## Random Forests

### Julia Perez

Libraries

Import & Clean Data

drug <- read\_csv("drug\_data-1.csv")

## Rows: 1885 Columns: 32

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (19): Column14, Column15, Column16, Column17, Column18, Column19, Column...  
## dbl (13): Column1, Column2, Column3, Column4, Column5, Column6, Column7, Col...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

names(drug) = c("ID", "Age", "Gender", "Education", "Country", "Ethnicity",  
"Nscore", "Escore", "Oscore", "Ascore", "Cscore", "Impulsive",  
"SS", "Alcohol", "Amphet", "Amyl", "Benzos", "Caff", "Cannabis",  
"Choc", "Coke", "Crack", "Ecstasy", "Heroin", "Ketamine", "Legalh",  
"LSD", "Meth", "Mushrooms", "Nicotine", "Semer", "VSA")  
  
#str(drug)  
  
drug[drug == "CL0"] = "No"  
drug[drug == "CL1"] = "No"  
drug[drug == "CL2"] = "Yes"  
drug[drug == "CL3"] = "Yes"  
drug[drug == "CL4"] = "Yes"  
drug[drug == "CL5"] = "Yes"  
drug[drug == "CL6"] = "Yes"  
  
drug\_clean = drug %>% mutate\_at(vars(Age:Ethnicity), funs(as\_factor)) %>%  
mutate(Age = factor(Age, labels = c("18\_24", "25\_34", "35\_44",  
"45\_54", "55\_64", "65\_"))) %>%  
mutate(Gender = factor(Gender, labels = c("Male", "Female"))) %>%  
mutate(Education = factor(Education, labels =  
c("Under16", "At16", "At17", "At18", "SomeCollege",  
"ProfessionalCert", "Bachelors", "Masters", "Doctorate"))) %>%  
mutate(Country = factor(Country,  
labels = c("USA", "NewZealand", "Other", "Australia",  
"Ireland","Canada","UK"))) %>%  
mutate(Ethnicity = factor(Ethnicity,  
labels = c("Black", "Asian", "White", "White/Black", "Other",  
"White/Asian", "Black/Asian"))) %>%  
 mutate\_at(vars(Alcohol:VSA), funs(as\_factor)) %>%  
select(-ID)

## Warning: `funs()` was deprecated in dplyr 0.8.0.  
## Please use a list of either functions or lambdas:   
##   
## # Simple named list:   
## list(mean = mean, median = median)  
##   
## # Auto named with `tibble::lst()`:   
## tibble::lst(mean, median)  
##   
## # Using lambdas  
## list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was generated.

#str(drug\_clean)  
  
drug\_clean = drug\_clean %>% select(!(Alcohol:Mushrooms)) %>% select(!(Semer:VSA))  
names(drug\_clean)

## [1] "Age" "Gender" "Education" "Country" "Ethnicity" "Nscore"   
## [7] "Escore" "Oscore" "Ascore" "Cscore" "Impulsive" "SS"   
## [13] "Nicotine"

### Check for Missing Data

summary(drug\_clean)

## Age Gender Education Country   
## 18\_24:643 Male :943 SomeCollege :506 USA : 557   
## 25\_34:481 Female:942 Bachelors :480 NewZealand: 5   
## 35\_44:356 Masters :283 Other : 118   
## 45\_54:294 ProfessionalCert:270 Australia : 54   
## 55\_64: 93 At18 :100 Ireland : 20   
## 65\_ : 18 At16 : 99 Canada : 87   
## (Other) :147 UK :1044   
## Ethnicity Nscore Escore Oscore   
## Black : 33 Min. :-3.464360 Min. :-3.273930 Min. :-3.273930   
## Asian : 26 1st Qu.:-0.678250 1st Qu.:-0.695090 1st Qu.:-0.717270   
## White :1720 Median : 0.042570 Median : 0.003320 Median :-0.019280   
## White/Black: 20 Mean : 0.000047 Mean :-0.000163 Mean :-0.000534   
## Other : 63 3rd Qu.: 0.629670 3rd Qu.: 0.637790 3rd Qu.: 0.723300   
## White/Asian: 20 Max. : 3.273930 Max. : 3.273930 Max. : 2.901610   
## Black/Asian: 3   
## Ascore Cscore Impulsive   
## Min. :-3.464360 Min. :-3.464360 Min. :-2.555240   
## 1st Qu.:-0.606330 1st Qu.:-0.652530 1st Qu.:-0.711260   
## Median :-0.017290 Median :-0.006650 Median :-0.217120   
## Mean :-0.000245 Mean :-0.000386 Mean : 0.007216   
## 3rd Qu.: 0.760960 3rd Qu.: 0.584890 3rd Qu.: 0.529750   
## Max. : 3.464360 Max. : 3.464360 Max. : 2.901610   
##   
## SS Nicotine   
## Min. :-2.078480 Yes:1264   
## 1st Qu.:-0.525930 No : 621   
## Median : 0.079870   
## Mean :-0.003292   
## 3rd Qu.: 0.765400   
## Max. : 1.921730   
##

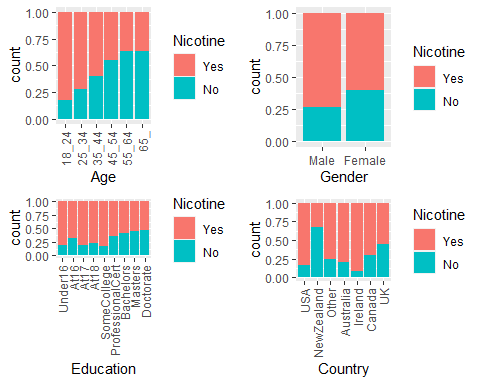
There is no missing data in drug\_clean.

### Split data

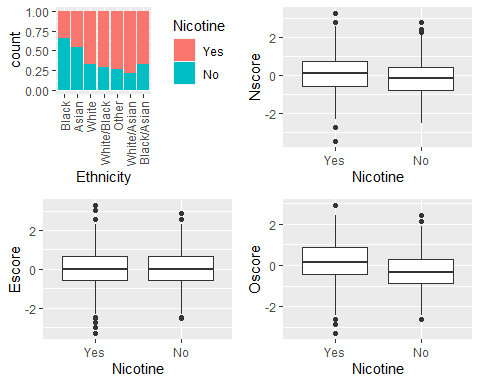
set.seed(1234)  
drug\_clean\_split <- initial\_split(drug\_clean, prop = 0.70, strata = Nicotine)  
train <- training(drug\_clean\_split)  
test <- testing(drug\_clean\_split)

### Exploratory Data Analysis

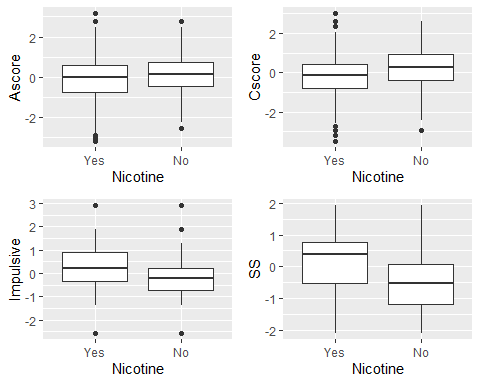
p1 <- ggplot(train, aes(Age, fill = Nicotine)) + geom\_bar(position = "fill") +  
 theme(axis.text.x = element\_text(angle=90, vjust=.5, hjust=1))  
p2 <- ggplot(train, aes(Gender, fill = Nicotine)) + geom\_bar(position = "fill")  
p3 <- ggplot(train, aes(Education, fill = Nicotine)) + geom\_bar(position = "fill") +  
 theme(axis.text.x = element\_text(angle=90, vjust=.5, hjust=1))  
p4 <- ggplot(train, aes(Country, fill = Nicotine)) + geom\_bar(position = "fill") +   
 theme(axis.text.x = element\_text(angle=90, vjust=.5, hjust=1))  
grid.arrange(p1, p2, p3, p4)



p5 <- ggplot(train, aes(Ethnicity, fill = Nicotine)) + geom\_bar(position = "fill") +  
 theme(axis.text.x = element\_text(angle=90, vjust=.5, hjust=1))  
p6 <- ggplot(train, aes(Nicotine, Nscore)) + geom\_boxplot()  
p7 <- ggplot(train, aes(Nicotine, Escore)) + geom\_boxplot()  
p8 <- ggplot(train, aes(Nicotine, Oscore)) + geom\_boxplot()  
grid.arrange(p5, p6, p7, p8)



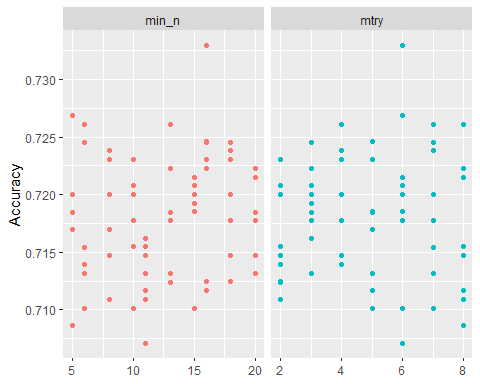
p9 <- ggplot(train, aes(Nicotine, Ascore)) + geom\_boxplot()  
p10 <- ggplot(train, aes(Nicotine, Cscore)) + geom\_boxplot()  
p11 <- ggplot(train, aes(Nicotine, Impulsive)) + geom\_boxplot()  
p12 <- ggplot(train, aes(Nicotine, SS)) + geom\_boxplot()  
grid.arrange(p9, p10, p11, p12)



* Age: Nicotine use decreases with age.
* Gender: Men use nicotine more than women.
* Education: College and higher educated individuals use less nicotine than individuals with less education.
* Country: Country does seem to impact nicotine use. New Zealanders are less likely to be nicotine users compared to the Irish.
* Ethnicity: Ethnicity looks like it impacts nicotine usage. Blacks and Asians are less likely to use nicotine when compared to other ethnicities.
* Nscore: Nicotine users appear to have a slightly higher Nscore compared to non-nicotine users. This difference appears slight and may not represent a strong impact.
* Escore: Escore does not appear to predict nicotine usage.
* Oscore: Nicotine users have a higher average Oscore compared to non-nicotine users. This variable looks to predict nicotine usage.
* Ascore: Non-nicotine users have a higher Ascore compared to nicotine users. This difference appears slight and may not predict nicotine usage.
* Cscore: Non-nicotine users have a higher Cscore compared to nicotine users. This looks to be a large difference and likely predicts nicotine usage.
* Impulsive: Non-nicotine users have lower impulsive scores than nicotine users. This is a likely predictor for nicotine usage.
* SS: Nicotine users appear to have a much higher SS compared to non-nicotine users. This is a likely predictor for nicotine usage.

### Random Forest Model

#Set folds  
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") %>%   
 set\_mode("classification")  
  
drug\_wflow <-   
 workflow() %>%   
 add\_model(rf\_model) %>%   
 add\_recipe(drug\_recipe)  
  
rf\_grid <- grid\_regular(  
 mtry(range = c(2,8)),  
 min\_n(range = c(5,20)),  
 levels = 10  
)  
  
set.seed(123)  
rf\_res <- tune\_grid(  
 drug\_wflow,  
 resamples = rf\_folds,  
 grid = rf\_grid)  
  
rf\_res %>%  
 collect\_metrics() %>%  
 filter(.metric == "accuracy") %>%  
 select(mean, min\_n, mtry) %>%  
 pivot\_longer(min\_n:mtry,  
 values\_to = "value",  
 names\_to = "parameter"  
 ) %>%  
 ggplot(aes(value, mean, color = parameter)) +  
 geom\_point(show.legend = FALSE) +  
 facet\_wrap(~parameter, scales = "free\_x") +  
 labs(x = NULL, y = "Accuracy")



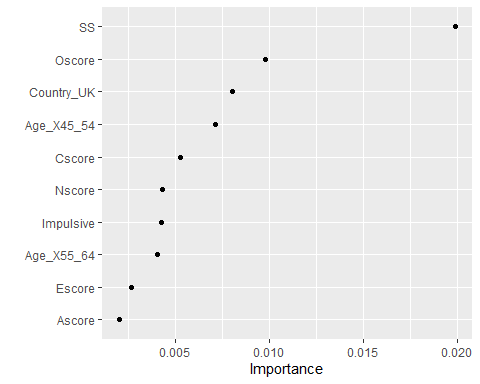
best\_rf <- select\_best(rf\_res, "accuracy")  
best\_rf

## # A tibble: 1 x 3  
## mtry min\_n .config   
## <int> <int> <chr>   
## 1 6 16 Preprocessor1\_Model54

The best mtry is 6, and the best min\_n is 16.

### Final Model and Variable Importance

final\_rf <- finalize\_workflow(drug\_wflow, best\_rf)  
  
final\_rf\_fit <- fit(final\_rf, train)  
final\_rf\_fit %>% extract\_fit\_parsnip() %>% vip(geom = "point")



Sensation seeing (SS) is the most important variable in this model by a good margin. An individual’s age being between 45 and 54 is the next most important variable followed by openess to experience (Oscore) then being located in the UK. Other important variables are conscientiousness (Cscore), impulsiveness, being between 55 and 64 years old, neuroticism (Nscore), being female, and having some college education.

### Evaluating Accuracy

#Training set predictions  
train\_pred <- predict(final\_rf\_fit, train)  
#head(train\_pred)  
  
#Confusion Matrix for Training  
confusionMatrix(train\_pred$.pred\_class, train$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   
##

#Testing set predictions  
test\_pred <- predict(final\_rf\_fit, test)  
  
#Confusion Matrix for Testing Set  
confusionMatrix(test\_pred$.pred\_class, test$Nicotine,   
 positive = "Yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Yes No  
## Yes 327 119  
## No 53 68  
##   
## Accuracy : 0.6966   
## 95% CI : (0.657, 0.7343)  
## No Information Rate : 0.6702   
## P-Value [Acc > NIR] : 0.09699   
##   
## Kappa : 0.2462   
##   
## Mcnemar's Test P-Value : 7.188e-07   
##   
## Sensitivity : 0.8605   
## Specificity : 0.3636   
## Pos Pred Value : 0.7332   
## Neg Pred Value : 0.5620   
## Prevalence : 0.6702   
## Detection Rate : 0.5767   
## Detection Prevalence : 0.7866   
## Balanced Accuracy : 0.6121   
##   
## 'Positive' Class : Yes   
##

On the training set, the model has an accuracy of 90.67%, compared to the No Information Rate of 67.07%. On the testing set, the model had a large drop in accuracy to 70.55%. When compared to the No Information Rate of 67.02%, this model did not perform very well on the testing set.

This model could be used to identify individuals who are likely to use nicotine before they start consuming it. However, I would not recommend this model for real world use. The model did not perform much better on the testing set than if we had assumed all individuals were nicotine users. In a real world scenario, it would likely be more cost and time effective to provide all individuals counseling regarding nicotine use than to collect the relevant data to use with this model.