World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:9, No:2, 2015

STATISTICA Software: A State of the Art Review

S. Sarumathi, N. Shanthi, S. Vidhya, P. Ranjetha

***Abstract*—**Data mining idea is mounting rapidly in admiration and also in their popularity. The foremost aspire of data mining method is to extract data from a huge data set into several forms that could be comprehended for additional use. The data mining is a technology that contains with rich potential resources which could be supportive for industries and businesses that pay attention to collect the necessary information of the data to discover their customer's performances. For extracting data there are several methods are available such as Classification, Clustering, Association, Discovering, and Visualization… etc., which has its individual and diverse algorithms towards the effort to fit an appropriate model to the data. STATISTICA mostly deals with excessive groups of data that imposes vast rigorous computational constraints. These results trials challenge cause the emergence of powerful STATISTICA Data Mining technologies. In this survey an overview of the STATISTICA software is illustrated along with their significant features.

***Keywords*—**Data Mining, STATISTICA Data Miner, Text Miner, Enterprise Server, Classification, Association, Clustering, Regression.

I. Introduction

FROM the past era, the data mining is a potent and effective novel technology that has been enhanced and also grown rapidly. Generally in database data mining is used to extract the formerly unidentified and potentially beneficial information from data. It is also said to be a part of knowledge discovery process. Data mining considered to a clever technique which can be used to extract their convenient patterns. They also contain analysis and prediction for collecting and handling data [1]. For extracting exact patterns and trends which already exists in data that generally uses complex algorithms and mathematical analysis. The foremost desire of data mining method is to create an efficient predictive and descriptive ideal for a massive amount of data. Numerous real world data mining difficulties include several skirmishing measures of enactment or objective where they want to be concurrently enhanced. The features of data mining usually deal with large and intricate datasets where its volume differs from gigabytes to terabytes. They entail some data mining processes and procedures to be sturdy, steady and scalable by means of all capabilities to work together with various research domains. The data mining tool like STATISTICA Data Miner, Text Miner and Enterprise Server can be developed due to the critical role played in each and every feature of information extraction in several tasks of data mining [2].

The STATISTICA Data Analysis plus Data Mining Platform includes the STATISTICA software that provides the complete and efficient system of user-friendly tools for the data mining process from querying databases to create final reports. Single workstation, multiple-user and Enterprise editions consist of StatSoft's data mining and predictive modeling software [3]. The STATISTICA software consists of STATISTICA Text Miner, Data Miner and Enterprise Analytics. Meanwhile from logical perception, the graphical interface used in the STATISTICA software leads to be more proficient, user friendly and at ease to work in where they are vastly favored by researchers and scientists [2].

Mrs.S.Sarumathi, Associate Professor, is with the Department of Information Technology, K. S. Rangasamy College of Technology, Tamil Nadu, India (phone: 9443321692; e-mail: rishi_saru20@rediffmail.com).

Dr.N.Shanthi, Professor and Dean, is with the Department of Computer Science Engineering, Nandha Engineering College, Tamil Nadu, India (e-mail: shanthimoorthi@yahoo.com).

Ms.S.Vidhya, PG Scholar, is with the Department of Information Technology, K. S. Rangasamy College of Technology, Tamil Nadu, India (phone: 9443960666; e-mail: vidhyapsubramani@gmail.com).

Ms.P.Ranjetha, PG Scholar, is with the Department of Information Technology, K.S.Rangasamy College of Technology, Tamil Nadu, India (phone: 8344789117; e-mail: ranjetha04@gmail.com).

II. STATISTICA Tool

A. STATISTICA Text Miner

STATISTICA Text Miner is an extension of STATISTICA Data Miner which is ideal for translating formless text data into meaningful information [3]. It is particularly modeled as a common and open-architecture tool for mining formless information. The feature extraction or selection and other analytic tools existing in STATISTICA Text Miner and they are not applicable to text documents or else web pages, but also used to index, cluster, classify which consist of our analysis formless information like bitmaps imported as data matrices [4]. Enterprise Analytics offers an effective server-platform for offloading resource-intensive model building tasks, central configurations of analyses, models, report templates, queries and web browser-based or Windows workstation clients [3].

STATISTICA Text Miner [4] is specially planned for common and open-architecture tool for mining formless information. The feature extraction or selection plus other analytic tools existing in STATISTICA Text Miner are not valid to text documents or else Web pages other than we can be used to cluster, classify, index or else consist of analyses formless information like bitmaps introduced as data matrices and so on.,

- Accessing Documents
- Processing Documents
- Analyzing Documents

1) Accessing Documents

The program includes huge choices for accessing text documents in dissimilar forms containing pdf, xml, html, txt, etc., Flexible user interface choices are offered for choosing

huge number of files through wild-cards. The program supports all Web-crawling abilities because documents can be separated from the Web and begins at a specific root Web page (URL). That the entire document connected to an exact page will be included and the documents connected to those sub-documents, till a user-specified level. In STATISTICA data files where File names plus URLs is placed in text variables. Likewise the program can not only perform original text kept in text variables, but correctly interpret references into text documents into text documents or as URLs. Hence numeric information and textual information that is big documents is stored on a per-case that is observation basics plus significant analyses is processed on data files anywhere for every observation numeric and formless textual information is available. Choices are offered to flexibly import like lists of file names or else URLs to the column of the STATISTICA spread sheet.

2) Processing Documents

Documents are pre-processed and prior to indexing all documents. Exclusion rules plus stub-lists are useful to take away general, but not words such as "a", "are", and “the, to". After that a stemming algorithm is applicable thus that English words such as "travelling", "travelled" together calculate as instances of "travel". STATISTICA Text Miner contains stub lists, plus stemming algorithms for all languages. A stub list is augmented through the user as required. The program is planned thus that support for extra languages that is added with the least effort. Then the program will index the stubbed-and-steamed documents that to develop a frequency adds up all words plus for all documents. This raw-data information is the base for all subsequent numerical analysis. By developing a STATISTICA Data File holds the counts to summarize the documents and several extra filters are applied. Examples like counts the most frequent words per documents can be normalized depend on the length of every document, transformed and optionally compressed. The outcome data file with numeric information like raw counts, most-frequent-word counts, etc., is prepared for further analyses. Several choices are given for writing the information extracted from text to the input data file or else straight to external databases.

3) Analyzing Documents

Every statistical analysis methods are useful to the numeric summaries denotes the texts. Simple summary statistics may take out the general word used in the documents. By mapping the documents to SVD dimensions, to calculate the similarity of documents, etc., by mapping documents to depend on transforming word counts, concurrent maps of documents and words are created. This reflects the "meaning" of documents. Clustering techniques are applied to find clusters of related documents. Predictive data mining techniques are used to compare the numerical summaries of documents into other indicators of interest. Key analytic components need extensive data processing are implemented through multi-threaded computing technology and to take out optimum performance from advance multiple-processor server hardware [4].

B. STATISTICA Enterprise Server

STATISTICA Enterprise Server [5] is an extremely scalable, completely Web-enabled data analysis, enterprise-class plus a database gateway application system that is developed in distributed processing technology plus supports multi-tier Client-Server architecture configurations. It represents the analytic, reporting, query and graphics functionality of STATISTICA by easy-to-use, standard Web or else Windows client interfaces. It allows users of the desktop version to offload computationally intensive analytics plus database operations into the Server. It is provided as a full, Internet browser-based user interfaces that enables users to interactively create data sets, etc., STATISTICA Enterprise Server is developed through open architecture plus contains .NET-compatible development kit tools that allow IT department personnel to modify every component of the system or else enlarge it by developing its foundations. It offered with Web browser-based user interface allowing us to specify analyses plus review outcomes. Tools are offered to customize these dialogs plus set up user interface or else to add up original functions. STATISTICA Enterprise Server applications insert a new dimension and a continuous array of possibilities to the whole line of STATISTICA Data Analysis, Quality Control or Six Sigma software and Data Mining. The System is friendly with every main Web server Software platform, works with Microsoft .NET and Sun or Java environments, plus does not need changes to the existing firewall plus Internet or Intranet security systems.

1) Additional Features of STATISTICA Enterprise Server

- A Broad Choice of Analytic Facilities and Configurations
- Two Common Categories of Web-based Analytics

a) A Broad Choice of Analytic Facilities and Configurations

The STATISTICA Enterprise Server is provided as a full solution that contains the analytic functionality of any STATISTICA product or else any mixture of products from STATISTICA Base to STATISTICA Data Miner applications.

b) Two Common Categories of Web-Based Analytics

- *Customer Web-Based Applications*

STATISTICA Enterprise Server support more customized Web-based analytic applications to set an organization's specific requirements. Users log on plus highly-targeted user interface customized for the exacting application requirements. Users have single-click access to the preferred collection of queries, reports and analysis outcomes were all viewed in their Web browser.

- *Server-Based Interactive Statistical Application*

STATITICA analytics is obtainable through the server based architecture and given that all of the advantages of client software to install, central configuration and continuing management, highly interactive user experiences and increased scalability. Example the most recent data and reports with choices is to interactively drill down to bring the outcomes and interactively attain additional and particular insights about the business.

C. STATISTICA Data Miner

STATISTICA Data Miner [6] can process, read, and write data from almost all standard file formats, they can access directly as well as a process or score databases (even without executing explicit import or effective export operation. It also delivers the most effective tools like data pre-processing, cleaning, and filtering for effective feature selection between thousands or millions of candidate predictors, options to merge multiple data sources, automatic optimal binning, outliers, deal with missing data, align data based on multiple criteria containing timestamps at unequal intervals (data aggregation), remove duplicate records, etc. They deliver effective wizards such as Data Miner Recipes to acquire the useful outcomes solutions quickly. They provide the conversant workspaces drag-and-drop interface to generate custom workflows, agree to interactive detailed drill down into particular intermediate and final results, and is completely programmable and customizable. STATISTICA Data Miner knows how to produce predictive models in many formats, including Java, PMML, C++ (C#) and other common programming or scoring languages such as SAS which has stored database procedures. The STATISTICA Rapid Deployment engine allows us to move directly from modeling to deployment as well as scoring of live data, databases, etc. STATISTICA Data Miner is totally integrated into the STATISTICA line of products, e.g., it is used to process optimization as well as advanced model-based process monitoring, automatic scoring of live data via STATISTICA Enterprise, etc. It is used as a desktop application or run in client-server architecture i.e. for server-based parallel processing of multiple analyses, with load balancing to maintain huge number of users, as well as options for scheduled batch. STATISTICA Data Miner contains the complete selection of data mining methods available on the market; e.g., by remote the complete selection of clustering techniques, neural networks architectures, association and sequence analysis (an optional add-on), classification or regression tress also called as recursive partitioning methods, multivariate modeling contains MARSplines, Support Vector Machines, and many other predictive techniques; even techniques for advanced or true simulation and optimization of models are delivered. The complete and effective system of user-friendly tools for the whole data mining techniques from querying databases is to create final reports. They offer the biggest selection of graphics and visualization methods of any competing products, to empower effective data exploration and visual data mining. It is a unique application in terms of its steep comprehensiveness, technology, power, and the flexibility of the accessible user interfaces. No other data mining application will get faster from "messy data" through the application of the advanced technologies and algorithms to actionable keys and knowledge [6].

Today, constantly increasing volumes of data are available where organization fights to make sense of it. Many projects like hard deadlines, shortage of highly skilled analysts frequently leave organizations where we will benefit from the dig data deluge.

STATISTICA Data Miner offers scientists, business analysts, data scientists, engineers, statisticians Ph.D.'s along with an intuitive statistics, data mining and predictive analytics software solution to allow us to make most of our big data. STATISTICA Data Miner provides more complete and inclusive set of data management and analytics packages in the industry in addition build-simplicity. More over 16,000 functions every access from one general user interface where it constantly position the easiest to use platforms on the market.

Streamline plus automate the data mining with rapid deployment, data mining recipes, customizable. STATISTICA Data Miner is a flexible, open, extensible software solution which makes use of industry-standard interfaces plus scripting languages. Determine hidden patterns and complex relationships among data to know what happened and why it happened to predict the future and optimize the possible results [7].

D. Data Mining Methods

STATISTICA Data Miner includes wide-ranging implementations of boosted tree, random forests for classification as well as regression problems, support vector machines, automated neural network searches, several clustering techniques, k-nearest neighbors, Kohonen networks, generalized linear models, partial least squares (PLS), algorithms for effectual association and sequence analysis for transactional databases. Techniques are provided for automatic competitive evaluation of the model, to calculate average prediction across models and so on. Apply advanced non-normal as well as multivariate simulation and optimization to final data mining techniques, for example for manufacturing or process optimization, campaign optimization, etc. From quality logging, integrated diverse methods and technologies are converted into the data mining projects and process Weibull analysis, power analysis, capability analysis, linear as well as nonlinear models. All STATISTICA procedures are used for doing projects in data mining and no programming or custom-development work is necessary to use these methods. All methods in the STATISTICA software can be scripted using the in-build STATISTICA Visual Basic macro or scripting environment and scripts can leverage third-party libraries and application, for example might include algorithm available in the most popular R language.

1) Visual Data Mining

All of STATISTICA's unique as well as unmatched graphical capabilities are available for data mining depends on the final results, tables, original data, or intermediate derived data. Apply zooming, highlighting, brushing through many graphs. Select from hundreds of graph kinds to visualize data after slicing, cleaning or drilling down. Produce graphical summaries and comparisons even for extremely huge data sources.

2) Data Access

STATISTICA Data Miner has the capability to handle or process concurrently several data streams in a single process

such as align data, merge data, aggregate data. It is optimized for processing extremely huge data sets which includes distinctive options to pre-screen, even over a various variables or parameters and/or draw stratified or simple unsystematic samples of records using DIEHARD-certified unsystematic sampling procedures. It offers highly optimized read and write access to huge databases, containing the IDP (In-Place Database Processing) method that quickly read data asynchronously directly from remote database servers (by distributed processing if supported by the server), such as bypassing the essential to "import" data and generate a local copy where more information is available here. Nearly all industry standard file formats are imported as well as exported, it may include text, SAS, Excel, SPSS and most database formats. Many special database formats are supported such as OSI PI, which is used to apply data mining techniques for optimizing continuous procedures.

3) Data Preprocessing

STATISTICA Data Miner delivers possibilities for automatic identification, recoding of outliers, unusual observations, and sparse classes and so on. Efficient and effective automatic Feature Selection is provided to fast identify important variables such as input parameters, even among over a million candidate parameters. Several options are available for optimal binning of predictors values and groups. Program offers dedicated functionality to merge, aggregated multiple data sources, align, e.g., time-stamped data from transaction databases, batch processes, etc. (STATISTICA ETL). An efficient transformation language as well as editor is available, to process single-pass through the data transformations such as date, lagging and time operations, expressions using logic operators, text operations. It implements on open architecture, through unlimited automation options and support for custom extensions.

4) User Interfaces

STATISTICA Data Miner provides a choice of user interfaces as well as options to flexibly switch among them based on the process. It offers dedicated interfaces for developing analytic workflows, next standard predictive data mining recipes, or else for processing, interactive ad-hoc analyses concurrently with numerous data inputs or intermediate outcomes, as well as using concurrently any mixture of the hundreds of techniques and graphs available in the program. It offers an easy to use, drag-and-drop depends user interface for creating analytic workflows that can be used even by novices. The program contains easy to use wizard-like user interfaces, alternate to develop models next to a recipe of common procedures as well as best-practices such as Data Miner Recipes. It offers powerful, interactive data exploration tools, as well as with the selection of interactive, examining graphics-visualization tools available in products. All aspects of the STATISTICA Data Miner as well as STATISTICA solutions are programmatically accessible, through scripting from inside the application or else from other applications such as C++, C#, VB.NET, etc. Outcomes can be stored as reports such as MS Word documents, Excel spreadsheets, PDF documents, or in the STATISTICA Workbook format [6].

5) Deployment of Models

STATISTICA Data Miner contains several options for flexible deployment of prediction techniques. The program can create XML syntax based PMML (Predictive Models Markup Language) files for predictive classification, prediction, or else predictive clustering of big data files. Choices for deploying predictive techniques in C, C++, C#, SAS language code, JAVA, or as stored database procedures are also available (Version 9.1 and higher). It can also write predictive values, classification probabilities, cluster assignments as well as probabilities, classifications, prediction residuals and additional results directly into external databases for subsequent analyses, selection, etc., by the efficient IDP (In-Place Database Processing) technology for reading as well as writing information from and to external databases. Using deployment tool, Rapid Deployment, ROC curve and profit charts can be generated to calculate the deployment of data mining techniques.

6) Platform and Integration in Statistica Solutions

STATISTICA Data Miner integrates with all STATISTICA solutions. It can run either on a desktop computer or on a server. User interfaces are available for server-based processing based on web-browser-only. Desktop analysis such as data mining workflows, data miner recipes can be offloaded to a server for server-based processing as well as results are retrieved from the analysis. Server side, it has the advantages of the STATISTICA Enterprise Server client-server architecture, for equivalent computing, advanced load balancing, etc. when deployed as a slice of the STATISTICA Enterprise solution, data mining predictions (scoring) techniques can be directly stored in the secure enterprise repository of reports, analysis templates, etc., and issued to authorized users throughout the enterprise. No extra result provider can perfectly integrate advanced data mining and data analysis methods into a secure and role-based enterprise analysis system to develop advanced systems like process monitoring, scoring engines, etc.

E. Data Miner Workspace Tools

STATISTICA Data Miner provides the complete selection of statistical, exploratory, and visualization methods available on the market, containing leading edge and highly efficient neural network or machine learning and classification techniques. The comprehensive analytic functionality of STATISTICA is available for data mining, they are encapsulated in over 300 nodes that can be designed with a structured and customizable Node Browser and dragged into the data mining workspace. The specialized workspace templates for data mining are optimized for speed as well as efficiency and can be categorized into the following five areas [6].

- General Slicer/Dicer and Drill-Down Explorer
- General Classifier
- General Modeler/Multivariate Explorer
- General Forecaster
- General Neural Networks Explorer

1) General Slicer/Dicer and Drill-Down Explorer

A huge number of analysis nodes are provided for creating exploratory graphs, to calculate descriptive statistics, tabulations, etc. These nodes are connected to input data sources, or to all intermediate outcomes. A STATISTICA Drill-Down Explorer is available for interactively exploring our data by drilling down on selected variables as well as categories or ranges of values in those variables. For examples, we can drill-down on Gender, to show the distribution for a variable Income for female only; next we could drill down on a specific income group, to create graphical summaries for preferred variables, for females in the chosen income group only. The characteristic of STATISTICA Drill-Down Explore is the capacity to select and deselect drill-down variables as well as groups in any order; next we could deselect variable Gender and thus shows selected graphs and statistics for the selected income group, for both males as well as females. Another feature is its range of slicing methods. Hence, they provide flexibility for slicing-and-dicing the data. It can be useful for raw data, database connections for in-place processing of data in remote databases, or to any intermediate outcome computed in a STATISTICA Data Miner project.

2) General Classifier

STATISTICA Data Miner provides the broadest selection of tools to process data mining, classification methods available on the market, containing classification trees, general CHAID models, generalized linear models, general classification and regression tree modeling (GTrees), cluster analysis techniques and general discriminate analysis models. Advanced Neural network classifiers available in STATISTICA Neural Networks are available in STATISTICA Data Miner and can be used in conjunction or competition with another classification method.

Deployment: The program containing options for generating C#, C++, C, STATICTICA Visual Basic, or else the PMML code for deployment of final solutions in your programs. Techniques are available for deployment after training, connect new data to the deployment node to calculate predicted classifications.

3) General Modeler/Multivariate Explorer

STATISTICA Data Miner provides the broadest selection of tools to develop deployable data mining techniques depends on neural network, linear, or nonlinear as well as tools to explore data. The user can develop predictive models depend on general multivariate methods. STATISTICA offers the whole range of techniques, as of linear and nonlinear regression models, generalized additive models and advanced generalized linear, regression trees and CHAID, to advanced neural networks techniques and multivariate adaptive regression splines (MARSplines).

STATISTICA Data Miner also contains methods that are not commonly found in data mining software, such as partial least squares methods for feature selection, to reduce the number of variables, survival analysis for analyzing data containing censored observations, structural equation modeling methods to develop and evaluate confirmatory linear models, factor analysis and multidimensional scaling for exploring structure in a huge number of variables, etc.

Deployment: The program containing options for generating C#, C++, C, STATICTICA Visual Basic, or else the PMML code for deployment of final solutions in your programs. Techniques are available for deployment after training, connect new data to the deployment node to calculate predicted values.

4) General Forecaster

STATISTICA Data Miner contains a wide selection of traditional that is non-neural networks-based and forecasting techniques contains an exponential smoothing with seasonal components, seasonal decomposition, ARIMA, Fourier spectral decomposition, polynomial lags analysis, etc., and neural network methods for time series data.

Deployment: Forecast can compute for numerous models in data mining projects, as well as plotted on a graph for comparative evaluation. For example, you can calculate as well as compute predictions from numerous ARIMA models, dissimilar methods for seasonal and non-seasonal time-series neural network architectures. This tool covers the most comprehensive selection of neural networks techniques available on the market. This component of STATISTICA Data Miner provides tools to approach virtually any data mining problem, including powerful forecasting, classification, and hidden structure detection). Features of the NN Explore is the selection of intelligent problem solvers as well as automatic wizards that uses Artificial Intelligence techniques to evaluate the demanding problems involved in the advanced NN analysis. The Explorer provides the broadest selection of cutting-edge NN architectures as well as procedures and highly optimized algorithms that contain radial basis function networks, multilayer perception, probabilistic neural networks, self-organizing feature maps, generalized regression neural networks, principle component network, linear models, and cluster networks. Networks bands of these architectures can also be calculated. Estimation methods contain conjugate gradient decent, Levenberg-Marquardt, back propagation, quasi-Newton, LVQ, delta-bar-delta, pruning algorithms, quick propagation, and more; choices are available for cross validation, sensitivity analysis, bootstrapping, sub sampling, etc.

5) General Neural Networks Explorer

This tool covers the comprehensive selection of neural network methods obtainable on the market. This component of STATISTICA Data Miner provides tools to approach virtually any data mining problem containing hidden structure detection, classification, and powerful forecasting. Features of

the NN Explorer are the selection of intelligent problem solvers as well as automatic wizards that make use of Artificial Intelligence techniques to solve the demanding problems involved in the advanced NN analysis. The Explorer provides the broadest selection of cutting edge NN architecture as well as procedures and highly optimized algorithms that contain multilayer perception, probabilistic neural networks, radial vital task networks, generalized regression neural networks, linear models, cluster networks, self-organizing feature maps, and principal component network. Network bands of this architecture can be calculated. Estimation techniques contains Levenberg-Marquardt, delta-bar-delta, back propagation, quasi-Newton, LVQ, conjugate gradient decent, quick propagation, pruning algorithms, and more choices are available for bootstrapping, sensitivity analysis, cross validation, sub sampling, etc.

Deployment: STATISTICA Neural Networks contain code generator options to produce C++, C, C# and STATISTICA Visual Basic code for one or more trained network and ensembles of networks. This policy can be rapidly included in our custom deployment programs. Additionally, trained neural networks as well as ensembles of neural networks can be stored, to be applied later for computing predicted responses or else classifications for new data. A deployment node can be hauled into the data miner workspaces to process prediction as well as predictive classification based on training neural networks, connect the data to deployment [6].

F. Data Mining Modules

1) Feature Selection and Variable Filtering

This module will mechanically choose subsets of variable from very large data files or else databases associated for in-place processing. The module can handle an infinite number of variables more than a million input variables is scanned to choose predictors for regression or else classification. Specially, the program contains various choices for choosing variables that are similar to be helpful or else informative in particular consequent analyses. The algorithm implements the Feature Selection and Variable Filtering module will choose continuous plus categorical predictor variables which explain a connection to the continuous or else categorical dependent variables of interest, in spite of whether that connection is simple or complex. Thus the program does not bias the choosing in support of any particular model that we can utilize to identify a final good rule, etc., for prediction or else classification. Several higher feature selection choices are obtainable. This module is mainly helpful in combination with the in-place processing of databases when it is used to examine large lists of input variables which choose expected candidates that have information related to the analyses of interest, plus automatically choose those variables for extra analyses with another node in the data miner project. Subset of variables depends on an early scan through this module can be accepted to further feature selection methods depends on neural networks, CHAID, MAR Splines, classifiers or linear regression. These choices permit STATISTICA Data Miner to handle data sets in the multiple giga and the terabyte range [6].

2) Association Rules

This module includes a full implementation of the Apriori algorithm for detecting association rules. The Association Rules module permits us to process quickly huge data sets for associations depends on pre-defined "threshold" values for detection. Especially the program will sense associations among exact values of categorical variables in huge data sets. This is a genera task in various data mining projects useful to databases, maintaining records of customer transactions. So the program will handle efficiently large analysis process [6].

3) Interactive Drill-Down Explorer

A primary step of data mining projects is to search the data to attain a first impression of the kinds of variables in the analyses plus their possible associations. The function of the interactive Drill-Down Explorer is to offer their relationships to other variables, exploratory data analysis, etc. [6].

4) Generalized EM & K-Means Analysis

The STATISTICA Generalized Expectation Maximization and k-Means Clustering module is an expansion of the techniques available in the common STATISTICA Cluster Analysis options where specially planned to handle huge data sets, plus to permit clustering or continuous or categorical variables plus to offer the functionality for full unsupervised clustering for pattern recognition with complete deployment choices for predictive clustering. Several cross-validation options are offered that will select and examine a most excellent solution for the clustering problems. They do not need to mention the number of clusters previous to an analysis instead program will use cross-validation based methods to select a number of clusters. The advanced EM techniques presented in this module are referred to as probability-based clustering or else statistical clustering. Several cross-validation choices are offered to permit us to select and examine a final result for the clustering problem outcomes and graphs plus detailed classification statistics are calculated for every observation. These techniques are optimized to manage huge data sets, plus several outcomes are offered to make easy subsequent analyses by the assignment of observations to clusters. Also contain the choices for deploying cluster outcome for categorizing observations.

5) Generalized Additive Models (GAM)

The STATISTICA Generalized Additive Models services are an implementation of techniques development. The program will maintain continuous and categorical predictor variables. STATISTICA contains a comprehensive selection of methods for fitting nonlinear models to data like General Classification and Regression Trees, Nonlinear Estimation module, etc.,

6) General Classification and Regression Trees (GTREES)

In the General Classification and Regression Tree (CC&RT) Model is a recursive partitioning method used to divide case depends on a collection of predictor variables. Not like linear or nonlinear regression algorithm these modules will identify hierarchical decision rules to offer optimal

separation among observations with regard to a categorical or continuous criterion variable depend on divide into one or more categorical predictor variables. The common Trees module includes several extensions plus options that are not found in the implementation of this algorithm and useful for data mining applications. Additionally, standard analyses implement these methods in STATISTICA enable us to identify ANOVA or ANCOVA-like design with categorical and continuous predictor variables. ANOVA or ANCOVA-like predictor designs are specified though dialogs, etc. The command syntax is compatible across modules to apply identical designs to dissimilar analyses.

7) Support Vector Machines (SVM)

The SVM method performs classification plus regression tasks to build nonlinear decision boundaries. Because of these features the space boundaries were established. SVM can show a huge degree of flexibility in handling regression plus classification tasks of various complexities. STATISTICA SVM support RBF, linear, sigmoid, and polynomial. It offers an ability to handle imbalanced data. Cross-validation is well established techniques used for formative the value of the several model parameters between a collection of given values. A huge number of graphs plus spreadsheet is computed to examine the value of the fit and to help with the interpretation of outcomes. Automatic techniques for deployment of KNN results for regression plus classification prediction are offered.

8) Native Bayes

The Native Bayes classifier depends on the Bayesian Theorem plus mainly matched when the dimensionality of the input is large owing to its simplifying statement of independence between the predictors. The assumption of independence is Naive Bayes normally outcomes more sophisticated classification methods. The predictor variables are independent and not accurate where it does make simple the classification task. It permits the class conditional densities to be designed for every variable to reduce a multidimensional task to a number of one-dimensional tasks. The assumption does not appear to involve the posterior probabilities, mainly in regions near decision boundaries. The classification task is unaffected. STATISTICA supports categorical predictors plus provide various options for modeling numeric predictors to analysis. These contain gamma, normal, Poisson density functions, and log normal. STATISTICA offers automatic methods for deployment of Naive Bayes model.

9) K-Nearest Neighbours (KNN)

STATISTICA KNN is a memory based method. In contrast to other statistical methods needs no training. It divides into the category of prototype methods. They function on the intuitive idea that secure objects are similar to be in the similar category. In KNN predictions are depends on a collection of prototype. These methods handle huge datasets plus both categorical and continuous predictors. Cross-validation is a technique used to examine a model parameters that are not known. A huge number of graphs plus spreadsheet is computed to examine the value of the fit plus to help with the interpretation of outcomes. Automatic methods for use of the KNN result of regression or else classification, prediction are offered [6].

G. Features

1. STATISTICA Data Mining provides a wealth of choices and methods are not available in challenging products.
2. These characteristics can be critical to maximize ROI in a competitive environment.
3. STATISTICA Data Miner can be used by beginners, and provides automatic model builder and wizard-like "Data Miner Recipes," yet provides the most comprehensive selection of methods as well as techniques for experts to solve the difficult problems.
4. STATISTICA Data Miner is the versatile data mining tool available, provides the right tools to gain critical business or process insights quickly, and to act on those deploy for instant ROI.
5. STATISTICA Live Score is an optional data mining tool, offers an efficient method of deploying data mining techniques.
6. STATISTICA Data Miner use with large data files as well as with the modern version has been improved to speed computation time and increase scalability and performance.

III. Overview of STATISTICA Software

Table I determines the overview of the STATISTICA data mining software which is based on their system, client and server requirements and also its benefits. The main aspire of this overview is not to scrutinize that STATISTICA is the best data mining software, but to epitomize the usage standards and awareness of this software in various fields.

TABLE I
OVERVIEW OF STATISTICA SOFTWARE

Name of the Software	STATISTICA			
System Requirement	Windows XP	Windows Server 2003 & 2008	Windows Vista	Windows 7 & 8
Client Requirement	Windows XP	512 MB RAM (1 GB recommended)	500 MHz processor (2.0 GHz, 64-Bit, dual core recommended)	
Server Requirement	Windows Server 2008 R2 or later	2 GB RAM (8 GB recommended)	1.0 GHz processor (2.0 GHz, 64-Bit, dual core recommended)	2.5 GB disk space
Benefits	Easier to use with data mining methods and customizable templates	Simple to use huge set of integrated analytics	Customized, Open and flexible	For adopting and integrating it is cost-effective

IV. Conclusion

STATISTICA software is explained with their usage of numerous tasks in this paper. Every single sub task of STATISTICA software is likely to be a highly important underpinning process for well-organized information extraction. This obligatory paves the method for the development of STATISTICA software. This STATISTICA software has the outstanding graphical interface, wide technical paradigm, and inbuilt multipart algorithms where it is very beneficial for treating substantial quantity of data more precisely and legibly. Thus the major role of this survey is to enhance knowledge about the STATISTICA software plus its use in some industries which will be very beneficial for the readers and it meets the desires of STATISTICA software researchers to innovate advanced tools in future.

References

[1] Ms. Aparna Raj, Mrs. Bincy G, Mrs. T.Madhu. "Survey on Common Data Mining Classification Techniques". International Journal of Wisdom Based Computing, Vol: 2(1), 2012
[2] S. Sarumathi, N. Shanthi, S. Vidhya M. Sharmila. "A Review: Comparative Study of Diverse Collection of Data Mining Tools". International Journal of Computer, Information, Systems and Control Engineering Vol: 8 No: 6, 2014
[3] Statistica Data Miner (Online). Available at: http://www.statsoft.com/Solutions/Cross-Industry/Data-Mining
[4] Statistica Text Miner (Online). Available at: http://www.statsoft.com/Products/STATISTICA/Text-Miner
[5] Statistica Enterprise Server (Online). Available at: http://www.statsoft.com/Products/STATISTICA/Enterprise-Server
[6] Statistica Data Miner (Online). Available at: http://www.statsoft.com/Products/STATISTICA/Data-Miner
[7] Statistica Data Miner (Online). Available at: http://www.statsoft.com/Portals/0/Support/Download/Brochures/STATISTICA Data Miner.pdf

Mrs. S. Sarumathi received B.E. degree in Electronics and Communication Engineering from Madras University, Madras, Tamil Nadu India in 1994 and the M.E. degree in Computer Science and Engineering from K. S. Rangasamy College of Technology, Namakkal Tamil Nadu, India in 2007. She is doing her Ph.D. programme under the area Data Mining in Anna University, Chennai. She has a teaching experience of about 16 years. At present she is working as Associate professor in Information Technology department at K.S. Rangasamy College of technology. She has published 5 reputed International Journals and two National journals. And also she has presented papers in three International conferences and four national Conferences. She has received many cash awards for producing cent percent results in university examination. She is a life member of ISTE.

Dr. N. Shanthi received the B.E. degree in Computer Science and Engineering from Bharathiyar University, Coimbatore, Tamil Nadu, India in 1994 and the M.E. degree in Computer Science and Engineering from Government College of Technology, Coimbatore, Tamil Nadu, and India in 2001. She has completed the Ph.D. degree in Periyar University, Salem in offline handwritten Tamil Character recognition. She worked as a HOD in department of Information Technology, at K. S. Rangasamy College of Technology, Tamil Nadu, India since 1994 to 2013, and currently working as a Professor & Dean in the department of Computer Science and Engineering at Nandha Engineering College Erode. She has published 29 papers in the reputed International journals and 9 papers in the National and International conferences. She has published 2 books. She is supervising 14 research scholars under Anna University, Chennai. She acts as the reviewer for 4 international journals. Her current research interest includes Document Analysis, Optical Character Recognition, and Pattern Recognition and Network security. She is a life member of ISTE.

Ms. S. Vidhya holds a B.Tech degree in Information Technology from N.P.R College of Engineering and Technology, affiliated to Anna University of Technology, Tiruchirappalli, Tamil Nadu, India in 2013. Now she is an M. Tech student of Information Technology department in K. S. Rangasamy College of Technology. She has published 3 international journals and presented three papers in National level technical symposium. Her Research interests include Data Mining, Wireless Networks. Most of her current work involves the development of efficient cluster ensemble algorithms for extracting accurate clusters in large dimensional database.

Ms. P. Ranjetha received B.Tech degree in Information Technology from K. S. Rangasamy College of Technology, affiliated to Anna University Chennai, Tamil Nadu, India in 2014. Now she is an M.Tech student of Information Technology department in K. S. Rangasamy College of Technology. She has presented two papers in National level technical symposium. Her Research interests include Mining Medical data, Opinion Mining and Web mining.