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)
#####gd_df = read.csv(file="/Users/emmaoo/Desktop/train.csv", header= TRUE, col.names = cnames)

# 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 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",
               "Place.of.birth")
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
Birth.asphyxia       character             9%     5
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
Genetic.Disorder     character             9%     4
```

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",
                 "Birth.asphyxia",
                 "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",
                 "Birth.asphyxia",
                 "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",
                 "Genetic.Disorder")
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",
                 "Birth_asphyxia",
                 "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",
                 "Genetic_Disorder")
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
```

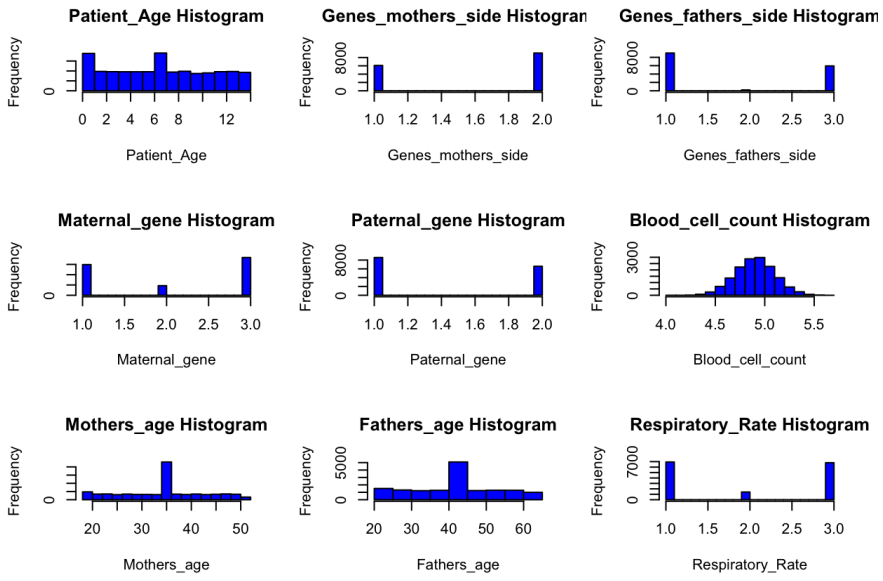
Statistical Characteristics (including distribution, skewness, outliers)

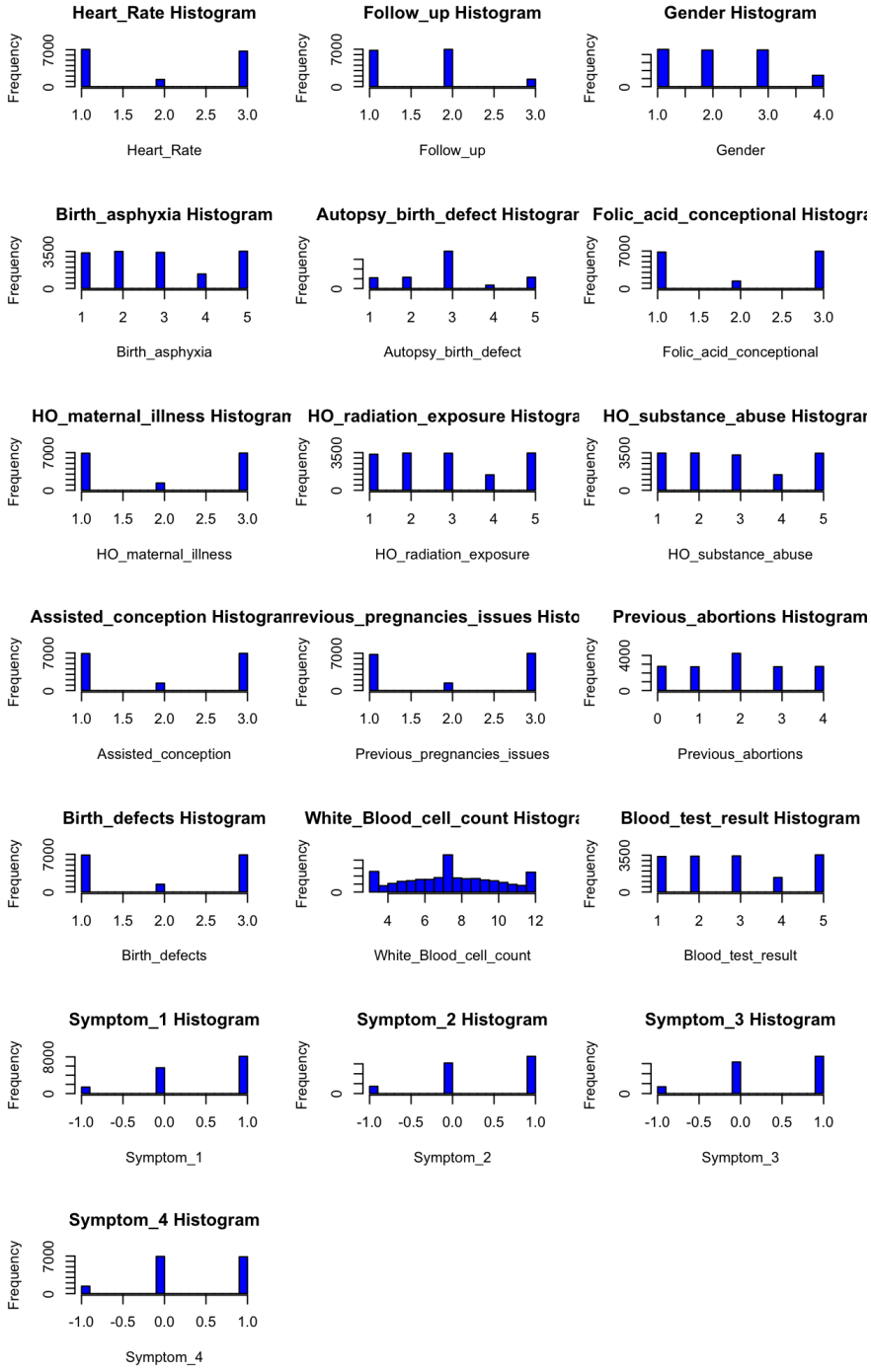
```
# Generate updated summary of base dataset which includes these characteristics
univariate(train_df)
```

```
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
Birth_asphyxia      factor                      5
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
Genetic_Disorder    factor                      4
```

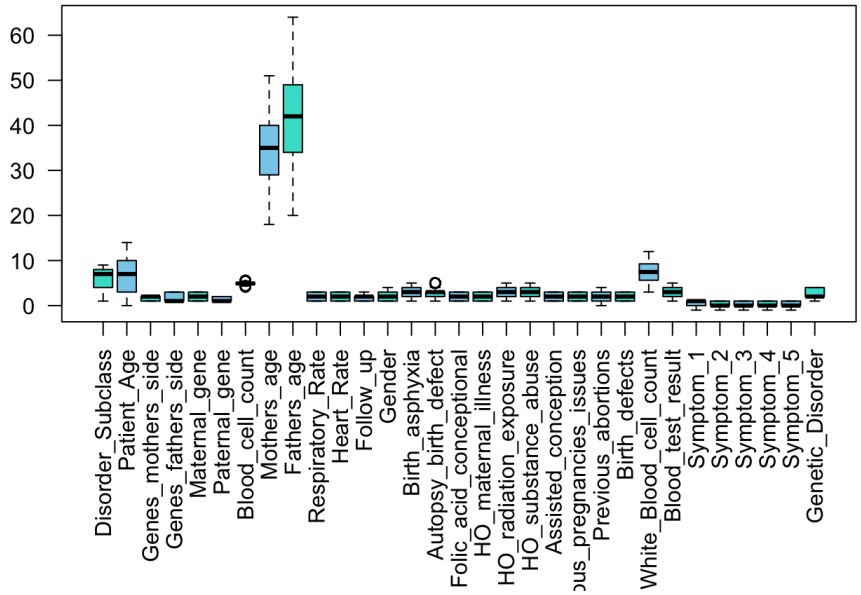
```
#summary(train_df)
```

```
# 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"
)
```






```
# Generate boxplot(s)
par(mar = c(10, 2, 1, 1))
boxplot(train_df, las = 2, col = c("turquoise", "skyblue"))
```



Other Feature Engineering (transformation, aggregation, enrichment)

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

Multivariate Analysis (and reduction)

Collinearity and Dependencies

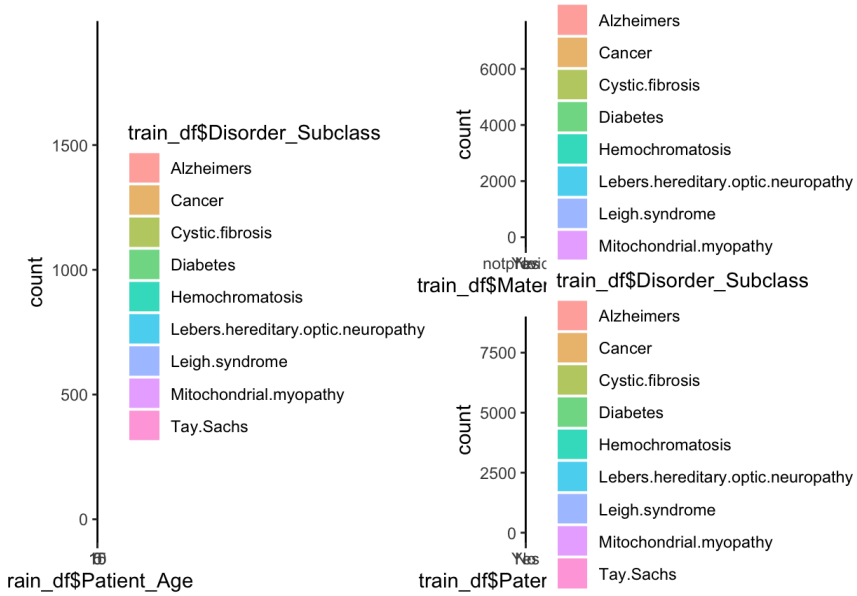
```
# Calculate Cramer's V "measure of association" between nominal factor variables (uses Chi-square statistic)
cscorr <- PairApply(train_df[ , sapply(train_df, is.factor)], CramerV, symmetric = TRUE)

# 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)
Birth_asphyxia (J)
Autopsy_birth_defect (K)
Folic_acid_conceptional (L)
HO_maternal_illness (M)
HO_radiation_exposure (N)
HO_substance_abuse (O)
Assisted_conception (P)
Previous_pregnancies_issues (Q)
Birth_defects (R)
Blood_test_result (S)
Genetic_Disorder (T)
```

```
cscorr
```

```
     A     B     C     D     E     F     G    H    I     J    K     L     M     N     O     P     Q     R     S    T
A 1.00 0.198 0.131 0.123 0.168 0.019 0.026 0.02 0.02 0.022 0.02 0.020 0.019 0.024 0.02 0.019 0.026 0.025 0.03 0.78
B 0.20 1.000 0.005 0.097 0.012 0.005 0.005 0.01 0.01 0.008 0.01 0.013 0.009 0.016 0.01 0.003 0.017 0.008 0.01 0.08
C 0.13 0.005 1.000 0.013 0.093 0.018 0.020 0.01 0.02 0.022 0.02 0.021 0.013 0.030 0.02 0.013 0.018 0.016 0.02 0.07
D 0.12 0.097 0.013 1.000 0.008 0.048 0.040 0.05 0.05 0.054 0.04 0.053 0.048 0.052 0.04 0.055 0.047 0.044 0.05 0.06
E 0.17 0.012 0.093 0.008 1.000 0.003 0.009 0.01 0.01 0.023 0.02 0.003 0.001 0.008 0.02 0.003 0.008 0.006 0.02 0.06
F 0.02 0.005 0.018 0.048 0.003 1.000 0.045 0.03 0.05 0.036 0.02 0.043 0.028 0.030 0.04 0.035 0.036 0.042 0.04 0.05
G 0.03 0.005 0.020 0.040 0.009 0.045 1.000 0.04 0.05 0.034 0.02 0.035 0.029 0.047 0.03 0.055 0.042 0.041 0.05 0.05
H 0.02 0.015 0.012 0.046 0.011 0.029 0.040 1.00 0.04 0.033 0.04 0.041 0.043 0.032 0.04 0.043 0.051 0.038 0.05 0.04
I 0.02 0.010 0.023 0.047 0.010 0.054 0.045 0.04 1.00 0.042 0.02 0.032 0.051 0.045 0.04 0.035 0.028 0.036 0.04 0.04
J 0.02 0.008 0.022 0.054 0.023 0.036 0.034 0.03 0.04 1.000 0.03 0.035 0.026 0.020 0.03 0.047 0.048 0.036 0.03 0.03
K 0.02 0.010 0.016 0.035 0.025 0.019 0.023 0.04 0.02 0.025 1.00 0.030 0.022 0.028 0.03 0.021 0.024 0.029 0.03 0.03
L 0.02 0.013 0.021 0.053 0.003 0.043 0.035 0.04 0.03 0.035 0.03 1.000 0.020 0.049 0.04 0.028 0.032 0.030 0.04 0.04
M 0.02 0.009 0.013 0.048 0.001 0.028 0.029 0.04 0.05 0.026 0.02 0.020 1.000 0.048 0.04 0.043 0.042 0.032 0.04 0.05
N 0.02 0.016 0.030 0.052 0.008 0.030 0.047 0.03 0.04 0.020 0.03 0.049 0.048 1.000 0.03 0.046 0.047 0.052 0.04 0.04
O 0.02 0.011 0.015 0.043 0.016 0.035 0.032 0.04 0.04 0.032 0.03 0.042 0.037 0.026 1.00 0.033 0.050 0.049 0.03 0.03
P 0.02 0.003 0.013 0.055 0.003 0.035 0.055 0.04 0.03 0.047 0.02 0.028 0.043 0.046 0.03 1.000 0.035 0.032 0.03 0.03
Q 0.03 0.017 0.018 0.047 0.008 0.036 0.042 0.05 0.03 0.048 0.02 0.032 0.042 0.047 0.05 0.035 1.000 0.032 0.04 0.03
R 0.02 0.008 0.016 0.044 0.006 0.042 0.041 0.04 0.04 0.036 0.03 0.030 0.032 0.052 0.05 0.032 0.032 1.000 0.04 0.05
S 0.03 0.013 0.018 0.052 0.016 0.036 0.046 0.05 0.04 0.029 0.03 0.042 0.041 0.037 0.03 0.031 0.041 0.044 1.00 0.04
T 0.78 0.082 0.065 0.063 0.064 0.054 0.045 0.04 0.04 0.028 0.03 0.035 0.046 0.042 0.03 0.034 0.030 0.053 0.04 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

```
# 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               0        0               0
0              0
  Cancer                                       0     61               0       31               0
0              0
  Cystic.fibrosis                              0      0            2032        0              50
0              0
  Diabetes                                   103      8               0     1239               0
0              0
  Hemochromatosis                              0      0              30        0              94
0              0
  Lebers.hereditary.optic.neuropathy           8      0              34       29               0
234             27
  Leigh.syndrome                               5      0             139       77              13
226           2868
  Mitochondrial.myopathy                       3      5              57       19              89
26           1020
  Tay.Sachs                                    0      0             330        0             787
0              0
                                     Reference
Prediction                            Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                               0         0
  Cancer                                                   0         0
  Cystic.fibrosis                                          0       347
  Diabetes                                                 0         0
  Hemochromatosis                                          0        99
  Lebers.hereditary.optic.neuropathy                       2         0
  Leigh.syndrome                                        1892        62
  Mitochondrial.myopathy                                1468       141
  Tay.Sachs                                                0      1503

 Accuracy (average) : 0.6267
```

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

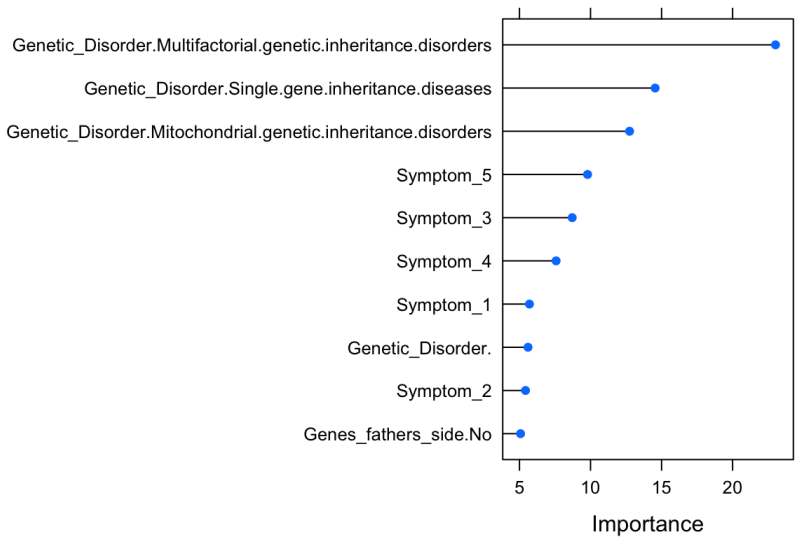
```
ROC curve variable importance

  variables are sorted by maximum importance across the classes
  only 20 most important variables shown (out of 78)

                                                      Alzheimers Cancer Cystic.fibrosis Diabetes Hemochromatosis
Genetic_Disorder.Multifactorial.genetic.inheritance.disorders      0.933  0.533           0.933    0.933           0.933
Genetic_Disorder.Single.gene.inheritance.diseases                  0.956  0.500           0.951    0.500           0.500
Genetic_Disorder.Mitochondrial.genetic.inheritance.disorders       0.500  0.500           0.500    0.952           0.953
Symptom_5                                                          0.931  0.931           0.931    0.931           0.931
Symptom_3                                                          0.914  0.914           0.914    0.914           0.914
Symptom_4                                                          0.884  0.884           0.884    0.884           0.884
Symptom_2                                                          0.844  0.844           0.844    0.844           0.844
Symptom_1                                                          0.808  0.808           0.808    0.808           0.808
Genes_mothers_side.Yes                                             0.787  0.787           0.787    0.787           0.787
Genes_mothers_side.No                                              0.787  0.787           0.787    0.787           0.787
Paternal_gene.Yes                                                  0.774  0.774           0.774    0.774           0.774
Paternal_gene.No                                                   0.774  0.774           0.774    0.774           0.774
Genes_fathers_side.Yes                                             0.764  0.764           0.764    0.764           0.764
Genes_fathers_side.No                                              0.757  0.757           0.757    0.757           0.757
Maternal_gene.Yes                                                  0.737  0.737           0.737    0.737           0.737
Maternal_gene.No                                                   0.711  0.711           0.711    0.711           0.711
Fathers_age                                                        0.570  0.570           0.570    0.570           0.570
Patient_Age                                                        0.562  0.562           0.562    0.562           0.562
Follow_up.High                                                     0.556  0.556           0.556    0.556           0.556
Blood_test_result.inconclusive                                     0.558  0.558           0.558    0.558           0.558
                                                      Lebers.hereditary.optic.neuropathy Leigh.syndrome Mitochondria
l.myopathy Tay.Sachs
Genetic_Disorder.Multifactorial.genetic.inheritance.disorders                      0.933          0.933
0.966     0.933
Genetic_Disorder.Single.gene.inheritance.diseases                                  0.500          0.953
0.956     0.956
Genetic_Disorder.Mitochondrial.genetic.inheritance.disorders                       0.950          0.500
0.500     0.500
Symptom_5                                                                          0.931          0.931
0.931     0.632
Symptom_3                                                                          0.914          0.914
0.914     0.629
Symptom_4                                                                          0.884          0.884
0.884     0.606
Symptom_2                                                                          0.844          0.844
0.844     0.593
Symptom_1                                                                          0.808          0.808
0.808     0.586
Genes_mothers_side.Yes                                                             0.787          0.787
0.787     0.578
Genes_mothers_side.No                                                              0.787          0.787
0.787     0.578
Paternal_gene.Yes                                                                  0.774          0.774
0.774     0.587
Paternal_gene.No                                                                   0.774          0.774
0.774     0.587
Genes_fathers_side.Yes                                                             0.764          0.764
0.764     0.595
Genes_fathers_side.No                                                              0.757          0.757
0.757     0.600
Maternal_gene.Yes                                                                  0.737          0.737
0.737     0.550
Maternal_gene.No                                                                   0.711          0.711
0.711     0.544
Fathers_age                                                                        0.570          0.570
0.570     0.530
Patient_Age                                                                        0.562          0.562
0.562     0.524
Follow_up.High                                                                     0.556          0.556
0.561     0.513
Blood_test_result.inconclusive                                                     0.558          0.558
0.558     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        6               0
1              1
  Cancer                                      0     31               0       11               0
0              0
  Cystic.fibrosis                            10      0            2126       63              56
26             62
  Diabetes                                  103     37               5     1258               0
1              5
  Hemochromatosis                             0      2              33        1             111
0              2
  Lebers.hereditary.optic.neuropathy          2      0               3        2               0
250             32
  Leigh.syndrome                              3      0              74       44              23
172          2505
  Mitochondrial.myopathy                      1      4              44        6              65
33           1288
  Tay.Sachs                                   0      0             337        4             778
3              20
                                    Reference
Prediction                           Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                              0         0
  Cancer                                                  0         0
  Cystic.fibrosis                                        21       382
  Diabetes                                                7         1
  Hemochromatosis                                        10       153
  Lebers.hereditary.optic.neuropathy                      1         0
  Leigh.syndrome                                       1703        67
  Mitochondrial.myopathy                               1597        92
  Tay.Sachs                                              23      1457

 Accuracy (average) : 0.6158
```

```
# Check variable importance
lr_imp <- varImp(lr_fit, scale = FALSE)
plot(lr_imp, top = 10)
```

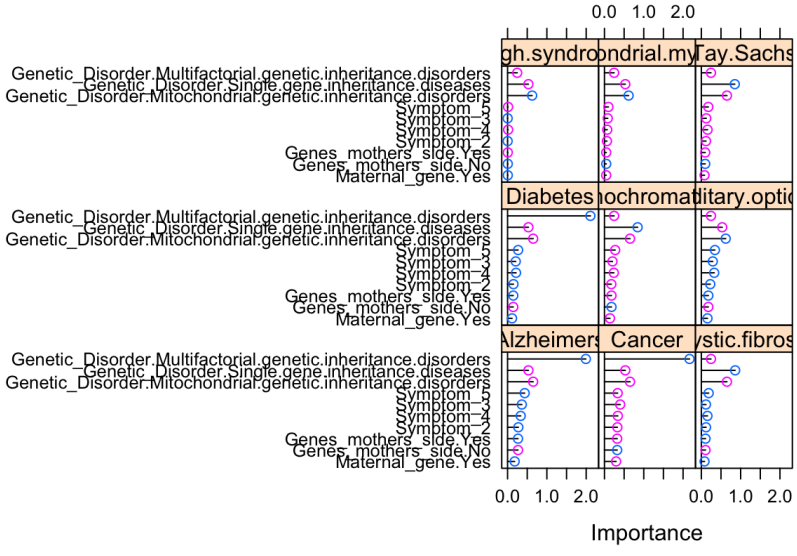


Nearest shrunken Centroids Model

```
# 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
```

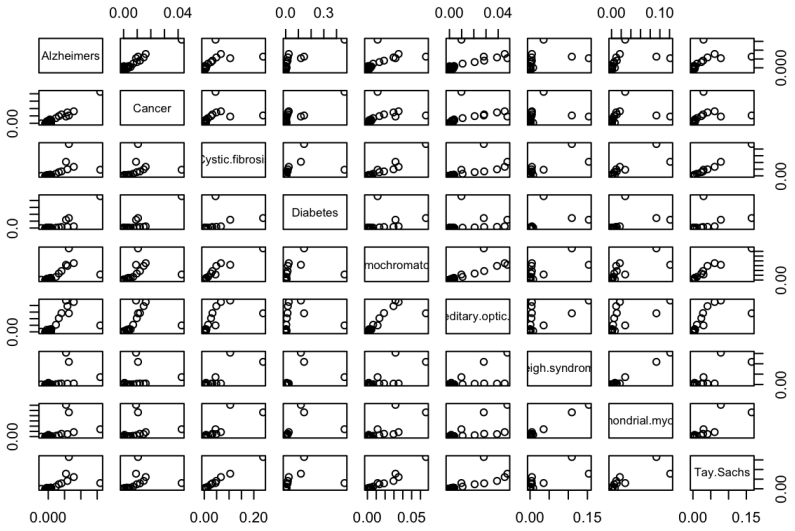
```
Cross-Validated (10 fold) Confusion Matrix

(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                             1      0            2123        0             203
3              0
  Diabetes                                  103     69               0     1270               0
0              0
  Hemochromatosis                             0      0               0        0               0
0              0
  Lebers.hereditary.optic.neuropathy          0      0               0        0               0
0              0
  Leigh.syndrome                             14      0             207      119              54
478           3580
  Mitochondrial.myopathy                      1      5              22        6              48
5            335
  Tay.Sachs                                   0      0             270        0             728
0              0
                                    Reference
Prediction                           Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                              0         0
  Cancer                                                  0         0
  Cystic.fibrosis                                         0       749
  Diabetes                                                0         0
  Hemochromatosis                                         0         0
  Lebers.hereditary.optic.neuropathy                      0         0
  Leigh.syndrome                                       2839       128
  Mitochondrial.myopathy                                523        75
  Tay.Sachs                                               0      1200

 Accuracy (average) : 0.5737
```

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



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)
```

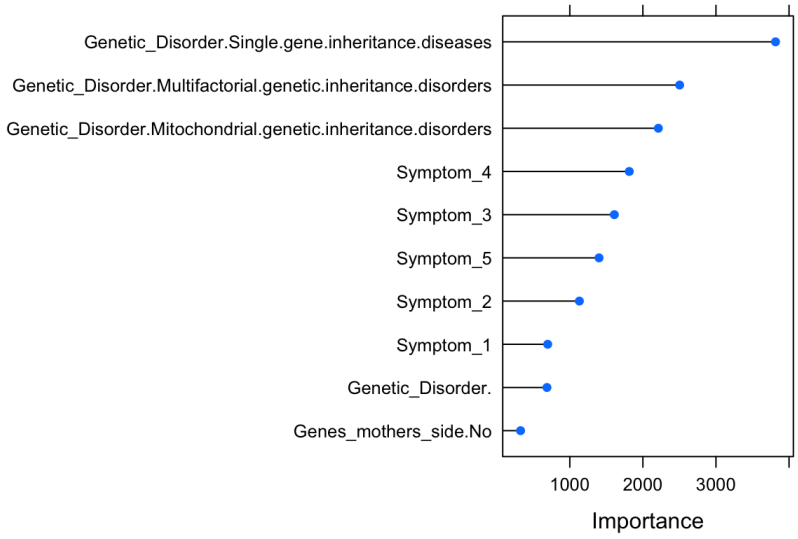


CART Model

```
# 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     62               0       13               0
0              0
  Cystic.fibrosis                             6      0            2208       56              35
24            59
  Diabetes                                  108      7              13     1267               0
8              6
  Hemochromatosis                             0      0               4        0             168
0              0
  Lebers.hereditary.optic.neuropathy          0      0               0        0               0
233             37
  Leigh.syndrome                              2      0              91       35              15
214          2529
  Mitochondrial.myopathy                      3      4              50       24              60
7           1272
  Tay.Sachs                                   0      1             256        0             755
0             12
                                    Reference
Prediction                           Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                              0         0
  Cancer                                                  0         0
  Cystic.fibrosis                                        25       249
  Diabetes                                                1         0
  Hemochromatosis                                         0       124
  Lebers.hereditary.optic.neuropathy                      1         0
  Leigh.syndrome                                       1647        60
  Mitochondrial.myopathy                               1657       123
  Tay.Sachs                                              31      1596

 Accuracy (average) : 0.6412
```

```
# Check variable importance
cart_imp <- varImp(cart_fit, scale = FALSE)
plot(cart_imp, top = 10)
```

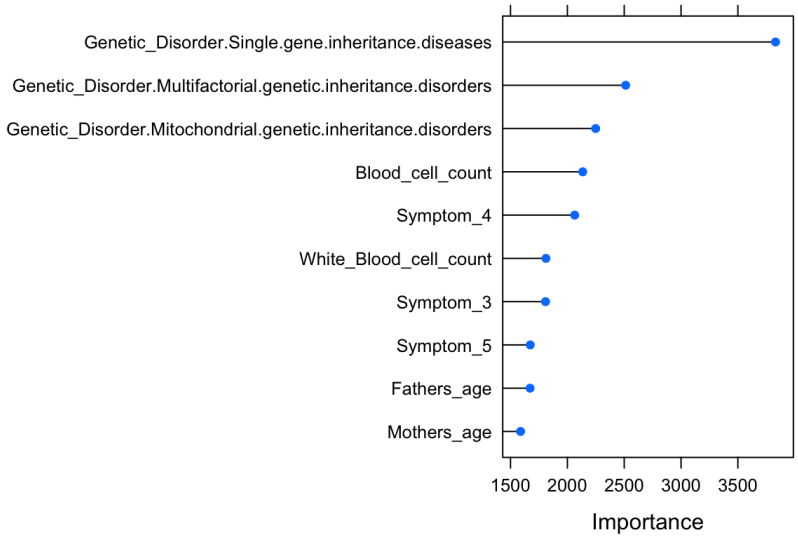


Bagged Trees Model

```
# 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                                   7      0               0       10               0
0              0
  Cancer                                       0     57               0        5               0
0              0
  Cystic.fibrosis                              7      0            2240       48              18
17             46
  Diabetes                                   102     12              24     1278               0
16             14
  Hemochromatosis                              0      1              17        0             276
0              8
  Lebers.hereditary.optic.neuropathy           2      0               3        2               0
286             38
  Leigh.syndrome                               1      0              80       41              21
150          2432
  Mitochondrial.myopathy                       0      1              39        9              40
16           1351
  Tay.Sachs                                    0      3             219        2             678
1              26
                                     Reference
Prediction                            Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                               0         0
  Cancer                                                   0         0
  Cystic.fibrosis                                         26       150
  Diabetes                                                 2         0
  Hemochromatosis                                         15       312
  Lebers.hereditary.optic.neuropathy                       2         0
  Leigh.syndrome                                        1546        52
  Mitochondrial.myopathy                                1726        83
  Tay.Sachs                                               45      1555

 Accuracy (average) : 0.6503
```

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

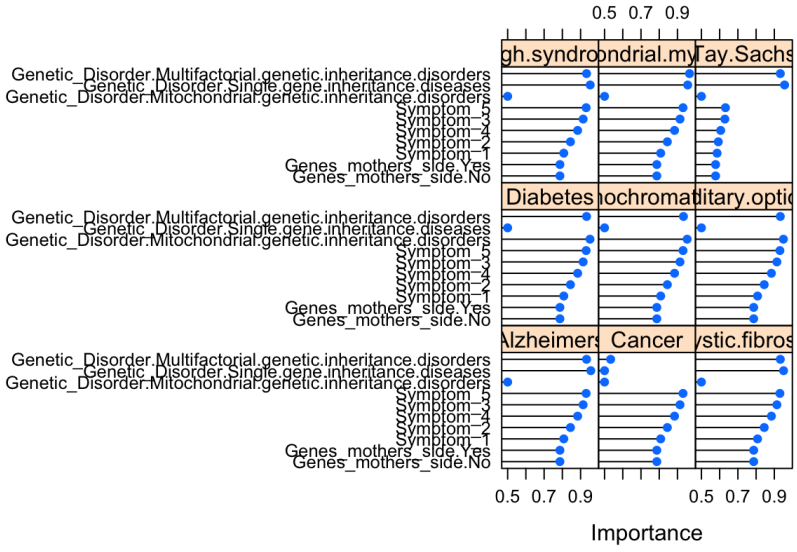


KNN Model

```
# 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               2        2               0
0              1
  Cancer                                       0      0               0        0               0
0              0
  Cystic.fibrosis                             27      7             822      260             180
75           345
  Diabetes                                    22      2             106      174              15
22            82
  Hemochromatosis                              2      5              95       18              75
3            57
  Lebers.hereditary.optic.neuropathy           0      0              17       18               5
7            20
  Leigh.syndrome                              35     20             660      483             250
218         1764
  Mitochondrial.myopathy                      24     26             497      328             239
142         1421
  Tay.Sachs                                    9     14             423      112             269
19           225
                                     Reference
Prediction                            Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                               0         0
  Cancer                                                   1         0
  Cystic.fibrosis                                        255       413
  Diabetes                                                53        24
  Hemochromatosis                                         50       144
  Lebers.hereditary.optic.neuropathy                      18         5
  Leigh.syndrome                                        1455       534
  Mitochondrial.myopathy                                1281       497
  Tay.Sachs                                              249       535

 Accuracy (average) : 0.3073
```

```
# Check variable importance
knn_imp <- varImp(knn_fit, scale = FALSE)
plot(knn_imp, top = 10)
```



Model Validation / Evaluation

```
# 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        0               0
0              0
  Cancer                                    0     16               0        6               0
0              0
  Cystic.fibrosis                           0      0             508        0              15
0              0
  Diabetes                                 26      2               0      310               0
0              0
  Hemochromatosis                           0      0               7        0              26
0              0
  Lebers.hereditary.optic.neuropathy        1      0               6        7               0
49              7
  Leigh.syndrome                            1      0              43       19               1
65            718
  Mitochondrial.myopathy                    0      0              13        6              16
7            253
  Tay.Sachs                                 0      0              78        0             200
0              0
                                  Reference
Prediction                         Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                            0         0
  Cancer                                                0         0
  Cystic.fibrosis                                       0        92
  Diabetes                                              0         0
  Hemochromatosis                                       0        31
  Lebers.hereditary.optic.neuropathy                    2         0
  Leigh.syndrome                                      467        19
  Mitochondrial.myopathy                              371        34
  Tay.Sachs                                             0       362

Overall Statistics

               Accuracy : 0.624
                 95% CI : (0.608, 0.639)
    No Information Rate : 0.258
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.533

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Alzheimers Class: Cancer Class: Cystic.fibrosis Class: Diabetes Class: Hemochromatosis
Sensitivity                   0.034483       0.88889                 0.776          0.8908                0.10078
Specificity                   1.000000       0.99841                 0.966          0.9919                0.98923
Pos Pred Value                1.000000       0.72727                 0.826          0.9172                0.40625
Neg Pred Value                0.992600       0.99947                 0.954          0.9890                0.93765
Prevalence                    0.007662       0.00476                 0.173          0.0919                0.06816
Detection Rate                0.000264       0.00423                 0.134          0.0819                0.00687
Detection Prevalence          0.000264       0.00581                 0.162          0.0893                0.01691
Balanced Accuracy             0.517241       0.94365                 0.871          0.9413                0.54500
                     Class: Lebers.hereditary.optic.neuropathy Class: Leigh.syndrome Class: Mitochondrial.myopathy Class: Ta
y.Sachs
Sensitivity                                         0.4050               0.734                       0.442
0.6729
Specificity                                         0.9937               0.781                       0.888
0.9144
Pos Pred Value                                      0.6806               0.539                       0.530
0.5656
Neg Pred Value                                      0.9806               0.894                       0.848
0.9440
Prevalence                                          0.0320               0.258                       0.222
0.1421
Detection Rate                                      0.0129               0.190                       0.098
0.0956
Detection Prevalence                                0.0190               0.352                       0.185
0.1691
Balanced Accuracy                                   0.6993               0.758                       0.665
0.7936
```

```
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     10               0        3               0
0              0
  Cystic.fibrosis                           1      0             537       21              16
7             22
  Diabetes                                 27      8               0      314               0
0              0
  Hemochromatosis                           0      0              11        0              38
0              1
  Lebers.hereditary.optic.neuropathy        0      0               3        3               0
52             7
  Leigh.syndrome                            1      0              17        3               3
55           595
  Mitochondrial.myopathy                    0      0               9        4               8
7            348
  Tay.Sachs                                 0      0              78        0             193
0              5
                                  Reference
Prediction                         Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                            0         0
  Cancer                                                0         0
  Cystic.fibrosis                                      12       103
  Diabetes                                              0         0
  Hemochromatosis                                       4        49
  Lebers.hereditary.optic.neuropathy                    0         0
  Leigh.syndrome                                      380        21
  Mitochondrial.myopathy                              440        21
  Tay.Sachs                                             4       344

Overall Statistics

               Accuracy : 0.616
                 95% CI : (0.6, 0.631)
    No Information Rate : 0.258
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.526

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Alzheimers Class: Cancer Class: Cystic.fibrosis Class: Diabetes Class: Hemochromatosis
Sensitivity                    0.00000       0.55556                 0.820          0.9023                 0.1473
Specificity                    1.00000       0.99920                 0.942          0.9898                 0.9816
Pos Pred Value                     NaN       0.76923                 0.747          0.8997                 0.3689
Neg Pred Value                 0.99234       0.99788                 0.962          0.9901                 0.9402
Prevalence                     0.00766       0.00476                 0.173          0.0919                 0.0682
Detection Rate                 0.00000       0.00264                 0.142          0.0830                 0.0100
Detection Prevalence           0.00000       0.00343                 0.190          0.0922                 0.0272
Balanced Accuracy              0.50000       0.77738                 0.881          0.9461                 0.5644
                     Class: Lebers.hereditary.optic.neuropathy Class: Leigh.syndrome Class: Mitochondrial.myopathy Class: Ta
y.Sachs
Sensitivity                                             0.4298                0.608                         0.524
0.6394
Specificity                                             0.9965                0.829                         0.865
0.9138
Pos Pred Value                                          0.8000                0.553                         0.526
0.5513
Neg Pred Value                                          0.9815                0.859                         0.864
0.9386
Prevalence                                              0.0320                0.258                         0.222
0.1421
Detection Rate                                          0.0137                0.157                         0.116
0.0909
Detection Prevalence                                    0.0172                0.284                         0.221
0.1649
Balanced Accuracy                                       0.7131                0.719                         0.695
0.7766
```

```
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                           0      0             528        0              45
0              0
  Diabetes                                 27     18               0      316               0
0              0
  Hemochromatosis                           0      0               0        0               0
0              0
  Lebers.hereditary.optic.neuropathy        0      0               0        0               0
0              0
  Leigh.syndrome                            2      0              57       32               7
120            907
  Mitochondrial.myopathy                    0      0               5        0              10
1             71
  Tay.Sachs                                 0      0              65        0             196
0              0
                                  Reference
Prediction                         Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                            0         0
  Cancer                                                0         0
  Cystic.fibrosis                                       0       176
  Diabetes                                              0         0
  Hemochromatosis                                       0         0
  Lebers.hereditary.optic.neuropathy                    0         0
  Leigh.syndrome                                      731        36
  Mitochondrial.myopathy                              109        17
  Tay.Sachs                                             0       309

Overall Statistics

               Accuracy : 0.573
                 95% CI : (0.557, 0.589)
    No Information Rate : 0.258
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.462

 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                 0.806          0.9080                 0.0000
Specificity                    1.00000       1.00000                 0.929          0.9869                 1.0000
Pos Pred Value                     NaN           NaN                 0.705          0.8753                    NaN
Neg Pred Value                 0.99234       0.99524                 0.958          0.9907                 0.9318
Prevalence                     0.00766       0.00476                 0.173          0.0919                 0.0682
Detection Rate                 0.00000       0.00000                 0.139          0.0835                 0.0000
Detection Prevalence           0.00000       0.00000                 0.198          0.0954                 0.0000
Balanced Accuracy              0.50000       0.50000                 0.868          0.9475                 0.5000
                     Class: Lebers.hereditary.optic.neuropathy Class: Leigh.syndrome Class: Mitochondrial.myopathy Class: Ta
y.Sachs
Sensitivity                                              0.000                0.927                        0.1298
0.5743
Specificity                                              1.000                0.649                        0.9647
0.9196
Pos Pred Value                                             NaN                0.479                        0.5117
0.5421
Neg Pred Value                                           0.968                0.962                        0.7954
0.9288
Prevalence                                               0.032                0.258                        0.2219
0.1421
Detection Rate                                           0.000                0.240                        0.0288
0.0816
Detection Prevalence                                     0.000                0.500                        0.0563
0.1506
Balanced Accuracy                                        0.500                0.788                        0.5472
0.7470
```

```
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: 37%
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      11     103       0       0       5       0       0        1.00
Cancer (B)                                0       0       0      69       0       0       0       5       0        1.00
Cystic.fibrosis (C)                       0       0    2302       1       0       0     158      33     128        0.12
Diabetes (D)                              0       0      48    1270       0       0      68       9       0        0.09
Hemochromatosis (E)                       0       0      20       0       0       0      34      66     913        1.00
Lebers.hereditary.optic.neuropathy (F)    0       0      24       1       0       0     460       1       0        1.00
Leigh.syndrome (G)                        0       0      18       0       0       0    3088     809       0        0.21
Mitochondrial.myopathy (H)                0       0       2       0       0       0    2125    1229       6        0.63
Tay.Sachs (I)                             0       0     222       0       0       0      98     104    1728        0.20
               Test set error rate: 36%
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       1      27       0       0       1       0       0        1.00
Cancer (B)                                0       0       0      18       0       0       0       0       0        1.00
Cystic.fibrosis (C)                       0       0     561       0       0       0      52       4      38        0.14
Diabetes (D)                              0       0      14     316       0       0      18       0       0        0.09
Hemochromatosis (E)                       0       0       4       0       0       0       5      12     237        1.00
Lebers.hereditary.optic.neuropathy (F)    0       0       4       0       0       0     116       1       0        1.00
Leigh.syndrome (G)                        0       0       4       0       0       0     803     171       0        0.18
Mitochondrial.myopathy (H)                0       0       0       0       0       0     541     299       0        0.64
Tay.Sachs (I)                             0       0      52       0       0       0      25      28     433        0.20
```

```
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     15               0        4               0
0              0
  Cystic.fibrosis                           0      0             549       18               5
2             21
  Diabetes                                 28      3              13      316               0
4              5
  Hemochromatosis                           0      0               0        0              42
0              0
  Lebers.hereditary.optic.neuropathy        0      0               0        0               0
49             8
  Leigh.syndrome                            1      0              21        9               2
64           689
  Mitochondrial.myopathy                    0      0              12        1              15
2            255
  Tay.Sachs                                 0      0              60        0             194
0              0
                                  Reference
Prediction                         Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                            0         0
  Cancer                                                0         0
  Cystic.fibrosis                                       8        62
  Diabetes                                              1         0
  Hemochromatosis                                       0        36
  Lebers.hereditary.optic.neuropathy                    4         0
  Leigh.syndrome                                      410        13
  Mitochondrial.myopathy                              417        40
  Tay.Sachs                                             0       387

Overall Statistics

               Accuracy : 0.651
                 95% CI : (0.636, 0.666)
    No Information Rate : 0.258
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.568

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Alzheimers Class: Cancer Class: Cystic.fibrosis Class: Diabetes Class: Hemochromatosis
Sensitivity                    0.00000       0.83333                0.838          0.9080                 0.1628
Specificity                    1.00000       0.99894                0.963          0.9843                 0.9898
Pos Pred Value                     NaN       0.78947                0.826          0.8541                 0.5385
Neg Pred Value                 0.99234       0.99920                0.966          0.9906                 0.9417
Prevalence                     0.00766       0.00476                0.173          0.0919                 0.0682
Detection Rate                 0.00000       0.00396                0.145          0.0835                 0.0111
Detection Prevalence           0.00000       0.00502                0.176          0.0978                 0.0206
Balanced Accuracy              0.50000       0.91614                0.901          0.9462                 0.5763
                     Class: Lebers.hereditary.optic.neuropathy Class: Leigh.syndrome Class: Mitochondrial.myopathy Class: Ta
y.Sachs
Sensitivity                                             0.4050                0.704                          0.496
0.719
Specificity                                             0.9967                0.815                          0.890
0.922
Pos Pred Value                                          0.8033                0.570                          0.562
0.604
Neg Pred Value                                          0.9807                0.888                          0.861
0.952
Prevalence                                              0.0320                0.258                          0.222
0.142
Detection Rate                                          0.0129                0.182                          0.110
0.102
Detection Prevalence                                    0.0161                0.319                          0.196
0.169
Balanced Accuracy                                       0.7008                0.760                          0.693
0.821
```

```
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                                3      0               0        1               0
0              0
  Cancer                                    0     15               0        0               0
0              0
  Cystic.fibrosis                           0      0             553       12               4
1             19
  Diabetes                                 25      3              10      323               0
3              4
  Hemochromatosis                           0      0               3        0              78
0              1
  Lebers.hereditary.optic.neuropathy        0      0               1        1               0
69             11
  Leigh.syndrome                            1      0              19       10               1
44            570
  Mitochondrial.myopathy                    0      0              12        0               8
4             368
  Tay.Sachs                                 0      0              57        1             167
0              5
                                  Reference
Prediction                         Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                            0         0
  Cancer                                                0         1
  Cystic.fibrosis                                       6        42
  Diabetes                                              1         0
  Hemochromatosis                                       5        91
  Lebers.hereditary.optic.neuropathy                    2         0
  Leigh.syndrome                                      379        17
  Mitochondrial.myopathy                              430        18
  Tay.Sachs                                            17       369

Overall Statistics

               Accuracy : 0.637
                 95% CI : (0.621, 0.652)
    No Information Rate : 0.258
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.554

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Alzheimers Class: Cancer Class: Cystic.fibrosis Class: Diabetes Class: Hemochromatosis
Sensitivity                   0.103448       0.83333                 0.844          0.9282                 0.3023
Specificity                   0.999734       0.99973                 0.973          0.9866                 0.9716
Pos Pred Value                0.750000       0.93750                 0.868          0.8753                 0.4382
Neg Pred Value                0.993124       0.99920                 0.968          0.9927                 0.9501
Prevalence                    0.007662       0.00476                 0.173          0.0919                 0.0682
Detection Rate                0.000793       0.00396                 0.146          0.0853                 0.0206
Detection Prevalence          0.001057       0.00423                 0.168          0.0975                 0.0470
Balanced Accuracy             0.551591       0.91653                 0.909          0.9574                 0.6370
                     Class: Lebers.hereditary.optic.neuropathy Class: Leigh.syndrome Class: Mitochondrial.myopathy Class: Ta
y.Sachs
Sensitivity                                             0.5702                0.583                         0.512
0.6859
Specificity                                             0.9959                0.832                         0.861
0.9239
Pos Pred Value                                          0.8214                0.548                         0.512
0.5990
Neg Pred Value                                          0.9859                0.851                         0.861
0.9467
Prevalence                                              0.0320                0.258                         0.222
0.1421
Detection Rate                                          0.0182                0.151                         0.114
0.0975
Detection Prevalence                                    0.0222                0.275                         0.222
0.1627
Balanced Accuracy                                       0.7831                0.708                         0.686
0.8049
```

```
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
}
```

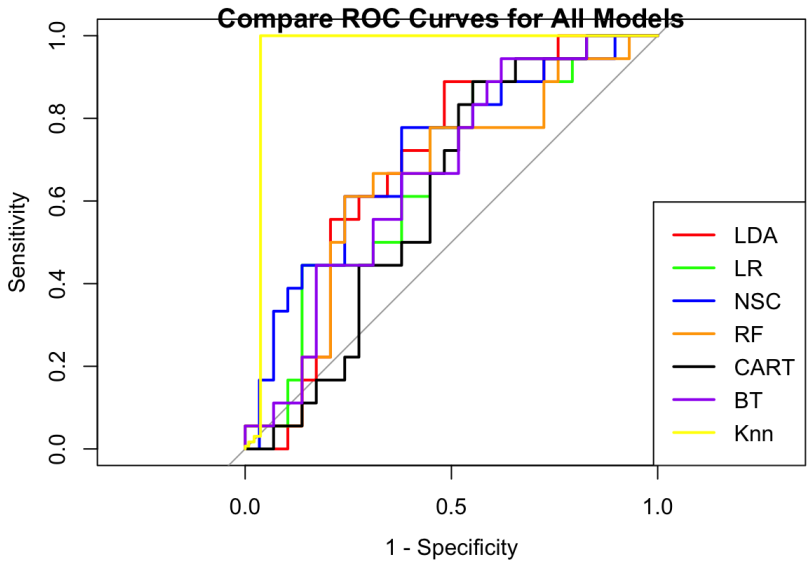
```
[1] "KNN"
```

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                           7      1             194       59              50
9             78
  Diabetes                                  5      0              30       45               2
6             20
  Hemochromatosis                           0      2              21        2              21
0             14
  Lebers.hereditary.optic.neuropathy        1      0               2        8               2
3              3
  Leigh.syndrome                           11      7             169      114              71
69            440
  Mitochondrial.myopathy                    4      8             114       90              53
31            362
  Tay.Sachs                                 1      0             125       30              59
3             61
                                  Reference
Prediction                         Mitochondrial.myopathy Tay.Sachs
  Alzheimers                                            0         0
  Cancer                                                0         0
  Cystic.fibrosis                                      54       106
  Diabetes                                             14        11
  Hemochromatosis                                      16        36
  Lebers.hereditary.optic.neuropathy                    4         2
  Leigh.syndrome                                      384       132
  Mitochondrial.myopathy                              306       107
  Tay.Sachs                                            62       144

Overall Statistics

               Accuracy : 0.305
                 95% CI : (0.29, 0.32)
    No Information Rate : 0.258
    P-Value [Acc > NIR] : 0.0000000000959

                  Kappa : 0.122

 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                0.2962          0.1293                0.08140
Specificity                    1.00000       1.00000                0.8837          0.9744                0.97420
Pos Pred Value                     NaN           NaN                0.3477          0.3383                0.18750
Neg Pred Value                 0.99234       0.99524                0.8571          0.9170                0.93548
Prevalence                     0.00766       0.00476                0.1731          0.0919                0.06816
Detection Rate                 0.00000       0.00000                0.0513          0.0119                0.00555
Detection Prevalence           0.00000       0.00000                0.1474          0.0351                0.02959
Balanced Accuracy              0.50000       0.50000                0.5899          0.5519                0.52780
                     Class: Lebers.hereditary.optic.neuropathy Class: Leigh.syndrome Class: Mitochondrial.myopathy Class: Ta
y.Sachs
Sensitivity                                          0.024793                0.450                        0.3643
0.268
Specificity                                          0.993996                0.659                        0.7389
0.895
Pos Pred Value                                       0.120000                0.315                        0.2847
0.297
Neg Pred Value                                       0.968617                0.775                        0.8030
0.881
Prevalence                                           0.031968                0.258                        0.2219
0.142
Detection Rate                                       0.000793                0.116                        0.0808
0.038
Detection Prevalence                                 0.006605                0.369                        0.2840
0.128
Balanced Accuracy                                    0.509395                0.554                        0.5516
0.581
```

```
## 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))

if (lr_eval)
  lr_roc <- multiclass.roc(response = test_df$Disorder_Subclass,
                           predictor = order(lr_pred))
#  lr_roc <- multiclass.roc(response = lr_fit$pred$obs,
#                predictor = lr_fit$pred$Cancer,
#                levels = rev(levels(lr_fit$pred$obs)))

if (nsc_eval)
  nsc_roc <- multiclass.roc(response = test_df$Disorder_Subclass,
                            predictor = order(nsc_pred))
#  nsc_roc <- multiclass.roc(response = nsc_fit$pred$obs,
#                predictor = nsc_fit$pred$Cancer,
#                levels = rev(levels(nsc_fit$pred$obs)))

if (rf_eval)
  rf_roc <- multiclass.roc(response = test_df$Disorder_Subclass,
                           predictor = order(rf_fit$test$predicted))

if (cart_eval)
  cart_roc <- multiclass.roc(response = test_df$Disorder_Subclass,
                             predictor = order(cart_pred))
#  cart_roc <- multiclass.roc(response = cart_fit$pred$obs,
#                predictor = cart_fit$pred$Cancer,
#                levels = rev(levels(cart_fit$pred$obs)))

if (bt_eval)
  bt_roc <- multiclass.roc(response = test_df$Disorder_Subclass,
                           predictor = order(bt_pred))
#  bt_roc <- multiclass.roc(response = bt_fit$pred$obs,
#                predictor = bt_fit$pred$Cancer,
#                levels = rev(levels(bt_fit$pred$obs)))

if (knn_eval)
  knn_roc <- multiclass.roc(response = knn_fit$pred$obs,
               predictor = knn_fit$pred$Cancer,
            levels = rev(levels(knn_fit$pred$obs)))
```

```
### Compare Models using ROC curve
if (lda_eval)
  plot.roc(lda_roc$rocs[[1]], type = "s", col = 'red', legacy.axes = TRUE)

if (lr_eval)
  plot.roc(lr_roc$rocs[[1]], type = "s", add = TRUE, col = 'green', legacy.axes = TRUE)

if (nsc_eval)
  plot.roc(nsc_roc$rocs[[1]], type = "s", add = TRUE, col = 'blue', legacy.axes = TRUE)

if (rf_eval)
  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)
  plot.roc(cart_roc$rocs[[1]], type = "s",col = 'black', add = TRUE, legacy.axes = TRUE)

if (bt_eval)
  plot.roc(bt_roc$rocs[[1]], type = "s", col = 'purple',add = TRUE, legacy.axes = TRUE)

if (knn_eval)
  plot.roc(knn_roc$rocs[[1]], type = "s", col = 'yellow', add = TRUE, legacy.axes = TRUE)

legend("bottomright", legend = c("LDA", "LR", "NSC", "RF", "CART", "BT", "Knn"),
       col = c("red", "green", "blue", "orange", "black", "purple", "yellow"), lwd = 2)
title(main = "Compare ROC Curves for All Models")
```



Confusion Matrix TABLE (NEEDS TO ADD THE CODES)

```
models <- c("LDA", "Logistic Regression","NSC", "Random Forest", "Bagging", "KNN")

Accuracy_train <- c(0.3562,0.3396, 0.3795, 1, 0.3586, 0.2265)

Accuracy_test <- c(0.35, 0.35,0.27,0.38,0.37,0.226 )

AUC_Score <- c(0.840, 0.810, 0.596, 0.660, 0.561, 0.494)

df_results <- data.frame(models, Accuracy_train,Accuracy_test, AUC_Score )
df_results
```

Top Ten Important Predictors for the optimal model

```
varimp <- varImp(ldaFit)
plot(varimp, top = 10)
```

---

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)↩