Predicting Genetic Disorders

Emma Oo[1], Sindhu Bhattarai[2], Dave Friesen[3]

06/27/2022

Data Load and Validation

```
# Load dataset(s)
gd_df <- read.csv("../data/train_genetic_disorders.csv", header = TRUE)

# Data validation and understanding, including structure, content, and statistical characteristics covered below
```

Data Structure Review

```
# Summarize base dataset and [optionally] sample rows
str(gd_df)
```

```
'data.frame':   22083 obs. of  45 variables:
 $ Patient.Id                                    : chr  "PID0x6418" "PID0x25d5" "PID0x4a82" "PID0x4ac8" ...
 $ Patient.Age                                   : int  2 4 6 12 11 14 3 3 11 4 ...
 $ Genes.in.mother.s.side                        : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ Inherited.from.father                         : chr  "No" "Yes" "No" "No" ...
 $ Maternal.gene                                 : chr  "Yes" "No" "No" "Yes" ...
 $ Paternal.gene                                 : chr  "No" "No" "No" "No" ...
 $ Blood.cell.count..mcL.                        : num  4.76 4.91 4.89 4.71 4.72 ...
 $ Patient.First.Name                            : chr  "Richard" "Mike" "Kimberly" "Jeffery" ...
 $ Family.Name                                   : chr  "" "" "" "Hoelscher" ...
 $ Father.s.name                                 : chr  "Larre" "Brycen" "Nashon" "Aayaan" ...
 $ Mother.s.age                                  : int  NA NA 41 21 32 NA NA 40 45 44 ...
 $ Father.s.age                                  : int  NA 23 22 NA NA NA 63 NA 44 42 ...
 $ Institute.Name                                : chr  "Boston Specialty & Rehabilitation Hospital" "St. Margaret's Hospi
tal For Women" "" "" ...
 $ Location.of.Institute                         : chr  "55 FRUIT ST\nCENTRAL, MA 02114\n(42.36247485742686, -71.069247245
45246)" "1515 COMMONWEALTH AV\nALLSTON/BRIGHTON, MA 02135\n(42.34665771451756, -71.14136122385321)" "-" "55 FRUIT ST\nCENTRA
L, MA 02114\n(42.36247485742686, -71.06924724545246)" ...
 $ Status                                        : chr  "Alive" "Deceased" "Alive" "Deceased" ...
 $ Respiratory.Rate..breaths.min.                : chr  "Normal (30-60)" "Tachypnea" "Normal (30-60)" "Tachypnea" ...
 $ Heart.Rate..rates.min                         : chr  "Normal" "Normal" "Tachycardia" "Normal" ...
 $ Test.1                                        : int  0 NA 0 0 0 0 NA 0 0 0 ...
 $ Test.2                                        : int  NA 0 0 0 0 0 0 0 0 0 ...
 $ Test.3                                        : int  NA 0 0 0 0 0 0 NA 0 0 ...
 $ Test.4                                        : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Test.5                                        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Parental.consent                              : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ Follow.up                                     : chr  "High" "High" "Low" "High" ...
 $ Gender                                        : chr  "" "" "" "Male" ...
 $ Birth.asphyxia                                : chr  "" "No" "No record" "Not available" ...
 $ Autopsy.shows.birth.defect..if.applicable.    : chr  "Not applicable" "None" "Not applicable" "No" ...
 $ Place.of.birth                                : chr  "Institute" "" "" "Institute" ...
 $ Folic.acid.details..peri.conceptional.        : chr  "No" "Yes" "Yes" "No" ...
 $ H.O.serious.maternal.illness                  : chr  "" "Yes" "No" "Yes" ...
 $ H.O.radiation.exposure..x.ray.                : chr  "No" "Not applicable" "Yes" "-" ...
 $ H.O.substance.abuse                           : chr  "No" "Not applicable" "" "Not applicable" ...
 $ Assisted.conception.IVF.ART                   : chr  "No" "No" "Yes" "" ...
 $ History.of.anomalies.in.previous.pregnancies  : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ No..of.previous.abortion                      : int  NA NA 4 1 4 0 3 1 0 1 ...
 $ Birth.defects                                 : chr  "" "Multiple" "Singular" "Singular" ...
 $ White.Blood.cell.count..thousand.per.microliter.: num  9.86 5.52 NA 7.92 4.1 ...
 $ Blood.test.result                             : chr  "" "normal" "normal" "inconclusive" ...
 $ Symptom.1                                     : int  1 1 0 0 0 1 0 0 1 0 ...
 $ Symptom.2                                     : int  1 NA 1 0 0 0 0 0 1 0 ...
 $ Symptom.3                                     : int  1 1 1 1 0 0 0 1 1 1 ...
 $ Symptom.4                                     : int  1 1 1 0 0 1 0 NA 0 1 ...
 $ Symptom.5                                     : int  1 0 1 0 NA 0 0 0 1 1 ...
 $ Genetic.Disorder                              : chr  "Mitochondrial genetic inheritance disorders" "" "Multifactorial g
enetic inheritance disorders" "Mitochondrial genetic inheritance disorders" ...
 $ Disorder.Subclass                             : chr  "Leber's hereditary optic neuropathy" "Cystic fibrosis" "Diabetes"
"Leigh syndrome" ...
```

```
#head(gd_df, 3)
```

Preliminary Feature Reduction (clearly n/a to Objective and Hypothesis)

```
# Define n/a columns and subset dataframe; Note retaining "some" informational variables like "Institute.Name" for
#   possible descriptive analytic purposes
drop_cols <- c("Patient.Id",
               "Patient.First.Name",
               "Family.Name",
               "Father.s.name",
               "Institute.Name",
               "Location.of.Institute",
               "Status",
               "Test.1",
               "Test.2",
               "Test.3",
               "Test.4",
               "Test.5",
               "Parental.consent",
               "Birth.asphyxia",
               "Place.of.birth",
               "Genetic.Disorder")
gd_df <- gd_df[ , !(names(gd_df) %in% drop_cols)]
```

Class Target and Label Review

```
# Check for missing labels; set aside where missing
missing_target <- which(is.na(gd_df$Disorder.Subclass) | (gd_df$Disorder.Subclass == ""))
cat("Rows pre-subset for missing labels: ", format(nrow(gd_df), format = "d", big.mark = ","), sep = "")
```

```
Rows pre-subset for missing labels: 22,083
```

```
gd_hold_df <- gd_df[missing_target, ]
gd_df <- gd_df[-missing_target, ]
cat("Deleted rows with missing labels: ", format(nrow(gd_hold_df), format = "d", big.mark = ","), sep = "")
```

```
Deleted rows with missing labels: 3,140
```

```
cat("Remaining rows (labeled): ", format(nrow(gd_df), format = "d", big.mark = ","), sep = "")
```

```
Remaining rows (labeled): 18,943
```

```
# Show frequency distribution for [prospective] target class(es)
show_frequency <- function(desc, c) {
  t <- as.data.frame(prop.table(table(c)))
  colnames(t) <- c("Class", "Frequency")
  cat(desc, "\n"); print(t[order(-t$Freq, t$Class), 1:2], row.names = FALSE)
}
show_frequency("Pre-Split Frequency Distribution", gd_df$Disorder.Subclass)
```

```
Pre-Split Frequency Distribution
                              Class Frequency
                       Leigh syndrome     0.258
               Mitochondrial myopathy     0.222
                      Cystic fibrosis     0.173
                            Tay-Sachs     0.142
                             Diabetes     0.092
                       Hemochromatosis     0.068
 Leber's hereditary optic neuropathy     0.032
                           Alzheimer's     0.008
                                Cancer     0.005
```

```
# Move the target class to "top" of dataframe so column removals don't impact
gd_df <- gd_df[ , c(ncol(gd_df), 1:(ncol(gd_df) - 1))]
target_col = 1

# Clean (prelim) target class values
gd_df$Disorder.Subclass <- gsub("'", "", gd_df$Disorder.Subclass, fixed = TRUE)
gd_df$Disorder.Subclass <- gsub(" ", ".", gd_df$Disorder.Subclass, fixed = TRUE)
gd_df$Disorder.Subclass <- gsub("-", ".", gd_df$Disorder.Subclass, fixed = TRUE)
```

Data Partitioning

```
# Split data 80/20 train/test, using caret's inherent stratified split to compensate for class imbalance
set.seed(1)
train_index <- createDataPartition(gd_df$Disorder.Subclass, times = 1, p = 0.80, list = FALSE)
train_df <- gd_df[train_index, ]
test_df <- gd_df[-train_index, ]
show_frequency("Post-Split Frequency Distribution (Train)", train_df$Disorder.Subclass)
```

```
Post-Split Frequency Distribution (Train)
                              Class Frequency
                      Leigh.syndrome     0.258
              Mitochondrial.myopathy     0.222
                     Cystic.fibrosis     0.173
                           Tay.Sachs     0.142
                            Diabetes     0.092
                      Hemochromatosis     0.068
 Lebers.hereditary.optic.neuropathy     0.032
                          Alzheimers     0.008
                              Cancer     0.005
```

Data Cleaning (and reduction)

Data (Sample) Characteristic Review for Pre-Processing

(Suppressing custom code for simplicity)

```
# Generate a summary (cursory) view of base dataset for initial understanding and pre-processing direction
univariate(train_df)
```

```
Summary Univariate Analysis (15,158 observations)
                     Type          NA BlankZ  Unique     Min     Max    Mean  Median Outlier< >Outlier Kurtosis Skewness
Disorder.Subclass    character                  9
Patient.Age          integer       6%     6%      15              14               7    No       Yes       0.017   -1.211
Genes.in.mother.s... character                  2
Inherited.from.fa... character            1%      3
Maternal.gene        character           12%      3
Paternal.gene        character                  2
Blood.cell.count.... numeric                15,158   4.093   5.610   4.900   4.902    No       Yes      -0.011   -0.037
Mother.s.age         integer      26%           34      18      51              35    No       Yes      -0.006   -1.219
Father.s.age         integer      25%           45      20      64              42    No       Yes      -0.002   -1.210
Respiratory.Rate.... character            9%      3
Heart.Rate..rates... character            9%      3
Follow.up            character            9%      3
Gender               character            9%      4
Autopsy.shows.bir... character            4%      5
Folic.acid.detail... character            9%      3
H.O.serious.mater... character            8%      3
H.O.radiation.exp... character            9%      5
H.O.substance.abuse  character            9%      5
Assisted.concepti... character            9%      3
History.of.anomal... character            9%      3
No..of.previous.a... integer       9%    18%      5               4               2    No       Yes       0.001   -1.292
Birth.defects        character            9%      3
White.Blood.cell.... numeric       9%          11,858   3.000  12.000   7.460   7.443    No       Yes       0.020   -0.979
Blood.test.result    character            9%      5
Symptom.1            integer       9%    37%      2               1               1    No       Yes      -0.369   -1.864
Symptom.2            integer       9%    40%      2               1               1    No       Yes      -0.197   -1.961
Symptom.3            integer       8%    41%      2               1               1    No       Yes      -0.166   -1.973
Symptom.4            integer       9%    45%      2               1                    No       Yes       0.010   -2.000
Symptom.5            integer       9%    48%      2               1                    No       Yes       0.146   -1.979
```

Missing Values

```
# Impute basic integer values with medians
medianf <- function(x) {
 result <- median(x, na.rm = TRUE)
 if (is.integer(x))
   result <- as.integer(result)
 return(result)
}
median_cols = c("Patient.Age", "Mother.s.age", "Father.s.age", "No..of.previous.abortion")
for (n in median_cols) {
  train_df[n][is.na(train_df[n])] <- apply(train_df[n], 2, medianf)
  test_df[n][is.na(test_df[n])] <- apply(test_df[n], 2, medianf)
}

# Impute categorical blanks with common "notprovided"; note we could also impute these with categorical mode,
#   or most frequent categorical value of each column using the cmode() function below
cols_tofill <- c("Inherited.from.father",
                 "Maternal.gene",
                 "Respiratory.Rate..breaths.min.",
                 "Heart.Rate..rates.min",
                 "Follow.up",
                 "Gender",
                 "Autopsy.shows.birth.defect..if.applicable.",
                 "Folic.acid.details..peri.conceptional.",
                 "H.O.serious.maternal.illness",
                 "H.O.radiation.exposure..x.ray.",
                 "H.O.substance.abuse",
                 "Assisted.conception.IVF.ART",
                 "History.of.anomalies.in.previous.pregnancies",
                 "Birth.defects",
                 "Blood.test.result")
train_df[cols_tofill][train_df[cols_tofill] == ""] <- "notprovided"
test_df[cols_tofill][test_df[cols_tofill] == ""] <- "notprovided"

cmode <- function(x) {
  uniqx <- unique(na.omit(x))
  uniqx[which.max(tabulate(match(x, uniqx)))]
}

# Impute what appear to be masked "flag" columns iwth placeholder -1 values. . .
flag_cols <- c("Symptom.1", "Symptom.2", "Symptom.3", "Symptom.4", "Symptom.5")
train_df[flag_cols][is.na(train_df[flag_cols])] <- as.integer(-1)
test_df[flag_cols][is.na(test_df[flag_cols])] <- as.integer(-1)

# Impute mean for one numeric column
train_df$White.Blood.cell.count..thousand.per.microliter.[is.na(train_df$White.Blood.cell.count..thousand.per.microliter.)]
 <-
  mean(train_df$White.Blood.cell.count..thousand.per.microliter., na.rm = TRUE)
test_df$White.Blood.cell.count..thousand.per.microliter.[is.na(test_df$White.Blood.cell.count..thousand.per.microliter.)] <-
  mean(test_df$White.Blood.cell.count..thousand.per.microliter., na.rm = TRUE)

# Note not using knnImpute for the limited number of numerical [prospective] features given that it
#   centers/scales, which is illogical for the values in this dataset
#pp <- preProcess(train_df[ , -target_col, drop = FALSE], method = "knnImpute", k = 10)
#train_df[ , -target_col] <- predict(pp, train_df[ , -target_col, drop = FALSE])
#test_df[ , -target_col] <- predict(pp, test_df[ , -target_col, drop = FALSE])

# Last on the list: Genetic.Disorder - we're not classifying to this but it is relevant/informational as a
#   superclass to the target Disorder.Subclass and shuold ultimately be imputed using similar Disorder.Subclass
#   observations which do have valid Genetic.Disorder values
```

Feature Updates (including variable types/formats, names)

```
# Re-type variables
factor_cols <- c("Disorder.Subclass",
                 "Genes.in.mother.s.side",
                 "Inherited.from.father",
                 "Maternal.gene",
                 "Paternal.gene",
                 "Respiratory.Rate..breaths.min.",
                 "Heart.Rate..rates.min",
                 "Follow.up",
                 "Gender",
                 "Autopsy.shows.birth.defect..if.applicable.",
                 "Folic.acid.details..peri.conceptional.",
                 "H.O.serious.maternal.illness",
                 "H.O.radiation.exposure..x.ray.",
                 "H.O.substance.abuse",
                 "Assisted.conception.IVF.ART",
                 "History.of.anomalies.in.previous.pregnancies",
                 "Birth.defects",
                 "Blood.test.result")
train_df[factor_cols] <- lapply(train_df[factor_cols], factor)
test_df[factor_cols] <- lapply(test_df[factor_cols], factor)
# Note dummy variables may be introduced below (model-dependent)

# Simplify variable naming
rename_cols <- c("Disorder_Subclass",
                 "Patient_Age",
                 "Genes_mothers_side",
                 "Genes_fathers_side",
                 "Maternal_gene",
                 "Paternal_gene",
                 "Blood_cell_count",
                 "Mothers_age",
                 "Fathers_age",
                 "Respiratory_Rate",
                 "Heart_Rate",
                 "Follow_up",
                 "Gender",
                 "Autopsy_birth_defect",
                 "Folic_acid_conceptional",
                 "HO_maternal_illness",
                 "HO_radiation_exposure",
                 "HO_substance_abuse",
                 "Assisted_conception",
                 "Previous_pregnancies_issues",
                 "Previous_abortions",
                 "Birth_defects",
                 "White_Blood_cell_count",
                 "Blood_test_result",
                 "Symptom_1",
                 "Symptom_2",
                 "Symptom_3",
                 "Symptom_4",
                 "Symptom_5")
colnames(train_df) <- rename_cols
colnames(test_df) <- rename_cols
```

Zero/Near-Zero Variances

```
# n/a for this dataset
```

Duplicate Values

```
# n/a for this dataset
```

“Noisy” Data

```
# n/a for this dataset
```

Data Transformation

Centering/Scaling (standardizing/normalizing)

```
# n/a for this dataset
```

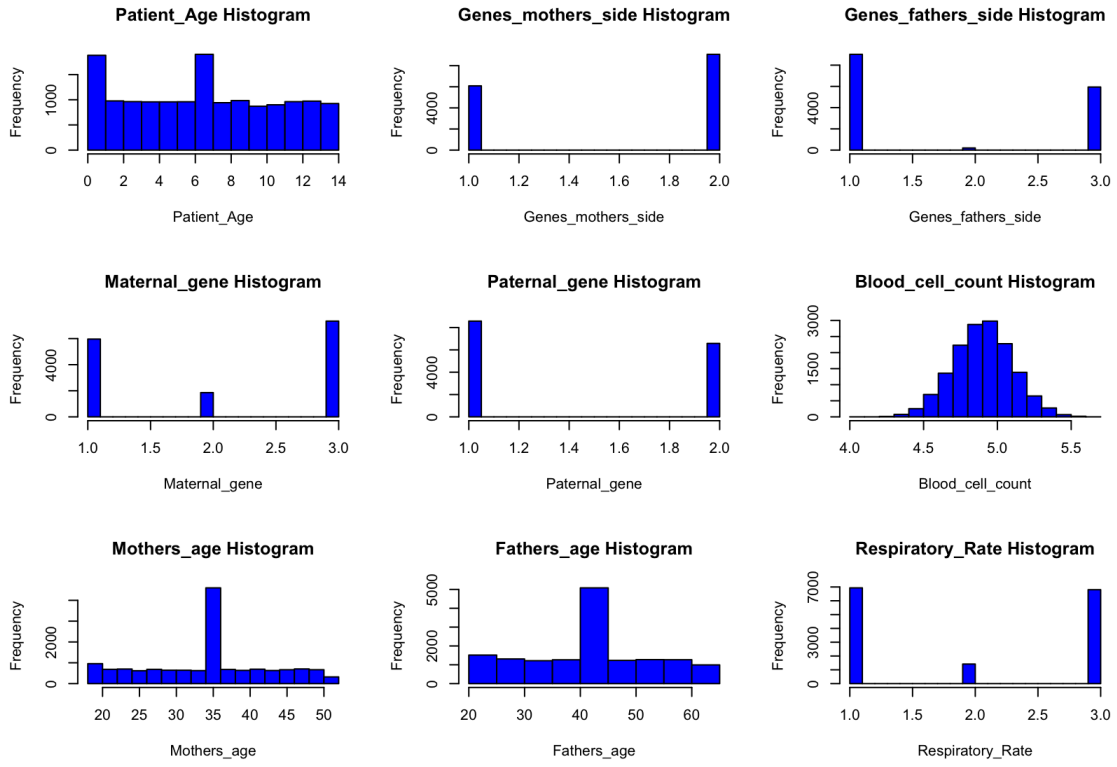
Statistical Characteristics (including distribution, skewness, outliers)

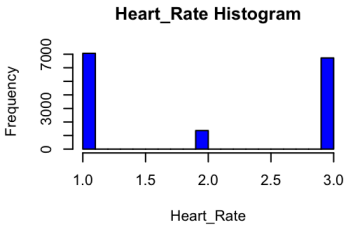
```
# Generate updated summary of base dataset which includes these characteristics
univariate(train_df)
```

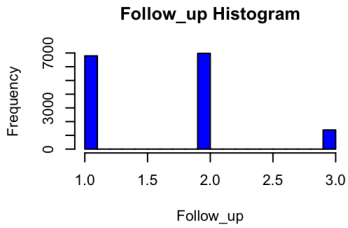
```
Summary Univariate Analysis (15,158 observations)
                     Type        NA BlankZ  Unique      Min      Max     Mean   Median Outlier< >Outlier Kurtosis Skewness
Disorder_Subclass    factor                      9
Patient_Age          integer             6%     15               14                7    No       Yes       0.016   -1.090
Genes_mothers_side   factor                      2
Genes_fathers_side   factor                      3
Maternal_gene        factor                      3
Paternal_gene        factor                      2
Blood_cell_count     numeric                15,158    4.093    5.610    4.900    4.902    No       Yes      -0.011   -0.037
Mothers_age          integer                    34       18       51               35    No       Yes      -0.048   -0.593
Fathers_age          integer                    45       20       64               42    No       Yes      -0.007   -0.600
Respiratory_Rate     factor                      3
Heart_Rate           factor                      3
Follow_up            factor                      3
Gender               factor                      4
Autopsy_birth_defect factor                      5
Folic_acid_concep... factor                      3
HO_maternal_illness  factor                      3
HO_radiation_expo... factor                      5
HO_substance_abuse   factor                      5
Assisted_conception  factor                      3
Previous_pregnanc... factor                      3
Previous_abortions   integer            18%       5                4                2    No       Yes               -1.116
Birth_defects        factor                      3
White_Blood_cell_... numeric                11,859    3.000   12.000    7.460    7.460    No       Yes       0.021   -0.768
Blood_test_result    factor                      5
Symptom_1            integer            37%       3       -1        1                1    No       Yes      -0.769   -0.496
Symptom_2            integer            40%       3       -1        1                     No       Yes      -0.643   -0.624
Symptom_3            integer            41%       3       -1        1                     No       Yes      -0.626   -0.613
Symptom_4            integer            45%       3       -1        1                     No       Yes      -0.502   -0.679
Symptom_5            integer            48%       3       -1        1                     No       Yes      -0.413   -0.702
```

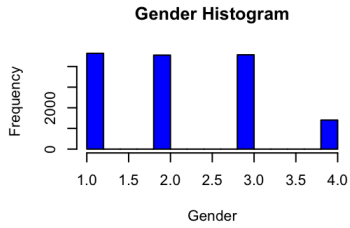
```
#summary(train_df)
```

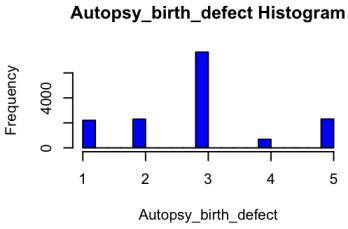
```
# Generate histograms across predictors and target
pred_for_hist <- train_df[ , 2:29]
pred_for_hist <- pred_for_hist %>% mutate_if(is.character, as.numeric)
pred_for_hist <- pred_for_hist %>% mutate_if(is.factor, as.numeric)
par(mfrow = c(3, 3))
for (i in 1:ncol(pred_for_hist))
  hist(pred_for_hist[ , i], xlab = names(pred_for_hist[i]), main = paste(names(pred_for_hist[i]), "Histogram"), col = "blue"
)
```

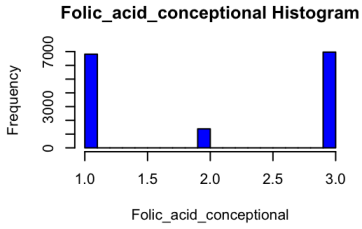


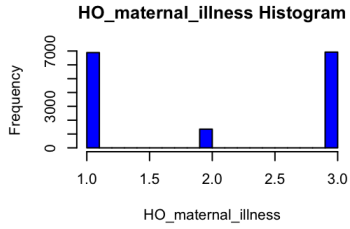

Heart_Rate Histogram
Frequency
0
3000
7000
1.0
1.5
2.0
2.5
3.0
Heart_Rate

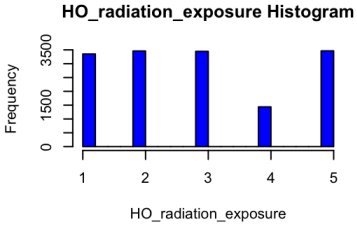

Follow_up Histogram
Frequency
0
3000
7000
1.0
1.5
2.0
2.5
3.0
Follow_up

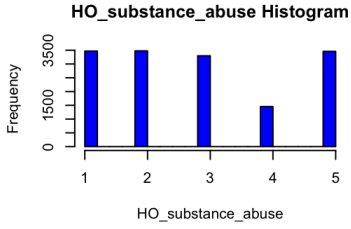

Gender Histogram
Frequency
0
2000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Gender

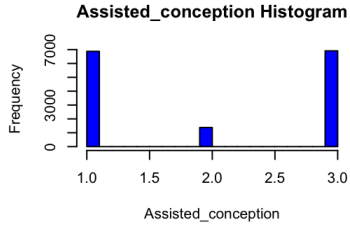

Autopsy_birth_defect Histogram
Frequency
0
4000
1
2
3
4
5
Autopsy_birth_defect

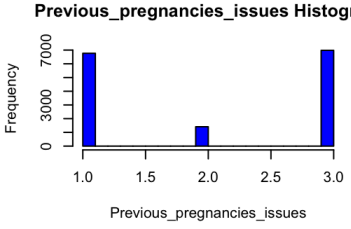

Folic_acid_conceptional Histogram
Frequency
0
3000
7000
1.0
1.5
2.0
2.5
3.0
Folic_acid_conceptional

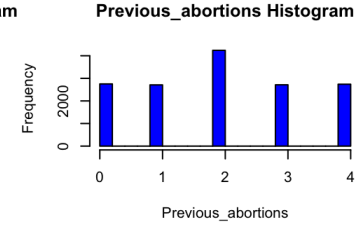

HO_maternal_illness Histogram
Frequency
0
3000
7000
1.0
1.5
2.0
2.5
3.0
HO_maternal_illness

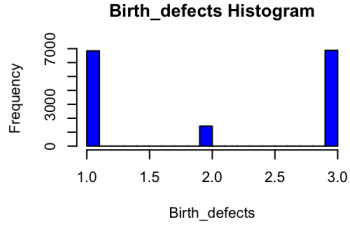

HO_radiation_exposure Histogram
Frequency
0
1500
3500
1
2
3
4
5
HO_radiation_exposure

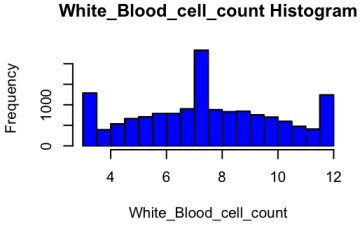

HO_substance_abuse Histogram
Frequency
0
1500
3500
1
2
3
4
5
HO_substance_abuse

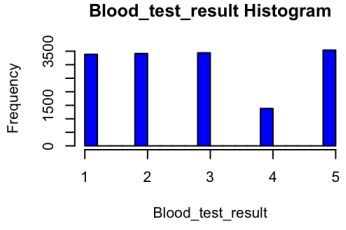

Assisted_conception Histogram
Frequency
0
3000
7000
1.0
1.5
2.0
2.5
3.0
Assisted_conception

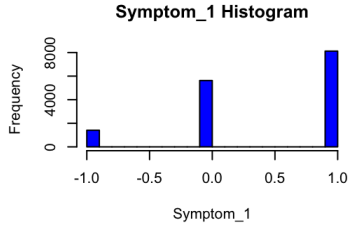

Previous_pregnancies_issues Histogram
Frequency
0
3000
7000
1.0
1.5
2.0
2.5
3.0
Previous_pregnancies_issues

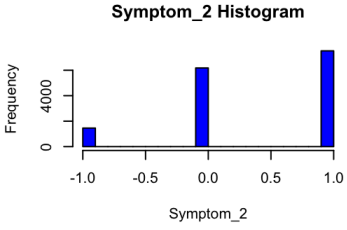

Previous_abortions Histogram
Frequency
0
2000
0
1
2
3
4
Previous_abortions

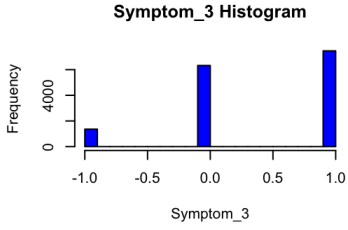

Birth_defects Histogram
Frequency
0
3000
7000
1.0
1.5
2.0
2.5
3.0
Birth_defects

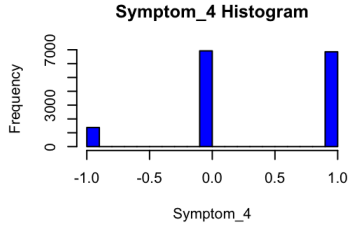

White_Blood_cell_count Histogram
Frequency
0
1000
4
6
8
10
12
White_Blood_cell_count

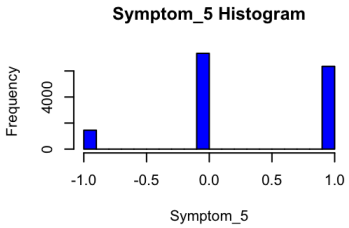

Blood_test_result Histogram
Frequency
0
1500
3500
1
2
3
4
5
Blood_test_result


Symptom_1 Histogram
Frequency
0
4000
8000
-1.0
-0.5
0.0
0.5
1.0
Symptom_1

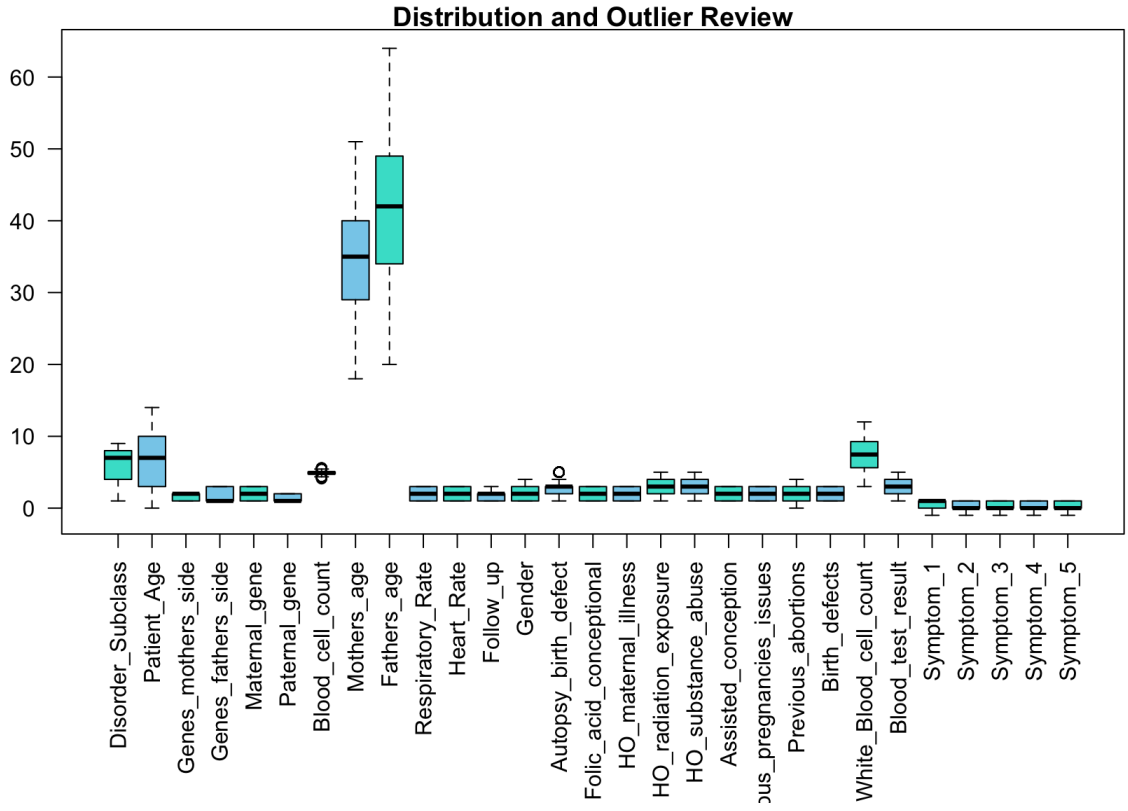

Symptom_2 Histogram
Frequency
0
4000
-1.0
-0.5
0.0
0.5
1.0
Symptom_2


Symptom_3 Histogram
Frequency
0
4000
-1.0
-0.5
0.0
0.5
1.0
Symptom_3


Symptom_4 Histogram
Frequency
0
3000
7000
-1.0
-0.5
0.0
0.5
1.0
Symptom_4




```
# Generate boxplot(s)
par(mar = c(10, 2, 1, 1))
boxplot(train_df, las = 2, col = c("turquoise", "skyblue"), main = "Distribution and Outlier Review", ylab = "Frequency")
```



Other Feature Engineering (transformation, aggregation, enrichment)

```
# n/a for this dataset
```

Multivariate Analysis (and reduction)

Collinearity and Dependencies

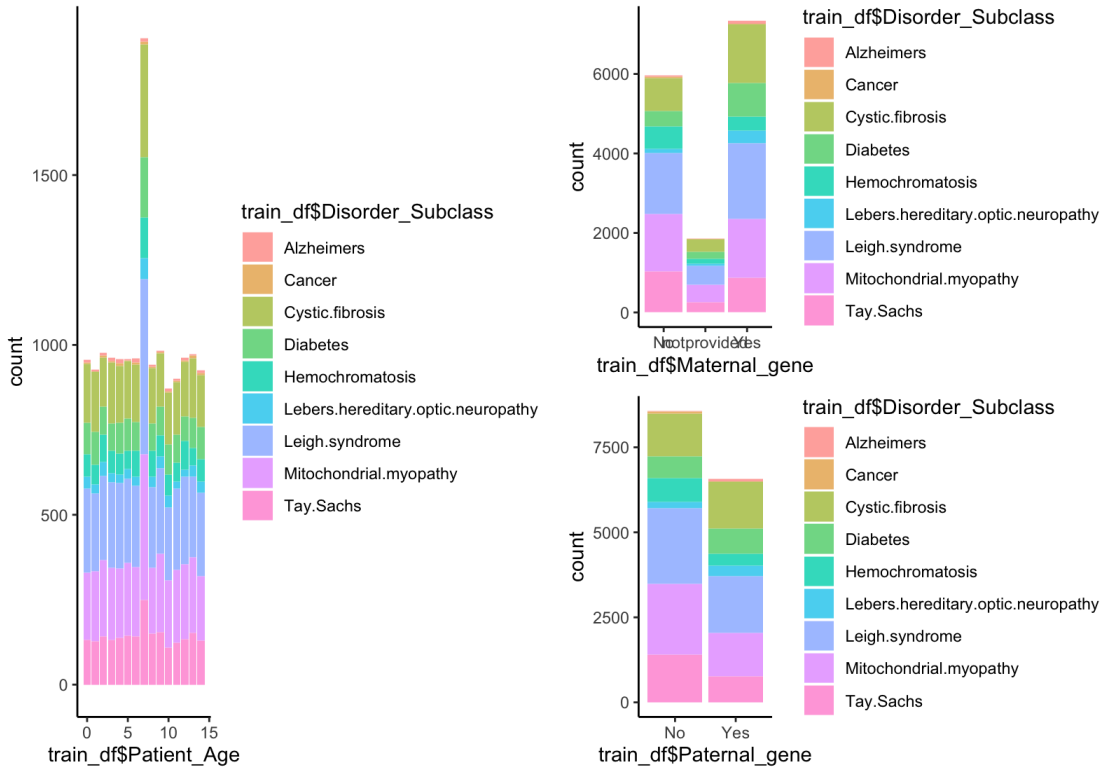
```
# Calculate Cramer's V "measure of association" between nominal factor variables (uses Chi-square statistic)
cscorr <- PairApply(train_df[ , sapply(train_df, is.factor)], CramerV, symmetric = TRUE)

# Shorten variable names for ease of reviewing output matrix
rn <- rownames(cscorr)
for (n in 1:length(rownames(cscorr))) {
  rn[n] <- paste(rownames(cscorr)[n], " (", AscToChar(64 + n), ")", sep = "")
  rownames(cscorr)[n] <- paste(AscToChar(64 + n))
}
for (n in 1:length(colnames(cscorr)))
  colnames(cscorr)[n] <- paste(AscToChar(64 + n))

# Show master list of variable names along with output ("correlation") matrix
cat(rn, sep = "\n")
```

```
Disorder_Subclass (A)
Genes_mothers_side (B)
Genes_fathers_side (C)
Maternal_gene (D)
Paternal_gene (E)
Respiratory_Rate (F)
Heart_Rate (G)
Follow_up (H)
Gender (I)
Autopsy_birth_defect (J)
Folic_acid_conceptional (K)
HO_maternal_illness (L)
HO_radiation_exposure (M)
HO_substance_abuse (N)
Assisted_conception (O)
Previous_pregnancies_issues (P)
Birth_defects (Q)
Blood_test_result (R)
```

```
cscorr
```

```
     A     B     C     D     E     F     G    H    I    J     K     L     M    N     O     P     Q    R
A 1.00 0.198 0.131 0.123 0.168 0.019 0.026 0.02 0.02 0.02 0.020 0.019 0.024 0.02 0.019 0.026 0.025 0.03
B 0.20 1.000 0.005 0.097 0.012 0.005 0.005 0.01 0.01 0.01 0.013 0.009 0.016 0.01 0.003 0.017 0.008 0.01
C 0.13 0.005 1.000 0.013 0.093 0.018 0.020 0.01 0.02 0.02 0.021 0.013 0.030 0.02 0.013 0.018 0.016 0.02
D 0.12 0.097 0.013 1.000 0.008 0.048 0.040 0.05 0.05 0.04 0.053 0.048 0.052 0.04 0.055 0.047 0.044 0.05
E 0.17 0.012 0.093 0.008 1.000 0.003 0.009 0.01 0.01 0.02 0.003 0.001 0.008 0.02 0.003 0.008 0.006 0.02
F 0.02 0.005 0.018 0.048 0.003 1.000 0.045 0.03 0.05 0.02 0.043 0.028 0.030 0.04 0.035 0.036 0.042 0.04
G 0.03 0.005 0.020 0.040 0.009 0.045 1.000 0.04 0.05 0.02 0.035 0.029 0.047 0.03 0.055 0.042 0.041 0.05
H 0.02 0.015 0.012 0.046 0.011 0.029 0.040 1.00 0.04 0.04 0.041 0.043 0.032 0.04 0.043 0.051 0.038 0.05
I 0.02 0.010 0.023 0.047 0.010 0.054 0.045 0.04 1.00 0.02 0.032 0.051 0.045 0.04 0.035 0.028 0.036 0.04
J 0.02 0.010 0.016 0.035 0.025 0.019 0.023 0.04 0.02 1.00 0.030 0.022 0.028 0.03 0.021 0.024 0.029 0.03
K 0.02 0.013 0.021 0.053 0.003 0.043 0.035 0.04 0.03 0.03 1.000 0.020 0.049 0.04 0.028 0.032 0.030 0.04
L 0.02 0.009 0.013 0.048 0.001 0.028 0.029 0.04 0.05 0.02 0.020 1.000 0.048 0.04 0.043 0.042 0.032 0.04
M 0.02 0.016 0.030 0.052 0.008 0.030 0.047 0.03 0.04 0.03 0.049 0.048 1.000 0.03 0.046 0.047 0.052 0.04
N 0.02 0.011 0.015 0.043 0.016 0.035 0.032 0.04 0.04 0.03 0.042 0.037 0.026 1.00 0.033 0.050 0.049 0.03
O 0.02 0.003 0.013 0.055 0.003 0.035 0.055 0.04 0.03 0.02 0.028 0.043 0.046 0.03 1.000 0.035 0.032 0.03
P 0.03 0.017 0.018 0.047 0.008 0.036 0.042 0.05 0.03 0.02 0.032 0.042 0.047 0.05 0.035 1.000 0.032 0.04
Q 0.02 0.008 0.016 0.044 0.006 0.042 0.041 0.04 0.04 0.03 0.030 0.032 0.052 0.05 0.032 0.032 1.000 0.04
R 0.03 0.013 0.018 0.052 0.016 0.036 0.046 0.05 0.04 0.03 0.042 0.041 0.037 0.03 0.031 0.041 0.044 1.00
```

```
# Per hypothesis, relate (visualize) target with maternal and paternal genes to understand more direct relationship
p1 <- ggplot(train_df, aes(x = train_df$Patient_Age, fill = train_df$Disorder_Subclass)) + geom_bar() + theme_classic() + sc
ale_fill_hue(c = 60, l = 80)
p2 <- ggplot(train_df, aes(x = train_df$Maternal_gene, fill = train_df$Disorder_Subclass)) + geom_bar() + theme_classic() +
 scale_fill_hue(c = 60, l = 80)
p3 <- ggplot(train_df, aes(x = train_df$Paternal_gene, fill = train_df$Disorder_Subclass)) + geom_bar() + theme_classic() +
 scale_fill_hue(c = 60, l = 80)
p1 + p2 / p3
```



Predictor Transformations (e.g., PCA)

Modeling

```
# Optional feature subset for experiment with only gene predictors
drop_cols <- c("Patient_Age",
               "Blood_cell_count",
               "Mothers_age",
               "Fathers_age",
               "Respiratory_Rate",
               "Heart_Rate",
               "Follow_up",
               "Gender",
               "Autopsy_birth_defect",
               "Folic_acid_conceptional",
               "HO_maternal_illness",
               "HO_radiation_exposure",
               "HO_substance_abuse",
               "Assisted_conception",
               "Previous_pregnancies_issues",
               "Previous_abortions",
               "Birth_defects",
               "White_Blood_cell_count",
               "Blood_test_result",
               "Symptom_1",
               "Symptom_2",
               "Symptom_3",
               "Symptom_4",
               "Symptom_5")
train_df <- train_df[ , !(names(train_df) %in% drop_cols)]
test_df <- test_df[ , !(names(test_df) %in% drop_cols)]
```

```
# Convert factors to dummies (retaining non-factors and also keeping the target as a factor)
dummies <- dummyVars(Disorder_Subclass ~. , data = train_df[ , sapply(train_df, is.factor)])
train_df <- cbind(Disorder_Subclass = train_df$Disorder_Subclass, train_df[ , !sapply(train_df, is.factor)], data.frame(pred
ict(dummies, newdata = train_df)))
dummies <- dummyVars(Disorder_Subclass ~. , data = test_df[ , sapply(test_df, is.factor)])
test_df <- cbind(Disorder_Subclass = test_df$Disorder_Subclass, test_df[ , !sapply(test_df, is.factor)], data.frame(predict
(dummies, newdata = test_df)))

# Create common control for models
set.seed(1)
fit_control <- trainControl(method = "cv",
                            savePredictions = "all",
                            classProbs = TRUE,
                            summaryFunction = multiClassSummary)
```

Linear Discriminate Analysis Model

```
# Train LDA model
set.seed(476)
lda_fit <- train(x = train_df[ , -target_col, drop = FALSE],
                 y = train_df$Disorder_Subclass,
                 method = "lda",
                 preProc = c("center", "scale"),
                 metric = "ROC",
                 trControl = fit_control)
lda_cm <- confusionMatrix(lda_fit, norm = "none")
lda_cm
```

```
Cross-Validated (10 fold) Confusion Matrix

(entries are un-normalized aggregated counts)

                                    Reference
Prediction                           Alzheimers Cancer Cystic.fibrosis Diabetes Hemochromatosis Lebers.hereditary.optic.neur
opathy Leigh.syndrome
  Alzheimers                                  0      0               3        2               0
1              0
  Cancer                                      0      0               1        1               4
0              3
  Cystic.fibrosis                            50      0            1379      843               1
299            713
  Diabetes                                   45      0             110      131               0
87             12
  Hemochromatosis                             0     29              28        3              90
2             83
  Lebers.hereditary.optic.neuropathy         10      0              10       15               0
6              0
  Leigh.syndrome                             13      1             840      301             191
71           2145
  Mitochondrial.myopathy                      1     22             206       77             506
19            764
  Tay.Sachs                                   0     22              45       22             241
1            195
                                    Reference
Prediction                           Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                              0         0
  Cancer                                                  6         6
  Cystic.fibrosis                                       193        33
  Diabetes                                                2         1
  Hemochromatosis                                       114        84
  Lebers.hereditary.optic.neuropathy                      0         0
  Leigh.syndrome                                       1683       781
  Mitochondrial.myopathy                               1102       947
  Tay.Sachs                                             262       300

 Accuracy (average) : 0.34
```

```
# Check variable importance
lda_imp <- varImp(lda_fit, scale = FALSE)
lda_imp
```

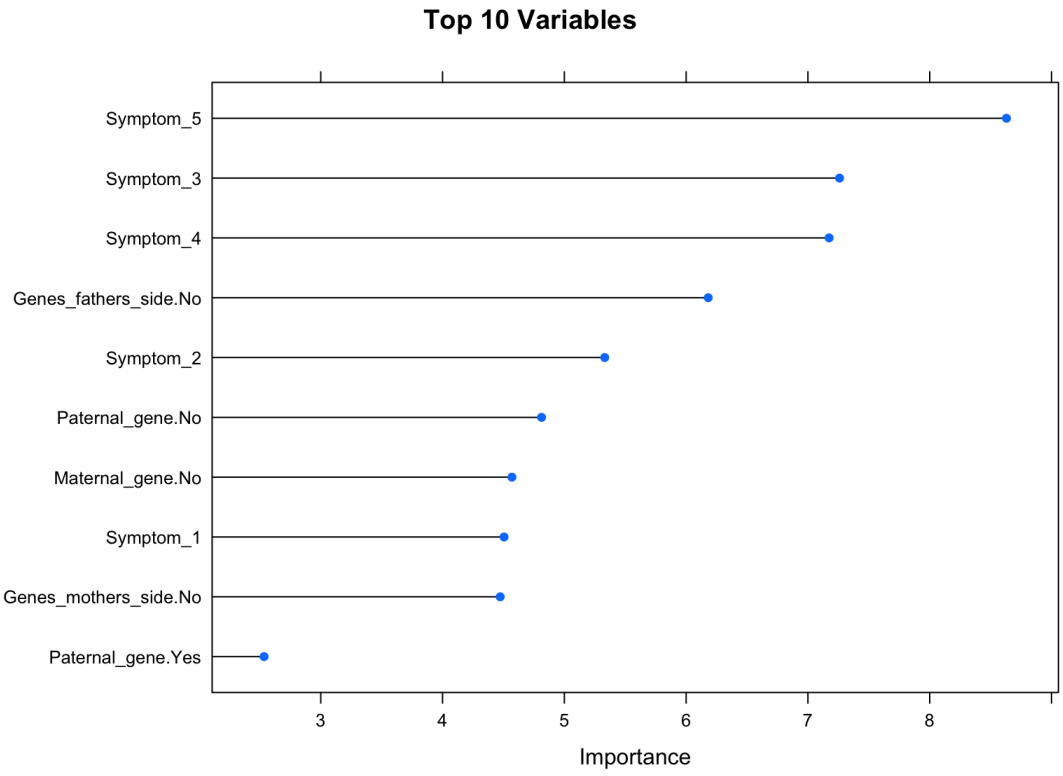
```
ROC curve variable importance

  variables are sorted by maximum importance across the classes
  only 20 most important variables shown (out of 69)

                                Alzheimers Cancer Cystic.fibrosis Diabetes Hemochromatosis Lebers.hereditary.optic.neuropathy
Leigh.syndrome
Symptom_5                            0.931  0.931           0.931    0.931           0.931                                0.931
0.931
Symptom_3                            0.914  0.914           0.914    0.914           0.914                                0.914
0.914
Symptom_4                            0.884  0.884           0.884    0.884           0.884                                0.884
0.884
Symptom_2                            0.844  0.844           0.844    0.844           0.844                                0.844
0.844
Symptom_1                            0.808  0.808           0.808    0.808           0.808                                0.808
0.808
Genes_mothers_side.Yes               0.787  0.787           0.787    0.787           0.787                                0.787
0.787
Genes_mothers_side.No                0.787  0.787           0.787    0.787           0.787                                0.787
0.787
Paternal_gene.Yes                    0.774  0.774           0.774    0.774           0.774                                0.774
0.774
Paternal_gene.No                     0.774  0.774           0.774    0.774           0.774                                0.774
0.774
Genes_fathers_side.Yes               0.764  0.764           0.764    0.764           0.764                                0.764
0.764
Genes_fathers_side.No                0.757  0.757           0.757    0.757           0.757                                0.757
0.757
Maternal_gene.Yes                    0.737  0.737           0.737    0.737           0.737                                0.737
0.737
Maternal_gene.No                     0.711  0.711           0.711    0.711           0.711                                0.711
0.711
Fathers_age                          0.570  0.570           0.570    0.570           0.570                                0.570
0.570
Patient_Age                          0.562  0.562           0.562    0.562           0.562                                0.562
0.562
Follow_up.High                       0.556  0.556           0.556    0.556           0.556                                0.556
0.556
Blood_test_result.inconclusive       0.558  0.558           0.558    0.558           0.558                                0.558
0.558
HO_substance_abuse.notprovided       0.556  0.556           0.556    0.556           0.556                                0.556
0.556
Blood_cell_count                     0.555  0.555           0.555    0.555           0.555                                0.555
0.555
Birth_defects.Multiple               0.526  0.523           0.523    0.523           0.523                                0.528
0.523
                                Mitochondrial.myopathy Tay.Sachs
Symptom_5                                        0.931     0.632
Symptom_3                                        0.914     0.629
Symptom_4                                        0.884     0.606
Symptom_2                                        0.844     0.593
Symptom_1                                        0.808     0.586
Genes_mothers_side.Yes                           0.787     0.578
Genes_mothers_side.No                            0.787     0.578
Paternal_gene.Yes                                0.774     0.587
Paternal_gene.No                                 0.774     0.587
Genes_fathers_side.Yes                           0.764     0.595
Genes_fathers_side.No                            0.757     0.600
Maternal_gene.Yes                                0.737     0.550
Maternal_gene.No                                 0.711     0.544
Fathers_age                                      0.570     0.530
Patient_Age                                      0.562     0.524
Follow_up.High                                   0.561     0.513
Blood_test_result.inconclusive                   0.558     0.526
HO_substance_abuse.notprovided                   0.556     0.516
Blood_cell_count                                 0.555     0.503
Birth_defects.Multiple                           0.550     0.526
```

Logistic Regression Model

```
# Train LR model
set.seed(476)
invisible(capture.output(
  lr_fit <- train(x = train_df[ , -target_col, drop = FALSE],
                  y = train_df$Disorder_Subclass,
                  method = "multinom",
                  metric = "ROC",
                  trControl = fit_control)
))
lr_cm <- confusionMatrix(lr_fit, norm = "none")
lr_cm
```

```
Cross-Validated (10 fold) Confusion Matrix

(entries are un-normalized aggregated counts)

                                     Reference
Prediction                            Alzheimers Cancer Cystic.fibrosis Diabetes Hemochromatosis Lebers.hereditary.optic.neur
opathy Leigh.syndrome
  Alzheimers                                   0      0               0        0               0
0              0
  Cancer                                       0      0               0        0               0
0              0
  Cystic.fibrosis                             70      0            1332      875               1
331            587
  Diabetes                                    34      0              81       87               0
52              4
  Hemochromatosis                              0     10               7        1              31
1             34
  Lebers.hereditary.optic.neuropathy           0      0               0        0               0
0              0
  Leigh.syndrome                              13      3             966      338             244
83           2378
  Mitochondrial.myopathy                       2     25             193       75             523
19            732
  Tay.Sachs                                    0     36              43       19             234
0            180
                                     Reference
Prediction                            Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                               0         0
  Cancer                                                   0         0
  Cystic.fibrosis                                        139        17
  Diabetes                                                 1         0
  Hemochromatosis                                         36        28
  Lebers.hereditary.optic.neuropathy                       0         0
  Leigh.syndrome                                        1886       906
  Mitochondrial.myopathy                                1057       922
  Tay.Sachs                                              243       279

 Accuracy (average) : 0.3407
```

```r
# Check variable importance
lr_imp <- varImp(lr_fit, scale = FALSE)
plot(lr_imp, top = 10, main = "Top 10 Variables")
```

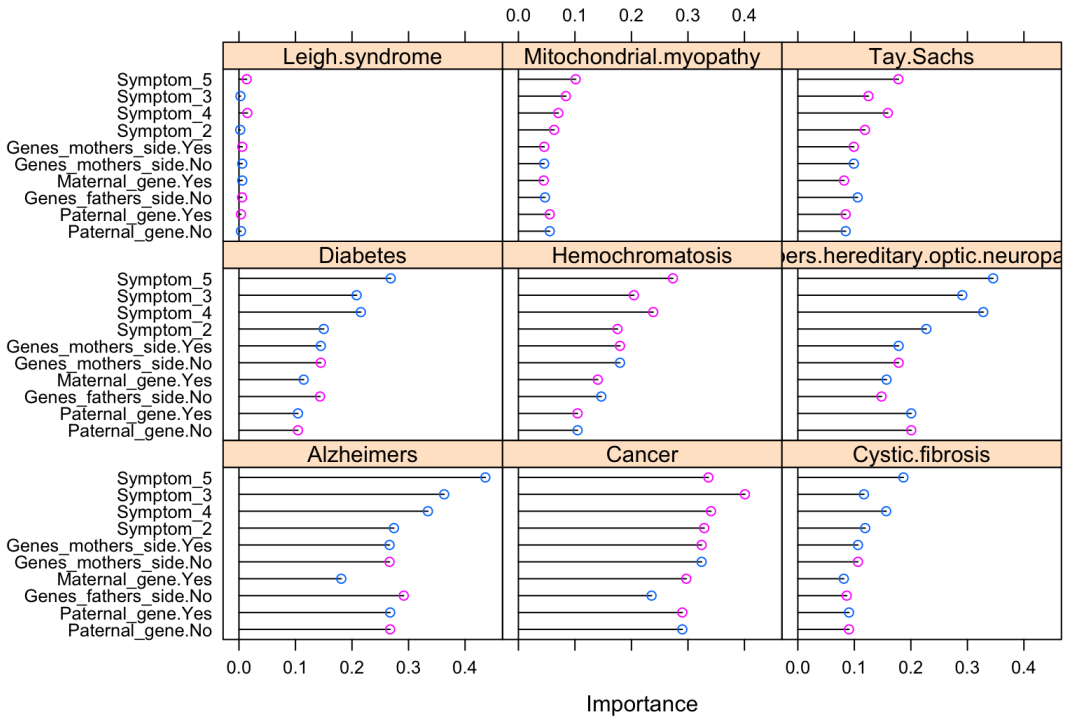


Nearest shrunken Centroids Model

```
# Train NSC model
set.seed(476)
invisible(capture.output(
  nsc_fit <- train(x = train_df[ , -target_col, drop = FALSE],
                   y = train_df$Disorder_Subclass,
                   method = "pam",
                   preProc = c("center", "scale"),
                   tuneGrid = data.frame(threshold = seq(0, 25, length = 30)),
                   metric = "ROC",
                   trControl = fit_control)
))
nsc_cm <- confusionMatrix(nsc_fit, norm = "none")
nsc_cm
```

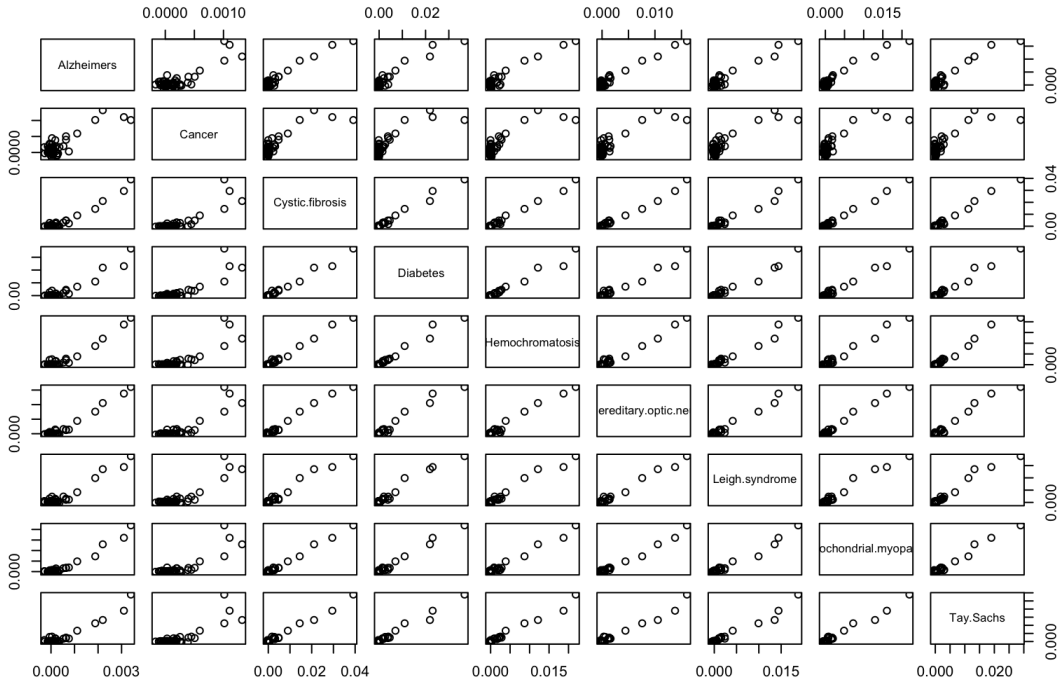
```
Cross-Validated (10 fold) Confusion Matrix

(entries are un-normalized aggregated counts)

                                    Reference
Prediction                           Alzheimers Cancer Cystic.fibrosis Diabetes Hemochromatosis Lebers.hereditary.optic.neur
opathy Leigh.syndrome
  Alzheimers                                  0      0               0        0               0
0              0
  Cancer                                      0      0               0        0               0
0              0
  Cystic.fibrosis                            50      0             162      172               0
132             25
  Diabetes                                    0      0               0        0               0
0              0
  Hemochromatosis                             0      0               0        0               0
0              0
  Lebers.hereditary.optic.neuropathy          0      0               0        0               0
0              0
  Leigh.syndrome                             69     20            2374     1201             698
350           3561
  Mitochondrial.myopathy                      0     54              86       22             334
4            328
  Tay.Sachs                                   0      0               0        0               1
0              1
                                    Reference
Prediction                           Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                              0         0
  Cancer                                                  0         0
  Cystic.fibrosis                                         0         0
  Diabetes                                                0         0
  Hemochromatosis                                         0         0
  Lebers.hereditary.optic.neuropathy                      0         0
  Leigh.syndrome                                       2847      1679
  Mitochondrial.myopathy                                513       472
  Tay.Sachs                                               2         1

 Accuracy (average) : 0.2795
```

```
# Check variable importance
nsc_imp <- varImp(nsc_fit, scale = FALSE)
plot(nsc_imp, top = 10, main = "Top 10 Variables")
```



Random Forest Model

```
# Create Random Forest weight vector based on class priors
priors <- as.list(prop.table(table(train_df$Disorder_Subclass)))
wts <- data.frame(Disorder_Subclass = train_df$Disorder_Subclass, w = 0.0)
for (n in 1:length(priors))
  wts[wts$Disorder_Subclass == names(priors[n]), ]$w <- priors[[n]]

# Train the model (using defaults)
rf_fit <- randomForest(x = train_df[ , -target_col, drop = FALSE],
                       y = train_df$Disorder_Subclass,
                       xtest = test_df[ , -target_col, drop = FALSE],
                       ytest = test_df$Disorder_Subclass,
                       weights = as.vector(wts$w),
                       importance = TRUE)

# Simplify class names for more coherent confusion matrix, and output
for (n in 1:length(rownames(rf_fit$confusion)))
  rownames(rf_fit$confusion)[n] <- paste(rownames(rf_fit$confusion)[n], " (", AscToChar(64 + n), ")", sep = "")
for (n in 1:length(rownames(rf_fit$confusion)))
  colnames(rf_fit$confusion)[n] <- paste("Class", AscToChar(64 + n))
for (n in 1:length(rownames(rf_fit$test$confusion)))
  rownames(rf_fit$test$confusion)[n] <- paste(rownames(rf_fit$test$confusion)[n], " (", AscToChar(64 + n), ")", sep = "")
for (n in 1:length(rownames(rf_fit$test$confusion)))
  colnames(rf_fit$test$confusion)[n] <- paste("Class", AscToChar(64 + n))

# Check variable importance
rf_imp <- varImp(rf_fit, scale = FALSE)
plot(rf_imp, top = 10, main = "Top 10 Variables")
```

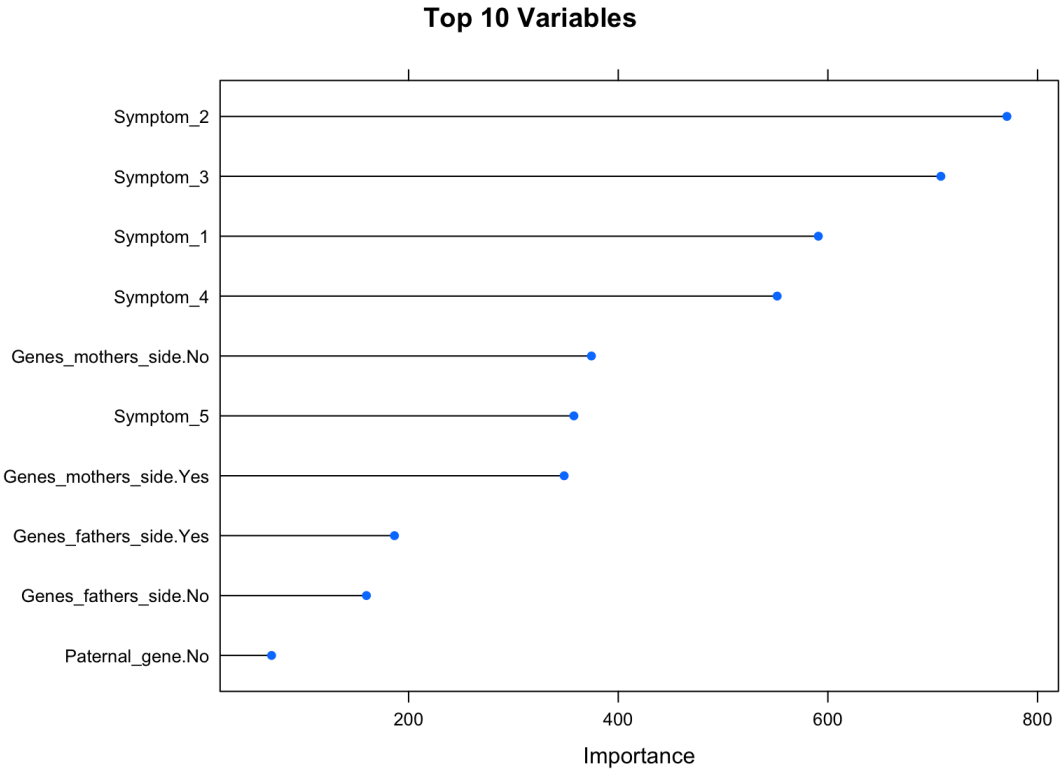


CART Model

```
# Train CART model
set.seed(476)
cart_fit <- train(x = train_df[ , -target_col, drop = FALSE],
                  y = train_df$Disorder_Subclass,
                  method = "rpart",
                  tuneLength = 30,
                  metric = "ROC",
                  trControl = fit_control)
cart_cm <- confusionMatrix(cart_fit, norm = "none")
cart_cm
```

```
Cross-Validated (10 fold) Confusion Matrix

(entries are un-normalized aggregated counts)

                                     Reference
Prediction                            Alzheimers Cancer Cystic.fibrosis Diabetes Hemochromatosis Lebers.hereditary.optic.neur
opathy Leigh.syndrome
  Alzheimers                                   0      0               0        0               0
0              0
  Cancer                                       0      0               0        0               0
0              0
  Cystic.fibrosis                             37      0            1201      743               0
198            527
  Diabetes                                    18      0             124      155               0
96              8
  Hemochromatosis                              0     33               2        1             151
0             13
  Lebers.hereditary.optic.neuropathy          56      0              49       94               0
116             1
  Leigh.syndrome                               7      0            1062      359             100
68           2154
  Mitochondrial.myopathy                       1     13             160       34             361
8             962
  Tay.Sachs                                    0     28              24        9             421
0             250
                                     Reference
Prediction                            Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                               0         0
  Cancer                                                   0         0
  Cystic.fibrosis                                        113         8
  Diabetes                                                 1         0
  Hemochromatosis                                         59       113
  Lebers.hereditary.optic.neuropathy                       0         0
  Leigh.syndrome                                        1549       586
  Mitochondrial.myopathy                                1180       877
  Tay.Sachs                                              460       568

 Accuracy (average) : 0.3645
```

```r
# Check variable importance
cart_imp <- varImp(cart_fit, scale = FALSE)
plot(cart_imp, top = 10, main = "Top 10 Variables")
```

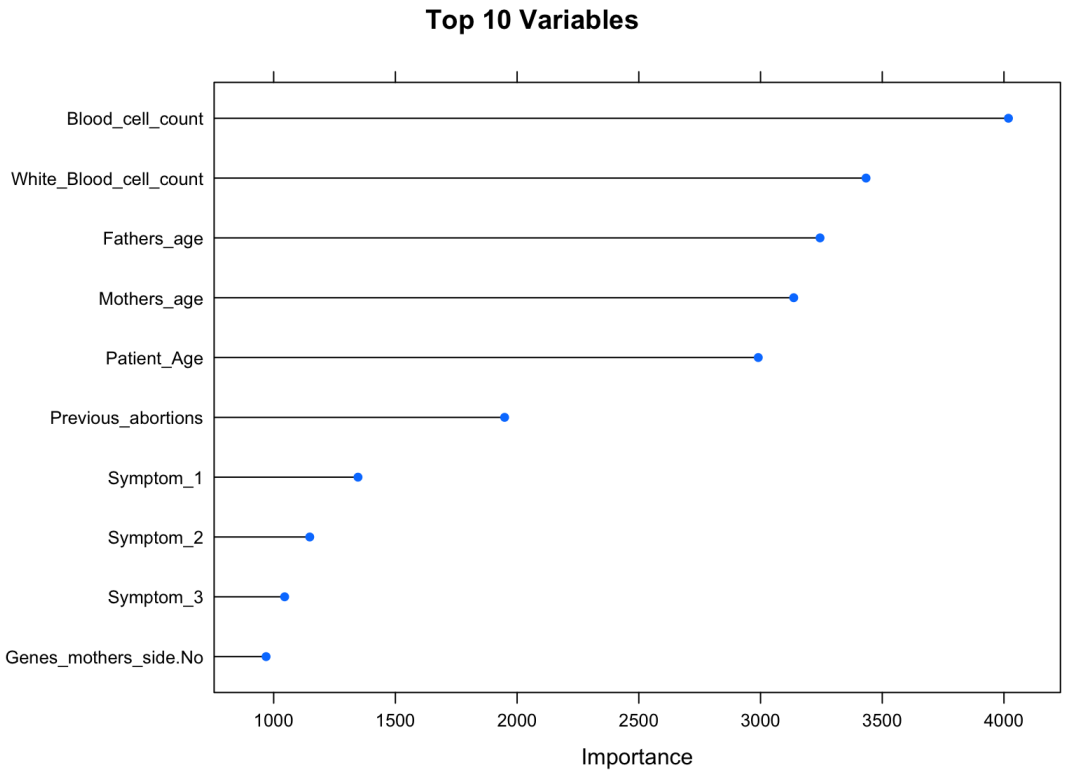


Bagged Trees Model

```
# Train BT model
bt_fit <- train(x = train_df[ , -target_col, drop = FALSE],
                y = train_df$Disorder_Subclass,
                method = "treebag",
                metric = "ROC",
                importance = TRUE,
                trControl=fit_control)
bt_cm <- confusionMatrix(bt_fit, norm = "none")
bt_cm
```

```
Cross-Validated (10 fold) Confusion Matrix

(entries are un-normalized aggregated counts)

                                    Reference
Prediction                           Alzheimers Cancer Cystic.fibrosis Diabetes Hemochromatosis Lebers.hereditary.optic.neur
opathy Leigh.syndrome
  Alzheimers                                  6      0               3        6               0
14              0
  Cancer                                      0      0               1        0               2
0               0
  Cystic.fibrosis                            24      0            1143      679               2
191           581
  Diabetes                                   46      0             401      339               0
168            89
  Hemochromatosis                             0     37               1        2             241
0             56
  Lebers.hereditary.optic.neuropathy         38      0              50       69               0
71              7
  Leigh.syndrome                              5      1             755      231             111
32           1760
  Mitochondrial.myopathy                      0     15             238       59             267
8           1108
  Tay.Sachs                                   0     21              30       10             410
2            314
                                    Reference
Prediction                           Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                              0         0
  Cancer                                                  0         2
  Cystic.fibrosis                                       149        16
  Diabetes                                               16         0
  Hemochromatosis                                       161       258
  Lebers.hereditary.optic.neuropathy                      0         0
  Leigh.syndrome                                       1243       469
  Mitochondrial.myopathy                               1204       804
  Tay.Sachs                                             589       603

 Accuracy (average) : 0.3541
```

```
# Check variable importance
bt_imp <- varImp(bt_fit, scale = FALSE)
plot(bt_imp, top = 10, main = "Top 10 Variables")
```

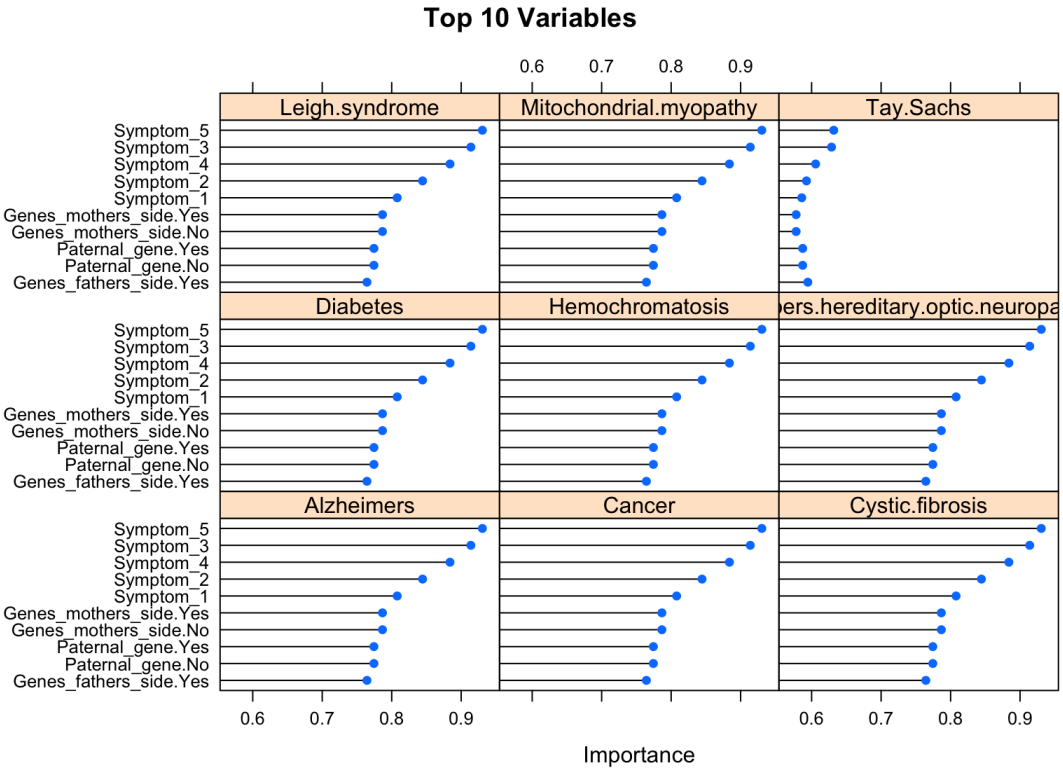


KNN Model

```
# Train KNN model
set.seed(476)
knn_fit <- train(x = train_df[ , -target_col, drop = FALSE],
                 y = train_df$Disorder_Subclass,
                 method = "knn",
                 metric = "ROC",
                 trControl = fit_control)
knn_cm <- confusionMatrix(knn_fit, norm = "none")
knn_cm
```

```
Cross-Validated (10 fold) Confusion Matrix

(entries are un-normalized aggregated counts)

                                     Reference
Prediction                            Alzheimers Cancer Cystic.fibrosis Diabetes Hemochromatosis Lebers.hereditary.optic.neur
opathy Leigh.syndrome
  Alzheimers                                   0      0               0        0               0
0              0
  Cancer                                       0      0               0        0               0
0              0
  Cystic.fibrosis                             30     10             523      273             117
129            597
  Diabetes                                    12      0             150       85              28
40            170
  Hemochromatosis                              0      6              49       24              49
4            114
  Lebers.hereditary.optic.neuropathy           1      0              14       14               2
4             18
  Leigh.syndrome                              46     21             932      533             356
165          1431
  Mitochondrial.myopathy                      23     23             709      334             325
103          1107
  Tay.Sachs                                    7     14             245      132             156
41            478
                                     Reference
Prediction                            Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                               0         0
  Cancer                                                   1         0
  Cystic.fibrosis                                        429       258
  Diabetes                                               103        42
  Hemochromatosis                                        112        65
  Lebers.hereditary.optic.neuropathy                       8         5
  Leigh.syndrome                                        1168       765
  Mitochondrial.myopathy                                1074       689
  Tay.Sachs                                              467       328

 Accuracy (average) : 0.2305
```

```r
# Check variable importance
knn_imp <- varImp(knn_fit, scale = FALSE)
plot(knn_imp, top = 10, main = "Top 10 Variables")
```



Model Validation / Evaluation

```r
# Validate models
if (lda_eval) {
  print("Linear Disrciminate Analysis")
  lda_pred <- predict(lda_fit, test_df[ , -target_col, drop = FALSE])
  lda_pred_cm = confusionMatrix(lda_pred, test_df$Disorder_Subclass)
  lda_pred_cm
}
```

```
[1] "Linear Disrciminate Analysis"
```

```
Confusion Matrix and Statistics

                                  Reference
Prediction                         Alzheimers Cancer Cystic.fibrosis Diabetes Hemochromatosis Lebers.hereditary.optic.neur
opathy Leigh.syndrome
  Alzheimers                                1      0               0        1               0
0              0
  Cancer                                    0      0               1        0               2
0              0
  Cystic.fibrosis                          12      0             340      229               0
77           137
  Diabetes                                 10      0              23       33               0
17             2
  Hemochromatosis                           0      4               4        3              24
0             15
  Lebers.hereditary.optic.neuropathy        2      0               2        2               0
2              0
  Leigh.syndrome                            4      1             208       61              42
18           582
  Mitochondrial.myopathy                    0      6              61       14             134
5            188
  Tay.Sachs                                 0      7              16        5              56
2             54
                                  Reference
Prediction                         Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                            0         0
  Cancer                                                1         0
  Cystic.fibrosis                                      53         5
  Diabetes                                              0         0
  Hemochromatosis                                      26        29
  Lebers.hereditary.optic.neuropathy                    0         0
  Leigh.syndrome                                      408       208
  Mitochondrial.myopathy                              279       217
  Tay.Sachs                                            73        79

Overall Statistics

               Accuracy : 0.354
                 95% CI : (0.339, 0.37)
    No Information Rate : 0.258
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.181

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Alzheimers Class: Cancer Class: Cystic.fibrosis Class: Diabetes Class: Hemochromatosis
Sensitivity                   0.034483       0.00000                0.5191         0.09483                0.09302
Specificity                   0.999734       0.99894                0.8361         0.98487                0.97703
Pos Pred Value                0.500000       0.00000                0.3986         0.38824                0.22857
Neg Pred Value                0.992598       0.99524                0.8926         0.91486                0.93641
Prevalence                    0.007662       0.00476                0.1731         0.09194                0.06816
Detection Rate                0.000264       0.00000                0.0898         0.00872                0.00634
Detection Prevalence          0.000528       0.00106                0.2254         0.02246                0.02774
Balanced Accuracy             0.517108       0.49947                0.6776         0.53985                0.53503
                     Class: Lebers.hereditary.optic.neuropathy Class: Leigh.syndrome Class: Mitochondrial.myopathy Class: Ta
y.Sachs
Sensitivity                                          0.016529                0.595                        0.3321
0.1468
Specificity                                          0.998362                0.662                        0.7878
0.9344
Pos Pred Value                                       0.250000                0.380                        0.3086
0.2705
Neg Pred Value                                       0.968494                0.824                        0.8053
0.8686
Prevalence                                           0.031968                0.258                        0.2219
0.1421
Detection Rate                                       0.000528                0.154                        0.0737
0.0209
Detection Prevalence                                 0.002114                0.405                        0.2388
0.0771
Balanced Accuracy                                    0.507446                0.628                        0.5600
0.5406
```

```
if (lr_eval) {
  print("Logistic Regression")
  lr_pred <- predict(lr_fit, test_df[ , -target_col, drop = FALSE])
  lr_pred_cm = confusionMatrix(lr_pred, test_df$Disorder_Subclass)
  lr_pred_cm
}
```

```
[1] "Logistic Regression"
```

```
Confusion Matrix and Statistics

                                     Reference
Prediction                            Alzheimers Cancer Cystic.fibrosis Diabetes Hemochromatosis Lebers.hereditary.optic.neur
opathy Leigh.syndrome
  Alzheimers                                   0      0               0        0               0
0              0
  Cancer                                       0      0               0        0               0
0              0
  Cystic.fibrosis                             17      0             334      233               0
82            133
  Diabetes                                     8      0              19       31               0
14              2
  Hemochromatosis                              0      1               2        1              14
0              9
  Lebers.hereditary.optic.neuropathy           0      0               0        0               0
0              0
  Leigh.syndrome                               4      1             217       60              40
18            569
  Mitochondrial.myopathy                       0      7              65       17             146
5            217
  Tay.Sachs                                    0      9              18        6              58
2             48
                                     Reference
Prediction                            Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                               0         0
  Cancer                                                   0         0
  Cystic.fibrosis                                         40         5
  Diabetes                                                 0         0
  Hemochromatosis                                         13        18
  Lebers.hereditary.optic.neuropathy                       0         0
  Leigh.syndrome                                         408       205
  Mitochondrial.myopathy                                 309       234
  Tay.Sachs                                               70        76

Overall Statistics

               Accuracy : 0.352
                 95% CI : (0.337, 0.368)
    No Information Rate : 0.258
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.175

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Alzheimers Class: Cancer Class: Cystic.fibrosis Class: Diabetes Class: Hemochromatosis
Sensitivity                    0.00000       0.00000                 0.5099         0.08908                 0.0543
Specificity                    1.00000       1.00000                 0.8371         0.98749                 0.9875
Pos Pred Value                     NaN           NaN                 0.3957         0.41892                 0.2414
Neg Pred Value                 0.99234       0.99524                 0.8909         0.91458                 0.9345
Prevalence                     0.00766       0.00476                 0.1731         0.09194                 0.0682
Detection Rate                 0.00000       0.00000                 0.0882         0.00819                 0.0037
Detection Prevalence           0.00000       0.00000                 0.2230         0.01955                 0.0153
Balanced Accuracy              0.50000       0.50000                 0.6735         0.53828                 0.5209
                     Class: Lebers.hereditary.optic.neuropathy Class: Leigh.syndrome Class: Mitochondrial.myopathy Class: Ta
y.Sachs
Sensitivity                                             0.000                0.582                        0.3679
0.1413
Specificity                                             1.000                0.660                        0.7654
0.9350
Pos Pred Value                                            NaN                0.374                        0.3090
0.2648
Neg Pred Value                                          0.968                0.819                        0.8093
0.8679
Prevalence                                              0.032                0.258                        0.2219
0.1421
Detection Rate                                          0.000                0.150                        0.0816
0.0201
Detection Prevalence                                    0.000                0.402                        0.2642
0.0758
Balanced Accuracy                                       0.500                0.621                        0.5666
0.5381
```

```
if (nsc_eval) {
  print("Nearest Shrunken Centroids")
  nsc_pred <- predict(nsc_fit, test_df[ , -target_col, drop = FALSE])
  nsc_pred_cm = confusionMatrix(nsc_pred, test_df$Disorder_Subclass)
  nsc_pred_cm
}
```

```
[1] "Nearest Shrunken Centroids"
```

```
Confusion Matrix and Statistics

                                  Reference
Prediction                         Alzheimers Cancer Cystic.fibrosis Diabetes Hemochromatosis Lebers.hereditary.optic.neur
opathy Leigh.syndrome
  Alzheimers                                0      0               0        0               0
0              0
  Cancer                                    0      0               0        0               0
0              0
  Cystic.fibrosis                          14      0              39       45               0
29              3
  Diabetes                                  0      0               0        0               0
0              0
  Hemochromatosis                           0      0               0        0               0
0              0
  Lebers.hereditary.optic.neuropathy        0      0               0        0               0
0              0
  Leigh.syndrome                           15      7             597      298             162
91            897
  Mitochondrial.myopathy                    0     11              19        5              96
1             78
  Tay.Sachs                                 0      0               0        0               0
0              0
                                  Reference
Prediction                         Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                            0         0
  Cancer                                                0         0
  Cystic.fibrosis                                       0         0
  Diabetes                                              0         0
  Hemochromatosis                                       0         0
  Lebers.hereditary.optic.neuropathy                    0         0
  Leigh.syndrome                                      738       418
  Mitochondrial.myopathy                              102       120
  Tay.Sachs                                             0         0

Overall Statistics

               Accuracy : 0.274
                 95% CI : (0.26, 0.289)
    No Information Rate : 0.258
    P-Value [Acc > NIR] : 0.014

                  Kappa : 0.031

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Alzheimers Class: Cancer Class: Cystic.fibrosis Class: Diabetes Class: Hemochromatosis
Sensitivity                    0.00000       0.00000                0.0595          0.0000                 0.0000
Specificity                    1.00000       1.00000                0.9709          1.0000                 1.0000
Pos Pred Value                     NaN           NaN                0.3000             NaN                    NaN
Neg Pred Value                 0.99234       0.99524                0.8315          0.9081                 0.9318
Prevalence                     0.00766       0.00476                0.1731          0.0919                 0.0682
Detection Rate                 0.00000       0.00000                0.0103          0.0000                 0.0000
Detection Prevalence           0.00000       0.00000                0.0343          0.0000                 0.0000
Balanced Accuracy              0.50000       0.50000                0.5152          0.5000                 0.5000
                     Class: Lebers.hereditary.optic.neuropathy Class: Leigh.syndrome Class: Mitochondrial.myopathy Class: Ta
y.Sachs
Sensitivity                                              0.000                0.917                        0.1214
0.000
Specificity                                              1.000                0.171                        0.8879
1.000
Pos Pred Value                                             NaN                0.278                        0.2361
NaN
Neg Pred Value                                           0.968                0.856                        0.7799
0.858
Prevalence                                               0.032                0.258                        0.2219
0.142
Detection Rate                                           0.000                0.237                        0.0269
0.000
Detection Prevalence                                     0.000                0.852                        0.1141
0.000
Balanced Accuracy                                        0.500                0.544                        0.5047
0.500
```

```
if (rf_eval) {
  print("Random Forest")
  rf_fit
}
```

```
[1] "Random Forest"
```

```
Call:
 randomForest(x = train_df[, -target_col, drop = FALSE], y = train_df$Disorder_Subclass,      xtest = test_df[, -target_col,
drop = FALSE], ytest = test_df$Disorder_Subclass,      weights = as.vector(wts$w), importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 8

        OOB estimate of  error rate: 66%
Confusion matrix:
                                    Class A Class B Class C Class D Class E Class F Class G Class H Class I class.error
Alzheimers (A)                            0       0     104       5       0       0      10       0       0         1.0
Cancer (B)                                0       0       0       0       0       0       2      47      25         1.0
Cystic.fibrosis (C)                       0       0    1167       4       0       0    1326     123       2         0.6
Diabetes (D)                              0       0     921      12       0       0     435      26       1         1.0
Hemochromatosis (E)                       0       0       0       0       0       0     218     687     128         1.0
Lebers.hereditary.optic.neuropathy (F)    0       0     403       8       0       0      71       4       0         1.0
Leigh.syndrome (G)                        0       0     327       0       0       0    2664     881      43         0.3
Mitochondrial.myopathy (H)                0       0      49       0       0       0    1950    1258     105         0.6
Tay.Sachs (I)                             0       0       0       0       0       0     832    1192     128         0.9
               Test set error rate: 64%
Confusion matrix:
                                    Class A Class B Class C Class D Class E Class F Class G Class H Class I class.error
Alzheimers (A)                            0       0      27       0       0       0       2       0       0         1.0
Cancer (B)                                0       0       0       0       0       0       0      13       5         1.0
Cystic.fibrosis (C)                       0       0     291       0       0       0     348      16       0         0.6
Diabetes (D)                              0       0     243       1       0       0     100       4       0         1.0
Hemochromatosis (E)                       0       0       0       0       0       0      47     188      23         1.0
Lebers.hereditary.optic.neuropathy (F)    0       0      93       3       0       0      23       2       0         1.0
Leigh.syndrome (G)                        0       0      48       0       0       0     735     193       2         0.2
Mitochondrial.myopathy (H)                0       0      10       0       0       0     496     316      18         0.6
Tay.Sachs (I)                             0       0       0       0       0       0     212     295      31         0.9
```

```
if (cart_eval) {
  print("CART")
  # Validate model vs. test data
  cart_pred <- predict(cart_fit, test_df[ , -target_col, drop = FALSE])
  cart_pred_cm = confusionMatrix(cart_pred, test_df$Disorder_Subclass)
  cart_pred_cm
}
```

```
[1] "CART"
```

```
Confusion Matrix and Statistics

                                  Reference
Prediction                         Alzheimers Cancer Cystic.fibrosis Diabetes Hemochromatosis Lebers.hereditary.optic.neur
opathy Leigh.syndrome
  Alzheimers                                0      0               0        0               0
0              0
  Cancer                                    0      0               0        0               0
0              0
  Cystic.fibrosis                          10      0             319      187               0
59           119
  Diabetes                                  3      0              25       41               0
19             1
  Hemochromatosis                           0      6               0        1              34
0              5
  Lebers.hereditary.optic.neuropathy       14      0               7       31               0
19             0
  Leigh.syndrome                            2      0             271       83              28
20           617
  Mitochondrial.myopathy                    0      2              28        4             103
4            183
  Tay.Sachs                                 0     10               5        1              93
0             53
                                  Reference
Prediction                         Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                            0         0
  Cancer                                                0         0
  Cystic.fibrosis                                      29         3
  Diabetes                                              0         0
  Hemochromatosis                                      16        29
  Lebers.hereditary.optic.neuropathy                    0         0
  Leigh.syndrome                                      414       144
  Mitochondrial.myopathy                              268       215
  Tay.Sachs                                           113       147

Overall Statistics

               Accuracy : 0.382
                 95% CI : (0.366, 0.397)
    No Information Rate : 0.258
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.219

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Alzheimers Class: Cancer Class: Cystic.fibrosis Class: Diabetes Class: Hemochromatosis
Sensitivity                    0.00000       0.00000                0.4870          0.1178                0.13178
Specificity                    1.00000       1.00000                0.8700          0.9860                0.98384
Pos Pred Value                     NaN           NaN                0.4394          0.4607                0.37363
Neg Pred Value                 0.99234       0.99524                0.8902          0.9169                0.93936
Prevalence                     0.00766       0.00476                0.1731          0.0919                0.06816
Detection Rate                 0.00000       0.00000                0.0843          0.0108                0.00898
Detection Prevalence           0.00000       0.00000                0.1918          0.0235                0.02404
Balanced Accuracy              0.50000       0.50000                0.6785          0.5519                0.55781
                     Class: Lebers.hereditary.optic.neuropathy Class: Leigh.syndrome Class: Mitochondrial.myopathy Class: Ta
y.Sachs
Sensitivity                                            0.15702                0.631                        0.3190
0.2732
Specificity                                            0.98581                0.657                        0.8170
0.9153
Pos Pred Value                                         0.26761                0.391                        0.3321
0.3483
Neg Pred Value                                         0.97254                0.836                        0.8079
0.8837
Prevalence                                             0.03197                0.258                        0.2219
0.1421
Detection Rate                                         0.00502                0.163                        0.0708
0.0388
Detection Prevalence                                   0.01876                0.417                        0.2132
0.1115
Balanced Accuracy                                      0.57142                0.644                        0.5680
0.5943
```

```
if (bt_eval) {
  print("Bagged Trees")
  bt_pred <- predict(bt_fit, test_df[ , -target_col, drop = FALSE])
  bt_pred_cm = confusionMatrix(bt_pred, test_df$Disorder_Subclass)
  bt_pred_cm
}
```

```
[1] "Bagged Trees"
```

```
Confusion Matrix and Statistics

                                     Reference
Prediction                            Alzheimers Cancer Cystic.fibrosis Diabetes Hemochromatosis Lebers.hereditary.optic.neur
opathy Leigh.syndrome
  Alzheimers                                   2      0               1        2               0
0              0
  Cancer                                       0      0               0        0               0
0              0
  Cystic.fibrosis                             11      0             303      162               0
45           127
  Diabetes                                     9      0              92      101               0
42            23
  Hemochromatosis                              0     12               0        2              60
0             13
  Lebers.hereditary.optic.neuropathy           7      0               8       24               0
18             0
  Leigh.syndrome                               0      0             189       40              20
11           458
  Mitochondrial.myopathy                       0      3              55       15              78
4            293
  Tay.Sachs                                    0      3               7        2             100
1             64
                                     Reference
Prediction                            Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                               0         0
  Cancer                                                   0         0
  Cystic.fibrosis                                         36         6
  Diabetes                                                 4         0
  Hemochromatosis                                         41        62
  Lebers.hereditary.optic.neuropathy                       0         0
  Leigh.syndrome                                         310       107
  Mitochondrial.myopathy                                 307       213
  Tay.Sachs                                              142       150

Overall Statistics

               Accuracy : 0.37
                 95% CI : (0.354, 0.385)
    No Information Rate : 0.258
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.218

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Alzheimers Class: Cancer Class: Cystic.fibrosis Class: Diabetes Class: Hemochromatosis
Sensitivity                   0.068966       0.00000                0.4626          0.2902                 0.2326
Specificity                   0.999201       1.00000                0.8764          0.9505                 0.9631
Pos Pred Value                0.400000           NaN                0.4391          0.3727                 0.3158
Neg Pred Value                0.992857       0.99524                0.8863          0.9297                 0.9449
Prevalence                    0.007662       0.00476                0.1731          0.0919                 0.0682
Detection Rate                0.000528       0.00000                0.0801          0.0267                 0.0159
Detection Prevalence          0.001321       0.00000                0.1823          0.0716                 0.0502
Balanced Accuracy             0.534083       0.50000                0.6695          0.6204                 0.5978
                     Class: Lebers.hereditary.optic.neuropathy Class: Leigh.syndrome Class: Mitochondrial.myopathy Class: Ta
y.Sachs
Sensitivity                                            0.14876                0.468                        0.3655
0.2788
Specificity                                            0.98936                0.759                        0.7756
0.9018
Pos Pred Value                                         0.31579                0.404                        0.3171
0.3198
Neg Pred Value                                         0.97237                0.804                        0.8108
0.8830
Prevalence                                             0.03197                0.258                        0.2219
0.1421
Detection Rate                                         0.00476                0.121                        0.0811
0.0396
Detection Prevalence                                   0.01506                0.300                        0.2557
0.1239
Balanced Accuracy                                      0.56906                0.614                        0.5705
0.5903
```

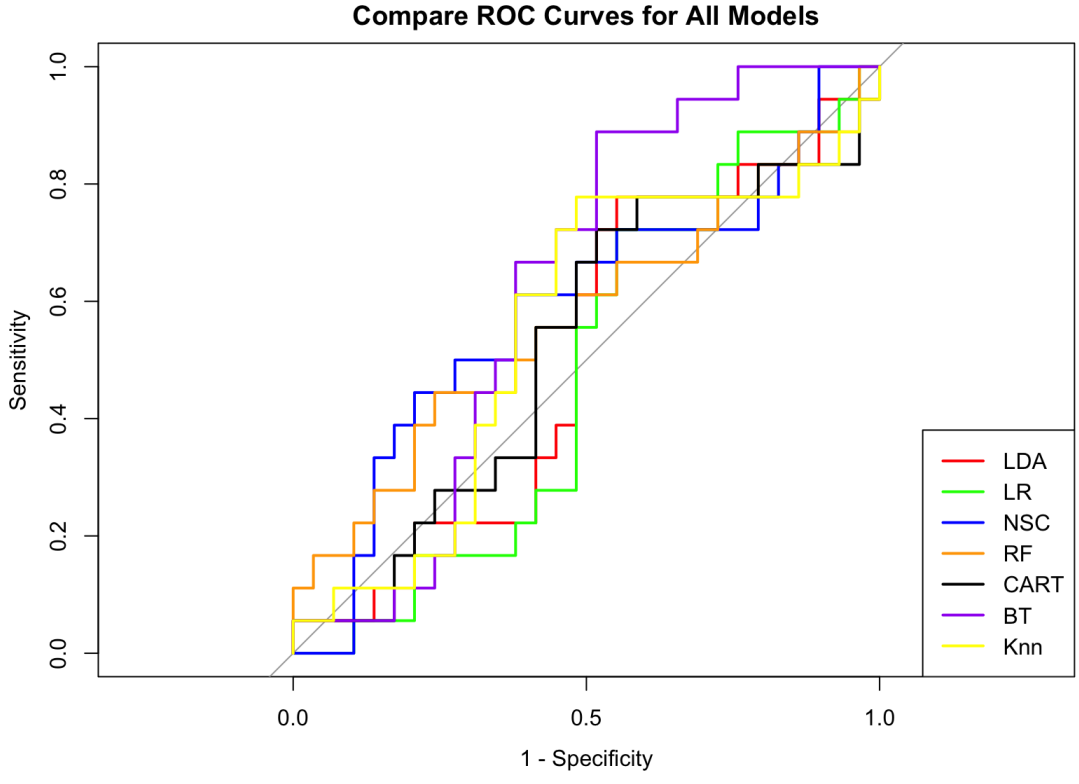
```
if (knn_eval) {
  print("KNN")
  knn_pred <- predict(knn_fit, test_df[ , -target_col, drop = FALSE])
  knn_pred_cm = confusionMatrix(knn_pred, test_df$Disorder_Subclass)
  knn_pred_cm
}
```

```
[1] "KNN"
```

Confusion Matrix and Statistics

```
                                  Reference
Prediction                         Alzheimers Cancer Cystic.fibrosis Diabetes Hemochromatosis Lebers.hereditary.optic.neur
opathy Leigh.syndrome
  Alzheimers                                0      0               0        0               0
0              0
  Cancer                                    0      0               0        0               0
0              0
  Cystic.fibrosis                           6      1             111       72              26
18           134
  Diabetes                                  4      0              36       20               3
9             39
  Hemochromatosis                           0      2              13        2              14
1             24
  Lebers.hereditary.optic.neuropathy        1      0               3        4               2
4              6
  Leigh.syndrome                           12      7             232      115              85
58           376
  Mitochondrial.myopathy                    5      8             186       92              91
22           276
  Tay.Sachs                                 1      0              74       43              37
9            123
                                  Reference
Prediction                         Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                            0         0
  Cancer                                                0         0
  Cystic.fibrosis                                     122        78
  Diabetes                                             27        11
  Hemochromatosis                                      34        18
  Lebers.hereditary.optic.neuropathy                    1         1
  Leigh.syndrome                                      287       180
  Mitochondrial.myopathy                              249       160
  Tay.Sachs                                           120        90

Overall Statistics

               Accuracy : 0.228
                 95% CI : (0.215, 0.242)
    No Information Rate : 0.258
    P-Value [Acc > NIR] : 1

                  Kappa : 0.027

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Alzheimers Class: Cancer Class: Cystic.fibrosis Class: Diabetes Class: Hemochromatosis
Sensitivity                    0.00000       0.00000                0.1695         0.05747                 0.0543
Specificity                    1.00000       1.00000                0.8540         0.96247                 0.9733
Pos Pred Value                     NaN           NaN                0.1954         0.13423                 0.1296
Neg Pred Value                 0.99234       0.99524                0.8309         0.90979                 0.9336
Prevalence                     0.00766       0.00476                0.1731         0.09194                 0.0682
Detection Rate                 0.00000       0.00000                0.0293         0.00528                 0.0037
Detection Prevalence           0.00000       0.00000                0.1501         0.03937                 0.0285
Balanced Accuracy              0.50000       0.50000                0.5117         0.50997                 0.5138
                     Class: Lebers.hereditary.optic.neuropathy Class: Leigh.syndrome Class: Mitochondrial.myopathy Class: Ta
y.Sachs
Sensitivity                                            0.03306                0.3845                         0.2964
0.1673
Specificity                                            0.99509                0.6523                         0.7148
0.8747
Pos Pred Value                                         0.18182                0.2781                         0.2287
0.1811
Neg Pred Value                                         0.96891                0.7526                         0.7808
0.8637
Prevalence                                             0.03197                0.2584                         0.2219
0.1421
Detection Rate                                         0.00106                0.0993                         0.0658
0.0238
Detection Prevalence                                   0.00581                0.3572                         0.2877
0.1313
Balanced Accuracy                                      0.51407                0.5184                         0.5056
0.5210
```

```r
## Plot the ROC curve for the hold-out set
if (lda_eval) {
  lda_roc <- multiclass.roc(response = test_df$Disorder_Subclass,
                            predictor = order(lda_pred))
  plot.roc(lda_roc$rocs[[1]], type = "s", col = 'red', legacy.axes = TRUE,
           main = "Compare ROC Curves for All Models")
}

if (lr_eval) {
  lr_roc <- multiclass.roc(response = test_df$Disorder_Subclass,
                           predictor = order(lr_pred))
  plot.roc(lr_roc$rocs[[1]], type = "s", add = TRUE, col = 'green', legacy.axes = TRUE)
}

if (nsc_eval) {
  nsc_roc <- multiclass.roc(response = test_df$Disorder_Subclass,
                            predictor = order(nsc_pred))
  plot.roc(nsc_roc$rocs[[1]], type = "s", add = TRUE, col = 'blue', legacy.axes = TRUE)
}

if (rf_eval) {
  rf_roc <- multiclass.roc(response = test_df$Disorder_Subclass,
                           predictor = order(rf_fit$test$predicted))
  plot.roc(rf_roc$rocs[[1]], type = "s", col = 'orange', add = TRUE, legacy.axes = TRUE)
  #par(pty = "s")
  #lines <- sapply(2:length(rocs), function(x) lines.roc(rocs[[x]], col = x))
  #dev <- dev.off()
}

if (cart_eval) {
  cart_roc <- multiclass.roc(response = test_df$Disorder_Subclass,
                             predictor = order(cart_pred))
  plot.roc(cart_roc$rocs[[1]], type = "s",col = 'black', add = TRUE, legacy.axes = TRUE)
}

if (bt_eval) {
  bt_roc <- multiclass.roc(response = test_df$Disorder_Subclass,
                           predictor = order(bt_pred))
  plot.roc(bt_roc$rocs[[1]], type = "s", col = 'purple',add = TRUE, legacy.axes = TRUE)
}

if (knn_eval) {
  knn_roc <- multiclass.roc(response = test_df$Disorder_Subclass,
                            predictor = order(knn_pred))
  plot.roc(knn_roc$rocs[[1]], type = "s", col = 'yellow', add = TRUE, legacy.axes = TRUE)
}

if (lda_eval | lr_eval | nsc_eval | rf_eval | cart_eval | bt_eval | knn_eval)
  legend("bottomright", legend = c("LDA", "LR", "NSC", "RF", "CART", "BT", "Knn"),
         col = c("red", "green", "blue", "orange", "black", "purple", "yellow"), lwd = 2)
```



1. University of San Diego, eoo@sandiego.edu (mailto:eoo@sandiego.edu)↩
2. University of San Diego, sbhattarai@sandiego.edu (mailto:sbhattarai@sandiego.edu)↩
3. University of San Diego, dfriesen@sandiego.edu (mailto:dfriesen@sandiego.edu)↩