The International Journal Of Engineering And Science (IJES)
|| Volume || 5 || Issue || 9 || Pages || PP -01-11 || 2016 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

A Survey and Analysis on Classification and Regression Data Mining Techniques for Diseases Outbreak Prediction in Datasets

Hakizimana Leopord[1], Dr. Wilson Kipruto Cheruiyot[2], Dr. Stephen Kimani[3]
[1]*Ph.D. Research Scholar, Computing Department, School of Computing and Information Technology, Jomo Kenyatta University of Agriculture and Technology, Company, Kigali, Rwanda*
[1,2]*Senior lecturer, Computing Department, School of Computing and Information Technology, Jomo Kenyatta University of Agriculture and Technology, Company,* Nairobi, *Kenya*

----------ABSTRACT----------
Classification and regression as data mining techniques for predicting the diseases outbreak has been permitted in the health institutions which have relative opportunities for conducting the treatment of diseases. But there is a need to develop a strong model for predicting disease outbreak in datasets based in various countries by filling the existing data mining technique gaps where the majority of models are relaying on single data mining techniques which their accuracies in prediction are not maximized for achieving expected results and also prediction are still few. This paper presents a survey and analysis for existing techniques on both classification and regression models techniques that have been applied for diseases outbreak prediction in datasets.
Keywords: *Classification, Regression, Data Mining, Prediction Model, Outbreak Diseases.*

Date of Submission: 22 August 2016 Date of Accepted: 05 September 2016

I. INTRODUCTION

Over the past years and today, there has been a rapid technological improvement in Computer Science which has led to the evolution and developments of data mining technologies in the health sector for the purposes of hidden pattern discovery such as disease prediction, detection, forecasting which has a paramount importance in health decision making. Thus, with the aging population on the rise in developed countries and the increasing cost of healthcare, governments and large health organizations are becoming very interested in the potential of health informatics to save time, money and human lives (Shukla, 2014).

In Africa, many people are dying because of the weakness of the disease outbreaks prediction techniques usage. Ebola outbreak in 2014 killed many people in West African countries like Guinea, Liberia, and Sierra Leone. On the other hand in 2011, another potentially fatal dengue hemorrhagic fever (DHF) was a crucial public health concern in Malaysia. This research intends to bridge that gap by predicting disease outbreaks using a classification and regression model for predicting disease outbreak in datasets as a hybrid approach. Moreover, research is proposing that after testing and evaluation of the research the outputs and impact will sound as the paramount importance in health dimension in decision making for the stake holders and policy makers where disease outbreaks can be prevented before affecting a big number of people.

The health care field contains a huge amounts of data which holds sensitive information about patient details , diagnosis, medical conditions and disease prognosis such as outbreaks and people have no awareness about it. Therefore, their prediction is resulting so difficult among health workers and the most diseases recognized or marked on the last stage. As well, a research has demonstrated that these diseases are the abnormal medical conditions of organisms that impair bodily functions, are associated with recognizable symptoms and signs. The causes of diseases may be related to external factors such as infectious disease or autoimmune disease in the case of internal dysfunctions (Kumar, 2007). Researchers like Huang (2004) explained that an outbreak or an epidemic is the occurrence of a health-related event (illness, disease complications and health-related behavior) clearly in excess of the normal expected. Again an epidemic may include any kind of disease, including noninfectious conditions.

Some researchers identified that among types of diseases considered as outbreaks includes more than 50 percent of new cancer cases occurred in developing countries (M, 2012). Communicable disease outbreaks cause millions of deaths throughout Sub-Saharan Africa each year; most of the diseases causing epidemics in the region have been nearly eradicated or brought under control in other parts of the world. Moreover, detailed discussions regarding the diseases outbreak are going to be mentioned and explained below.

In 1998, the World Health Organization African Regional Office (WHO/AFRO) presented the integrated disease surveillance and response (IDSR) strategy. In this view, research shows that the most commonly reported

epidemic outbreaks in Africa included: cholera, dysentery, malaria and hemorrhagic fevers (e.g., Ebola, Rift Valley fever, Crimean-Congo fever and yellow fever) and the cyclic meningococcal meningitis outbreak that affects countries along the "meningitis belt" (spanning Sub-Saharan Africa from Senegal, Gambia to Kenya and Ethiopia) accounts for other major epidemics in the region (Senait Kebede, 2010).

More specifically, in July 2011 – June 2012, Rwanda Biomendical Centre (RBC) has conducted an Outbreak Management. The RBC/EID provided the appropriate drugs, including other consumables in case of outbreak, supervised affected health system components with management guidelines for infection control, recommendations for prevention and control measures. This was renewed each year to facilitate quick response in case of an outbreak (KAYUMBA, July 2011 – June 2012).

Early prediction mechanisms are critical in reducing the impact of epidemics and preventing the epidemics from becoming unmanageable by making a rapid response. For example, the cholera epidemic killed over 100,000 people worldwide and sickened 35 million people during the year 2010 (Enserink, 2010). Some researchers estimated that 17 million people die of cardiovascular diseases (CVD) every year (Mackay, 2004).

Abdoulaye (2015) revealed that, during 2014 Ebola outbreak in West Africa which was the longest, largest, and deadliest and the most complex outbreak ever witnessed globally. The Ebola virus disease (EVD) epidemic in Guinea, Liberia and Sierra Leone was the longest, largest, deadliest, the most complex and challenging Ebola outbreak in history. It was unprecedented in terms of its duration; size of infections and fatality, geographical spread unlike the past outbreaks which lasted for a very short time, the West African Ebola case has lasted for more than one year and has not yet fully abated. As of 11 February 2015, there were 22,859 EVD cases in total: 3,044 in Guinea, 8,881 in Liberia, and 10,934 in Sierra Leone with cumulative deaths of 9,162 victims.

Fortunately, the disease outbreak prediction models are increasingly gaining popularity since these models are developed to predict the disease outbreaks which are becoming increasingly common world-wide. A variety of data mining models are in use around the world, with coverage of national, international and global diseases outbreaks. These models use different technologies for predicting and prioritization of potential disease outbreaks. The ultimate purposes of these models are to ensure quick prediction of possible disease outbreaks.

Therefore, data mining is considered as predominant, motivating areas of research with the hope of discovering meaningful information from huge data sets. At this juncture, data mining is becoming very popular in the healthcare field but the efficient strategies for predicting unknown and valuable information in health data are still an issue to be dealt with. In the health industry, data mining provides extreme benefits such as detection of the fraud in health insurance, availability of medical solutions to the patients at lower cost, detection of causes of diseases and identification of good medical treatment methods. Furthermore, it helps the healthcare researchers for making efficient health care policies, constructing drug recommendation systems and developing health profiles of individuals (H. C. Koh, 2005).

Just years ago, researchers showed that data mining techniques are used to analyze the various factors that are responsible for diseases for example, types of food the body needs, different working environments, education levels, living conditions, availability of pure water, health care services, cultural ,environmental and agricultural factors

Moreover, medical data mining has a great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis and predictions. Even though the available raw medical data is widely distributed, heterogeneous in nature and voluminous. This data needs to be collected in an organized form. This collected data can be then integrated to form a hospital information system. Data mining technology provides a user-oriented approach to the novel and discovers hidden patterns in the data (Jyoti Soni U. A., 2011).

Recently, the huge amount of data being collected and stored in databases or dataset format has recently increased due to the advancements of interest to researchers in data mining, machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for modeling ultimate purposes.

The researcher Richards G (2001) has discussed in the knowledge discovery and data mining (KDD), he stated that KDD is an interdisciplinary area focusing upon methodologies for extracting useful knowledge from data. Again the term knowledge discovery in databases or KDD for short refers to the broad process of finding knowledge in data and emphasizes the "high-level" application of particular data mining methods. The unifying goal of the KDD process is to extract knowledge from data in the context of large databases. Applications of data mining have already been proven to provide benefits to many areas of medicine including diagnosis, prognosis and treatment.

Jiawei (2009) stated that, data mining is an interdisciplinary field merging ideas from statistics, machine learning, information science, visualization and other disciplines. This symbolizes that it is a very useful approach to integrate information and theory for knowledge discovery from any informatics such as Bioinformatics, Chemo informatics, Nano informatics and materials informatics. Data mining consists of a set of techniques that can be used to extract relevant and interesting knowledge from data. Data mining has several

tasks such as association rules, classification, predictions and clustering etc. Classification techniques are supervised learning techniques that are to classify data item into the predefined class label. It is one of the most useful techniques in data mining models building by relaying on the datasets. Using classification techniques commonly to build models that are used to predict future data trends (Mahendra Tiwari, 2013).

The purpose of this paper is to provide the review and analysis on classification and regression techniques models and the main purpose of this survey is to help a research for establishing a best model disease outbreak prediction with critical enhancement for achieving the maximum accuracy advantages or reinforcing prediction as data mining technologies.

The structure adopted for this paper is statement of the methodology adopted for the paper, review of the related work, critique of the various classification and regression techniques models in related work, presentation of conclusion of the state of the art and the conducting of this analysis and survey, the methodology used are to select papers done within the last five years and others done many years ago to ensure that we get the recent state of the art paper and some data mining history so that it would provide relevant information. The books, conferences or journal papers were considered and all of resources should be indexed. A critical analysis citing weaknesses and strengths of the journals on classification and regression prediction model techniques are then presented in the paper and a conclusion drawn in the paper on the basis of observations made on the analysis and reviews.

II. RELATED WORK

2.1 Data Mining and KDD Process

Furthermore before conducting a review and analysis work, we first have to understand what data mining is as the main area of the study. Hand (2001) declared that data mining came into existence in the middle of 1990's and appeared as a powerful tool that is suitable for fetching previously unknown pattern and useful information from huge dataset. Various studies highlighted that data mining techniques help the data holder to analyze and discover unsuspected relationships among their data which in turn helpful for decisions making.

Taneja (2014) stated that a data mining is a technique that deals with the extraction of hidden predictive information from a large database. It uses sophisticated algorithms for the process of sorting through large amounts of data sets and picking out relevant information. Data mining (the Analysis step of the Knowledge Discovery in Databases process, or KDD), a relatively young and interdisciplinary field of computer science, is the process of extracting Patterns from large data sets by combining methods from statistics and artificial intelligence with database management (Gaurav Taneja, 2014).

Fayyad (1996) coined that the term Knowledge Discovery in Databases, or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems and data visualization. Furthermore researcher had presented an outline of the steps of the KDD process as shown below.

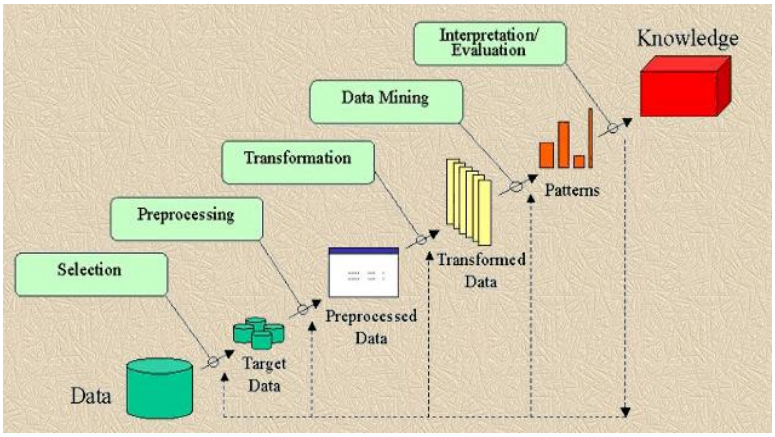

Figure 2.1 Steps of the KDD Process (Fayyad, 1996).

Generally, Data Mining and Knowledge Discovery in Databases (KDD) are related terms and are used interchangeably but many researchers assume that both terms are different as Data Mining is one of the most important stages of the KDD process (U. Fayyad, 1996).

Data mining techniques has been used as hybrid model in term of outbreak diseases prediction, this means that it way of combining two data mining techniques for filling up the weakness that caused by one technique.

2.2. Classification and Regression Data Mining Techniques

A classification technique is a systematic approach to building classification models for training and testing data sets. Several classification models such as Decision Tree Classifier, Rule-Based Classifier, Neural Network Classifier, naive Bayesian Classifier, Neuro-Fuzzy classifier, Support Vector Machines, regression and so on. As reported in the literature so that this section discusses the various data mining techniques which are used in the prediction models in both of the health domain also the application domain. Data mining (DM) is the extraction of useful information from large data sets that results in predicting or describing the data using techniques such as classification, clustering, association, etc. Data mining has found extensive applicability in the healthcare industry such as in classifying optimum treatment methods, predicting disease risk factors, and finding efficient cost structures of patient care. Research using data mining models have been applied to diseases such as diabetes, asthma, cardiovascular diseases, AIDS, etc. Various techniques of data mining such as naïve Bayesian classification, artificial neural networks, support vector machines, decision trees, logistic regression, etc. have been used to develop models in healthcare research (Mythili T., 2014).

Classification divides data samples into target classes. The classification technique predicts the target class for each data point. For example, patients can be classified as "high risk" or "low risk" patient on the basis of their disease pattern using data classification approach. It is a supervised learning approach having known class categories. Binary and multilevel are the two methods of classification. In binary classification, only two possible classes such as, "high" or "low" risk patient may be considered while the multiclass approach has more than two targets for example, "high", "medium" and "low" risk patient. Data set is partitioned as training and testing dataset. Using training dataset we trained the classifier. Correctness of the classifier could be tested using test a dataset. Classification is one of the most widely used methods of Data Mining in Healthcare organization (Divya Tomar, 2013).

Noor Diana Ahmad Tarmizi (2013) stated that a classification technique, also known as a classifier is a systematic approach in developing a classification model from a set of input data. There are many classifiers such as decision tree (DT), Neural networks (NN), naive Bayesian, support vector machine (SVM) and rough set theory (RST). Each classifier uses learning algorithms to discover the most appropriate model for the relationship between an attribute set and class labels of input data. The model produced by the learning algorithm should both fit the input data well and correctly predict the class labels of records that it has never seen before.

The main advantages and disadvantages of different classification techniques summary was reviewed as indicated in the table 2.1 and some of them will be considered in this research conduction process (Divya Tomar, 2013).

Table 2.1 Advantages and disadvantages of different classification techniques

Methods	**Advantage**	**Disadvantage**
K-NN	1. It is easy to implement. 2. Training is done in faster manner	1. It requires large space. 2. Sensitive to noise 3. Testing is slow.
Decision Tree	1.There are no requirements of domain knowledge in the construction of decision tree. 2.It minimizes the ambiguity of complicated decisions and assigns exact values to outcomes of various actions. 3. It can easily process the data with high dimension. 4. It is easy to interpret. 5. Decision tree also handles both numerical and categorical data.	1. It is restricted to one output attribute. 2. It generates categorical output. 3. It is an unstable classifier i.e. performance of classifier is depend upon the type of dataset. 4. If the type of dataset is numeric than it generates a complex decision tree
Support Vector Machine	1. Better Accuracy as compare to other classifier. 2. Easily handle complex nonlinear data points. 3.Over fitting problem is not as much as other methods.	1. Computationally expensive. 2. The main problem is the selection of right kernel function. For every dataset different kernel function shows different results. 3. As compare to other methods training process take more time. 4. SVM was designed to solve the problem of binary class. It solves the problem of multi class by breaking it into pair of two classes

		such as one-against-one and one-against-all.
Neural Network	1. Easily identify complex Relationships between dependent and independent variables. 2. Able to handle noisy data.	1. Local minima. 2. Over-fitting. 3. The processing of ANN network is difficult to interpret and require high processing time if there are large neural networks.
Bayesian Belief Network	1. It makes computations process easier. 2. Have better speed and accuracy for huge datasets.	1. It does not give accurate results in some cases where there exists dependency among variables.

2.2.1 Naive Bayes Classifier

Naïve Bayes is a data mining technique that shows success in classification of diagnosing heart disease patients (Sitar-Taut, 2009). Naïve Bayes is based on probability theory to find the most likely possible classifications (Yadav, 2012). Bayesian classifier calculates conditional probability of an instance belonging to each class, and based on such conditional probability data, the instance is classified as the class with the highest conditional probability. In knowledge expression, it has the excellent interpretability same as decision tree, and is able to use previous data to build analysis model for future prediction or classification (Gustafson, 1993). As reported by Shubpreet Kaur D. R (2015) Navies Bayes is the most common technique that is used in data mining. It gives maximum accuracy of 96.5% in curing heart patients as shown in the Figure indicated below and has coined that Naïve Bayes is widely used technique in prediction of various diseases and has the maximum accuracy of 96.5%.

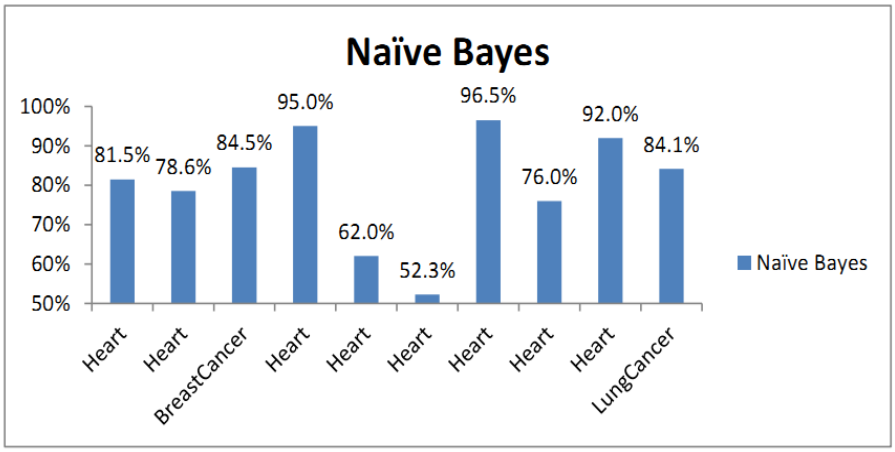

Figure 2.2 Comparison of Various Diseases on Naïve Bayes Technique

The Decision Tree is one of classification techniques I.S.Jenzi (2013) clarified that Decision Tree classifier takes the training set, attribute list and the splitting criteria method as inputs. A Decision tree is generated from which rules are predicted. Different attribute selection measures like Information Gain, Gain ratio, Chi square statistics, and Gini Index and Distance measure can be used. The attributes can be reduced and then given to the model. By reducing the number of attributes, the algorithm can perform faster and efficiently. In this work, the Information gain ratio is used as the splitting criteria. The attribute with the highest information gain is taken as the root of the tree. Information gain is based on Claude Shannon's work on information theory. Info Gain of an attribute A is used to select the best splitting criterion attribute. The highest Info Gain is selected to build the decision tree .The formula is given as follows- Info Gain (A) = Info (D) − Info A (D)

Where

A is the attribute investigated.

Info (D) = $-\sum_{i=1}^{m} p_i \log_2(p_i)$

Where

= probability (class i in dataset D);

m = number of class values.

Info A (D) = $\sum_{j=1}^{v} \frac{|D|}{|I|}$ Info (Dj)

Where

|D j | = number of observations with attribute

Value j in Dataset D;

|D| = total number of observations in dataset D;
D j = sub dataset of D that contains attribute
Value j;
v = all attribute values
The accuracy measure in decision trees was clarified by Shubpreet Kaur D. R (2015) that decision trees are most popular data mining technique which applied and used everywhere that it gives maximum optimal results as shown on the following table.

Table 2.2 Accuracy Measure in Decision Trees

Disease Considered	Author	Year of Publication	Accuracies in DTrees
Heart	Cheung,	2001	81.11%
Skin Diseases	Bojarczuk	2001	89.12%
Breast Cancer	Dursun Delen et al.	2005	93.62%
Heart	Andreeva, P	2006	75.73%
Breast Cancer	Bellaachia et al	2006	86.70%
Heart	Palaniappan, et al.	2007	94.93%
Heart	Sitar-Taut et al.	2009	60.40%
Heart	Tu, et al.,	2009	78.90%
Skin Diseases	Polat and Gunes	2009	96.71%
Heart	Asha Rajkumar et al	2010	52.00%
Heart	Jyoti Soni	2011	99.20%
Heart	Akhil jabbar et al	2012	80.00%
Heart	Abhishek Taneja	2013	94.29%
Liver	Syeda Farha Shazmeen et al.	2013	69.58%
Kidney	K R Lakshmi et al.	2014	78.44%

2.2.2 Artificial neural network
Neural Networks are analytic techniques modeled after the (hypothesized) processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations (on specific variables) from other observations (on the same or other variables) after executing a process of so-called learning from existing data. Neural Networks is one of the Data Mining techniques. The first step is to design a specific network architecture (that includes a specific number of "layers" each consisting of a certain number of "neurons"). The Network is then subjected to the process of "training." In that phase, neurons apply an iterative process to the number of inputs to adjust the weights of the network in order to optimally predict the sample data on which the "training" is performed. After the phase of learning from an existing data set, the new network is ready and it can then be used to generate predictions. The resulting "network" developed in the process of "learning" represents a pattern detected in the data (Megha Gupta, 2010).

2.2.3 Regression
Logistic regression is an approach to prediction, like Ordinary Least Squares (OLS) regression Boshra Bahrami (2015) and Regression is a very important technique of data mining. With the help of it, we can easily identify those functions that are useful in order to demonstrate the correlation among different various variables. It is mainly a mathematical tool. With the help of training dataset we can easily construct it. Consider two variables 'P' and 'Q'. These two types of variables are mainly used in the field of statistics. One of them is known dependence and another one is independent variables. The maximum number of dependent variables cannot be more than one while independent can be exceeds one. Regression is mostly used in order to inspect the certain relationship between variables (Parvez Ahmad, 2015).
Fox (1997) expressed that a regression is used to find out the functions that explain the correlation among different variables. A mathematical model is constructed using training dataset. Also showed that in statistical modeling two kinds of variables are used where one is called dependent variable and another one is called independent variable and usually represented using 'Y' and 'X'. There is always one dependent variable while independent variable may be one or more than one. Regression is a statistical method which investigates the

relationships between variables. By using Regression dependences of one variable upon others may be established. Based on a number of independent variables regression is of two types, one is Linear and another one is Non-linear. Linear regression identifies the relation of a dependent variable and one or more independent variables. It is based on a model which utilizes linear function for its construction. Linear regression finds out a line and calculates vertical distances of the points from the line and minimizes the sum of square of vertical distance (Fox, 1997).

2.3 Classification and Regression Prediction Models for Diseases Outbreak

As stated in the introduction that the related classification or regression prediction Models as data mining techniques should be reviewed and analyzed as follows for furthermore enhancement. Researchers like Divya (2011) indicated that both classification and regression are used for predicting the class or the outcome of a function. The only difference between them is the nature of attributes. If the attributes are categorical then one can use classification algorithms such as Naïve Bayes, SVM, etc., and if the attributes are continuous then regression model using SVM or linear regression achieves great performance.

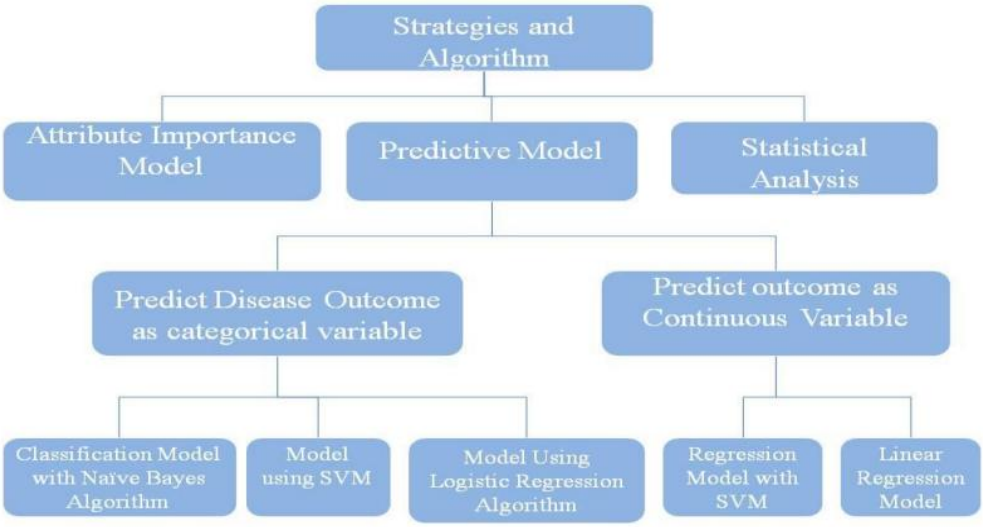

Figure 2.3 Functioning of Classification and Regression Techniques

Sahana (2013) in his research work that the statistics, logistic regression is a type of regression analysis used for predicting the outcome of a categorical dependent variable (a dependent variable that can take on a limited number of values, whose magnitudes are not meaningful but whose ordering of magnitudes may or may not be meaningful) based on one or more predictor variables. An explanation of logistic regression begins with an explanation of the logistic function, which always takes on values between zero and one:

$$f(t)=\frac{1}{1+e^{-t}}$$

Data Mining Researchers like Yuhanis Yusof Z. M (2011) has elaborated prediction model for incorporating Least Squares Support Vector Machines (LS-SVM) in predicting future dengue outbreaks. Data sets used in the undertaken study includes data on dengue cases and rainfall level collected in five districts in Selangor. Data were preprocessed using the Decimal Point Normalization before being fed into the training model. Predicted results of unseen data show that the LS-SVM prediction model outperformed the Neural Network model in terms of prediction accuracy and computation time in their conclusion coined that a hybrid algorithm could be applied in the future to obtain the optimal parameter, thus improve the prediction accuracy. Over the few years ago Islem Ouanes MD (2012) presented a model to predict short-term death or readmission after intensive care unit discharge and he said that many critically ill patients experience clinical deterioration or death shortly after discharge from the intensive care unit (ICU). Sangeeta Grover (2014) introduced a machine learning model for Predicting an Influenza Epidemic Based on Twitter Data, this model is a time series classification and prediction for identification of the epidemic stage based probabilistic model of vocabulary (BOWs) used in different stages of the epidemic shared online by the act of tweeting as described in below;

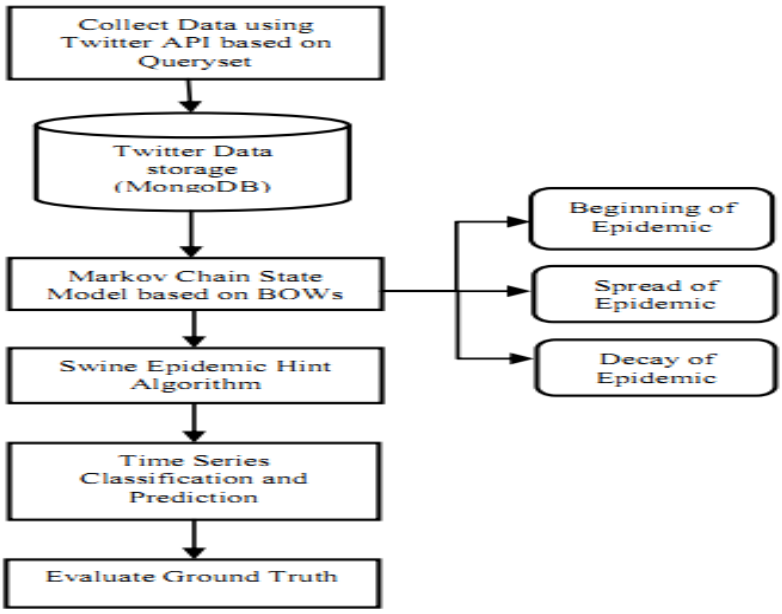

Figure 2.4 Proposed model and its components example (Sangeeta Grover, 2014).

Orlando P. Zacarias (2013) proposed a model for Predicting the Incidence of Malaria Cases in Mozambique Using Regression Trees and Forests. Researcher had stated that Malaria remains a significant public health concern in Mozambique with disease cases reported in almost every province. The research was aimed to investigate the prediction models of the number of malaria cases in districts of Maputo province. By using the data including administrative districts, malaria cases, indoor residual spray and climatic variables temperature, rainfall and humidity. The regression trees and random forest models were developed using the statistical tool R, and applied to predict the number of malaria cases during one year, based on observations from preceding years. Models were compared with respect to the mean squared error (MSE) and correlation coefficient. Indoor Residual Spray (IRS), month of January, minimal temperature and rainfall variables were found to be the most important factors when predicting the number of malaria cases, with some districts showing high malaria incidence. Additionally, by reducing the time window for what historical data to take into account, predictive performance can be increased substantially.

In the same range of period the Jyoti Soni U. A (2011) conducted a research which intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use in today's medical research particularly in Heart Disease Prediction and has concluded that the number experiments that have been conducted to compare the performance of predictive data mining technique on the same dataset and the outcome reveals that Decision Tree outperforms and sometime Bayesian classification is having similar accuracy as of decision tree but other predictive methods like KNN, Neural Networks, Classification based on clustering are not performing well. In his second conclusion said that the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction.

Currently, Boshra Bahrami (2015) carries a research on evaluating the different classification techniques in heart disease diagnosis. Classifiers like J48 Decision Tree, K Nearest Neighbors (KNN), Naive Bayes (NB), and SMO are used to classify dataset. After classification, some performance evaluation measures like accuracy, precision, sensitivity, specificity, F-measure and area under ROC curve are evaluated and compared. In his comparison results showed that J48 Decision tree is the best classifier for heart disease diagnosis on the existing dataset. I.S.Jenzi (2013) Conducted a study to develop a heart disease prediction system by using data mining techniques with the aimed of identify useful patterns of information from the medical data for quality decision making.

The Decision Tree and Naïve Bayes as Data mining techniques were used on building his classifier model used for predicting heart disease so that the result obtained from the classifier enabled to establish significant patterns and relationships between the medical factors relating to heart disease. there are also some model developed using regression such as as dental caries prediction model by data mining regression model Constructed using regression as data mining technique but he described that logistic regression models and linear discriminant analysis are not always appropriate because of it well unknown that when applying logistic regression analysis to small samples or unbalanced independent variables, asymptotic solution of the odd ratio obtained by the maximum like hood estimate is not guaranteed (Tamaki, 2009) .

Boshra Bahrami (2015) states that the millions of people die of heart disease annually, application of data mining techniques in heart disease diagnosis seems to be essential and because of that researchers presented a Prediction and Diagnosis of Heart Disease by Data Mining Techniques with the objective to evaluate different classification techniques in heart disease diagnosis where Classifiers like J48 Decision Tree, K Nearest Neighbors(KNN), Naive Bayes(NB), and SMO are used to classify datasets. After his classification, some performance evaluation measures like accuracy, precision, sensitivity, specificity, F-measure and area under ROC curve are evaluated and compared. The comparison results showed that J48 Decision tree is the best classifier for heart disease diagnosis on the existing dataset for having good understanding about their research you would like to respect the following Sequential overview of proposed approach for Prediction and Diagnosis of Heart Disease

Figure 2.5 Sequential overview of proposed approach (Boshra Bahrami M. H., 2015).

Shubpreet (2015) in his research where he demonstrated the data mining prediction models were developed such as a proposed a system for heart disease prediction using data mining techniques, statistical and data mining aspects on Kidney stones: a systematic review and met analysis, design of a predictive model for heart disease detection to enhance their liability of heart disease diagnosis, early prevention and detection of skin cancer and lung cancer risk using data mining, performance evaluation of different data mining classification algorithm and predictive analysis.

In this researcher, there is a dream to have a hybrid model. In this regard the most of hybrid done must be reviewed , this is paramount important to the researcher to be sure that there is no other work done purely similar to this and all technologies applied in hybrid models will be observed so that this will be guided.

The idea behind combined (Hybrid) prediction models is to derive advantages of individual model's best features to obtain the best possible results in a given problem/situation (Shubpreet Kaur a. D., 2015).

(N K Kameswara Rao, 2014) proposed as new hybrid algorithm, Epidemic Disease Prediction and Identification (EDPI) algorithm for combining the decision tree and association rule mining to predict the changes of getting Epidemic Disease in some selected areas. He notified that the prediction of disease in this algorithm can be shows by the relationships between desired parameters.

III. DISCUSSION

The classification and regression techniques in the literature are enormous as clarified by theoretical, methodological and historical evidence, more views on the research done such as Tamaki (2009) presented a dental caries prediction through the data mining regression model using regression as the one of data mining technique, but he observed that logistic regression models and linear discriminant analysis are not always appropriate because of its well unknown that when applying a logistic regression analysis of small samples or unbalanced independent variables, asymptotic solution of the odd ratio obtained by the maximum like hood estimate is not guaranteed. Generally, the most researches work conducted the outbreak disease prediction via mining models and had paid attention in models comparison with the purposes of looking the best one in term of accuracy not to identify a hybrid model. For instance, a model for predicting future dengue outbreaks incorporating LS-SVM has been presented and proved that LS-SVM was capable to obtain the better generalization ability compared to NNM, thus improving the prediction accuracy Yuhanis Yusof (2011) but concluded by urging that the hybrid algorithms will be needed in the future for obtaining the optimal parameter and improve the prediction accuracy in respect of new factors can happen anytime.

Furthermore, as cited by Shubpreet Kaur D. R (2015) in his research that the data mining prediction models were developed such as a system for heart disease prediction using data mining techniques; statistical and data mining aspects on Kidney stones; a systematic review and met analysis; design of a predictive model for heart disease detection to enhance their liability of heart disease diagnosis; early prevention and detection of skin cancer and lung cancer risk using data mining; performance evaluation of different data mining classification algorithm and predictive analysis etc. unfortunately, all these models mentioned none of them is considered as a hybrid model so that some of the outputs were not powerful as presented in their results and recommendations. The main advantages and disadvantages of different classification techniques were identified Divya Tomar (2013) so that single technique is not considered purely perfect because of these advantages and disadvantages clarified.

IV. CONCLUSION

In this paper review and analysis has been conducted on enormous classification and regression prediction models techniques and the objective is to contribute in the theoretical, methodological and historical gaps observed during related work done. The researcher noticed that many classification and regression prediction models techniques are available but relied on single techniques but also the available hybrid techniques are still few and more are needed so that there is need to include information on the theoretical, methodological and historical by establishing more complex model to increase the accuracy of predicting disease outbreaks. An analysis critique has been presented detailing the weaknesses of the varies classification and regression models where after the establishment of a comprehensive data mining hybrid model are mostly that it will be able to predict disease outbreaks with a maximum accuracy. This therefore is the most of previous models focused on using single data mining techniques which are classification or regression etc. Finally, we request for further research to address the various weaknesses as stated in this review paper and the idea behind combined (Hybrid) prediction models is to derive advantages of individual model's best features to obtain the best possible results in a given problem/situation (Shubpreet Kaur a. D., 2015) .

REFERENCES

[1]. A.Martin, V. G. (2011, February). A HYBRID MODEL FOR BANKRUPTCY PREDICTION USING GENETIC ALGORITHM, FUZZY C-MEANS AND MARS . *International Journal on Soft Computing (IJSC), 2*(1).

[2]. Abdoulaye Mar Dieye, J. O. (2015). *Socio-Economic Impact of Ebola Virus Disease in West African Countries.* Lagos: United Nations Development Group – Western and Central Africa (UNDG-WCA) Regional Directors Team (RDT).

[3]. Benning, S. D. (n.d.). *class/6k180_park/Student-Reports/sbenning/* . Retrieved May 25, 2015, from http://www.biz.uiowa.edu.

[4]. Berndtsson, M. &. (2008). *Thesis Projects: Aguide for Students in Computer Science and Information Systems.* London: Springer-Verlag London.

[5]. Boshra Bahrami, M. H. (2015, Feb). Prediction and Diagnosis of Heart Disease by Data Mining Techniques. *Journal of Multidisciplinary Engineering Science and Technology (JMEST), 2*(2).

[6]. Boshra Bahrami, M. H. (2015). Prediction and Diagnosis of Heart Disease by Data Mining Techniques. *ournal of Multidisciplinary Engineering Science and Technology (JMEST), 2*(2).

[7]. Chandra, S. D. (2013). *Software Defect Prediction Based on Classification Rule Mining.* Department of Computer Science and Engineering National Institute of Technology Rourkela Rourkela { 769 008, India.

[8]. D. Hand, H. M. (2001). *Principles of data mining.* MIT.

[9]. Divya Tomar, S. A. (2013). A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology, 5*(5), 241-266.

[10]. Divya, S. A. (2011, June 3-4). Weighted Support Vector Regression approach for Remote Healthcare monitoring. *IEEE-International Conference on Recent Trends in Information Technology* (pp. 978-1-4577). Chennai: 0590-8/11/$26.00 © 2011 IEEE MIT.

[11]. Enserink, M. (2010). *No vaccines in the time of cholera Science.*

[12]. Fayyad, P.-S. S. (1996). Advances in Knowledge Discovery and Data Mining. *AAAI Press / The MIT Press*, 1-34.

[13]. Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods.*

[14]. Frontières, M. S. (2008). *Filovirus Haemorrhagic Fever Guideline.* Barcelona, Spain: MS.

[15]. Gaurav Taneja, A. S. (2014, Semptember 9). STUDY OF CLASSIFIERS IN DATA MINING. *International Journal of Computer Science and Mobile Computing, 3*(9), 263 – 269.

[16]. Gustafson, D. H. (1993). Measuring quality of care in psychiatric emergencies: Construction and evaluation of a Bayesian index. *Health Services Research*.

[17]. H. C. Koh, G. T. (2005). "Data Mining Application in Healthcare. *ournal of Healthcare Information Management, 19*(2).

[18]. Hailu, T. G. (2015). *Comparing Data Mining Techniques in HIV Testing Prediction* . Addis Ababa, Ethiopia : Intelligent Information Management.

[19]. Han, K. (2006). *Data Mining: Concepts and Techniques* (2nd Edition ed.). Morgan Kaufmann Publishers,San Francisco.

[20]. Huang, F. I. (2004). *Disease Outbreak Investigation.*

[21]. I.S.Jenzi, P. P. (2013, March). A Reliable Classifier Model Using Data Mining Approach for Heart Disease Prediction. *International Journal of Advanced Research in Computer Science and Software Engineering, 3*(3).

[22]. Islem Ouanes MD, C. S. (2012). A model to predict short-term death or readmission after intensive care unit discharge. *Journal of Critical Care*, 422.

[23]. Jiawei Han ., M. K. (2009). *Data Mining Concepts and Techniques.* Morgan Kaufamann Pubisher.

[24]. Jyoti Soni, U. A. (2011). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *International Journal of Computer Applications (0975 – 8887), 17*(8).

[25]. Jyoti Soni, U. A. (2011). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction . *International Journal of Computer Applications*, 0975 – 8887.

[26]. KAYUMBA, D. p. (July 2011 – June 2012). *RWANDA BIOMEDICAL CENTER ,ANNUAL REPORT*. Kigali: Ministry of Health.
[27]. Kumar, P. C. (2007). *Kumar and Clark"s clinical medicine.* England : Saunders Ltd.
[28]. Lewnard J. A., M. L.-M. (2014). Dynamics and Control of Ebola Virus Transmission in Montserrado. *a mathematical modelling analysis, Lancet Infectious Diseases* .
[29]. M, S. (2012). Cancer Control in Developing Countries: Need for Epidemiological Surveillance based on Health Information Systems and Health Services Research. . *Journal of Health Informatics in Developing Countries*, 356- 360.
[30]. Mackay, J. M. (2004). Atlas of Heart Disease and. *Nonserial Publication*.
[31]. Mahendra Tiwari, R. S. (2013, Feb). An Empirical Study of Applications of Data Mining Techniques for Predicting Student Performance in Higher Education. *International Journal of Computer Science and Mobile Computing , 2*(2), 53 – 57.
[32]. Megha Gupta, N. A. (2010). CLASSIFICATION TECHNIQUES ANALYSIS . *NCCI 2010 -National Conference on Computational Instrumentation.* CSIO Chandigarh, INDIA.
[33]. Mythili T, D. M. (2013, 4 16). A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL). *68*(16), 0975 – 8887.
[34]. Mythili T., D. M. (2014, April). A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL). *International Journal of Computer Applications (0975 – 8887), 68*.
[35]. N K Kameswara Rao, D. G. (2014). A hybrid Algorithm for Epidemic Disease Prediction with Multi Dimensional Data. *nternational Journal of Advanced Research in Computer Science and Software Engineering, 4*(3).
[36]. Noor Diana Ahmad Tarmizi, F. J. (2013, August). Malaysia Dengue Outbreak Detection Using Data Mining Models. *Journal of Next Generation Information Technology(JNIT), 4*(6).
[37]. Organization, W. H. (2014, 1 13). *Clinical management of patients with viral haemorrhagic fever: a pocket guide for the front-line health worker.* Retrieved December 26, 2015, from Available at: http://apps.who.int/iris/bitstream/10665/.
[38]. Orlando P. Zacarias, H. B. (2013). Predicting the Incidence of Malaria Cases in Mozambique Using Regression Trees and Forests. *International Journal of Computer Science and Electronics Engineering (IJCSEE), 1*(1), 2320–4028.
[39]. Parvez Ahmad, S. Q. (2015, June). Techniques of Data Mining In Healthcare: A Review. *International Journal of Computer Applications (0975 – 8887), 120*.
[40]. Richards G, R.-S. V. (2001). *Data mining for indicators of early* . Artif Intell Med.
[41]. Sahana, D. C. (2013). *Software Defect Prediction Based Classification Rule Mining.* Rourkela: epartment of Computer Science and Engineering National Institute of Technology Rourkela.
[42]. Sangeeta Grover, G. S. (2014, July). Prediction Model for Influenza Epidemic Based on Twitter Data. *International Journal of Advanced Research in Computer and Communication Engineering, 3*(7).
[43]. Senait Kebede, S. D. (2010, march 1). TRENDS OF MAJOR DISEASE OUTBREAKS IN THE AFRICAN REGION, 2003-2007. *East Africa/1 Joumal ofPublic Health, 7*(1).
[44]. Shubpreet Kaur, a. D. (2015). Future Trends of Data Mining in Predicting the Various Diseases in Medical Healthcare System . *International Journal of Energy, Information and Communications* , 17-34 .
[45]. Shubpreet Kaur, D. R. (2015). Future Trends of Data Mining in Predicting the Various Diseases in Medical Healthcare System . *International Journal of Energy, Information and Communications*, 17-34.
[46]. Shukla, S. B. (2014). A Literature Review in Health Informatics Using Data Mining Techniques. *International journal of software and hardware research engineering, 2*(2).
[47]. Sitar-Taut, V. e. (2009). "Using machine earning algorithms in cardiovascular disease risk evaluation. *"Journal of Applied Computer Science & Mathematics*.
[48]. U. Fayyad, G. P.-S. (1996). The KDD process of extracting useful knowledge form volumes of data.commun. *ACM, 39*(11), 27-34.
[49]. WHO. (n.d.). Retrieved May 25, 2015, from http://www.who.int/influenza/human_animal_interface/en.
[50]. Yadav, S. K. (2012). "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification. *World of Computer Science and Information Technology (WCSIT)*.
[51]. Yo Tamaki, Y. N. (2009). Construct of dental caries prediction model by data mining. *Journal of Oral Science, 51*(1), 61-68.
[52]. Yuhanis Yusof, Z. M. (2011, August). Dengue Outbreak Prediction: A Least Squares Support Vector Machines Approach. *International Journal of Computer Theory and Engineering,, 3*(4).
[53]. Yuhanis Yusof, Z. M. (2011). Dengue Outbreak Prediction: A Least Squares Support Vector Machines Approach. *International Journal of Computer Theory and Engineering.*

Biographies and Photographs

Hakizimana Leopord is a PhD student in the Department of Computing at the Jomo Kenyatta University of Agriculture and Technology He has a B.Sc. In Information Technology (University of Rwanda/Umutara Polytechnic), Graduate Diploma in Leadership Development in ICT and the Knowledge Society (Dublin City University, Europe); MSc in Computer Science (Sikkim Manipal University, India). His research interest is in the field of Data Mining.

Dr. Wilson Kipruto Cheruiyot is the current director of Jomo Kenyatta University of Agriculture and Technology Kigali Campus-Rwanda. He has the following academic qualifications Bsc (Hons); PGD-E (Egerton University, Kenya); Msc in Computer Application and Technology (central south university of technology Hunan, china); PhD in Computer Application and Technology (Central South University, China)

Dr. Stephen Kimani is the current director School of Computing and Information Technology of the Jomo Kenyatta University of Agriculture and Technology in Kenya. He has previously been a researcher at CSIRO (Australia) and Sapienza University of Rome (Italy). He has the following academic qualifications: BSc in Mathematics and Computer Science (JKUAT); MSc in Advanced Computing (University of Bristol, UK); PhD in Computer Engineering (Sapienza University of Rome, Italy).