Proyecto parte 2

María Rodríguez

Limpieza de datos

Librerias a utilizar

```
library(tidyverse)
library(readxl)
library(dplyr)
```

carga de datos

```
C2016 <- read_excel("2016.xlsx")
C2017 <- read_excel("2017.xlsx")
C2018 <- read_excel("2018.xlsx")
C2019 <- read_excel("2019.xlsx")
C2020 <- read_excel("2020.xlsx")
C2021 <- read_excel("2021.xlsx")
C2022 <- read_excel("2022.xlsx")
C2023 <- read_excel("2023.xlsx")
C2024 <- read_excel("2024.xlsx")
```

Se elimina la columna fechaCierreRecepciónOfertas

```
C2016 <- C2016 %>% select(-fechaCierreRecepciónOfertas)
C2017 <- C2017 %>% select(-fechaCierreRecepciónOfertas)
C2018 <- C2018 %>% select(-fechaCierreRecepciónOfertas)
C2019 <- C2019 %>% select(-fechaCierreRecepciónOfertas)
C2020 <- C2020 %>% select(-fechaCierreRecepciónOfertas)
C2021 <- C2021 %>% select(-fechaCierreRecepciónOfertas)
C2022 <- C2022 %>% select(-fechaCierreRecepciónOfertas)
C2023 <- C2023 %>% select(-fechaCierreRecepciónOfertas)
C2024 <- C2024 %>% select(-fechaCierreRecepciónOfertas)
```

Combinar todos los dataframes

```
concursos_g <- bind_rows(C2016, C2017, C2018, C2019, C2020, C2021, C2022, C2023,
                         C2024)
```

Pasar el nombre de las columnas a solo minúsculas

```
colnames(concursos_g) <- tolower(colnames(concursos_g))
```

seleccionar columnas que se van a utilizar

```
concursos <- concursos_g %>%
  select(tipodeentidadpadre,tipoentidad,entidadcompradora,modalidad,nombre,categorías)
```

Eliminar los archivos que ya no se necesitan

```
remove(C2016,C2017,C2018,C2019,C2020,C2021,C2022,C2023,C2024)
```

Convertir todas las columnas no categóricas automáticamente

```
concursos <- concursos%>%
  mutate(across(where(is.character), as.factor))
```

seleccionar un subset

Se realizó un subset con las variables "categorías" y "entidadcompradora" debido a que tenían demasiados niveles y R no lograba procesarlos.

```
concursos_s <- subset(concursos,
                      categorías %in% c("Salud e insumos hospitalarios",
                                        "Construcción y materiales afines",
                                        "Seguros, fianzas y servicios bancarios",
                                        "Alimentos y semillas",
                                        "Transporte, repuestos y combustibles") &
                        entidadcompradora %in% c("MINISTERIO DE SALUD PÚBLICA",
                                                 "MINISTERIO DE FINANZAS PÚBLICAS",
                                                 "INSTITUTO GUATEMALTECO DE SEGURIDAD SOCIAL -IGSS-",
                                                 "INSTITUTO NACIONAL DE ELECTRIFICACIÓN -INDE",
                                                 "MINISTERIO PÚBLICO")
)

concursos_s <- droplevels(concursos_s)
```

Se asigna un número a las categorías seleccionadas para mejorar la vista de las gráficas

```
concursos_s2 <- concursos_s %>%
  mutate(
    categorias_num = case_when(
      categorías == "Salud e insumos hospitalarios" ~ 1,
      categorías == "Construcción y materiales afines" ~ 2,
      categorías == "Seguros, fianzas y servicios bancarios" ~ 3,
      categorías == "Alimentos y semillas" ~ 4,
      categorías == "Transporte, repuestos y combustibles" ~ 5,
      TRUE ~ NA_integer_
    ),
    entidad_compradora_num = case_when(
      entidadcompradora == "MINISTERIO DE SALUD PÚBLICA" ~ 1,
      entidadcompradora == "MINISTERIO DE FINANZAS PÚBLICAS" ~ 2,
      entidadcompradora == "INSTITUTO GUATEMALTECO DE SEGURIDAD SOCIAL -IGSS-" ~ 3,
      entidadcompradora == "INSTITUTO NACIONAL DE ELECTRIFICACIÓN -INDE" ~ 4,
      entidadcompradora == "MINISTERIO PÚBLICO" ~ 5,
      TRUE ~ NA_integer_
    )
  )
```

ARBOL DE DECISIÓN

Librerias a utilizar

```
library(rpart)
library(rpart.plot)
```

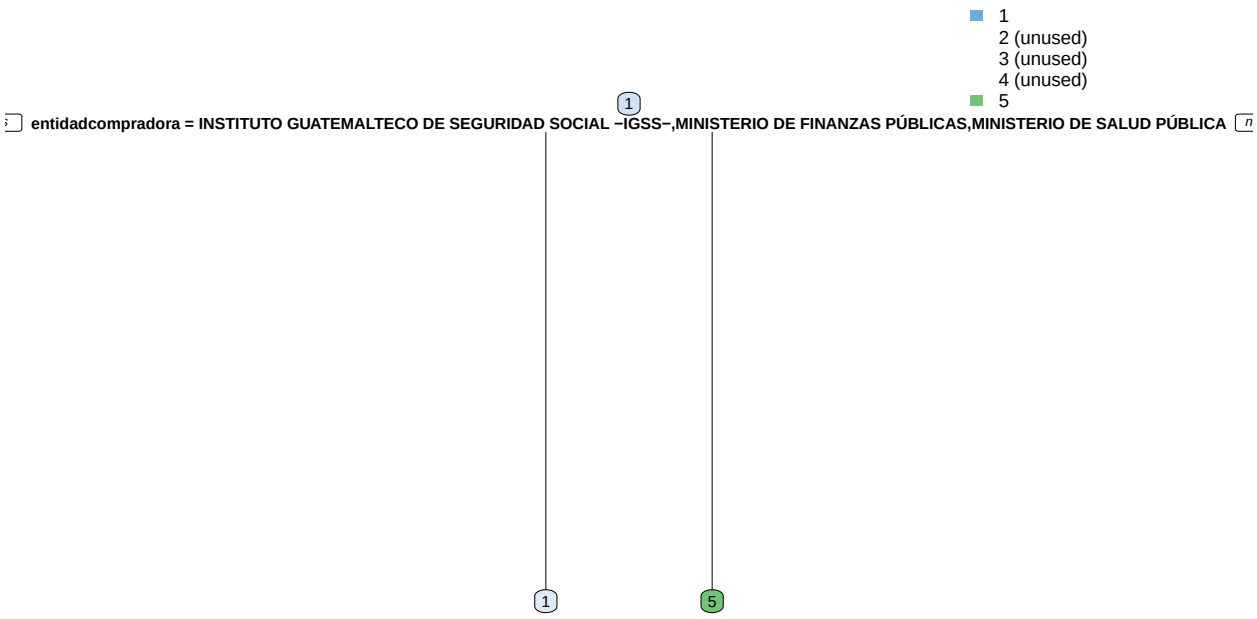
Arbol 1

Se crea el primer árbol, seleccionando la variable a predecir, las variabls predictoras y el dataset

```
arbol1 <- rpart(categorias_num ~ tipodeentidadpadre + tipoentidad + entidadcompradora,
                data = concursos_s2, method = "class")
```

a continuación se genera la gráfica

```
rpart.plot(arbol1, type=2, extra=0, under = TRUE, fallen.leaves = TRUE,
           box.palette = "BuGn", main ="Predicción de categoría", cex = 0.5)
```



se crea un nuevo dataframe que servira para predecir

```
categoria <- data.frame(
  tipodeentidadpadre="Sector Público",
  tipoentidad="Administración Central",
  entidadcompradora="MINISTERIO DE SALUD PÚBLICA"
)
```

para predecir se utiliza el arbol y el dataframe creados

```
result <- predict(arbol1,categoria, type="class")

result
```

```
## 1
## 1
## Levels: 1 2 3 4 5
```

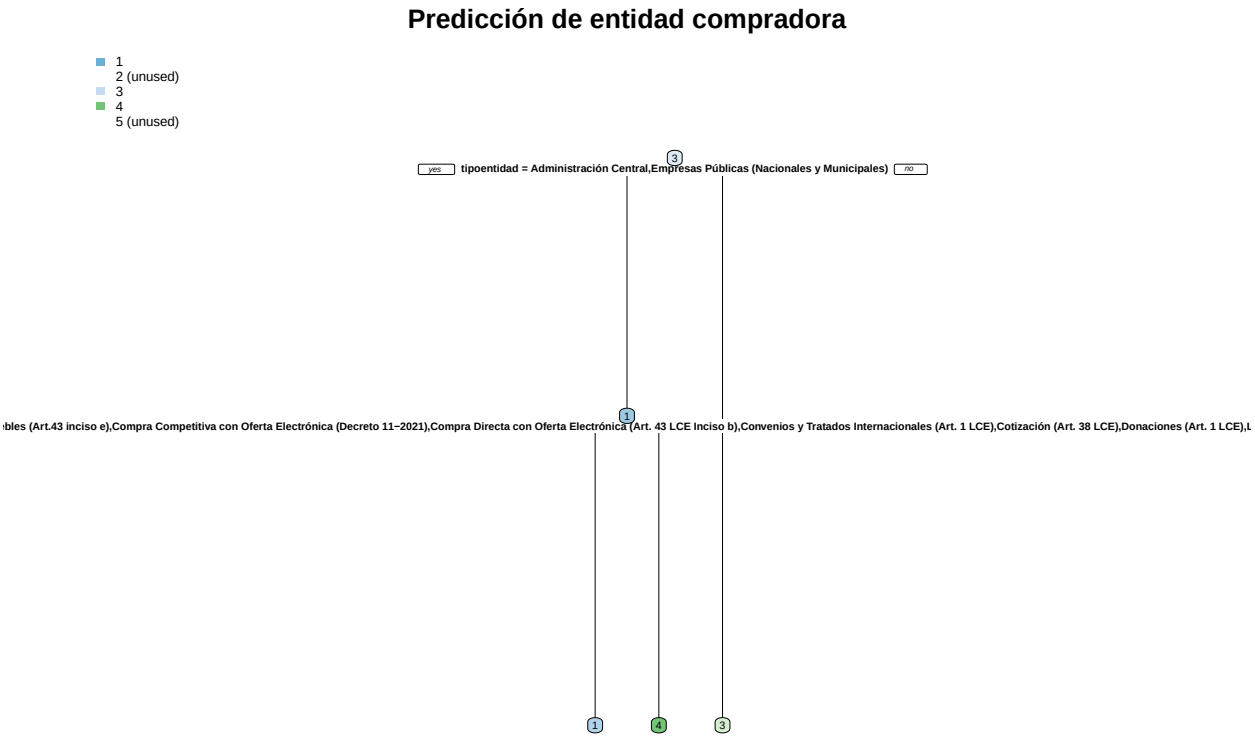
Arbol 2

Se crea el arbol2

```
arbol2 <- rpart(entidad_compradora_num ~ tipodeentidadpadre + tipoentidad +
                  modalidad + categorías, data = concursos_s2, method = "class")
```

Graficar

```
rpart.plot(arbol2, type=2, extra=0, under = TRUE, fallen.leaves = TRUE,
           box.palette = "BuGn",
           main ="Predicción de entidad compradora", cex = 0.35)
```



nuevo dataframe

```
categoria2 <- data.frame(
  tipodeentidadpadre="Sector Público",
  modalidad="Compra Directa con Oferta Electrónica (Art. 43 LCE Inciso b)",
  tipoentidad="Administración Central",
  categorías="Salud e insumos hospitalarios"
)
```

Predecir

```
result2 <- predict(arbol2,categoria2, type="class")

result2
```

```
## 1
## 1
## Levels: 1 2 3 4 5
```

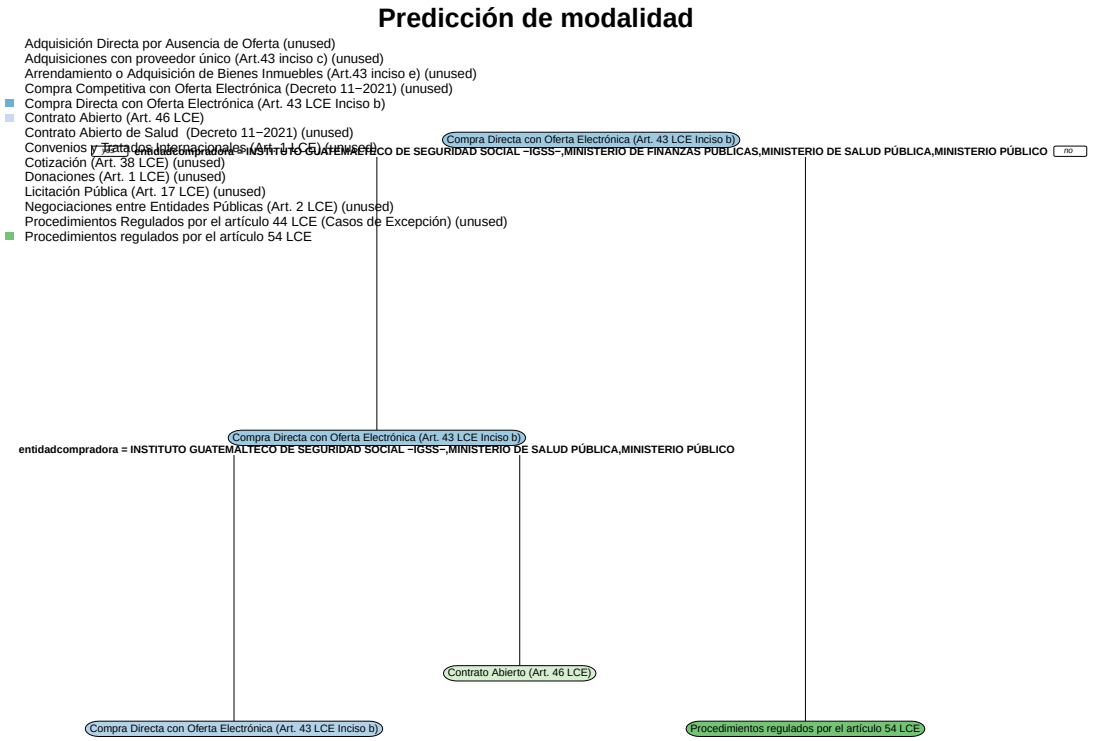
Arbol 3

Se crea el arbol3

```
arbol3 <- rpart(modalidad ~ tipodeentidadpadre + tipoentidad + entidadcompradora + categorías,
                data = concursos_s2, method = "class")
```

Graficar

```
rpart.plot(arbol3, type=2, extra=0, under = TRUE, fallen.leaves = TRUE,
           box.palette = "BuGn",
           main ="Predicción de modalidad", cex = 0.35)
```



Crear dataframe

```
categoria3 <- data.frame(
  tipodeentidadpadre="Sector Público",
  tipoentidad="Administración Central",
  entidadcompradora="MINISTERIO PÚBLICO",
  categorías="Salud e insumos hospitalarios"
)
```

Predecir

```
result3 <- predict(arbol3,categoria3, type="class")

result3
```

```
##                                                                1
## Compra Directa con Oferta Electrónica (Art. 43 LCE Inciso b)
## 14 Levels: Adquisición Directa por Ausencia de Oferta ...
```

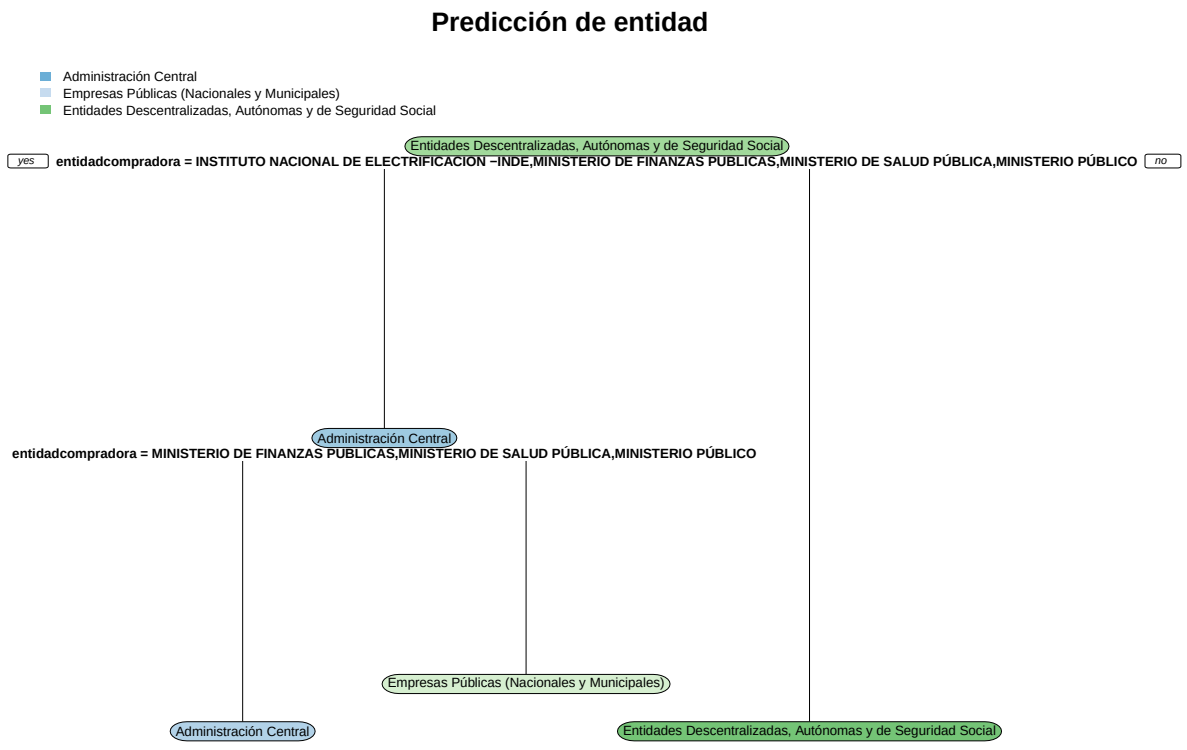
Arbol 4

crear arbol4

```
arbol4 <- rpart( tipoentidad  ~ tipodeentidadpadre + entidadcompradora + modalidad +
                   categorías,
                 data = concursos_s2, method = "class")
```

Graficar

```
rpart.plot(arbol4, type=2, extra=0, under = TRUE, fallen.leaves = TRUE,
           box.palette = "BuGn",
           main ="Predicción de entidad", cex = 0.4)
```



crear nuevo data frame

```
categoria4 <- data.frame(
  tipodeentidadpadre="Sector Público",
  entidadcompradora="INSTITUTO NACIONAL DE ELECTRIFICACIÓN -INDE",
  modalidad="Contrato Abierto (Art. 46 LCE)",
  categorías="Transporte, repuestos y combustibles"
)
```

Predecir

```
result4 <- predict(arbol4,categoria4, type="class")

result4
```

```
##                                           1
## Empresas Públicas (Nacionales y Municipales)
## 3 Levels: Administración Central ...
```

RANDOM FOREST

Libreria a utilizar

```
library(randomForest)
```

Primeros pasos

Establecer la semilla

```
set.seed(200)
```

Reordenar aleatoriamente las filas del dataframe

```
concursos_s <- concursos_s[sample(1:nrow(concursos_s)),]
```

Crear un índice para dividir los datos

```
index <- sample(1:nrow(concursos_s),0.8*nrow(concursos_s))
```

Crear el conjunto de datos de entrenamiento y el conjunto de datos de prueba

```
TRAIN <- concursos_s[index,]
test <- concursos_s[-index,]
```

Random forest 1

Se crea el primer bosque, seleccionando la variable a predecir, las variabls predictoras y el dataset

```
bosque1 <- randomForest(entidadcompradora ~ tipodeentidadpadre + tipoentidad +
                          categorías + modalidad,
                        data = TRAIN,
                        ntree = 30
)
```

Se realiza la prueba

Se crea el dato nuevo

```
dato_nuevo1 <- data.frame(
  tipodeentidadpadre = "Sector Público",
  tipoentidad = "Entidades Descentralizadas, Autónomas y de Seguridad Social",
  categorías = "Salud e insumos hospitalarios",
  modalidad = "Compra Directa con Oferta Electrónica (Art. 43 LCE Inciso b)"
)

str(dato_nuevo1)
```

```
## 'data.frame':    1 obs. of  4 variables:
##  $ tipodeentidadpadre: chr "Sector Público"
##  $ tipoentidad       : chr "Entidades Descentralizadas, Autónomas y de Seguridad Social"
##  $ categorías        : chr "Salud e insumos hospitalarios"
##  $ modalidad         : chr "Compra Directa con Oferta Electrónica (Art. 43 LCE Inciso b)"
```

Convertir las columnas categóricas a factores con los mismos niveles usados en el entrenamiento

```
levels_tipodeentidadpadre <- levels(TRAIN$tipodeentidadpadre)
levels_tipoentidad <- levels(TRAIN$tipoentidad)
levels_categorias <- levels(TRAIN$categorías)
levels_modalidad <- levels(TRAIN$modalidad)

dato_nuevo1$tipodeentidadpadre <- factor(dato_nuevo1$tipodeentidadpadre, levels = levels_tipodeentidadpa
dato_nuevo1$tipoentidad <- factor(dato_nuevo1$tipoentidad, levels = levels_tipoentidad)
dato_nuevo1$categorías <- factor(dato_nuevo1$categorías, levels = levels_categorias)
dato_nuevo1$modalidad <- factor(dato_nuevo1$modalidad, levels = levels_modalidad)
```

Realizar predicción

```
prediccion1 <- predict(bosque1, dato_nuevo1)
prediccion1
```

```
##                                                1
## INSTITUTO GUATEMALTECO DE SEGURIDAD SOCIAL -IGSS-
## 5 Levels: INSTITUTO GUATEMALTECO DE SEGURIDAD SOCIAL -IGSS- ...
```

crear otro dato nuevo

```
dato_nuevo2 <- data.frame(
  tipodeentidadpadre = "Sector Público",
  tipoentidad = "Administración Central",
  categorías = "Salud e insumos hospitalarios",
  modalidad = "Compra Directa con Oferta Electrónica (Art. 43 LCE Inciso b)"
)
```

Convertir las columnas categóricas a factores con los mismos niveles usados en el entrenamiento

```
levels_tipodeentidadpadre <- levels(TRAIN$tipodeentidadpadre)
levels_tipoentidad <- levels(TRAIN$tipoentidad)
levels_categorias <- levels(TRAIN$categorías)
levels_modalidad <- levels(TRAIN$modalidad)

dato_nuevo2$tipodeentidadpadre <- factor(dato_nuevo2$tipodeentidadpadre, levels = levels_tipodeentidadpa
dato_nuevo2$tipoentidad <- factor(dato_nuevo2$tipoentidad, levels = levels_tipoentidad)
dato_nuevo2$categorías <- factor(dato_nuevo2$categorías, levels = levels_categorias)
dato_nuevo2$modalidad <- factor(dato_nuevo2$modalidad, levels = levels_modalidad)
```

Predicción

Random forest 2

Crear segundo bosque

```
bosque2 <- randomForest(tipoentidad  ~ tipodeentidadpadre + entidadcompradora +
                          modalidad + categorías,
                        data = TRAIN,
                        ntree = 30
)
```

Realizar prueba

Crear dato nuevo

```
dato_nuevo3 <- data.frame(
  tipodeentidadpadre = "Sector Público",
  entidadcompradora = "INSTITUTO NACIONAL DE ELECTRIFICACIÓN -INDE",
  categorías = "Alimentos y semillas",
  modalidad = "Compra Directa con Oferta Electrónica (Art. 43 LCE Inciso b)"
)
```

Convertir las columnas categóricas a factores con los mismos niveles usados en el entrenamiento

```
levels_tipodeentidadpadre <- levels(TRAIN$tipodeentidadpadre)
levels_entidadcompradora <- levels(TRAIN$entidadcompradora)
levels_categorias <- levels(TRAIN$categorías)
levels_modalidad <- levels(TRAIN$modalidad)

dato_nuevo3$tipodeentidadpadre <- factor(dato_nuevo3$tipodeentidadpadre, levels = levels_tipodeentidadpa
dato_nuevo3$entidadcompradora <- factor(dato_nuevo3$entidadcompradora,
                                        levels = levels_entidadcompradora)
dato_nuevo3$categorías <- factor(dato_nuevo3$categorías, levels = levels_categorias)
dato_nuevo3$modalidad <- factor(dato_nuevo3$modalidad, levels = levels_modalidad)
```

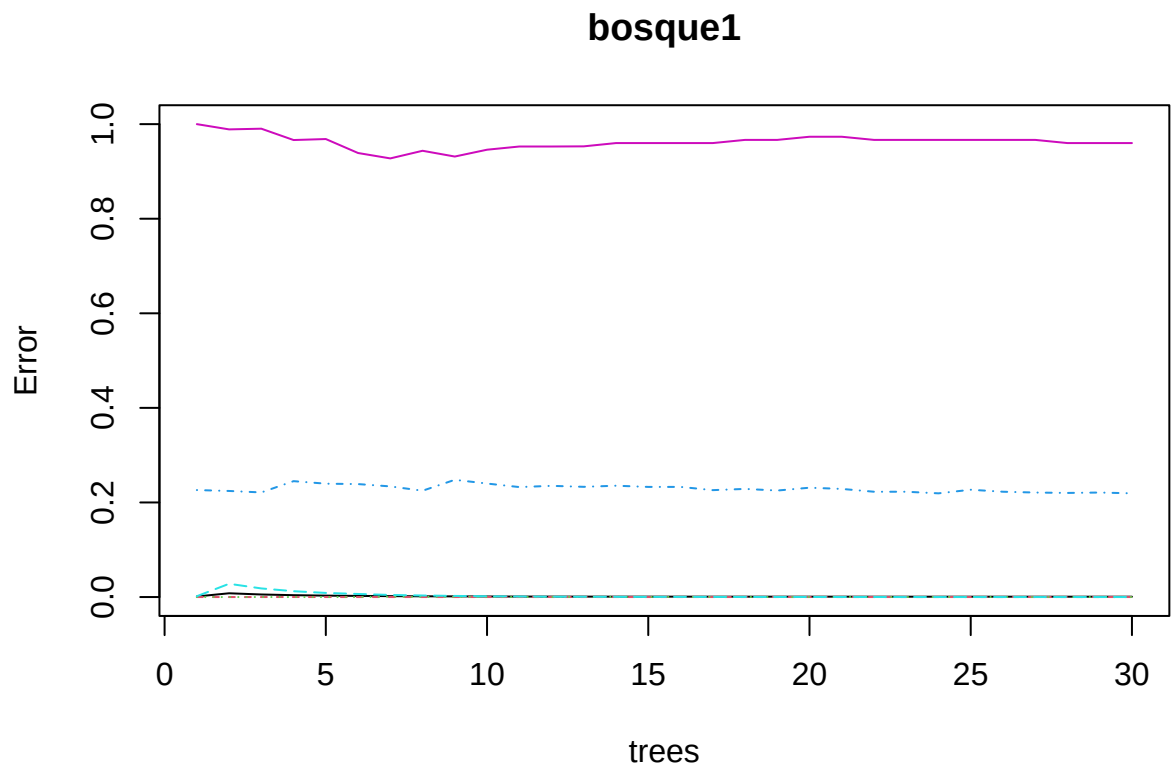
Predicción

```r
prediccion3 <- predict(bosque2, dato_nuevo3)
prediccion3
```

```
##                                           1
## Empresas Públicas (Nacionales y Municipales)
## 3 Levels: Administración Central ...
```

Gráfico

```r
plot(bosque1)
```



Creando otro dato nuevo

```r
dato_nuevo4 <- data.frame(
  tipodeentidadpadre = "Sector Público",
  entidadcompradora = "MINISTERIO DE FINANZAS PÚBLICAS",
  categorías = "Alimentos y semillas",
  modalidad = "Compra Directa con Oferta Electrónica (Art. 43 LCE Inciso b)"
)
```

Convertir las columnas categóricas a factores con los mismos niveles usados en el entrenamiento

```r
levels_tipodeentidadpadre <- levels(TRAIN$tipodeentidadpadre)
levels_entidadcompradora <- levels(TRAIN$entidadcompradora)
```

```
levels_categorias <- levels(TRAIN$categorías)
levels_modalidad <- levels(TRAIN$modalidad)

dato_nuevo4$tipodeentidadpadre <- factor(dato_nuevo4$tipodeentidadpadre, levels = levels_tipodeentidadpa
dato_nuevo4$entidadcompradora <- factor(dato_nuevo4$entidadcompradora,
                                        levels = levels_entidadcompradora)
dato_nuevo4$categorías <- factor(dato_nuevo4$categorías, levels = levels_categorias)
dato_nuevo4$modalidad <- factor(dato_nuevo4$modalidad, levels = levels_modalidad)
```

Predicción

```
prediccion4 <- predict(bosque2, dato_nuevo4)
prediccion4
```

```
##                       1
## Administración Central
## 3 Levels: Administración Central ...
```

Gráfico

```
plot(bosque2)
```