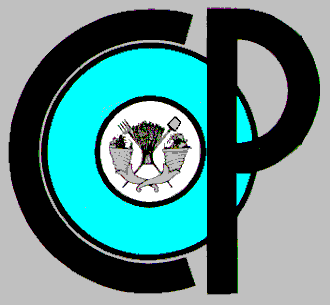
**COLEGIO DE POSTGRADUADOS**

**INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN EN CIENCIAS AGRÍCOLAS**

**CAMPUS MONTECILLO**

**COA-501** **HERRAMIENTAS DE CÓMPUTO PARA INVESTIGADORES**

**(ANÁLISIS ESTADÍSTICO DE DATOS)**



**"Análisis de clasificación de la transparencia de copa en *Pseudotsuga menziesii* con técnicas de aprendizaje automatizado"**

**PRESENTA:** IVÁN FERMÍN QUIROZ IBÁÑEZ

**PROFESOR:** PhD. JUAN MANUEL GONZÁLEZ CAMACHO

**MONTECILLO, TEXCOCO, EDO. DE MÉXICO**

20 de noviembre de 2022

OTOÑO 2022

**Contenido**

[1. Introducción 3](#_Toc119858191)

[2. Objetivo del proyecto de análisis de datos 3](#_Toc119858192)

[3. Planteamiento del problema 4](#_Toc119858193)

[4. Solución propuesta 4](#_Toc119858194)

[4.1. Análisis exploratorio de datos 4](#_Toc119858195)

[4.2. Random Forest 9](#_Toc119858196)

[4.3. Naive Bayes 13](#_Toc119858197)

[4.4. K-NN 13](#_Toc119858198)

[4.5. Curvas Receiver Operating Characteristic (ROC) 15](#_Toc119858199)

[5. Conclusiones 16](#_Toc119858200)

[6. Bibliografía 16](#_Toc119858201)

**Figuras**

[Figura 1. Matriz de diagramas de dispersión de la base de datos original. 7](#_Toc119794915)

[Figura 2. PCA1 vs PCA2 de la base original por Edad del follaje colectado. 8](#_Toc119794916)

[Figura 3. Gráfica de sedimentación y PCA1 vs PCA2 por transparencia de copa. 9](#_Toc119794917)

[Figura 4. Gráficas del error por el número de árboles utilizados con la técnica de RF. 11](#_Toc119794918)

[Figura 5. Mtry vs error observado con la técnica de RF. 11](#_Toc119794919)

[Figura 6. Histograma del Número de nodos por árboles con la técnica de RF. 12](#_Toc119794920)

[Figura 7. Relación de la importancia de las variables utilizadas en la técnica de RF. 12](#_Toc119794921)

[Figura 8. Curvas ROC para la técnica de Random Forest, Naive Bayes y K-NN. 15](#_Toc119794922)

**Cuadros**

[Cuadro 1. Métricas utilizadas para comparar las técnicas de ML. 15](#_Toc119794478)

# Introducción

Las plantaciones de árboles de navidad con *Pseudotsuga menziesii* (Mirb.) Franco ocupan actualmente 10 % de la superficie total de este tipo de plantaciones en México. El tizón suizo es una enfermedad propia de este género, originada por el hongo *Phaeocryptopus gaeumannii* (T. Rohde) Petrak. Los principales síntomas incluyen follaje clorótico y abscisión prematura de acículas infectadas.

Se tomaron los datos de la tesis de posgrado en ciencias forestales de Quiroz (2019), donde el objetivo fue determinar la relación e impacto del tizón suizo con la transparencia de copa, evaluar el efecto de aplicaciones de fungicidas químicos y fertilizantes inorgánicos en el nivel de variación del grado y dinámica temporal de infestación del tizón suizo y evaluar la relación entre la fertilización química y el estado nutrimental de los árboles con los niveles de severidad e incidencia de la enfermedad. El estudio se estableció en el predio “Reserva Multifuncional El Manantial” en Aquixtla, Puebla. Los tratamientos aplicados fueron fertilización con urea y sulfato de potasio y control químico con propiconazol y prochloraz. Se utilizó un diseño experimental factorial 32 con un arreglo en bloques al azar generalizado tomando como factor de confusión o bloqueo el nivel visual de transparencia de copa asociado al tizón suizo.

# Objetivo del proyecto de análisis de datos

Realizar el análisis exploratorio de datos y aplicar las técnicas de aprendizaje automatizado Naive bayes, K-NN y Random forest para generar un modelo de clasificación de la transparencia de copa de árboles de *Pseudotsuga menziesii* asociada con la enfermedad conocida como tizón suizo.

# Planteamiento del problema

Derivado de un experimento realizado en campo, se visualiza la posibilidad de que, con variables asociadas al tizón suizo, se pueda generar algún modelo útil para predecir la categoría de transparencia de copa en *P. menziesii*.

# Solución propuesta

Se realizó el análisis exploratorio de datos, con el fin de determinar que variable se incluirían en los modelos de ML (i.e. transparencia de copa, severidad, incidencia, longitud de brotes, área de acículas, etc). A continuación, se incluyen los resultados más relevantes de la implementación de Naive Bayes, K-NN y Random Forest para clasificar la transparencia de copa en Douglas-fir.

## Análisis exploratorio de datos

#base de datos  
summary(base)

## Arbol Bloque Longitud Latitud X\_UTM   
## 1 : 12 Alta :216 Min. :-97.99 Min. :19.72 Min. :605708   
## 2 : 12 Baja :216 1st Qu.:-97.99 1st Qu.:19.72 1st Qu.:605724   
## 3 : 12 Media:216 Median :-97.99 Median :19.72 Median :605732   
## 4 : 12 Mean :-97.99 Mean :19.72 Mean :605732   
## 5 : 12 3rd Qu.:-97.99 3rd Qu.:19.72 3rd Qu.:605741   
## 6 : 12 Max. :-97.99 Max. :19.72 Max. :605759   
## (Other):576   
## Y\_UTM Altitud AltRan Fungicida Nutrimento Rep   
## Min. :2180985 Min. :2973 Alta:312 1:216 1:216 Min. :1.0   
## 1st Qu.:2181009 1st Qu.:2979 Baja:336 2:216 2:216 1st Qu.:1.0   
## Median :2181025 Median :2982 3:216 3:216 Median :1.5   
## Mean :2181027 Mean :2985 Mean :1.5   
## 3rd Qu.:2181039 3rd Qu.:2991 3rd Qu.:2.0   
## Max. :2181087 Max. :2994 Max. :2.0   
##   
## Muestreo DDA Anio AcicR Abs   
## 1:108 0 :108 2016:324 Min. : 30.00 Min. : 0.00   
## 2:108 63 :108 2017:324 1st Qu.: 83.75 1st Qu.: 0.00   
## 3:108 91 :108 Median :101.00 Median : 2.00   
## 4:108 119:108 Mean :103.14 Mean : 6.91   
## 5:108 147:108 3rd Qu.:121.00 3rd Qu.: 8.00   
## 6:108 175:108 Max. :225.00 Max. :104.00   
## NA's :24 NA's :24   
## TotalAc Inc Sevmed Sevmin   
## Min. : 36.0 Min. :0.0000 Min. :0.00000 Min. :0.00000   
## 1st Qu.: 91.0 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.00000   
## Median :108.0 Median :0.2667 Median :0.01483 Median :0.00000   
## Mean :110.1 Mean :0.4584 Mean :0.12498 Mean :0.06102   
## 3rd Qu.:128.0 3rd Qu.:0.9667 3rd Qu.:0.24258 3rd Qu.:0.10333   
## Max. :225.0 Max. :1.0000 Max. :0.61900 Max. :0.50667   
## NA's :24 NA's :24 NA's :24 NA's :24   
## Sevmax LonBrot indcol Afmed   
## Min. :0.00000 Min. : 3.300 Min. : 0.00 Min. :0.1322   
## 1st Qu.:0.00000 1st Qu.: 7.000 1st Qu.: 0.00 1st Qu.:0.2610   
## Median :0.07333 Median : 8.450 Median : 0.46 Median :0.3096   
## Mean :0.20214 Mean : 8.629 Mean :11.72 Mean :0.3117   
## 3rd Qu.:0.40083 3rd Qu.: 9.963 3rd Qu.:22.85 3rd Qu.:0.3577   
## Max. :0.90667 Max. :23.400 Max. :61.90 Max. :0.6133   
## NA's :24 NA's :24 NA's :24 NA's :24   
## Afmax Afmin Aftotal CA   
## Min. :0.152 Min. :0.0910 Min. : 3.966 Min. : 0.000   
## 1st Qu.:0.300 1st Qu.:0.2087 1st Qu.: 7.831 1st Qu.: 0.000   
## Median :0.352 Median :0.2470 Median : 9.287 Median : 1.646   
## Mean :0.358 Mean :0.2486 Mean : 9.350 Mean : 6.178   
## 3rd Qu.:0.411 3rd Qu.:0.2870 3rd Qu.:10.730 3rd Qu.: 7.692   
## Max. :0.849 Max. :0.4440 Max. :18.398 Max. :64.662   
## NA's :24 NA's :24 NA's :24 NA's :24   
## RA Color trat ABCPEIC   
## Min. : 35.34 Min. :1.000 11 : 72 Min. : 0.0000   
## 1st Qu.: 92.31 1st Qu.:1.000 12 : 72 1st Qu.: 0.0000   
## Median : 98.35 Median :2.000 13 : 72 Median : 0.1925   
## Mean : 93.82 Mean :1.694 21 : 72 Mean : 326.8269   
## 3rd Qu.:100.00 3rd Qu.:2.000 22 : 72 3rd Qu.: 683.7950   
## Max. :100.00 Max. :3.000 23 : 72 Max. :2850.9250   
## NA's :24 NA's :24 (Other):216

class(base)

## [1] "data.frame"

str(base)

## 'data.frame': 648 obs. of 32 variables:  
## $ Arbol : Factor w/ 54 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ Bloque : Factor w/ 3 levels "Alta","Baja",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Longitud : num -98 -98 -98 -98 -98 ...  
## $ Latitud : num 19.7 19.7 19.7 19.7 19.7 ...  
## $ X\_UTM : num 605748 605744 605740 605754 605742 ...  
## $ Y\_UTM : num 2181070 2181057 2181039 2181034 2181026 ...  
## $ Altitud : num 2979 2979 2980 2980 2984 ...  
## $ AltRan : Factor w/ 2 levels "Alta","Baja": 2 2 2 2 2 1 1 1 1 1 ...  
## $ Fungicida : Factor w/ 3 levels "1","2","3": 1 1 2 1 2 1 1 2 2 3 ...  
## $ Nutrimento: Factor w/ 3 levels "1","2","3": 1 2 2 2 1 3 3 3 2 1 ...  
## $ Rep : int 1 1 1 2 1 1 2 1 2 1 ...  
## $ Muestreo : Factor w/ 6 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ DDA : Factor w/ 6 levels "0","63","91",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Anio : Factor w/ 2 levels "2016","2017": 1 1 1 1 1 1 1 1 1 1 ...  
## $ AcicR : int 74 129 69 33 98 71 68 44 92 92 ...  
## $ Abs : int 25 2 2 6 20 15 3 5 4 15 ...  
## $ TotalAc : int 99 131 71 39 118 86 71 49 96 107 ...  
## $ Inc : num 0.9 1 0.767 1 1 ...  
## $ Sevmed : num 0.15 0.331 0.347 0.351 0.384 ...  
## $ Sevmin : num 0 0.137 0 0.21 0.197 ...  
## $ Sevmax : num 0.4 0.667 0.583 0.573 0.56 ...  
## $ LonBrot : num 7.25 12.9 5.1 3.9 9.25 5 6.8 4.5 7.7 9.55 ...  
## $ indcol : num 13.5 33.1 26.6 35.1 38.4 ...  
## $ Afmed : num 0.239 0.374 0.196 0.37 0.362 ...  
## $ Afmax : num 0.275 0.436 0.217 0.411 0.482 ...  
## $ Afmin : num 0.194 0.308 0.163 0.286 0.29 ...  
## $ Aftotal : num 7.18 11.23 5.88 11.1 10.87 ...  
## $ CA : num 25.25 1.53 2.82 15.38 16.95 ...  
## $ RA : num 74.7 98.5 97.2 84.6 83.1 ...  
## $ Color : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ trat : Factor w/ 9 levels "11","12","13",..: 1 2 5 2 4 3 3 6 5 7 ...  
## $ ABCPEIC : num 0 0 0 0 0 0 0 0 0 0 ...

#matriz de diagramas de dispersion  
psych::pairs.panels(base[,c(5,6,7,18,19,22,24,28:30)])

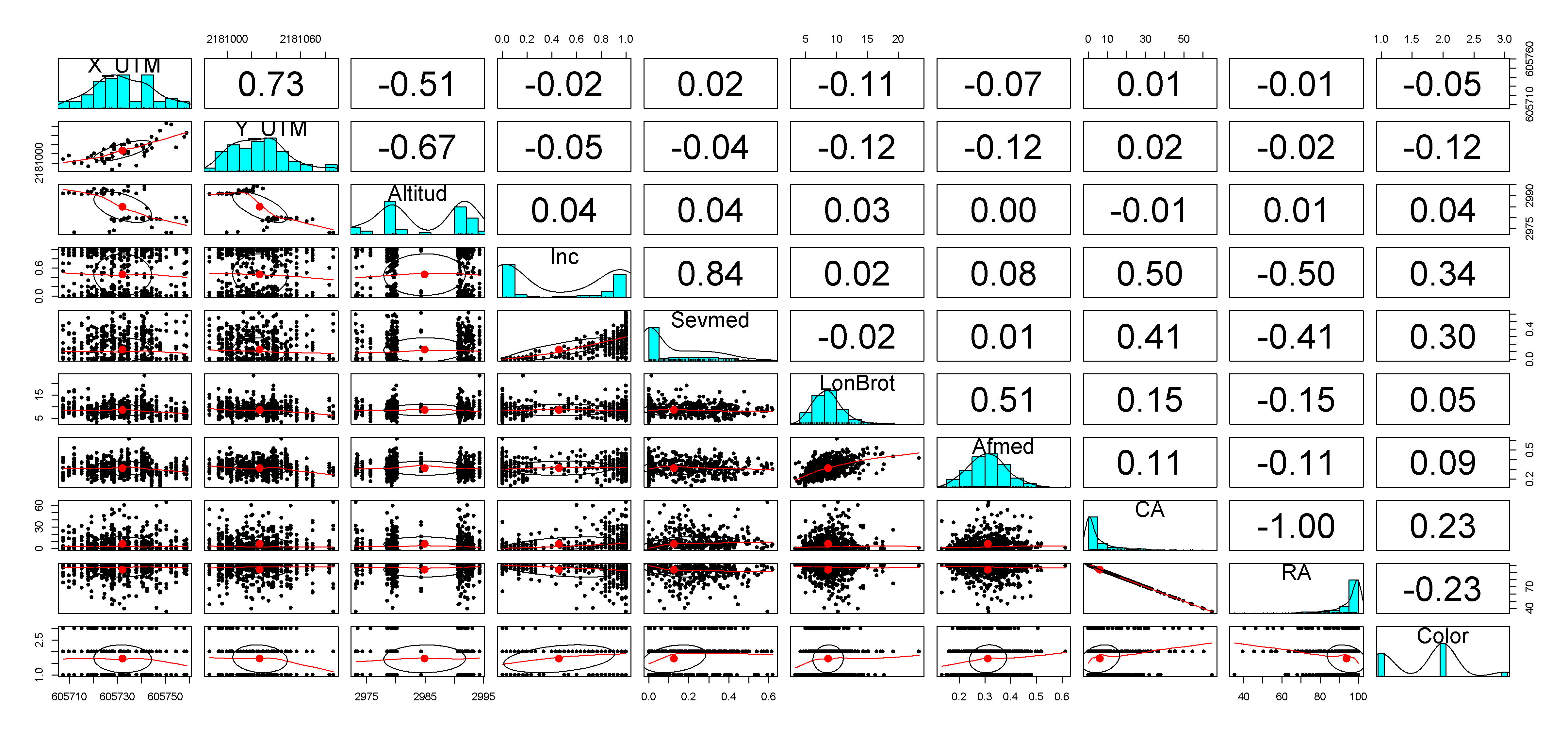


Figura 1. Matriz de diagramas de dispersión de la base de datos original.

base.pc<- na.omit(base[,c(14,2,5,6,7,18,19,22,24,28:30)])  
  
#Análisis de componentes principales   
pca <- prcomp(base.pc[,-c(1,2)], scale = T, center=T)  
summary(pca)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Standard deviation 1.7451 1.5196 1.2216 1.0107 0.88496 0.70241 0.69614  
## Proportion of Variance 0.3045 0.2309 0.1492 0.1021 0.07832 0.04934 0.04846  
## Cumulative Proportion 0.3045 0.5354 0.6847 0.7868 0.86512 0.91446 0.96292  
## PC8 PC9 PC10  
## Standard deviation 0.46896 0.38840 1.265e-10  
## Proportion of Variance 0.02199 0.01509 0.000e+00  
## Cumulative Proportion 0.98491 1.00000 1.000e+00

#biplot  
library(ggplot2)  
Edad <- as.factor(base.pc$Anio)  
plot4 <- ggplot(base.pc, aes(x = pca$x[,1], y = pca$x[,2], colour = Edad)) +  
 geom\_point(size=3) + xlab("PC1") + ylab("PC2")+  
 ggtitle("PCA por Edad del follaje")  
plot4

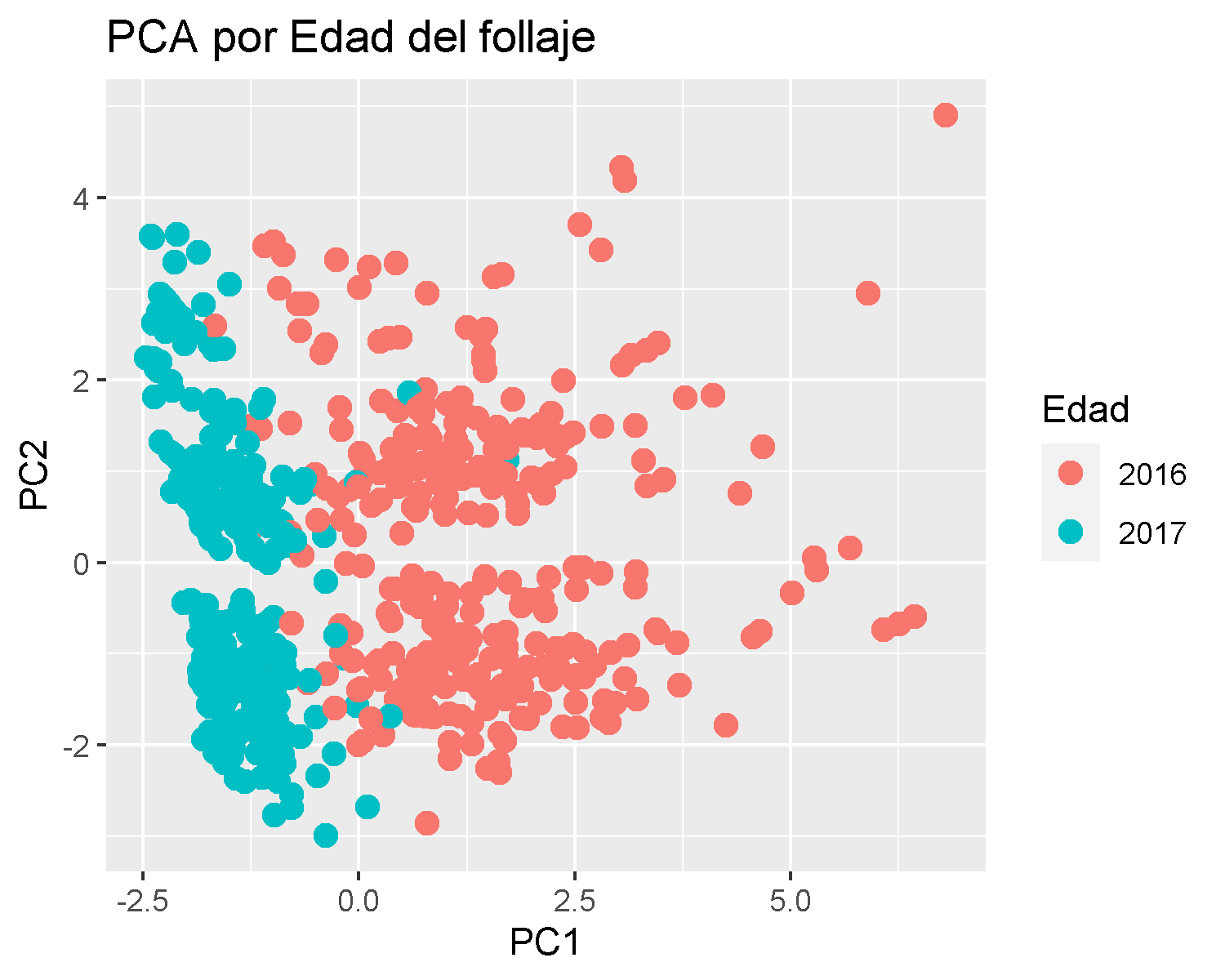


Figura 2. PCA1 vs PCA2 de la base original por Edad del follaje colectado.

par(mfrow = c(1,2))  
plot(pca, main="Sedimentación")  
  
colores <- function(vec){  
 # la funci?n rainbow() devuelve un vector que contiene el n?mero de colores distintos  
 col <- rainbow(length(unique(vec)))  
 return(col[as.numeric(as.factor(vec))])  
}  
  
plot(pca$x[,c(1, 2)], col = colores(base.pc$Bloque),  
 pch = 19,   
 xlab = "PC1",   
 ylab = "PC2", main="PCA por Transparencia")

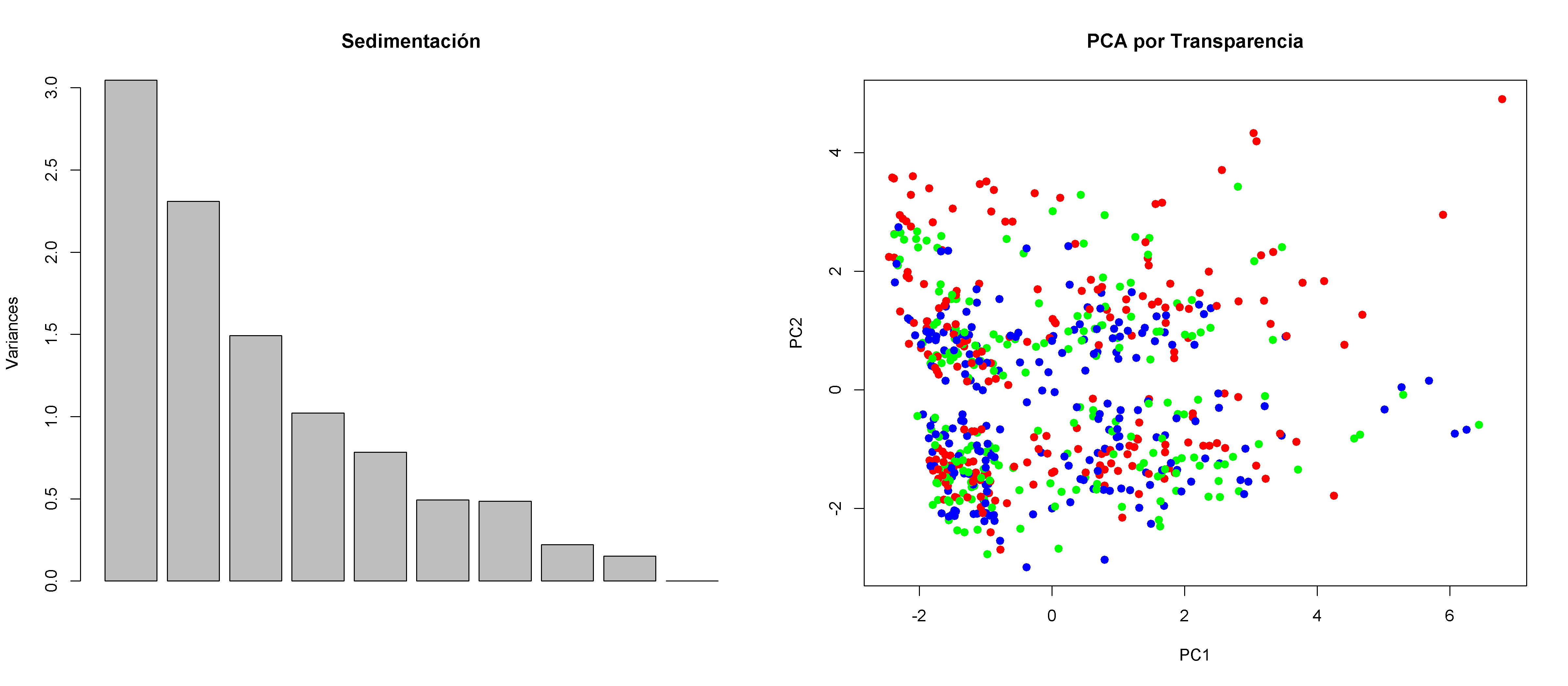


Figura 3. Gráfica de sedimentación y PCA1 vs PCA2 por transparencia de copa.

**Interpretación:** De acuerdo con el AED y ACP se seleccionaron las variables transparencia de copa, coordenadas UTM (X y Y), altitud, incidencia, severidad, longitud de brote, área de acícula, acículas caídas, acículas retenidas y color de acícula, para implementar las técnicas de ML.

## Random Forest

library(randomForest)  
library(caret)  
  
#Getting Data  
base\_rf <- na.omit(base[,c(2,5,6,7,18,19,22,24,28:30)])  
  
#Data Partition  
set.seed(123)  
ind <- sample(2, nrow(base\_rf), replace = TRUE, prob = c(0.7, 0.3))  
train <- base\_rf[ind==1,]  
test <- base\_rf[ind==2,]  
  
  
  
p2 <- predict(rf, test)  
confusionMatrix(p2, test$Bloque)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Alta Baja Media  
## Alta 60 0 0  
## Baja 2 56 0  
## Media 1 0 67  
##   
## Overall Statistics  
##   
## Accuracy : 0.9839   
## 95% CI : (0.9536, 0.9967)  
## No Information Rate : 0.3602   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.9758   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: Alta Class: Baja Class: Media  
## Sensitivity 0.9524 1.0000 1.0000  
## Specificity 1.0000 0.9846 0.9916  
## Pos Pred Value 1.0000 0.9655 0.9853  
## Neg Pred Value 0.9762 1.0000 1.0000  
## Prevalence 0.3387 0.3011 0.3602  
## Detection Rate 0.3226 0.3011 0.3602  
## Detection Prevalence 0.3226 0.3118 0.3656  
## Balanced Accuracy 0.9762 0.9923 0.9958

(tab2 <- table(p2, test$Bloque))

##   
## p2 Alta Baja Media  
## Alta 60 0 0  
## Baja 2 56 0  
## Media 1 0 67

1 - sum(diag(tab2)) / sum(tab2)

## [1] 0.01612903

#error del 0%

#Error rate of Random Forest  
plot(rf)

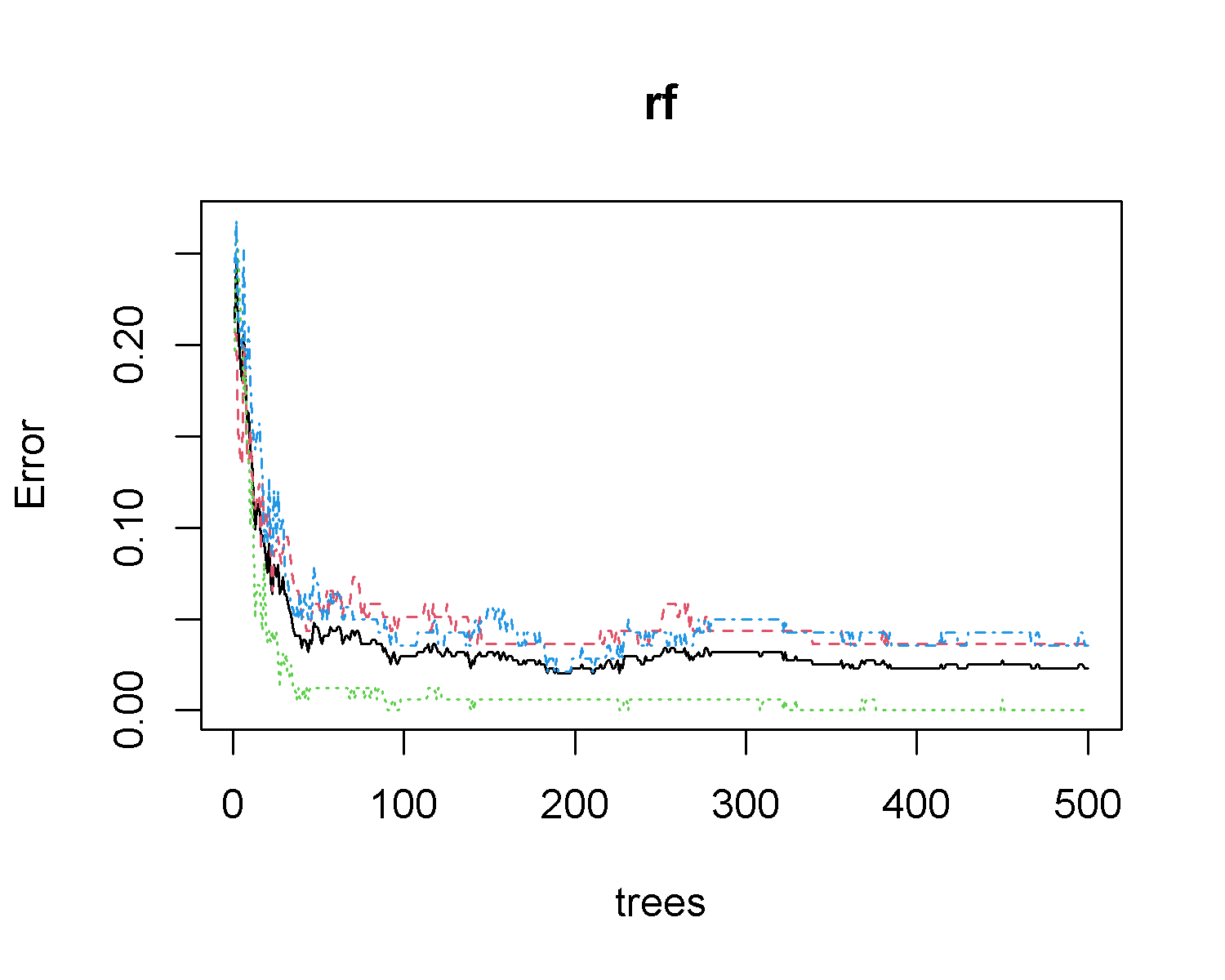


Figura 4. Gráficas del error por el número de árboles utilizados con la técnica de RF.

#Tune mtry (Número de variables aleatorias utilizadas en cada árbol)  
t <- tuneRF(train[,-1], train[,1],  
 stepFactor = 0.5,  
 plot = TRUE,  
 ntreeTry = 5,  
 trace = TRUE,  
 improve = 0.05)

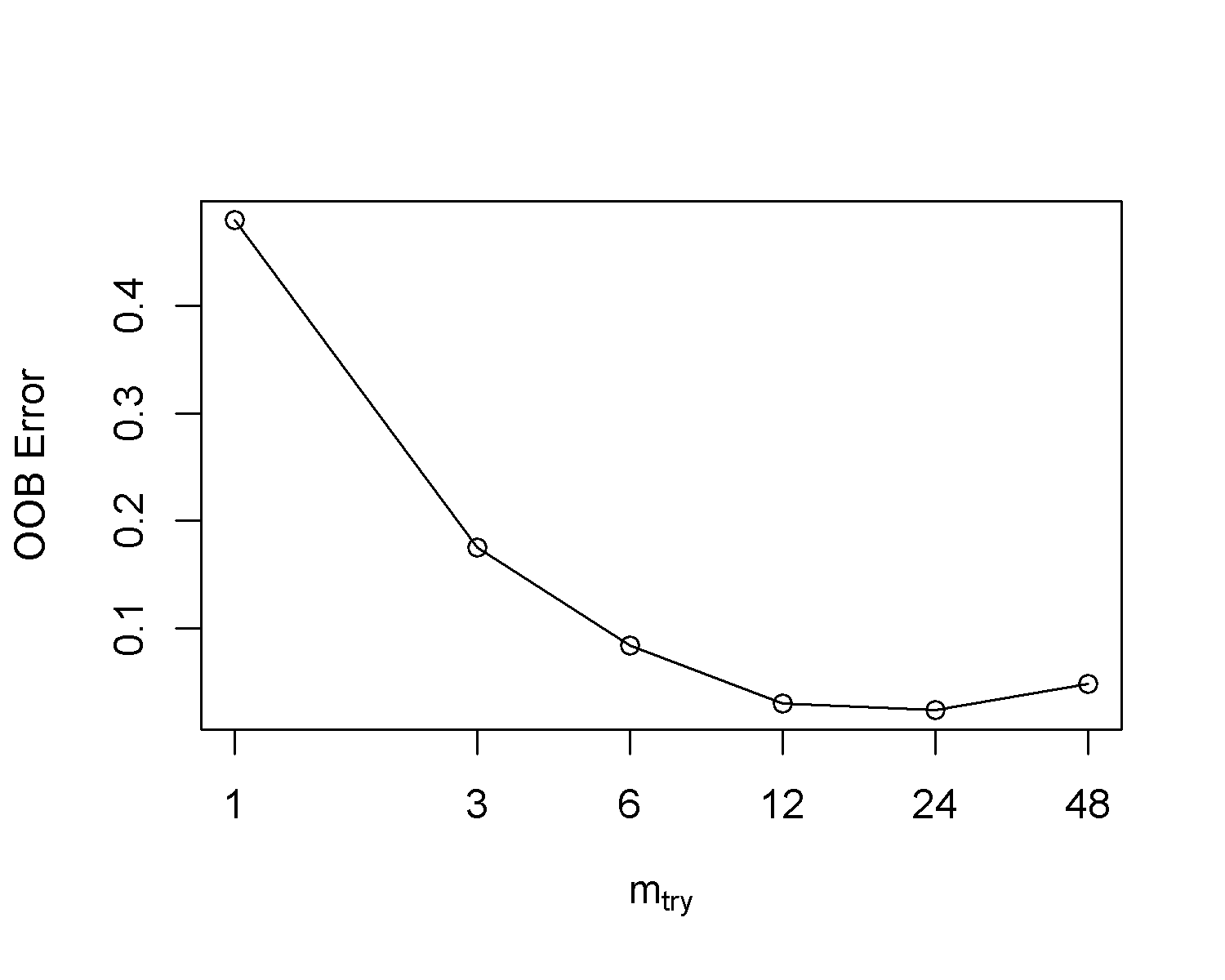


Figura 5. Mtry vs error observado con la técnica de RF.

#mtry=6  
  
#No. of nodes for the trees  
hist(treesize(rf),  
 main = "No. of Nodes for the Trees",  
 col = "blue")

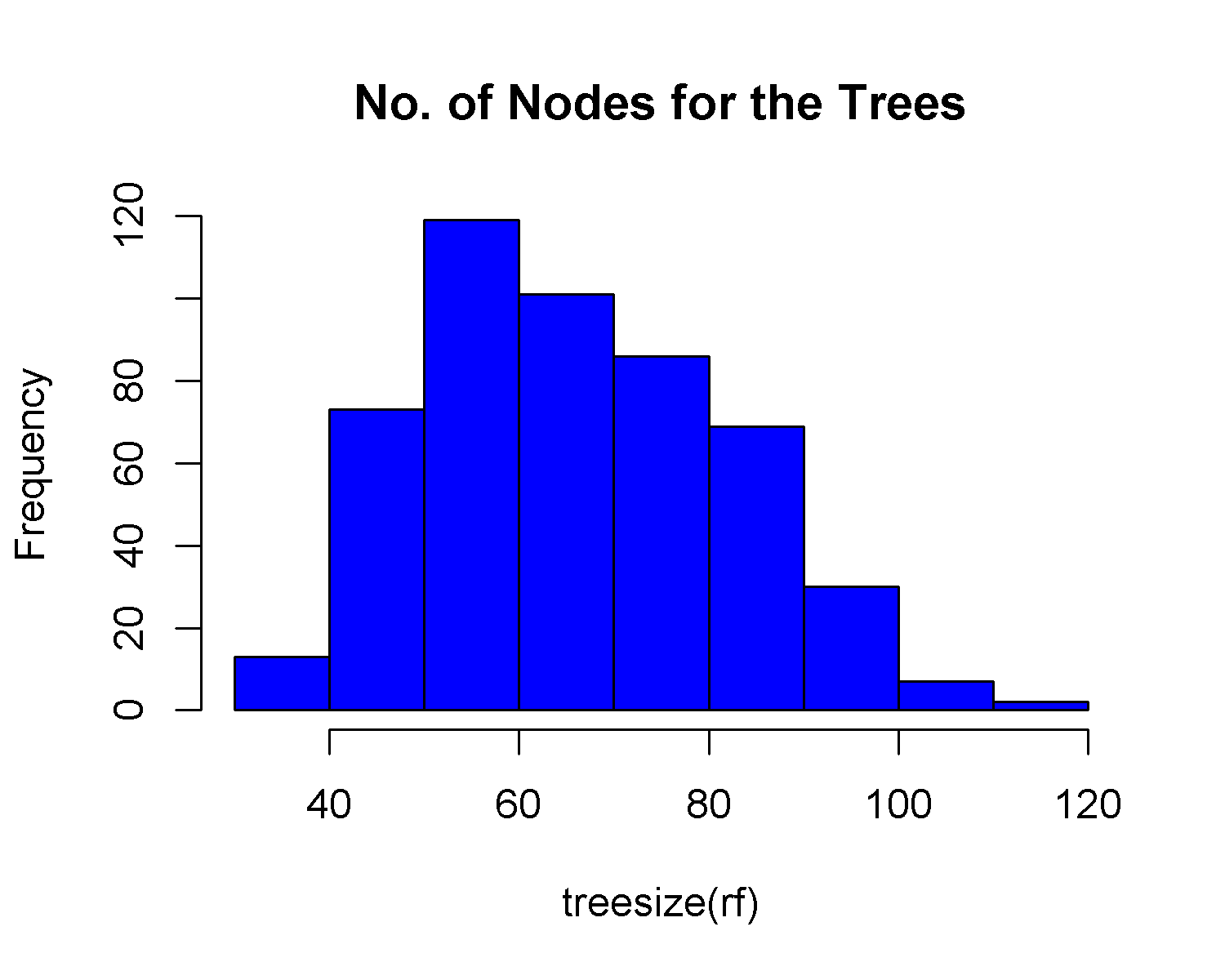


Figura 6. Histograma del Número de nodos por árboles con la técnica de RF.

#media de 60 árboles  
  
#Variable Importance  
varImpPlot(rf,  
 sort = T,  
 n.var = 10,  
 main = "Top 10 - Importancia de Variables")

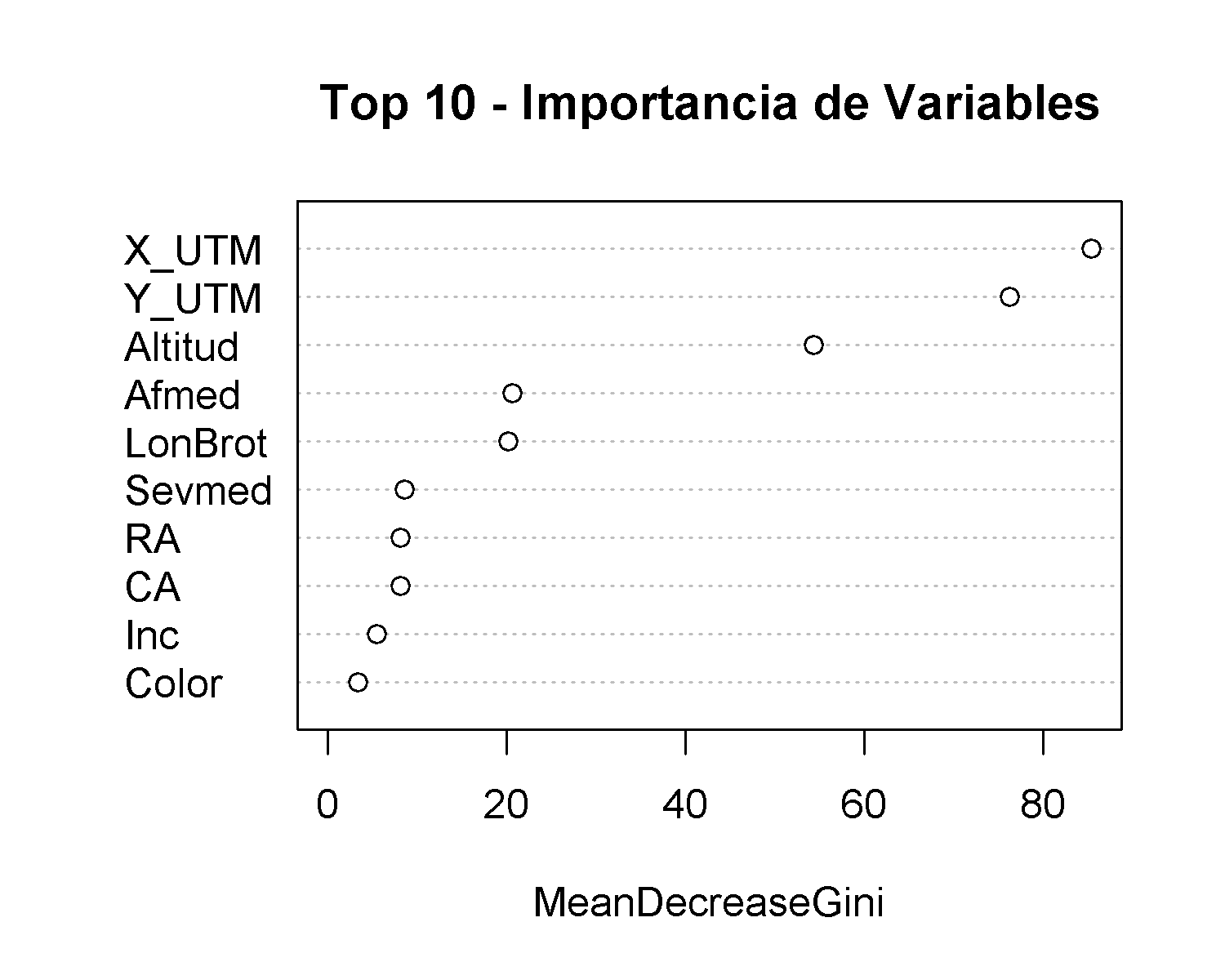


Figura 7. Relación de la importancia de las variables utilizadas en la técnica de RF.

**Interpretación:** se obtuvo una precisión del 98% y un valor = 0.98, una mtry (Número de variables aleatorias utilizadas en cada árbol) de 6, 60 nodos promedio por árbol, las variables de mayor peso son las coordenadas UTM (X y Y).

## Naive Bayes

#clasificacion en datos poco correlacionados  
#Prediccion  
library(caret)  
confusionMatrix(p2,test\_nb$Bloque)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Alta Baja Media  
## Alta 28 14 13  
## Baja 12 30 22  
## Media 17 19 35  
##   
## Overall Statistics  
##   
## Accuracy : 0.4895   
## 95% CI : (0.4164, 0.5629)  
## No Information Rate : 0.3684   
## P-Value [Acc > NIR] : 0.0004302   
##   
## Kappa : 0.2309   
##   
## Mcnemar's Test P-Value : 0.8238127   
##   
## Statistics by Class:  
##   
## Class: Alta Class: Baja Class: Media  
## Sensitivity 0.4912 0.4762 0.5000  
## Specificity 0.7970 0.7323 0.7000  
## Pos Pred Value 0.5091 0.4688 0.4930  
## Neg Pred Value 0.7852 0.7381 0.7059  
## Prevalence 0.3000 0.3316 0.3684  
## Detection Rate 0.1474 0.1579 0.1842  
## Detection Prevalence 0.2895 0.3368 0.3737  
## Balanced Accuracy 0.6441 0.6042 0.6000

**Interpretación:** se obtuvo una precisión de 49% y un valor de =0.23. En general esta técnica no fue tan buena para clasificar la transparencia de copa.

## K-NN

library(kknn)

#Precisión test de prueba  
pred <- predict(knn, test\_knn[,-1])  
table <- table(test\_knn[,1],pred)  
table

## pred  
## Alta Baja Media  
## Alta 35 13 13  
## Baja 13 29 21  
## Media 12 16 35

clas <- (sum(diag(table)))/sum(table)  
clas

## [1] 0.5294118

#Precisión del 53% de datos de prueba  
  
#matriz de confusion con la prueba  
  
library(caret)  
confusionMatrix(pred,test\_knn$Bloque)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Alta Baja Media  
## Alta 35 13 12  
## Baja 13 29 16  
## Media 13 21 35  
##   
## Overall Statistics  
##   
## Accuracy : 0.5294   
## 95% CI : (0.4552, 0.6026)  
## No Information Rate : 0.3369   
## P-Value [Acc > NIR] : 4.914e-08   
##   
## Kappa : 0.294   
##   
## Mcnemar's Test P-Value : 0.8695   
##   
## Statistics by Class:  
##   
## Class: Alta Class: Baja Class: Media  
## Sensitivity 0.5738 0.4603 0.5556  
## Specificity 0.8016 0.7661 0.7258  
## Pos Pred Value 0.5833 0.5000 0.5072  
## Neg Pred Value 0.7953 0.7364 0.7627  
## Prevalence 0.3262 0.3369 0.3369  
## Detection Rate 0.1872 0.1551 0.1872  
## Detection Prevalence 0.3209 0.3102 0.3690  
## Balanced Accuracy 0.6877 0.6132 0.6407

**Interpretación:** se obtuvo una precisión de 53% y un valor de =0.30. En general esta técnica no fue tan buena para clasificar la transparencia de copa.

## Curvas Receiver Operating Characteristic (ROC)

# Validation set assessment #2: ROC curves and AUC  
# Needs to import ROCR package for ROC curve plotting:  
library(ROCR)

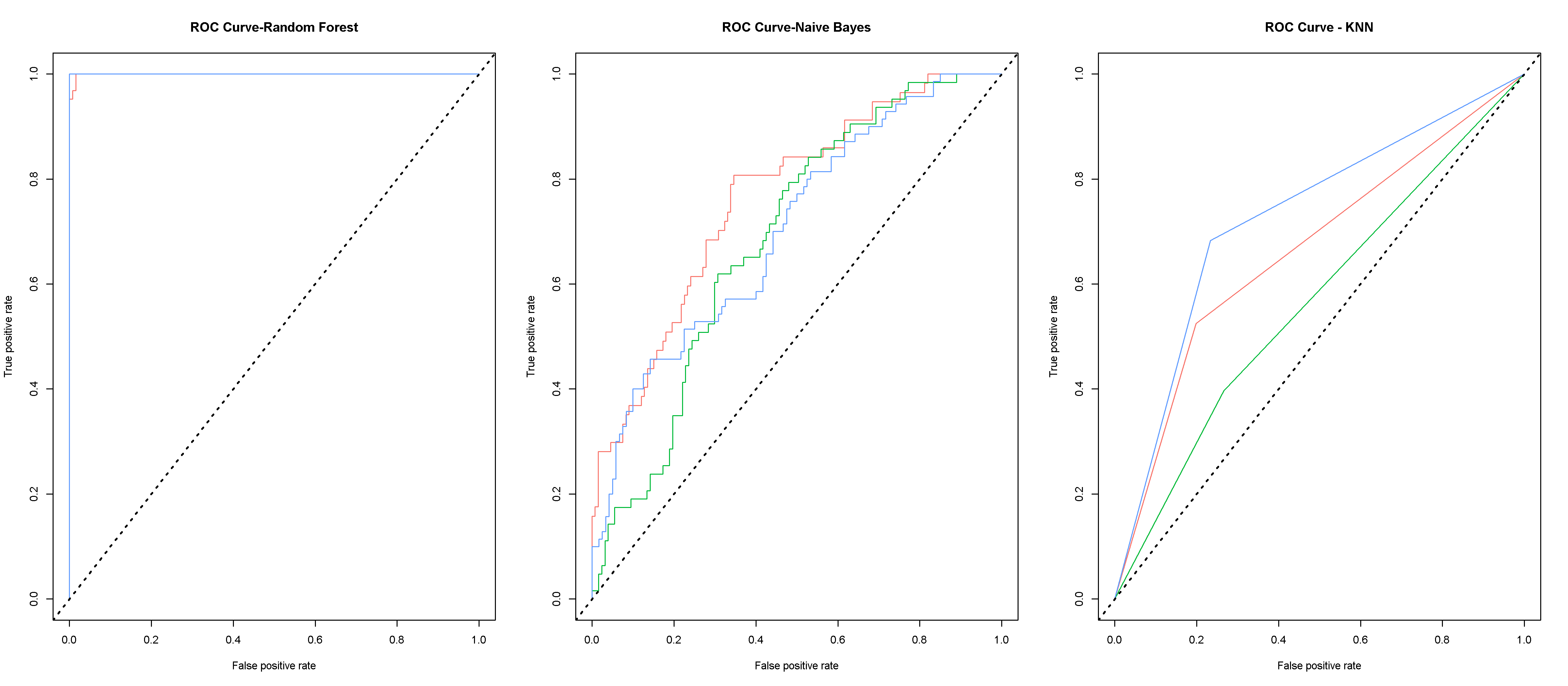


Figura 8. Curvas ROC para la técnica de Random Forest, Naive Bayes y K-NN.

#cuadro con valores de precision  
  
library(gt)  
pres %>% gt()

Cuadro 1. Métricas utilizadas para comparar las técnicas de ML.

| **Métrica** | **Random Forest** | **Naive Bayes** | **K-NN** |
| --- | --- | --- | --- |
| **Precisión (%)** | 0.98 | 0.49 | 0.53 |
| **Kappa** | 0.98 | 0.23 | 0.30 |
| **AUC-ROC** | 1.00 | 0.70 | 0.72 |

# Conclusiones

De acuerdo la evaluación de las curvas ROC, precisión y valor de kappa, es evidente que el mejor modelo para clasificar a la transparencia de copa de Douglas-fir fue Random Forest, además tuvo mayor cantidad de positivos verdaderos con sensibilidad, especificidad y precisión equilibrada en comparación de Naive Bayes y K-NN, donde se pueden considerar que son modelos de regulares a malos para clasificar la transparencia de copa.

# Bibliografía

finnstats. Naive Bayes Classification in R | R-Bloggers. 9 de abril de 2021, <https://www.r-bloggers.com/2021/04/naive-bayes-classification-in-r/>.

finnstats. Random Forest in R | R-Bloggers. 13 de abril de 2021, <https://www.r-bloggers.com/2021/04/random-forest-in-r/>.

Gandhi, Rohith. Naive Bayes Classifier. Medium, 17 de mayo de 2018, <https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c>.

K-NN Classifier in R Programming». GeeksforGeeks, 18 de junio de 2020, <https://www.geeksforgeeks.org/k-nn-classifier-in-r-programming/>.

Naive Bayes Classifier in R Programming. GeeksforGeeks, 18 de junio de 2020, <https://www.geeksforgeeks.org/naive-bayes-classifier-in-r-programming/>.

Quiroz, I. I. F. 2019. Tolerancia al tizón suizo en una plantación de árboles de navidad en Aquixtla, Puebla. Tesis de Maestría. Programa en Ciencias Forestales. Colegio de Postgraduados. 101p.

RPubs - KNN. <https://rpubs.com/JairoAyala/601703>. Accedido 20 de noviembre de 2022.