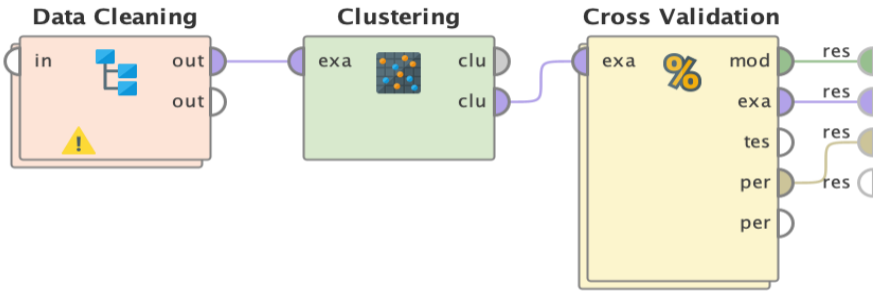
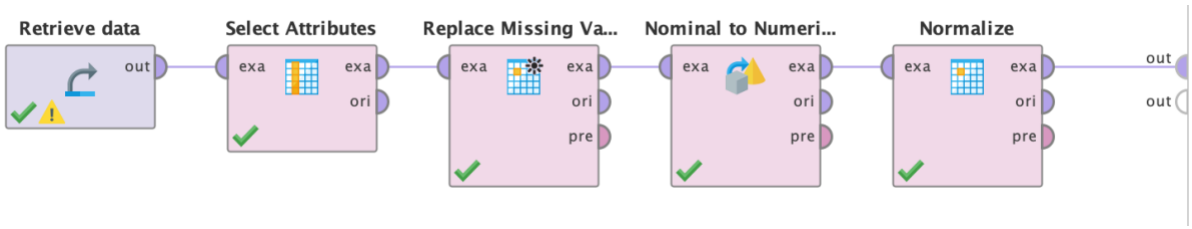
Kobe Bryant's Shots Prediction

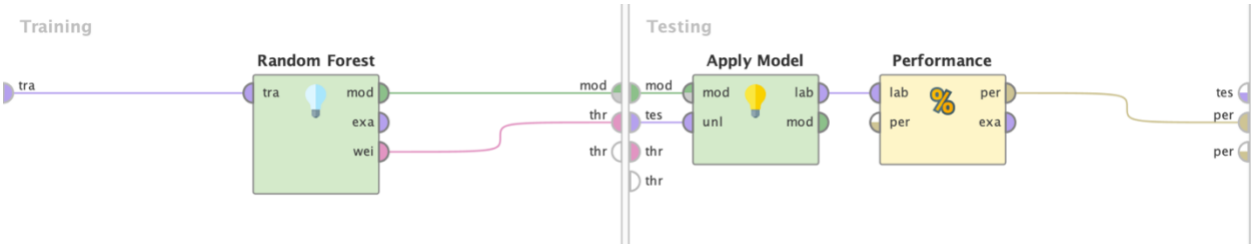
Process:



Data Cleaning Subset:



Random Forest: (K=3)

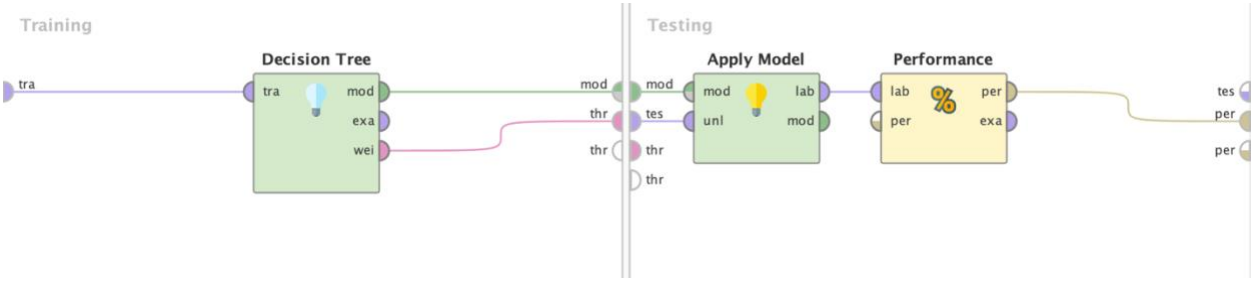


accuracy: 99.32% +/- 0.20% (micro average: 99.32%)

	true cluster_2	true cluster_1	true cluster_0	class precision
pred. cluster_2	14445	209	0	98.57%
pred. cluster_1	0	9685	0	100.00%
pred. cluster_0	0	0	6358	100.00%
class recall	100.00%	97.89%	100.00%	

Row No.	id	label	shot_made...	action_type	combined_...	season	shot_type	shot_zone_...	shot_zone_...	shot_z
1	1	cluster_2	-0.772	-0.473	-0.515	-1.462	-0.515	-1.738	-0.937	-1.318
2	2	cluster_2	-0.772	-0.473	-0.515	-1.462	-0.515	-1.075	-0.937	-0.410
3	3	cluster_2	1.295	-0.473	-0.515	-1.462	-0.515	-0.411	-0.937	-1.318
4	4	cluster_2	-0.772	-0.473	-0.515	-1.462	-0.515	0.253	-0.937	-1.318
5	5	cluster_1	1.295	-0.362	0.610	-1.462	-0.515	0.916	-0.154	0.497
6	6	cluster_2	-0.772	-0.473	-0.515	-1.462	-0.515	-1.075	-0.937	-0.410
7	7	cluster_1	1.295	-0.250	1.734	-1.462	-0.515	0.916	-0.154	0.497
8	8	cluster_1	-0.772	-0.473	-0.515	-1.462	-0.515	0.916	-0.154	0.497
9	9	cluster_2	1.295	-0.473	-0.515	-1.462	-0.515	-1.075	0.629	-0.410
10	10	cluster_2	-0.772	-0.138	-0.515	-1.462	-0.515	0.916	0.629	-0.410
11	11	cluster_0	-0.772	-0.473	-0.515	-1.462	1.943	-0.411	1.411	1.404
12	12	cluster_2	1.295	-0.473	-0.515	-1.462	-0.515	0.253	-0.937	-1.318
13	13	cluster_2	1.295	-0.138	-0.515	-1.462	-0.515	-1.075	0.629	-0.410
14	14	cluster_2	-0.772	-0.473	-0.515	-1.462	-0.515	-1.075	-0.937	-0.410
15	15	cluster_1	-0.772	-0.473	-0.515	-1.462	-0.515	0.916	0.629	0.497
16	16	cluster_2	-0.772	-0.473	-0.515	-1.462	-0.515	0.916	-0.937	-1.318
17	17	cluster_1	-0.772	-0.026	1.734	-1.462	-0.515	0.916	-0.154	0.497
18	18	cluster_0	1.295	-0.473	-0.515	-1.462	1.943	-0.411	1.411	1.404
19	19	cluster_2	-0.772	-0.473	-0.515	-1.462	-0.515	-0.411	-0.937	-1.318

Decision tree: (K=3)

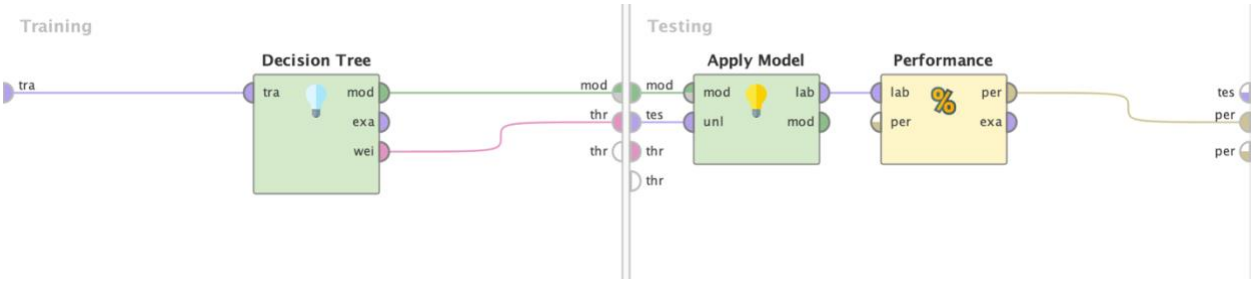


accuracy: 99.48% +/- 0.23% (micro average: 99.48%)

	true cluster_2	true cluster_1	true cluster_0	class precision
pred. cluster_2	14413	128	0	99.12%
pred. cluster_1	32	9766	0	99.67%
pred. cluster_0	0	0	6358	100.00%
class recall	99.78%	98.71%	100.00%	

Row No.	id	label	shot_made...	action_type	combined_...	season	shot_type	shot_zone_...	shot_zone_...	shot_z
1	1	cluster_2	-0.772	-0.473	-0.515	-1.462	-0.515	-1.738	-0.937	-1.318
2	2	cluster_2	-0.772	-0.473	-0.515	-1.462	-0.515	-1.075	-0.937	-0.410
3	3	cluster_2	1.295	-0.473	-0.515	-1.462	-0.515	-0.411	-0.937	-1.318
4	4	cluster_2	-0.772	-0.473	-0.515	-1.462	-0.515	0.253	-0.937	-1.318
5	5	cluster_1	1.295	-0.362	0.610	-1.462	-0.515	0.916	-0.154	0.497
6	6	cluster_2	-0.772	-0.473	-0.515	-1.462	-0.515	-1.075	-0.937	-0.410
7	7	cluster_1	1.295	-0.250	1.734	-1.462	-0.515	0.916	-0.154	0.497
8	8	cluster_1	-0.772	-0.473	-0.515	-1.462	-0.515	0.916	-0.154	0.497
9	9	cluster_2	1.295	-0.473	-0.515	-1.462	-0.515	-1.075	0.629	-0.410
10	10	cluster_2	-0.772	-0.138	-0.515	-1.462	-0.515	0.916	0.629	-0.410
11	11	cluster_0	-0.772	-0.473	-0.515	-1.462	1.943	-0.411	1.411	1.404
12	12	cluster_2	1.295	-0.473	-0.515	-1.462	-0.515	0.253	-0.937	-1.318
13	13	cluster_2	1.295	-0.138	-0.515	-1.462	-0.515	-1.075	0.629	-0.410
14	14	cluster_2	-0.772	-0.473	-0.515	-1.462	-0.515	-1.075	-0.937	-0.410
15	15	cluster_1	-0.772	-0.473	-0.515	-1.462	-0.515	0.916	0.629	0.497
16	16	cluster_2	-0.772	-0.473	-0.515	-1.462	-0.515	0.916	-0.937	-1.318
17	17	cluster_1	-0.772	-0.026	1.734	-1.462	-0.515	0.916	-0.154	0.497
18	18	cluster_0	1.295	-0.473	-0.515	-1.462	1.943	-0.411	1.411	1.404
19	19	cluster_2	-0.772	-0.473	-0.515	-1.462	-0.515	-0.411	-0.937	-1.318

Decision Tree: (K = 2)

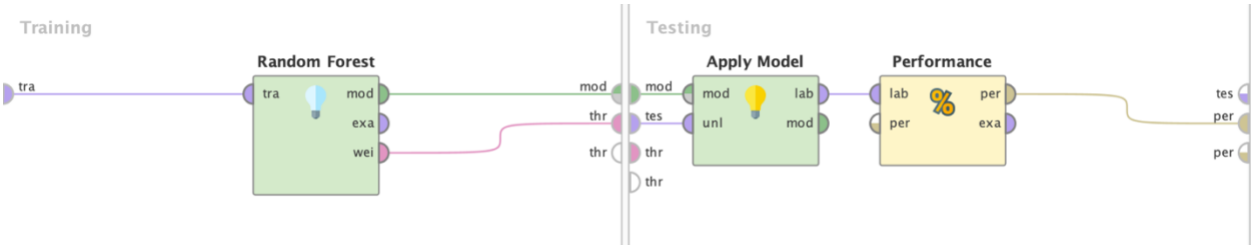


accuracy: 99.98% +/- 0.03% (micro average: 99.98%)

	true cluster_0	true cluster_1	class precision
pred. cluster_0	24329	1	100.00%
pred. cluster_1	4	6363	99.94%
class recall	99.98%	99.98%	

Row No.	id	label	shot_made...	action_type	combined_...	season	shot_type	shot_zone_...	shot_zone_...	shot_z
1	1	cluster_0	-0.772	-0.473	-0.515	-1.462	-0.515	-1.738	-0.937	-1.318
2	2	cluster_0	-0.772	-0.473	-0.515	-1.462	-0.515	-1.075	-0.937	-0.410
3	3	cluster_0	1.295	-0.473	-0.515	-1.462	-0.515	-0.411	-0.937	-1.318
4	4	cluster_0	-0.772	-0.473	-0.515	-1.462	-0.515	0.253	-0.937	-1.318
5	5	cluster_0	1.295	-0.362	0.610	-1.462	-0.515	0.916	-0.154	0.497
6	6	cluster_0	-0.772	-0.473	-0.515	-1.462	-0.515	-1.075	-0.937	-0.410
7	7	cluster_0	1.295	-0.250	1.734	-1.462	-0.515	0.916	-0.154	0.497
8	8	cluster_0	-0.772	-0.473	-0.515	-1.462	-0.515	0.916	-0.154	0.497
9	9	cluster_0	1.295	-0.473	-0.515	-1.462	-0.515	-1.075	0.629	-0.410
10	10	cluster_0	-0.772	-0.138	-0.515	-1.462	-0.515	0.916	0.629	-0.410
11	11	cluster_1	-0.772	-0.473	-0.515	-1.462	1.943	-0.411	1.411	1.404
12	12	cluster_0	1.295	-0.473	-0.515	-1.462	-0.515	0.253	-0.937	-1.318
13	13	cluster_0	1.295	-0.138	-0.515	-1.462	-0.515	-1.075	0.629	-0.410
14	14	cluster_0	-0.772	-0.473	-0.515	-1.462	-0.515	-1.075	-0.937	-0.410
15	15	cluster_0	-0.772	-0.473	-0.515	-1.462	-0.515	0.916	0.629	0.497
16	16	cluster_0	-0.772	-0.473	-0.515	-1.462	-0.515	0.916	-0.937	-1.318
17	17	cluster_0	-0.772	-0.026	1.734	-1.462	-0.515	0.916	-0.154	0.497
18	18	cluster_1	1.295	-0.473	-0.515	-1.462	1.943	-0.411	1.411	1.404
19	19	cluster_0	-0.772	-0.473	-0.515	-1.462	-0.515	-0.411	-0.937	-1.318

Random Forest: (K=2)



accuracy: 99.98% +/- 0.02% (micro average: 99.98%)

	true cluster_0	true cluster_1	class precision
pred. cluster_0	24333	6	99.98%
pred. cluster_1	0	6358	100.00%
class recall	100.00%	99.91%	

Row No.	id	label	shot_made...	action_type	combined_...	season	shot_type	shot_zone_...	shot_zone_...	shot_z
1	1	cluster_0	-0.772	-0.473	-0.515	-1.462	-0.515	-1.738	-0.937	-1.318
2	2	cluster_0	-0.772	-0.473	-0.515	-1.462	-0.515	-1.075	-0.937	-0.410
3	3	cluster_0	1.295	-0.473	-0.515	-1.462	-0.515	-0.411	-0.937	-1.318
4	4	cluster_0	-0.772	-0.473	-0.515	-1.462	-0.515	0.253	-0.937	-1.318
5	5	cluster_0	1.295	-0.362	0.610	-1.462	-0.515	0.916	-0.154	0.497
6	6	cluster_0	-0.772	-0.473	-0.515	-1.462	-0.515	-1.075	-0.937	-0.410
7	7	cluster_0	1.295	-0.250	1.734	-1.462	-0.515	0.916	-0.154	0.497
8	8	cluster_0	-0.772	-0.473	-0.515	-1.462	-0.515	0.916	-0.154	0.497
9	9	cluster_0	1.295	-0.473	-0.515	-1.462	-0.515	-1.075	0.629	-0.410
10	10	cluster_0	-0.772	-0.138	-0.515	-1.462	-0.515	0.916	0.629	-0.410
11	11	cluster_1	-0.772	-0.473	-0.515	-1.462	1.943	-0.411	1.411	1.404
12	12	cluster_0	1.295	-0.473	-0.515	-1.462	-0.515	0.253	-0.937	-1.318
13	13	cluster_0	1.295	-0.138	-0.515	-1.462	-0.515	-1.075	0.629	-0.410
14	14	cluster_0	-0.772	-0.473	-0.515	-1.462	-0.515	-1.075	-0.937	-0.410
15	15	cluster_0	-0.772	-0.473	-0.515	-1.462	-0.515	0.916	0.629	0.497
16	16	cluster_0	-0.772	-0.473	-0.515	-1.462	-0.515	0.916	-0.937	-1.318
17	17	cluster_0	-0.772	-0.026	1.734	-1.462	-0.515	0.916	-0.154	0.497
18	18	cluster_1	1.295	-0.473	-0.515	-1.462	1.943	-0.411	1.411	1.404
19	19	cluster_0	-0.772	-0.473	-0.515	-1.462	-0.515	-0.411	-0.937	-1.318

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.