Project 2

David Cruz

11/20/2020

Introduction

The purpose of this project is to classify stars as giants or dwarfs based on their visual apparent magnitude, distance between the star and earth, B.V color index, and absolute magnitude. I built a random forest model and a variable importance plot for this classification type problem.

Importing the Data Set

```
library(tidyverse)
```

```
## ── Attaching packages ──────────────────────────────────────── tidyverse 1.3.0 ──
```

```
## ✓ ggplot2 3.3.2     ✓ purrr   0.3.4
## ✓ tibble  3.0.3     ✓ dplyr   1.0.2
## ✓ tidyr   1.1.2     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.5.0
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(randomForest)
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
data <- read.csv("Star39552_balanced.csv")
```

Cleaning the data

Changing bland column names into meaningful column names.

```
names(data)[names(data) == 'Vmag'] <- 'VisualAppMagnitude'
names(data)[names(data) == 'Plx'] <- 'DistanceStarEarth'
names(data)[names(data) == 'Amag'] <- 'AbsoluteMagnitude'
names(data)[names(data) == 'TargetClass'] <- 'StarClassification'
```

Changing StarClassification into a factor variable and renaming levels into meaningful names.

```
data$StarClassification <- as.factor(data$StarClassification)
levels(data$StarClassification) <- c('Dwarf', 'Giant')
```

Removing the e_Plx and SpType columns. The e_Plx column displays the standard error of DistanceStarEarth. I removed these columns because I did not use this in my exploratory data analysis or predictive models section.

```
data$e_Plx <- NULL
data$SpType <- NULL
```

Checking for Unusual/NA Values:

```
sum(is.na(data))
```

```
## [1] 0
```

Data Dictionary

1. **VisualAppMagnitude (numeric)** - Visual Apparent Magnitude of the Star. This is a measure of luminosity of a celestial object.
2. **DistanceStarEarth (numeric)** - Distance between the Star and the Earth.
3. **B.V (numeric)** - B-V Color index. A hot star has a B-V color index that is negative or close to 0. A cool star has a B-V color index that is close to 2.0.
4. **AbsoluteMagnitude (numeric)** - Absolute Magnitude of the Star.
5. **StarClassification (factor: “Dwarf”, “Giant”)** - Classifies whether the star is a Dwarf or Giant.

Exploratory Data Analysis

Table: Amount of Dwarf and Giant Stars

```
table(data$StarClassification)
```

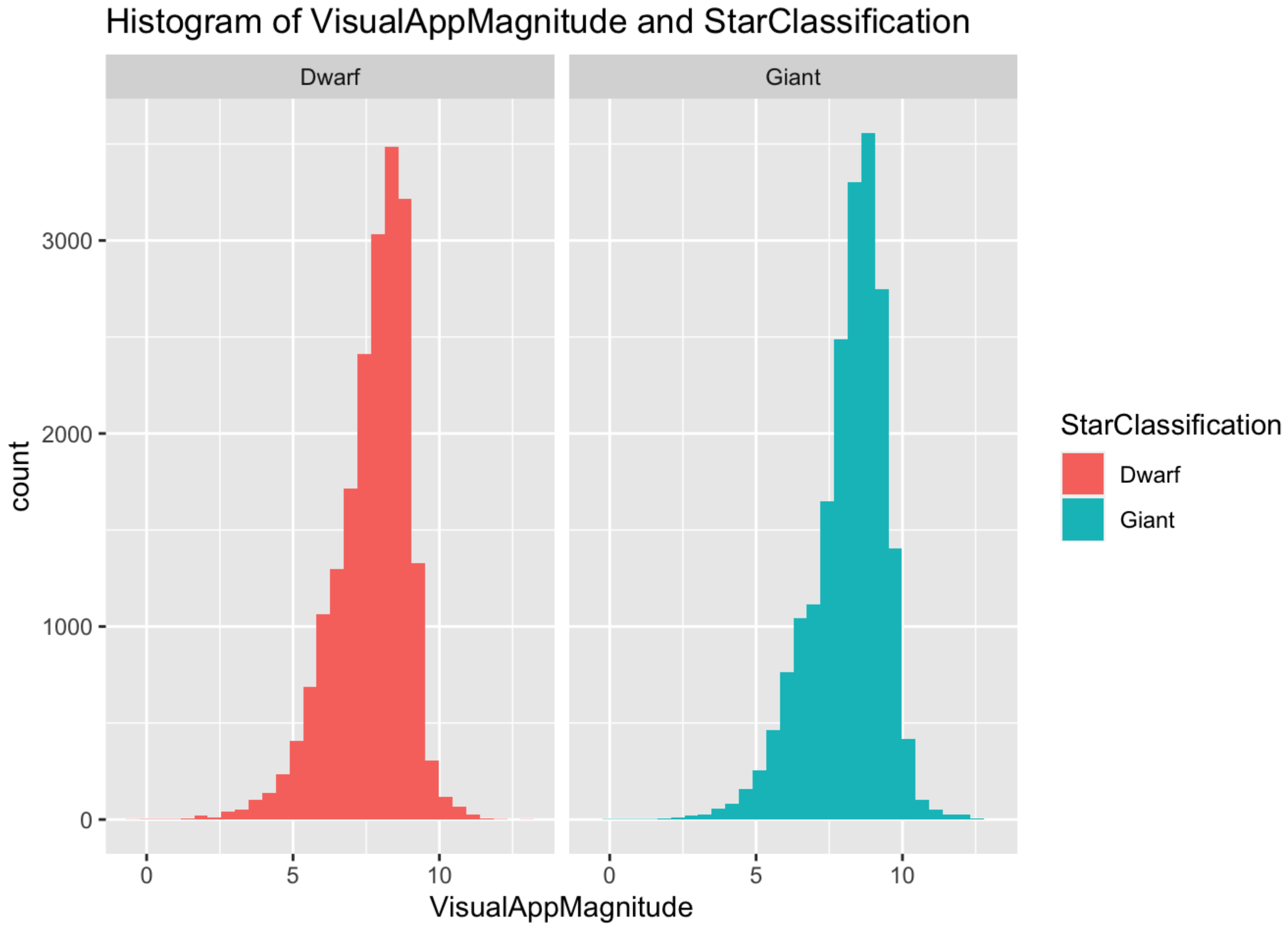
```
##
## Dwarf Giant
## 19776 19776
```

Histogram of VisualAppMagnitude and StarClassification

Very similar histogram shape here. The varImpPlot will probably say that this variable is not extremely important for classifying stars.

```
ggplot() + geom_histogram(data=data, aes(x=VisualAppMagnitude, fill=StarClassification))
+
  facet_wrap(~StarClassification) +
  ggtitle("Histogram of VisualAppMagnitude and StarClassification")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

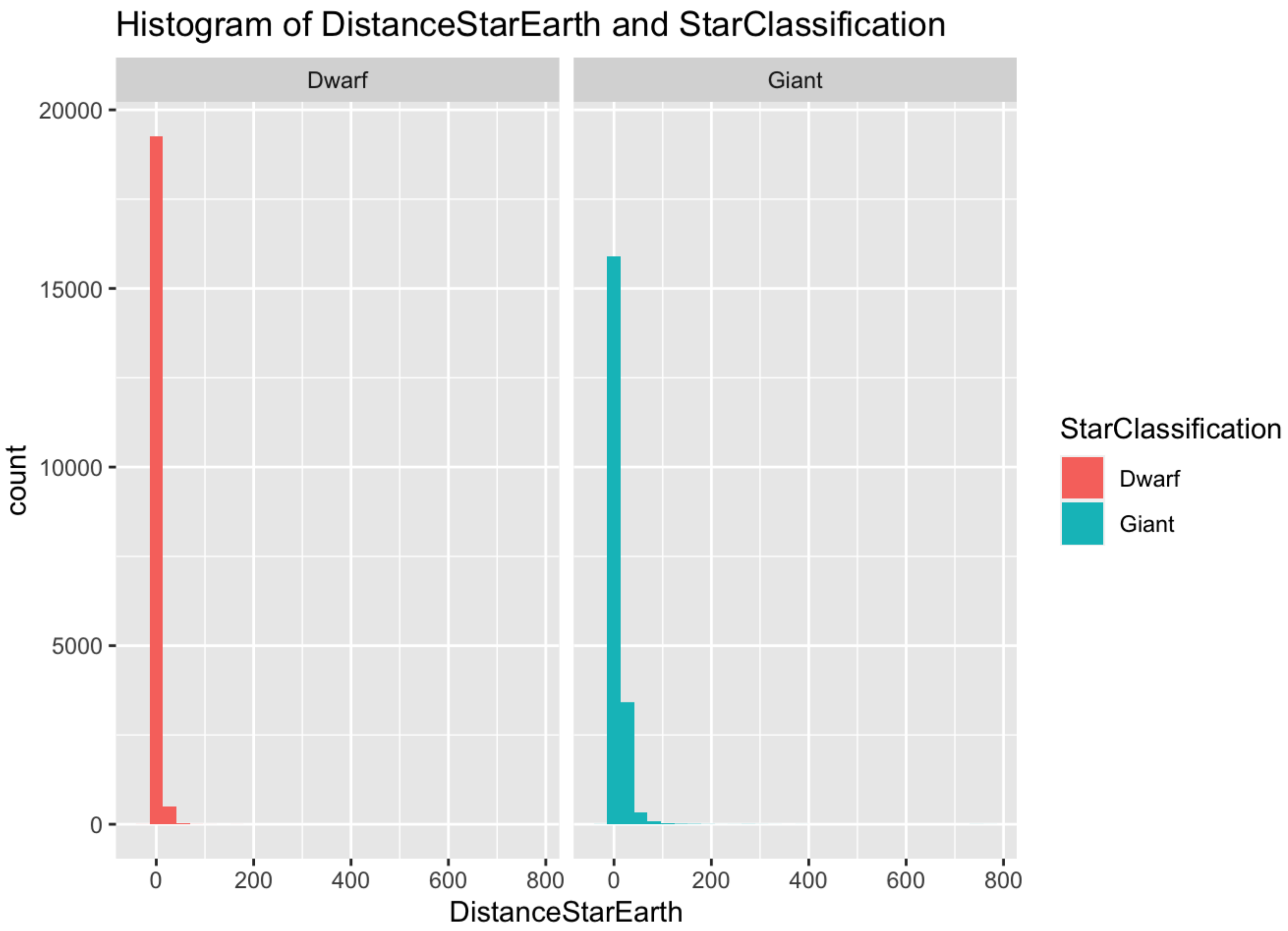


Histogram of DistanceStarEarth and StarClassification

Both star classifications have a similar shape here, except that much more dwarf stars are closer to Earth than Giant stars; nothing special to look at. The varImpPlot will also probably rate this variable with lower importance when classifying stars.

```
ggplot() + geom_histogram(data=data, aes(x=DistanceStarEarth,
  fill=StarClassification)) +
  facet_wrap(~StarClassification) +
  ggtitle("Histogram of DistanceStarEarth and StarClassification")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

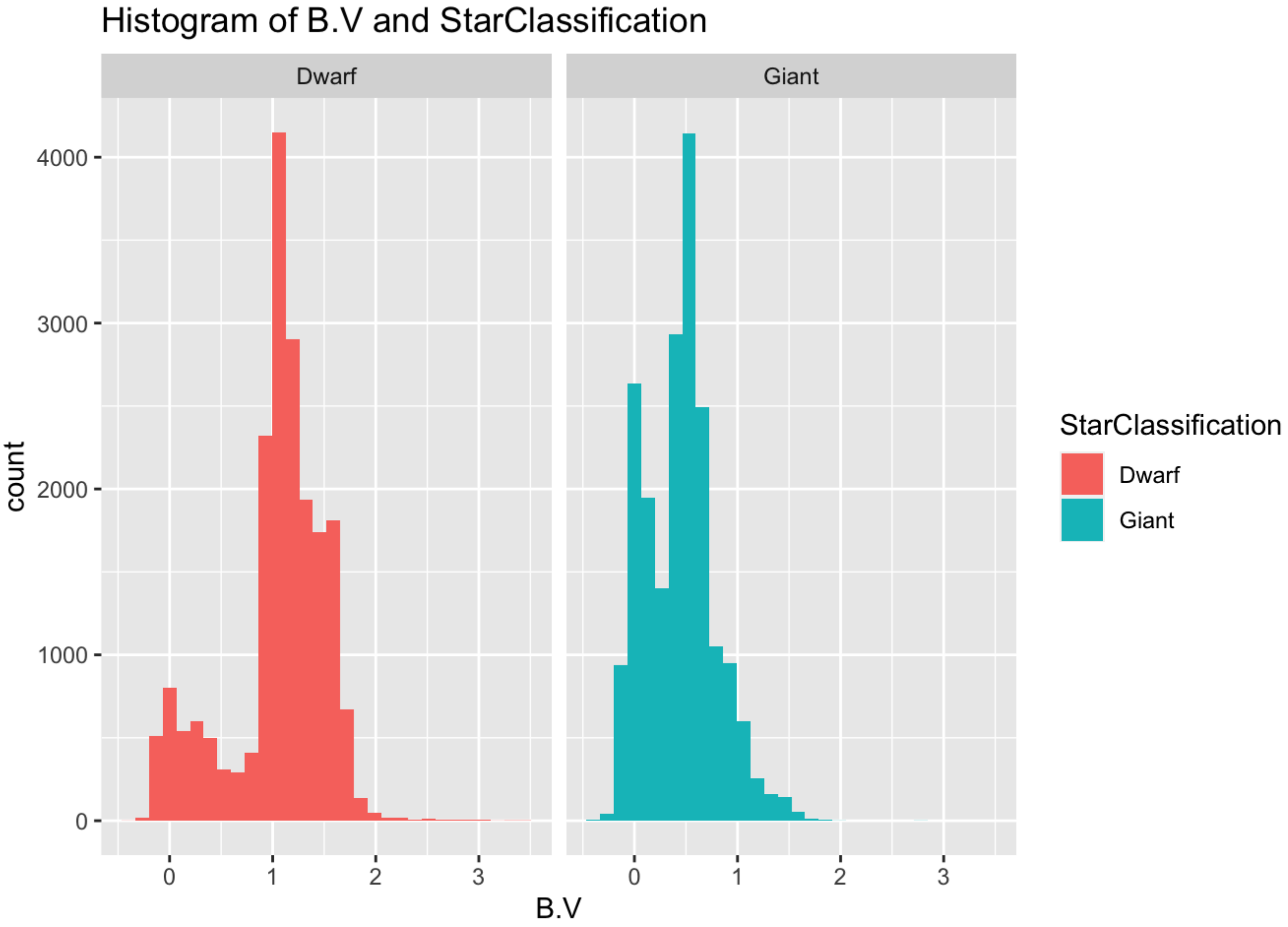


Histogram of B.V. and StarClassification

Notice that dwarf stars are closer to 2.0 on the B.V scale, which indicates that dwarf stars are cool. Also notice that giant stars are closer to 0 on the B.V scale, indicating that giant stars are hot. This variable will probably hold significant importance on the varImpPlot.

```
ggplot() + geom_histogram(data=data, aes(x=B.V,
  fill=StarClassification)) +
  facet_wrap(~StarClassification) +
  ggtitle("Histogram of B.V and StarClassification")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

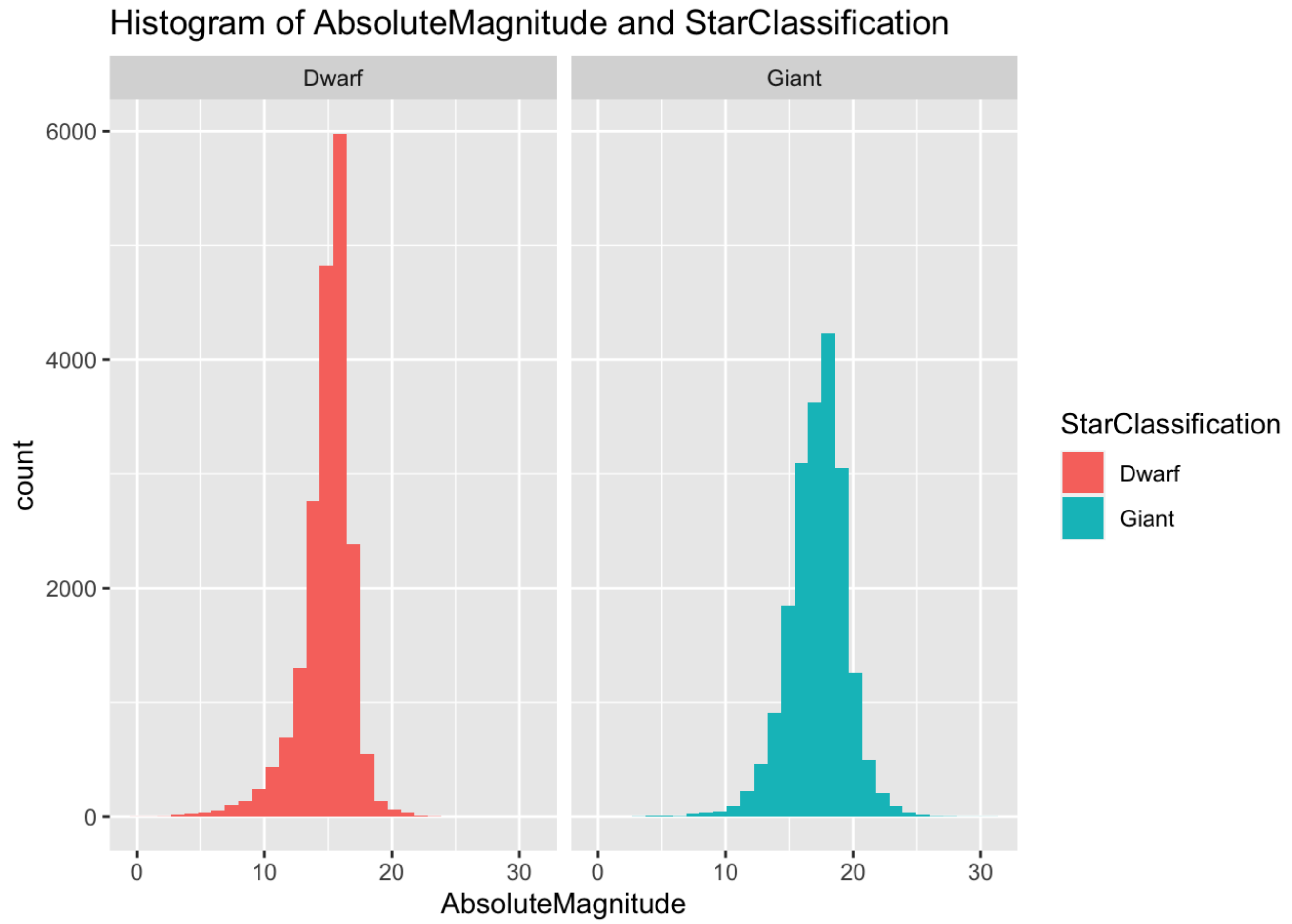


Histogram of AbsoluteMagnitude and StarClassification

Giant stars have a slightly larger absolute magnitude than dwarf stars. This may hold some importance on the varImpPlot.

```
ggplot() + geom_histogram(data=data, aes(x=AbsoluteMagnitude,
  fill=StarClassification)) +
  facet_wrap(~StarClassification) +
  ggtitle("Histogram of AbsoluteMagnitude and StarClassification")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Modeling and Model Analysis

For the random forest model, I wanted to see which variables are most helpful for classifying stars. Therefore, the response variable is StarClassification, and the predictor variables are VisualAppMagnitude, DistanceStarEarth, B.V, and AbsoluteMagnitude. The random forest model and variable importance plot shown below are particularly useful for astronomers who want to know which variables are important for correctly identifying stars. These models will provide additional evidence to already existing knowledge of star classification with these variables.

Train and Test Set:

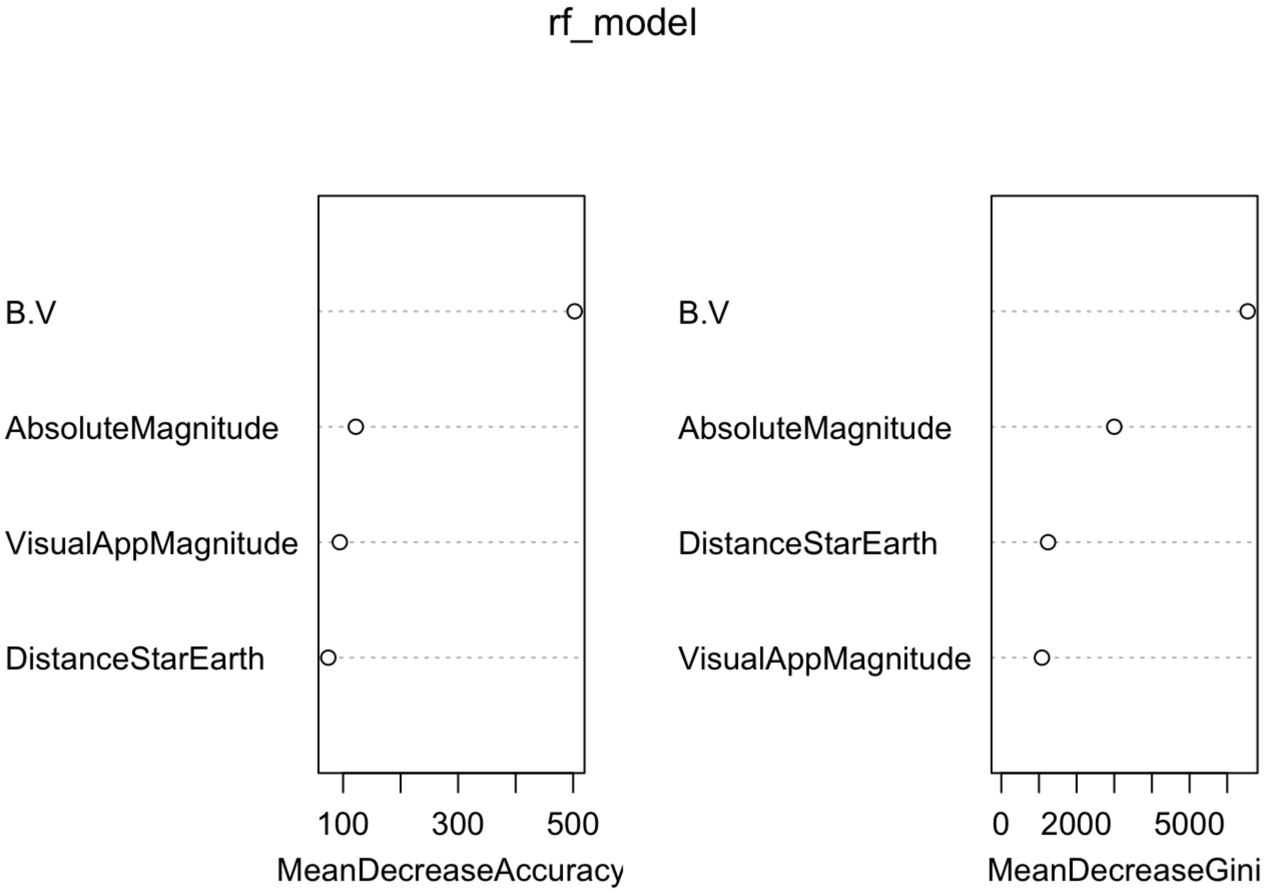
```
set.seed(1234)
data$ID <- 1:nrow(data)
train <- sample_frac(data, .6)
test <- anti_join(data, train,  by="ID")
train$ID <- NULL
test$ID <- NULL
```

Creating the random forest model:

```
rf_model <- randomForest(StarClassification ~ ., data=train,
                         ntree = 500,
                         mtry = 3,
                         importance = TRUE)
```

Looking at the varImpPlot below, this suggests that B.V and AbsoluteMagnitude are the two most important variables to analyze when classifying stars.

```
varImpPlot(rf_model)
```



Out of 15821 observations, 1908 of those were incorrectly classified. This is about a 12% error, which is a solid error rate. I have not tuned any parameters to achieve this error rate. Overall, the random forest performed well!

```
test_predict <- predict(rf_model,test)
test_actual <- test$StarClassification
table(test_predict,test_actual)
```

```
##             test_actual
## test_predict Dwarf Giant
##        Dwarf  6844   875
##        Giant  1034  7068
```

Conclusions and Questions for Further Study

Conclusions:

Of the four predictor variables, two of those variables–AbsoluteMagnitude and B.V.–rated highly on the variable importance plot. I investigated the relationship between these two variables and star classification by analyzing the histograms plotted for each variable. I observed that the AbsoluteMagnitude for Giant Stars were slightly larger than Dwarf Stars. This concurs with existing astronomy knowledge; a higher absolute magnitude (measure of luminosity) yields hotter temperatures. This also supports my observations on the B.V variable for its respective histogram. I observed that the giant stars are hotter (negative or close to zero on the B.V scale), whereas dwarf stars are cooler (closer to 2.0 on the B.V scale). The B.V variable rated extremely high on the variable importance plot, and this also concurs with existing astronomy knowledge for star classification.

As for my random forest model, the model performed well. There is only a 12% error for 15821 observations. This is expected since existing astronomy knowledge (i.e. the Hertzsprung-Russell diagram) shows the relationship of a stars' absolute magnitude/luminosity with their star classifications and temperatures. I built a much more simplified model of the Hertzsprung-Russell diagram.

Questions:

1. Should the SpType variable be simplified into general categories rather than specific sub-categories?

Spectral Type is determined by spectra and temperature, so it may be more beneficial and usuable to identify general patterns instead of extremely specific sub-categories of spectral types. This would probably provide additional evidence for my conclusions about the B.V variable.

2. Would adding a color variable be beneficial for star classification?

Color is important for identifying a star's spectral type and temperature. Blue is hot, and red is cool. This would be similar to the B.V variable, but it would describe the color gradient (from Blue-white to red) rather than its numerical value on the B.V scale. This is useful for star classification.

3. Which approach–clustering, knn algorithm, or random forest–would be the best appraoch for classifying specific stars?

The random forest model I built only classifies stars as dwarfs or giants. However, there are more specific stars such as white dwarfs, red supergiants, blue giants, etc. On the Hertzsprung-Russell diagram, you will see that these specific stars are in clusters.