Spring
2017

Wine Region Classification

SWEN 5931 RESEARCH TOPICS IN SOFTWARE ENGINEERING

RUSHIKESH MANGRULKAR

Classification of Wine Maker based on chemical attributes

Problem Statement:

Classify the wine maker depending on the analysis of the 13 chemical constituents found in each of the three types of wines.

Overview:

We follow the CRISP – DM methodology for achieving the prediction, the steps involved in the CRISP – DM methodology are:

a. Business Understanding
b. Data Understanding
c. Data Preparation
d. Modeling
e. Evaluation
f. Deployment

We will primarily focus only on steps 2 – 5 for this project.

a. **Data Understanding:** On observing the data at first instance, the good thing was none of the attributes had any missing values.
b. Data Preparation: The data seems simple and straightforward. Thus, not much data preparation was needed.
c. Modeling: The first operator used for modeling the data was assigning role to the dependent variable. Thus, we used a '**Set Role**' operator to assign label to Class which will help classify the wine maker. The next challenge here is converting the label data into acceptable format for 'Decision Tree' or 'K-NN' operators used for classification. We are thus using the '**Generate Attribute**' operator here. This operator is used to add an additional column or attribute. This attribute here will be used to convert the numerical data structure of Class to nominal with help of functional expression and store back to the new attribute. [1]

Furthermore, we will be needing the data stream to be split for training and testing. The '**Split data**' operator will be used by us for splitting the data into 70 percent for training and 30 percent for testing. The classification method we will be using here is K-NN (K Nearest Neighbor) which accepts the training data from the Split Data operator.

Lastly, we use '**Performance Classification**' operator to check the performance of our prediction. This operator accepts input from the '**Apply Model**' operator which is the labeled data from the 'Apply Model' in this case. We are using two criterions of the 'Performance Classification' to check monitor the performance namely, 'classification error' and 'accuracy'.

The performance output from the 'Performance Classification' operator connects to the result. In order to have a prediction versus actual data overview from the test dataset, we connect the examples from the performance operator to the result.

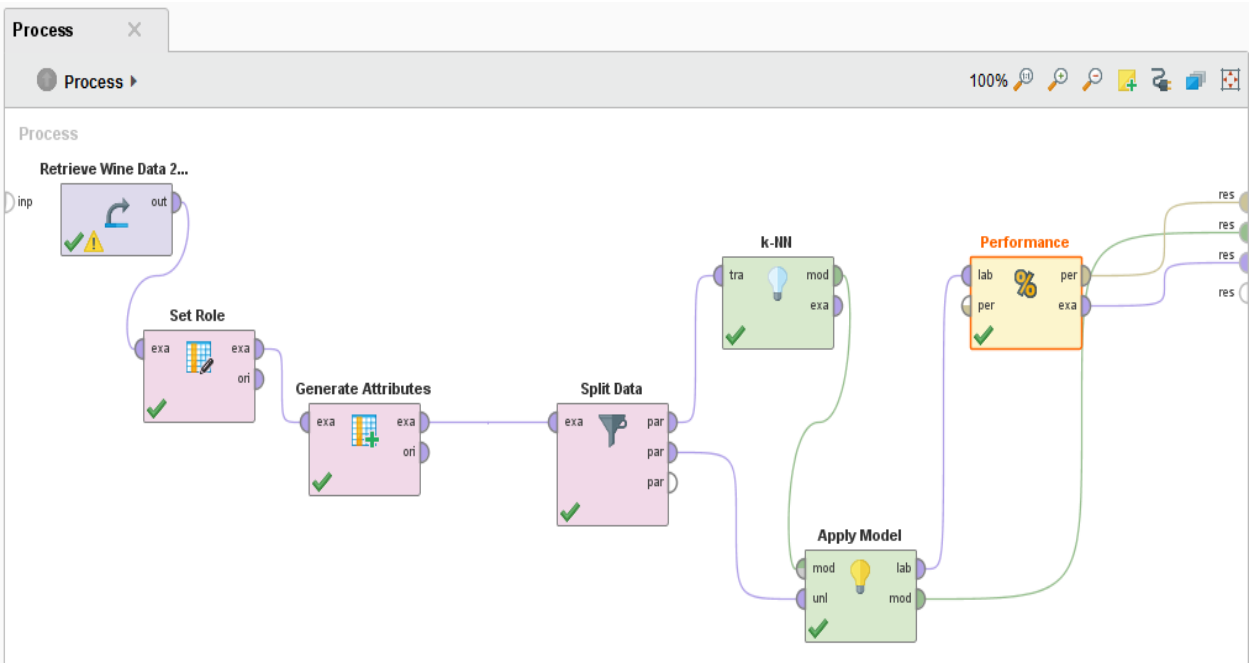

Figure 1.0: Process to classify the wine maker

d. **Data Evaluation:**

Results:

Table View Plot View

accuracy: 66.04%

	true 2	true 1	true 3	class precision
pred. 2	13	2	8	56.52%
pred. 1	1	16	2	84.21%
pred. 3	4	1	6	54.55%
class recall	72.22%	84.21%	37.50%	

Figure 2.0: Performance Vector

KNNClassification

```
1-Nearest Neighbour model for classification.
The model contains 125 examples with 13 dimensions of the following classes:
  2
  1
  3
```

Figure 3.0: KNN Classification

References:

[1] Usage of Excel, CSV and SPSS Data Files.

http://community.rapidminer.com/t5/RapidMiner-Studio/Usage-of-Excel-CSV-and-SPSS-Daa-Files/td-p/669

[2] Predicting quality of wine based on chemical attributes.

Author: Amelia Lemionet, Yi Liu and Zhenxiang Zhou

http://cs229.stanford.edu/proj2015/245_report.pdf