EasyChair Preprint

№ 3656

Heart Patients Datasets Analysis using Weka Tool

Fazeela Maryum

June 21, 2020

Heart Patients Datasets analysis using Weka tool

Fazeela Maryum

Fazeelamaryum7@gmail.com

Riphah International University Lahore

Abstract- In health department human task are increase and different techniques are available for doctors to cure patients. In Medical business is a worldwide business and people want research on latest patient's data. Health field become an outstanding field in the wide spread territory of restorative science and they're very popular field now a days. Huge amount of raw data in medical side are need to be convert use full data this data has hidden information and patterns to which we can learn and make accurate decisions in future. These decisions are applying to patients and cure the patients, with the help of data mining they reduce patients' tests, money and time. Lack of analysis of data mining tool as indicated by successful test results together with the covered-up data, so and such a framework is created utilizing information-digging calculations for arranging the data patterns with help to more, identify heart patients' issues. The data need to be classified and make visualization using best available tool. Using data mining tools, they provide better solutions. in my case study i use 5 major classification algorithms to find accuracy in heart patient's diseases, these algorithms are: k-Nearest Neighbors (KNN), Linear Regression (LR), Random Forest (RM), Naive Bayes (NB), Support Vector Machine (SVM). These data mining methods can provide solution for heart patients. In this paper they analysis some few prediction and parameters for heart patients also suggests HDPS (Heart Diseases Prediction System) put together aggregate with respect to the data mining approaches.

Keywords: **Naïve Bayes, Data Mining, Heart Disease, Random Forest, Support Vector Machine, Weka Tool, k-Nearest Neighbors, Linear Regression.**

I. INTRODUCTION

Data mining method is just to extract useful or meaningful data from raw data or huge data. These meaning full data is use to make decision making and we visualization techniques are implement to learn patterns. In health department they provide solution to heart patients and predict before time. Social insurance information mining has a very huge potential as indicated by find the covered up designs among the informational indexes about the medicinal space. There are many algorithms are available to learn health department to learn better way of heart patient's data. DM applications in social insurance can have great potential and It computerizes the procedure of finding prescient data in enormous databases. DM provides automate the system which they provide predictive information in huge database. In DM disease prediction plays an important role. Sicknesses under the coronary illness umbrella incorporate vein infections, for example, coronary corridor ailment, heart musicality issues (arrhythmias) and heart surrender you're brought into the world with (inborn heart absconds), among others. Heart conditions, for example, those that influence your heart's muscle, valves or cadence, likewise are viewed as types of coronary illness. Cardiovascular illness, for the most part, alludes to conditions that include limited or blocked veins that can prompt a coronary failure, chest torment (angina) or stroke. Data Mining health care department is an very important task because for doctors it is very easy to find attributes that can which are more important for patients diagnosis like patients symptoms , age, weight etc. [2]. This will make the doctors are diagnose the patients' disease are more efficiently and accurately. Finding patient's data history they male patterns and implement decisions on it. It makes utilization of calculations after concentrate the data

and designs determined by the information disclosure in databases process and different phases of information revelation in databases process are featured Using data mining Knowledge discovery are with in databases. In this case study I follow KDD process which mention figure 1. KDD process is Knowledge Discovery are with in Databases.

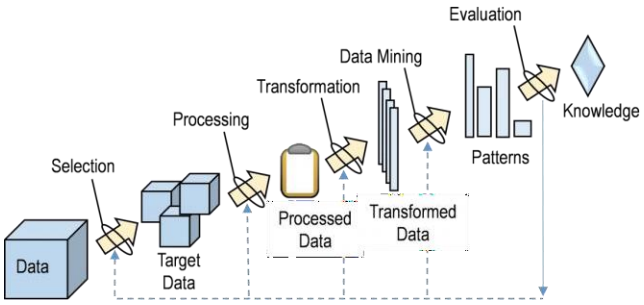

Figure 1: KDD (Knowledge Discovery within Databases) Process [1]

There are various steps which is followed by KDD process and their methods. In different data bases there is selection stages that obtain different resources of data. This process they remove blank unused data and clean the data this is called preprocessing after that I applied different techniques to get desired output. At long last into the between the implication organize and that will show the outcome after end client in a significant way. Preprocessing of databases comprises of Data cleaning and Data Integration.

II. *DATA MINING* TECHNIQUES

There are many types of Data Mining Techniques but some of are mention below

1. **Association**
2. **Clustering**
3. **Classification**
4. **Numeric**

Association rule mining:

Association is an information mining capacity that finds the likelihood of the co-event of things in an assortment. Association and Correlation Patterns Simultaneously in Database. Dynamic: Data mining means to find valuable examples in enormous datasets. An Association rule having support. certainty more noteworthy than or equivalent to a client indicated least help edge and separately a base certainty limit and information mining process connection certainty measures are embraced at the same time and the connections between co-happening things are communicated as Association rules. Association rules are frequently used to dissect deals exchanges.

Clustering: In data mining classification is item assigns into a collection of classes or mark categories. The main goal is to classified item and accurately predicting the targeted class in each class of data. For instance, an arrangement model could be utilized to recognize advance candidates as low, medium, or high credit dangers. Clustering is just a process that make the objects into same class and the data in one class is more alike into each other groups.

Classification: In Statistics, ML Machine Learning is a supervised learning technique in which they learn program of computer where data is input given to it after that they use to learn the classify new observations.

Numeric prediction: In Data Mining, the expression "Expectation" refers to determined suppositions of specific unforeseen developments made based on accessible handled information. The expectation itself is determined from the accessible information and demonstrated as per the current elements. The expectation in information mining is to recognize information focuses simply on the depiction of another related information esteem. Expectation determines the connection between a thing you know and a thing you have to anticipate for future reference.

Out of this Five Algos learning techniques in my case study I identify performed better and which accuracy are good. Data type is depending on application or data mining methods. Which is fitted to be utilized in the strategies and measuring data mining embarrassments rely upon the kinds of information to stand utilized and the choice about data mining procedure which is generally proper for the data used.

III. MACHINE LEARNING (ML)

Machine Learning (ML) are data science method and it working is depends on past experience and learning. ML are developed algo that data directly with small little and no human interfere. ML do many tasks like classification, decision making or predication. Now days people want to study on ML [3] and make their research on its Ml is most

important aspect in data science. First input values are training set of data and begin in ML. In this stage, the ML Algo utilizes the preparation dataset in the wake of gaining from the information and structure designs. The learning stage yields a model so much is utilized by method for the testing stage. In data testing sets are and apply in training phase the resulting for analysis. The general execution in regards to the test dataset exhibits the model's capacity in similarity with playing out its errand against information.ML they provide statement in to regarding stamens.

AI [4] calculations fabricate a numerical model dependent on test information, known as making information and so as to settle on forecasts or choices without being clearly modified to play out the commission. The investigation of scientific reorganization conveys strategies, hypothesis and request areas to the field of AI. AI is firmly identified with computational sizes, which centers on making forecasts utilizing Machines. AI is the logical study of calculations and factual models that PC frameworks use to play out a particular task without utilizing express guidelines, depending on examples and deduction. AI calculations are utilized in a wide range of uses, for example, email disentanglement and PC vision, where it is troublesome or infeasible to build up a regular calculation for adequately playing out the job.

IV. OPEN SOURCE SOFTWARE

Open source software is nothing just a free package software. Open source software is available publicly, people are use freely. Planned subsequent to denying everybody the privilege as per misuse the product. Its modification they are freely use and everyone are freely use. Most of companies are provide this software are freely available and its modifications. In open source software are freely there and make modification to develop extensions through which a freely available and publicly license is available.

Open-source software is a kind of program where in source code is discharged under a license in which the copyright holder awards clients the rights to study, appropriate and change the product to anybody and for any reason.

V. HEART DISEASES

Most of death around the world is due to heart disease and it is very important to learn this issue in the beginning or initial stages to cure patients. Initial stages cure patient is very tough for heart doctor to cure this type of patients. For diagnosing and identification doctors are adopting many scientific methodologies and technologies. The productive treatment is consistently official to right also, exact conclusion. In few times many doctors are fail to analysis of current patients' situation. That is way we use ML to generate accurate result and make better decision on it [3].

VI. HEART DISEASE DATASET

I used heart patient's data sets download from kaggal website [4] which provide free data sets of patents to analysis classification techniques on it. Our data sets have 260 instance and 50 attributes. However, I use only 14 attributes are use in my case study. For accurate prediction 14 attributes are enough for heart disease patients [5]. The data sets are huge but our used data is just 260 and its 260 there is also missing values, blank values and prepressed it to use 14 attributes.

VII. OVERVIEW OF DATA MINING TOOLS

In Data mining there is many tools are available on internet. They ranging from marketing which use in AL field, Products measures and different fields of science. Form the decades there are many tool are developed to solve data mining's problems there is wide range of tools and this tool are also freely available. Every tool have their advantages and disadvantages and which data they easily compute its tools dependency. Inside Data mining, there is a gathering of devices that have been created by an exploration network and information investigation lovers; he is sans given the value utilizing one on the current open-source licenses. In data mining's every type of data mining tool are develop some of our text mining specialist some are image or some of video mining expert. There is large number paid tools are also available or some of are open source tools. Because of its boundless utilization and multifaceted nature engaged with building data mining applications, countless Data mining instruments hold been created over decades. Open source tools are freely available not only single person or due to one organization this work done due to all over the world and international organization to work on it and contributes to development team and tools. This development style is providing the means of extracting data from databases. Data Mining's tools are predicting futures patterns, trends behaviors and allowing to business make proactive and accurate result. For accurate result and accurate patterns for future result then use powerful tool and algorithms to find results. As the quantity of open apparatuses proceeds by developing the decision of the most reasonable instrument turns out to be progressively troublesome. Many open source tools are available for data mining some of are given below.

- **Weka**
- **Orange**
- **DataMelt**

- **Apache Mahout**
- **ELKI**

We use Weka tool in this paper. Data mining tools like Weka are user friendly and its accuracy are best and this tool are freely available on internet. The first step is to selecting open source tool for our datasets which we want to check his accuracy or being tested. There are lot of data mining tool are available on internet. In my case study I use weka for analysis. Weka is open source tool.

VIII. WEKA

Weka is an open source tool of Machine Learning ML his environment is for Waikato environment [10]. This ML tool are introduced by Waikato University, New Zealand. In Data Mining is processing data which is classification, association regression and data visualization. Its also help to selecting features section of new algorithms are implementing in Weka. Weka provide user-friendly GUI which loaded data sets from file and from URL source. His data set file accepted CSV, Arff format, C4.5's format and LIB SVM's format. Weka provide evaluation criteria for their calculations like recall, precision, confusion matrix, false negative, true positive, etc. There some advantages of Weka tool are portable, GUI, platform independent and pen source which contained very fast collection of different other data mining algorithms.

IX. COMPARATIVE STUDY

The study methodology is to collecting regarding a set of free sources of data mining. The discovery knowledge tools are being tested and select the data which you want to use. After selecting datasets classification algorithms are implementing on and test on selecting tool. Parallel checking of tool performance. Shows the general technique followed for satisfying the objective of its examination. Which i am followed in shown in Figure 2

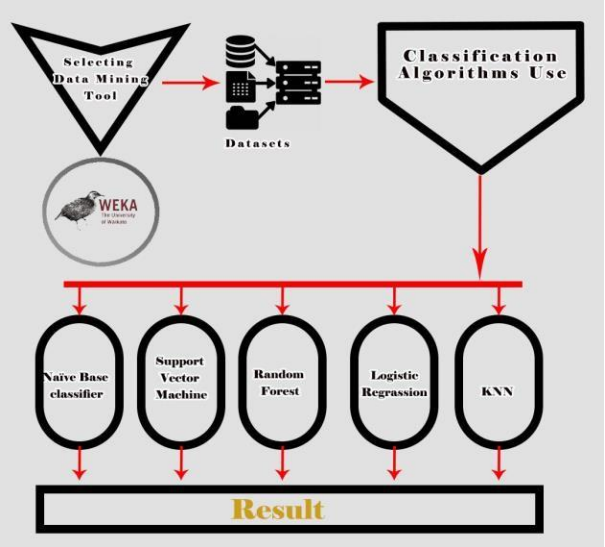

Figure 2: Tools Implementation Methodology [2]

X. RESEARCH QUESTION

In this paper, our research depends on heart patients disease with utilizing Weka Tool and Classification are implementing on it.

- **Which Classification algorithm are more accurate than other 5 algorithms?**
- **Our data sets which algorithm are suitable for our problem?**

Recall and Precision

a) Naïve Bayes

Dimensionality is high inputs then the Naive Bayes classifier method is mostly suited. Most of the problem is occur that the attribute is generally depending on each other and they cannot be applied. Naive Bayes Algorithm are hard to debugging and it is very difficult to understand [6]. In robotics field Naive Bayes are used. For our data sets precession 0.87 and recall 0.76 in Weka tool its accuracy is 83.13%.

b) Support Vector Machine

SVM stands for Support Vector Machines. Its ML Classification algorithm and very powerful algorithm. they provide classification pattern. Most of people use this algorithm for great deal of attention now a days. SVM Is supervised Learning Algorithm technique. It is also a good in medical field of heart patients. The Support Vector Machines calculation predicts the event about coronary illness by capacity on plotting the sickness. Predicting

characteristics in regards to the multidimensional hyper plane or orders the classes ideally by making the organization among two info clusters [5]. SVM algorithm have high accuracy and nonlinear regarding use features called kernels. For our data sets precession 0.874 and recall 0.836 in Weka tool its accuracy 86.15%.

c) Random Forest

RF RM is unpruned classification trees and it gave accurate and best performance in practical problems. First calculation for arbitrary choice woods was made by Tin Kam Ho and Its work fast but not noisy and blank data's sets. Its generally best performance over other tree-based algorithms. RM Build their performance on a number of trees. Random forests or irregular choice timberland are a troupe learning strategy for order, relapse and different undertakings that works by developing a huge number of choice trees at preparing time. Random choice woodlands right for choice trees' propensity for over fitting to their preparation set. The Outputting the class that is the method of the classes (grouping) or mean forecast (relapse) of the individual trees. For our data sets precession 0.889 and recall 0.834 in Weka tool its accuracy 86.83%.

d) Logistic Regression

Logistic regression examination is a important tool for representative and breaking down data. For model, connection between rash driving and number of street accidents by a driver is best concentrated through regression. It's a S-formed curve that can take any genuine valued number and guide it into an incentive somewhere in the range of 0 and 1, however never precisely at those points of detention. Logistic regression is utilized to depict information and to clarify the connection between one ward twofold factor and at least one supposed, ordinal, and temporary or amount level free factors. For our data sets precession 0.878 and recall 0.792 in Weka tool its accuracy 84.49%.

e) KNN (K-nearest neighbor)

KNN stands for K-nearest neighbor and there are under the classification algorithm that's find K objects in group of training sets which close to the test values. It classifies unlabeled object. Compute the distance between label object and object after that KNN are identified. Mostly depends on classification accuracy is the choice value of k. It is better to using KNN classifier [8]. For huge data sets KNN reduce the error from datasets. Selecting KNN should be possible tentatively and where a number concerning designs taken out from the preparation set can be classified applying the rest of the preparation designs for various abilities over KNN. KNN give least error. If KKN provides and shares various KNN then very-neighbor of class weights are added together. The result of these weights is sum and the score of that class according to the test Doc [9]. For our data sets precession 0.795 and recall 0.935 in Weka tool its accuracy is 85.83%.

f) Precision

Precision is called Positive Predictive values and it's defined as the total Average of Probability relevant.

Formula:

Precision = (No. Total True Positive)/ (No. Total True Positive + False Positive).

g) Recall

Recall is defined the total Avg of probability of retrieval complete.

Formula:

Recall= (True Positive)/ (True Positive + False Negative)

Result Analysis:

In our dataset we implement 5 classification algorithms and we see Random forest accuracy have more accurate than other.

Classification Algorithms	Precession	Recall	Accuracy
Naïve base classifier	0.872	0.766	83.12%
SMO or Support Vector Machine	0.874	0.836	86.15%
Random Forest	0.889	0.834	86.83%
Logistic	0.878	0.792	84.49%
KNN	0.795	0.935	85.83%

Table 1: Classification Algorithms with orange tool result

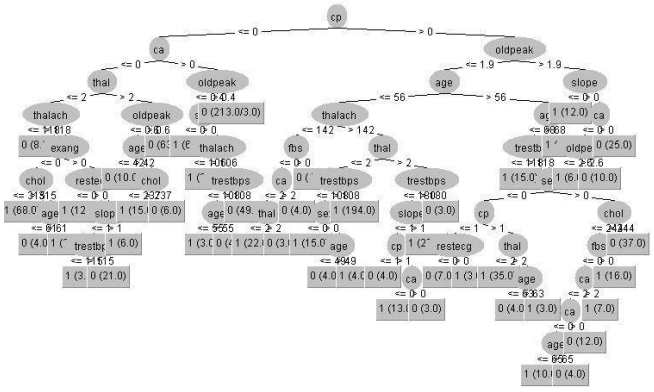

Figure 3: Tree Visualization In using J48 KNN

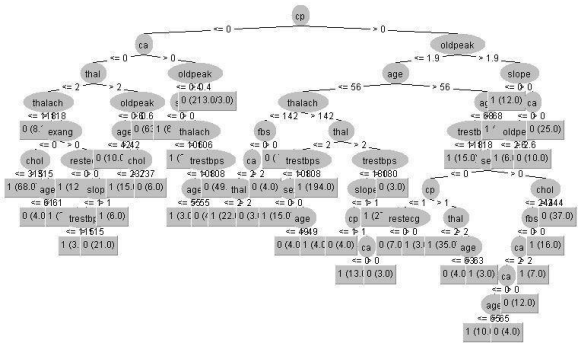

Figure 6: Tree Visualization in using J48 SVM

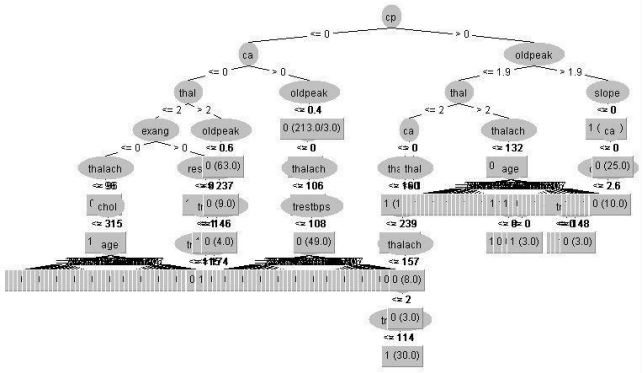

Figure 4: Tree Visualization in using J48 RF

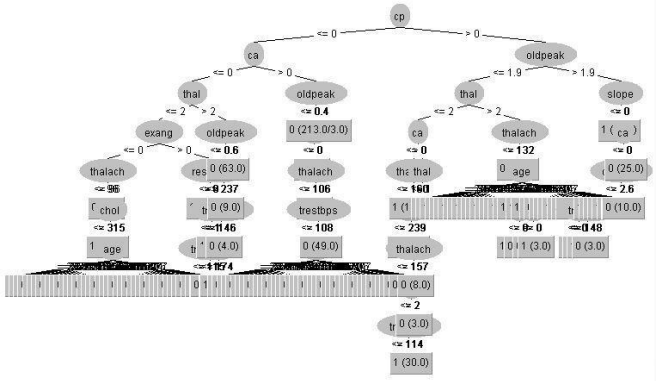

Figure 7: Tree Visualization in using J48 NB

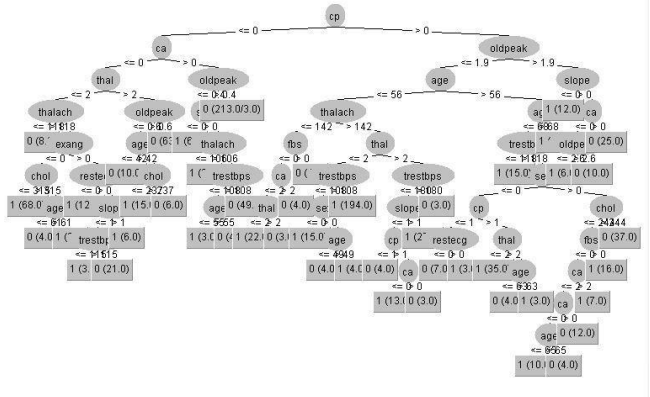

Figure 5: Tree Visualization in using J48 LR

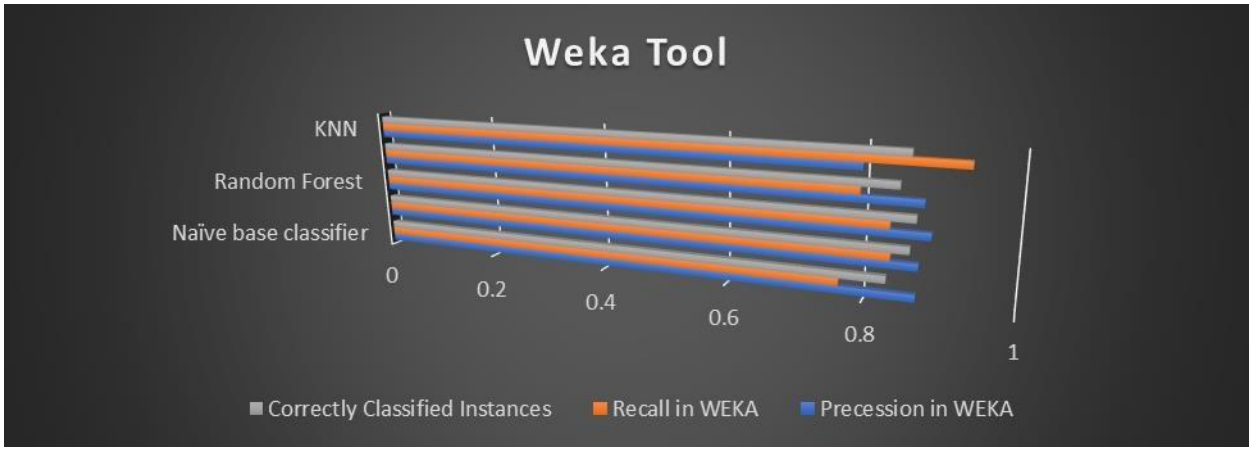

Figure 8: Using Weka tool Classification Algorithm Graph for Precession and Recall of Heart Disease

XI. CONCLUSION AND FUTURE SCOPE

For Hidden Knowledge in term of diseases data mining are very helpful. It can remain to predict and analyze in the future and predict the future behavior of diseases. To class label the set of unclassified cases the classification is one the data mining method. In Weka recall and precession are accurate in Weka tool. Weka tool provide best performance in our data sets. The main objective is our case study to compare these 5 classification algorithms with each other which one is best performance and his accuracy is more than other. I use 5 algorithms like KNN, LR, RF, SVM, NB, I implement same data sets on these algorithms and sort out which one is best. in my result RF is more accurate than other. Random Forest is a gathering regulated AI strategy.

This paper introduces a deliberate study of work done in Random Forest territory and Random Forest will quicken investigate in the field of Machine Learning. Random Forest has colossal capability of turning into a prevalent method for future classifiers and its presentation has been seen as similar with group strategies discharge and boosting. In future not only classification use I use clustering and association techniques to compare my result also check the accuracy and compare the performance of various data mining tools and techniques.

REFERENCES

[1] Iyer A, Jeyalatha S, Sumblay R. Diagnosis of diabetes using classification mining techniques. IJDKP. 2015; 5(1):1–14.

[2] Reclaiming Liberalism (ISBN 1-86197-797-2) is a book written by a group of prominent British Liberal Democrat politicians and edited by David Laws and Paul Marshall in 2004.

[3] Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine, Single Volume: Expert Consult Premium Edition Originally published: February 25, 2011.

[4] G. KDD '18: Proceedings of the 24th ACM SIGKDD International Conference .November 16, 2018. 24th ACM Conference on Knowledge Discovery and Data Mining - KDD 2018.

[5] Gosain, A.; Kumar, A., "Analysis of health care data using different data mining techniques," Intelligent Agent & Multi-Agent Systems, 2009. IAMA 2009. International Conference on , vol., no., pp.1,6, 22- 24 July 2009.

[6] Data Science from Scratch Book by Joel Grus. understanding data science. Originally published: 2015Author: Joel Grus. Data Science from Scratch Book by Joel Grus. understanding data science. Originally published: 2015Author: Joel Grus.

[7] Data Mining online user-generated content: using sentiment analysis technique to study hotel service quality.Proceedings of the 46th Hawaii International Conference on System Sciences.

[8] Data Science from Scratch Book by Joel Grus. understanding data science. Originally published: 2015Author: Joel Grus.

[9] Iyer A, Jeyalatha S, Sumblay R. Diagnosis of diabetes using classification mining techniques. IJDKP. 2015; 5(1):1–14.

[10] Competitions, Kaggle Kernels, Kaggle Datasets, Kaggle Learn, Jobs Board.Owner Alphabet Inc.On 8 March 2017, Google announced that they were acquiring Kaggle.

[11] Identifying Compromised Accounts on Social Media UsingStatistical Text Analysis
Auther:Dominic Seyler, Lunan Li, ChengXiang Zhai.

[12] Mining online user-generated content: using sentiment analysis technique to study hotel service quality. Proceedings of the 46th Hawaii International Conference on System Sciences.

[13] Stieglitz S, Dang-Xuan L. 2013. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. J Manage Inf Syst. 29:217–248.

[14] Crush It!: Why NOW Is the Time to Cash In on Your Passion.

[15] Get Found Using Google, Social Media :Originally published: November 18, 2009
Authors: Brian Halligan, Dharmesh Shah