ACSIJ Advances in Computer Science: an International Journal, Vol. 3, Issue 1, No.7 , January 2014
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Theoretical Factors underlying Data Mining Techniques in Developing Countries: a case of Tanzania

Salim Amour Diwani[1], Anael Sam[2]

[1] Information Communication Science and Engineering, Nelson Mandela African Institution of Science and Technology, Arusha, P.O.BOX 447, Tanzania
diwania@nm-aist.ac.tz

[2] Information Communication Science and Engineering, Arusha, P.O.BOX 447, Tanzania
Anael.sam@nm-aist.ac.tz

Abstract

Just as the mining of Tanzanite is the process of extracting large block of hard rock's by using sophisticated hard rock mining techniques to find valuable tanzanite glamour, data mining is the process of extracting useful information or knowledge from large un-organized data to enable effective decision making. Although data mining technology is growing rapidly, many IT experts and business consultants may not have a clue about the term. The purpose of this paper is to introduce data mining techniques, tools, a survey of data mining applications, data mining ethics and data mining process.

Keywords: Data Mining, Knowledge Discovery, Mererani , KDD, J48, RandomForest, UserClassifier, SimpleCART, RandomTree, BFTree, DecisionStump, LADTree, logit transform, Naive Bayes

1. Introduction

The amount of data kept in computer files and databases are growing rapidly. At the same time users of these data are expecting to get sophisticated knowledge from them. Banking industry wants to give better service to customers and cross sell bank services, banks need to mine the huge amount of data generated by different banking operations so that new approaches and ideas are generated. Data Mining is an interdisciplinary field that combines artificial intelligence, computer science, machine learning, database management, data visualization, mathematic algorithms, and statistics. It is a technology for knowledge extraction from huge databases[1]. This technology provides different methodologies for decision making, problem solving, analysis, planning, diagnosis, detection, integration, prevention, learning and innovation. Data mining technology involves different algorithms to accomplish different functions. All these algorithms attempt to fit the model to the data. Data mining algorithms is divided into two parts:- predictive model which makes a prediction about values of data using known results found from different data based on historical data, like how Netflix recommends movies to its customers based on the films themselves which are arranged as groups of common movies and how Amazon recommends books to its users based on types of books user requested. The other is a descriptive model which identifies patterns or relationships concealed in a database. Unlike the predictive model, a descriptive model serves as a way to explore the properties of the data examined, not to predict new properties. For instance we can find the relationship in the employees database between age and lunch patterns. Assume that in Tanzania most employees in their thirties like to eat Ugali with Sukumawiki for lunch break and employees in their forties prefer to carry a home cooked lunch from their homes. Therefore we can find this pattern in the database by using descriptive model. Data mining can be applied in strategic decision making, wealth generation, analyzing events and for security purpose by mining network streaming data or finding for abnormal behavior which might occur in the network . Data mining helps organizations to achieve various business objectives such as low costs, increase revenue generation while maintaining the quality of their products. Also data mining can improves the customer

service, better target marketing campaigns, identify high risk clients and improve production processes.

2. Methodology

2.1 Identification of Publications

In order to identify selected publications in the area of knowledge discovery in health care system, articles were selected from various databases and resources linked to Nelson Mandela African Institute of Science and Technology (NM-AIST), such as African Digital Library, Organization for economic corporation and development(OECD), Hinari, Agora, Oare, Emerald, Institute of Physics and IEEE. Keywords such as database, data mining, data warehousing, healthcare and knowledge discovery were used to facilitate the searches.

2.2 Selection of Publications

This literature reviews considered the papers and articles published between 2008 and 2012 in the areas of data mining, knowledge discovery, and health care. The literature published in English were selected with specific emphasis placed on literature covering the relationship between knowledge discovery and data mining, applications of data mining in health care in different countries in the developed world and constraints and requirements for the set up of data mining in health care.

2.3 Analysis Strategy for Selected Publication

The selected articles will then be analyzed for trends in data mining over the period of review stated above. The selected literatures were categorized according to areas of emphasis including frame work for knowledge discovery and data mining techniques, methods and algorithm for knowledge discovery and data mining and specific instances of application of data mining in health care.

3. Data Mining Techniques

Data mining make the use of various techniques in order to perform different types of tasks. These techniques examine the sample data of a problem and select the appropriate model that fits closely to the problem to be tackled. Data mining is described as the automated analysis of large amounts of data to find patterns and trends that may have otherwise gone undiscovered[2]. What data mining techniques does is to find the hidden pattern within data and then to build the model to predict the behaviors based on data collected. Hence data mining is about finding the pattern and building the best model. Pattern is an event or combination of events in a database that occurs more often than expected. Typically this means that its actual occurrences is significantly different than what would be expected by random choice and a model is a description of the original historical database from which it was built that can be successfully applied to new data in order to make predictions about missing values or to make statements about expected values.

The goal of a predictive model is to predict future outcomes based on past records with known answers. Two approaches are commonly used to generate models: supervised and unsupervised.

Supervised or directed modeling is goal-oriented. The task of supervising modeling is to explain the value of some particular field. The user selects the target field and directs the computer to tell how to estimate, classify, or predict it. Unsupervised modeling is used to explain those relationships once they have been found. [3, 4]

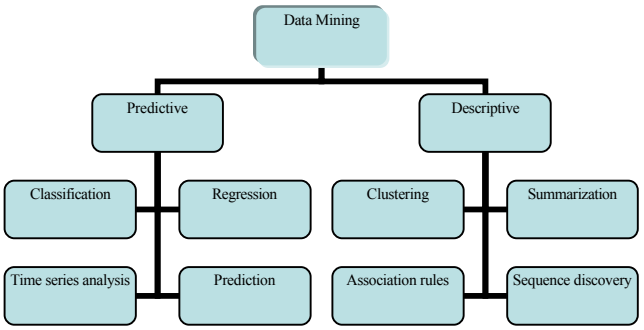

Fig. 1 Data Mining Models and Tasks

4. Predictive Techniques

A predictive model is the process of analyzing historical data using identifying and quantifying relationships between predictive inputs and outcomes found during knowledge and discovery. Some of the techniques of predictive model are classification, regression, generalized linear models, neural networks, genetic algorithms, stochastic gradient boosted trees and support vector machines.

4.1 Classification

Classification maps each data element to one of a set of pre-determined classes based on the difference among data elements belonging to different classes. In

classification, data is divided into groups of objects and each object is assigned to a class of unknown sample, then we identify the attributes that are common to each class. For example, classifying patients with recurring health problem to understand major diseases and make necessary arrangement to give proper treatments. Data mining searches healthcare databases and classifies the patients into symptoms or problems of patients and classify into groups so that they can be easily referred at the time of emergency.

4.2 Regression

Regression is a data mining application which map data into a real valued prediction variable. In regression the variable of interest is continuous in nature. The goal of regression is to develop a predictive model where a set of variables are used to predict the variable of interest. For instance a patient with high blood pressure wishing to reduce his cholesterol level by taking medication. The patients want to predict his cholesterol level before medication and after medication. He uses a simple linear regression formula to predict this value by fitting past behavior to a linear function and then use the same function to predict the values at points in the future based on the values , he then alters his medical improvement. Some of the functions of regression are logistic regression, simple logistic, isotonic regression and sequential minimal optimization (SMO).

4.3 Time Series Analysis

Time series analysis is a collection of observations of well defined data items obtained through repeated measurements over time. The values are usually obtained hourly, daily, weekly or monthly. Time series analysis can be used to identify the nature of the phenomenon represented by the sequence of observation as well as forecasting or predicting future values of time series variable. Example the tidal charts are prediction based upon tidal heights in the past. The known component of the tides are built into models that can be employed to predict future values of the tidal heights.

4.4 Predictions

Prediction is the process of predicting future events, it will predict what will happen next based on the available input data. For instance if Salim turns on the TV in the evening then he will 80% of the time go to kitchen to make coffee. Prediction techniques includes Nearest Neighbor, Neural Network, Bayesian Classifier, Decision Tree, Hidden Markov Model and Temporal Belief Network. Prediction applications includes flooding, speech recognition, machine learning and pattern recognition.

4.5 Decision Tree

Decision tree is a popular structure of supervised learning technique used for classification and regression. The goal is to create a model that predict the value of the target variable. Some of the popular algorithms for decision tree are J48, RandomForest, UserClassifier, SimpleCART, Random Tree, BFTree, DecisionStump, and LADTree. Decision tree can be used in business to quantify decision making and to allow comparison of different possible decision to be made. Below is the example of Fisher's Iris data set. Starting from the top if petal width is less than or equal to 0.6cm the Iris is Setosa. Next we see that if Petal width is greater than 0.6 also greater than 1.7cm then the Iris is Varginica.

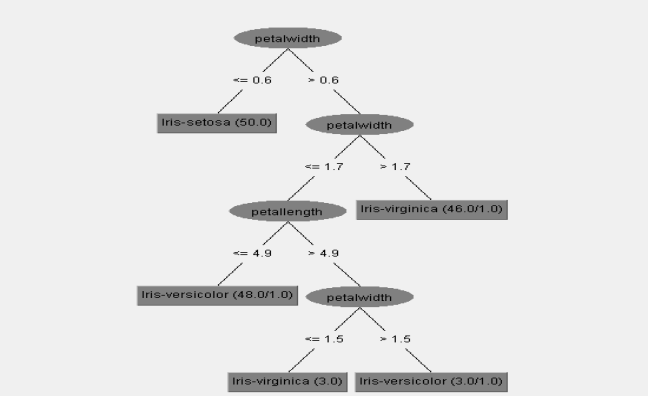

Fig. 2 Example of Decision Tree

4.6 Naive Bayes

Naive Bayes is the supervised learning algorithms which analyzes the relationships between independent and dependent variables. This method goes by the name of *Naïve Bayes* because it's based on Bayes' rule and "naïvely" assumes independence—it is only valid to multiply probabilities when the events are independent. The assumption that attributes are independent (given the class) in real life certainly is a simplistic one. But despite the disparaging name, Naïve Bayes works very effectively when tested on actual datasets, particularly when combined with some of the attribute selection procedures[5].

4.7 K-Nearest Neighbor

Nearest neighbor method is a technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset. Sometimes it is called the k-nearest neighbor technique [6]. *K*-nearest neighbor is asymptotically optimal for large k and n with $k/n \rightarrow 0$. Nearest-neighbor methods gained popularity in machine learning through the work of [7], who showed that instance-based learning can be combined with noisy exemplar pruning and attribute weighting and that the resulting methods perform well in comparison with other learning methods.

4.8 Logistic Regression

Logistic regression is a popular and powerful data mining technique that enable to determine the impact of multiple independent variables presented simultaneously to predict membership of one or other of the two dependent variable categories, which uses logit transform to predict probabilities directly. Logistic regression does not assume a linear relationship between dependent and independent variables. The dependent variable must be dichotomy while independent variable need not to be interval nor normally distributed, nor linearly related, nor of equal variance within each group. Logistic regression attempts to produce accurate probability estimates by maximizing the probability of the training data. Hence, probability estimates lead to accurate classification.

4.9 Neural Network

Neural Network is very powerful and complicated data mining technique based on the models of the brain and nervous systems. Neural network is highly parallel which processes information much more like a brain rather than a serial computer. It is particularly effective for predicting events when a network has a large database. Neural networks are typically organized in layers and each layer is made of interconnected nodes. Input layers are interconnected with a number of hidden layers where the actual processing is done via system of weighted connections. These hidden layers can be connected to other hidden layers which finally link to the output layer. Neural network can be applied in voice recognition system, image recognition system, industrial robotics, medical imaging, data mining and aerospace application.

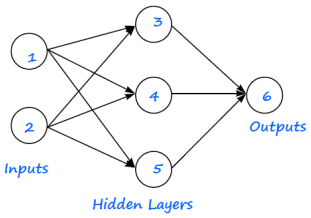

Fig. 3 Example of Neural Network

5. Descriptive Data Mining Techniques

Descriptive data mining techniques are typically unsupervised which describes data set in a concise way and presents interesting characteristics of the data without having any predefined target. They are used to induce interesting patterns from unlabelled data. The induced patterns are useful in exploratory data analysis. Some of the descriptive techniques are clustering, summarization, association rules and sequence discovery.

5.1 Association Rules

Association rules is a process to search relationships among data items in a given data set, which helps in managing all the data items. For instance association rule can be used in retail sales community to identify items that are frequently purchased together. For example people buying school uniforms in December also buy school bags. An association is a rule of the form if X

then Y it is denoted as $X \longrightarrow Y$.
Any rule if $X \longrightarrow Y \longrightarrow Y \longrightarrow X$, then X and Y are called an interesting item set.

5.2 Clustering

Clustering is a process of partitioning or segmenting a set of data or objects into a set of meaningful sub classes, called clusters. Clustering is similar to classification except that the group are not predefined, but rather defined by data alone. Clustering is alternatively referred to as unsupervised learning or segmentation. It can be thought of as partitioning or segmentation the data into groups that might or might not disjointed. Clustering can be used in organizing computing clusters, market segmentation, social network analysis and astronomical data analysis.

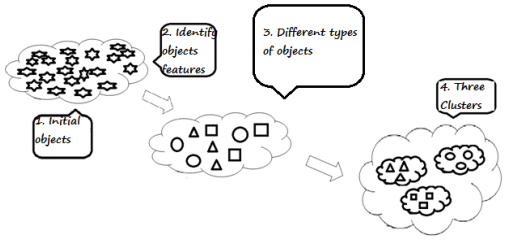

Fig. 4 Clustering Process

In clustering, there are no predefined classes and no examples. The records are grouped together on the basis of self-similarity. It is up to the user to determine what meaning, if any, to attach to the resulting clusters. Clusters of symptoms might indicate different diseases. Clusters of customer attributes might indicate different market segments. Clustering is often done as a prelude to some other form of data mining or modeling. For example, clustering might be the first step in a market segmentation effort: Instead of trying to come up with a one-size-fits-all rule for "what kind of promotion do customers respond to best," first divide the customer base into clusters or people with similar buying habits, and then ask what kind of promotion works best for each cluster[8].

5.3 Sequence Discovery

Sequence discovery is an ability to determine sequential pattern in the data. The input data is the set of sequences called data sequences. Each data sequence is an ordered list of item sets, where each item set is a set of literal. Unlike market basket analysis which requires the items to be purchased over time in some order. For instance, in the medical domain, a data sequence may correspond to the symptoms or diseases diagnosed during the visit to the doctor. The patterns discovered using this data could be used in diseases research to help identify symptoms or diseases that precede certain diseases.

6. Data Mining Process

In order to systematically conduct any data mining analysis, certain procedures should be eventually followed. There is a standard process called Cross-Industry Standard Process for Data Mining (CRISP-DM) widely used by industry members.

6.1 Understanding the Business

This is the first phase which its aim is to understand the objectives and requirements of the business problem and generating data mining definition for the related problem.

6.2 Understanding the Data

The objective of this phase is to analyze the data collected in the first phase and study its characteristics. Models such as cluster analysis can also be applied in this phase so that the patterns can matched to propose hypothesis for solving the problem.

6.3 Data Preparation

In this phase raw data are first transformed and cleaned to generate the data sets that are in desired format. Therefore, this phase creates final datasets that are input to various modeling tools which providing the opportunity to see patterns based on business understanding.

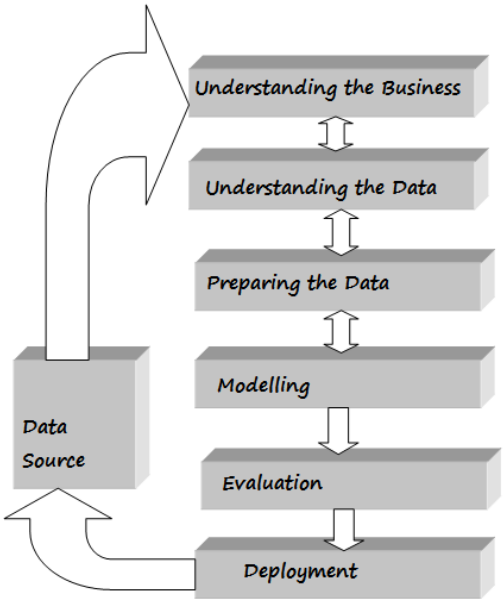

Fig. 5 CRISP DM-PROCESS

6.4 Modeling

In this phase different data mining techniques for modeling such as visualization and clustering analysis can be selected and applied depending on the nature of the data. The data collected from previous phase can be analyzed and predicted the generated output.

6.5 Evaluation

In this phase model results or set of models that you generate in the previous phase should be evaluated for better analysis of the refined data.

6.6 Deployment

The objective of this phase is to organized and implement the knowledge discovered from the previous phase in such a way that it is easy for end users to understand and use in prediction or identification of key situations. Also models need to be monitored for changes in operating conditions, because what might be true today may not be true a year from now. If significant changes do occur, the model should be redone. It's also wise to record the results of data mining projects so documented evidence is available for future studies[8].

7. Data Mining Tools

Data mining is the process that uses a variety of data analysis tools which originates from machine learning and statistics. It uses a variety of data analysis and modeling techniques to discover patterns and relationships in data that may be used to make valid predictions. Most frequently, many business end users are lacking quantitative skills which enables them to analyze their data more effectively by using statistical methods analysis tools. To bridge this gap, software companies have developed data mining application software to make the job a lot easier than they think. These data mining tools allows users to analyze data from different angles by summarizing, predicting and identify the relationships.

Table 1: Data Mining Tools

Product	Vendor	Functions
CART	Salford Systems	Classification
Clementine	SPSS Inc	Association rules, classification, clustering, factor analysis, forecasting, prediction, sequence discovery
Darwin	Oracle Corporation	Clustering, prediction, classification, association rules
Enterprise Miner	SAS Institute Inc.	Association rules, classification, clustering, prediction, time series
Intelligent Miner	IBM Corporation	Association rules, clustering, classification, prediction, sequential patterns, time series
LOGIT	Salford Systems	Forecasting, hypothesis testing
JDA Intellect	JDA Software Group, Inc.	Association rules, Classification, Clustering, Prediction
WEKA	The University of Waikato	Association rules, Classification, Clustering, Visualization

8. Application of Data Mining

Data mining is an emerging technology which finds its application in various fields to identify buying patterns from customers, determine the distribution schedule among outlets, identify successful medical therapies for different illness, detect patterns of fraudulent credit card use and other innovative applications in order to solve the social problem and improve the quality of life. Hence data mining application can be used by small, medium and large organizations to achieve business objectives such as low costs, increase revenue generation while maintaining high standard of living. Below are some of the applications of data mining:

8.1 Business

Mining enables established business organizations to consolidate their business setup by providing them with reduced cost of doing the business, improved profit, and enhanced quality of service to the consumer.

8.2 Electronic Commerce

Data mining can be used in web design and promotion depending on the user's needs and wants. Also it can be used in cross selling by suggesting to a web customers items that he/she may be interested in, through correlating properties about the customer, or the items the person has ordered.

8.3 Computer Security

Data mining enables network administrators and computer security experts to combine its analytical techniques with your business knowledge to identify probable instances of fraud and abuse that compromises the security of computer or a network.

8.4 Health Care

Healthcare organization generates large amount of data in its clinical and diagnostic activities. Data mining enables such organization to use machine learning techniques to analyze healthcare data and discovered new knowledge that might be useful in developing new drugs.

8.5 Telecommunication

Telecommunication industry can use data mining to enable telecommunication analyst to consolidate telecommunication setup by providing them with reduced cost of doing the business, improving profit and enhancing the quality of service to consumers.

8.6 Banking

Data mining enables banking authorities to study and analyze the credit patterns of their consumers and prevent any kind of bad credits or fraud detection in any kind of banking transactions. It also enables them to find hidden correlations between different financial indicators and identify stock trading from historical market data.

8.7 Bioinformatics

Data mining enables biological scientists to analyze large amount of data being generated in the field of bioinformatics studies using the techniques of data visualization.

8.8 Stocks and Investment

Data mining enables you to first study the specific patterns of growth or downslides of various companies and then intelligently invest in a company that shows the most stable growth for a specific period.

8.9 Crime Analysis

Data mining enables security agencies and police organizations to analyze the crime rate of a city or a locality by studying the past and current attributes that leads to the crime. The study and analysis of these crime reports helps prevent the reoccurrence of such incidences and enables concerned authorities to take preventive measures too

9. Discussion

Since the conception of data mining, data mining has achieved tremendous success in today's business world. Many new problems have emerged and have been solved by data mining researchers. However, we still have big challenges in front of us. Some of the most difficult challenges faced by data miners are individual privacy, anonymization, discrimination and integration.
Issue of privacy of an individual is an ethical issue, before collecting any personal information the purpose must be stated and such information must not be disclosed to others without consent. Also personal information must

not be transmitted to location where equivalent data protection cannot be assured and some data are too sensitive to be collected except in extreme circumstances such as sexual orientation or religion.

Anonymization of an individual is another ethical issue which is harder than you think. For instance the hospital data are very sensitive before the release of medical data to any researcher should first be anonymized by removing all identifying information of an individual such as name, address, and national identification number.

The main purpose of data mining is discrimination. In banking industry data mining is used to discriminate customers who apply for bank loan and check if they are eligible or not. The ones who are eligible for bank loan in what amount they can apply for the loan and for how long. Also data mining can be used to discriminate customers by identifying who can get special offer. Some kind of discrimination are unethical and illegal hence cannot be allowed such as racial, sexual and religious.

Maintaining Data integrity is very important in any organizations, if the data does not have proper integrity, it could be false or wrong. If the data is wrong, the results will be wrong. Hence conclusions and interpretations of the results will be wrong. Everything will have been a complete waste of time and money, and it will never get work as a researcher again. The key challenge here is the integration of data from different sources. For instance CRDB bank in Tanzania have many branches across the country and each branch require different database which needed to be integrated to the main branch. Each customer has more than one ATM card and each ATM card has more than one address. Therefore mining these type of data is very difficult since the application software need to translate data from one location to another location and select most address which recently entered hence it is difficult to extract the correct information.

Recently the cost of hardware system has dropped dramatically, hence data mining and data warehousing becoming extremely expensive to maintain . Small, medium and large organizations are accumulating a lot of data in day to day activities through their local branches, government agencies and companies that are dealing with data collection electronically. Therefore there is no other choice rather than to implement data mining technology in their organization no matter what so that they can achieve their business goals, increase revenue generation and reduce labor cost. Therefore data mining is like a pregnant woman whatever she eats nourishes the baby.

10. Conclusion

This paper has defined data mining as a tool of extracting useful information or knowledge from large un-organized data base which enable organization to make effective decision. Most of the organization uses data warehouse to store their data and later be extracted and mined in order to discover new knowledge from their databases in the acceptable format such as ARFF and CSV format. The data is then analyzed using data mining techniques and the best model will be built which help organizations in their effective decision making. Some of the techniques have been discussed includes classification, regression, time series analysis, prediction, decision tree, naive bayes, k nearest neighbor, logistic regression and neural network.

Data mining process called CRISP-DM model have been discussed. In order to carry out any data mining task some procedure need to be followed included business understanding, data understanding, data preparation modeling, testing, evaluation and deployment. Therefore before doing any data mining task you need to ask yourself what do you need to know and why and how much data do you need to collect, collect your required data and clean the data, data cleaning is the most hardest part of data mining task which need high understanding of data mining techniques. Lastly, you define your new features and deploy your data by convincing your boss if the data you mined make any sense and can be used in decision making.

Many industries use data mining applications in day to day activities such as banking, healthcare, Telecommunication, Marketing and police. In marketing data mining can be used to select best profitable market to launch new products and to maintain competitive edge over competitors. In banking industries data mining can be used to segment data, determine customers preferences, detect frauds, cross selling banking services and retention of customers. In healthcare data mining can be used to segment patients into groups, identifying the frequent patients and their recurring health problems, relationships between diseases and symptoms, curbing the treatment costs and predicting the medical diagnosis. Therefore we conclude that those industries who take full advantage of data mining application in their day to day activities have the ability to enhance competitive advantage.

Acknowledgements

I would like to thank almighty and my family for their constant support during this work. I would also like to

thank my supervisors Dr. Anael Sam from Nelson Mandela African Institute of Science and Technology and Dr. Muhammad Abulaish from Jamia Milia Islamia for their guidance and valuable support during this work.

References

[1]Liao, S.-h. (2003). 'Knowledge Management Technologies and applications-Literature review from 1995 to 2002'. *Expert System with Application,* 25, 155-164.
[2] Fabris, P. 1998. Advanced Navigation. CIO, May15.
[3] Berson, A., Smith, S., and Thearling, K., Building Data Mining Applications for CRM (McGraw-Hill, New York, 2000).
[4] J. Shanmugasundarum, M.V.Nagendra-Prasad, S. Vadhavkar, A. Gupta, Use Of Recurrent Neural Networks For Strategic Data Mining Of Sales, MIT Sloan School of Management, Working Paper 4347- 02, 2002.
[5] Zhang, H., Jiang, L., & Su, J. (2005). Hidden Naïve Bayes. In *Proceedings of the 20th National Conference on Artificial Intelligence* (pp. 919–924). Pittsburgh. Menlo Park, CA: AAAI Press.
[6] Devroye, L., Györfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. New York: Springer-Verlag.
[7] Aha, D. (1992). Tolerating noisy, irrelevant, and novel attributes in instance-based learning algorithms. *International Journal of Man-Machine Studies*, 36(2), 267–287.
[8]Michael Berry, Gordon Linoff., Data Mining Techniques (WILLEY Publishing, Inc., Indianapolis, Indiana, 2004)

Salim Amour Diwani received his BS degree in computer science at Jamia Hamdard University, New Delhi, India in 2006 and Ms degree in computer science at Jamia Hamdard University in New Delhi, India in 2008. He is currently a PhD scholar in Information communication Science and Engineering at Nelson Mandela African Institution of Science and Technology in Arusha, Tanzania. His primary research interests are in Data Mining, Machine Learning and Database Management Systems. He published two papers in the area of Data Mining.