Multivariate statistics: Assignment 1

Team 27: **Raïsa Carmen** *s0204278* **Marco Chi Chung Fong** *r0865521*
Wenting Jiang *r0824739* **Chin Wei Ma** *r0877202*

1 Task 1

All R scripts and the data can be found on this GitHub repository.

The data consists of 18 variables of which the categorical variable *Private* denotes whether the school is private (*Private* == *Yes*) or not (*Private* == *No*). There are 777 schools in the data; 565 (72.72%) private schools and 212 (27.28%) non-private schools.

Before we start to build models to distinguish private and non-private schools, Wilk's lambda test is performed where the null hypothesis states that the centroid (mean) in both groups is the same. Both in the log-transformed and the original (centered or standardized) data, this null hypothesis is rejected. This means that the centroids of private and non-private schools differ significantly.

The test of Box to test for equality of within-group covariance matrices shows that H_0 of equal covariance matrices across groups is not supported by the data. However, we still report on the results from LDA (which assumes equal covariances) since the alternative, QDA, does not always perform better because lower bias of the QDA classifier may not outweigh its higher model complexity.

Table 1 shows all results for the different performance measures (hit rate, sensitivity and specificity) and the different datasets. overall, sensitivity (proportion of positive observations that are being correctly classified) tends to be higher than specificity (proportion of negative observations that are being correctly classified) meaning it's generally easier to detect Private schools than non-private schools.

As expected it doesn't matter whether the data is standardized or centered for the QDA and LDA models and log-transforming skewed variables yields a higher hit rate in LDA and QDA. The hit rate is higher in LDA than in QDA which suggests that the true decision boundary is likely close to linear. The extra complexity of QDA is not worthwhile.

It is curious that KNN works so badly on the centered data with log-transformed skewed variables. The highest overall hit rate, for all datasets, is achieved in the random forest models. Bagging achieves approximately the same performance as LDA although it sacrifices specificity for higher sensitivity.

```r
College_logtranfskewed <- College %>%
  mutate(Apps = log(Apps),
         Accept = log(Accept),
         Enroll = log(Enroll),
         Top10perc = log(Top10perc),
         F.Undergrad = log(F.Undergrad),
         P.Undergrad = log(P.Undergrad),
         Books = log(Books),
         Personal = log(Personal),
         Expend = log(Expend))
College_centered <- College %>%
  dplyr::select(-Private) %>%
  scale(center = TRUE, scale = FALSE) %>%
  as.data.frame() %>%
  mutate(Private = College$Private)
College_logtranfskewed_centered <- College_logtranfskewed %>%
  dplyr::select(-Private) %>%
  scale(center = TRUE, scale = FALSE) %>%
  as.data.frame() %>%
  mutate(Private = College$Private)
College_std <- College %>%
  dplyr::select(-Private) %>%
  scale(center = TRUE, scale = TRUE) %>%
  as.data.frame() %>%
  mutate(Private = College$Private)
College_logtranfskewed_std <- College_logtranfskewed %>%
  dplyr::select(-Private) %>%
  scale(center = TRUE, scale = TRUE) %>%
  as.data.frame() %>%
  mutate(Private = College$Private)
testlist <- list(center = College_centered,
                 center_logtransfskewed = College_logtranfskewed_centered,
                 std = College_std,
                 std_logtransfskewed = College_logtranfskewed_std
                 )

#test the difference between centroids
wilks_results <- testlist %>% map(function(x) Wilks_manova(x, "Private"))
#In each dataset, H0:mu_{yes} = mu_{no} is rejected -->
# centroids of Private and non-private schools differ significantly

#test the difference between centroids
boxm_results <- testlist %>% map(function(x) boxM_test(x, "Private"))

#LDA
lda_results <- map2(testlist, names(testlist),
                    function(x,y){lda_analysis(x,"Private", y)}) %>%
  do.call("rbind", .)

#QDA
qda_results <- map2(testlist, names(testlist),
                    function(x,y){qda_analysis(x,"Private", y)}) %>%
  do.call("rbind", .)

#KNN
knn_output <- map2(testlist, names(testlist),
                    function(x,y){knn_analysis(x,"Private", y, 1:100)})
knn_results <- lapply(knn_output, function(element){
  element[[2]] # The first element contains graphs
})
knn_results <- knn_results %>%
  do.call("rbind", .)

#bagging
bagging_output <- map2(testlist, names(testlist),
                    function(x,y){bagging_analysis(x, "Private", y,
                                              ntree = 2000)})
```

Table 1: A performance comparison of the different methods

	Hit rate				Sensitivity				Specificity			
Method	Centered	Centered Logtransf	Standardized	Standardized Logtransf	Centered	Centered Logtransf	Standardized	Standardized Logtransf	Centered	Centered Logtransf	Standardized	Standardized Logtransf
LDA	0.941	0.942	0.941	0.942	0.952	0.949	0.952	0.949	0.910	0.925	0.910	0.925
QDA	0.909	0.929	0.909	0.929	0.961	0.940	0.961	0.940	0.769	0.901	0.769	0.901
KNN	0.940	0.852	0.932	0.943	0.972	0.901	0.958	0.975	0.854	0.722	0.863	0.858
bagging	0.942	0.941	0.942	0.941	0.977	0.973	0.973	0.973	0.849	0.854	0.858	0.854
randomForest	0.947	0.949	0.947	0.947	0.973	0.975	0.973	0.972	0.877	0.877	0.877	0.882

```r
bagging_results <- lapply(bagging_output, function(element){
  element[[2]] # The first element contains graphs
})
bagging_results <- bagging_results %>%
  do.call("rbind", .)

#random forests
rf_output <- map2(testlist, names(testlist),
                    function(x,y){RF_analysis(x, "Private", y,
                                                          ntree = 2000, mtry = "optimize")})
rf_results <- lapply(rf_output, function(element){
  element[[2]] # The first element contains graphs
})
rf_results <- rf_results %>%
  do.call("rbind", .)
```

For the K-nearest neighbour (KNN) approach, the function knn_cv from the *class* package is used to execute KNN with LOOCV and test values of K between 1 and 100. The model with the highest hit rate is chosen as the final model in Table 1. If there are multiple models with the same hit rate, the smallest K is chosen, Figure 1 shows the results in each of the datasets. The chosen K is 10 for the centered data, 71 for the centered data where the skewed variables are log-transformed, 4 for the standardized data, 11 for the standardized data where the skewed variables are log-transformed.

For the bagging and Random forest (RF) approach, the function randomForest from the *randomForest* package is used. 2000 bootstrapped data and trees are drawn in all models. Figure 2 and 3 show the importance of different variables in the final models. *F.Undergrad* and *outstate* are the most important one in all bagging and RF models. The difference in MeanDecreseGini (the mean decrease in the Gini index) between the most important variables (*F.Undergrad* and *Outstate*) and the other variables is much larger in the bagging model than in the RF models. *Enroll*, which is not at all important in the bagging models, is consistently the third most important variable in the RF models.

For the RF results, we try different values for the mtry parameter which is the number of variables randomly sampled as candidates at each split. The model with the highest hit rate is chosen as the final model in Table 1. If there are multiple models with the same hit rate, the smallest mtry is chosen, Figure 4 shows the results in each of

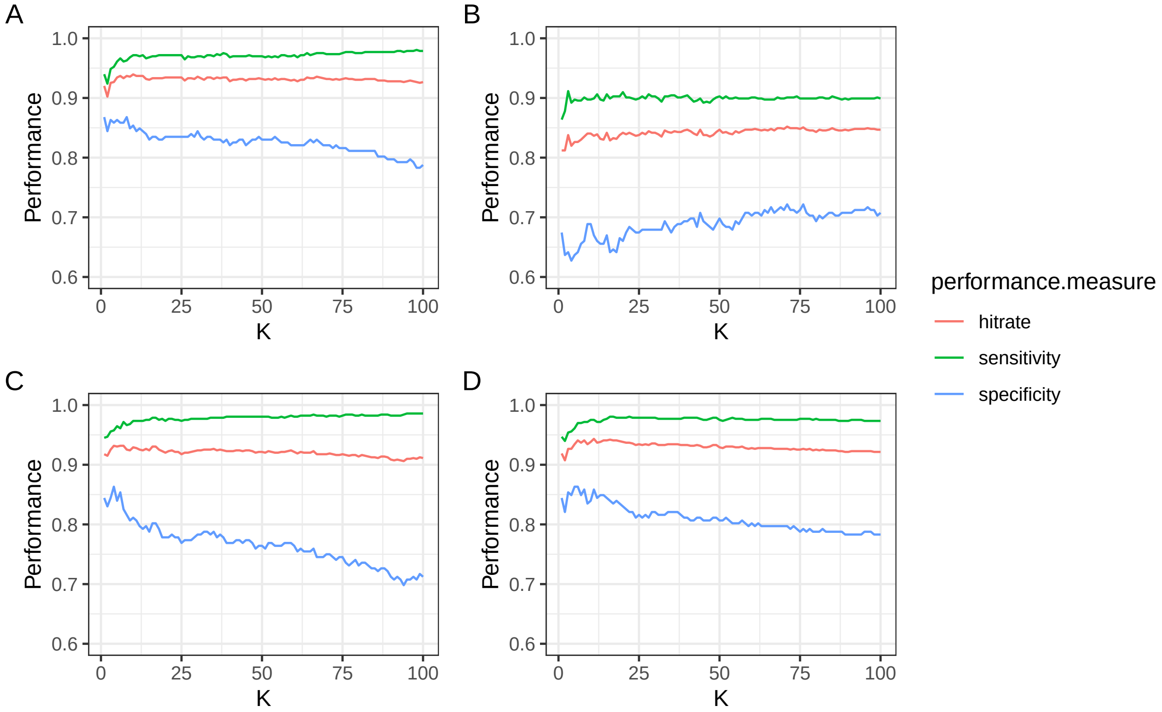

Figure 1: Evolution of hit rate, sensitivity and specificity for different K in the KNN approach. A shows the centered data, B the centered data with log-transformed skewed variables, C shows the standardized data, and D the standardized data with log-transformed skewed variables.

the datasets. The chosen value for mtry is 5 for the centered data (4A), 10 for the centered data where the skewed variables are log-transformed (4B), 8 for the standardized data (4C), 10 for the standardized data where the skewed variables are log-transformed (4D).

2 Task 2

3 Task 3

4 Appendix

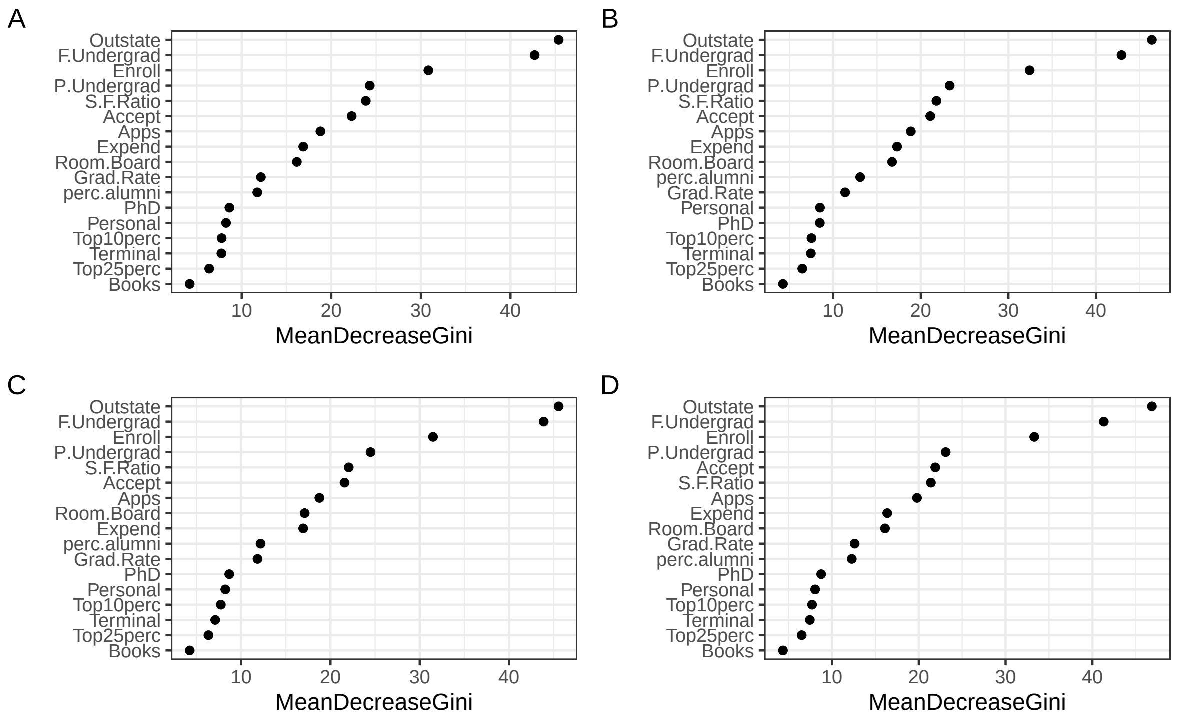

Figure 2: Variance importance plots for the bagging models. A shows the centered data, B the centered data with log-transformed skewed variables, C shows the standardized data, and D the standardized data with log-transformed skewed variables.

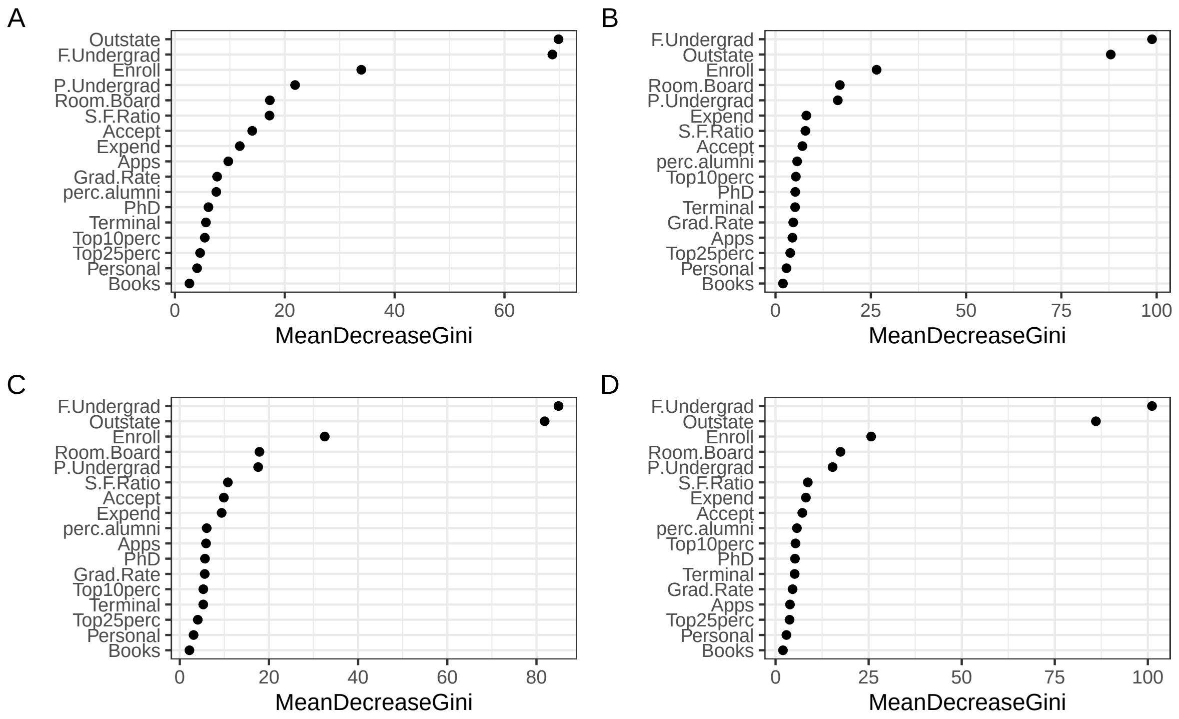

Figure 3: Variance importance plots for the random forest models. A shows the centered data, B the centered data with log-transformed skewed variables, C shows the standardized data, and D the standardized data with log-transformed skewed variables.

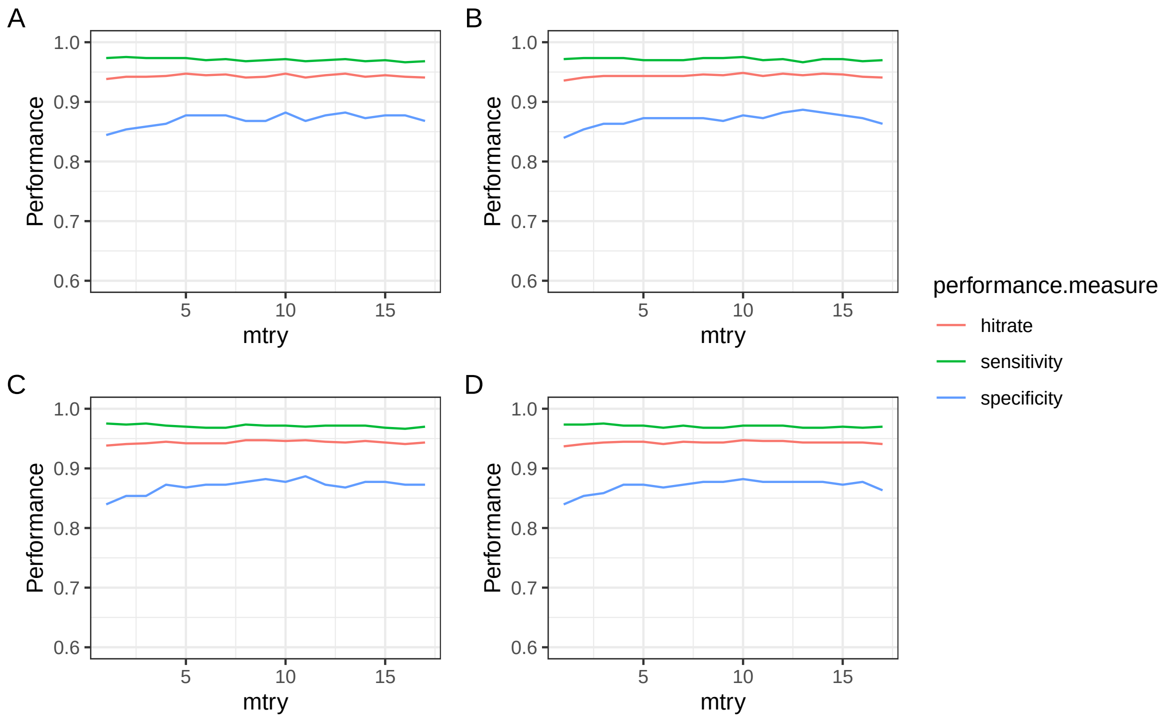

Figure 4: Optimization of the mtry parameter for the random forest models. A shows the centered data, B the centered data with log-transformed skewed variables, C shows the standardized data, and D the standardized data with log-transformed skewed variables.