Practical:-5

Aim: To Study about WEKA Datamining Tool.

Waikato Environment for Knowledge Analysis (WEKA) is developed on the Java platform that contained a collection of machine learning and DM algorithms that widely used for data classification, clustering, association rule, and evaluation. The WEKA tool provides the interface that allows user to apply the DM methods directly to the dataset or user can embed their own programming Java code on WEKA to suit with their project. This tool also supports the variety file formats for mining include ARFF, CSV, LibSVM, and C4.5.

Weka is a collection of machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform. The algorithms can either be applied directly to a dataset or called from your own Java code. The original non-Java version of Weka was a TCL/TK front-end to (mostly third-party) modeling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and a Makefile-based system for running machine learning experiments. This original version was primarily designed as a tool for analyzing data from agricultural domains, but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research. Advantages of Weka include:

- Free availability under the GNU General Public License.
- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- A Comprehensive collection of Data preprocessing and modeling techniques.
- Ease of use due to its graphical user interfaces.

WEKA supports several standard data mining tasks, more specifically, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicted on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes. Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-dimensional data mining, but there is separate software for converting a collection of lined database tables into a single table that is suitable for processing using Weka. The primary available data such as census (2001), socio-economic data, and few basic information of Latur district are collected from National Informatics Centre (NIC), Latur, which is mainly required to design and develop the database for Latur district of Maharashtra state of India. The database is designed in MS-Access 2003 database management system to store the collected data. The data is formed according to the required format and structures.

Further, the data is converted to ARFF (Attribute Relation File Format) format to process in WEKA. An ARFF file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software. This document describes the version of ARFF used with Weka versions 3.2 to 3.3; this is an extension of the ARFF format as described in the data mining book written by Ian H. Witten and Eibe Frank

. After processing the ARFF file in WEKA the list of all attributes, statistics and other parameters can be utilized as shown in Figure 1.

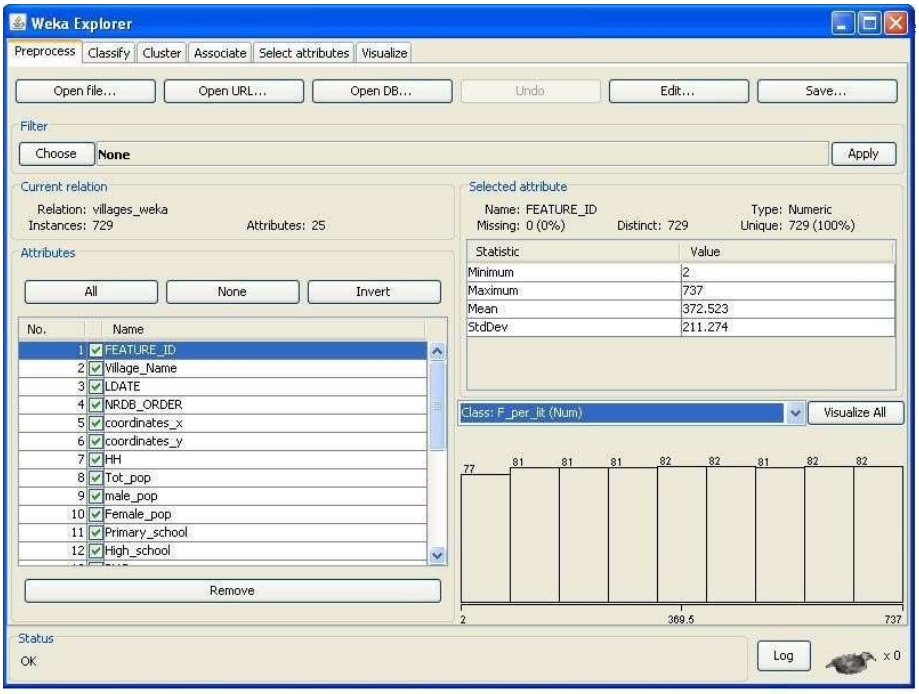

Fig.1 Processed ARFF file in WEKA

In the above shown file, there are 729 villages data is processed with different attributes (25) like population, health, literacy, village locations etc. Among all these, few of them are preprocessed attributes generated by census data like, percent_male_literacy, total_percent_literacy, total_percent_illiteracy, sex_ratio etc. The processed data in Weka can be analyzed using different data mining techniques like, Classification, Clustering, Association rule mining, Visualization etc. algorithms. The Figure 2 shows the few processed attributes which are visualized into a 2dimensional graphical representation.

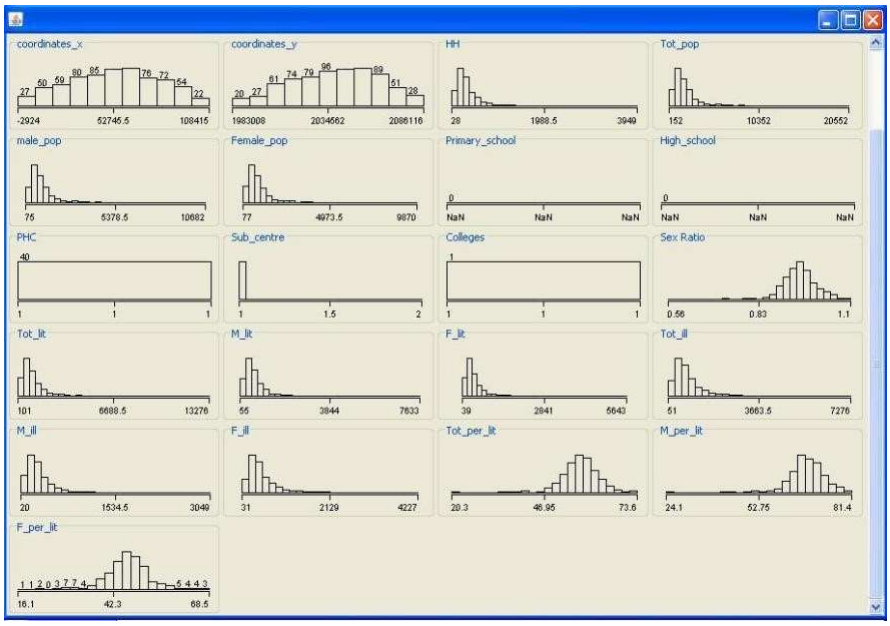

Fig. 2 Graphical visualization of processed attributes

The information can be extracted with respect to two or more associative relation of data set. In this process, we have made an attempt to visualize the impact of male and female literacy on the gender inequality. The literacy related and population data is processed and computed the percent wise male and female literacy. Accordingly we have computed the sex ratio attribute from the given male and female population data. The new attributes like, male_percent_literacy, female_percent_literacy and sex_ratio are compared each other to extract the impact of literacy on gender inequality. The Figure 3 and Figure 4 are the extracted results of sex ratio values with male and female literacy.

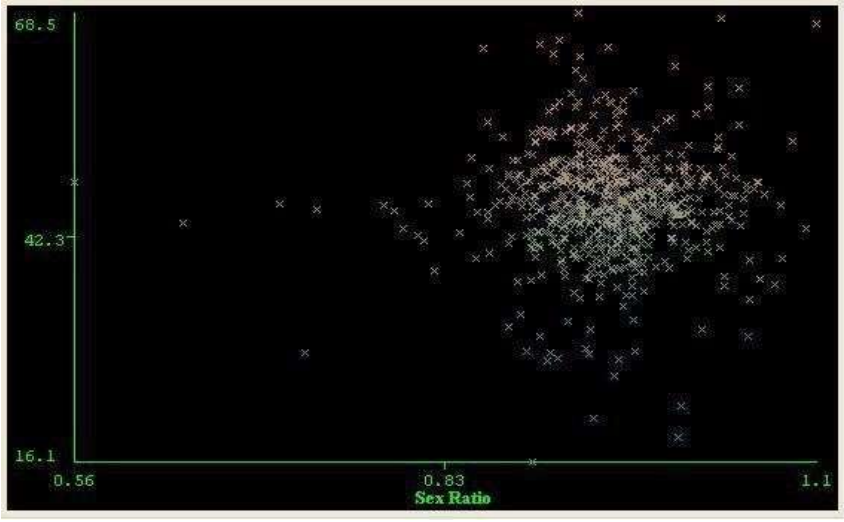

Fig. 3 Female literacy and Sex ratio values

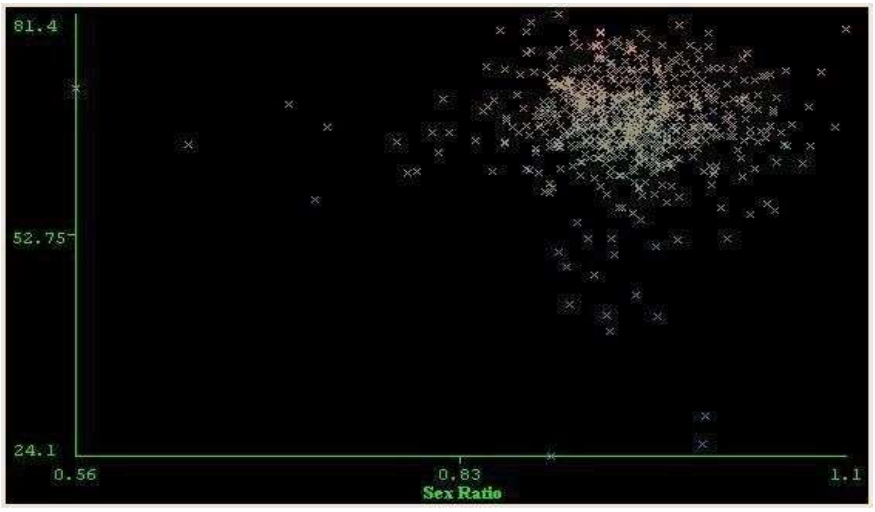

Fig. 4 Male literacy and Sex ratio values

On the Y-axis, the female percent literacy values are shown in Figure 3, and the male percent literacy values are shown in Figure 4. By considering both the results, the female percent literacy is poor than the male percent literacy in the district. The sex ratio values are higher in male percent literacy than the female percent literacy. The results are purely showing that the literacy is very much important to manage the gender inequality of any region.

CONCLUSION:

Knowledge extraction from database is become one of the key process of each and every organization for their development issues. This is not only important in the commercial industries but also plays a vital role in e-governance for future planning and development issues. This paper shows one of the small importance's of Weka to utilization and analysis for census data mining issues and knowledge discovery. It's for E-Governance.