

```

library(caret)
library(gbm)
data(scot)
ds<-scot

#1.Set the species column as outcome/target and converting to numeric
ds$Species<-as.numeric(factor(ds$Species))

#2. Remove the Month, Year, Site, Location features
ds$Month<-NULL
ds$Year<-NULL
ds$Site<-NULL
ds$Location<-NULL
str(ds)

#3. Check if any values are null. If there are, impute missing values using
KNN
sum(is.na(ds))
preProcValues <- preProcess(ds, method = c("knnImpute","center","scale"))

library('RANN')
ds_processed <- predict(preProcValues, ds)
sum(is.na(ds_processed))

#4. Converting every categorical variable to numerical (if needed)
dmy <- dummyVars(" ~ .", data = ds_processed,fullRank = T)
ds_transformed <- data.frame(predict(dmy, newdata = ds_processed))
str(ds_transformed)
ds_transformed$Species<-as.factor(ds_transformed$Species)

#5. With a seed of 100, 75% training, 25% testing. Build the following models:
randomforest, neural net, naive bayes and GBM
#Splitting ds into 75% training and 25% testing
set.seed(100)
index <- createDataPartition(ds_transformed$Species, p=0.75, list=FALSE)
trainSet <- ds_transformed[index,]

```