AssessAI

Ayan Bashir Sheikh

2024-09-06

```r
library(ggplot2)
```

```
## Warning: package ’ggplot2’ was built under R version 4.3.3
```

```r
library(dplyr)
```

```
##
## Attaching package: ’dplyr’
```

```
## The following objects are masked from ’package:stats’:
##
##     filter, lag
```

```
## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```r
library(caret)
```

```
## Warning: package ’caret’ was built under R version 4.3.3
```

```
## Loading required package: lattice
```

```r
library(randomForest)
```

```
## Warning: package ’randomForest’ was built under R version 4.3.3
```

```
## randomForest 4.7-1.1
```

```
## 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
```

```
library(cluster)
library(e1071)
```

```
## Warning: package ’e1071’ was built under R version 4.3.3
```

```
library(tidyr)

data =read.csv("C:\\Users\\ayans\\Downloads\\exams.csv")

head(data)
```

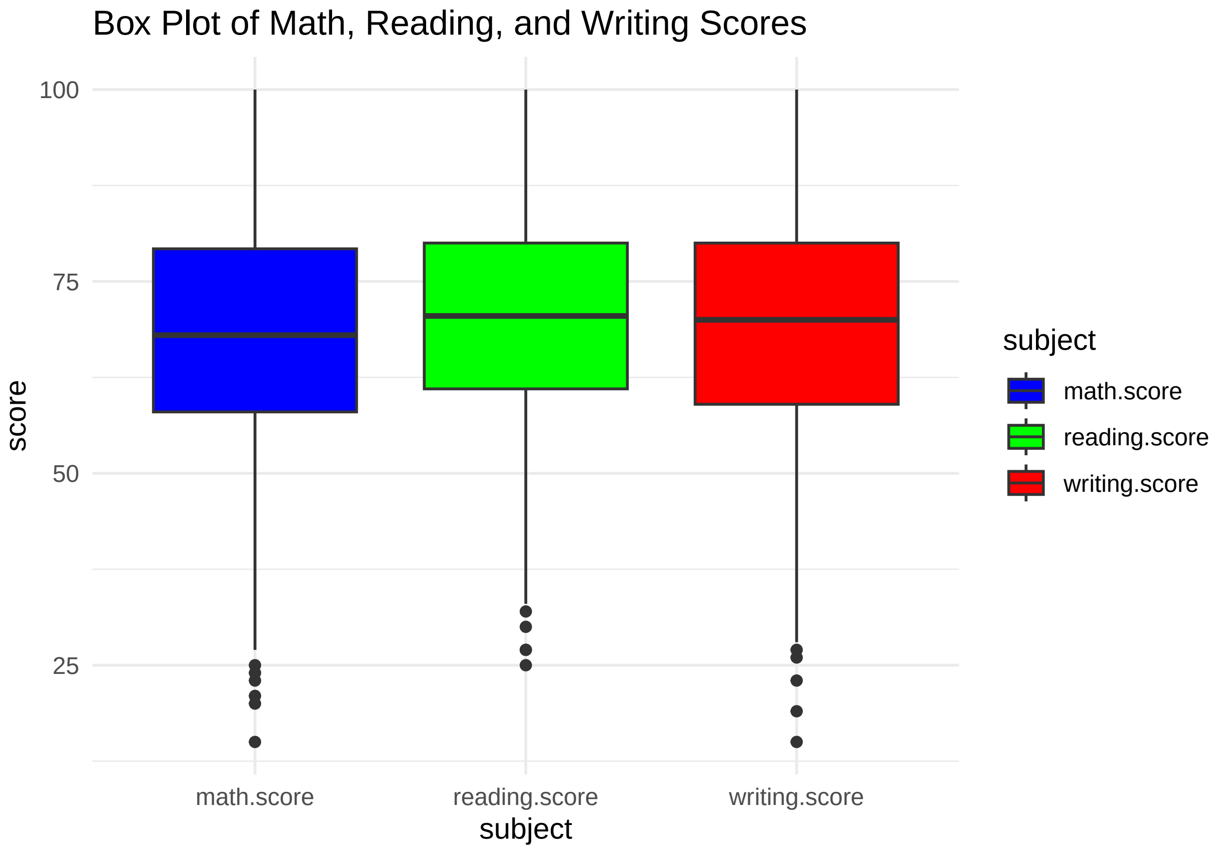
```
##   gender race.ethnicity parental.level.of.education        lunch
## 1 female        group D                some college     standard
## 2   male        group D          associate’s degree     standard
## 3 female        group D                some college free/reduced
## 4   male        group B                some college free/reduced
## 5 female        group D          associate’s degree     standard
## 6   male        group C            some high school     standard
##   test.preparation.course math.score reading.score writing.score
## 1               completed         59            70            78
## 2                    none         96            93            87
## 3                    none         57            76            77
## 4                    none         70            70            63
## 5                    none         83            85            86
## 6                    none         68            57            54
```

Exploratory Data Analysis

The box plot provides a visual summary of the distribution of scores for three subjects: Math, Reading, and Writing. Each box plot displays the median, quartiles, and potential outliers for the respective scores.

```
 scores=data[,c(6,7,8)]
scores_long = scores %>%
  pivot_longer(cols = everything(), names_to = "subject", values_to = "score")


ggplot(scores_long, aes(x = subject, y = score, fill = subject)) +
  geom_boxplot() +
  ggtitle("Box Plot of Math, Reading, and Writing Scores") +
  theme_minimal() +
  scale_fill_manual(values = c("blue", "green", "red"))
```



The box plot reveals that students tend to perform better in reading and writing compared to math. The presence of outliers in math and writing scores suggests that targeted interventions may be needed to support students who are struggling in these areas. The consistent performance in reading, with no significant outliers, indicates a more uniform level of achievement in this subject.

Correlation Analysis

The correlation matrix provides insights into the linear relationships between the three variables. The values in the matrix range from -1 to 1, where:

1) 1 indicates a perfect positive linear relationship. 2) -1 indicates a perfect negative linear relationship. 3) 0 indicates no linear relationship.

The correlation coefficients are all above 0.79, indicating strong to very strong positive relationships between the variables. These high values suggest that the variables are closely related and that improvements in one area are likely to be associated with improvements in the others.

The strong correlations between reading, writing, and math scores highlight the importance of a well-rounded educational approach. Enhancing reading and writing skills can have a positive impact on math performance, and vice versa.

```
cor(data[, c("math.score", "reading.score", "writing.score")])
```

```
##               math.score reading.score writing.score
## math.score     1.0000000     0.8117671     0.7900549
## reading.score  0.8117671     1.0000000     0.9489088
## writing.score  0.7900549     0.9489088     1.0000000
```

The correlation matrix reveals strong positive relationships between math, reading, and writing scores. These findings underscore the interconnectedness of these academic skills and suggest that interventions aimed at improving one area may have beneficial effects on the others. This information is crucial for developing comprehensive educational strategies that support overall student achievement.

Regression Analysis

The linear regression model was used to predict students' math scores based on various factors, including reading and writing scores, gender, race/ethnicity, parental education level, lunch status, and test preparation course.

```r
model = lm(math.score ~ reading.score + writing.score + gender + race.ethnicity + parental.level.of.edu
summary(model)
```

```
## 
## Call:
## lm(formula = math.score ~ reading.score + writing.score + gender + 
##     race.ethnicity + parental.level.of.education + lunch + test.preparation.course, 
##     data = data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -18.4719  -3.6811  -0.0503   3.6779  16.4876 
## 
## Coefficients:
##                                               Estimate Std. Error t value
## (Intercept)                                  -10.41977    1.30051  -8.012
## reading.score                                  0.25834    0.04220   6.121
## writing.score                                  0.68467    0.04243  16.137
## gendermale                                    13.16949    0.37359  35.251
## race.ethnicitygroup B                          0.22610    0.72468   0.312
## race.ethnicitygroup C                          0.07152    0.68338   0.105
## race.ethnicitygroup D                          0.01258    0.70325   0.018
## race.ethnicitygroup E                          4.58241    0.77048   5.947
## parental.level.of.educationbachelor's degree  -0.40038    0.65392  -0.612
## parental.level.of.educationhigh school        -0.04344    0.53929  -0.081
## parental.level.of.educationmaster's degree    -0.24105    0.73352  -0.329
## parental.level.of.educationsome college       -0.09054    0.52741  -0.172
## parental.level.of.educationsome high school   -0.72925    0.56633  -1.288
## lunchstandard                                  4.42927    0.38294  11.566
## test.preparation.coursenone                    3.94626    0.41022   9.620
##                                              Pr(>|t|)    
## (Intercept)                                  3.17e-15 ***
## reading.score                                1.34e-09 ***
## writing.score                                 < 2e-16 ***
## gendermale                                    < 2e-16 ***
## race.ethnicitygroup B                           0.755    
## race.ethnicitygroup C                           0.917    
## race.ethnicitygroup D                           0.986    
## race.ethnicitygroup E                        3.78e-09 ***
## parental.level.of.educationbachelor's degree    0.540    
## parental.level.of.educationhigh school          0.936    
## parental.level.of.educationmaster’s degree      0.743
## parental.level.of.educationsome college         0.864
## parental.level.of.educationsome high school     0.198
## lunchstandard                                 < 2e-16 ***
## test.preparation.coursenone                   < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 5.414 on 985 degrees of freedom
## Multiple R-squared:  0.8757, Adjusted R-squared:  0.874
## F-statistic: 495.8 on 14 and 985 DF,  p-value: < 2.2e-16
```

1)Reading Score Each additional point in reading score increases the math score by 0.258 points. 2) Writing Score Each additional point in writing score increases the math score by 0.685 points. 3)Gender (Male) Male students score 13.169 points higher in math compared to female students. 4)Race/Ethnicity (Group E): Students in group E score 4.582 points higher in math. 5)Lunch (Standard) Students with standard lunch score 4.429 points higher in math. 6)Test Preparation Course (None) Students who did not take a test preparation course score 3.946 points higher in math.

Model Fit

R-squared 87.57% of the variability in math scores is explained by the model. Adjusted R-squared 87.4%, indicating a strong fit even after adjusting for the number of predictors. F-statistic: The model is statistically significant (p < 2.2e-16).

The model highlights that reading and writing scores, gender, race/ethnicity, lunch status, and test preparation course are significant predictors of math scores. These findings suggest that improving reading and writing skills, along with considering gender and lunch status, can positively impact math performance. The high R-squared value indicates that the model explains a substantial portion of the variability in math scores, making it a reliable tool for predicting student performance.

```r
library(randomForest)
library(ggplot2)
library(dplyr)


data =read.csv("C:\\Users\\ayans\\Downloads\\exams.csv")
set.seed(123)

trainIndex =sample(1:1000,size=800,replace=T)
dataTrain = data[ trainIndex,]
dataTest  = data[-trainIndex,]

rf_model = randomForest(math.score ~., data = dataTrain)
predictions = predict(rf_model, dataTest)

# Calculate Mean Squared Error (MSE)
mse = mean((dataTest$math.score - predictions)^2)
mse
```

```
## [1] 37.86789
```

```r
var(data$math.score)
```

```
## [1] 232.5685
```

The MSE (37.86789) is significantly lower than the variance (232.5685). This indicates that your Random Forest model is capturing a substantial amount of the variability in the math scores, which is a positive sign. lower MSE compared to the variance suggests that the model's predictions are relatively accurate. The model is effectively reducing the error compared to simply using the mean of the scores as a predictor.

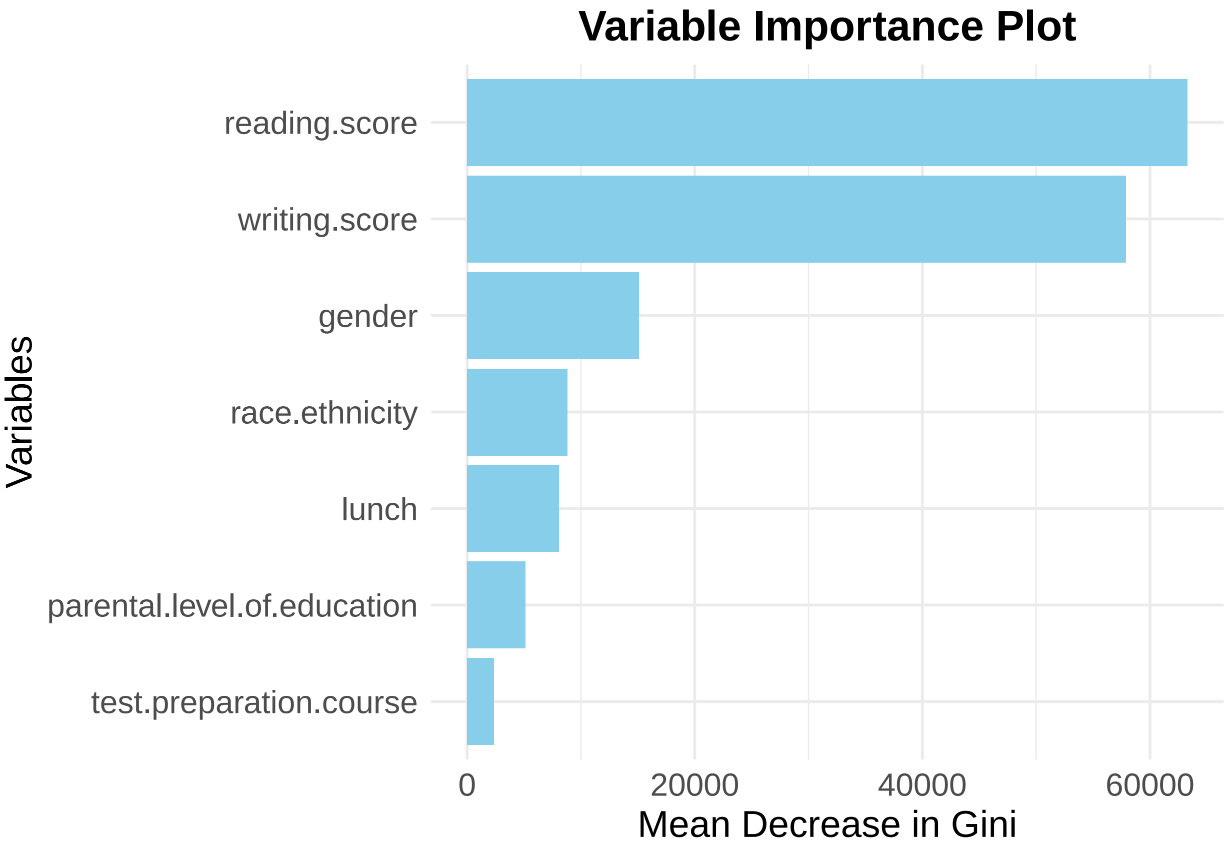
Variable Importance plot

The x-axis represents the "Mean Decrease in Gini," which measures how each variable contributes to the homogeneity of the nodes and leaves in the Random Forest model.A higher value indicates that the variable is more important for making accurate predictions.

```r
rf_model = randomForest(math.score ~ ., data = dataTrain, importance = TRUE)
importance_df = as.data.frame(importance(rf_model))
importance_df$Variable = rownames(importance_df)


colnames(importance_df) = c("MeanDecreaseAccuracy", "MeanDecreaseGini", "Variable")

ggplot(importance_df, aes(x = reorder(Variable, MeanDecreaseGini), y = MeanDecreaseGini)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  coord_flip() +
  ggtitle("Variable Importance Plot") +
  xlab("Variables") +
  ylab("Mean Decrease in Gini") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    axis.title.x = element_text(size = 14),
    axis.title.y = element_text(size = 14),
    axis.text = element_text(size = 12)
  )
```



1)Writing and Reading Scores Significant predictors of math scores. Improving these areas can positively impact math performance.

2)Feature Selection Top Predictors: Focus on writing and reading scores in future models to maintain predictive power while simplifying the model.

3)Resource Allocation Targeted Interventions: Allocate resources to improve writing and reading skills, and consider factors like gender, lunch status, and race/ethnicity to maximize student performance.

It provides a clear roadmap for improving student performance in math by focusing on key predictors, simplifying future models, and strategically allocating resources. These findings can inform evidence-based educational strategies and policies aimed at fostering academic success across various subjects.

```r
data$gender = as.factor(data$gender)
data$race.ethnicity = as.factor(data$race.ethnicity)
data$parental.level.of.education= as.factor(data$parental.level.of.education)
data$lunch =as.factor(data$lunch)
data$test.preparation.course = as.factor(data$test.preparation.course)
```

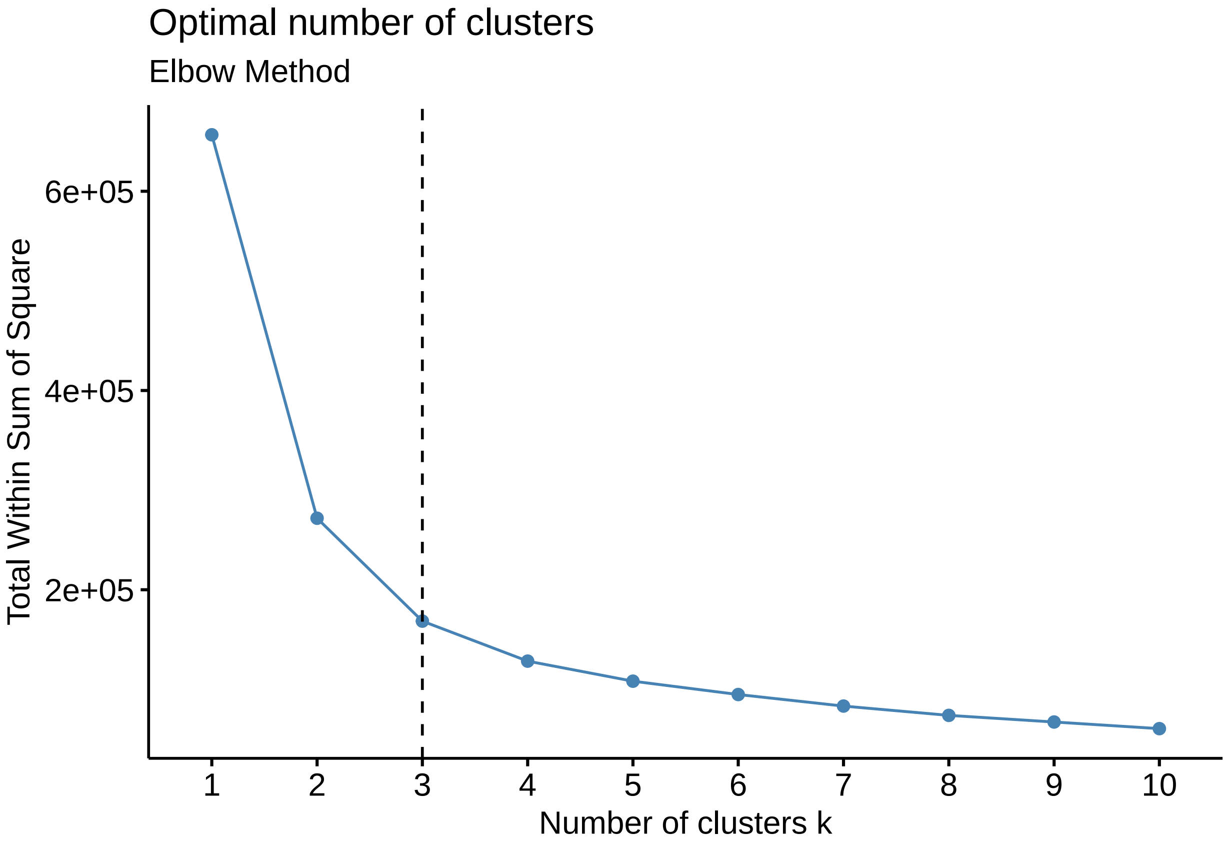
Clustering using K-Means

```r
library(factoextra)
```

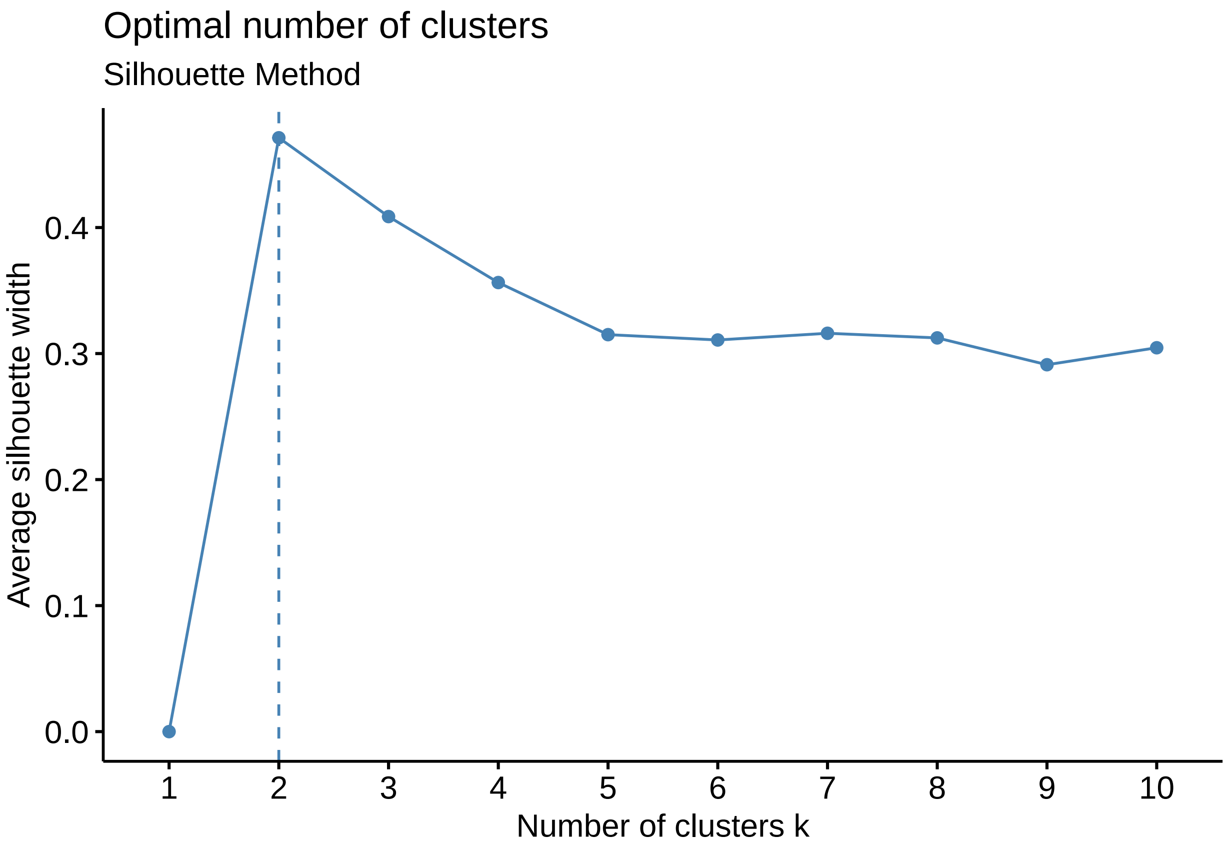
```
## Warning: package 'factoextra' was built under R version 4.3.3
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(cluster)


#Elbow Method
elbow_plot = fviz_nbclust(scores, kmeans, method = "wss") +
  geom_vline(xintercept = 3, linetype = 2) +
  labs(subtitle = "Elbow Method");elbow_plot
```



```
# Silhouette Statistics
silhouette_plot = fviz_nbclust(scores, kmeans, method = "silhouette") +
  labs(subtitle = "Silhouette Method");silhouette_plot
```

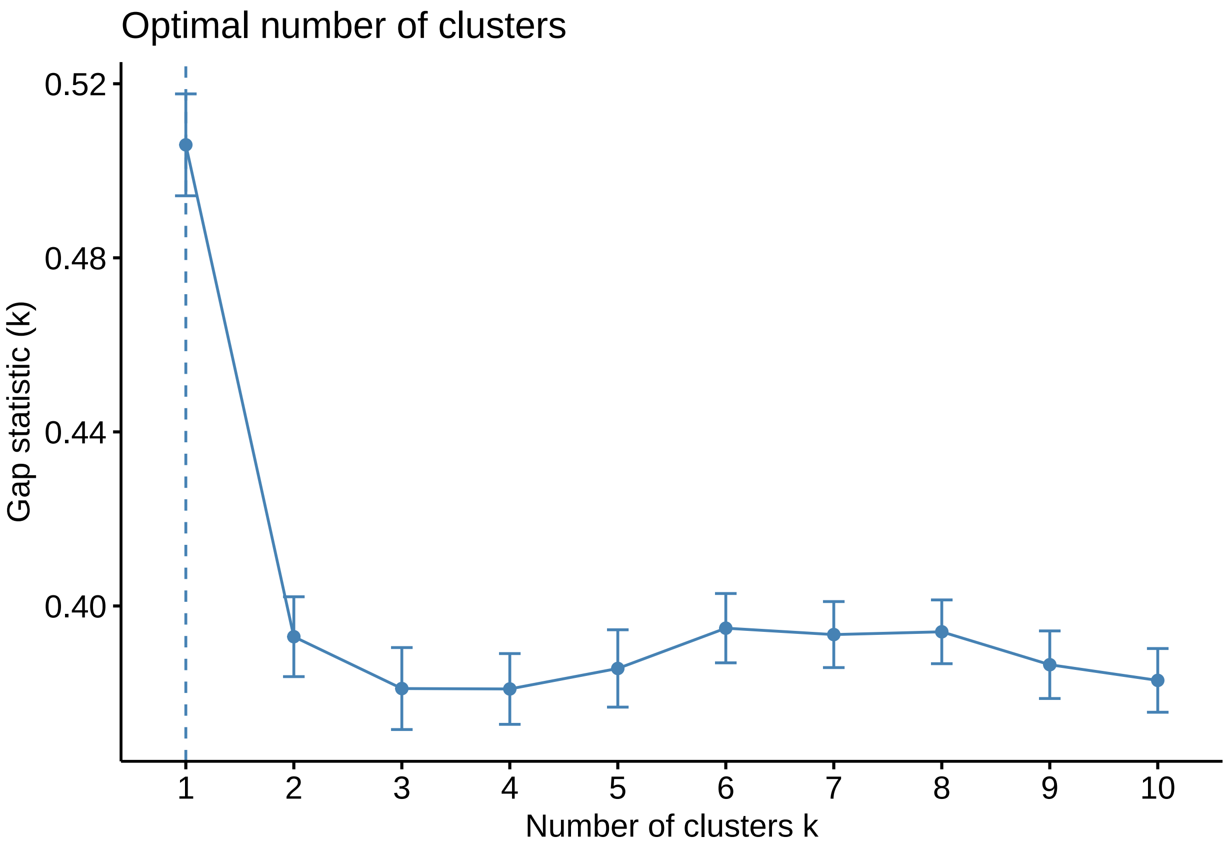


```
gap_stat = clusGap(scores, FUN = kmeans, nstart = 25, K.max = 10, B = 50)
```

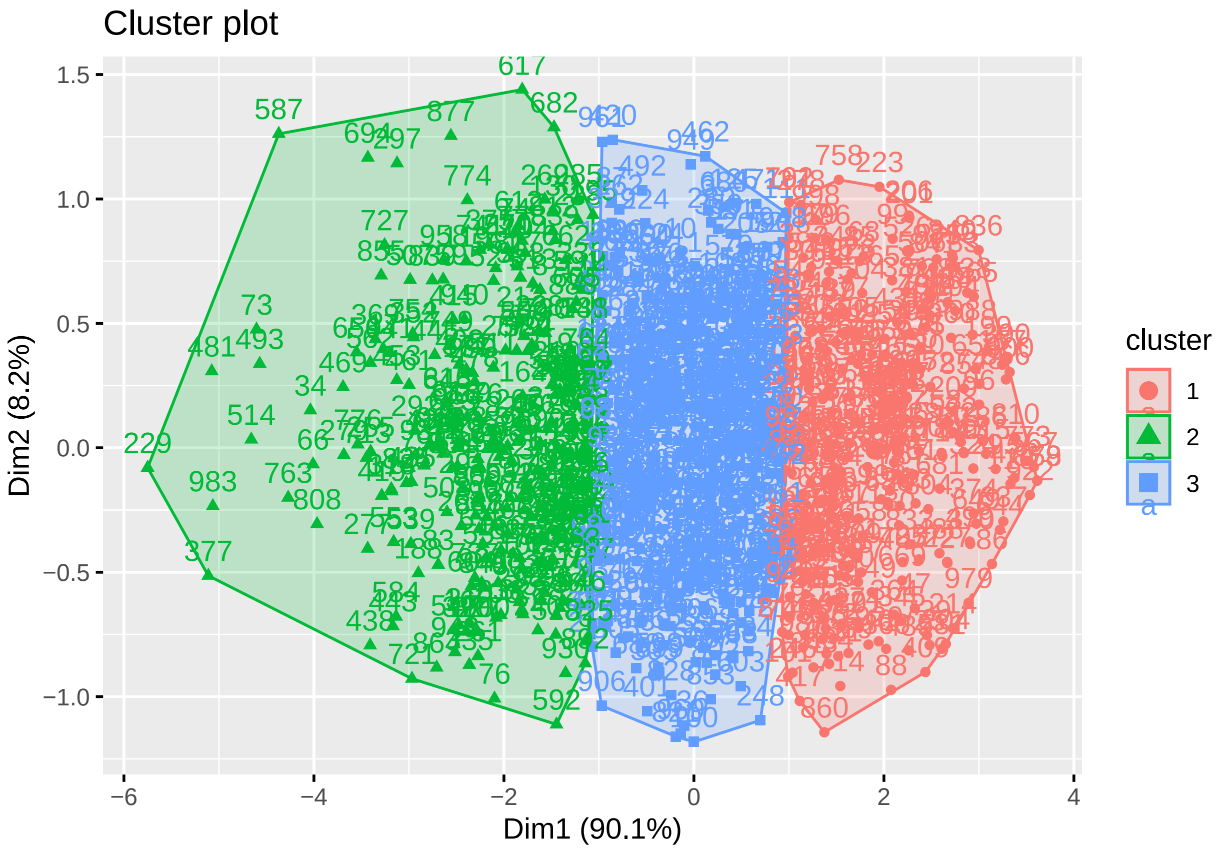
```
## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations
```

```
gap_stat_plot = fviz_gap_stat(gap_stat);gap_stat_plot
```



```
kmeans_result = kmeans(scores, centers = 3, nstart = 25)
fviz_cluster(kmeans_result, data = scores)
```



Grouping Students:The K-Means clustering algorithm has grouped students into three clusters based on their scores or learning characteristics. This allows for the creation of virtual classrooms where students with similar levels of understanding are grouped together.

Personalized Learning Paths: Each cluster can have customized learning paths, resources, and activities designed to address the specific needs of the students within that group. For example, Cluster 1 (red) may consist of students who need additional support, while Cluster 3 (blue) may include students who require more advanced materials.

Scalability and Integration: This clustering approach is scalable and can be integrated with other educational technologies like VR and AR to create immersive learning environments. Students in the same cluster could participate in a VR-based collaborative project that matches their skill level, enhancing their learning experience.

The application of K-Means clustering provides a robust framework for grouping students based on their learning characteristics. This approach facilitates the creation of virtual classrooms tailored to the specific needs of each group, thereby enhancing the overall educational experience.

SVM

AssessAI is an advanced educational platform that leverages AI and machine learning to enhance the learning experience by offering tailored recommendations, immersive learning environments, and comprehensive progress tracking. The SVM model's performance in categorizing students based on their characteristics and predicting outcomes is crucial for the platform's personalized learning path recommendations.

```r
library(e1071)
library(caret)
```

```
set.seed(123)

data$gender = as.factor(data$gender)
data$race.ethnicity = as.factor(data$race.ethnicity)
data$parental.level.of.education = as.factor(data$parental.level.of.education)
data$lunch = as.factor(data$lunch)
data$test.preparation.course = as.factor(data$test.preparation.course)

data$math.score.cat = cut(data$math.score, breaks = c(-Inf, 60, 80, Inf), labels = c("low", "medium", "



trainIndex = createDataPartition(data$math.score.cat, p = 0.8, list = FALSE, times = 1)
dataTrain = data[trainIndex,]
dataTest = data[-trainIndex,]


svm_model = svm(math.score.cat ~ ., data = dataTrain, kernel = "linear")


svm_predictions = predict(svm_model, dataTest)


confusion_matrix = table(dataTest$math.score.cat, svm_predictions);confusion_matrix
```

```
##         svm_predictions
##          low medium high
##   low     59      1    0
##   medium   0     93    0
##   high     0      0   45
```

```
accuracy = sum(diag(confusion_matrix)) / sum(confusion_matrix)
accuracy
```

```
## [1] 0.9949495
```

The overall accuracy of the model is 0.9949495, which means the model correctly classified approximately 99.5% of the instances.

1)True Positives (TP): The diagonal elements (59, 93, 45) represent the correctly classified instances for each class.

2)False Positives (FP): The off-diagonal elements in each column (0, 1, 0) represent the instances incorrectly classified as that class.

3)False Negatives (FN): The off-diagonal elements in each row (1, 0, 0) represent the instances that belong to that class but were incorrectly classified as another class.

#Relevance to AssessAI

Personalized Learning Path Recommendations

The high accuracy and precision of the SVM model ensure that the personalized learning paths recommended by AssessAI are reliable and tailored to the individual needs of each student.

Skill and Expertise Suggestions

Accurate classification of students' performance allows AssessAI to provide precise skill and expertise suggestions, helping learners identify and acquire necessary skills to achieve their academic and professional goals.

Progress Tracking and Feedback

The model's ability to correctly classify students' performance enables effective progress tracking and real-time feedback, allowing for adjustments to learning paths as needed.

Enhanced Learning Features

The robust performance of the SVM model supports the platform's enhanced learning features, such as VR-based geographical explanations and 3D mathematical geometry, by ensuring that learners are accurately categorized and receive appropriate content.

The SVM model demonstrates excellent performance with high accuracy and F1-scores across all classes. The confusion matrix shows that the model has a very low misclassification rate, indicating its robustness in categorizing students based on their characteristics.

Collaborative Filtering

In AssessAI, Collaborative Filtering can be used to recommend topics and courses to learners based on their past interactions and the preferences of similar learners. This personalized approach enhances the learning experience by suggesting content that is most relevant and engaging for each learner.

Advantages of Collaborative Filtering in AssessAI

Personalization: Provides highly personalized recommendations based on individual learner behavior.

Scalability: Can handle large datasets effectively, making it suitable for platforms with many learners and courses.

Adaptability: Can adapt to changing learner preferences over time, ensuring that recommendations remain relevant.

Collaborative Filtering is a powerful technique for recommending topics and courses in AssessAI. By leveraging the preferences of similar learners, it provides personalized and relevant recommendations, enhancing the overall learning experience.