Program – 5

AIM: To perform pre-processing and use filters in Weka.

Introduction and Theory

Data preprocessing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: −100), impossible data combinations (e.g., Sex: Male, Pregnant: Yes), missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. Often, data preprocessing is the most important phase of a machine learning project, especially in computational biology.

If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data preprocessing includes cleaning, Instance selection, normalization, transformation, feature extraction and selection, etc. The product of data preprocessing is the final training set.

Steps in Data Preprocessing

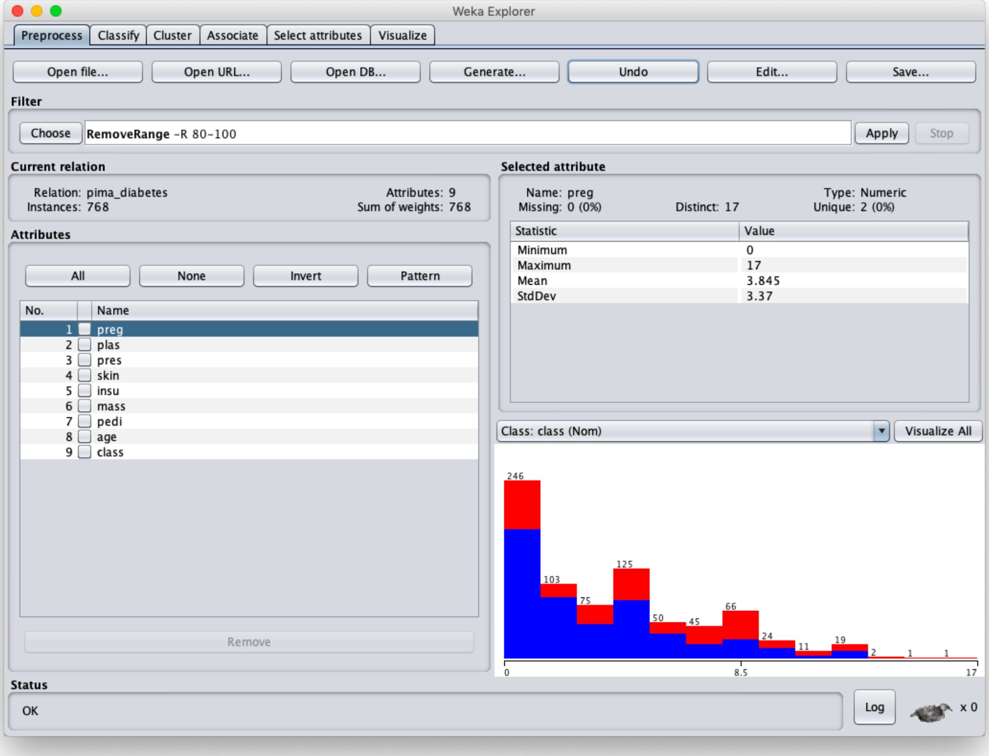
1. Import libraries
2. Read data
3. Checking for missing values
4. Checking for categorical data
5. Standardize the data
6. PCA transformation
7. Data splitting

Procedure

Pre-Processing

1. Open the Weka Explorer.
2. Load the Pima Indians onset of diabetes dataset.
3. Choose->unsupervised->instance and select any filter like RemoveRange
4. Select it then click the text box side to choose button.
5. Add range value in the dialog box
6. Click apply button.

Program – 5



Filter

RemoveRange -R 80-100

Current relation

Relation: pima_diabetes
Instances: 768
Attributes: 9
Sum of weights: 768

Selected attribute

Name: preg
Missing: 0 (0%)
Distinct: 17
Type: Numeric
Unique: 2 (0%)

Statistic	Value
Minimum	0
Maximum	17
Mean	3.845
StdDev	3.37

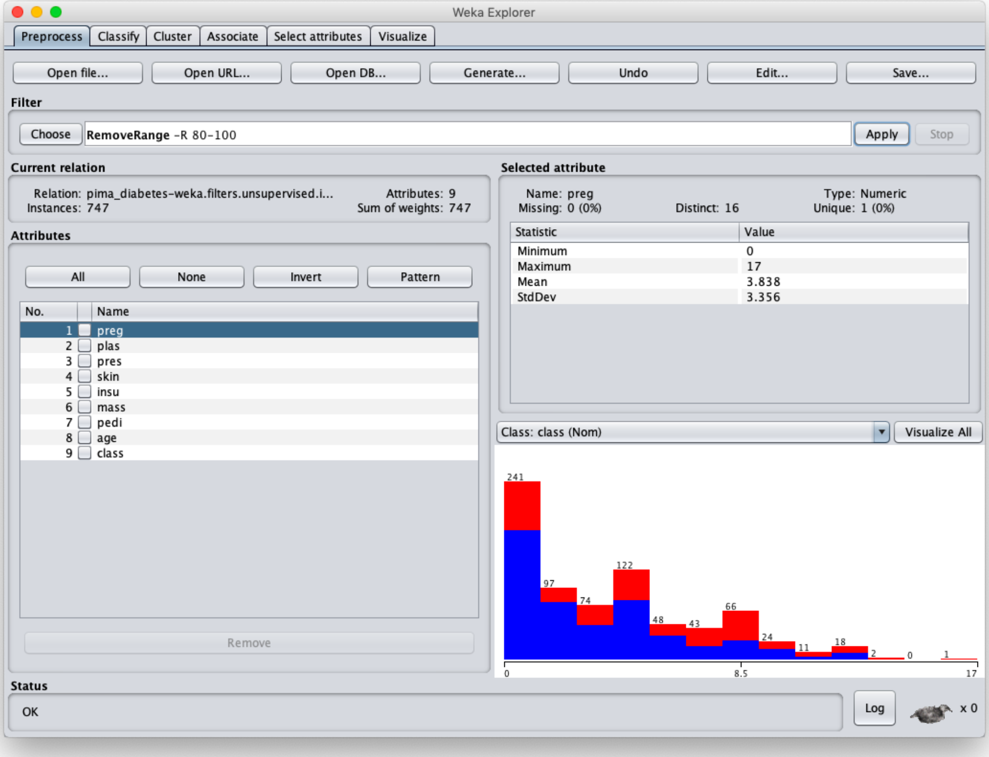
Attributes

No.	Name
1 □	preg
2 □	plas
3 □	pres
4 □	skin
5 □	insu
6 □	mass
7 □	pedi
8 □	age
9 □	class

Status

OK

Before remove_range



Filter

RemoveRange -R 80-100

Current relation

Relation: pima_diabetes-weka.filters.unsupervised.i...
Instances: 747
Attributes: 9
Sum of weights: 747

Selected attribute

Name: preg
Missing: 0 (0%)
Distinct: 16
Type: Numeric
Unique: 1 (0%)

Statistic	Value
Minimum	0
Maximum	17
Mean	3.838
StdDev	3.356

Attributes

No.	Name
1 □	preg
2 □	plas
3 □	pres
4 □	skin
5 □	insu
6 □	mass
7 □	pedi
8 □	age
9 □	class

Status

OK

After remove_range

Program – 5

Findings and Learnings:

1. Data preprocessing and filtering is an important step that needs to looked into before formal statistical modeling and analysis.
2. Weka Software provides a good interface and set of functions to preprocess and filter the data according to our needs.