Classification: Predicting Wheat Variety Using Kernel Geometrical Attributes

Prithviraj Lakkakula

03/18/2022

Contents

Goal

The project's goal is to accurately predict the wheat variety (`Kama`, `Rosa`, `Canadian`) using the seven attributes corresponding to each of wheat variety.

Data

In this project, I classify wheat variety based on the wheat kernel's geometrical properties. There are three varieties of wheat (`Kama`, `Rosa`, and `Canadian`), which is my class variable. Each variety has 70 observations accounting for a total of 210 observations. The features (X) are seven attributes, including area, perimeter, compactness, length of the kernel, width of the kernel, asymmetry coefficient, and length of kernel groove. Data are collected from UC Irvine Machine Learning Repository at https://archive-beta.ics.uci.edu/ml/datasets/seeds.

Data Preprocessing

```
library(dplyr)
wht_data <- read.csv("wheat_var_data.csv")
glimpse(wht_data)
```

```
## Rows: 210
## Columns: 8
## $ area                <dbl> 15.26, 14.88, 14.29, 13.84, 16.14, 14.38, 14.69, ~
## $ perimeter           <dbl> 14.84, 14.57, 14.09, 13.94, 14.99, 14.21, 14.49, ~
## $ compactness         <dbl> 0.8710, 0.8811, 0.9050, 0.8955, 0.9034, 0.8951, 0~
## $ length_kernel       <dbl> 5.763, 5.554, 5.291, 5.324, 5.658, 5.386, 5.563, ~
## $ width_kernel        <dbl> 3.312, 3.333, 3.337, 3.379, 3.562, 3.312, 3.259, ~
## $ asymmetry_coef      <dbl> 2.2210, 1.0180, 2.6990, 2.2590, 1.3550, 2.4620, 3~
## $ length_kernel_groove <dbl> 5.220, 4.956, 4.825, 4.805, 5.175, 4.956, 5.219, ~
## $ wheat_variety       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

```
summary(wht_data)
```

```
##       area         perimeter      compactness     length_kernel
##  Min.   :10.59   Min.   :12.41   Min.   :0.8081   Min.   :4.899
##  1st Qu.:12.27   1st Qu.:13.45   1st Qu.:0.8569   1st Qu.:5.262
##  Median :14.36   Median :14.32   Median :0.8734   Median :5.524
##  Mean   :14.85   Mean   :14.56   Mean   :0.8710   Mean   :5.629
##  3rd Qu.:17.30   3rd Qu.:15.71   3rd Qu.:0.8878   3rd Qu.:5.980
##  Max.   :21.18   Max.   :17.25   Max.   :0.9183   Max.   :6.675
##   width_kernel   asymmetry_coef   length_kernel_groove wheat_variety
##  Min.   :2.630   Min.   :0.7651   Min.   :4.519        Min.   :1
##  1st Qu.:2.944   1st Qu.:2.5615   1st Qu.:5.045        1st Qu.:1
##  Median :3.237   Median :3.5990   Median :5.223        Median :2
##  Mean   :3.259   Mean   :3.7002   Mean   :5.408        Mean   :2
##  3rd Qu.:3.562   3rd Qu.:4.7687   3rd Qu.:5.877        3rd Qu.:3
##  Max.   :4.033   Max.   :8.4560   Max.   :6.550        Max.   :3
```

By inspecting mean and median of all seven attributes, one can conclude that there are no outliers/anomalies. Also, we need to convert the `wheat_variety` variable into categorical or qualitative or class variable instead of an integer.

```
library(dplyr)
wht_data$wheat_variety <- as.factor(wht_data$wheat_variety)
wht_data <- wht_data %>%
  mutate(wheat_var =
         ifelse(wheat_variety == "1", "Kama",
                ifelse(wheat_variety == "2", "Rosa", "Canadian"))) %>%
  select(-wheat_variety)
str(wht_data)
```

```
## 'data.frame':    210 obs. of  8 variables:
##  $ area                : num  15.3 14.9 14.3 13.8 16.1 ...
##  $ perimeter           : num  14.8 14.6 14.1 13.9 15 ...
##  $ compactness         : num  0.871 0.881 0.905 0.895 0.903 ...
##  $ length_kernel       : num  5.76 5.55 5.29 5.32 5.66 ...
##  $ width_kernel        : num  3.31 3.33 3.34 3.38 3.56 ...
##  $ asymmetry_coef      : num  2.22 1.02 2.7 2.26 1.35 ...
##  $ length_kernel_groove: num  5.22 4.96 4.83 4.8 5.17 ...
##  $ wheat_var           : chr  "Kama" "Kama" "Kama" "Kama" ...
```

Exploratory Data Analysis

Let us now look at the relationships of the three wheat varieties with each of the seven features.

```
library(dplyr)
wht_data %>%
  group_by(wheat_var) %>%
  summarise_all(mean)
```

```
## # A tibble: 3 x 8
##   wheat_var  area perimeter compactness length_kernel width_kernel
##   <chr>     <dbl>     <dbl>       <dbl>         <dbl>        <dbl>
## 1 Canadian   11.9      13.2       0.849          5.23         2.85
## 2 Kama       14.3      14.3       0.880          5.51         3.24
## 3 Rosa       18.3      16.1       0.884          6.15         3.68
## # ... with 2 more variables: asymmetry_coef <dbl>, length_kernel_groove <dbl>
```

On average, `Rosa` wheat variety seem to have higher length, width, area, perimeter and compactness, followed by `Kama` variety. However, `Canadian` variety has the highest average asymmetry coefficient compared with other wheat varieties.

```
library(ggplot2)
library(geomtextpath)

ggplot(wht_data, aes(x = length_kernel, colour = wheat_var, label = wheat_var)) +
  geom_textdensity(size = 6, fontface = 2, hjust = 0.2, vjust = 0.3) +
  theme(legend.position = "none") + theme_bw()
```

```
library(ggplot2)
library(geomtextpath)
ggplot(wht_data, aes(x = asymmetry_coef, colour = wheat_var,
                     label = wheat_var)) +
       theme(legend.position = "none") +
       geom_textdensity(size = 6, fontface = 2, spacing = 50,
                   vjust = -0.2, hjust = "ymax") + ylim(c(0, 0.4)) + theme_minimal()
```

```
ggplot(wht_data, aes(x = length_kernel, y = width_kernel,
                     color = wheat_var)) +
  geom_point(alpha = 0.3) + theme(legend.position = "bottom") +
  geom_labelsmooth(aes(label = wheat_var), text_smoothing = 30,
                   fill = "#F6F6FF",
                method = "loess", formula = y ~ x,
                size = 4, linewidth = 1, boxlinewidth = 0.3) +
  scale_colour_manual(values = c("forestgreen", "deepskyblue4", "tomato4")) +
   theme_bw()
```

Correlation Pairs

Now, let us look at the range of all the variables except the response variable.

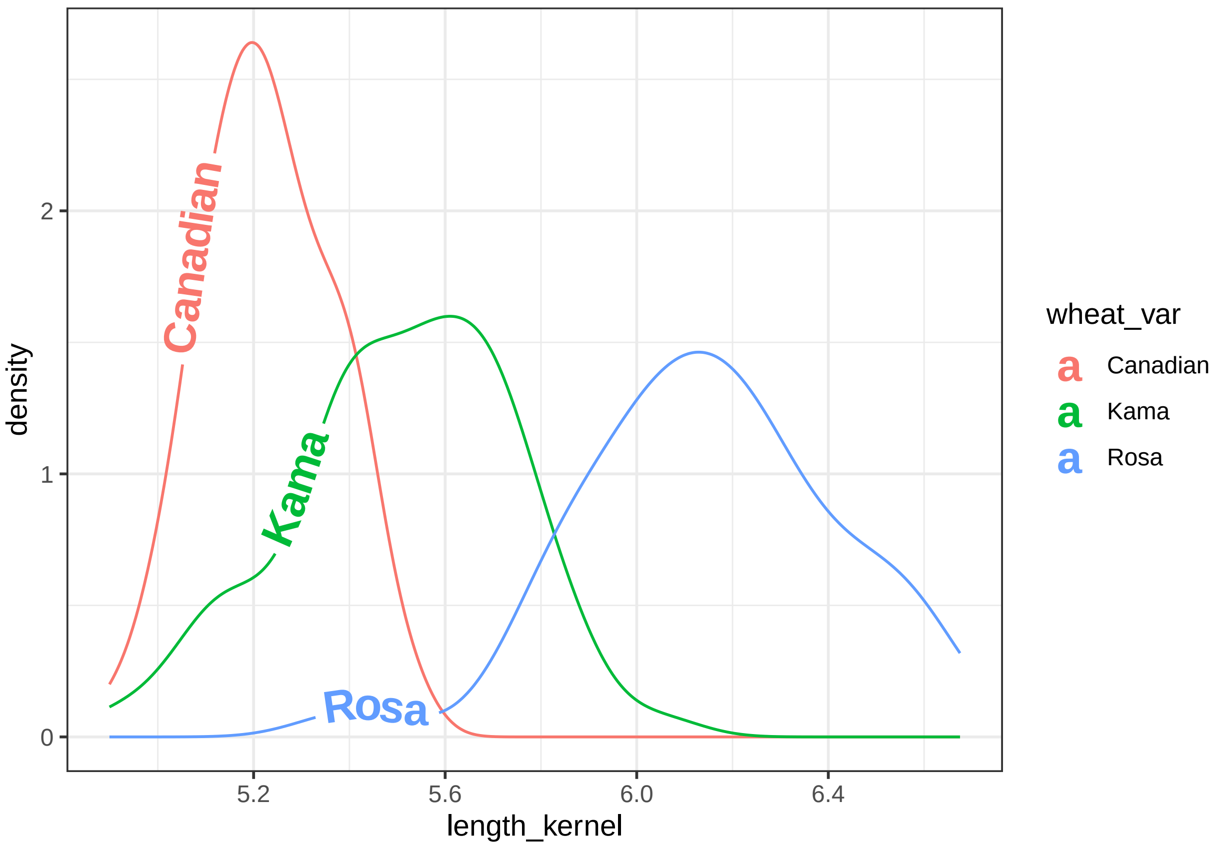


Figure 1: Density plot of kernel length of three wheat varieties

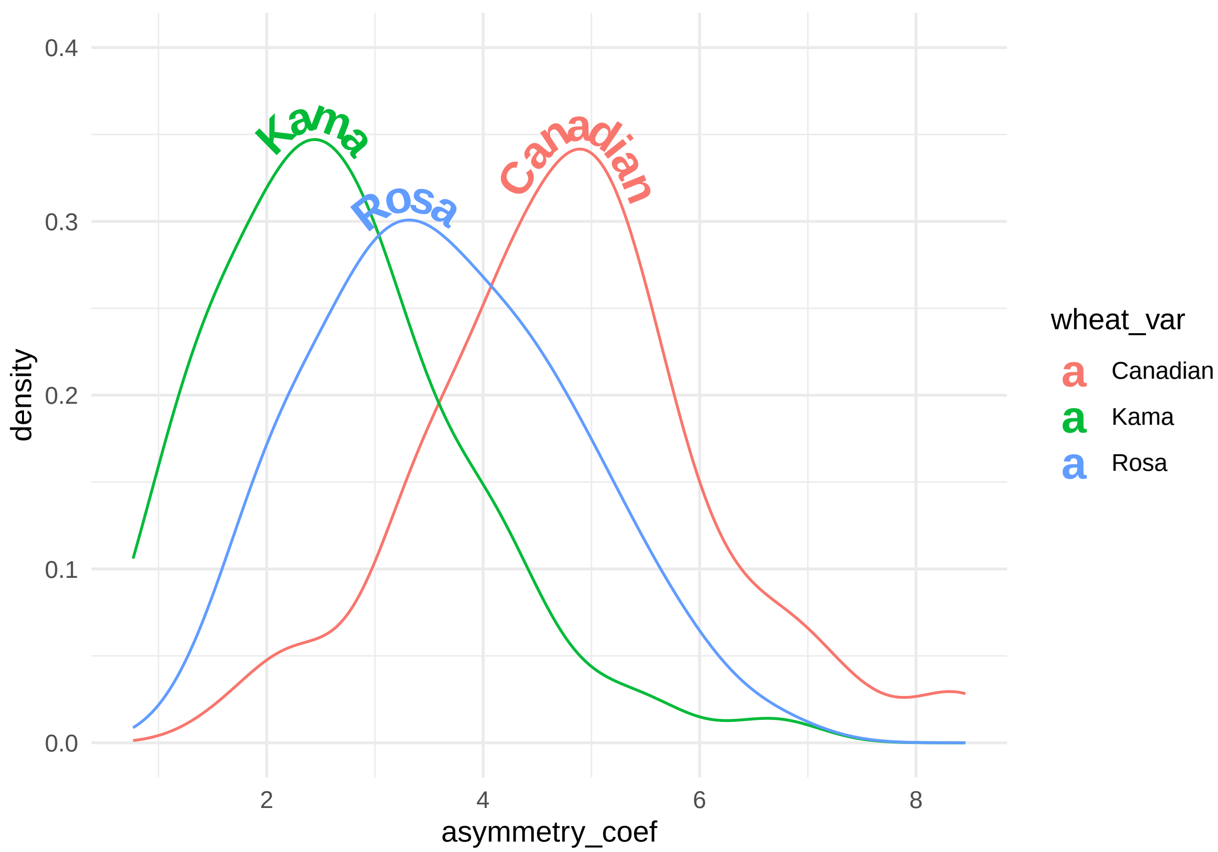


Figure 2: Density plot of asymmetry coefficient of three wheat varieties

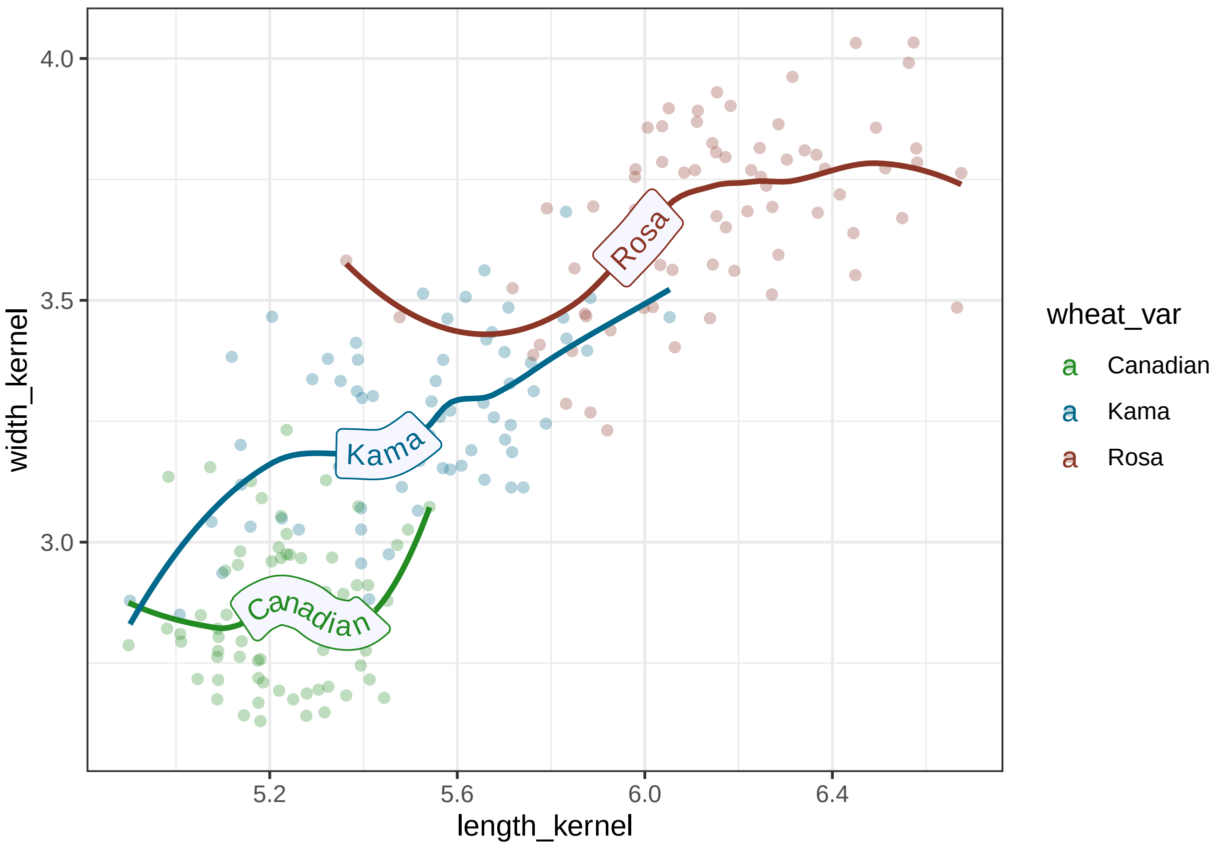


Figure 3: Trend Lines through scatter plot of length and width of wheat varieties

```r
wht_data %>%
  select(-wheat_var) %>%
  summarise_all(range)
```

```
##    area perimeter compactness length_kernel width_kernel asymmetry_coef
## 1 10.59     12.41      0.8081         4.899        2.630         0.7651
## 2 21.18     17.25      0.9183         6.675        4.033         8.4560
##   length_kernel_groove
## 1                4.519
## 2                6.550
```

Standardization of the features

Since some of the variables are in different range than the others. Let us do Z-score normalization or standardization the `scale()` function in R. When applying the decision trees (random forests and gradient boosting) and KNN machine leanrning algorithms, we may need not scale.

```r
library(dplyr)
##Z-score normalization
wht_data_scaled <- wht_data %>% mutate_each_(list(~scale(.) %>% as.vector),
vars = c("area","perimeter", "compactness",
         "length_kernel", "width_kernel",
         "asymmetry_coef", "length_kernel_groove"))
head(wht_data_scaled)
```

```
##          area    perimeter  compactness length_kernel width_kernel
## 1  0.14175904  0.214948819 6.045733e-05    0.30349301    0.1413640
## 2  0.01116136  0.008204153 4.274938e-01   -0.16822270    0.1969616
## 3 -0.19160873 -0.359341919 1.438945e+00   -0.76181710    0.2075516
## 4 -0.34626388 -0.474200066 1.036904e+00   -0.68733567    0.3187467
## 5  0.44419577  0.329806966 1.371233e+00    0.06650665    0.8032397
## 6 -0.16067770 -0.267455401 1.019976e+00   -0.54740087    0.1413640
##   asymmetry_coef length_kernel_groove wheat_var
## 1     -0.9838010           -0.3826631      Kama
## 2     -1.7839036           -0.9198156      Kama
## 3     -0.6658882           -1.1863572      Kama
## 4     -0.9585276           -1.2270506      Kama
## 5     -1.5597684           -0.4742231      Kama
## 6     -0.8235144           -0.9198156      Kama
```

Near-zero variance features

```r
library(caret)
near_0_var <- nearZeroVar(wht_data, names = TRUE)
print(near_0_var)
```

```
## character(0)
```

It seems like there are no zero variance features, which is good. Therefore, we can use all the features to predict the the class of wheat variety.

Checking for class imbalance

We already know that there are equal observations for each of the wheat variety in our dataset. That is, each variety has 70 observations for a total of 210 observations. Therefore, our data set do not suffer with class imbalance

```
table(wht_data$wheat_var)
```

```
##
## Canadian     Kama     Rosa
##       70       70       70
```

Ensemble Models

Splitting the data

```
library(caret)

set.seed(4321)
wht_data_scaled$wheat_var <- as.factor(wht_data_scaled$wheat_var)
in_train <- createDataPartition(y = wht_data_scaled$wheat_var,
                                p = 0.80, list = FALSE)

training <- wht_data_scaled[in_train,]
testing <- wht_data_scaled[-in_train,]

table(training$wheat_var)
```

```
##
## Canadian     Kama     Rosa
##       56       56       56
```

```
table(testing$wheat_var)
```

```
##
## Canadian     Kama     Rosa
##       14       14       14
```

```
head(training)
```

```
##          area    perimeter compactness length_kernel width_kernel
## 2  0.01116136  0.008204153   0.4274938   -0.16822270  0.196961591
## 3 -0.19160873 -0.359341919   1.4389449   -0.76181710  0.207551602
## 4 -0.34626388 -0.474200066   1.0369037   -0.68733567  0.318746714
## 5  0.44419577  0.329806966   1.3712327    0.06650665  0.803239702
## 7 -0.05413749 -0.053053525   0.3767096   -0.14790958  0.001046394
## 8 -0.25347079 -0.351684709   0.8506951   -0.47066243  0.114889009
##   asymmetry_coef length_kernel_groove wheat_var
## 2    -1.78390358           -0.9198156      Kama
```

```
## 3    -0.66588820           -1.1863572      Kama
## 4    -0.95852756           -1.2270506      Kama
## 5    -1.55976843           -0.4742231      Kama
## 7    -0.07595385           -0.3846977      Kama
## 8    -0.66522311           -0.8302902      Kama
```

```r
str(training)
```

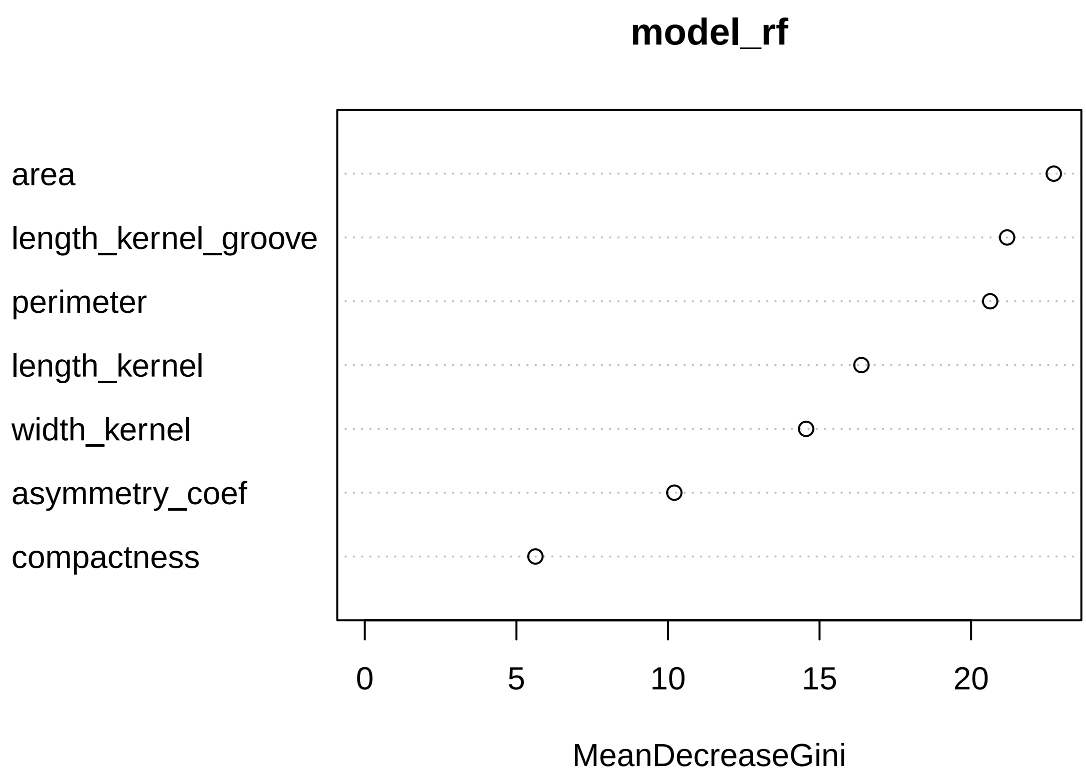
```
## 'data.frame':    168 obs. of  8 variables:
##  $ area                : num  0.0112 -0.1916 -0.3463 0.4442 -0.0541 ...
##  $ perimeter           : num  0.0082 -0.3593 -0.4742 0.3298 -0.0531 ...
##  $ compactness         : num  0.427 1.439 1.037 1.371 0.377 ...
##  $ length_kernel       : num  -0.1682 -0.7618 -0.6873 0.0665 -0.1479 ...
##  $ width_kernel        : num  0.19696 0.20755 0.31875 0.80324 0.00105 ...
##  $ asymmetry_coef      : num  -1.784 -0.666 -0.959 -1.56 -0.076 ...
##  $ length_kernel_groove: num  -0.92 -1.186 -1.227 -0.474 -0.385 ...
##  $ wheat_var           : Factor w/ 3 levels "Canadian","Kama",..: 2 2 2 2 2 2 2 2 2 2 ...
```

Classficiation: Random Forest

```r
### load the randomForest package
library(randomForest)

### train the random forest model: model_rf
model_rf <- randomForest(formula = wheat_var ~.,
                         data = training,
                         ntree = 500)

### print the rf model
print(model_rf)
```

```
##
## Call:
##  randomForest(formula = wheat_var ~ ., data = training, ntree = 500)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 7.74%
## Confusion matrix:
##          Canadian Kama Rosa class.error
## Canadian       53    3    0  0.05357143
## Kama            5   49    2  0.12500000
## Rosa            0    3   53  0.05357143
```

```r
### variable importance plots
varImpPlot(model_rf)
```




```r
print(model_rf$importance)
```

```
##                      MeanDecreaseGini
## area                        22.727897
## perimeter                   20.628287
## compactness                  5.628345
## length_kernel               16.382484
## width_kernel                14.556988
## asymmetry_coef              10.210161
## length_kernel_groove        21.187266
```

Classification : Gradient Boosting Model

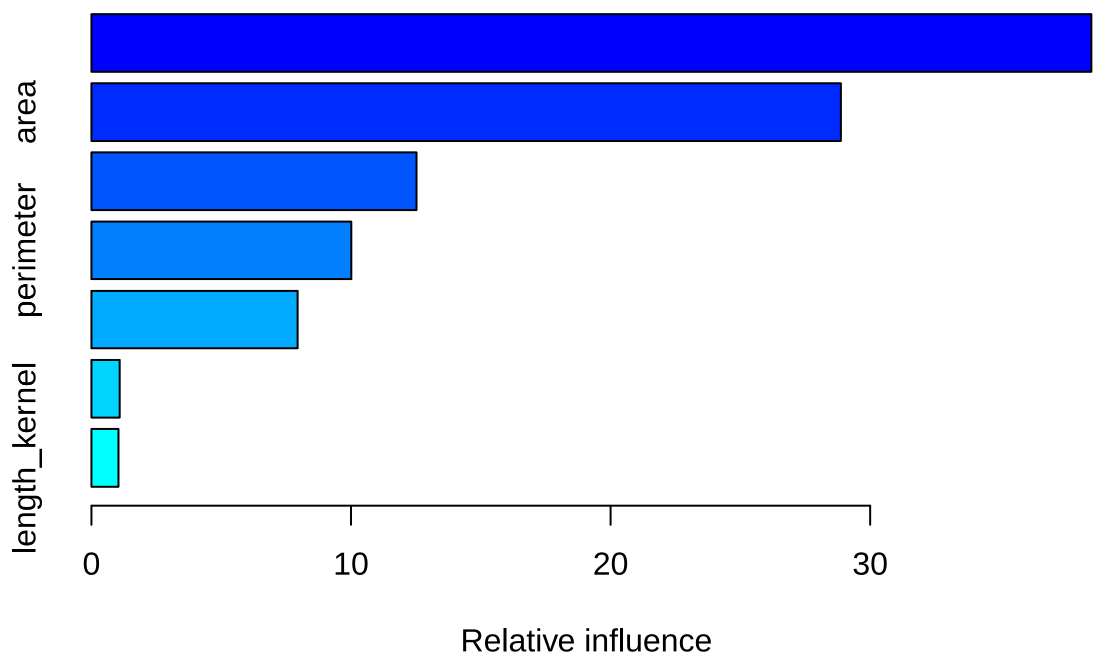
```r
### load the gradient boosting model package
library(gbm)

### train the gradient boosting model: model_gbm
model_gbm <- gbm(formula = wheat_var ~.,
                         data = training,
                         n.trees = 500)
```

```
## Distribution not specified, assuming multinomial ...
```

```r
### print the gbm model
print(model_gbm)
```

```
## gbm(formula = wheat_var ~ ., data = training, n.trees = 500)
## A gradient boosted model with multinomial loss function.
## 500 iterations were performed.
## There were 7 predictors of which 7 had non-zero influence.
```

```r
### summarize gbm's variable importance plots
summary(model_gbm)
```




```
##                                       var   rel.inf
## length_kernel_groove length_kernel_groove 38.521547
## area                                 area 28.871845
## asymmetry_coef             asymmetry_coef 12.522778
## perimeter                       perimeter 10.011092
## width_kernel                 width_kernel  7.942892
## compactness                   compactness  1.087771
## length_kernel               length_kernel  1.042073
```

Evaluating both Random Forest and Gradient Boosting Algorithms

```r
library(Metrics)

preds_rf <- predict(model_rf, newdata = testing)
preds_gbm <- predict(model_gbm, n.trees = 500, newdata = testing, type = "response")
## compute confusion matrix



classes <- colnames(preds_gbm)[apply(preds_gbm, 1, which.max)]
result_gbm <- data.frame(testing$wheat_var, classes)

#print(result_gbm)
(cm_rf <- confusionMatrix(preds_rf, testing$wheat_var))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Canadian Kama Rosa
##   Canadian       14    0    0
##   Kama            0   12    0
##   Rosa            0    2   14
##
## Overall Statistics
##
##                Accuracy : 0.9524
##                  95% CI : (0.8384, 0.9942)
##     No Information Rate : 0.3333
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9286
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: Canadian Class: Kama Class: Rosa
## Sensitivity                   1.0000      0.8571      1.0000
## Specificity                   1.0000      1.0000      0.9286
## Pos Pred Value                1.0000      1.0000      0.8750
## Neg Pred Value                1.0000      0.9333      1.0000
## Prevalence                    0.3333      0.3333      0.3333
## Detection Rate                0.3333      0.2857      0.3333
## Detection Prevalence          0.3333      0.2857      0.3810
## Balanced Accuracy             1.0000      0.9286      0.9643
```

```r
(cm_gbm <- confusionMatrix(as.factor(classes), testing$wheat_var))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Canadian Kama Rosa
##   Canadian       13    0    0
```

```
##   Kama            1   12    0
##   Rosa            0    2   14
##
## Overall Statistics
##
##                Accuracy : 0.9286
##                  95% CI : (0.8052, 0.985)
##     No Information Rate : 0.3333
##     P-Value [Acc > NIR] : 8.716e-16
##
##                   Kappa : 0.8929
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: Canadian Class: Kama Class: Rosa
## Sensitivity                   0.9286      0.8571      1.0000
## Specificity                   1.0000      0.9643      0.9286
## Pos Pred Value                1.0000      0.9231      0.8750
## Neg Pred Value                0.9655      0.9310      1.0000
## Prevalence                    0.3333      0.3333      0.3333
## Detection Rate                0.3095      0.2857      0.3333
## Detection Prevalence          0.3095      0.3095      0.3810
## Balanced Accuracy             0.9643      0.9107      0.9643
```

Conclusions

The ensemble models suggest that there is an accuracy of about 95% using Random Forest and 93% using GBM in predicting the correct wheat variety using a set of features. In the UC Irvine's data repository, it was indicated that there was some critical features that they could not provide due to proprietary issues associated with those data. Therefore, given those additional features, there is a scope for improving accuracy rate. Overall, the classification results show that accuracy of predicting the correct wheat variety is high.