RESEARCH PAPER

SHIVALI SHARMA

EXECUTIVE SUMMARY

Data mining is a process which is used by organizations or entities to turn raw data into some useful information. It is the method of finding patterns in large sets of data which includes methods at the intersection of machine learning, database systems and statistics. Data mining approaches are a way to drive competency and predict client behavior. However, a business may emerge from other businesses in the industry using predictive analysis by using the data mining strategies accurately. Data classification involves categorizing data or useful information so that it can be used in the most efficient and effective way. Decision Trees are considered as one of most well-known methods of representing classifiers.

This paper focusses on the use of three data mining software that implements the decision tree induction techniques. The three-software covered in this paper are SAS Enterprise Miner, IBM SPSS Modeler, and Rapid Miner. The data mining software is evaluated based on four categories – functionality, performance, usability and ancillary tasks support. Each category is subcategorized into different criteria which are discussed in this paper.

TABLE OF CONTENTS

EXECUTIVE SUMMARY 1
INTRODUCTION 3
OVERVIEW ON DECISION TREE INDUCTION 4
EVALUATION CRITERIA 5
SAS Enterprise Miner 7
Functionality Criteria 8
Computational Performance Criteria 9
Ancillary Tasks Report Criteria 9
Usability Criteria 11
SPSS IBM Modeler 12
Functionality Criteria 14
Computational Performance Criteria 15
Ancillary Tasks Report Criteria 15
Usability Criteria 17
RAPIDMINER 18
Functionality Criteria 18
Computational Performance Criteria 20
Ancillary Tasks Report Criteria 20
Usability Criteria 21
COMPARITIVE ANALYSIS 22
CONCLUSION 24
REFERENCES 25

INTRODUCTION

Data Mining is a process where brilliant and intelligent methods are used to excerpt data patterns. Data mining enables entities to learn more about their clients or customers by looking for patterns in large data sets and develop more effective & efficient strategies. Data mining strategies are utilized as a part of many research areas, including artificial intelligence, mathematics, genetics, and advertising.

A decision tree is like a sequential or flow-chart structure where each internal node represents an inquiry on an attribute, each branch represents a result of the inquiry, and each leaf node denotes a class label.

The objective of the paper is to evaluate and analyze three data mining software that implements the decision tree induction techniques. This paper gives an insight as to how SAS Enterprise Miner, Rapid Miner, and SPSS IBM Modeler builds decision trees using algorithms, their features, advantages and disadvantages. The data mining software are also evaluated based on functionality, performance, usability, and ancillary tasks support. These evaluations or analysis are categorized into criteria such as algorithmic variety, data cleansing, user interface, platform variety, etc. This paper provides an in-depth comparative analysis of SAS Enterprise Miner, Rapid Miner, and SPSS IBM Modeler and should not be used for marketing, selling or buying decision making of the above-mentioned data mining software. This paper is inclusive of the limitations of decision tree induction such as Binning, missing values, and overfitting. This paper only covers Decision Tree analysis of the above-mentioned software and no other techniques. This paper is also limited to my knowledge of the three-mentioned software and thus lacks expertise in the field.

OVERVIEW ON DECISION TREE INDUCTION

Decision tree constructs or builds classification or regression models in the form of a tree structure. It separates a data set into smaller subset, while in the meantime a related decision tree is additionally built. The result is represented by a tree with decision and leaf nodes. The advantages of having a decision tree are as follows:

- They perform variable screening or feature selection
- They require relatively less effort from users for data preparation
- The decision tree's performance is not affected by nonlinear relationships between parameters.
- They are easy to comprehend

Decision tree is a crucial step towards distribution technique. Decision trees are generated by distribution or segmentation dwells in distribution of data or observation into various subset based on partitioning methods or parameters. Decision trees are used in various industries to determine the relationship amongst variables which helps a business in analyzing the data. For instance, decision trees can be used to predict the stock trends by extracting the features from the daily stock market data and then selecting all the relevant features using decision trees. A rough set based classifier is then used to predict the next day's trend using the selected features.

EVALUATION CRITERIA

SAS Enterprise Miner, Rapid Miner and SPSS IBM Modeler are evaluated based on four categories:

- Functionality
- Performance
- Usability
- Ancillary Tasks Support

The above-mentioned categories further have different criteria based on which the software is evaluated and analyzed.

Functionality

1. Algorithmic Variety: This criterion sees if the software provides an adequate variety of mining techniques and algorithms including neural networks, rule induction, decision trees, clustering, etc.
2. Prescribed Methodology: This criterion sees if the software aids the user by presenting a sound, step-by-step mining methodology to help avoid spurious results.
3. Model Validation: This criterion sees if the tool supports model validation apart from model creation and if the tool encourages model validation as a part of methodology.
4. Data Type Flexibility: This criterion considers if the implementation of supported algorithms can handle a variety of data types, continuous data without binning, etc.
5. Algorithm Modifiability: This criterion sees if the user can modify and fine-tune the modeling algorithms.
6. Reporting: This criterion makes sure if the results of a mining analysis are reported in a variety of ways, if the tool provides both summary and detailed results, and if the tool can select actual records that fit a target profile.
7. Model Exporting: This criterion sees if the tool provides a variety of ways to export the tool for an ongoing case, after the data is validated.

Performance

1. Platform Variety: This criterion sees if the software can run on a wide variety of computer platforms and a typical business user platform.

Usability

1. User Interface: This criterion sees if the interface is not complicated and can be navigated easily, or the results presented by the interface are meaningful.

Ancillary Tasks Support

1. Data Cleansing: Data cleansing criterion considers if the tool allows user to modify false values in the data set or perform other data cleaning processes effectively.
2. Data Filtering: This criterion checks if the tool allows the selection of subsets of the data based on user-defined selection criteria.
3. Deriving Attributes: This is criterion sees if the tool can derive attributes based on inherited attributes, and if a wide variety of methods are available for the derivation of attributes.
4. Randomization: This criterion considers if the tool allows randomization of data before data modeling and sees if randomization is effective and efficient.
5. Record Deletion: This criterion sees if the tool allows the user to delete records which may be incomplete or may bias the results, or if the user can delete records from entire segments of population, or if the deleted records can be incorporated easily if needed.

SAS Enterprise Miner

SAS Enterprise Miner is an advanced analytics data mining tool that assists users to establish predictive and descriptive models effectively and efficiently through a streamlined data mining process. SAS Enterprise Miner's graphical interface allows users to move logically through the SAS SEMMA approach, which is sampling, exploration, modification, modeling, and assessment.

- **Sample**: Data Sampling involves selection of data set to create a model. Data partitioning is a part of this phase.
- **Explore**: This phase deals with data visualization where expected and unexpected relationships are determined amongst the variables which helps in understanding the data.
- **Modify**: Data is modified by creating, selecting and transforming the variables to prepare for data modeling.
- **Model:** Modeling phase emphasizes on applying of different data mining techniques to find the best possible result.
- **Assess**: This phase focuses on determining the efficiency and effectiveness of the results.

Advantages of SAS Enterprise Miner:

- It is flexible and can cater to complex or large data.
- It can be used by analytical and business users since the algorithms used in SAS strengthens the accuracy and stability of the predictions.
- Since it's graphical user interface is easy to use, the users can build models faster.

Functionality Criteria

Algorithmic Variety:	SAS Enterprise Miner supports for various algorithms such as Decision Trees (CHAID, CARD and C4.5), DM Neural Mode, Memory Based Reasoning (k – nearest neighbor). These are used based on the requirements of situation.
Prescribed Methodology	SAS Enterprise Miner follows the SEMMA approach for Data Mining Processes.
Model Validation	SAS Enterprise Miner allows a user to segregate the data into training, validation and testing. This segregation is based on the percentages provided by the user.
Data Type Flexibility	SAS Enterprise Miner is flexible to different data types. Users can import data from any file type into excel files and excel files can be used as raw data for analysis.
Algorithm Modifiability	SAS Enterprise Miner supports for setting up of target criteria for various nodes such as decision trees, Memory Based Reasoning, Support Vector Machine, etc.
Reporting and Model Exporting	Once the user has completed a process flow diagram, they can see their results after running the diagram. The results are presented in tabular and graphical format. Users can also share their SAS Enterprise Miner process flow diagrams by converting them into .xml files.

Computational Performance Criteria

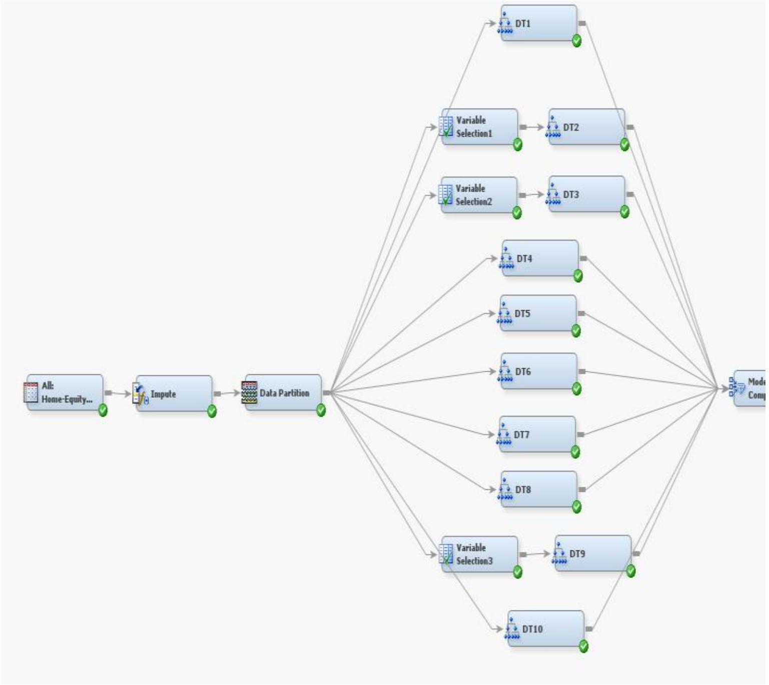
Platform Variety	SAS Enterprise Miner's client software functions on Microsoft Windows 7,8, and 10. Server host options includes: • HP-UX on Itanium 11i version 3 (11.31); • IBM AIX R64 on Power architecture 7.1; • IBM z/OS V1R12 and higher; • Linux x64 (64-bit), including SUSE Enterprise 11 SP1, Red Hat Enterprise Linux 6.1 and 6.7, and Oracle Linux 6.1; • Microsoft Windows on x64 (64-bit), including desktop Windows 7 x64 SP1, Windows 8 x64, and Windows 10 x64 or Server -- Windows Server 2008 x64 SP2 family, Windows Server 2008 R2 SP1 family, Windows Server 2012 family; • Solaris on Sparc version 10, update 9; and • Solaris on x64 (x64-x86): version 10, update 9; version 11

Ancillary Tasks Report Criteria

Data Cleansing	SAS Enterprise Miner provides for a replacement node which can either remove variables with missing data or calculate the missing data. If the user is using decision trees, missing values may be acceptable however, at the time of clustering the user can either exclude data with missing values or replace the missing values with data using Impute node. Impute node can replace data using any of the imputation methods such as Mean of the Nearest Cluster, Seed of Nearest Cluster, and Conditional Mean.

Data Filtering	SAS Enterprise Miner supports for data filtering using the filter node. The filter can be applied to a dataset to ignore a few outliers or any other observation.
Randomization	This can be done by using the sample node in SAS Enterprise Miner. A user can extract random sample using the sample node from a given dataset. Sometimes, randomization becomes very important to validate the results of an analysis. It could be used for either neutralizing any biasness or provide a base for the assumption for analysis.
Deriving Variables	Users can use Transform Node to derive variables in SAS Enterprise Miner. Occasionally, users may have to transform data for it become more informative. Transformation can be done for one or more variables. For instance, variable transformation might be used for correct non-normality, stabilize variance, improve additivity, etc.
Record Deletion	SAS Enterprise Miner does not support for record deletion. Users may exclude observations from a dataset using filter node.

Usability Criteria

User Interface	SAS Enterprise Miner has a user-friendly interface where users can select appropriate tabs from the Enterprise Miner’s toolbar and build process flow by dragging and dropping specific nodes. Below is an example of a process flow diagram.  DT1 Variable Selection1 DT2 Variable Selection2 DT3 DT4 DT5 DT6 AlI: Home-Equity... Impute Data Partition Mode Comp DT7 DT8 Variable Selection3 DT9 DT10

SPSS IBM Modeler

SPSS IBM Modeler is a data mining and text analytics software from IBM which is used to create predictive models and conduct analytical tasks. This application brings predictive intelligence to everyday business problems. SPSS IBM Modeler lets users to take advantage of its visual interface to perform data mining actions without any programming language. SPSS IBM Modeler aims at making complex predictive models easier to understand and use. The first version of SPSS included decision trees (ID3) and neural networks (backprop). It follows the Cross Industry Standard Process for Data Mining, also known as CRISP -DM, approach:

- **Business Understanding:** The first phase of this methodology emphasizes on understanding the business objectives and requirements, and defining the data mining objective.
- **Data Understanding:** The second phase begins with data collection, identification of problems with the data, and developing of hypothesis based on the overview of the data set.
- **Data Preparation:** The third phase includes all the activities to create the final dataset, which is inclusive of cleaning, derivation and transformation of data. The final data is prepared for constructing a model.
- **Modeling:** In the modeling phase, the user applies various parameters and techniques to get the best possible result.
- **Evaluation:** This phase begins when the model is created, and the user has results. The user analyses the results and sees if the business and data mining objectives were met after this phase.
- **Deployment:** Sometimes, the creation and evaluation of a model is not enough. The data may have to be presented to the clients. Deployment phase can begin from generation of reports or implementation of the data mining process.

Advantages of SPSS IBM Modeler:

- It offers expansive range of capabilities and algorithms.
- It is easy to use. The users can be guided through the analytical process with the help of its drag and drop interface.
- It supports the data analysis from business understanding to deployment of the data mining technique.

Functionality Criteria

Algorithmic Variety:	SPSS IBM Modeler supports for classification models. These require one or more input variables to predict one or more output variable, also known as the target variable. Decision Trees, regression, neural networks, SVM, and Bayesian networks are some of the examples of the classification models.
Prescribed Methodology	It follows the Cross Industry Standard Process for Data Mining, also known as CRISP -DM, methodology and includes Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.
Model Validation	SPSS IBM Modeler uses Holdout method for the computation of accuracy and stability of a predictive model. The data set is segregated randomly into two or three independent sets, known as training, validation and testing.
Data Type Flexibility	SPSS IBM Modeler accepts string, integer, date, timestamp or real number as datatypes for their discrete, interval or categorical values.
Reporting and Model Exporting	SPSS IBM Modeler allows a user to see the results and a summary on a separate screen. Users can save and share their models in Excel, SAS, TM1 or extension export.

Computational Performance Criteria

Platform Variety	SPSS IBM Modeler supports the following: • Windows Server 2008 Standard Edition for 32-bit x86 systems; • Windows Server 2008 Enterprise Edition for 64-bit x64 systems; • Windows Server 2008 Enterprise Edition R2 for 64-bit x64 systems; • Windows Server 2003 Standard Edition R2 for 32-bit x86 or 64-bit x64 systems IBM AIX® 6.1 or 7.1 for 64-bit POWER systems; • Oracle Solaris™ 9.x or 10 for 64-bit SPARC systems; • Red Hat Enterprise Linux 5.x for 32-bit x86 systems; • Red Hat Enterprise Linux 5.x or 6.x for 64-bit x64 or IBM System z systems; • Red Hat Enterprise Linux Advanced Platform 5.x for 32-bit x86 systems; • Red Hat Enterprise Linux Advanced Platform 5.x or 6.x for 64-bit x64 or IBM System z systems; SuSE Linux Enterprise Server 10 or 11 for 64-bit x64 or IBM System z systems

Ancillary Tasks Report Criteria

Data Cleansing	SPSS IBM Modeler can replace any missing values by using a simple or difficult imputation method. The outliers can be removed by Anomaly node, found during clustering.
Data Filtering	Users can select the data that is relevant to their analysis using Filter Node in SPSS IBM Modeler.

Randomization	SPSS IBM Modeler allows randomization of data using their regular random functions. Users may have to create separate random attributes within the two intervals and then use its random function.
Deriving Variables	SPSS IBM Modeler provides with a Derive node that can be used for deriving variables. The user may enter a formula for derivation of any variables.
Record Deletion	One can delete certain observations directly from a data set on permanent basis or select/highlight data that is required for an analysis.

Usability Criteria

User Interface	SPSS IBM Modeler is an easy to use interface for data mining process. It helps in constructing powerful and accurate models by using modeling algorithms, for instance, classification, segmentation, prediction, and association detection. The results can also be easily read, understood and deployed into databases.

RAPIDMINER

RapidMiner was developed in 2001 by Ralf Klinkenberg, Ingo Mierswa, and Simon Fischer and was known as YALE or Yet Another Learning Environment. In the year 2006, Ralf Klinkenberg, and Ingo Mierswa founded a company named Rapid-I and changed the name of the software to RapidMiner in 2007. "RapidMiner is a data science software platform developed by the company of the same name that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics." It can also be framed as, RapidMiner is a software platform that integrates data preparation, machine learning and predictive model deployment, for data scientists.

Advantages of RapidMiner:

- RapidMiner's template based framework ensures fast delivery and nearly eliminates for the user's need to code.
- It can be used to develop a strong model using the functions and algorithms offered by RapidMiner.

Functionality Criteria

Algorithmic Variety:	RapidMiner supports for predictive and clustering models. These require one or more input variables to predict one or more output variable, also known as the target variable. Decision Trees (CHAID, Decision Stump, ID3, etc.), Neural Nets, Logistic Regression, SVM, and Ensembles are some of the examples of predictive modelling. RapidMiner has a rich database for various algorithms.
Model Validation	RapidMiner provides with various operators to perform a validation of a model. These are: • Cross Validation • Bootstrapping Validation • Split Validation • Wrapper Split Validation

	• Wrapper X- Validation
Data Type Flexibility	RapidMiner supports the following data types: • Attribute • Date • Date_Time • Binomial • File_path • Integer • Nominal, etc
Reporting and Model Exporting	RapidMiner provides for automatic opening of results after running a process. The user can also report results from their repositories. RapidMiner also allows the users to share their results/repositories, or model. This can be done by converting it into a zip file.

Computational Performance Criteria

Platform Variety	RapidMiner can be used on various platforms and is ideal for small, medium and large enterprises; and supports the following: • 64-bit recommended • Windows 7, Windows 8, Windows 8.1, Windows 10 • Linux • MacOS X 10.8 or newer

Ancillary Tasks Report Criteria

Data Cleansing	Data Cleansing can be done in RapidMiner using Impute Missing Values, Replace Missing Values, Fill Data Gaps, Declare Missing Value, and Replace Infinite Values. Every operator has a different function and can be used for respective requirement.
Data Filtering	RapidMiner allows the users to filter their data set using Outliers, and Normalization.
Randomization	The operator 'Select by Random' in RapidMiner allows a user to select a random subset from a given dataset.
Deriving Variables	The users can derive variable using different operators in RapidMiner. RapidMiner has seven different operators to perform this function with pre-defined formulae.
Record Deletion	RapidMiner allows the user to delete the duplicate variables or observations using the operator 'Remove Duplicates'

Usability Criteria

User Interface	RapidMiner has an intuitive user interface. The user is faced with transformations or corrections as and when they build a model. The operators in RapidMiner are easy to find. It is easy for the beginners and very powerful for the experienced.

COMPARITIVE ANALYSIS

	Criterion	SAS	RapidMiner	SPSS Modeler
Functionality	Algorithmic Variety	Decision Trees (CHAID, CARD and C4.5), DM Neural Mode, Memory Based Reasoning (k – nearest neighbor)	Decision Trees (CHAID, Decision Stump, ID3, etc.), Neural Nets, Logistic Regression, SVM, and Ensembles	Decision Trees, regression, neural networks, SVM, and Bayesian networks
	Prescribed Methodology	SEMMA	CRISP - DM	CRISP – DM
	Model Validation	Allows segregation into training, validation and testing.	Allows segregation into new data, validation and testing.	Allows segregation into training, validation and testing.
	Reporting	Reporting Possible	Reporting Possible	Reporting Possible
	Model Exporting	Sharing Possible	Sharing Possible	Sharing Possible
	Data Type Flexibility	Very Flexible	Very Flexible	Very Flexible
Computational Performance Criteria	**Platform Variety**	Works on only Microsoft Windows.	Works on all platforms.	Works on all platforms.
Usability Criteria	**User Interface**	User Friendly	User Friendly	User Friendly

Ancillary Tasks Support	**Data Cleansing**	Outliers are removed using filter mode and missing values are imputed.	Uses Impute Missing Values, Replace Missing Values, Fill Data Gaps, Declare Missing Value, and Replace Infinite Values	Anomaly node is used to find outliers and missing values are imputed.
	Data Filtering	Filter node may be used.	Data can be filtered using normalization and outliers.	Filter Node may be used.
	Deriving Attributes	SAS uses transform variable node.	There are various formulae which can be used to drive attributes.	SPSS uses Derive Node.
	Randomization	This can be done using the sample data.	This can be done using 'Select by Random'.	This can be done using the sample data.
	Record Deletion	Filter node	Allows deletion using Remove Duplicates operator.	Records can be deleted using select node.

CONCLUSION

This paper focused on evaluating and analyzing three data mining software that implements the decision tree induction techniques. The paper gave brief description of the software, their features, advantages and evaluation criteria, i.e., functionality, performance, usability, and ancillary tasks support. After comparing the three software, we can easily distinguish between the three. However, which software would suit the best for a business would eventually depend upon the business requirements, data mining project requirements, team, etc.

REFERENCES

1. Roberta Sicillano and Claudia Conversano. Decision Tree Induction. University of Naples Federico II, Italy and University of Cassino, Italy
2. Sweta Tiwari, Prof Rekha Pandit, and Prof Vineet Richhariya. Predicting future trends in stock market by decision tree rough-set based hybrid system with HHMM. Dept. of Computer Science & Engineering, RGPV University Bhopal, India.
3. en.wikipedia.org
4. www.support.sas.com
5. www.ibm.support.com
6. www.docs.rapidminer.com