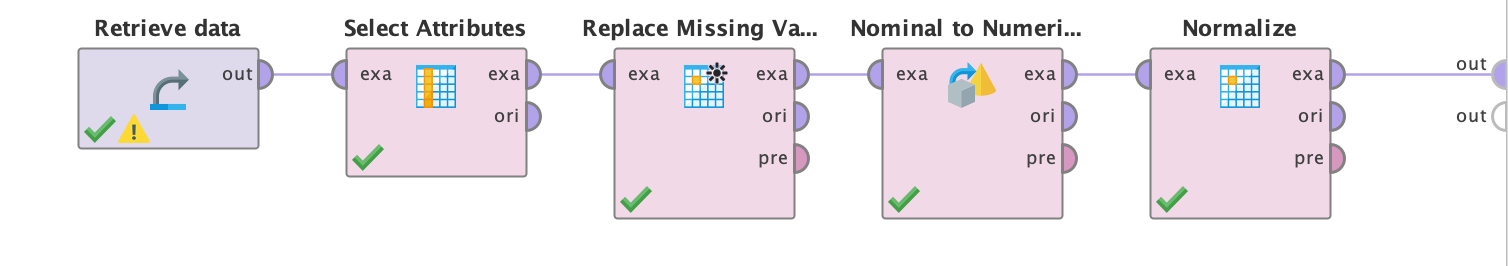
**Machine Learning Assignment**

**Process:**

Diagram

Description automatically generated

**Data Cleaning Subset:**



**Random Forest: (K=3)**

Diagram

Description automatically generated

Table

Description automatically generated

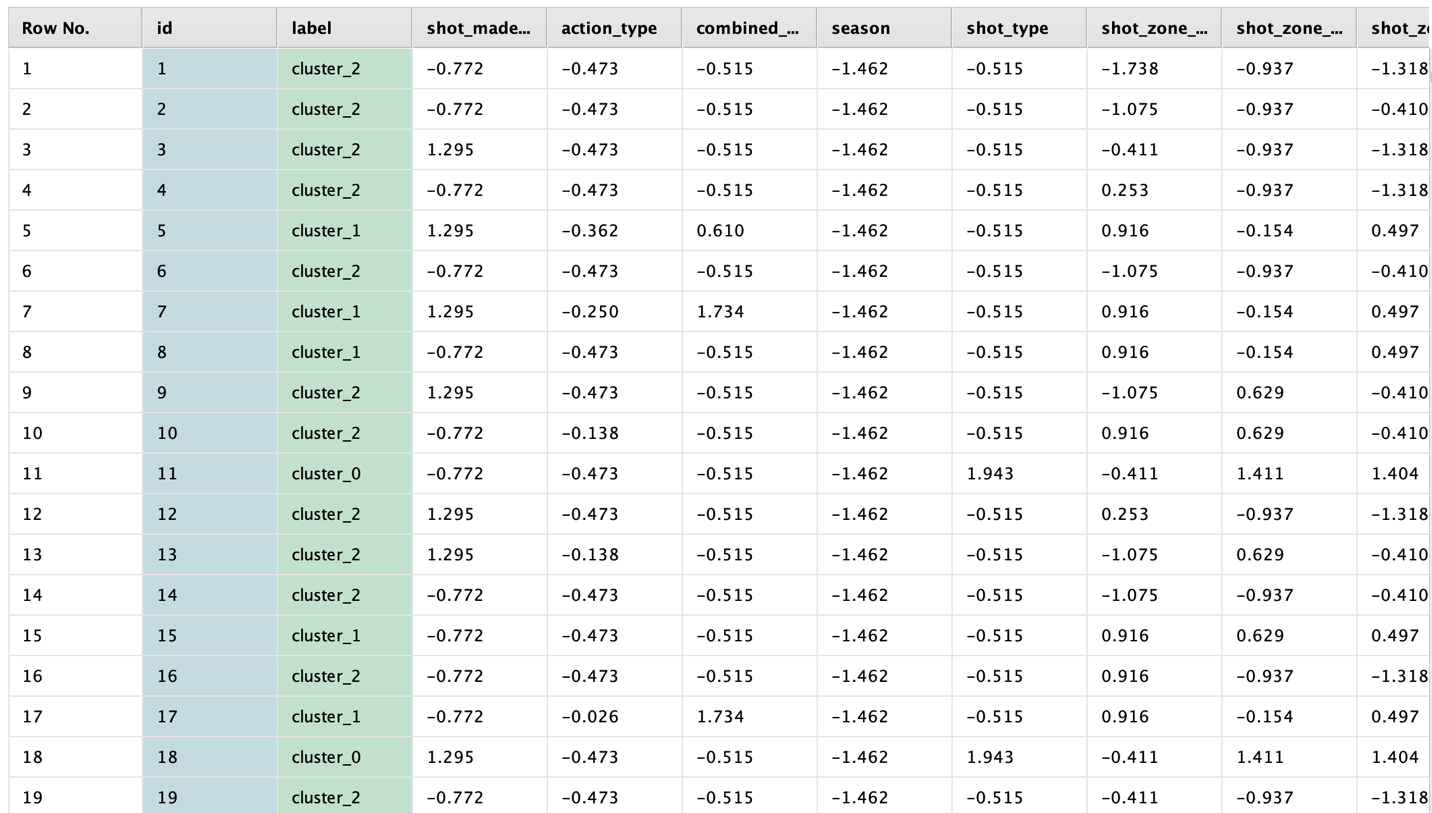
Table

Description automatically generated

**Decision tree: (K=3)**

Diagram

Description automatically generatedTable

Description automatically generated 

**Decision Tree: (K = 2)**

Diagram

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

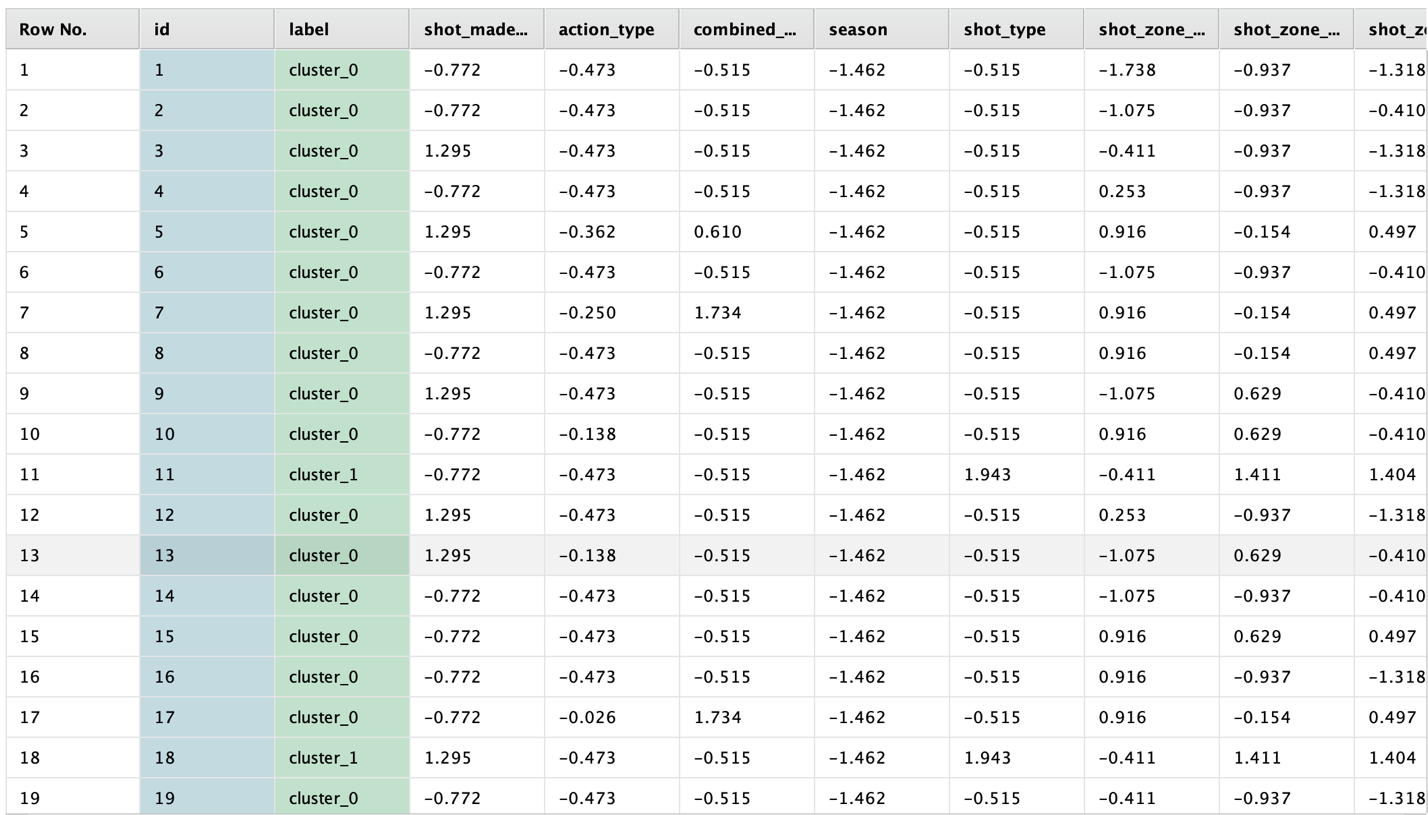
**Random Forest: (K=2)**

Diagram

Description automatically generated

Table

Description automatically generated



**REPORT:**

**Problem statement:**

The goal of the project is to evaluate Kobe Bryant's shots from his NBA career and provide predictions about whether a shot will be made or missed based on variables including position, timing, and game circumstances. The dataset, which has 25,698 observations and 25 variables, was obtained via a Kaggle competition.

**Data preprocessing:**

**Select attributes:** To remove unnecessary columns: We may exclude several columns from the dataset because they don't give any helpful data for clustering, such as game\_id, team\_id, and team\_name.

**Replacing** **missing values**: In the shot\_made\_flag column of the dataset, there are 5000 missing values. The shot\_made\_flag column's missing values can be replaced with average.

Converting nominal to numerical: By using **Nominal to Numerical** operator, we must convert categorical columns like action\_type, combined\_shot\_type, opponent, season, shot\_type, shot\_zone\_area, shot\_zone\_basic, and shot\_zone\_range to numerical format before we can do clustering.

Data scaling: Because the dataset's characteristics have varied scales, we may scale the data so that it has a zero mean and a single variance. To do this, we may use **Normalize operator.**

**Modeling:**

**Clustering:** Using the k-means clustering technique, we may group similar shots together. To do this, we may use the **k-Means clustering operator**. We may change the cluster size from 2 to 5, then assess how well the clustering method performs.

**Decision tree and random forest** classification: To forecast whether a shot will be successful or unsuccessful, we also used decision tree and random forest classifiers. The full dataset may be used to train the models, and cross-validation can be used to assess how well they perform.

**Analysis:**

After training decision tree and random forest classifiers on the entire dataset, we found that for K-Means Clustering with K=2, the results were very good, with an accuracy of 99.98% +/- 0.03% for Decision Tree and 99.98% +/- 0.02% for Random Forest. However, for K=3, the results were slightly lower, with an accuracy of 99.48% +/- 0.23% for Decision Tree and 99.32% +/- 0.20% for Random Forest. So, the accuracy was highest (99.98% +/- 0.03%) when the number of clusters was set to 2 and the random forest algorithm was used. This suggests that the clustering algorithm successfully distinguished between the two clusters of data, which correspond to successful and unsuccessful shots.

The confusion matrix also demonstrates that, with just a few exceptions, most data points were accurately categorized. This suggests that the clustering algorithm was able to effectively group similar data points together based on their attributes, resulting in reliable clustering results.

Therefore, the Kobe Bryant shot selection dataset's clustering results are reliable.