Department of computer Engineering

Subject : DWM **Year/ Sem: TE/ V**

Experiment Number: 6

Aim: Demonstration of exploratory analysis such as missing values and data discretization using WEKA tool.

Objective:

The objective of this experiment is to demonstrate data analysis like handling the missing values and data discretization.

Theory:

In real world data, there are some instances where a particular element is absent because of various reasons, such as, corrupt data, failure to load the information, or incomplete extraction.

It is important to identify, mark and handle missing data when developing machine learning models in order to get the very best performance. Pre-processing tools in WEKA are called "filters". WEKA contains filters for discretization, normalization, resampling, attribute selection, transformation and combination of attributes.

These are some ways to handle the missing values

1. Ignore observations of missing values if you are dealing with large datasets and a smaller number of records has missing values.

2. Replace with the most frequent values

3. Apply classify algorithm to predict

4. Replace with the mean values

To remove the missing data: -

1. Open the Weka Explorer
2. Load the dataset.

Weka→filters→unsupervized→attribute→ReplaceMissingValues.

Data discretization refers to a method of converting a huge number of data values into smaller ones so that the evaluation and management of data become easy. In other words, data discretization is a method of converting attributes values of continuous data into a finite set of intervals with minimum data loss. There are two forms of data discretization first is supervised discretization, and the second is unsupervised discretization. Supervised discretization refers

to a method in which the class data is used. Unsupervised discretization refers to a method depending upon the way which operation proceeds. It means it works on the top-down splitting strategy and bottom-up merging strategy.

Weka→filters→supervised→attribute→Discretize

Conclusion :- Hence studied the handling missing values and discretization.