```r
# import dataset
clinvar_variants <- read.csv("~/Documents/GitHub/ECL298_2026/Describe_data/clinvar_variants.csv",
                             na.strings = c("", "NA", ".", " "))
```

```r
colnames(clinvar_variants)
```

```
##  [1] "CHROM"       "POS"         "REF"         "ALT"         "CLASS"
##  [6] "Consequence" "IMPACT"      "Amino_acids" "AF_EXAC"     "SIFT"
## [11] "PolyPhen"    "LoFtool"     "CADD_PHRED"  "BLOSUM62"
```

```r
clinvar_variants$CLASS <- factor(
  clinvar_variants$CLASS,
  levels = c(0, 1),
  labels = c("Benign", "Pathogenic")
)
```

```r
df <- clinvar_variants[complete.cases(clinvar_variants), ] # no missing values
```

```r
set.seed(123)

n <- nrow(df)
train_id <- sample(n, size = 0.7 * n)

# split the dataset
train_data <- df[train_id, ]
test_data  <- df[-train_id, ]
```

```r
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.4.1
```

```
## randomForest 4.7-1.2
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```r
rf_fit <- randomForest(
  CLASS ~ Consequence + IMPACT + Amino_acids + AF_EXAC + SIFT + PolyPhen +
    LoFtool + CADD_PHRED + BLOSUM62,
  data = train_data,
  importance = TRUE
)
```

```r
rf_fit
```

```
##
## Call:
##  randomForest(formula = CLASS ~ Consequence + IMPACT + Amino_acids +      AF_EXAC + SIFT + PolyPhen +
##                Type of random forest: classification
##                      Number of trees: 500
```

```
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 25.57%
## Confusion matrix:
##            Benign Pathogenic class.error
## Benign       8624        730  0.07804148
## Pathogenic   2492        754  0.76771411
```

```
pred_class <- predict(rf_fit, newdata = test_data)
confusion_matrix <- table(Predicted = pred_class, True = test_data$CLASS)
confusion_matrix
```

```
##             True
## Predicted    Benign Pathogenic
##   Benign       3770       1017
##   Pathogenic    291        322
```

```
# Accuracy
mean(pred_class == test_data$CLASS)
```

```
## [1] 0.7577778
```

```
# variants importance
varImpPlot(rf_fit, type = 2)
```