CHAPTER 4

METHODOLOGY AND TOOLS

4.1 RESEARCH METHODOLOGY

In an effort to test empirically the suggested data mining technique, the data processing quality, it is important to find a real-world for effective data processing of database in application domain. In data mining technique, the application domain, data processing applications that have matured to become relatively stable. The data processing through data mining technique method chosen is mainly sophomore and finds the complexity of knowledge discovery data processing in computer applications. The generation process, existing data processing through data mining technique proposal approaches can be spitted into the following categories:

1. Content-based methods recommend for data processing using data mining technique to the user items similar to the ones the user chosen in the past relying on product features and textual item descriptions for processing of data.

2. Collaborative Filtering recommends to the user data processing items that people used with similar preferences knowledge discovery in data processing, liked in the past, taking the behavior, opinions and tastes of a large community of users into account.

In this thesis the performance of different data mining algorithm is evaluated empirically by conducting experiments using three different software tools viz. RapidMiner, Weka, and IBM SPSS statistics.

4.2 RAPID MINER TOOL

For a successful classification implementation, RapidMiner Studio 6 was used to perform experiments. RapidMiner is one of the world’s most widespread and
most used open source data mining solutions [16]. The project was born at the University of Dortmund in 2001 and has been developed further by Rapid-I GmbH since 2007. With this academic background, RapidMiner continues to not only address business clients, but also universities and researchers from the most diverse disciplines.

Fig. 4.1. RapidMiner User Interface

RapidMiner has a comfortable user interface (Fig. 4.1), where analyses are configured in a process view. RapidMiner uses a modular concept for this, where each step of an analysis (e.g. a pre-processing step or a learning procedure) is illustrated by an operator in the analysis process. These operators have input and output ports via which they can communicate with the other operators in order to receive input data or pass the hanged data and generated models over to the operators that follow.
Thus a data flow is created through the entire analysis process, as shown in fig. 4.2. The most complex analysis situations and needs can be handled by so-called super-operators, which in turn can contain a complete sub process. A well-known example is the cross-validation, which contains two sub processes.

A sub process is responsible for producing a model from the respective training data while the second sub process is given this model and any other generated results in order to apply these to the test data and measure the quality of the model in each case. A typical application is shown in fig. 4.3.

**4.3 WEKA**

Weka is the product of the University of Waikato (New Zealand) and was first implemented in its modern form in 1997[17]. It uses the GNU General Public
License (GPL). The figure of weka is shown in the figure 4.4. The software is written in the Java™ language and contains a GUI for interacting with data files and producing visual results.

![Weka GUI](image)

**Figure 4.4: Weka GUI**

It also has a general API, so we can embed WEKA, like any other library, in our own applications to such things as automated server-side datamining tasks. For working of weka we not need the deep knowledge of data mining that’s reason it is very popular data mining tool. Weka also provides the graphical user interface of the user and provides many facilities [4, 7].

### 4.4 IBM SPSS STATISTICS 20

SPSS was first developed in 1948 by social science researchers at Stanford University as a tool to help them with quantitative research[18]. In fact, the acronym SPSS initially stood for “Statistical Package for the Social Sciences”. As with “IBM” and “AT&T”, the company (and its software) is simply known by its initials, in part as a testament to its diverse user base. Although the software is most heavily used in social science contexts—particularly in psychology, political science and in academia— it is also used in medicine, marketing, and many other contexts. SPSS is appealing to many users from
less technical and/or mathematical disciplines because it has a particularly user-friendly interface (Figure 4.5) consisting of an Excel-like spreadsheet for the data and menus and buttons for manipulations and analyses. Although this “point and click” interface makes SPSS particularly attractive for statistical computing novices, individuals who require greater statistical functionality may find the application limiting.

Figure 4.5: IBM SPSS Statistics GUI

Between 2009 and 2010, the premier vendor for SPSS was called PASW (Predictive Analytics SoftWare) Statistics. The company announced on July 28, 2009 that it was being acquired by IBM. Versions 19.0 and 20.0 are named IBM SPSS Statistics[18].