ABSTRACT

In Knowledge Discovery, Data Mining is a relatively new concept. The past couple of decades have seen a dramatic increase in the amount of information or data being stored in electronic format. This accumulation of data has taken place at an explosive rate. It has been estimated that the amount of information in the world doubles every 20 months and the sizes as well as the number of databases are increasing even faster. There are many examples that can be cited. Point of sale data in retail, policy and claim data in insurance, medical history data in health care, financial data in banking and securities, are some instances of the types of data that is being collected.

This thesis is intended to evaluate the different data mining techniques empirically. 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. The classification algorithms k-nearest neighborhood, Naive Bayes, Decision Tree, Decision Stump and Rule Induction have their implemented source code in the RapidMiner Studio 6 version upon which experiments have carried out in order to measure the performance parameters of the algorithms over the datasets, their accuracy was evaluated. The decision tree algorithms BF Tree, FT Tree, J48 Tree, and LAD Tree were applied on the datasets. For a successful decision tree implementation, Weka 3.6.8 was used to aid the
investigation. The performance of neural network algorithms viz Multilayer Perceptron and Radial Basis Function was evaluated and compared using IBM SPSS Statistics software. The purpose of the experiments was twofold. The first aspect was to verify that existing data mining algorithm networks did in fact provide consistently better results. The second purpose was to investigate the effect of dataset variation on the performance of the algorithm. The performance of Association Rule Mining Algorithms viz Apriori and FP-Growth was evaluated and compared using Weka 3.6.8 software. Again the purpose of the experiments was twofold. The first aspect was to compare the performance in term of execution time, to find which algorithm is better then other. The second purpose was to investigate the effect of number of instance variation on the performance of the two algorithms.