Chapter 3

Methodology

3.1 Introduction

In this research an adapted version of KDD has been developed. KDD is reviewed in chapter 2. KDD is a general framework. Therefore, KDD steps that have no relation with the study in this thesis have been eliminated. These steps are (eliminated steps): Developing and understanding of the application domain in section 3.2; Preprocessing and creating operation database in section 3.3; Data mining in section 3.4 Interpreting mined pattern in section 3.5.

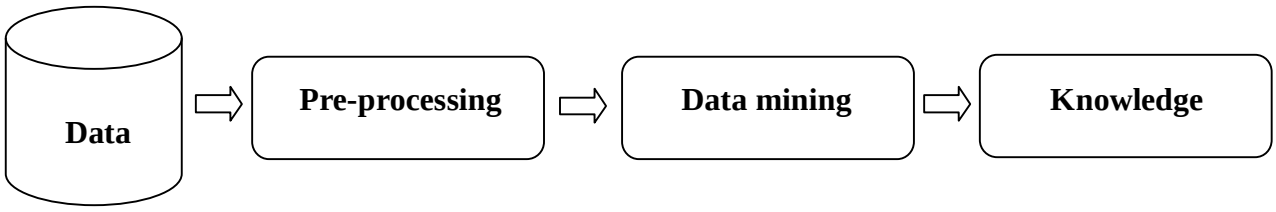

Figure 3.1: Knowledge discovery from data in data mining [17]

3.2 Developing an Understanding of the Application Domain

The primary goal of this step is to collect and gather all knowledge related to the educational area and the study in this thesis.

3.3 Preprocessing and Creating Operation Database

The primary goal of this step is to prepare the data and create operational database form the heterogeneous databases. Data have been used in our research is selected, constructing the data set and create the operational database that has been using in the following steps.

There are two datasets each one consist the grades of number of subjects:

1. Introduction to physics,
2. Special topics,
3. Digital logic,

4. Degrees processors,
5. Electricity and Magnetism,
6. Degrees process,
7. Nuclear physics.

The datasets were taken from students enrolled at x university during couple of academic years. The grades of 500 students were included in the dataset. The students' grades were described grades and number beside the attendance performance.

Database tables are:

- T-learning table: contains data about students who learned by traditional learning methods (all personal information needed about students) example of attributes: (Student number, Grades of the student …etc).
- E-learning table: contains data about students who followed or enrolled in e-learning cases or methods, example of attributes (Student number, Grades of the student …etc).
- To prepare this data in appropriate format several steps applied, these steps illustrate in the following:

i. All variables described above stored in one table in csv file format.

	no	Attendanc	Degree	l-type
2	1	9.6	47.4	E
3	2	2.9	72.9	E
4	3	9.3	74.6	E
5	4	3.3	53.8	E
6	5	1.9	39.1	E
7	6	5.6	44.8	E
8	7	6.2	76.1	E
9	8	6.2	85.2	E
10	9	4	47.1	E
11	10	0	65.1	E
12	11	6	37.4	E
13	12	1.2	22.8	E
14	13	2.6	49.8	E
15	14	2.4	87.8	E
16	15	6.6	65.2	E
17	16	0	65	E
18	17	4.2	70.5	E
19	18	4.1	80.8	E
20	19	8.2	94	E
21	20	3.1	27.5	E
22	21	6.8	95.7	E
23	22	8.2	80.5	E
24	23	7.7	82	E
25	24	3.8	35.1	E
26	25	3	47.9	E
27	26	4.1	60.3	E
28	27	2.3	69.9	E
29	28	0.9	22.4	E
30	29	1.7	46.1	E

Figure 3.2: Sample from Dataset

ii. Removed the duplicated attribute and select only the academically relevant attributes (such as id attribute in our case).

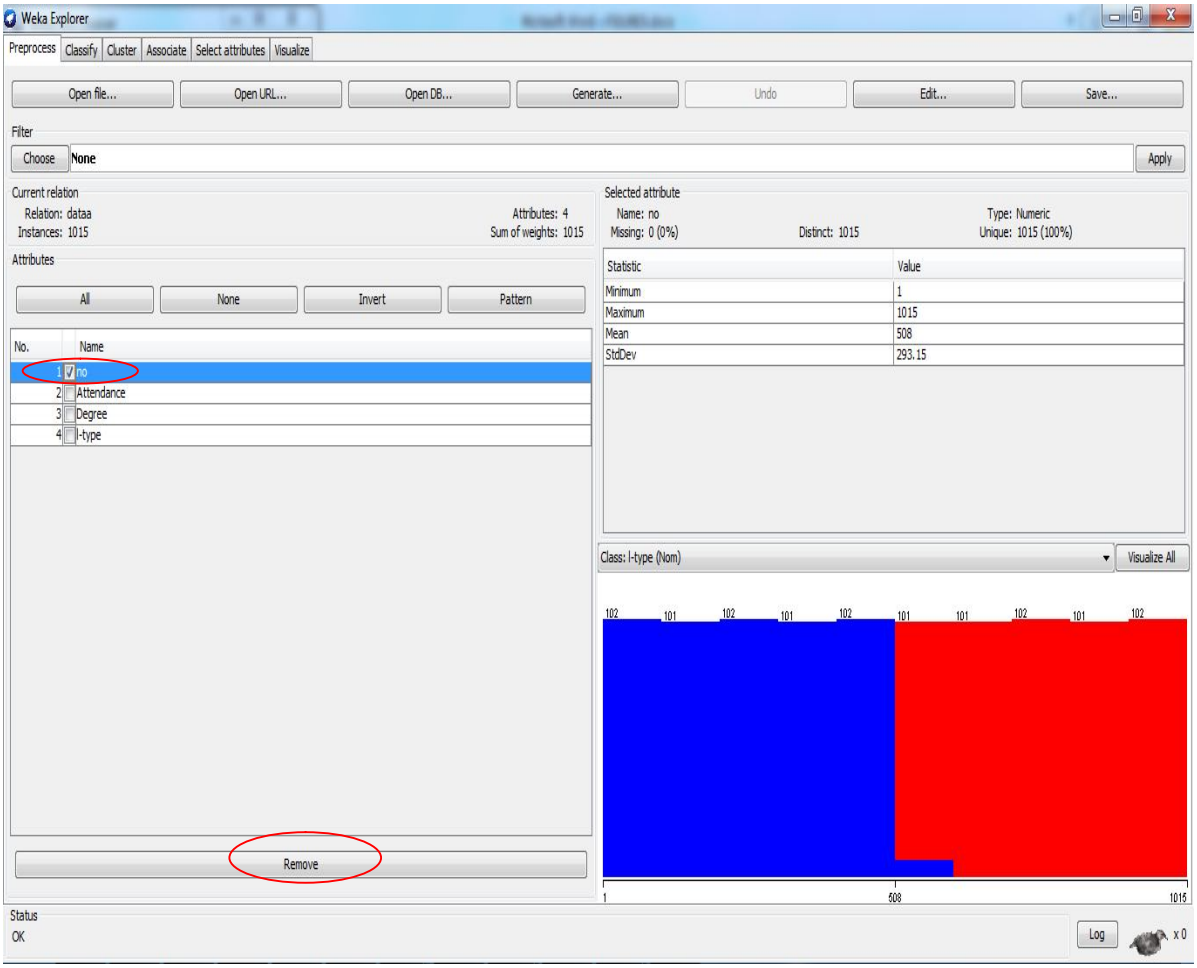

Figure 3.3: Remove irrelevant attributes

iii. Handle the records contains missing value, this was done using the most common value.

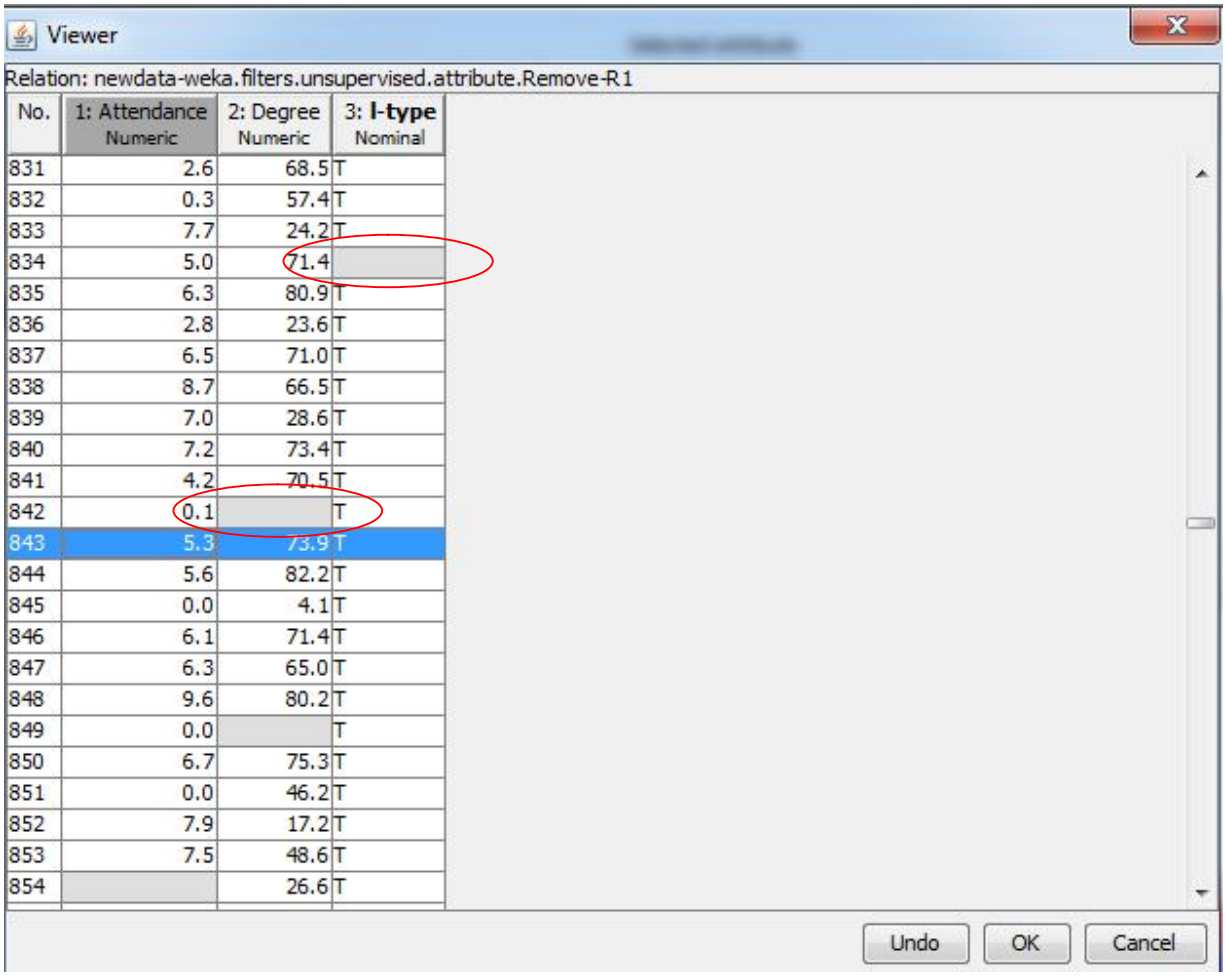

Figure 3.4: Handling Missing Values

iv. Transform the aggregate data, by getting the details data of the student from manual files which stored in the university database.

v. Add learning type record by applied the binning to summarized it. Table 3.1 illustrated this category.

E	Electronic- learning
T	Traditional-learning

Table 3.1: Normalize

Finally, after using these data cleansing and data preparation strategies, the final dataset, consisted of 3 variables (2 predictor variables "**Attendance degree, Student degree**" and 1 dependent variable "**Educational Type**") and 1500 records (instances).

Final Data sets description represent in following table.

S.NO	**Attribute Name**	**Description and class value**
1	Student number	Integer value represent the number of student file
2	Attendance degree	Integer value represent the student degree
3	Student degree	Integer value represent the student final degree
3	Educational Type	Categorical attribute, this data item identifies the type of learning.

Table 3.2: Dataset of case study Description

no	Attendance	Degree	l-type
1	9.6	47.4	E
2	2.9	72.9	E
3	9.3	74.6	E
4	3.3	53.8	T
5	1.9	39.1	T
6	5.6	44.8	E
7	6.2	76.1	E
8	6.2	85.2	E
9	4	47.1	E
10	0	65.1	E
11	6	37.4	E
12	1.2	22.8	E
13	2.6	49.8	E
14	2.4	87.8	E
15	6.6	65.2	E
16	0	65	E
17	4.2	70.5	E
18	4.1	80.8	E
19	8.2	14.8	E
20	3.1	27.5	E
21	6.8	95.7	E
22	8.2	80.5	E
23	7.7	28.5	E
24	3.8	35.1	E

Figure 3.5: Dataset CSV file

3.4 Data Mining

- Mining is core step of the KDD. There are many mining methods: classification, rule association and clustering. Clustering was used in this study to deal with the data that did not specify their relations from the beginning, and it partitioning large data sets into groups according to their similarity.
- There are many clustering techniques but we used K-mean in the thesis because it have simplest algorithms, popularity, flexibility, and it can deal with complex data, and the massive data, and considered as the major methods in clustering. The algorithm assembles multiple data depending on the characteristics of the K gathering, and the process of the assembly by reducing the distances between the data center and assembly (cluster centroid).
- Weka has been selected as a platform to perform the experiments.

3.5 Interpreting Mined Pattern

The results and the patterns mined are discussed in chapter 4.