STATISTICAL ANALYSIS ON DRINKING DATA

GOKUL VARATHARASAN
ABHIGNA VALAMBATLA
SAI GOUTHAM MANUKONDA

TABLE OF CONTENTS

Introduction

In this project, we worked on building a classification model to predict whether a person drinks alcohol based on different factors. We used the drinking dataset,This dataset is collected from National Health Insurance Service in Korea. , which contains various attributes related to individuals and their drinking habits. The main goal was to apply different machine learning models like LDA (Linear Discriminant Analysis), KNN (K-nearest neighbors), Random Forest, and SVM (Support Vector Machine), and compare their performance using accuracy and confusion matrices. We also created visualizations to explore how different factors, such as serum creatinine levels and smoking status, influence drinking behavior. The findings from this study help in understanding key patterns and risk factors related to alcohol consumption.

Why this Dataset?

We chose this dataset because it provides valuable insights into drinking behavior and its associated factors. Understanding patterns in alcohol consumption can help identify potential health risks and behavioral trends. By applying machine learning models, we aimed to explore how different factors, such as smoking status and serum creatinine levels, contribute to predicting whether an individual drinks alcohol. This dataset allowed us to apply data preprocessing, model training, and evaluation techniques while gaining practical experience in classification problems.

What are we trying to model ?

We are trying to predict if a person drinks alcohol based on their personal and health-related information. The goal is to find the key factors that influence drinking behavior and build a model that can classify individuals accurately. To do this, we train machine learning models to identify patterns in the data. We also analyze important features that may contribute to alcohol consumption, helping us understand the factors linked to drinking habits.

Data

The dataset consists of 1,000 data instances (rows) and 24 features (columns). These features include a mix of demographic, lifestyle, and health-related variables such as age, weight, smoking status, and serum creatinine levels. These attributes help in understanding potential correlations with alcohol consumption.Upon analyzing the dataset, we found that the data is imbalanced, with a higher number of non-drinkers compared to drinkers. This imbalance may impact the model's ability to classify drinking behavior accurately, requiring potential adjustments such as resampling techniques or weighted classification to improve performance.

Data Description

The dataset contains various health-related attributes that help analyze drinking behavior. The key columns and their descriptions are as follows:

Demographic and Physical Attributes:

- **Sex** – Male or female
- **Age** – Rounded up to 5 years
- **Height** – Rounded up to the nearest 5 cm
- **Weight** – Recorded in kilograms

Health and Medical Measurements:

- **Sight Left / Right** – Vision strength in left and right eyes
- **Hearing Left / Right** – Hearing ability (1 = normal, 2 = abnormal)
- **SBP (Systolic Blood Pressure)** – Measured in mmHg
- **DBP (Diastolic Blood Pressure)** – Measured in mmHg
- **BLDS (Fasting Blood Glucose Level)** – Measured in mg/dL

Blood and Biochemical Markers:

- **Total Cholesterol** – Measured in mg/dL
- **HDL / LDL Cholesterol** – Levels of good and bad cholesterol in mg/dL
- **Triglycerides** – Fat levels in the blood (mg/dL)
- **Hemoglobin** – Hemoglobin concentration in g/dL
- **Urine Protein** – Protein levels in urine (graded from -1 to 6)
- **Serum Creatinine** – Kidney function indicator (mg/dL)

Liver Function Indicators:

- **SGOT (AST) & SGOT (ALT)** – Liver enzyme levels (IU/L)
- **Gamma-GTP** – Liver enzyme indicator for alcohol consumption (IU/L)

Lifestyle and Target Variables:

- **SMK_stat_type_cd** – Smoking status (1 = never, 2 = quit, 3 = still smokes)
- **DRK_YN** – Drinking status (1 = drinks, 0 = does not drink)

This dataset provides a comprehensive overview of personal and health-related factors that may influence drinking behavior. These attributes are used to train and evaluate machine learning models for classification.

Exploratory Data Analysis

Summary Statistics

We started by finding some summary statistics for the data in order to determine the type of data and its values. Below is the output summary statistics for the fields we chose for the model.

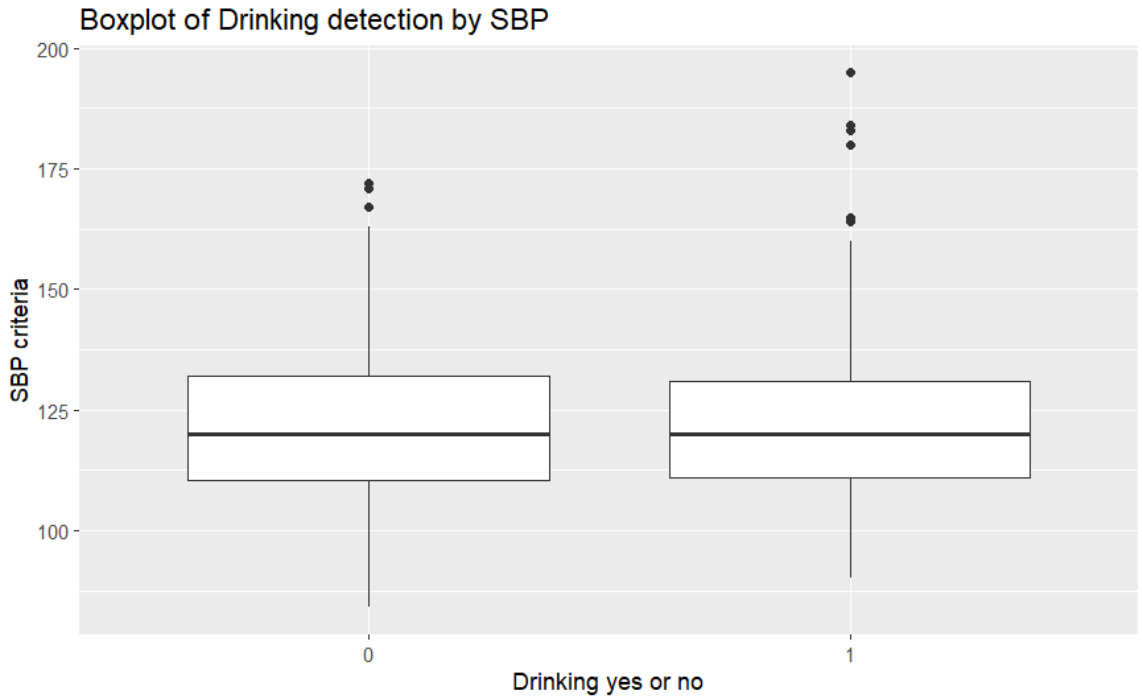
```
 sex          age            height          weight         waistline
Female:454   Min.   :20.00   Min.   :135.0   Min.   : 35.00   Min.   : 54.00
Male  :546   1st Qu.:35.00   1st Qu.:155.0   1st Qu.: 55.00   1st Qu.: 74.88
              Median :45.00   Median :165.0   Median : 65.00   Median : 81.55
              Mean   :46.88   Mean   :162.6   Mean   : 63.72   Mean   : 81.45
              3rd Qu.:55.00   3rd Qu.:170.0   3rd Qu.: 70.00   3rd Qu.: 88.00
              Max.   :85.00   Max.   :185.0   Max.   :120.00   Max.   :116.00
  sight_left      sight_right     hear_left       hear_right         SBP
Min.   :0.1000   Min.   :0.1   Min.   :1.000   Min.   :1.000   Min.   : 84.0
1st Qu.:0.7000   1st Qu.:0.7   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:111.0
Median :1.0000   Median :1.0   Median :1.000   Median :1.000   Median :120.0
Mean   :0.9835   Mean   :1.0   Mean   :1.025   Mean   :1.025   Mean   :122.4
3rd Qu.:1.2000   3rd Qu.:1.2   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:132.0
Max.   :9.9000   Max.   :9.9   Max.   :2.000   Max.   :2.000   Max.   :195.0
     DBP             BLDS          tot_chole       HDL_chole
Min.   : 51.00   Min.   : 62.00   Min.   : 90.0   Min.   : 22.00
1st Qu.: 70.00   1st Qu.: 87.00   1st Qu.:169.0   1st Qu.: 46.00
Median : 76.00   Median : 95.00   Median :195.0   Median : 54.00
Mean   : 76.06   Mean   : 99.92   Mean   :194.5   Mean   : 56.27
3rd Qu.: 82.00   3rd Qu.:105.00   3rd Qu.:218.0   3rd Qu.: 65.00
Max.   :115.00   Max.   :425.00   Max.   :384.0   Max.   :119.00
  LDL_chole       triglyceride      hemoglobin     urine_protein
Min.   : 25.0   Min.   :  16.0   Min.   : 9.30   Min.   :1.000
1st Qu.: 89.0   1st Qu.:  72.0   1st Qu.:13.10   1st Qu.:1.000
Median :112.0   Median : 105.0   Median :14.40   Median :1.000
Mean   :112.6   Mean   : 132.6   Mean   :14.24   Mean   :1.083
3rd Qu.:134.0   3rd Qu.: 167.0   3rd Qu.:15.50   3rd Qu.:1.000
Max.   :325.0   Max.   :1027.0   Max.   :18.10   Max.   :5.000
serum_creatinine    SGOT_AST         SGOT_ALT         gamma_GTP
Min.   :0.2000   Min.   :  4.00   Min.   :  5.00   Min.   :  3.00
1st Qu.:0.7000   1st Qu.: 19.00   1st Qu.: 14.75   1st Qu.: 15.00
Median :0.8000   Median : 22.00   Median : 20.00   Median : 23.00
Mean   :0.8654   Mean   : 25.38   Mean   : 25.39   Mean   : 36.15
```

3rd Qu.:1.0000 3rd Qu.: 28.00 3rd Qu.: 30.00 3rd Qu.: 39.25
Max. :5.0000 Max. :278.00 Max. :311.00 Max. :521.00
SMK_stat_type_cd DRK_YN
Min. :1.000 Min. :0.000
1st Qu.:1.000 1st Qu.:0.000
Median :1.000 Median :1.000
Mean :1.642 Mean :0.514
3rd Qu.:2.000 3rd Qu.:1.000
Max. :3.000 Max. :1.000

Plots

Box plots

Drinking va SBP

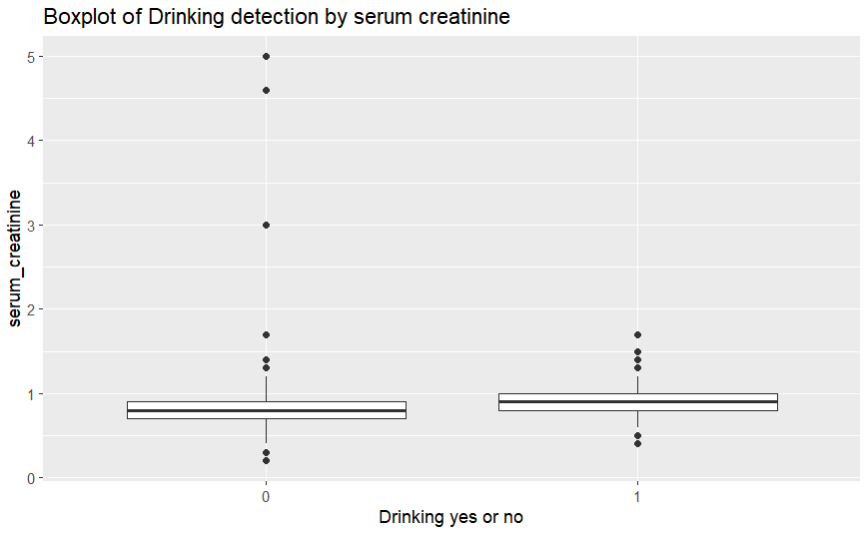


In the box plot, There is a similar distribution of blood pressure values for both drinkers (1) and non-drinkers (0). Outliers are present in both groups. The median values appear close, indicating no strong difference in blood pressure between drinkers and non-drinkers.

Drinking vs Total Cholesterol

Box plot shows that Drinkers and non-drinkers have similar median cholesterol levels. However, drinkers seem to have a slightly higher spread (variance), with more extreme outliers in the higher cholesterol range.

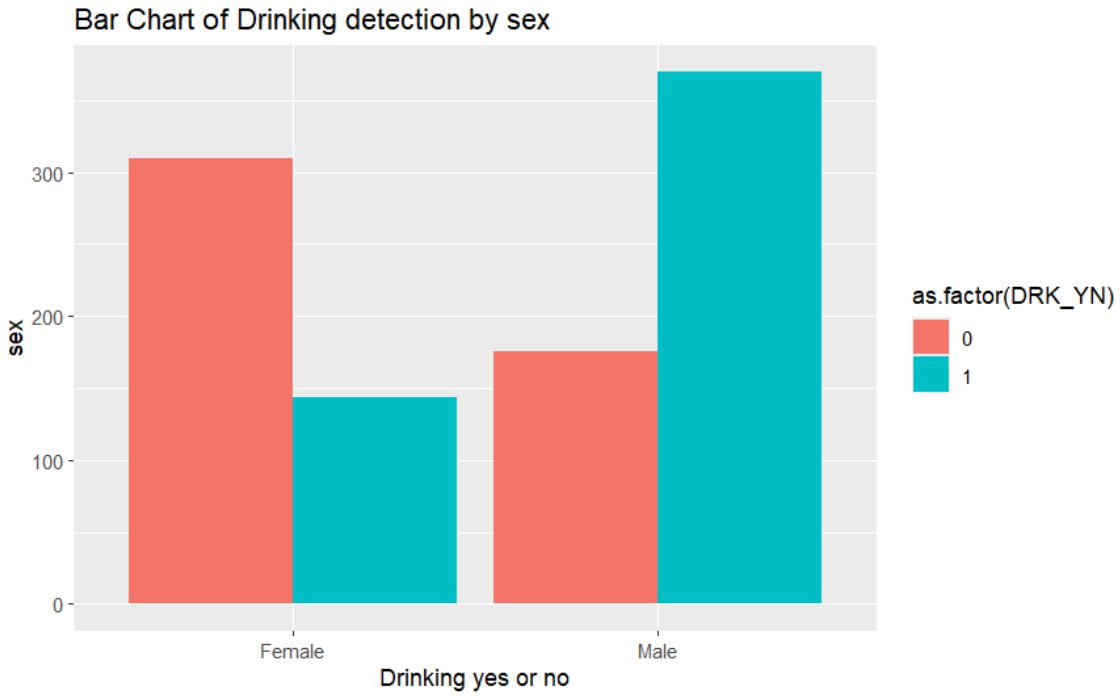
Drinking vs Serum Creatinine



Serum creatinine levels are almost identical between drinkers and non-drinkers. The presence of outliers suggests some individuals have very high levels, but the main distribution remains similar.

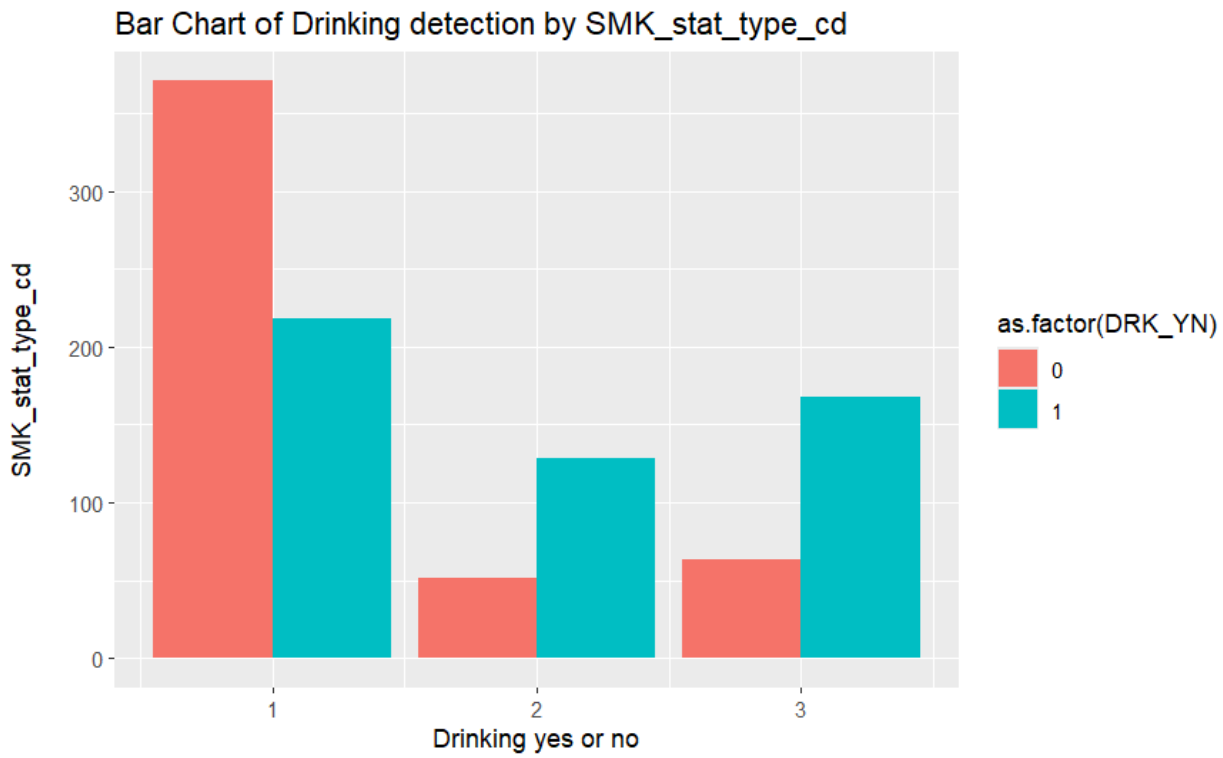
Bar Charts

Drinking habit vs sex



More females are non-drinkers (red bar is significantly taller than blue). Among males, drinkers (blue) are more prevalent or at least equal to non-drinkers. This suggests that men are more likely to drink than women, based on the dataset.

Drinking habit vs smoking status



Most non-smoker are non-drinkers (the red bar is taller). The balance between drinkers (blue) and non-drinkers (red) for former smokers is more, which suggests that quitting smoking doesn't necessarily mean quitting drinking, but there is still a high proportion of non-drinkers. Also, current smokers are more likely to be drinkers.

Modelling the data

Firstly, the data frame is splitted into two parts like 80% into train and 20% into test data. Our dataset has 1000 rows of data which has been split into 800 rows of training data and 200 rows of testing data.

LDA Model

Summary of model:

Call:
lda(DRK_YN ~ ., data = train)

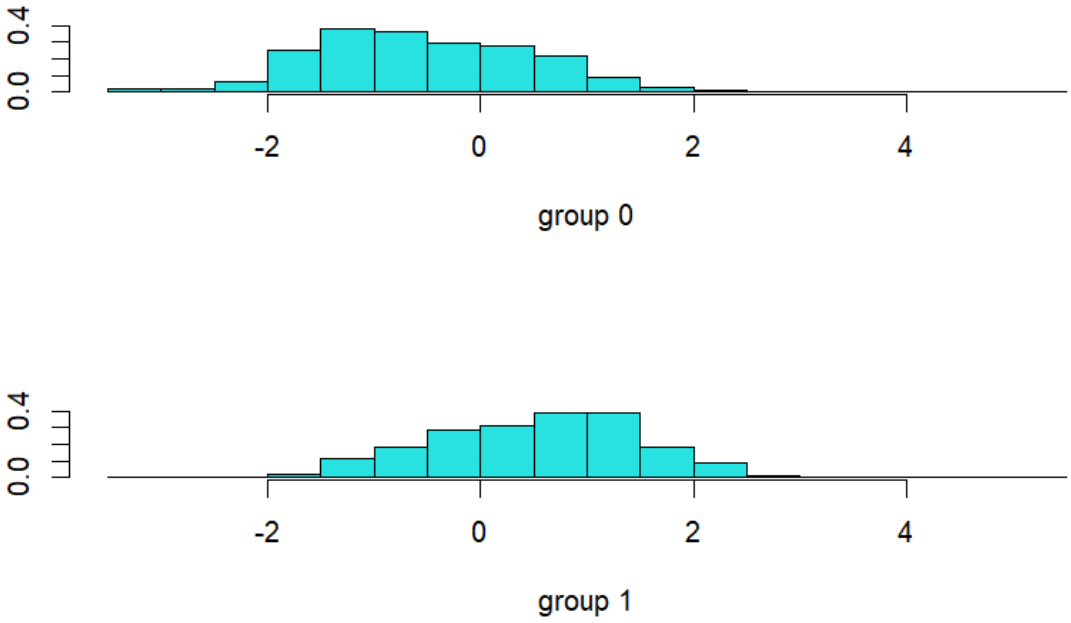
Prior probabilities of groups:
  0 1
0.49 0.51

Group means:
   sexMale   age   height   weight waistline sight_left sight_right
0 0.3852041 50.25510 159.5153 60.76531 80.48597   0.940051   0.9760204
1 0.7230392 43.66422 165.8333 67.08333 82.60319   1.040196   1.0438725
  hear_left hear_right   SBP   DBP   BLDS tot_chole HDL_chole LDL_chole
0 1.030612   1.035714 122.5969 75.20153 100.1607 191.9056 55.29592 112.2117
1 1.019608   1.022059 122.5098 76.68382 101.1127 197.7132 57.04657 113.6961
  triglyceride hemoglobin urine_protein serum_creatinine SGOT_AST SGOT_ALT
0   123.6505   13.88495   1.112245   0.8459184 24.60714 24.17857
1   142.1593   14.65490   1.066176   0.8948529 26.41912 27.25245
  gamma_GTP SMK_stat_type_cd
0 26.63010   1.403061
1 46.23775   1.879902

Coefficients of linear discriminants:
                      LD1
sexMale   1.1528828166
age   -0.0305659250
height   -0.0016269748
weight   0.0206672864
waistline   -0.0019381634
sight_left   0.1065683963
sight_right   0.0267345340
hear_left   -0.0292751927
hear_right   -0.0284299110
SBP   -0.0087401963
DBP   0.0081776974
BLDS   0.0009943276
tot_chole   0.0045897701
HDL_chole   0.0255004821
LDL_chole   -0.0038681687
triglyceride   0.0005248971
hemoglobin   -0.0185931928
urine_protein -0.2428177896
serum_creatinine -0.2788916745
SGOT_AST   0.0227371904
SGOT_ALT   -0.0253077365
gamma_GTP   0.0086785693
SMK_stat_type_cd 0.3323780042

LDA works by finding patterns in the data that best separate the two groups. Looking at the group means, class "1" had higher values for height, weight, cholesterol levels, and smoking status, while class "0" had slightly lower values. The linear discriminant coefficients showed that sex, hearing ability, smoking status, and serum creatinine levels played the biggest role in distinguishing between the two classes.

LDA Plot:



The histogram plot above shows how well LDA separates the groups. The top graph represents group 0, and the bottom one represents group 1. Since both distributions have similar spread along the x-axis and lack a clear gap between them, it means some misclassification is likely.

Confusion matrix

The model achieved 73% accuracy, correctly identifying 73.03% of class "0" cases and 67.57% of class "1" cases. However, 36 false positives and 24 false negatives suggest that some cases were hard to classify. Overall, LDA worked well, offering a structured way to separate the groups, but the overlapping data made it difficult to fully distinguish between them.

```
Confusion Matrix and Statistics


    0  1
  0 72 22
  1 32 74

               Accuracy : 0.73
                 95% CI : (0.6628, 0.7902)
    No Information Rate : 0.52
    P-Value [Acc > NIR] : 1.007e-09

                  Kappa : 0.4613

 Mcnemar's Test P-Value : 0.2207

            Sensitivity : 0.6923
            Specificity : 0.7708
         Pos Pred Value : 0.7660
         Neg Pred Value : 0.6981
             Prevalence : 0.5200
         Detection Rate : 0.3600
   Detection Prevalence : 0.4700
      Balanced Accuracy : 0.7316

       'Positive' Class : 0

[1] 0.27
```

Error rate in this model: 0.27

k-Nearest Neighbors(KNN) Model

Finding k for the model

We applied the k-Nearest Neighbors (KNN) model to our data set, using an 80-20 train-test split to ensure a fair evaluation. Since KNN is a distance-based algorithm, we first scaled the predictor variables to maintain uniform influence across all features. We tested multiple values of k ranging from 1 to 100 and identified k = 16 as the optimal choice, as it minimized classification error.

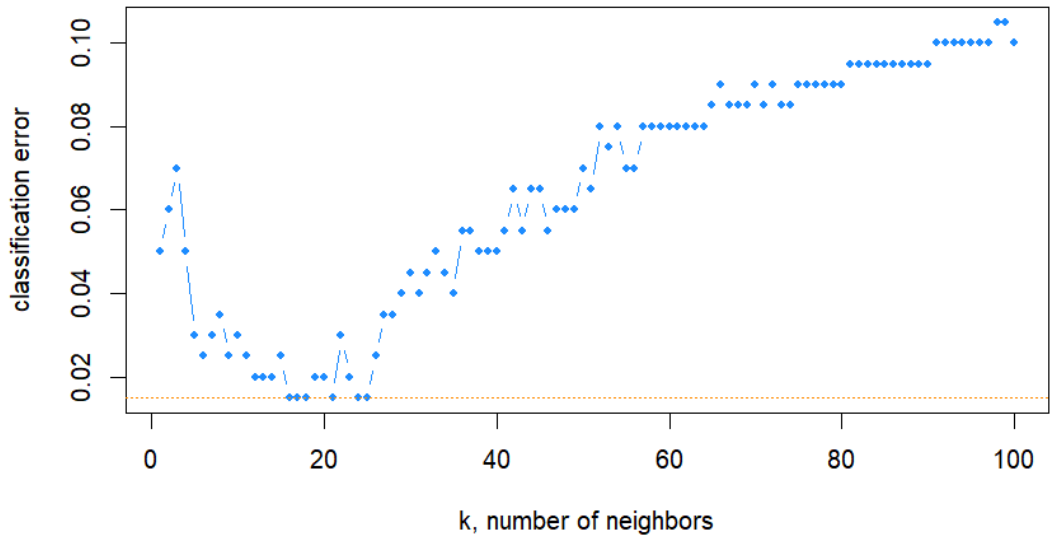
```
# finding k

set.seed(1234)
k.grid=1:100
error=rep(0, length(k.grid))

for (i in seq_along(k.grid)) {
  pred = knn(train = scale(knn.train),
             test  = scale(knn.test),
             cl    = knn.trainLabels,
             k     = k.grid[i])
  error[i] = mean(knn.testLabels !=pred)
}

min(error)
```

```
[1] 0.015
```

```
best_k = k.grid[which.min(error)]
best_k
```

```
[1] 16
```

Plot



The classification error plot showed that error rates fluctuated as K increased, with the lowest error observed at k = 16. However, there were some misclassifications, with 30 false positives and 45 false negatives, suggesting that the model had difficulty distinguishing between classes. This could be due to overlapping feature distributions, making it harder for KNN to form distinct decision boundaries.

Confusion Matrix

We evaluated the model using a confusion matrix, which showed an overall accuracy of 62.5%. The model performed better at detecting class “1” (specificity: 68.75%) than class “0” (sensitivity: 56.73%), meaning it struggled more with correctly identifying class “0.”

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 59 30
         1 45 66

               Accuracy : 0.625
                 95% CI : (0.5539, 0.6923)
    No Information Rate : 0.52
    P-Value [Acc > NIR] : 0.001763

                  Kappa : 0.2533

 Mcnemar's Test P-Value : 0.105969

            Sensitivity : 0.5673
            Specificity : 0.6875
         Pos Pred Value : 0.6629
         Neg Pred Value : 0.5946
             Prevalence : 0.5200
         Detection Rate : 0.2950
   Detection Prevalence : 0.4450
      Balanced Accuracy : 0.6274

       'Positive' Class : 0
```

While KNN provided reasonable results, the results indicate that KNN may not be the best model for this data set. But it can still be improved by using weighted KNN, tuning the choice of k with cross-validation, or considering alternative distance metrics.

Random Forest Model

Summary of model

```
Call:
 randomForest(formula = DRK_YN ~ ., data = train, importance = TRUE,      proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 31.62%
Confusion matrix:
    0   1 class.error
0 244 141   0.3662338
1 112 303   0.2698795
```

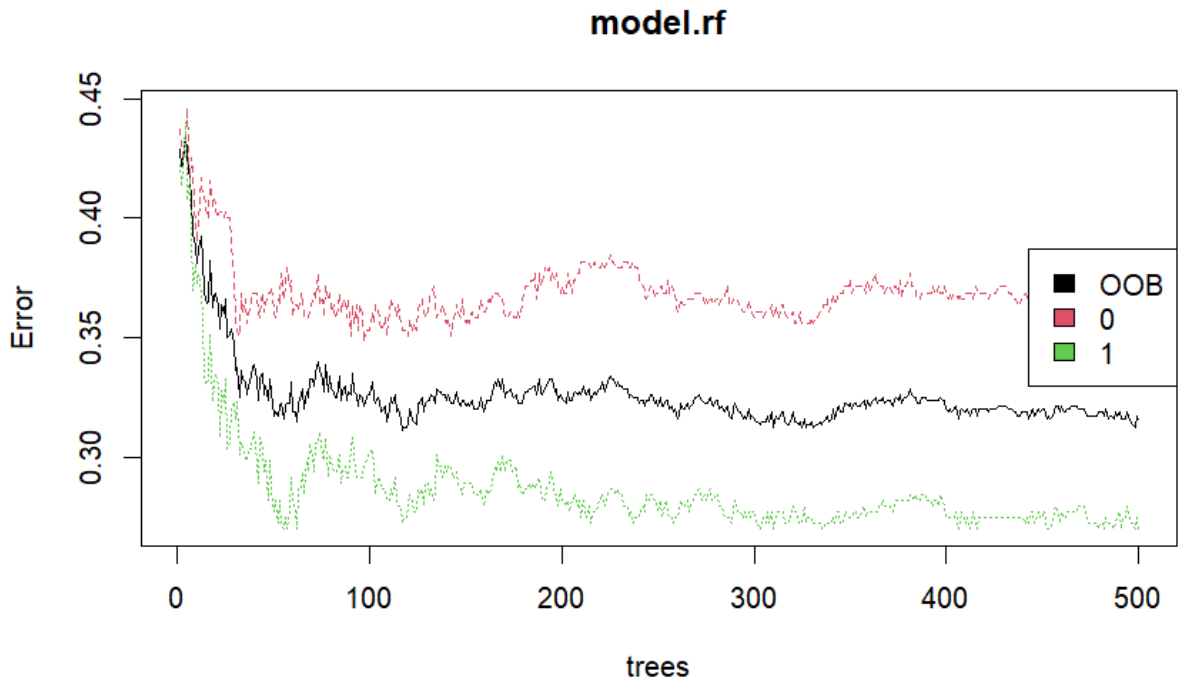
Confusion Matrix

```
 Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 65 23
         1 36 76

               Accuracy : 0.705
                 95% CI : (0.6366, 0.7672)
    No Information Rate : 0.505
    P-Value [Acc > NIR] : 7.188e-09

                  Kappa : 0.4107

 Mcnemar's Test P-Value : 0.1182

            Sensitivity : 0.6436
            Specificity : 0.7677
         Pos Pred Value : 0.7386
         Neg Pred Value : 0.6786
             Prevalence : 0.5050
         Detection Rate : 0.3250
   Detection Prevalence : 0.4400
      Balanced Accuracy : 0.7056

       'Positive' Class : 0
```

The Error Rate in this model: 31.62%

We applied the Random Forest model to our dataset, ensuring a proper 80-20 train-test split once again. The model was built with a default setting of 500 trees, selecting 4 features at each split from the Random Forest package, optimizing for both performance and stability. After training, we evaluated the accuracy of the model using a confusion matrix, which showed a 70.5% overall accuracy. The model correctly identified class “1” (specificity: 76.77%) more often than class “0” (sensitivity: 64.36%), indicating it performs better at detecting one category over the other.

Error rate plot

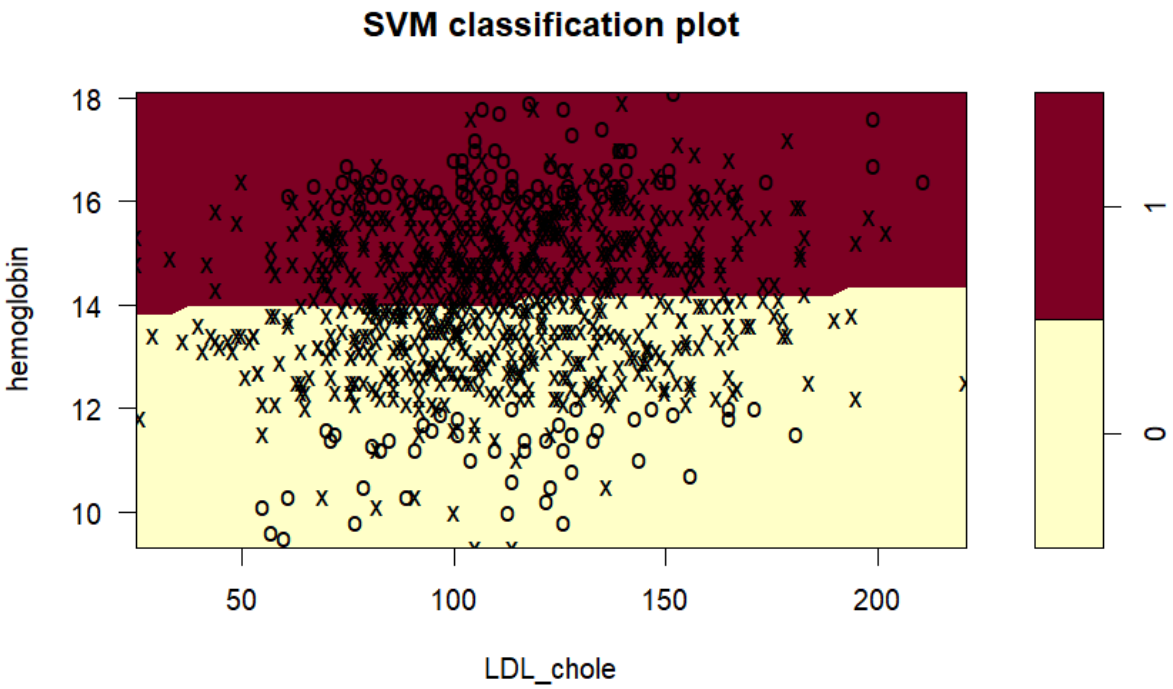


The error rate plot showed that after a certain number of trees, the out-of-bag (OOB) error stabilized at 31.62%, showing that the model is consistent but adding more trees did not significantly improve performance. This suggests that while the model generalizes well, further tuning might be needed to reduce classification errors. Also, the class-wise error rates from the confusion matrix indicate that class "0" has a higher misclassification rate (36.62%) compared to class "1" (26.98%), meaning the model struggles more with "0" class once again. This could be due to imbalanced data or overlapping features. Overall, the Random Forest model is a best choice for predicting our data set, given its robustness and accuracy.

SVM Model

```
Call:
svm(formula = DRK_YN ~ hemoglobin + LDL_chole, data = train_svm,
    type = "C-classification", kernel = "linear", cost = 0.1, scale = FALSE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  0.1

Number of Support Vectors:  675
```

Plot



The model was trained using all features, while for visualization, we selected two features (hemoglobin and LDL_chole) to create a decision boundary plot. The cost parameter was set to 0.1, allowing for some misclassifications while preventing overfitting. The decision boundary plot clearly shows that class “1” occupies the upper region (dark red), while class “0” is in the lower region (yellow), with some overlapping points indicating some challenges in the classification.

Confusion Matrix

```
[1] 0.64
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 52 23
         1 49 76

               Accuracy : 0.64
                 95% CI : (0.5693, 0.7065)
    No Information Rate : 0.505
    P-Value [Acc > NIR] : 8.098e-05

                  Kappa : 0.2818

 Mcnemar's Test P-Value : 0.003216

            Sensitivity : 0.5149
            Specificity : 0.7677
         Pos Pred Value : 0.6933
         Neg Pred Value : 0.6080
             Prevalence : 0.5050
         Detection Rate : 0.2600
   Detection Prevalence : 0.3750
      Balanced Accuracy : 0.6413

       'Positive' Class : 0
```

Error rate in this model: 0.64

The confusion matrix revealed an overall accuracy of 64%, which means the model correctly classified about 64% of test cases. The specificity (76.77%) shows the model performed well in identifying class “1”, but the sensitivity (51.49%) was lower, meaning it struggled more with correctly identifying class “0”. The presence of 675 support vectors suggests that model required many data points to define the decision boundary, which may indicate some overlap between the two classes. To get better results for the model, we could try scaling the features or fine-tuning hyperparameters.

Conclusion

After analyzing different classification models in this project, we found that the Random Forest model is the best for predicting DRK_YN. It achieved the second highest accuracy of 70.5%, performing better at identifying class “1” (specificity: 76.77%) than class “0” (sensitivity: 64.36%). This suggests that while it effectively detects positive cases, it has some difficulty classifying negative ones but performed much better than other models.. Compared to other models, Linear Discriminant Analysis (LDA) had a slightly higher accuracy of 73%, but misclassification occurred due to overlapping data distributions. K-Nearest Neighbors (KNN) showed 62.5% accuracy, struggling more with class “0” (sensitivity: 56.73%), while Support Vector Machine (SVM) had 64% accuracy but required 675 support vectors, indicating significant class overlap.

Although LDA had the highest accuracy, Random Forest produced more balanced predictions across both classes, making it a more reliable model overall. KNN and SVM had lower performance, particularly in predicting class "0". The misclassification rate for class "0" in Random Forest was 36.62%, which means there is still room for improvement. However, the model remains the best choice due to its stability across different subsets, ability to handle complex relationships, and robustness compared to other models.

To further improve the Random Forest model, we can address the class imbalance issue, as class "0" had a higher misclassification rate. Methods like SMOTE (Synthetic Minority Over-sampling Technique) or class-weight adjustments could help improve classification accuracy for the minority class. Additionally, we can explore alternative models like Gradient Boosting (GBM) or XGBoost, which may perform better by reducing bias and variance. Finally, implementing 10-fold cross-validation instead of a single train-test split would help ensure the model generalizes better to new data. With these adjustments, we can make Random Forest even more accurate and reliable for predictions.

References:

https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset