Predicción

Sabina Alejandra Castillo Trujillo	1804029
Melany Salazar Mata	1679234
Gabriel Adrián Contreras García	1752950
Víctor Alanís Mares	1821920
Jesús Ramon Castro Hernández	1887860

Metodologia de la partición de datos

Antes de empezar...

ELEMENTOS PARA HACER UN BUEN MODELO DE PREDICCIÓN

ELEMENTOS PREVIOS

Definir adecuadamente nuestro problema (objetivo, salidas deseadas…..).

Recopilar datos.

Elegir una medida o indicador de éxito.

Preparar los datos (tratar con campos vacíos, con valores categóricos..)

Dividir los datos

70% CONJUNTO DE ENTRENAMIENTO

15% CONJUNTO DE VALIDACIÓN

15% CONJUNTO DE PRUEBAS.

Árboles aleatorios

Árbol de decisión

Modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta o dependiente.

Para dividir el espacio muestral en subregiones es preciso aplicar una serie de reglas o decisiones, para que cada subregión contenga la mayor proporción posible de individuos de una de las poblaciones.

Si una subregión contiene datos de diferentes clases, se subdivide en regiones más pequeñas hasta fragmentar el espacio en subregiones menores que integran datos de la misma clase.

Los árboles se pueden clasificar en dos tipos que son:

1. Árboles de regresión en los cuales la variable respuesta y es cuantitativa.
2. Árboles de clasificación en los cuales la variable respuesta y es cualitativa.

Estructura básica de un árbol de decisión

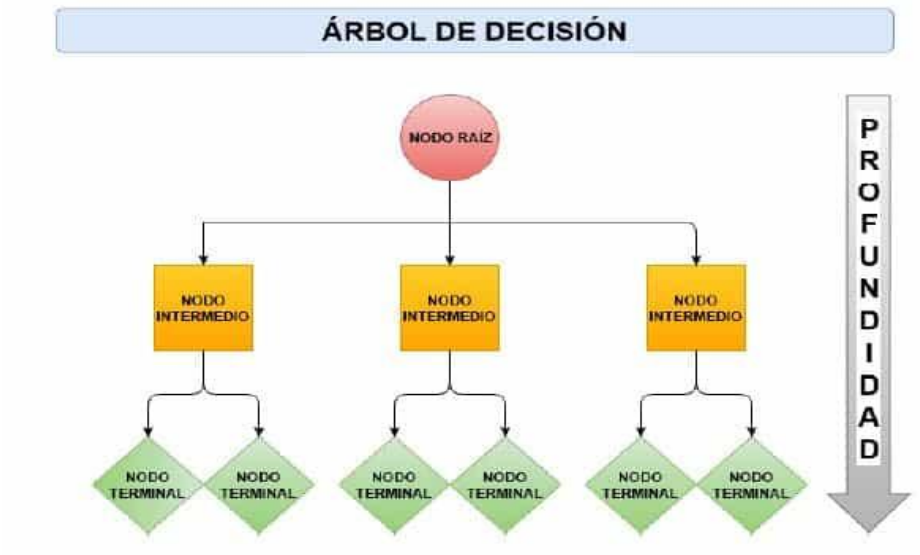
Los árboles de decisión están formados por nodos y su lectura se realiza de arriba hacia abajo.

Dentro de un árbol de decisión distinguimos diferentes tipos de nodos:

- ✓ Primer nodo o nodo raíz: en él se produce la primera división en función de la variable más importante.
- ✓ Nodos internos o intermedios: tras la primera división encontramos estos nodos, que vuelven a dividir el conjunto de datos en función de las variables.
- ✓ Nodos terminales u hojas: se ubican en la parte inferior del esquema y su función es indicar la clasificación definitiva.

Otro concepto que debes tener claro es la profundidad de un árbol, que viene determinada por el número máximo de nodos de una rama.

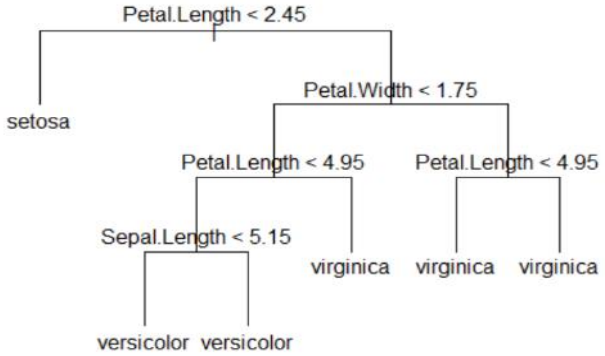
A continuación, te mostramos un ejemplo gráfico:

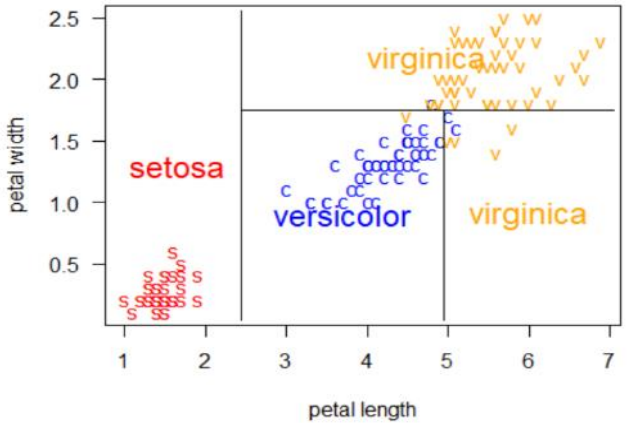


Árbol de clasificación

Consiste en hacer preguntas del tipo $¿x_k \leq c?$ para las covariables cuantitativas o preguntas del tipo $¿x_k = nivel_j?$ para las covariables cualitativas, de esta forma el espacio de las covariables es divido en hiper-rectángulos y todas las observaciones que queden dentro de un hiper-rectángulo tendrán el mismo valor grupo estimado.

En la siguiente figura se ilustra el árbol en el lado izquierdo y la partición del espacio en el lado derecho. La partición del espacio se hace de manera repetitiva para encontrar las variables y los valores de corte de tal manera que se minimice la función de costos.





Ejemplo

Iris setosa

Iris versicolor

Iris virginica

Si le damos las 150 flores del conjunto de datos Iris a un árbol de decisión para que lo clasifique, nos quedaría un árbol como el que se muestra a continuación. Vamos a aprender a leerlo:

- Cada color representa a una clase. El marrón para setosa, el verde para versicolor y el lila para virginica.
- El color es más intenso cuanto más seguros estamos que la clasificación es correcta
- Los nodos blancos, por tanto, evidencia la falta de certeza

Hay dos tipos de nodo:

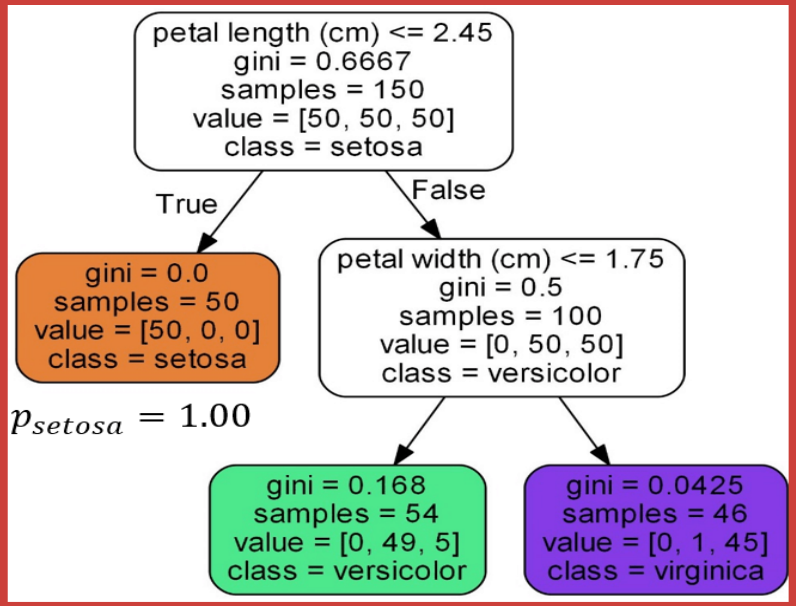
- Nodos de decisión: tienen una condición al principio y tienen más nodos debajo de ellos
- Nodos de predicción: no tienen ninguna condición ni nodos debajo de ellos. También se denominan «nodos hijo»

La información de cada nodo es la siguiente:

- Condición: Si es un nodo donde se toma alguna decisión.
- Gini: Es una medida de impureza. A continuación, veremos cómo se calcula.
- Samples: Número de muestras que satisfacen las condiciones necesarias para llegar a este nodo.
- Value: Cuántas muestras de cada clase llegan a este nodo.
- Class: Qué clase se les asigna a las muestras que llegan a este nodo.

Interpretación

La interpretación de este árbol de decisión sería: si la longitud del pétalo es menos de 2.45 centímetros, entonces la flor iris pertenece a la variedad setosa. Si, por el contrario, la longitud del pétalo es mayor que 2.45 centímetros, habría que mirar al ancho del pétalo. Cuando el ancho del pétalo es menor o igual a 1.75 centímetros, pertenece a la variedad versicolor con un 91% de probabilidad. Si no, parece que sería virginica con un 98% de probabilidad.



$p_{setosa} = 1.00$

$p_{versicolor} = 0.91$ $p_{virginica} = 0.98$

Gini: medida de limpieza

Gini es una medida de impureza. Cuando Gini vale 0, significa que ese nodo es totalmente puro. La impureza se refiere a cómo de mezcladas están las clases en cada nodo. Para calcular la impureza Gini, usamos la siguiente fórmula:

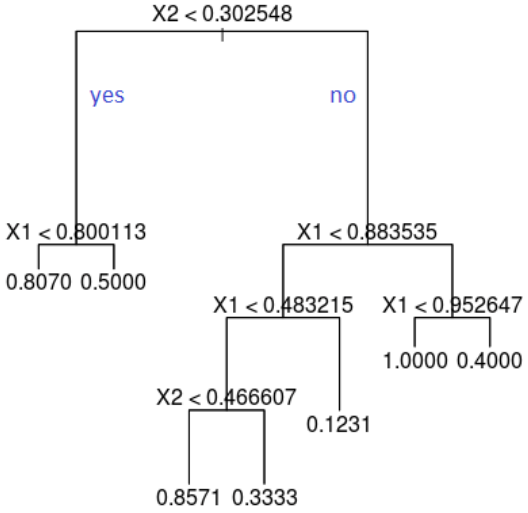
$$gini = 1 - \sum_{k=1}^{n} p_c^2$$

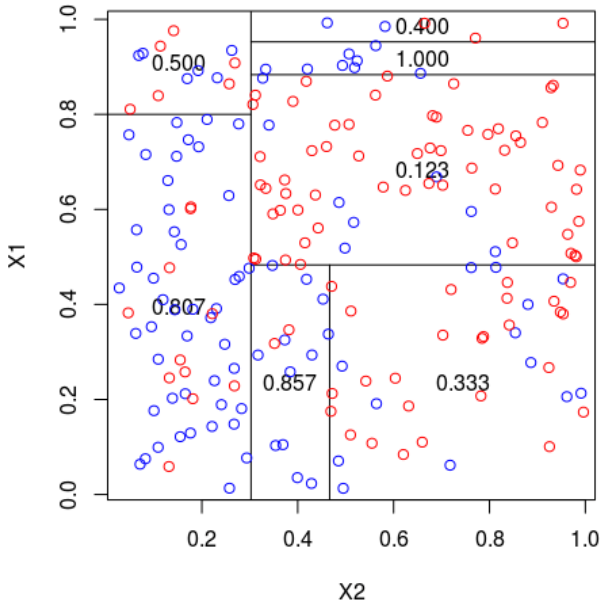
p_c se refiere a la probabilidad de cada clase. Podemos calcularla dividiendo el número de muestras de cada clase en cada nodo por el número de muestras totales por nodo.

Árbol de regresión

Consiste en hacer preguntas de tipo $¿x_k \leq c?$ para cada una de las covariables, de esta forma el espacio de las covariables es divido en hiper-rectángulos y todas las observaciones que queden dentro de un hiper-rectángulo tendrán el mismo valor estimado $\hat{y}$.

En la siguiente figura se ilustra el árbol en el lado izquierdo y la partición del espacio en el lado derecho.





Pasos

Los pasos para realizar la partición del espacio son:
1. Dado un conjunto de covariables (características), encontrar la covariable que permita predecir mejor la variable respuesta.
2. Encontrar el punto de corte sobre esa covariable que permita predecir mejor la variable respuesta.
3.Repetir los pasos anteriores hasta que se alcance el criterio de parada.

Ventajas

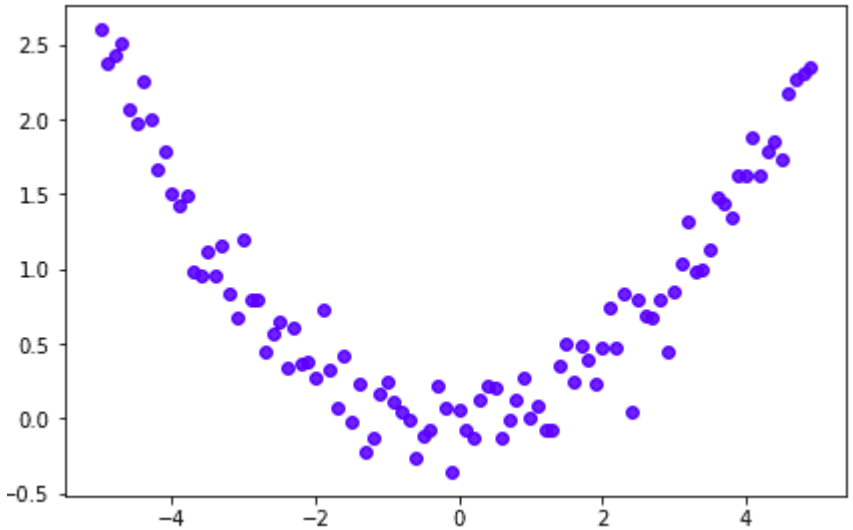
Algunas de las ventajas de los árboles de regresión son:

- Fácil de entender e interpretar.
- Requiere poca preparación de los datos.
- Las covariables pueden ser cualitativas o cuantitativas.
- No exige supuestos distribucionales.

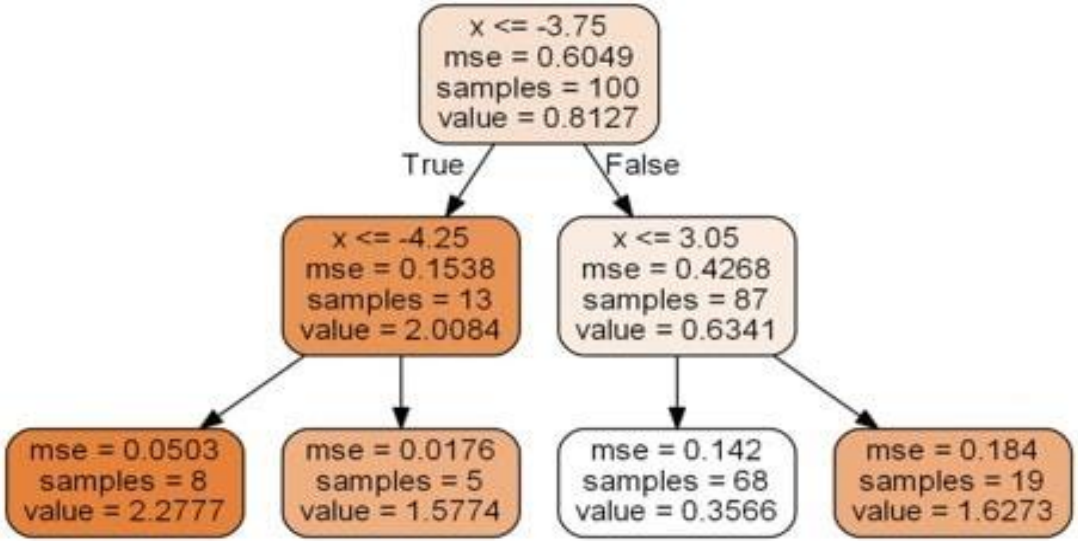
Ejemplo

Para explicar cómo funcionan los árboles de decisión para problemas de regresión vamos a usar los datos que se presentan en la siguiente gráfica. Para generarlos he usado la siguiente fórmula en el intervalo [-5, 5]:

$$y = 0.1x^2 + 0.2(Ruido\ Gaussiano)$$

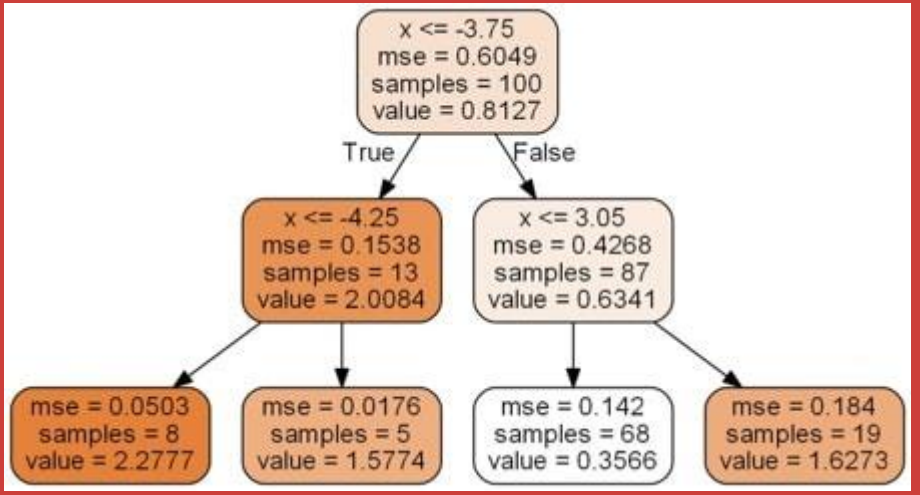


En el caso de regresión, en lugar de usar Gini como medida de impureza, usamos MSE, el error cuadrático medio. Para este problema, si usamos un árbol de decisión de profundidad 2, obtenemos el siguiente árbol.



Interpretación

La interpretación de este árbol de decisión sería: si el valor de x es menor que -4.25, predice 2.2777; si está en el intervalo (-4.25, -3.75] predice 1.5774; si está en el intervalo (-3.75, 3.05] predice 0.3566 y si es mayor que 3.05 predice 1.6273.



Bosques aleatorios

Random Forest (Bosque Aleatorio)

Random Forest

Técnica de aprendizaje automático supervisada basada en árboles de decisión. Su principal ventaja es que obtiene un mejor rendimiento de generalización para un rendimiento durante entrenamiento similar. Esta mejora en la generalización la consigue compensando los errores de las predicciones de los distintos árboles de decisión. Para asegurarnos que los árboles sean distintos, lo que hacemos es que cada uno se entrena con una muestra aleatoria de los datos de entrenamiento. Esta estrategia se denomina bagging.

Bagging

Una forma de mejorar un modelo predictivo es usando la técnica creada por Leo Breiman que denominó Bagging (o Bootstrap Aggregating). Esta técnica consiste en crear diferentes modelos usando muestras aleatorias con reemplazo y luego combinar o ensamblar los resultados.

La técnica de Bagging sigue estos pasos:

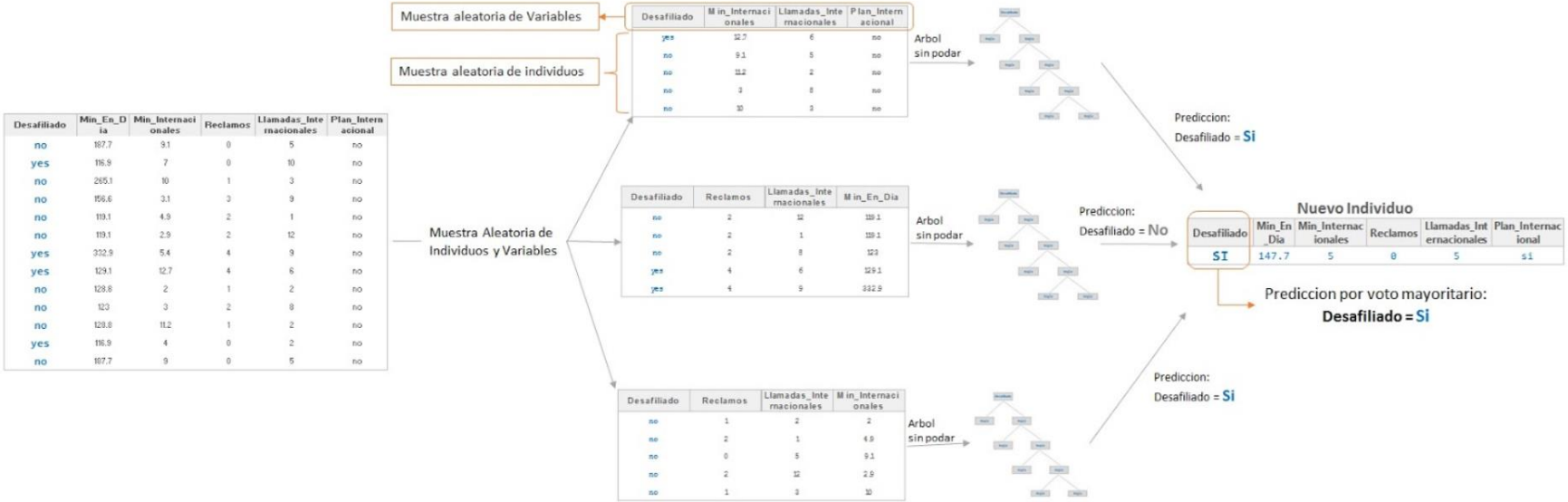
1. Divide el set de Entrenamiento en distintos sub set de datos, obteniendo como resultado diferentes muestras aleatorias con las siguientes características:
    - Muestra uniforme (misma cantidad de individuos en cada set)
    - Muestras con reemplazo (los individuos pueden repetirse en el mismo set de datos).
    - El tamaño de la muestra es igual al tamaño del set de entrenamiento, pero no contiene a todos los individuos ya que algunos se repiten.
    - Si se usan muestras sin reemplazo, suele elegirse el 50% de los datos como tamaño de muestra
2. Luego se crea un modelo predictivo con cada set, obteniendo modelos diferentes
3. Luego se construye o ensambla un único modelo predictivo, que es el promedio de todos los modelos.

¿Cómo funciona el algoritmo?

En forma resumida sigue este proceso:

- Selecciona individuos al azar (usando muestreo con reemplazo) para crear diferentes sets de datos.
- Crea un árbol de decisión con cada set de datos, obteniendo diferentes arboles, ya que cada set contiene diferentes individuos y diferentes variables en cada nodo.
- Al crear los arboles se eligen variables al azar en cada nodo del árbol, dejando crecer el árbol en profundidad (es decir, sin podar).
- Predice los nuevos datos usando el "voto mayoritario", donde clasificará como "positivo" si la mayoría de los arboles predicen la observación como positiva.

¿Cómo funciona el algoritmo?



Desafiliado	Min_En_Dia	Min_Internacionales	Reclamos	Llamadas_Internacionales	Plan_Internacional
no	187.7	9.1	0	5	no
yes	116.9	7	0	10	no
no	265.1	10	1	3	no
no	156.6	3.1	3	9	no
no	119.1	4.9	2	1	no
no	119.1	2.9	2	12	no
yes	332.9	5.4	4	9	no
yes	129.1	12.7	4	6	no
no	128.8	2	1	2	no
no	123	3	2	8	no
no	128.8	11.2	1	2	no
yes	116.9	4	0	2	no
no	187.7	9	0	5	no

Desafiliado	Min_Internacionales	Llamadas_Internacionales	Plan_Internacional
yes	12.7	6	no
no	9.1	5	no
no	11.2	2	no
no	3	8	no
no	10	3	no

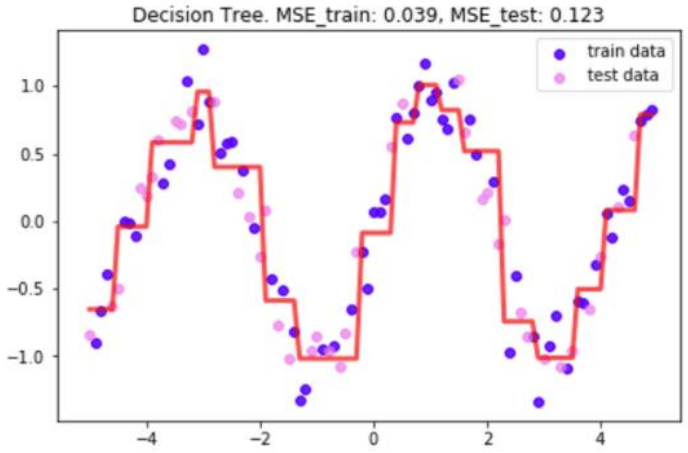
Desafiliado	Reclamos	Llamadas_Internacionales	Min_En_Dia
no	2	12	119.1
no	2	1	119.1
no	2	8	123
yes	4	6	129.1
yes	4	9	332.9

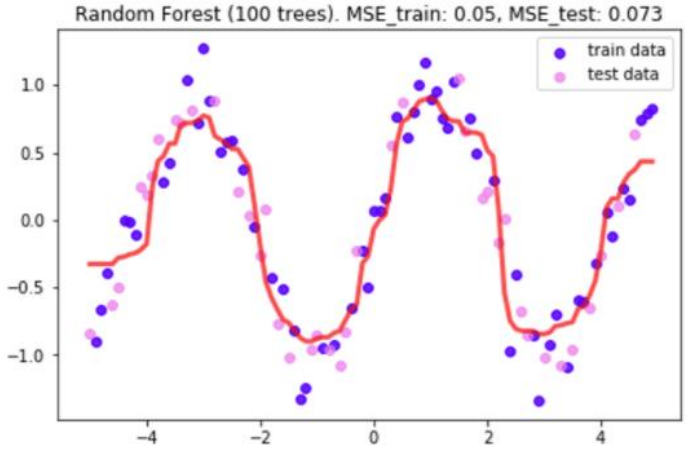
Desafiliado	Reclamos	Llamadas_Internacionales	Min_Internacionales
no	1	2	2
no	2	1	4.9
no	0	5	9.1
no	2	12	2.9
no	1	3	10

Desafiliado	Min_En_Dia	Min_Internacionales	Reclamos	Llamadas_Internacionales	Plan_Internacional
SI	147.7	5	0	5	si

Diferencia entre un árbol de decisión y un Random forest

En la siguiente imagen puedes ver la diferencia entre el modelo aprendido por un árbol de decisión y un random forest cuando resuelven el mismo problema de regresión. Este random forest en particular, utiliza 100 árboles.

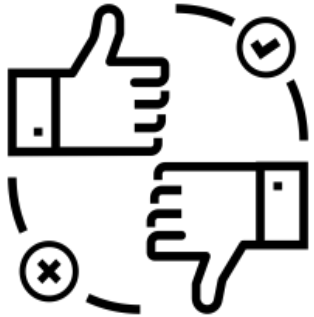




Ventajas y Desventajas

Dato que un random forest es un conjunto de árboles de decisión, y los árboles son modelos no-paramétricos, los random forests tienen las mismas ventajas y desventajas de los modelos no-paramétricos:

- Ventaja: pueden aprender cualquier correspondencia entre datos de entrada y resultado a predecir
- Desventaja: no son buenos extrapolando… porque no siguen un modelo conocido



Validación cruzada

Validación cruzada

Se emplea para estimar el ***test error rate*** de un modelo y así evaluar su capacidad predictiva, a este proceso se le conoce como ***model assessment***. También se puede emplear para seleccionar el nivel de flexibilidad adecuado.

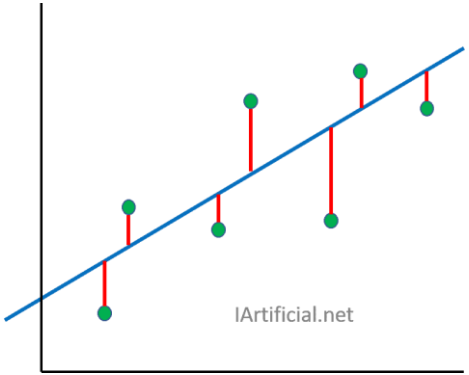
Validación simple
Leave One Out Cross-Validation (LOOCV)K-Fold Cross Validation

MÉTRICAS DE EFICACIA

PARA DATOS NUMÉRICOS Y CATEGÓRICOS

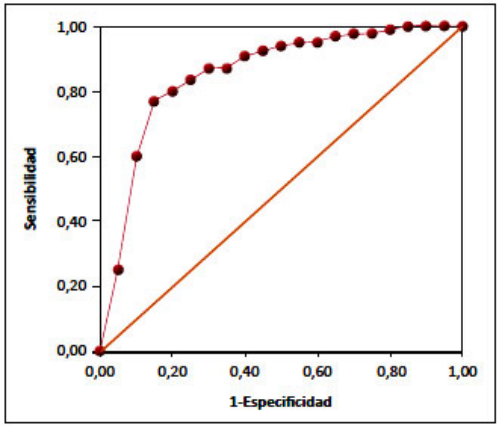
ERROR CUADRÁTICO MEDIO:

- Mide el promedio de los errores al cuadrado, es decir, la diferencia entre el estimador y lo que se estima.
- Valor esperado de la pérdida del error al cuadrado



CURVA ROC:

- Nos sirve para conocer el rendimiento global de la prueba (área bajo la curva)
- Eje X – Falsos positivos
- Eje Y – Verdaderos positivos



Ejemplo:

- **Planeación del problema:**
  - ✓ Se tienen los datos de un conjunto de personas a las que se les medidores de actividad física en piernas, brazos y torso, posteriormente se les pidió que levantaran cierto peso y se calificó cómo hicieron el ejercicio de la A a la E (siendo la A la mejor ejecución y E la peor ejecución)
- **Objetivo:**
  - ✓ Sabiendo la cantidad de actividad física que se hizo y en qué parte del cuerpo se concentró, predecir qué tan bien o mal una persona hizo dicho ejercicio

More information is available from the website here: http://groupware.les.inf.puc-rio.br/har (see the section on the Weight Lifting Exercise Dataset).

Carga, exploración y depuración de datos

El siguiente paso es cargar el conjunto de datos desde la URL proporcionada anteriormente. Luego, el conjunto de datos se divide en 2 para crear un conjunto de entrenamiento (70% de los datos) para el proceso de modelado y un conjunto de prueba (con el 30% restante) para las validaciones.

Se eliminaron varias variables del conjunto debido a que muchas sólo contenían valores NA, además se hizo un análisis de correlación entre variables

CONSTRUCCIÓN DEL MODELO

Se probarán dos métodos:

- Árboles de decisión
- Bosques aleatorios

Se analizará la eficacia de cada uno para este ejemplo, además de poder visualizar su matriz de confusión

BOSQUE ALEATORIO

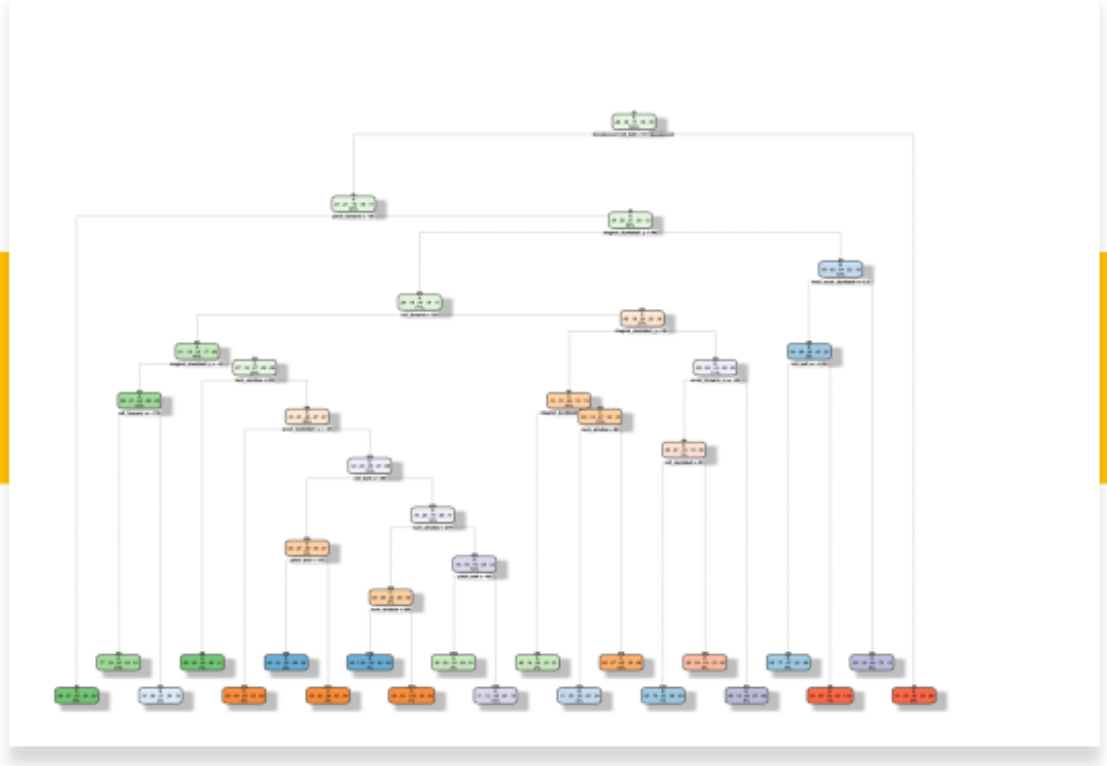
```r
# model fit
set.seed(12345)
controlRF <- trainControl(method="cv", number=3, verboseIter=FALSE)
modFitRandForest <- train(classe ~ ., data=TrainSet, method="rf",
                          trControl=controlRF)
modFitRandForest$finalModel
```

CONJUNTO DE ENTRENAMIENTO

```
Confusion matrix:
     A    B    C    D    E  class.error
A 3904    2    0    0    0 0.0005120328
B    6 2647    4    1    0 0.0041384500
C    0    5 2391    0    0 0.0020868114
D    0    0    9 2243    0 0.0039964476
E    0    0    0    5 2520 0.0019801980
```

CONJUNTO DE PRUEBA

```
          Reference
Prediction    A    B    C    D    E
         A 1674    1    0    0    0
         B    0 1138    2    0    0
         C    0    0 1024    2    0
         D    0    0    0  962    1
         E    0    0    0    0 1081
```



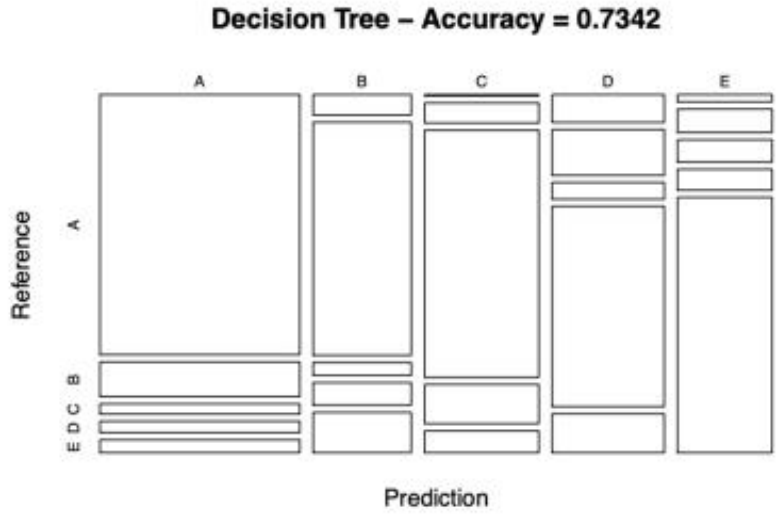
ÁRBOL DE DECISÓN

ÁRBOL DE DECISIÓN

```r
# model fit
set.seed(12345)
modFitDecTree <- rpart(classe ~ ., data=TrainSet, method="class")
fancyRpartPlot(modFitDecTree)
```

CONJUNTO DE PRUEBA

Prediction	Reference A	B	C	D	E
A	1502	201	59	66	74
B	58	660	37	64	114
C	4	66	815	129	72
D	90	148	54	648	126
E	20	64	61	57	696



ESTADÍSTICAS GENERALES

BOSQUES ALEATORIOS

```
Overall Statistics

               Accuracy : 0.999
                 95% CI : (0.9978, 0.9996)
    No Information Rate : 0.2845
    P-Value [Acc > NIR] : < 2.2e-16
```

ÁRBOLES DE DECISÓN

```
Overall Statistics

               Accuracy : 0.7342
                 95% CI : (0.7228, 0.7455)
    No Information Rate : 0.2845
    P-Value [Acc > NIR] : < 2.2e-16
```

Referencias

https://www.aprendemachinelearning.com/sets-de-entrenamiento-test-validacion-cruzada/
https://medium.com/datos-y-ciencia/machine-learning-c%C3%B3mo-desarrollar-un-modelo-desde-cero-cc17654f0d48
https://www.cienciadedatos.net/documentos/30_cross-validation_oneleaveout_bootstrap#k-fold_cross-validation
http://www.hrc.es/bioest/roc_1.html
https://es.wikipedia.org/wiki/Error_cuadr%C3%A1tico_medio
* Santana, E. (2014, 14 diciembre). Bagging para mejorar un modelo predictivo. Recuperado 21 de septiembre de 2020, de http://apuntes-r.blogspot.com/2014/12/bagging-para-mejorar-un-modelo.html
*Heras, J. M. (2020, 18 septiembre). Random Forest (Bosque Aleatorio): combinando árboles. Recuperado 21 de septiembre de 2020, de https://www.iartificial.net/random-forest-bosque-aleatorio/