# Niez, Gilly

## Predictive Analytics

### 06/06/2021

### Task 1

# summary(drug\_clean)  
# skim(drug\_clean)

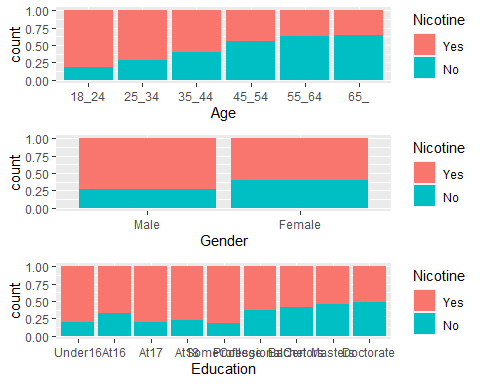
* No missingness found.

### Task 2

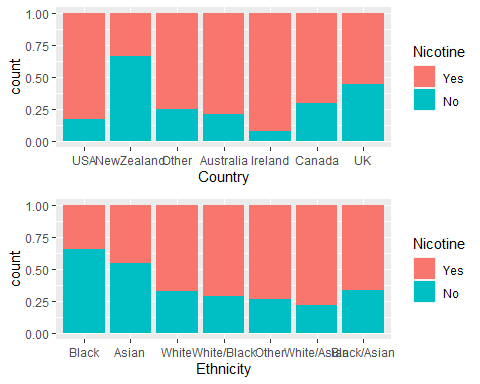
set.seed(1234)   
drug\_split = initial\_split(drug\_clean, prop = 0.70, strata = Nicotine)  
train = training(drug\_split)  
test = testing(drug\_split)

### Task 3

p1 = ggplot(train, aes(x = Age, fill = Nicotine)) + geom\_bar(position = "fill")  
p2 = ggplot(train, aes(x = Gender, fill = Nicotine)) + geom\_bar(position = "fill")  
p3 = ggplot(train, aes(x = Education, fill = Nicotine)) + geom\_bar(position = "fill")  
grid.arrange(p1,p2,p3)

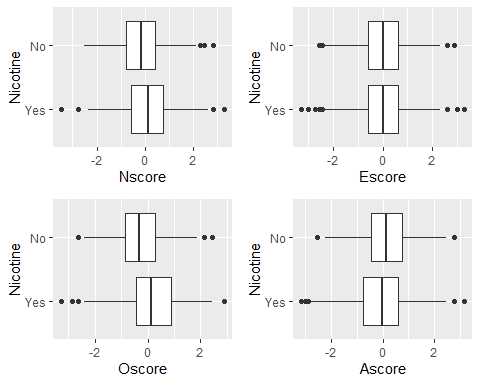


p1 = ggplot(train, aes(x = Country, fill = Nicotine)) + geom\_bar(position = "fill")  
p2= ggplot(train, aes(x= Ethnicity, fill= Nicotine)) + geom\_bar(position= "fill")  
grid.arrange(p1,p2)



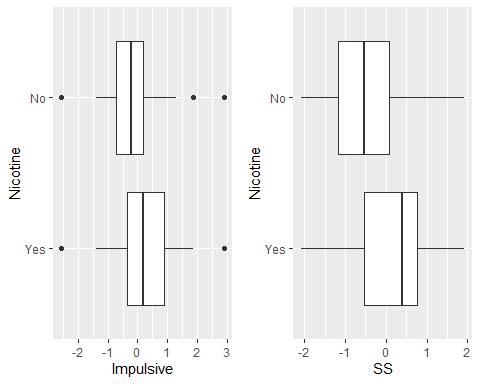
* Age shows that nicotine use tends to lessen as a person gets older. It is much higher from ages betweer 18-24 yrs old.
* Gender shows that there are more females within this data set that tends to use Nicotine.
* Education shows that there is a tendency of Nicotine used depending on levels of education.
* Certain countries has a tendency to use more Nicotine than others.
* Ethnicity variable shows that there are certain ethnicity that are susceptible to use more nicotine than others.

p1 = ggplot(train, aes(x = Nscore, y = Nicotine)) + geom\_boxplot()  
p2 = ggplot(train, aes(x = Escore, y = Nicotine)) + geom\_boxplot()  
p3 = ggplot(train, aes(x = Oscore, y = Nicotine)) + geom\_boxplot()  
p4= ggplot(train, aes(x = Ascore, y = Nicotine)) + geom\_boxplot()  
grid.arrange(p1,p2,p3,p4, ncol = 2)



* People with higher Neuroticism scores are tends to lean more towards nicotine use
* Extraversion scores does not show any tendencies towards the use of nicotine or lack there of.
* People with a higher Oscore tends to show their openness to try nicotine.
* However, it shows more people are decling the use of nicotine in Ascore.

p1 = ggplot(train, aes(x = Impulsive, y = Nicotine)) + geom\_boxplot()  
p2 = ggplot(train, aes(x = SS, y = Nicotine)) + geom\_boxplot()  
grid.arrange(p1,p2, ncol = 2)



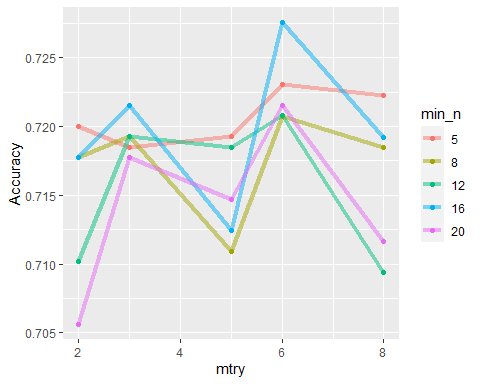
* This data shows that impulsive people are tend to use/try Nicotine.
* There are more people with higher SS score tends to use/try Nicotine.

### Task 4

set.seed(123)  
rf\_folds = vfold\_cv(train, v = 5)

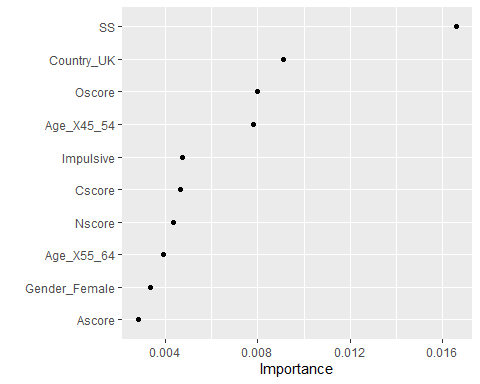
drug\_recipe = recipe(Nicotine ~., train) %>%  
 step\_dummy(all\_nominal(), -all\_outcomes())  
  
rf\_model = rand\_forest(mtry = tune(), min\_n = tune(), trees = 100) %>%   
 set\_engine("ranger", importance = "permutation") %>% #added importance metric  
 set\_mode("classification")  
  
drug\_wflow =   
 workflow() %>%   
 add\_model(rf\_model) %>%   
 add\_recipe(drug\_recipe)  
  
rf\_grid = grid\_regular(  
 mtry(range = c(2, 8)), #these values determined through significant trial and error  
 min\_n(range = c(5, 20)), #these values determined through significant trial and error  
 levels = 5  
)  
  
set.seed(123)  
rf\_res\_tuned = tune\_grid(  
 drug\_wflow,  
 resamples = rf\_folds,  
 grid = rf\_grid   
)

rf\_res\_tuned %>%  
 collect\_metrics() %>%  
 filter(.metric == "accuracy") %>%  
 mutate(min\_n = factor(min\_n)) %>%  
 ggplot(aes(mtry, mean, color = min\_n)) +  
 geom\_line(alpha = 0.5, size = 1.5) +  
 geom\_point() +  
 labs(y = "Accuracy")



best\_rf = select\_best(rf\_res\_tuned, "accuracy")  
  
final\_rf = finalize\_workflow(  
 drug\_wflow,  
 best\_rf  
)  
  
# final\_rf  
  
final\_rf\_fit = fit(final\_rf, train)

final\_rf\_fit %>% pull\_workflow\_fit() %>% vip(geom = "point")



* People with higher sensation seeing is deemed to be of top importance followed in a descending order: Country, Oscore, Age, impulsiveness, etc..

trainpredrf = predict(final\_rf\_fit, train)  
  
confusionMatrix(trainpredrf$.pred\_class, train$Nicotine,   
 positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Yes No  
## Yes 871 97  
## No 13 337  
##   
## Accuracy : 0.9165   
## 95% CI : (0.9003, 0.9309)  
## No Information Rate : 0.6707   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.8013   
##   
## Mcnemar's Test P-Value : 2.498e-15   
##   
## Sensitivity : 0.9853   
## Specificity : 0.7765   
## Pos Pred Value : 0.8998   
## Neg Pred Value : 0.9629   
## Prevalence : 0.6707   
## Detection Rate : 0.6608   
## Detection Prevalence : 0.7344   
## Balanced Accuracy : 0.8809   
##   
## 'Positive' Class : Yes   
##

testpredrf = predict(final\_rf\_fit, test)  
  
confusionMatrix(testpredrf$.pred\_class, test$Nicotine,   
 positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Yes No  
## Yes 327 115  
## No 53 72  
##   
## Accuracy : 0.7037   
## 95% CI : (0.6642, 0.741)  
## No Information Rate : 0.6702   
## P-Value [Acc > NIR] : 0.04826   
##   
## Kappa : 0.2681   
##   
## Mcnemar's Test P-Value : 2.523e-06   
##   
## Sensitivity : 0.8605   
## Specificity : 0.3850   
## Pos Pred Value : 0.7398   
## Neg Pred Value : 0.5760   
## Prevalence : 0.6702   
## Detection Rate : 0.5767   
## Detection Prevalence : 0.7795   
## Balanced Accuracy : 0.6228   
##   
## 'Positive' Class : Yes   
##

* Accuracy for the training set is : 92%
* Accuracy for the testing set is : 70%
* There is a significant gap between the performance of accuracies between the training and testing set.
* This model can be applicable in the real world to see which personality trait is more susceptible to the use of Nicotine.
* I would not advise using this model since it shows an imbalance of accuracy in testing and training set.