
- Se seleccionan muestras aleatorias del dataset de entrenamiento y se recrea un DataSet para cada árbol
- El mayor beneficio del Boostrapping es que garantizamos que cada línea del dataset de entrenamiento este incluida al menos en uno de los árboles de decisión.
- El dataset de entrenamiento original tiene un tamaño en cantidad de observaciones (filas) N, cada dataset creado para cada árbol individual también será de tamaño N.
- En los datasets recreados para cada árbol, una misma observación puede aparecer múltiples ocasiones. Esto es llamada "Reemplazo".

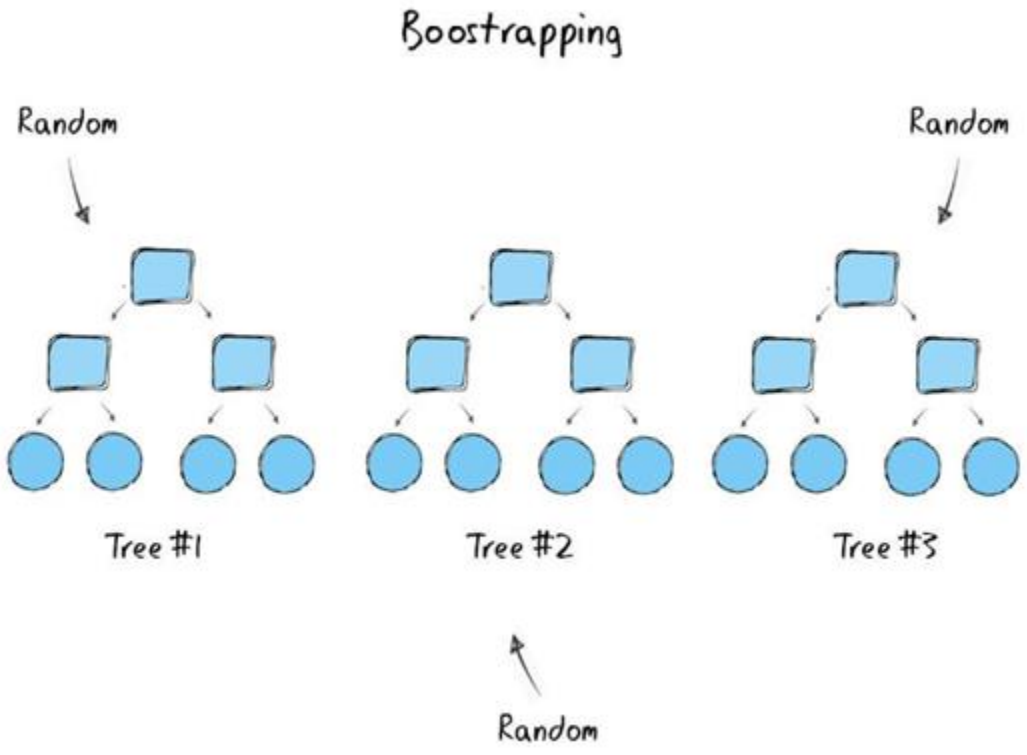

Figura 1-4 Bootstrapping

Movie Name	Key Actor	Category	Rating	Watched?
Inglourious Basterds	Brad Pitt	Action	Average	No
Snatch	Brad Pitt	Action	Average	No
Troy	Brad Pitt	Action	Low	No
The Curious Case of Benjamin Button	Brad Pitt	Drama	Average	No
Mr & Mrs. Smith	Brad Pitt	Romance	Low	No
Once Upon a Time in Mexico	Brad Pitt	Romance	Low	No
The Magnificent Seven	Denzel Washington	Action	Average	No
The Book of Eli	Denzel Washington	Action	Low	No
Flight	Denzel Washington	Drama	Average	No
Inferno	Tom Hanks	Drama	Low	No
You've Got Mail	Tom Hanks	Romance	Low	No
Fight Club	Brad Pitt	Action	Average	Yes
Allied	Brad Pitt	Drama	Average	Yes
Fury	Brad Pitt	Drama	Average	Yes
Moneyball	Brad Pitt	Drama	Exceptional	Yes
Angels and Demons	Tom Hanks	Drama	Exceptional	Yes
Captain Phillips	Tom Hanks	Drama	Average	Yes
Apollo 13	Tom Hanks	Drama	Exceptional	Yes
Forrest Gump	Tom Hanks	Drama	Exceptional	Yes
Cast Away	Tom Hanks	Drama	Average	Yes
Sleepless in Seattle	Tom Hanks	Romance	Exceptional	Yes

Each row is an example. There are 21 in total.

Random Selection of Training Samples

Movie Name	Key Actor	Category	Rating	Watched?
Inglourious Basterds	Brad Pitt	Action	Average	No
Inglourious Basterds	Brad Pitt	Action	Average	No
Once Upon a Time in Mexico	Brad Pitt	Romance	Low	No
The Magnificent Seven	Denzel Washington	Action	Average	No
Fight Club	Brad Pitt	Action	Average	Yes
Allied	Brad Pitt	Drama	Average	Yes
Allied	Brad Pitt	Drama	Average	Yes
Fury	Brad Pitt	Drama	Average	Yes
Fury	Brad Pitt	Drama	Average	Yes
Fury	Brad Pitt	Drama	Average	Yes
Forrest Gump	Tom Hanks	Drama	Exceptional	Yes
Cast Away	Tom Hanks	Drama	Average	Yes
The Magnificent Seven	Denzel Washington	Action	Average	No
Troy	Brad Pitt	Action	Low	No
Troy	Brad Pitt	Action	Low	No
Mr & Mrs. Smith	Brad Pitt	Romance	Low	No
Mr & Mrs. Smith	Brad Pitt	Romance	Low	No
Captain Phillips	Tom Hanks	Drama	Average	Yes
Forrest Gump	Tom Hanks	Drama	Exceptional	Yes
Captain Phillips	Tom Hanks	Drama	Average	Yes
Flight	Denzel Washington	Drama	Average	No

Figura 1-5: Ejemplo generación New Dataset

RF- ¿Cómo predice?

Existen 2 maneras para predecir:

- Por voto: es la estrategia más popular y se basa en el conteo de votos, es decir cada árbol emite un voto y la mayoría de votos es la ganadora y la que el modelo devuelve.

- Por promedio: Se basa en el promedio, es decir se calcula el porcentaje de cada uno de los resultados posibles y se muestra el mayoritario.

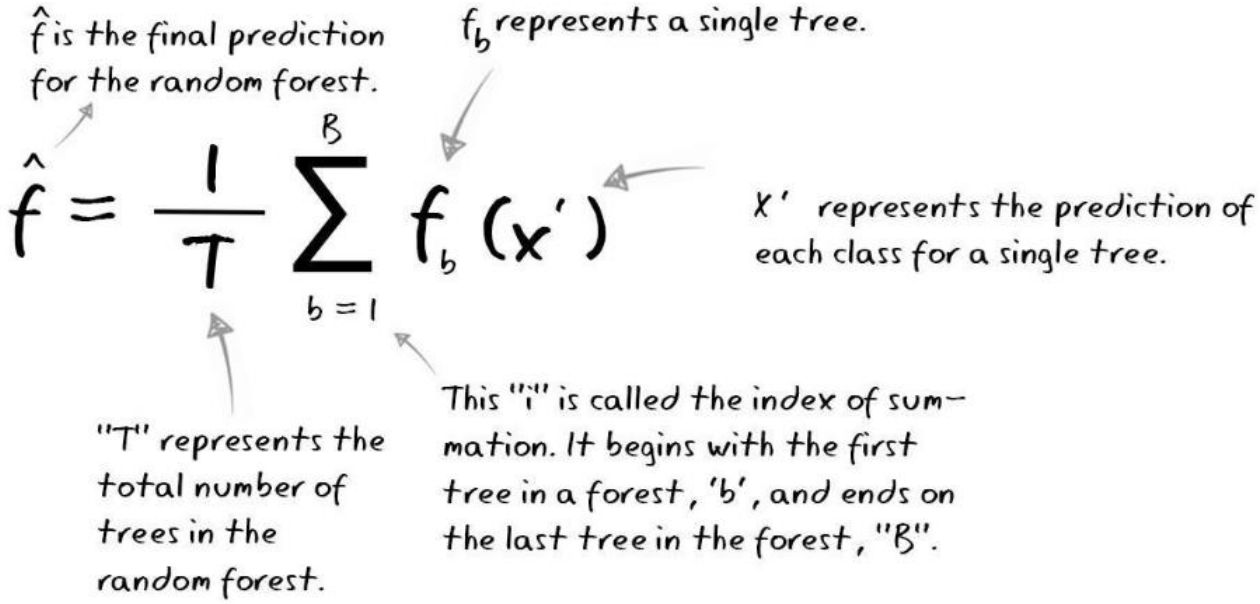

Figura 1-6 Calculando la predicción de acuerdo a todos los árboles.

Supongamos que nuestro modelo RF está compuesto por 3 árboles y obtenemos el siguiente resultado:

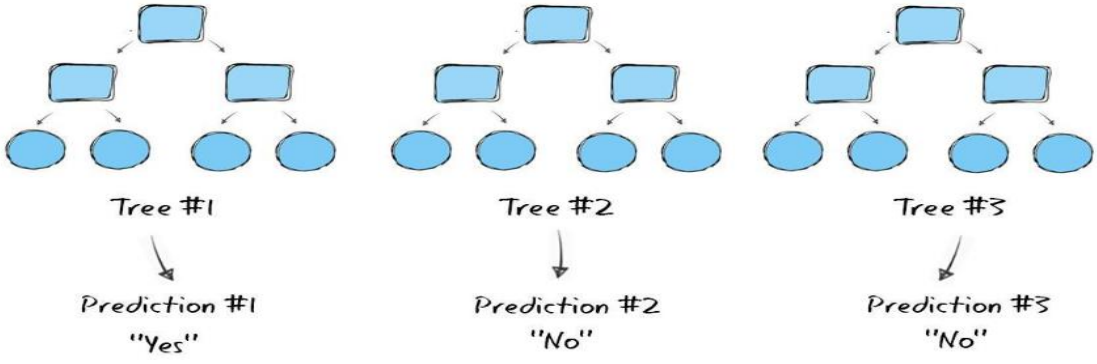

Figura 1-7 Ejemplo de votación

O.O.B.

El Out Of Bag Mean Squared Error es una medición típica de los modelos Random Forests y de otros algoritmos que emplean la técnica del bootstrapping.

En la elección aleatoria con reemplazamiento de los N datos sobre las N observaciones que forman la muestra inicial, realizada antes de la construcción de cada uno de los árboles, se suele quedar fuera de esta sub muestra en torno al 37% del total de las observaciones.

El MSE-OOB estima el error de predicción del modelo teniendo en cuenta las observaciones que han quedado "fuera de la bolsa"

$$MSE - OOB = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_{iOOB})^2$$

Siendo Yioob la predicción para la observación i obtenida promediando las predicciones individuales de los árboles para los que esta observación se ha quedado fuera de la bolsa e Yi el valor real de la variable de respuesta.

El MSE-OOB tiene una dependencia importante con los parámetros del modelo: cantidad de atributos, cantidad de árboles.

Sobre-entrenamiento

El sobreajuste o sobreentrenamiento es un término muy utilizado en estadística. Es un fenómeno que se produce cuando el algoritmo de tratamiento de datos generado se ajusta con mucha precisión a los datos de partida con los que se ha creado, pero es incapaz de predecir con suficiente precisión datos que se encuentran fuera de la muestra inicial.

En los arboles de decisión este fenómeno se mitiga con la poda, los RF son menos sensibles a este fenómeno, siendo muy improbable el sobreajuste.

Esto se debe al algoritmo de generación de los árboles, en los que la aleatoriedad en la elección de la muestra de cada árbol y de las variables candidatas a provocar la partición de cada nodo, contribuye de manera significativa a crear árboles diferentes que después serán promediados.

Sin embargo, es posible que, en determinados casos, una elección incorrecta de los parámetros fundamentales produzcan cierto grado de sobreajuste en el modelo, perdiendo precisión en las predicciones posteriores al entrenamiento.

Bibliografía

- Decision Trees
  By Chris Smith, 2017

- Modelado mediante RF de las emisiones de autobuses urbanos en función de los ciclos cinemáticos
  By Victor Pita González-Campos , 2017

- R Data Analysis Cookbook
  by Kuntal Ganguly,2017