# Batch: 3 Experiment Number: 3

**Roll Number:** 16010421119 **Name: Aarya Tiwari**

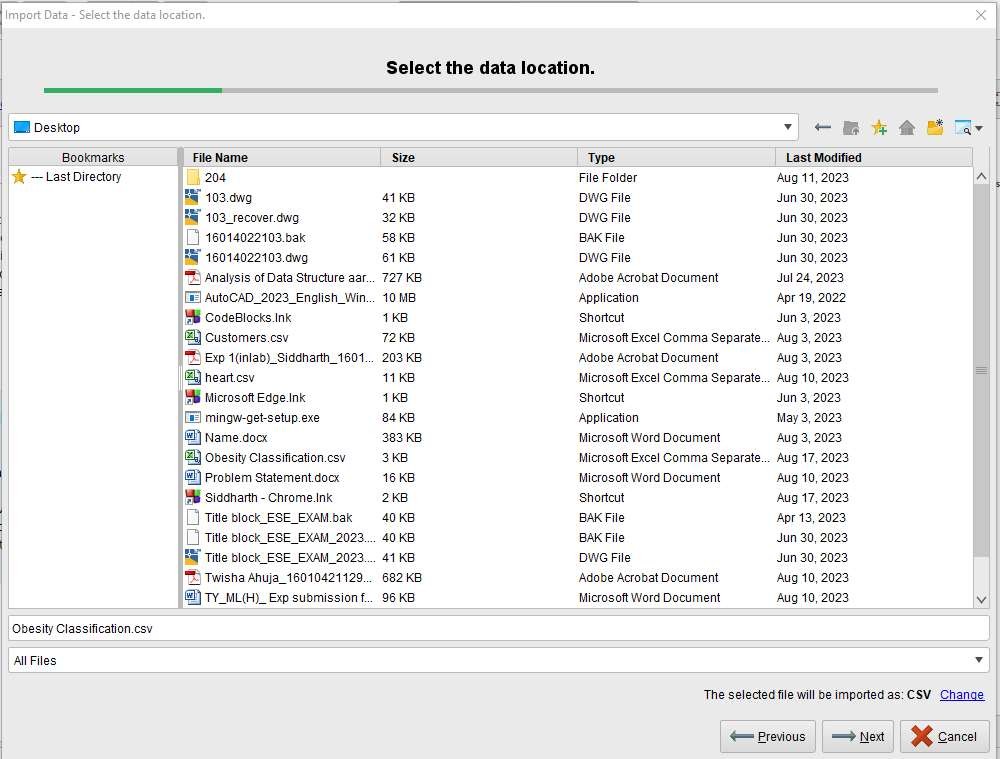
**Aim of the Experiment:** To analyze the data and execute any two classification algorithms using RapidMiner.

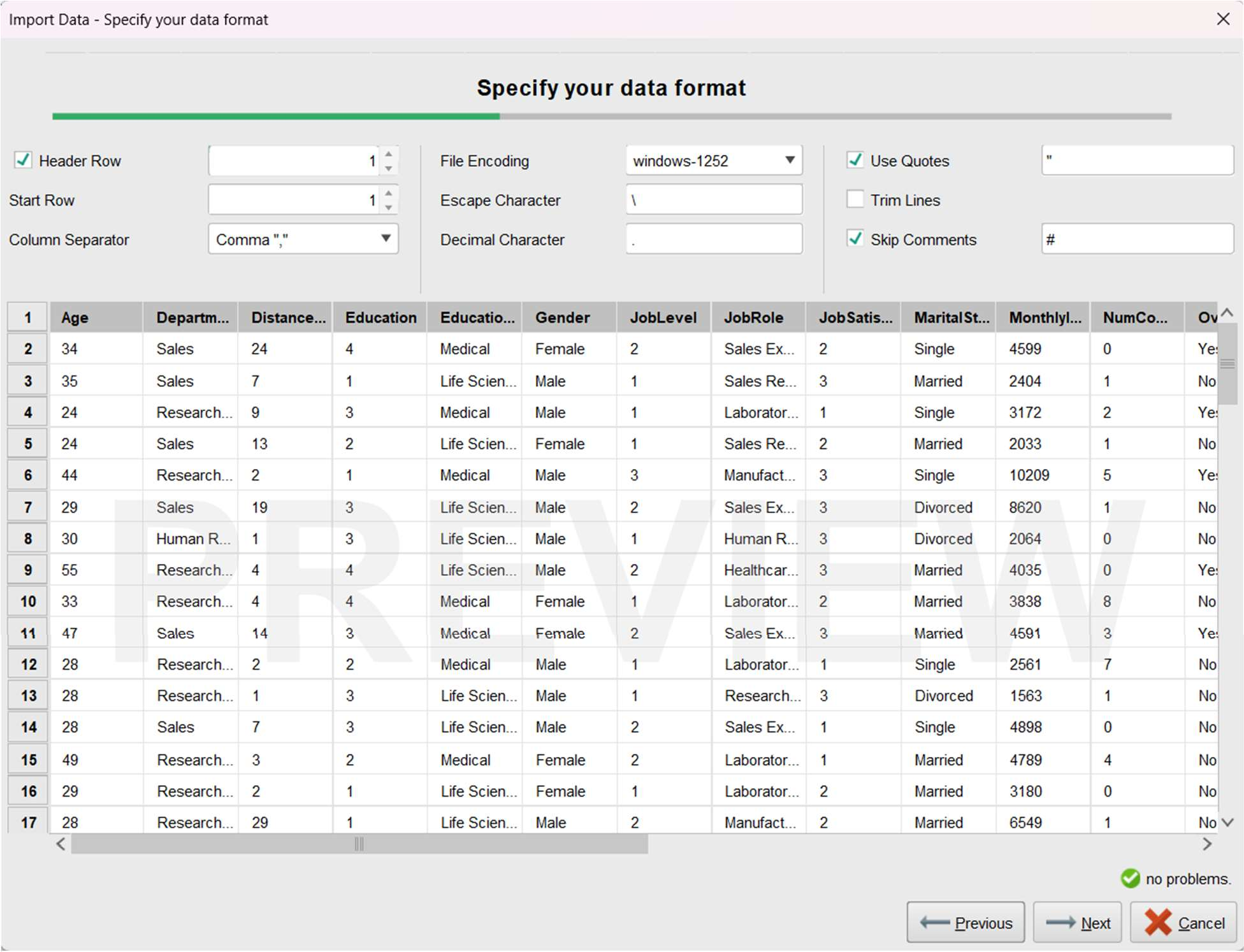
# Program/ Steps:

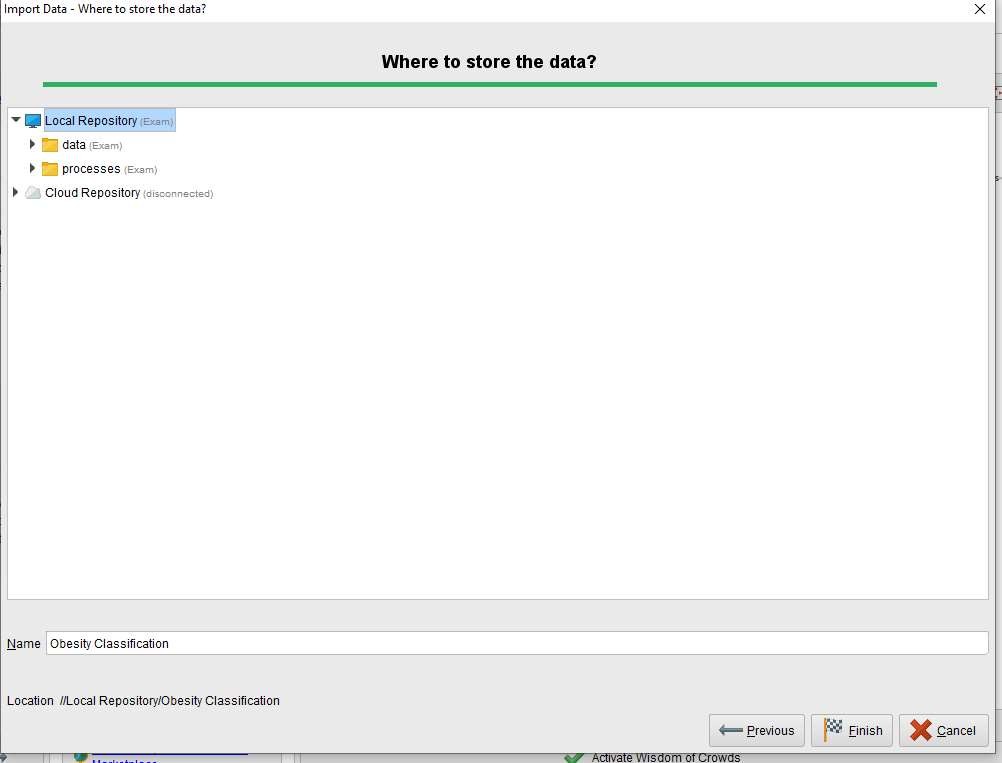
1. **Dataset Selection:** I opted for a dataset that includes classification tasks from Kaggle. Kaggle is a widely used platform that offers various datasets for machine learning and data analysis projects. The dataset I selected likely contains information relevant to the classification problem I intend to solve.
2. **Data Import in RapidMiner:** I utilized RapidMiner, a data science platform, to import the chosen dataset. This platform streamlines the data preprocessing and modeling process. Since no preprocessing was necessary, I directly loaded the dataset into RapidMiner, allowing me to proceed to the next steps.
3. **Target Attribute Definition:** In the dataset, I identified a key attribute called 'label' that serves as the target variable. This attribute contains information indicating whether an individual is categorized as underweight, normal weight, overweight, or obese. Designating this attribute as the target is crucial as it defines the objective of our classification task.
4. **Dataset Splitting for Training and Testing:** To ensure reliable model evaluation, I divided the dataset into two subsets: a training set and a testing set. Approximately 70% of the data was allocated for training the models, while the remaining 30% was reserved for evaluating their performance. This separation helps assess how well the models generalize to unseen data.
5. **Model Creation (Decision Tree and Naive Bayes):** I proceeded to construct two distinct machine learning models: one using the decision tree algorithm and the other employing the Naive Bayes algorithm. Each model is designed to predict the target attribute (weight category) based on the dataset's features. These models will later be evaluated to determine their effectiveness.
6. **Performance Analysis and Results Examination:** After training and evaluating both models, I conducted an in-depth analysis of their performance. This involved assessing various metrics such as accuracy, precision, recall, F1-score, and possibly AUC-ROC. I also studied the confusion matrix to understand the types of predictions the models made. Visualizations like ROC curves and precision-recall curves provided further insights into model behavior. Additionally, I considered factors like model interpretability, robustness, computational efficiency, and alignment with project goals to make an informed decision about selecting the most suitable model for the classification task. The analysis and evaluation process allowed me to

understand how well each model performed and which one is better suited for the specific problem at hand.

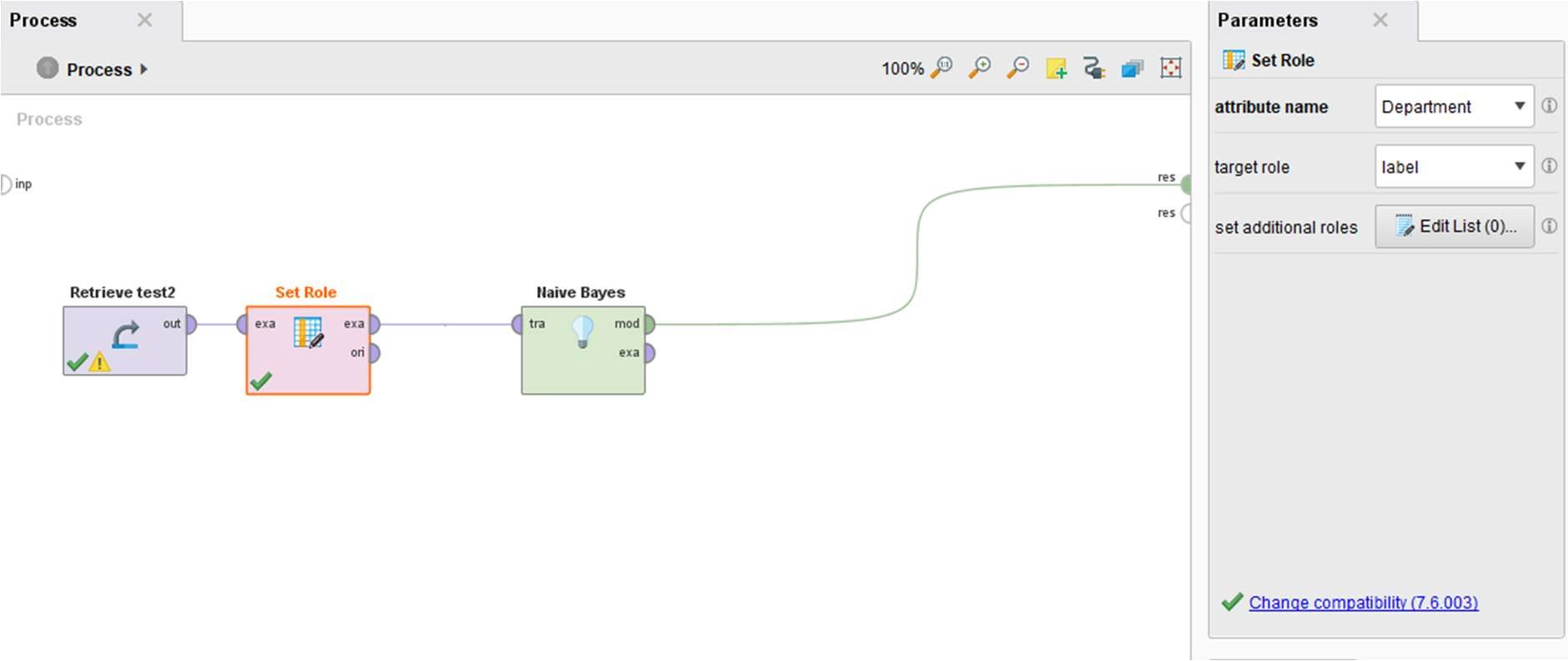
# Output/Result:

1. **Explore the dataset**

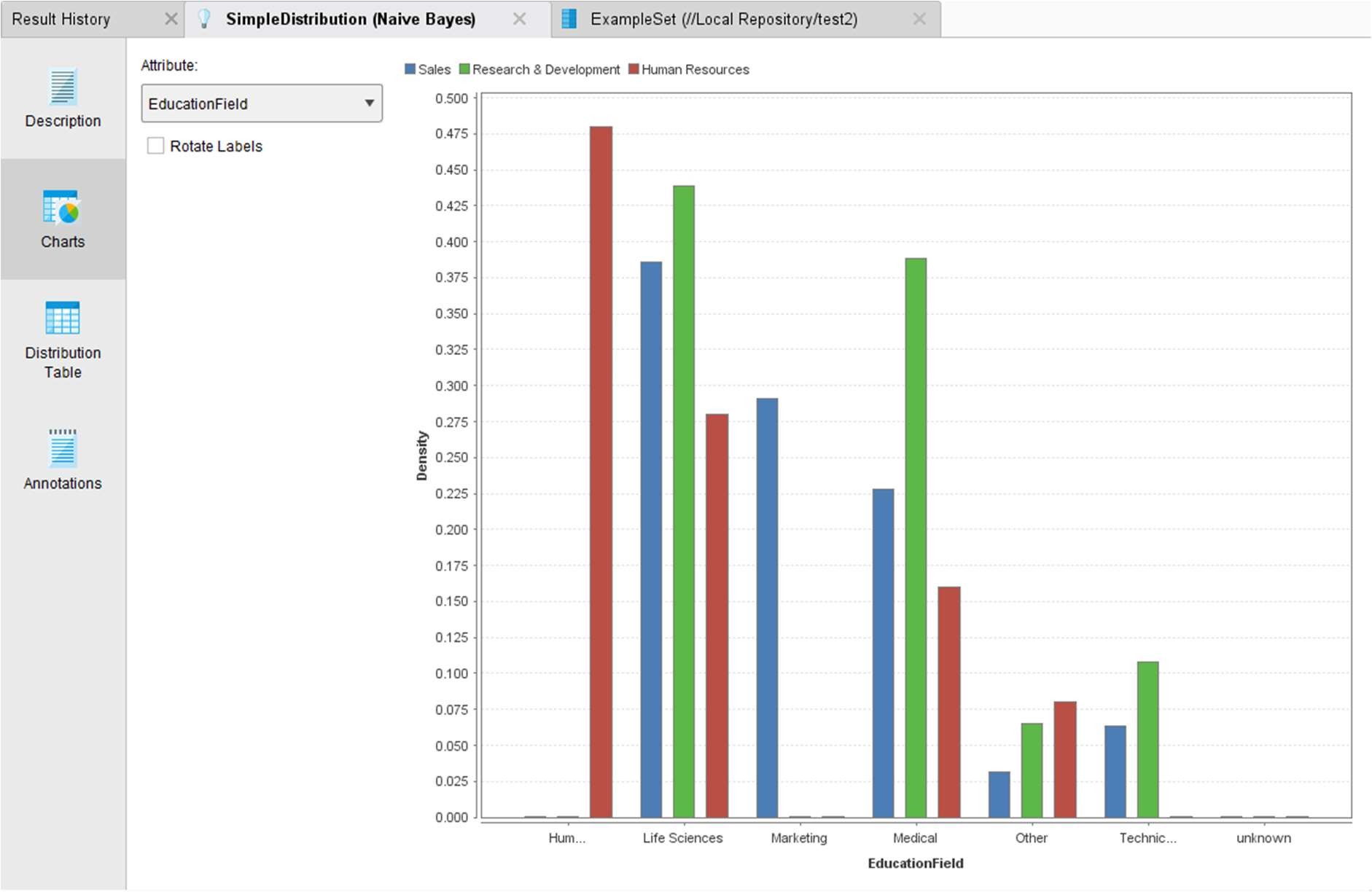


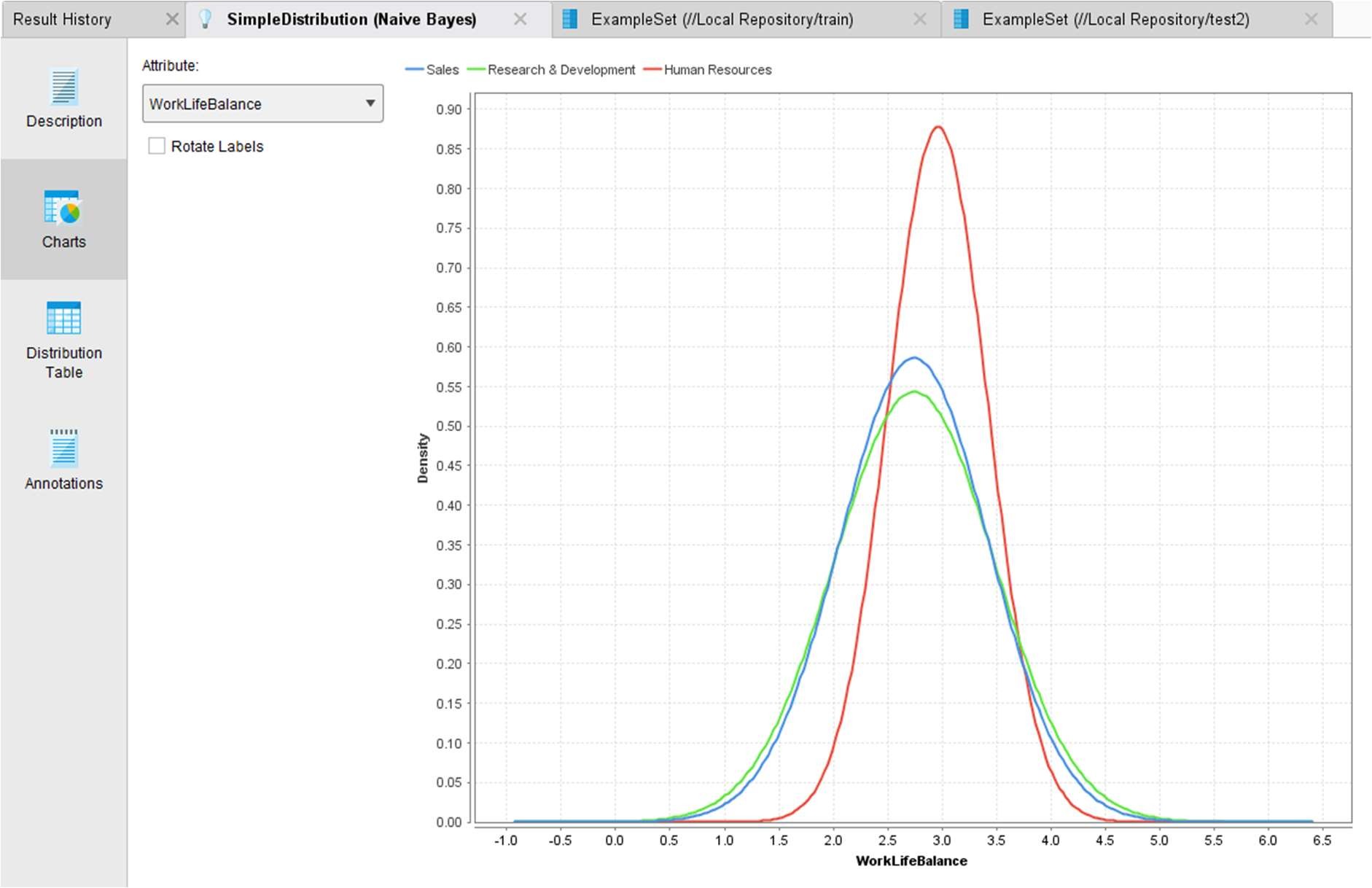


**Naïve Bayes**

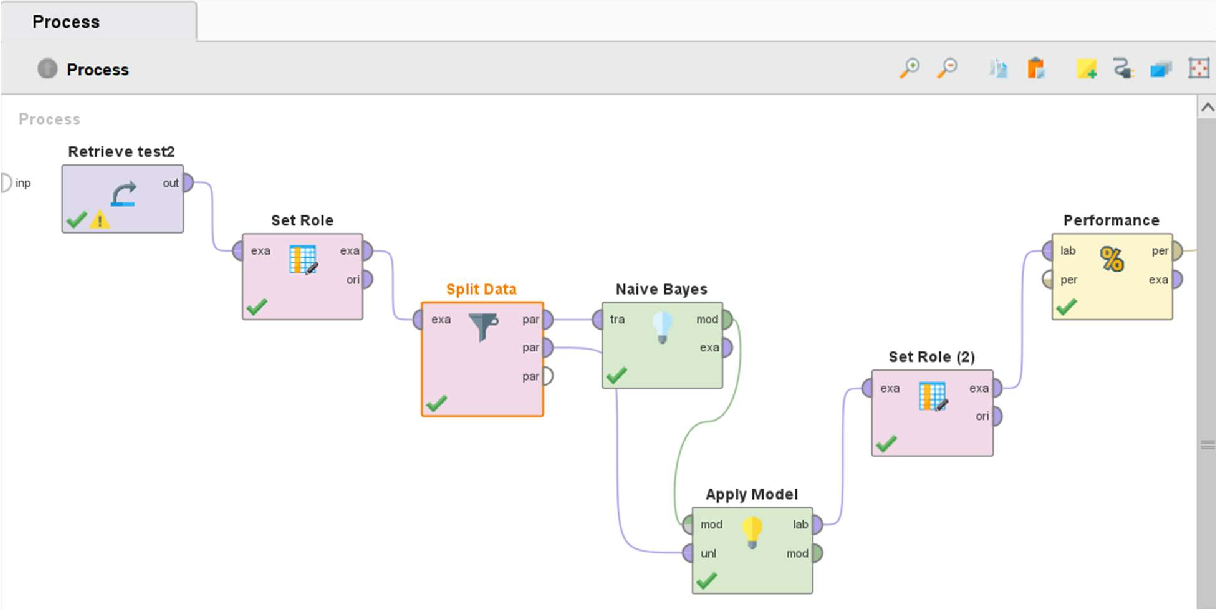


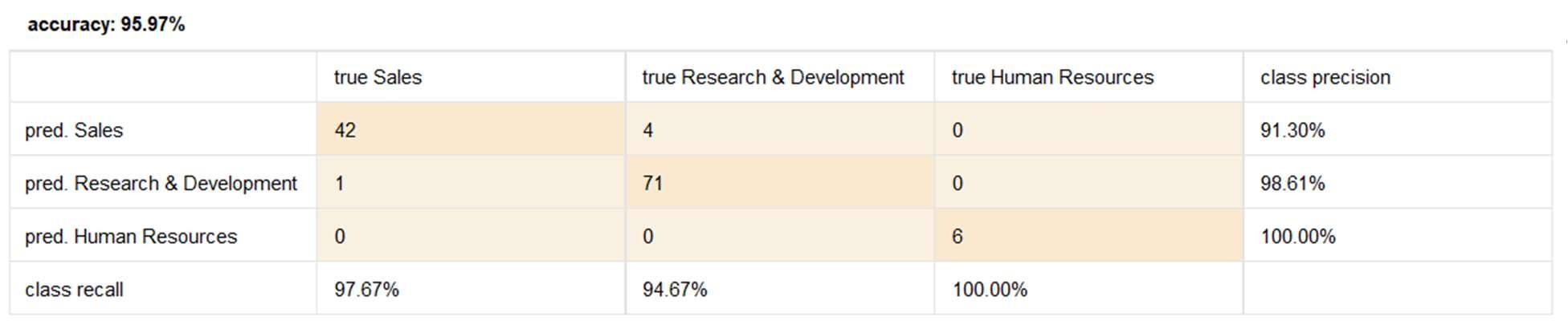
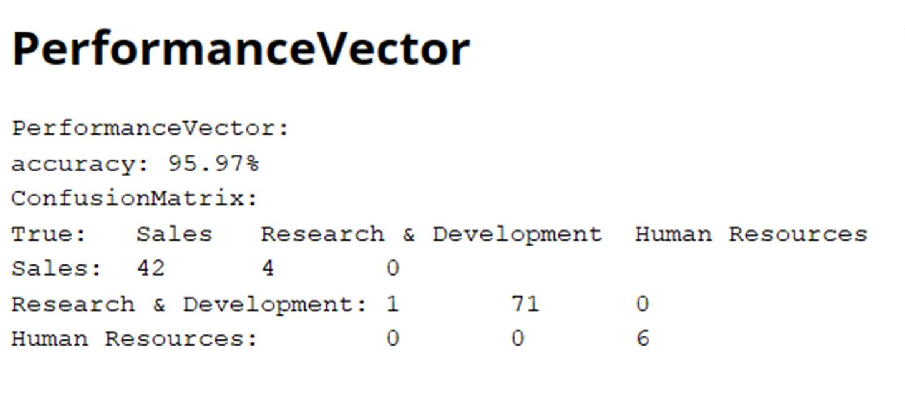


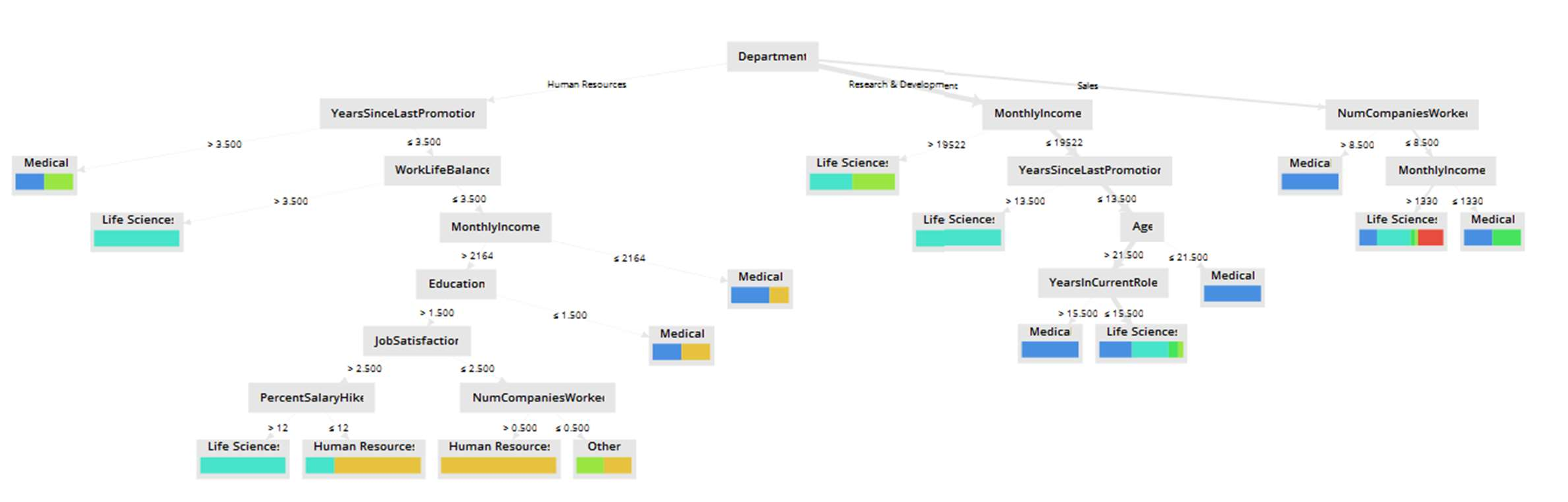




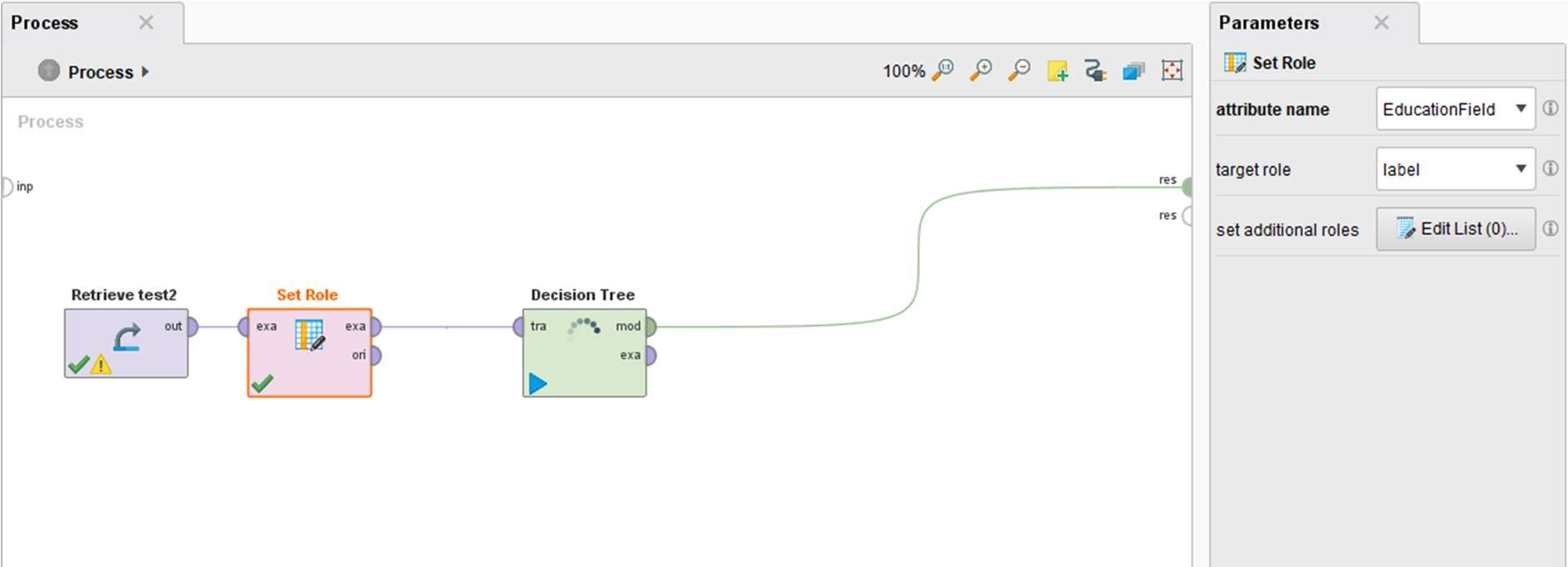
**Calculating performance:**



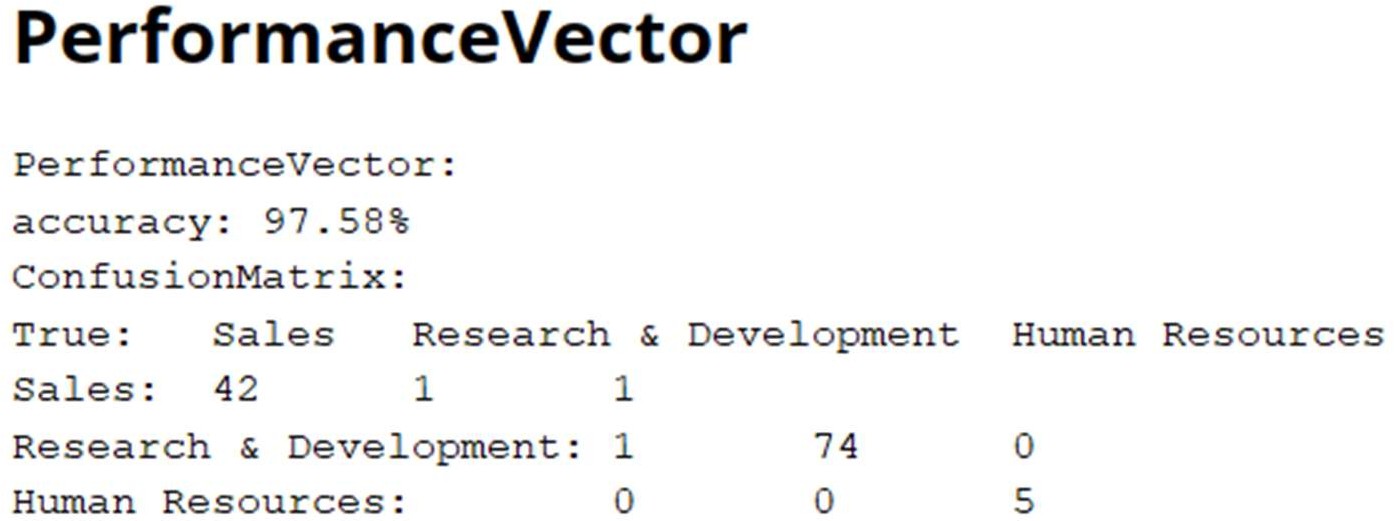
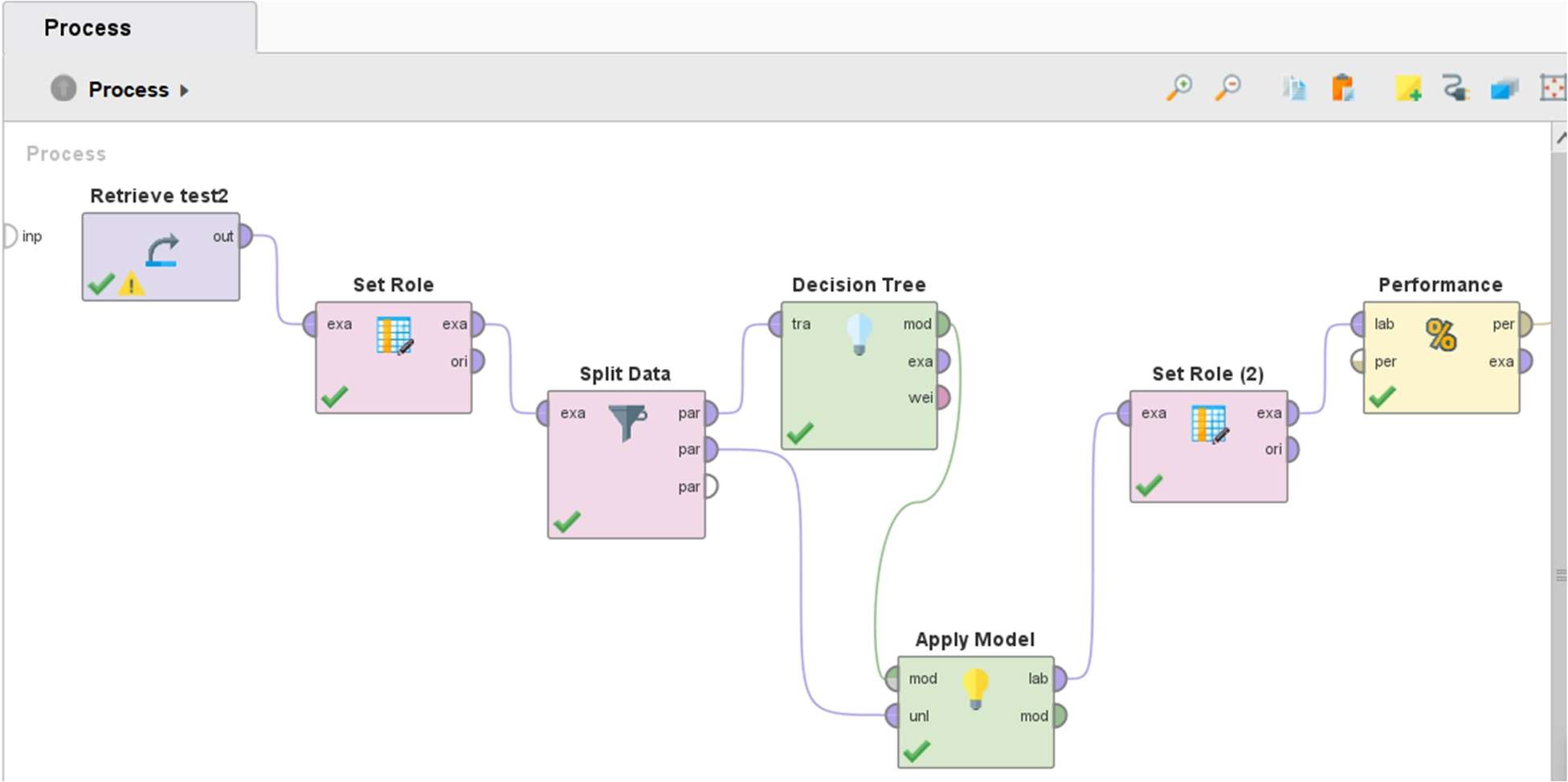


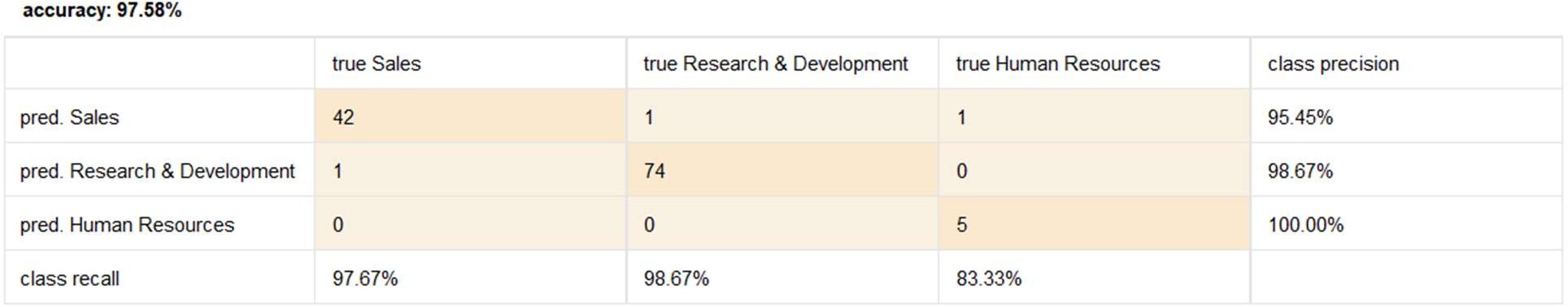


**Decision Trees**



**Calculating Performance:**





1. **Analyze the results produced by Rapidminer**

Results of Naive Bayes:

* Accuracy: 95.97%
* Confusion Matrix:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | true Sales | true Research & Development | true Human Resources | class precision |
| pred. Sales | 42 | 4 | 0 | 91.30% |
| pred. Research & Development | 1 | 71 | 0 | 98.61% |
| pred. Human Resources | 0 | 0 | 6 | 100.00% |
| class recall | 97.67% | 94.67% | 100.00% |  |

Results of Decision Tree:

* Accuracy: 97.58%
* Confusion Matrix:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | true Sales | true Research & | true Human | class |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Development | Resources | precision |
| pred. Sales | 42 | 1 | 1 | 95.45% |
| pred. Research & Development | 1 | 74 | 0 | 98.67% |
| pred. Human Resources | 0 | 0 | 5 | 100.00% |
| class recall | 97.67% | 98.67% | 83.33% |  |

Interpretation:

* The Naive Bayes model achieved an accuracy of 95.97%, indicating it correctly classified instances about 96% of the time on the testing data.
* Looking at the confusion matrix, the model also performs well, but there are a few more misclassifications compared to the Decision Tree model.

Interpretation:

* The Decision Tree model achieved an accuracy of 97.58%, indicating that it correctly classified instances nearly 97% of the time on the testing data.
* Looking at the confusion matrix, the model performs very well across all department categories. The majority of instances are correctly classified, with only a few instances being misclassified.

Comparison and Selection: Both models have shown strong performance with high accuracy rates. The Decision Tree model has a slightly higher accuracy compared to Naive Bayes. However, the choice between the two models could also depend on other factors such as interpretability, computational efficiency, and how the specific types of misclassifications impact the application.

**Some more points:**

1. Interpretability:

• Decision Tree: Think of it like a flowchart; decision trees provide a clear and easy-to-follow structure, making it simple to see how the model makes decisions.

• Naive Bayes: This one's a bit more like a mysterious black box. It's probabilistic and doesn't show its workings as clearly.

1. Robustness and Overfitting:

• Decision Tree: It can get overly obsessed with the training data if it's allowed to grow too complicated. Think of it like memorizing the answers rather than understanding the material.

• Naive Bayes: It's like a chill, easygoing model. Less prone to overthinking the training data, but it might miss subtle connections in the data.

1. Feature Importance:

• Decision Tree: Imagine it as a detective; it readily points out which clues (features) are crucial in solving the case (making predictions).

• Naive Bayes: This detective is a bit more secretive. It won't tell you which clues it values the most.

1. Handling Missing Data:

• Decision Tree: It's like a detective who can work with incomplete evidence. If some clues are missing, it can still crack the case using what's available.

• Naive Bayes: This detective prefers to have all the clues. If something's missing, you'll need to fill in the gaps to get a complete picture.

# Post Lab Question-Answers:

1. List any five open sources / freeware tools available for data mining.

Ans: Five open sources / freeware tools available for data mining:

1. Apache Mahout
2. DataMelt
3. ELKI
4. KNIME
5. Orange

**Outcomes:** CO2: Apply concepts of different types of Learning and Neural Network.

# Conclusion (based on the Results and outcomes achieved):

We successfully analyzed the data and executed two classification algorithms using RapidMiner and analyzed their performances.

# References:

Books/ Journals/ Websites: 1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3rd Edition