Experimento 1

Tomas Glauberman, Natalia Massri y Juan Silvestri

Experimento 1

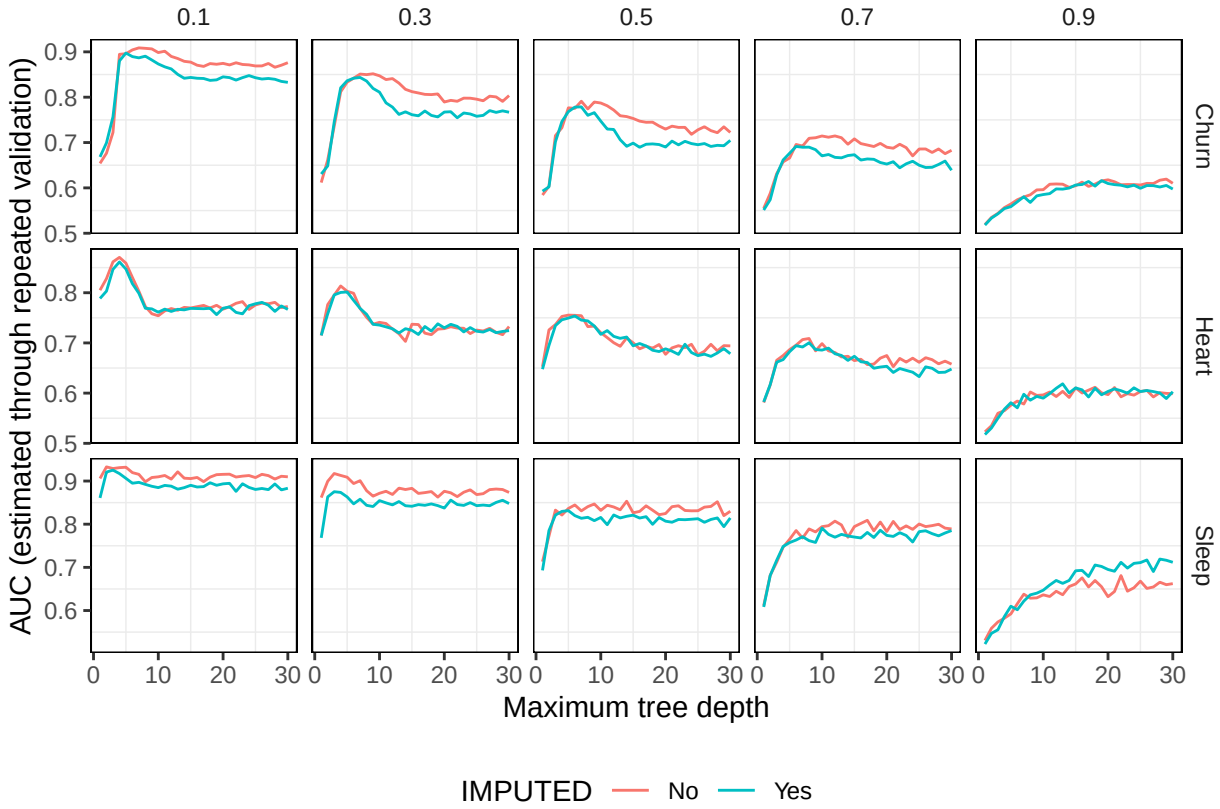
Se realizó un experimento en el que se analiza el impacto que tienen los datos faltantes en 3 datasets distintos para un modelo de Árbol de decisión. Para ello, se entrenaron múltiples modelos para cada uno de los datasets, lo que nos permite estimar y evaluar la performance.

Las variables independientes del experimento son:

- Profundidad del árbol (1-30)
- Proporción de datos faltantes (0-1)
- Valores faltantes imputados o no

En la siguiente figura se puede observar un resumen de los resultados.

```
source("exp_1.R")
```



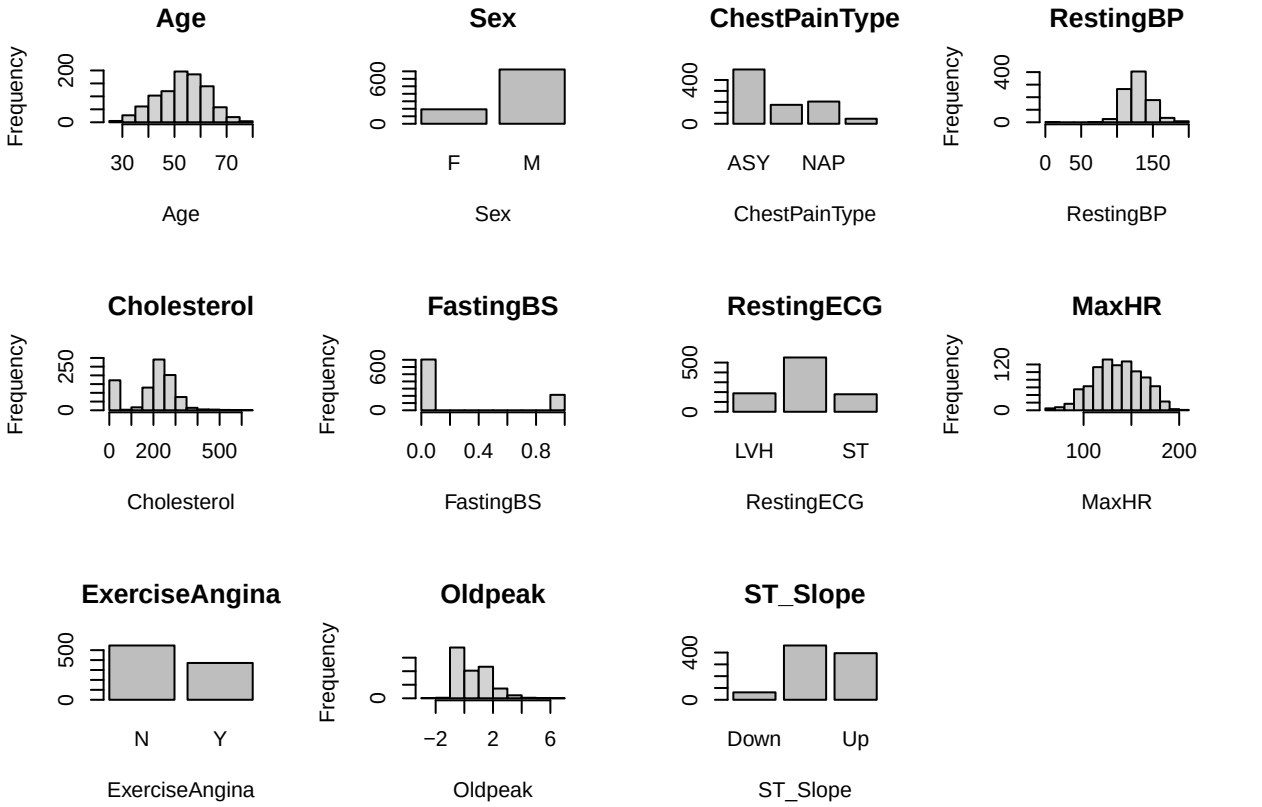
Se observa que a medida que aumenta la proporción de NAs la performance de los modelos disminuye. Esto ocurre ya que cuando hay una proporción significativa de NAs, tendremos una menor cantidad de datos reales

para entrenar los modelos, lo cual puede generar dificultades para capturar patrones y relaciones importantes en los datos reales y, por lo tanto, la predicción de datos nuevos será menos efectiva.

Por otro lado, en los datasets de Churn y Sleep Dissorder la performance sin imputar los datos faltantes es mejor que al imputar con la media. Creemos que esto se debe al eficiente manejo de los datos faltantes por parte de los árboles con el método de surrogate variables frente a los simples imputs que hicimos con la media y moda. En el dataset de Heart Disease, la performance imputando y sin imputar es similar. Creemos que esto ocurre ya que, como podemos ver a continuación, la distribución de las variables se centra en la media, por lo tanto, al tomar observaciones al azar para reemplazar con NAs e imputarlas con la media o moda (dependiendo de si es númerica o categórica), es problable que ese valor esté cerca del real por lo que no se ve una gran diferencia.

```
heart = read.table("./data/heart.csv", header = TRUE, sep=",", na.strings="",
                   stringsAsFactors=TRUE)

par(mfrow = c(3, 4))
for (col in names(heart)) {
  if(col != "HeartDisease"){
    if (is.numeric(heart[[col]])) {
      hist(heart[[col]], main = paste(col), xlab = col)
    }else{
      barplot(table(heart[[col]]), main = paste(col), xlab = col)
    }
  }
}
par(mfrow = c(1, 1))
```



Como podemos observar, en la mayoría de los casos, mayor profundidad en los árboles no parece mejorar la preformance. Esto ocurre porque los modelos son demasiado flexibles para la cantidad de datos o complejidad de las relaciones entre ellos, por lo que se ajustan por demás a los datos de entrenamiento (y su ruido),

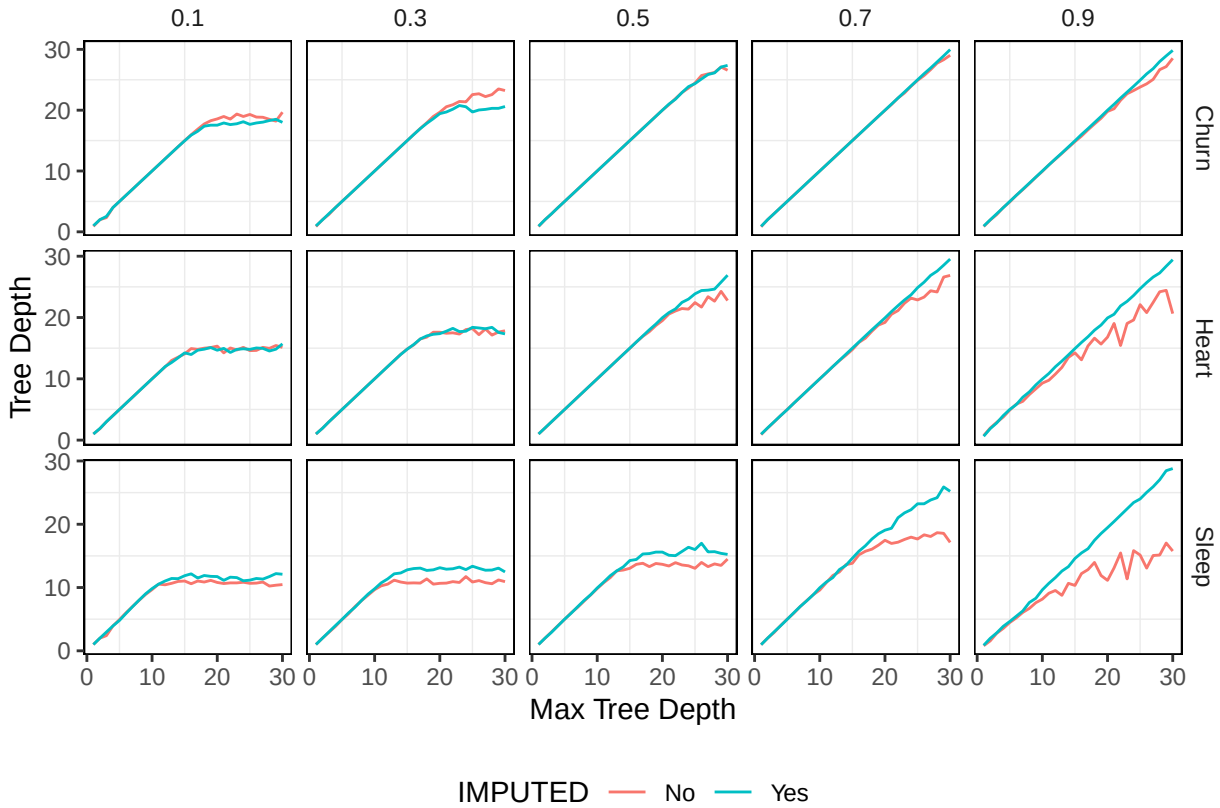
haciendo así que su performance no sea muy buena con datos nuevos. Sin embargo, cuando la proporción de datos faltantes es muy alta, el comportamiento es distinto. Suponemos que puede ocurrir debido a que no hay suficientes datos reales con los cuales se hace overfitting.

```
raw_data = read.table("./outputs/tables/exp_1.txt", header = TRUE)

# Calculate mean AUC values for different groups of experimental results
data_for_plot <- raw_data %>%
  group_by(dataset_name, prop_NAs, IMPUTED, maxdepth) %>%
  summarize(mean_depth=mean(tree_depth), .groups='drop')

# Create a ggplot object for the line plot
g <- ggplot(data_for_plot, aes(x=maxdepth, y=mean_depth, color=IMPUTED)) +
  geom_line() +
  theme_bw() +
  xlab("Max Tree Depth") +
  ylab("Tree Depth") +
  facet_grid(dataset_name ~ prop_NAs, scales="free_y") +
  theme(legend.position="bottom",
        panel.grid.major=element_blank(),
        strip.background=element_blank(),
        panel.border=element_rect(colour="black", fill=NA))

print(g)
```



Nos pareció interesante observar cómo varia la profundidad de los árboles dada la profundidad máxima, para los diferentes porp_NAs (imputados o no) en cada dataset y ver si encontrábamos alguna tendencia. Podemos notar que para proporciones bajas de NAs, como lo son 0.1 y 0.3, hay un momento en el que la profundidad real a la que llega el árbol se mantiene constante, y por lo tanto, la profundidad maxima no

influye. Sin embargo, cuando empieza a haber muchos missing en los datos (es decir, mayor proporción de NAs), la profundidad del arbol sigue creciendo, ya que necesita fijarse en variables que cuando hay muchos datos reales quizas no debería fijarse, pero al haber menos datos tiene en cuenta en cosas más específicas.