# A Multiple Classifier Approach to Improving Classification Accuracy Using Big Data Analytics Tool

Geetha. P
Asst Professor, Dept of CSE.
Cambridge Institute of Technology
Bangalore
*geetha.cse@citech.edu.in*

Dr. Chandrakant Naikodi
Visiting Professor, Dept of CSE
Cambridge Institute of Technology
Bangalore
*nadhachandra@gmail.com*

Dr. Suresh L
Principal.&Professor
Cambridge Institute of Technology
Bangalore
*suriakls@gmail.com*

***Abstract:*** At the heart of analytics is data. Data analytics has become an indispensable part of intelligent decision making in the current digital scenario. Applications today generate a large amount of data. Associated with the data deluge, data analytics field has seen an onset of a large number of open source tools and software to expedite large scale analytics. Data science community is robust with numerous options of tools available for storing, processing and analysing data. This research paper makes use of KNIME, one of the popular tools for big data analytics, to perform an investigative study of the key classification algorithms in machine learning. The comparative study shows that the classification accuracy can be enhanced by using a combination of the learning techniques and proposes an ensemble technique on publicly available datasets.

***Keywords:*** *Big Data, Analytics, Knowledge Mining, Ensemble Learning*

*****

## 1. INTRODUCTION

With widespread advancement in the Information and Communications Technology industry, there is an ever growing dissemination and availability of big data. Big data generated as a by-product of digitization originates from a number of sources like online banking transactions, search queries, online log data, sensor data, and click streams. This data deluge, marked by huge size, complexity and variety has transformed the face of analytics. Big data has revolutionized the field of data analytics, due to which, more novel and intelligent decision support systems are required to uncover interesting aspects of such complex and massive data. Big data analytics enables the discovery of hidden aspects of data, by studying their complex interrelationships and behaviour [2]. These insights can in turn be utilized by organizations for strategic decision making and provide smarter growth opportunities. Big data has embarked an analytics spree, bringing it with a number of different technologies and tools that assist in processing such large amounts of data. These datasets require processing on a large scale, marked by their challenging characteristics. A number of analytics tools, catering to different types of data have emerged; needless to say that the analytics rage has set its foot onto a myriad of fields ranging from business, medicine, banking and a lot more. These automated tools can facilitate the process of analysis and encourage intelligent decision support systems [1]. The recent past has seen a surge in the number of automated tools whose usage has grown over the years , and has proved to be even better than their commercial counterparts, In particular, WEKA[6],Rapid Miner[17],Orange[4],KNIME[3], R[18] are tools that have support for powerful data analytics encompassing well known machine learning algorithms. This research paper uses the KNIME analytics platform to compare classification algorithms on publicly available datasets. While demonstrating the potential of the tool, the study also shows that the accuracy of the classified instances can be enhanced by training a model on a combination of these algorithms. This is proposed through a combination model supported by the KNIME ensemble machine learning.

## 2. TOOL DESCRIPTION - KNIME

The Konstanz Information Miner (KNIME)[3] is equipped with an open API system that allows for new nodes to be added to the application in a way that makes integration not only fairly easy, but also allows for an efficient means of adding information and functionality to the application [3].

It is an effective analytics tool that has been used by a number of naive users for an initiator in analytics. A number of tools are in use today, equipped with all learning algorithms. KNIME is easy to use with its clear and easy interface. It provides support for workflows to be integrated from other tools as well. For instance, it allows WEKA workflows to be easily integrated with the analytics platform for performing a comparative analysis of the performance of the tool, with respect to the efficiency of the tool with the same algorithm.

## 3. RESEARCH APPROACH

The method of study involves the following phases:
(1)A theoretical survey of the freely available open source software tools [1][7][8] for data mining and selection of KNIME for the study.
(2)Collection of different type of the datasets to be used for analysis.
(3)Analogusly pre-processing the datasets so as to reduce any preconceived notion and ensure equality of comparison
(4)Selecting a set of classification algorithms for the study.
(5)Study improved performance through a hybrid model for the same datasets.
(6)Evaluate and compare results

KNIME is an acclaimed tool for big data analytics. The graphical and comprehensible interactive interface makes it an easy choice for the study by considering its suitability of usage for novice users.

## 4. DATA MINING FUNCTIONALITY

Data classification [2] is a supervised learning technique in data mining. The first step is the learning phase. During the learning phase, a a classifier model is constructed based on the training data. The training data is a collection which has pre-defined labels of data classes, which helps in learning. In the second step, the trained model is used for classification. The accuracy of the classifier can be learnt from an unknown dataset known as the test dataset. The number of correctly classified instances can measure the accuracy of the classifier. The classifier model can be then used to predict the labels of unknown future data, once the accuracy is adequate.

## 5. DATASET DESCRIPTION

A set of three different datasets [5] are selected for the study to reduce bias and make a justifiable analysis of the variation in performance of the tool with different classification techniques. The datasets are characterised by the type of data, the number of attributes in each dataset, the type of the attributes (numeric, ordinal, categorical etc...), the intended data mining task and the number of instances in the dataset. Table 1 shows the datasets selected and downloaded from the UCI machine repository [5].The datasets have investigative acceptability with variation in the number of instances (150 to 32,500),[5] the type of the target attribute (nominal Vs numeric)[5] and the number of attributes(3 to 15)[5].This variation in characteristics of the datasets will offer a better comparative performance analysis of the different classification algorithms. This in turn can generate an arguable measure of accuracy statistics.

## 6. COMPREHENSIVE STUDY

This section of the research study focuses on the selection of the customary classification algorithms deep-rooted in the field of data mining and knowledge discovery and a practical analysis of algorithms performance on KNIME. Although the field of knowledge discovery and data mining (KDD)[2] has seen a tremendous invasion in the field of advanced data analytics and machine learning, the literature speaks of certain well known algorithms that is always a fore runner for any performance study. These algorithms can be scaled to the current big data era[15][16], to mine data from large datasets in a parallel architectural framework that is prevalent in all data processing environments. Notable of them for classification task include, Decision tree classifier [2], Naive Bayesian Classifier [2], K-Nearest Neighbour Classifier [2], and Neural network [2].These algorithms and its certain variants are available in almost all open source tools and hence the research choice in this study

### 6.1 Scoring Classification Algorithms

The algorithms are assessed and tested using two known techniques, namely the k-fold Cross Validation method and the Percentage Split method. The two different test modes help in capturing the improvements in accuracy with the tool when there is a transition from one test mode to another.

The n-fold CV [2] [8] is an extensively used experimental testing method. The dataset under study is split into n disjoint sets randomly. Then the algorithm is trained for n-1 sets of data and the remaining set is used to test the performance of the algorithm. This process is repeated n times and the average of the values recorded is used to measure the accuracy of the classifier. In percentage split method [2][9], the complete dataset is randomly split into two disjoint datasets. The first set which the data mining system tries to extract knowledge from, is called the training set. The extracted knowledge may be tested against the second set which is called the test set. The common heuristic is to apply a 60/40 split [2]. The accuracy measure refers to the percentage of the correctly classified instances from the test data. Once the tests are carried out using the selected data sets, the accuracy of the different classifiers is compared. The accuracy is further tested with a multiple classifier combination.

### 6.2 Experimental Evaluation and Performance Analysis

The KNIME performance is evaluated experimentally with the selected datasets and the accuracy measure is tabulated using both cross validation and percentage split. (Table II & Table III). The percentage split uses the 60/40 criteria as per the standard conceived notions of the performance of the split. It is found that KNIME has a reasonably good performance for all well known classifiers, except for certain cases, where the accuracy is compromised due to the non suitability of the datasets for a specific algorithm. The MLP [2][3], for instance cannot be applied to nominal datasets (shown by NA) while MLP gives superior accuracy with the iris dataset and a reasonable accuracy measure with diabetes dataset. MLP can achieve proportional increase in accuracy by increasing the number of iterations. A merger of MLP and other mining algorithms requires a greater deal of pre-processing applied to datasets, than is normally required for other data mining techniques and hence knowledge about the technique is imperative. Dealing with missing attributes and string attributes are of significant importance for MLP training and hence require normalisation and filtering techniques functional at the pre-processing stage. The KNN[2][10] classification has been tested with the k value as 3.The tool shows good results for numerical datasets as is the case with KNN, considering the non suitability of the algorithm for continuous attributes as well as a numeric/categorical mix of attributes. The DT[2] and Naive Bayesian[2] models generate comparable accuracy levels for all datasets. In general, the algorithms under study give analogus results for the iris dataset due to the simple nature of the dataset and the identical type of the attributes. The comparative scoring between percentage split and cross validation technique does not show much of a variation as can be seen from the tabulation results. (Table II and Table III)

**6.3 Performance Enhancement–Multiple Classifier**

Multiple learning algorithms can produce better prediction results than the individual component algorithms. This kind of collective training of data is called ensemble learning [3][12]. KNIME supports ensemble machine learning through Bagging [3] and Boosting [3][14]. In addition, the tool provides capabilities for building user-defined collective modelling. This study proposes a model which is trained with a serial combination of the two popular data mining techniques, the Naive Bayesian and Decision Tree. The choice of the two techniques is purely heuristic based on multiple tried and tested combinations. It is found that a model that is trained on a combination of mining algorithms performs exceptionally well for all the datasets. Specifically, the census dataset on which individual counterparts have reduced throughput, a merger variant is able to do much better as is seen from the graphs (Fig 1 and Fig 2).The hybrid model generates comparable accuracy using both the test methods.

| Dataset | Dataset Properties | Number of instances | Number of attributes | Attribute type | Data Mining Task |
|---|---|---|---|---|---|
| Census Income | Multivariate | 48842 | 14 | Categorical, Integer | Classification |
| Diabetes | Multivariate | 20 | 9 | Categorical, Integer | Classification |
| Iris | Multivariate | 150 | 4 | Real | Classification |

**Table I Dataset Description**

| CLASSIFIER | DATASET | | |
|---|---|---|---|
| | Census Income | Diabetes(ARFF) | Iris |
| Decision Tree | 83.71% | 72.73% | 91.67% |
| Naive Bayesian | 81.51% | 73.38% | 95% |
| K-Nearest Neighbour | 75.49% | 69.15% | 95% |
| MLP | NA | 72.40% | 93.33% |
| Decision –Naive Ensemble | 97.344% | 81% | 96.356% |

**Table II Accuracy Statistics using Percentage Split**

| CLASSIFIER | DATASET | | |
|---|---|---|---|
| | Census Income | Diabetes(ARFF) | Iris |
| Decision Tree | 83.89% | 73.96% | 95.33% |
| Naive Bayesian | 81.71% | 73.31% | 94.67% |
| K-Nearest Neighbour | 76.04% | 68.49% | 96.67% |
| MLP | NA | 76.30% | 87.33% |
| Decision –Naive Ensemble | 97.23% | 79.049% | 98.45%+ |

**Table III Accuracy Statistics using Cross Validation**

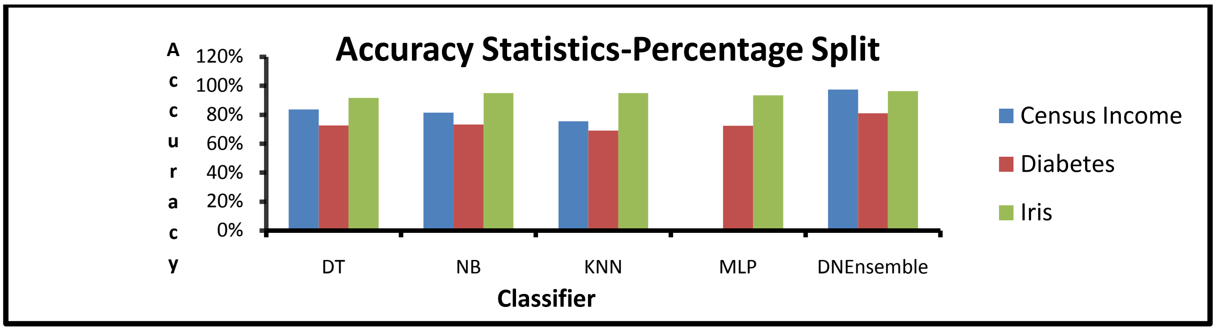

**Fig 1: Graphical representation of the accuracy statistics for Percentage Split method**

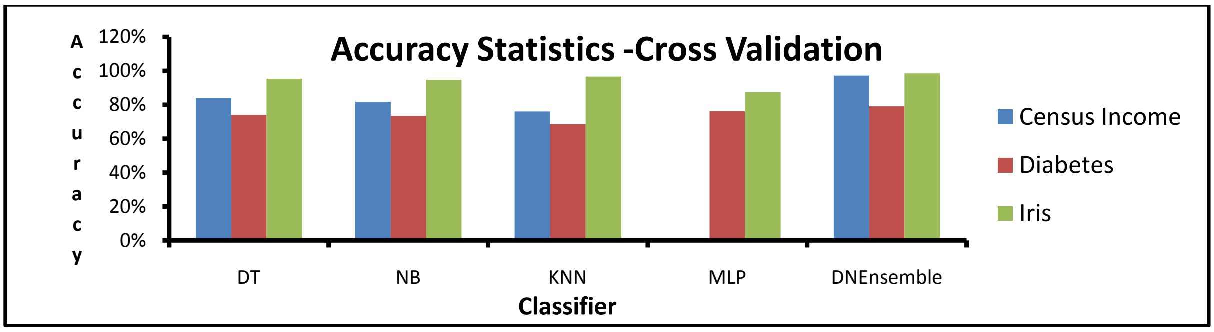

**Fig 2: Graphical Representation of the accuracy statistics for Cross validation Method**

## 7. CONCLUSIONS

Rapid advances in the field of data analytics have triggered the onset of a number of tools for effective analytics. There are a number of exceptionally powerful, open source and enterprise tools available for facilitating large scale analytics. These tools support a number of machine learning and mining techniques and hence can aid in advanced analytics. This research work has focused on experimental analysis on datasets of different sizes with four different classification algorithms using KNIME analytics platform. The results demonstrate the potential of the tool for the classification task, and studies enhanced performance through collective machine learning technique. The KNIME analytics platform can be seamlessly integrated with all big data extensions, providing research direction in advanced analytics and sophisticated data mining.

## REFERENCES

[1] Goebel, M., Gruenwald, L., A survey of data mining and knowledge discovery software tools, ACM SIGKDD Explorations Newsletter, v.1 n.1, p.20-33, June 1999 doi>10.1145/846170.846172.

[2] Han, J., Kamber, M., Jian P., Data Mining Concepts and Techniques. San Francisco, CA: Morgan Kaufmann Publishers, 2011.

[3] KNIME (Konstanz Information Miner), Available at: http://www.knime.org

[4] Orange – Data Mining Fruitful and Fun, Available at: http:// http://orange.biolab.si

[5] UCI Machine Learning Repository, Available at: http:// http://archive.ics.uci.edu/ml

[6] WEKA, the University of Waikato, Available at: http://www.cs.waikato.ac.nz/ml/weka

[7] A. Jović*, K. Brkić* and N. Bogunović*,An overview of free software tools for general data mining

[8] G. Piatetsky, KDnuggets Annual Software Poll: RapidMiner and R vie for first place, 2013, http://www.kdnuggets.com/2013/06/kdnuggets-annualsoftware-poll-rapidminer-r-vie-for-first-place.html

[9] Udaigiri Chandrasekhar Amareswar Reddy,Rohan Rath, "A comparative study of enterprise and open source big data analytical tools ",Journal of Big data 2011

**[10]** Yunjuan L, Lijun Z, Lijuan M, Qinglin M. Research and application of information retrieval techniques in intelligent question answering system. 2011 IEEE 3rd International Conference on Computer Research and Development (ICCRD); Shanghai.

[11] Arora S Chana I. A survey of clustering techniques for big data analysis. 2014 IEEE 5th International Conference Confluence the Next Generation Information Technology Summit

[12] Wu X, Zhu X, Wu G-Q, Ding W. Data mining with big data. IEEE Transactions on Knowledge and Data Engineering. 2014; 26(1):97–107. DOI: 10.1109/TKDE.2013.109.

[13] Han Hu, Yonggang Wen Tat-Seng Chua, Xuelong Li, "Toward Scalable Systems for Big Data Analytics:A Technology Tutorial",IEEE Transactions,2014

[14] V. A. AymaR. S. FerreiraP. HappD. Oliveira, R. Feitosa,G. Costa A. Plaza P. Gamba,Classification algorithms for big

data analysis, a map reduce Approach the International Archives Of the Photogrammetry,Remote Sensing and Spatial Information Sciences, Volume XL-3/W2, 2015

[15] Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao and Athanasios V. Vasilakos,"Big data analytics: a survey, Journal of Big Data 2015.

[16] Sara Landset,Taghi M. Khoshgoftaar,Aaron N. Richter, and Tawfiq Hasanin, A survey of open source tools for machine learning with big data in the Hadoop ecosystem, Journal of Big Data2015

*[17]* RapidMiner ,Available at https://rapidminer.com