Predicting Post-Graduation Salaries using Decision Tree and Random Forest

Graduate Tracer Study
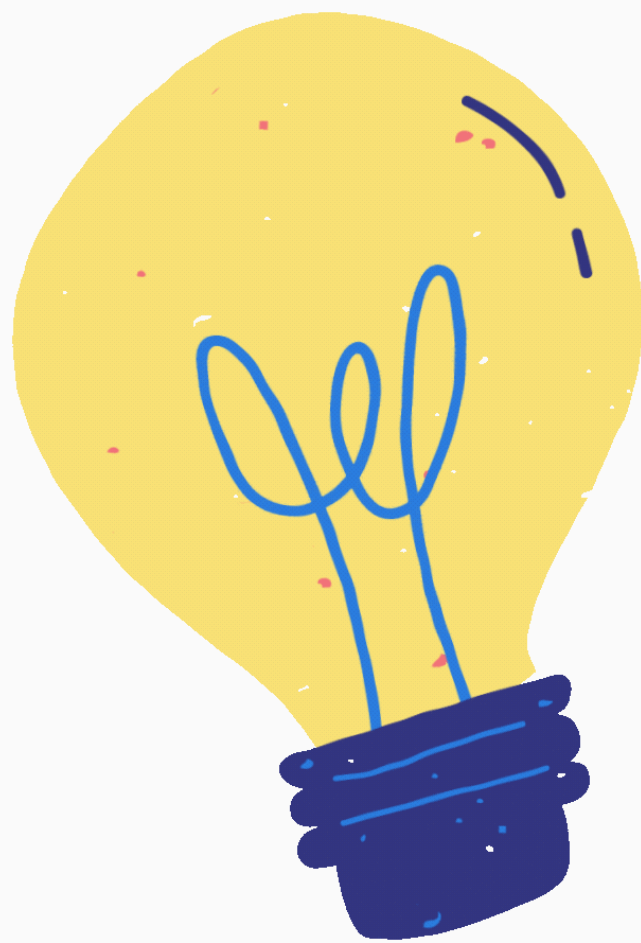
PREPARED BY

AINA FARZANA ZULKIFLI | 2021817142

NUR EISYATIN RADHIAH ANNUAR | 2021610238

SITI MUKHLISAH MUSKAMAL | 2022815898

YASMIN NABILA OTHMAN | 2021619506

CS2705B

PROBLEM STATEMENT

Limited exploration of the intricate relationship between program code, job occupation, company nature, years of working, and current salary in the existing dataset.

OBJECTIVES

- To study the correlation between program codes and post-graduation salaries.
- To build predictive models using classification technique to forecast post-graduation salaries based on the courses taken by graduates.
- To evaluate the accuracy of the predictive models.

Dataset Used

Data Source	Graduate Tracer Study
Total Data	902
Total Attributes	12
Target Class Attribute	Current Salary Midpoint
List of Attributes	Gender, Age, Graduation Year, Program Code, Current Occupation, Job category, Department, Company Name, Type of Company, Company's nature of business, Years of work experience (after graduation), Current Salary Midpoint

DECISION TREE

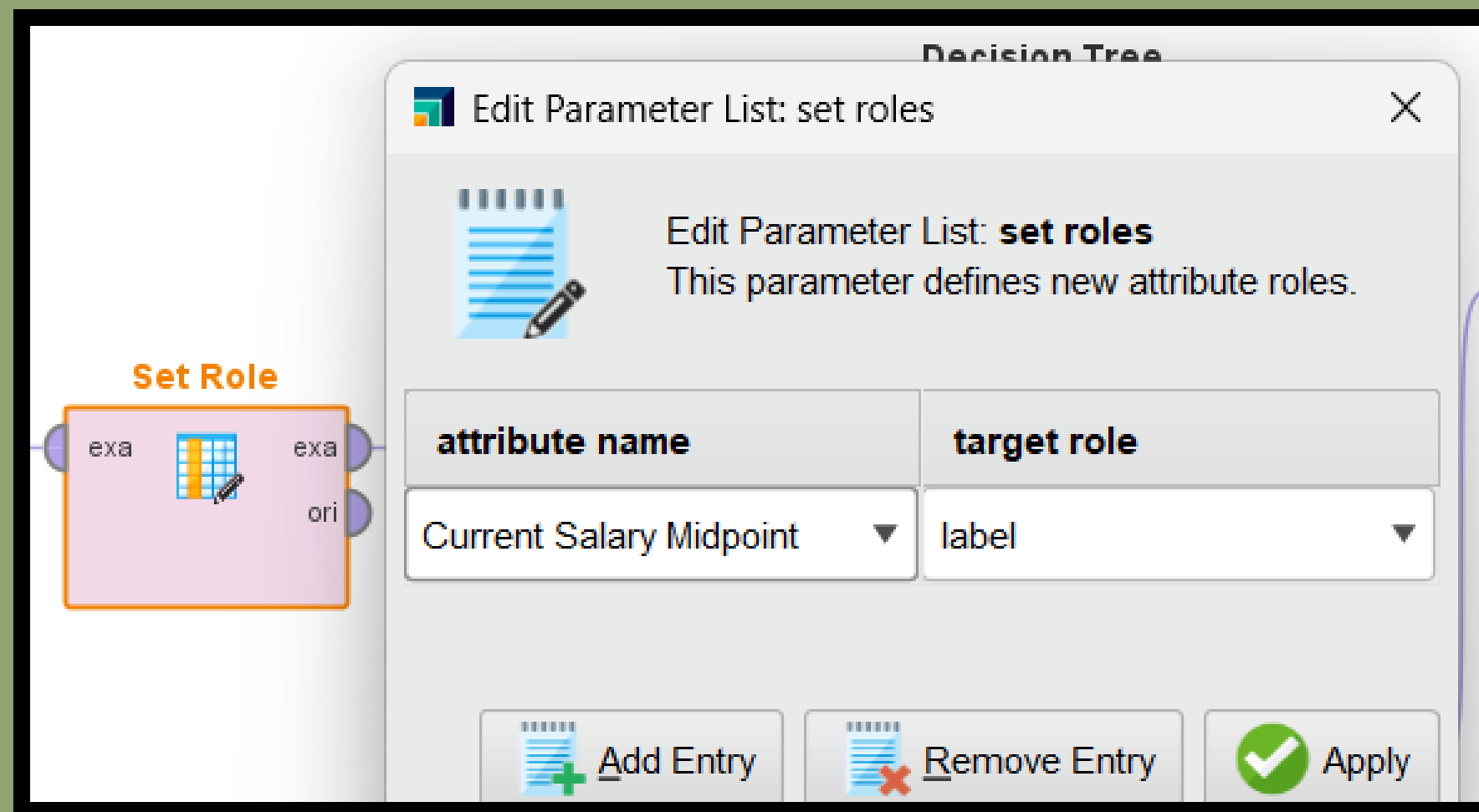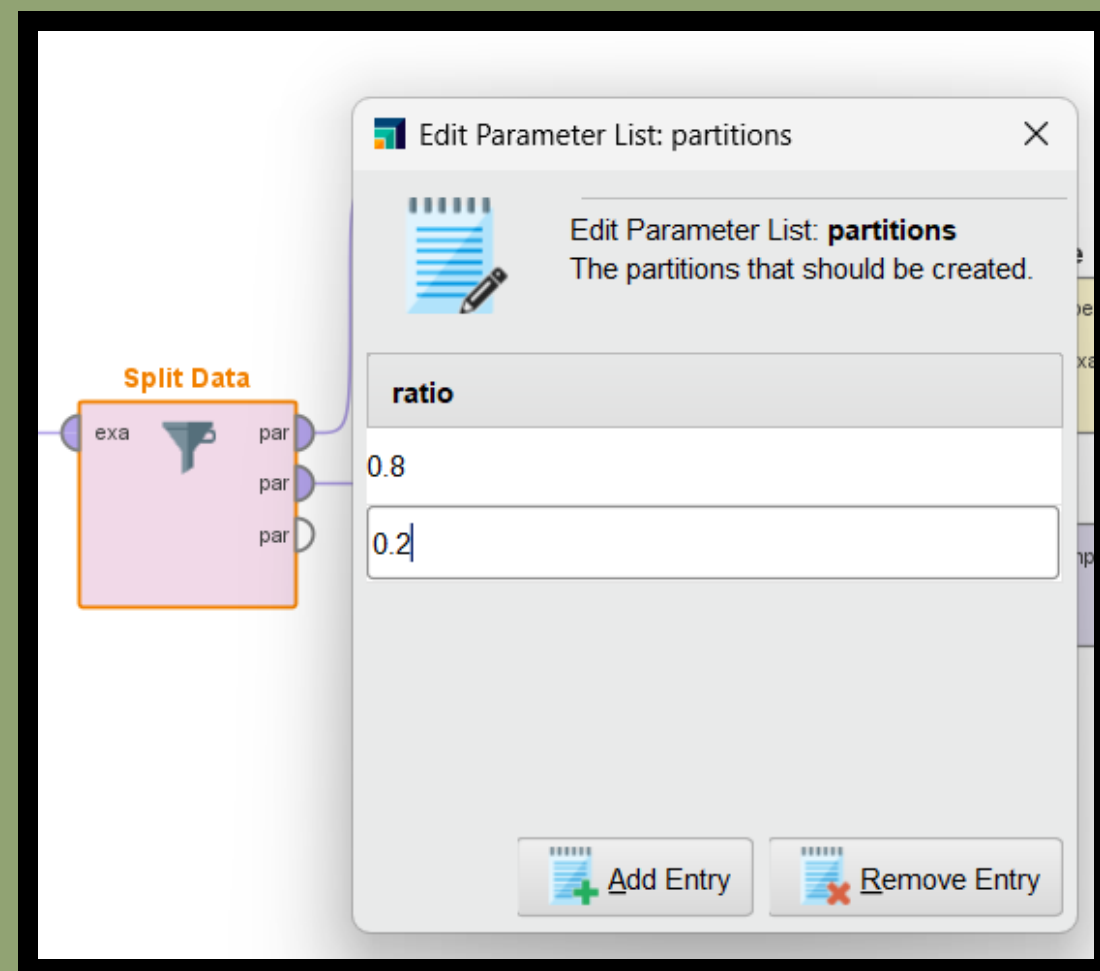## Modelling

Process in RapidMiner
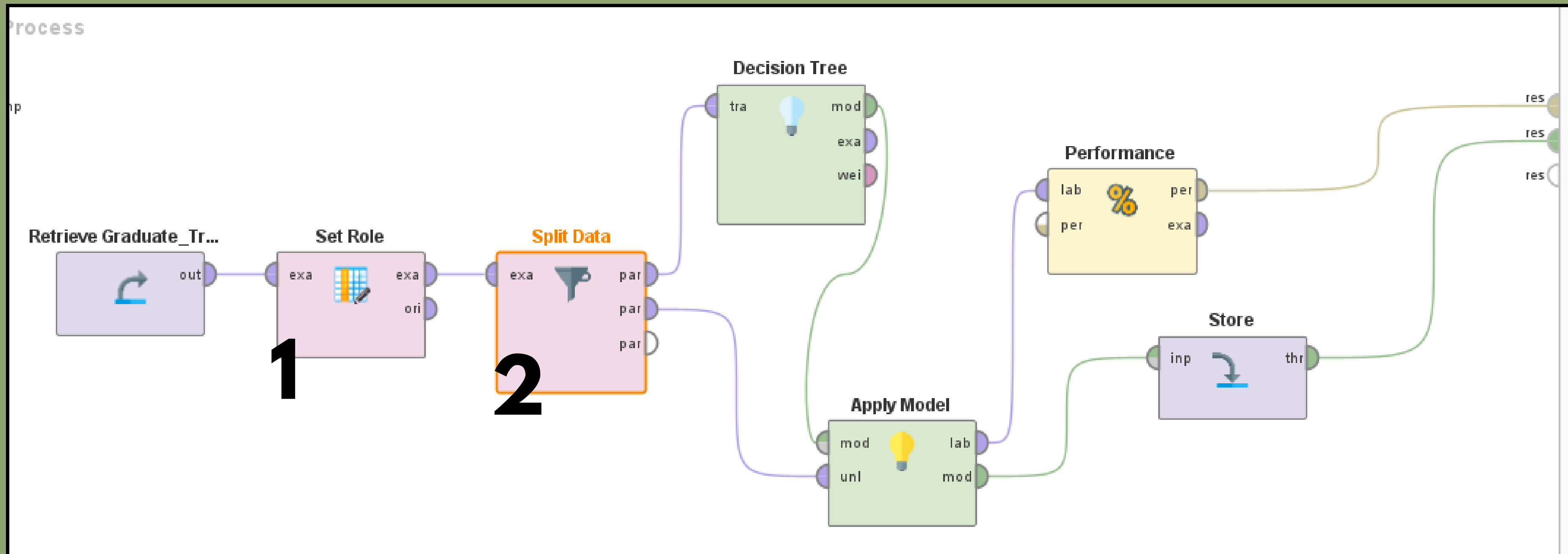
DECISION TREE

Setting Parameters

1 **Set Role**
Class Label - Current Salary Midpoint

2 **Split Data**
0.8 - Training
0.2 - Testing
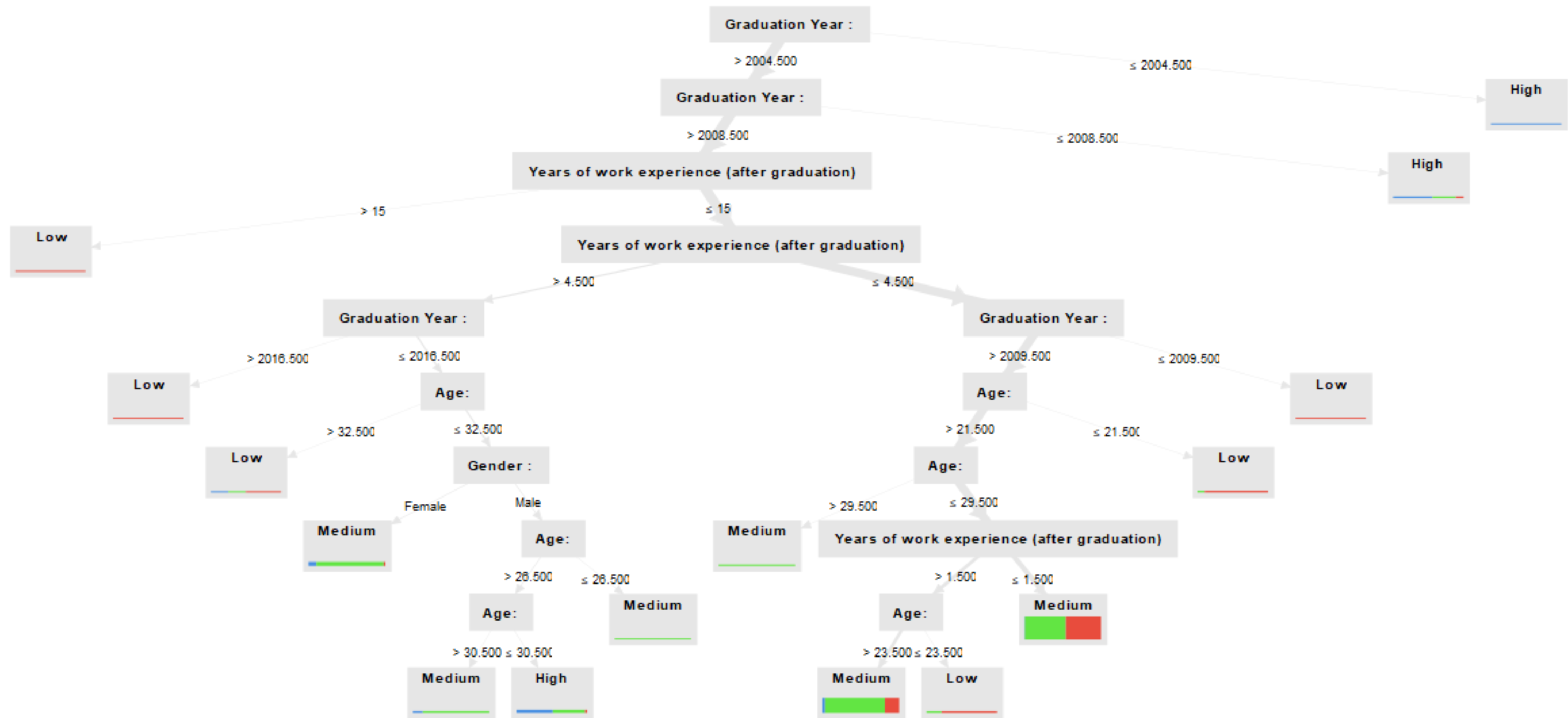

DECISION TREE

DECISION TREE

Evaluation

Confusion Matrix and Accuracy Percentage

accuracy: 62.15%

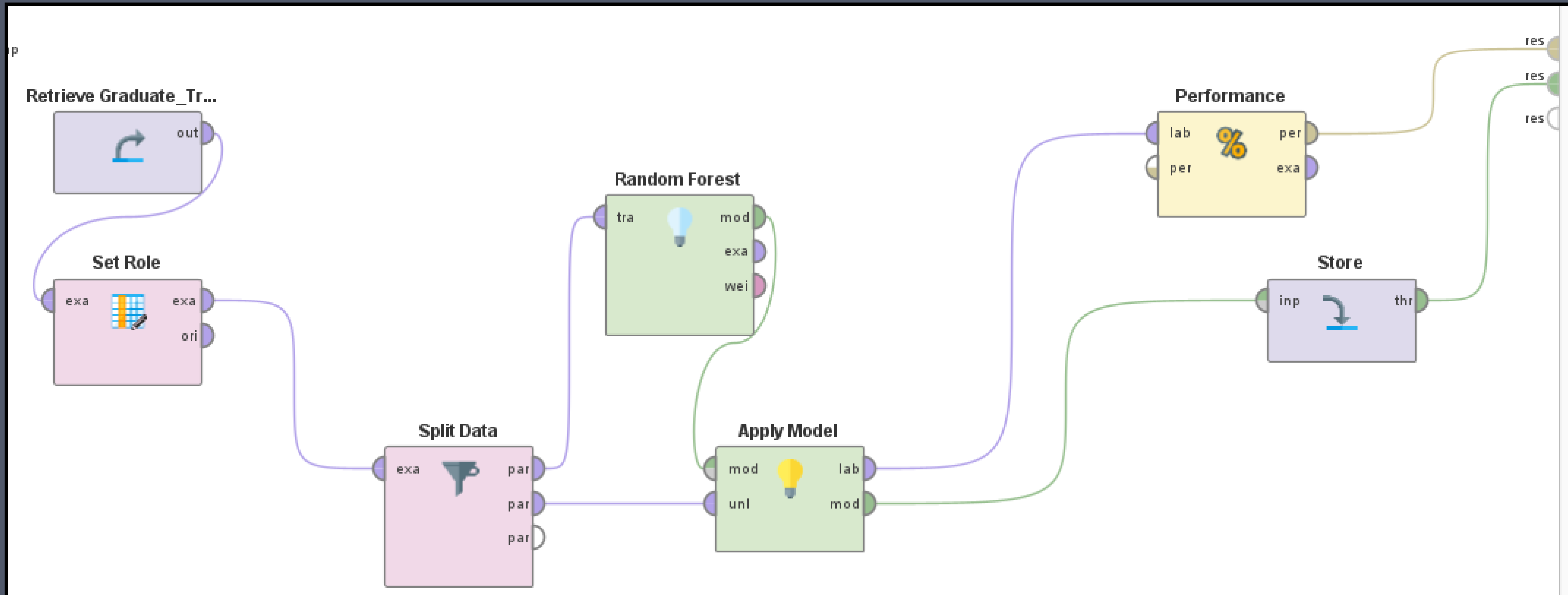|	true High	true Medium	true Low	class precision
pred. High	7	6	0	53.85%
pred. Medium	3	101	55	63.52%
pred. Low	1	2	2	40.00%
class recall	63.64%	92.66%	3.51%	

DECISION TREE

Tree

```
Graduation Year : > 2004.500
|    Graduation Year : > 2008.500
|    |    Years of work experience (after graduation): > 15: Low {High=0, Medium=0, Low=2}
|    |    Years of work experience (after graduation): ≤ 15
|    |    |    Years of work experience (after graduation): > 4.500
|    |    |    |    Graduation Year : > 2016.500: Low {High=0, Medium=0, Low=2}
|    |    |    |    Graduation Year : ≤ 2016.500
|    |    |    |    |    Age: > 32.500: Low {High=1, Medium=1, Low=2}
|    |    |    |    |    Age: ≤ 32.500
|    |    |    |    |    |    Gender : = Female: Medium {High=5, Medium=42, Low=1}
|    |    |    |    |    |    Gender : = Male
|    |    |    |    |    |    |    Age: > 26.500
|    |    |    |    |    |    |    |    Age: > 30.500: Medium {High=2, Medium=13, Low=0}
|    |    |    |    |    |    |    |    Age: ≤ 30.500: High {High=18, Medium=16, Low=1}
|    |    |    |    |    |    |    Age: ≤ 26.500: Medium {High=0, Medium=2, Low=0}
|    |    |    Years of work experience (after graduation): ≤ 4.500
|    |    |    |    Graduation Year : > 2009.500
|    |    |    |    |    Age: > 21.500
|    |    |    |    |    |    Age: > 29.500: Medium {High=0, Medium=6, Low=0}
|    |    |    |    |    |    Age: ≤ 29.500
|    |    |    |    |    |    |    Years of work experience (after graduation): > 1.500
|    |    |    |    |    |    |    |    Age: > 23.500: Medium {High=7, Medium=168, Low=40}
|    |    |    |    |    |    |    |    Age: ≤ 23.500: Low {High=0, Medium=3, Low=11}
|    |    |    |    |    |    |    Years of work experience (after graduation): ≤ 1.500: Medium {High=3, Medium=180, Low=155}
|    |    |    |    |    Age: ≤ 21.500: Low {High=0, Medium=1, Low=9}
|    |    |    |    Graduation Year : ≤ 2009.500: Low {High=0, Medium=0, Low=2}
|    Graduation Year : ≤ 2008.500: High {High=5, Medium=3, Low=1}
Graduation Year : ≤ 2004.500: High {High=3, Medium=0, Low=0}
```
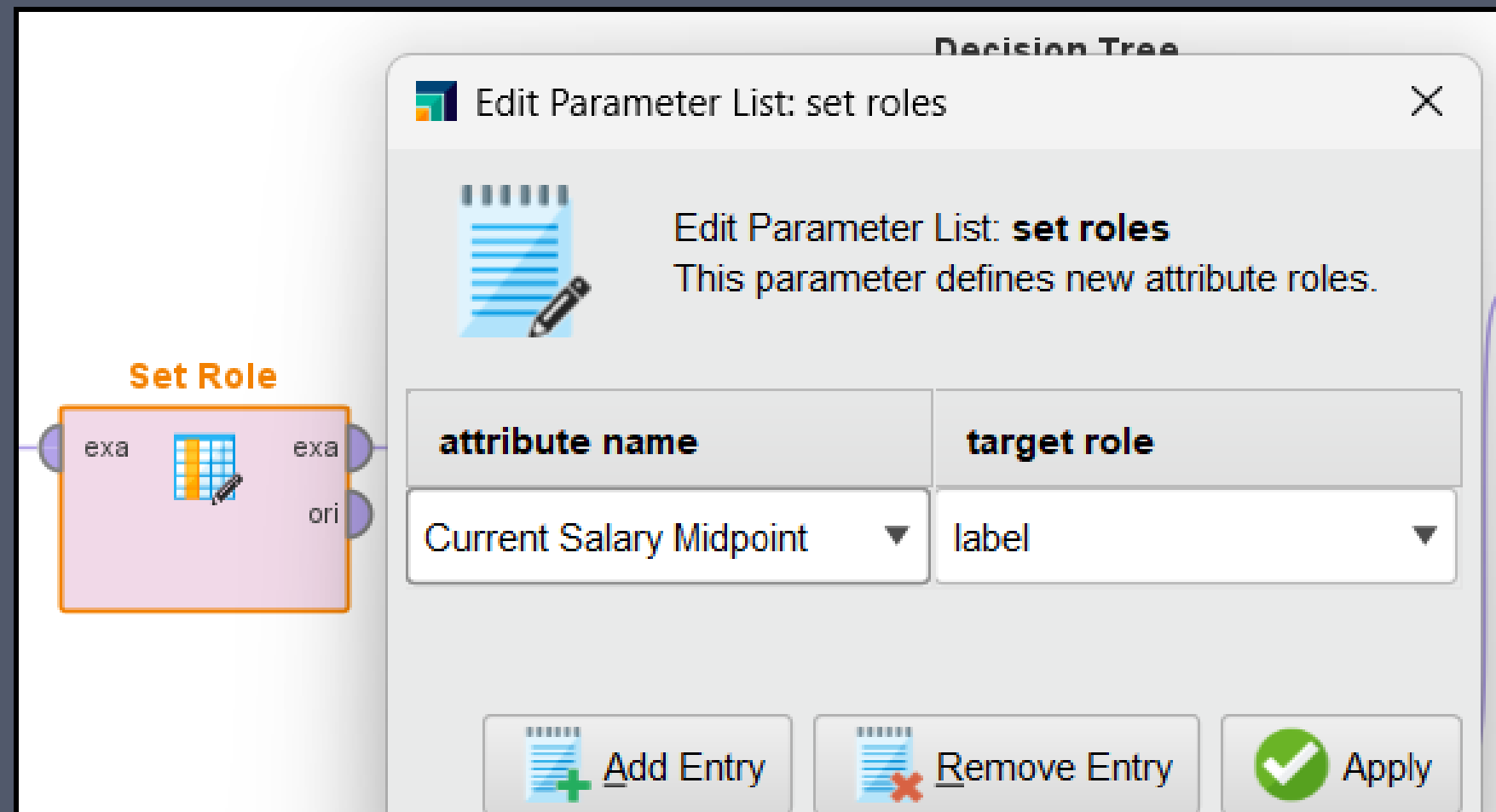
RANDOM FOREST

Modelling

Process in RapidMiner

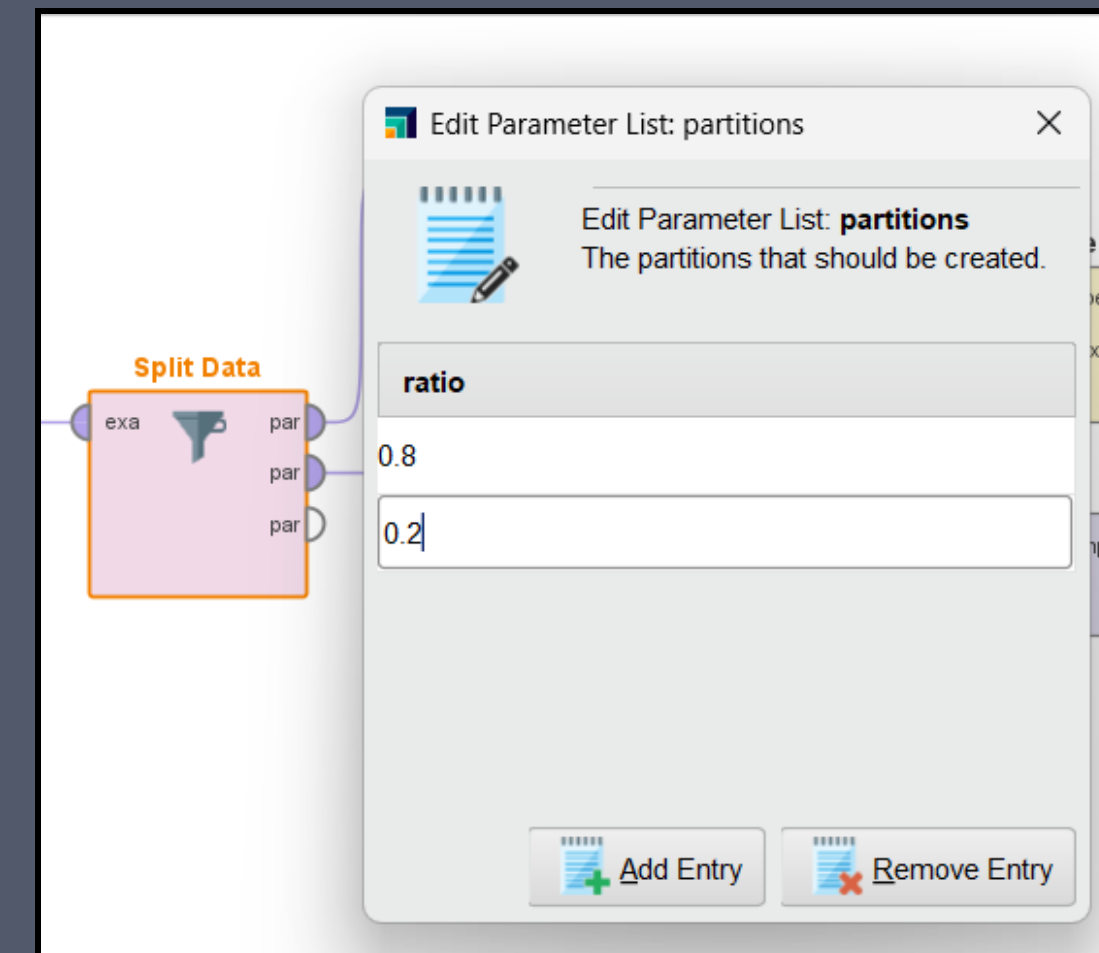**RANDOM FOREST**

Set Role
Class Label - Current Salary Midpoint

Split Data
Training: Testing
0.8:0.2

RANDOM FOREST

RANDOM FOREST

Evaluation

Confusion Matrix and Accuracy Percentage

accuracy: 64.41%

	true High	true Medium	true Low	class precision
pred. High	2	1	1	50.00%
pred. Medium	9	93	37	66.91%
pred. Low	0	15	19	55.88%
class recall	18.18%	85.32%	33.33%	

RANDOM FOREST

Comparison between Decision Tree and Random Forest
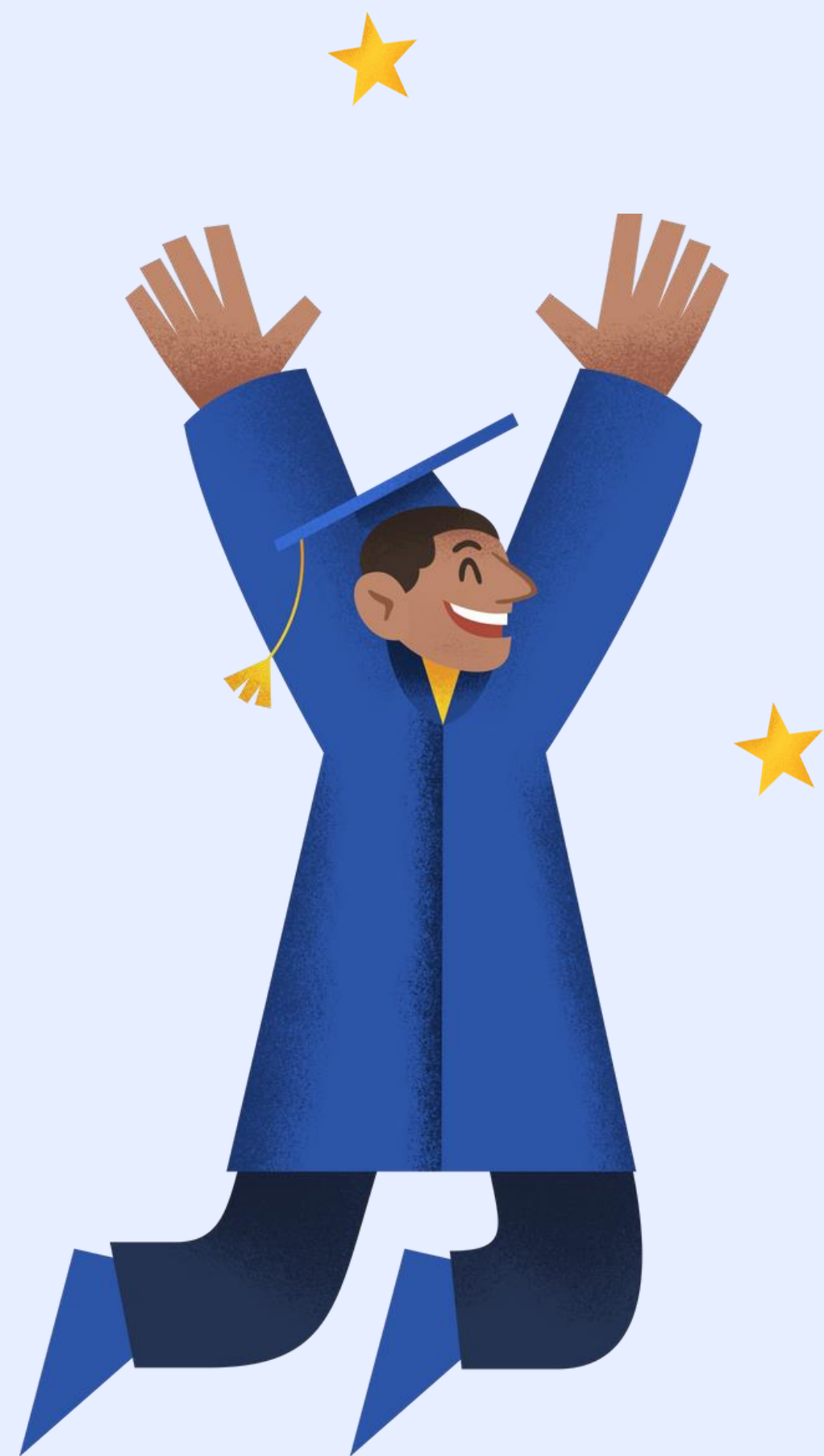
Decision tree

accuracy: 62.15%

	true High	true Medium	true Low	class precision
pred. High	7	6	0	53.85%
pred. Medium	3	101	55	63.52%
pred. Low	1	2	2	40.00%
class recall	63.64%	92.66%	3.51%	

Random Forest

accuracy: 64.41%

	true High	true Medium	true Low	class precision
pred. High	2	1	1	50.00%
pred. Medium	9	93	37	66.91%
pred. Low	0	15	19	55.88%
class recall	18.18%	85.32%	33.33%	

Deployment

DEPLOYMENT OF RANDOM FOREST
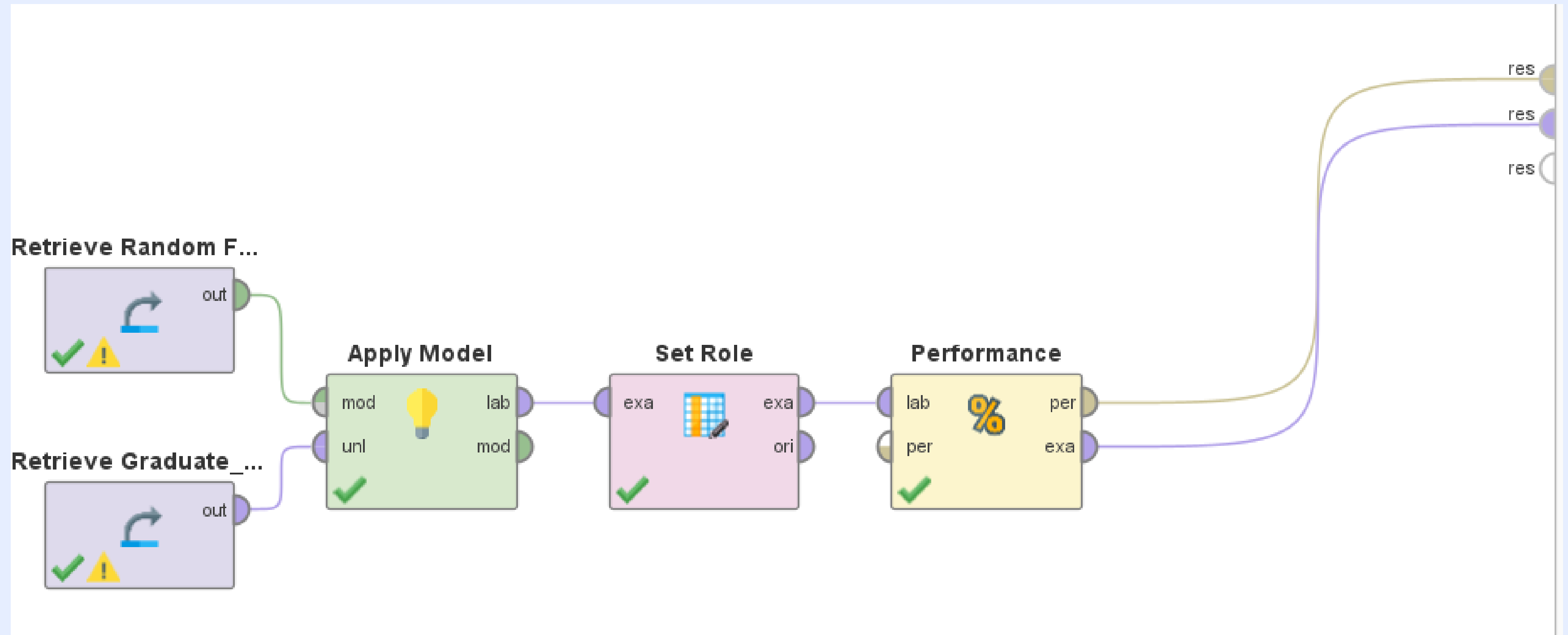
PERFORMANCE RESULT

accuracy: 75.00%

	true Medium	true Low	true High	class precision
pred. Medium	13	2	1	81.25%
pred. Low	2	1	0	33.33%
pred. High	0	0	1	100.00%
class recall	86.67%	33.33%	50.00%	

Precision: 71.53%

Recall: 56.67

F1-Score: 0.6782

Correlation between Program Code, Years of Work Experience and Prediction Salary
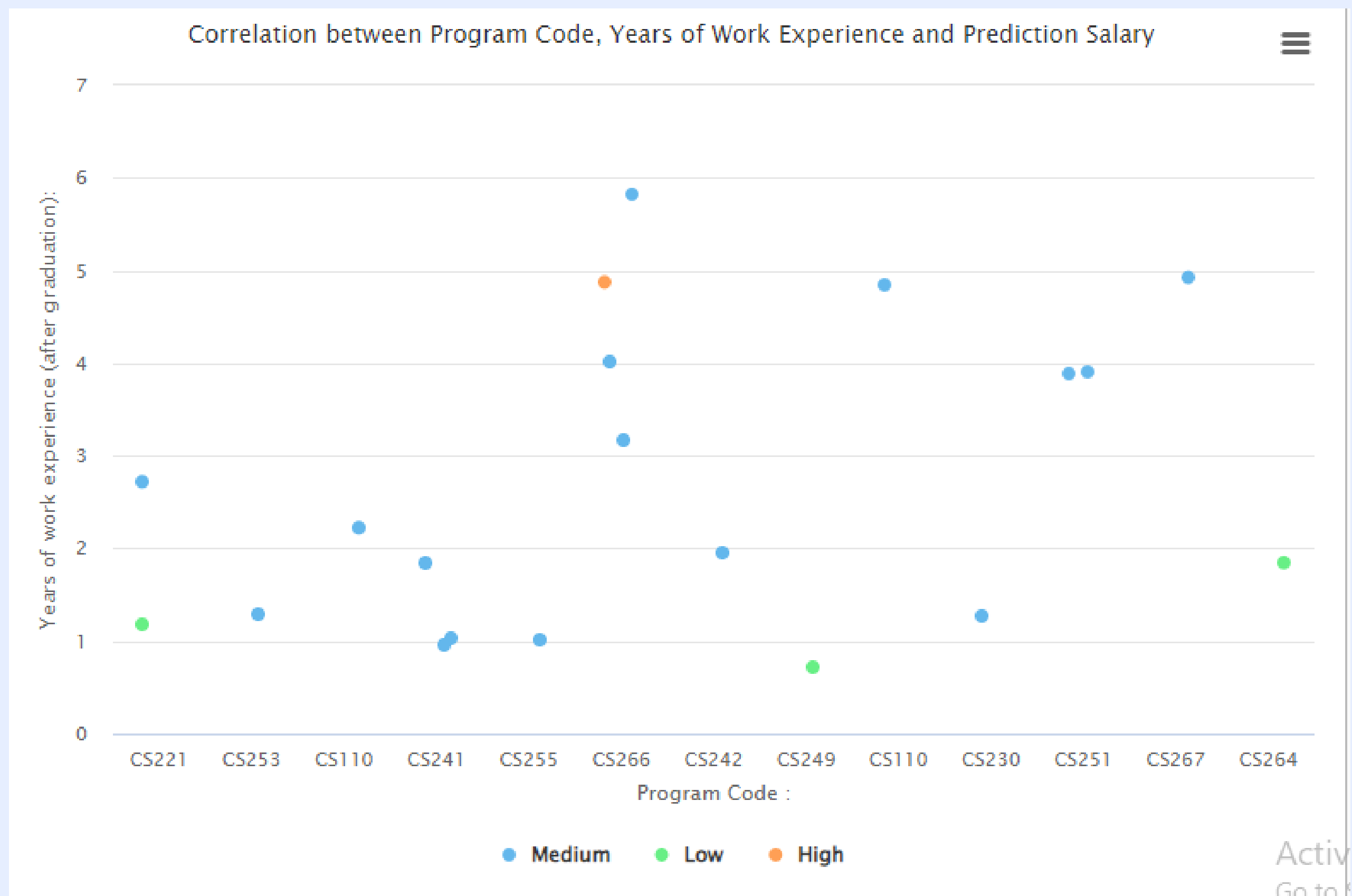
LOW: < RM3000
MEDIUM: RM3000 - RM5000
HIGH: > RM5000 HIGH

RECOMMENDATION

Tailored Career Guidance for Graduates

Skill Enhancement Programs

Data Collection Enhancement

Conclusion

Random Forest model has higher accuracy of 62.15%

Students are able to make more informed decisions about their educational paths

THANK YOU
and have a nice day!