Empirical Software Engineering (SE-404)

LAB A1-G2

Laboratory Manual

Department of Software Engineering

DELHI TECHNOLOGICAL UNIVERSITY(DTU)

Shahbad Daulatpur, Bawana Road, Delhi-110042

Submitted to: -

Ms. Priya Singh

Submitted by:-

Name-: ASHISH KUMAR

Roll number-: 2K18/SE/041

INDEX

S.No	EXPERIMENT	DATE	REMARKS
10.	Perform a comparison of the following data analysis tools. WEKA, KEEL, SPSS, MATLAB, R	04-01-2021	

Empirical Software Engineering LAB – A1 G2
EXPERIMENT 10

- ASHISH KUMAR
- 2K18/SE/041

Experiment Objective:- Perform a comparison of the following data analysis tools.

a. WEKA
b. KEEL
c. SPSS
d. MATLAB
e. R

Introduction:-

1. **WEKA:** Also known as Waikato Environment is a machine learning software developed at the University of Waikato in New Zealand. WEKA is a free and open source software package that assembles a wide range of data mining and model building algorithms.
   It is best suited for data analysis and predictive modeling. It contains algorithms and visualization tools that support machine learning. WEKA has a GUI that facilitates easy access to all its features. It is written in JAVA programming language and runs on almost any platform. WEKA supports major data mining tasks including data mining, processing, visualization, regression etc. It works on the assumption that data is available in the form of a flat file. WEKA can provide access to SQL Databases through database connectivity and can further process the data/results returned by the query.

[Source: screenshot of software]

WEKA supports four different frameworks for developing and executing the methods for DM tasks:

- Explorer: An environment for exploring data with WEKA.
- Experimenter: An environment for performing experiments and conducting statistical tests between learning schemes.
- KnowledgeFlow: This environment supports essentially the same functions as the Explorer but with a drag-and-drop interface. One advantage is that it supports incremental learning.
- SimpleCLI: Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface

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.

Disadvantages of WEKA include:

- It is not capable of multi-relational data mining.
- It can only handle small datasets. Whenever a set is bigger than a few megabytes an "Out Of Memory" error occurs.

2. **KEEL:** KEEL is a free software Java tool which empowers the user to assess the behavior of evolutionary learning and soft computing based techniques for different kind of data mining problems: regression, classification, clustering, pattern mining and so on. KEEL is a data mining tool used by many EDM researchers. For instance, KEEL has extremely extensive support for discretization algorithms, but has limited support for other methods for engineering new features out of existing features. It has excellent support for feature selection, with a wider range of algorithms than any other package. It also has extensive support for imputation of missing data, and considerable support for data re-sampling. KEEL is open-source and free for use under a GNU license.

   For modeling, KEEL has an extensive set of classification and regression algorithms; with a large focus on evolutionary algorithms. Its support for other types of data mining algorithms, such as clustering and factor analysis, is more limited than other packages. Support for association rule mining is decent, though not as extensive as some other packages.

[Source: screenshot of software]

The main features of KEEL are:

- It contains a large collection of evolutionary algorithms for predicting models, preprocessing methods (evolutionary feature and instance selection among others) and postprocessing procedures (evolutionary tuning of fuzzy rules). It also presents many state-of-the-art methods for different areas of data mining such as decision trees, fuzzy rule based systems or crisp rule learning.
- It includes around 100 data preprocessing algorithms proposed in the specialized literature: data transformation, discretization, instance and feature selection, noise filtering and so forth.
- It incorporates a statistical library to analyze the results of the algorithms.
- It comprises a set of statistical tests for analyzing the suitability of the results and for performing parametric and nonparametric comparisons among the algorithms.
- It provides a user-friendly interface, oriented to the analysis of algorithms.
- KEEL also allows creating experiments in on-line mode, aiming an educational support in order to learn the operation of the algorithms included.
- The software is aimed to create experimentations containing multiple datasets and algorithms to obtain results. Experiments are independently script-generated from the user interface for an off-line run in any machine that supports a Java Virtual Machine.

3. **SPSS:** SPSS stands for "Statistical Package for the Social Sciences". It is an IBM tool. This tool first launched in 1968. This is a software package which is mainly used for statistical analysis of the data.
SPSS is mainly used in the following areas like healthcare, marketing, and educational research, market researchers, health researchers, survey companies, education researchers, government, marketing organizations, data miners, and many others. It provides data analysis for descriptive statistics, numeral outcome predictions, and identifying groups. This software also gives data transformation, graphing and direct marketing features to manage data smoothly.

Features of SPSS are:

• The data from any survey collected via Survey Gizmo gets easily exported to SPSS for detailed and good analysis.
• In SPSS, data gets stored in .SAV format. These data mostly comes from surveys. This makes the process of manipulating, analyzing and pulling data very simple.
• SPSS have easy access to data with different variable types. These variable data is easy to understand. SPSS helps researchers to set up model easily because most of the process is automated.
• SPSS allows Opening data files, either in SPSS' own file format or many others.
• SPSS allows editing data such as computing sums and means over columns or rows of data. SPSS has outstanding options for more complex operations as well.
• SPSS has option for creating tables and charts containing frequency counts or summary statistics over (groups of) cases and variables.
• SPSS has a unique way to get data from critical data also. Trend analysis, assumptions, and predictive models are some of the characteristics of SPSS.
• SPSS is easy for you to learn, use and apply.
• It helps in to get data management system and editing tools handy.
• SPSS offers you in-depth statistical capabilities for analyzing the exact outcome.
• SPSS helps us to design, plotting, reporting and presentation features for more clarity.

Limitations are:
• SPSS is expensive to purchase for students.
• Usually involves added training to completely exploit all the available features.
• The graph features are not as simple as of Microsoft Excel.
• Documentation about algorithms is sometimes difficult or impossible to find.
• Information about effect size and confidence intervals is missing for many techniques.

4. **MATLAB:** MATLAB stands for Matrix laboratory. It was developed by Mathworks, and it is a multipurpose (or as we say it Multi-paradigm) programming language. It allows matrix manipulations and helps us to plot different types of functions and data. It can also be used for the analysis and design as such as the control systems.

Features of MATLAB:

- It is a high-level language for numerical computation, visualization and application development.
- It also provides an interactive environment for iterative exploration, design and problem solving.
- It provides vast library of mathematical functions for linear algebra, statistics, Fourier analysis, filtering, optimization, numerical integration and solving ordinary differential equations.
- It provides built-in graphics for visualizing data and tools for creating custom plots.
- MATLAB's programming interface gives development tools for improving code quality maintainability and maximizing performance.
- It provides tools for building applications with custom graphical interfaces.
- It provides functions for integrating MATLAB based algorithms with external applications and languages such as C, Java, .NET and Microsoft Excel.

Advantages of MATLAB are:

- Easy to use interface: A user-friendly interface with features you want to use is one click away.
- A large inbuilt database of algorithms: MATLAB has numerous important algorithms you want to use already built-in, and you just have to call them in your code.
- Extensive data visualization and processing: We can process a large amount of data in MATLAB and visualize them using plots and figures.
- Debugging of codes easy: There are many inbuilt tools like analyzer and debugger for analysis and debugging of codes written in MATLAB.
- Easy symbolic manipulation: We can perform symbolic math operations in MATLAB using the symbolic manipulation algorithms and tools in MATLAB

Disadvantages of MATLAB are:

- MATLAB is slow since it is an interpreted language that is MATLAB programs are not converted into Machine language but are run by external software, so it can sometimes be slow.
- We cannot create the OUTPUT file in MATLAB.
- One cannot use graphics in MATLAB with -nojvm option, on doing so, we will get a runtime error.
- We cannot make functions in one single .m file as we have in the case of other programming languages. We have to create different files for different functions.
- Sometimes, the error messages are not much informative, so you have to figure out the error yourself.

5. **R:** R is a popular and powerful open source programming language and software environment for statistical computing and graphics representation. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team. R implements various statistical techniques like linear and non-linear modeling, machine learning algorithms, time series analysis, and classical statistical tests and so on. R consists of a language and a run-time environment with graphics, a debugger, access to certain system functions, and the ability to run programs stored in script files. R allows integration with the procedures written in the C, C++, .Net, Python or FORTRAN languages for efficiency.

[Source: Wikipedia]

Features of R:

- R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.
- R has an effective data handling and storage facility,
- R provides a suite of operators for calculations on arrays, lists, vectors and matrices.
- R provides a large, coherent and integrated collection of tools for data analysis.
- R provides graphical facilities for data analysis and display either directly at the computer or printing at the papers.

Advantages of R:

1) Open Source: R is an open-source language. We can contribute to the development of R by optimizing our packages, developing new ones, and resolving issues.

2) Platform Independent: R is a platform-independent language or cross-platform programming language which means its code can run on all operating systems. R can run quite easily on Windows, Linux, and Mac.

3) Machine Learning Operations: R allows us to do various machine learning operations such as classification and regression. R is used by the best data scientists in the world.

4) Exemplary support for data wrangling: R allows us to perform data wrangling. R provides packages such as dplyr, readr which are capable of transforming messy data into a structured form.

5) Quality plotting and graphing: R simplifies quality plotting and graphing. R libraries such as ggplot2 and plotly advocates for visually appealing and aesthetic graphs which set R apart from other programming languages.

6) Statistics: R is mainly known as the language of statistics. It is the main reason why R is predominant than other programming languages for the development of statistical tools.

7) Highly Compatible: R is highly compatible and can be paired with many other programming languages like C, C++, Java, and Python. It can also be integrated with technologies like Hadoop and various other database management systems as well.

Disadvantages of R:

1) Data Handling: In R, objects are stored in physical memory. It is in contrast with other programming languages like Python. R utilizes more memory as compared to Python. It requires the entire data in one single place which is in the memory. It is not an ideal option when we deal with Big Data.

2) Basic Security: R lacks basic security. Because of this, there are many restrictions with R as it cannot be embedded in a web-application.

3) Complicated Language: R is a very complicated language, and it has a steep learning curve. The people who don't have prior knowledge or programming experience may find it difficult to learn R.

4) Weak Origin: The another disadvantage of R is, it does not have support for dynamic or 3D graphics. The reason behind this is its origin. It shares its origin with a much older programming language "S."

5) Lesser Speed: R programming language is much slower than other programming languages such as MATLAB and Python. In comparison to other programming language, R packages are much slower.

Learning from experiment:- We have successfully discussed and learnt various data analysis tools. We have compared 5 different analysis tools WEKA, KEEL, SPSS, MATLAB, R and mentioned their advantages and disadvantages as well.