IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Relative Efficacy of Weka Lazy Classifier IB1 and IBk for Different Test Mode Using Weka on the Dataset of Car Reviews

Sushilkumar Rameshpant Kalmegh

Professor
Department of Computer Science, Sant Gadge Baba Amravati University,
Amravati (M.S.) 444 602, India.

Abstract: The Size of data base is increasing day by day with rapid speed. The WEKA is data processing tool contain organized collection of state of art machine learning algorithm. However, convenient interactive graphical user interfaces are provided for data exploration, for setting up large-scale experiments on distributed computing platforms, and for designing configurations for streamed data processing. This paper has been carried out to make a performance evaluation of Weka Lazy Classifier IB1 and IBk. The paper sets out to make comparative evaluation of two classifiers from WEKA IB1 and IBk in the context of dataset of Car Reviews to maximize true positive rate and minimize false positive rate using Different Test Mode. The WEKA tool used for result processing. These algorithms are compared on the basis of accuracy.

Index Terms – **Classification, IB1, IBK, Lazy Classifier, Weka.**

I. INTRODUCTION

The amount of data in the world and in our lives seems ever-increasing and there's no end to it. We are overwhelmed with data. Today Computers make it too easy to save things. Inexpensive disks and online storage make it too easy to postpone decisions about what to do with all this stuff, we simply get more memory and keep it all. The World Wide Web (WWW) overwhelms us with information; meanwhile, every choice we make is recorded. As the volume of data increases, inexorably, the proportion of it that people understand decreases alarmingly. Lying hidden in all this data is information.

In data mining, the data is stored electronically and the search is automated or at least augmented by computer. Even this is not particularly new. Economists, statisticians, and communication engineers have long worked with the idea that patterns in data can be sought automatically, identified, validated, and used for prediction. What is new is the staggering increase in opportunities for finding patterns in data. Data mining is a topic that involves learning in a practical, non theoretical sense. We are interested in techniques for finding and describing structural patterns in data, as a tool for helping to explain that data and make predictions from it. Experience shows that in many applications of machine learning to data mining, the explicit knowledge structures that are acquired, the structural descriptions, are at least as important as the ability to perform well on new examples. People frequently use data mining to gain knowledge, not just predictions. In this given research paper Dataset of Car Reviews was used. Comparative analysis Weka Lazy Classifier IB1 and IBk with test mode Use Training set & 10-fold Cross-Validation was done using Weka Classifier. This paper is organized into Six parts. First part discusses the Introduction followed by the literature required for analysis of methods implemented. Third one is System Design followed by datasets used for analysis. Fifth is the Performance Analysis and then Conclusion.

II. LITERATURE SURVEY

1.1 WEKA

Weka was developed at the University of Waikato in New Zealand; the name stands for Waikato Environment for Knowledge Analysis The system is written in Java. It runs on almost any platform and has been tested under Linux, Windows, and Macintosh operating systems and even on a personal digital assistant. The Weka workbench is a collection of state-of-the-art machine learning algorithms and data pre processing tools. It is designed so that we can quickly try out existing methods on new datasets in flexible ways. It provides extensive support for the whole process of experimental data mining, including preparing the input data, evaluating learning schemes statistically, and visualizing the input data and the result of learning. As well as a variety of learning algorithms, it includes a wide range of pre processing tools. This diverse and comprehensive toolkit is accessed through a common interface so that its users can compare different methods and identify those that are most appropriate for the problem at hand. All algorithms take their input in the form of a single relational table in the ARFF format. The easiest way to use Weka is through a graphical user interface called Explorer as shown in figure 1. This gives access to all of its facilities using menu selection and form filling.

The Weka contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality. Advantages of Weka include:

- Free availability under the GNU General Public License
- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- A comprehensive collection of data preprocessing and modeling techniques.
- Ease of use due to its graphical user interfaces.

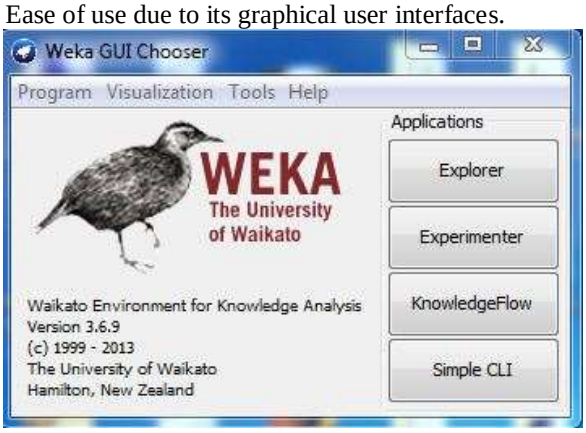

Figure 1: WEKA GUI Explorer

Weka supports several standard data mining tasks, more specifically, data pre-processing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes. Weka provides access to SQL databases using Java Database Connectivity. Weka's main user interface is the Explorer, but essentially the same functionality can be accessed through the component-based Knowledge Flow interface and from the command line. The Explorer interface features several panels providing access to the main components of the workbench. Fig. 2 shows Opening of file dataset of Car Reviews.arff file by Weka Explorer and Fig. 3 shows processing of arff file for IB1 Classifier. [1], [11].

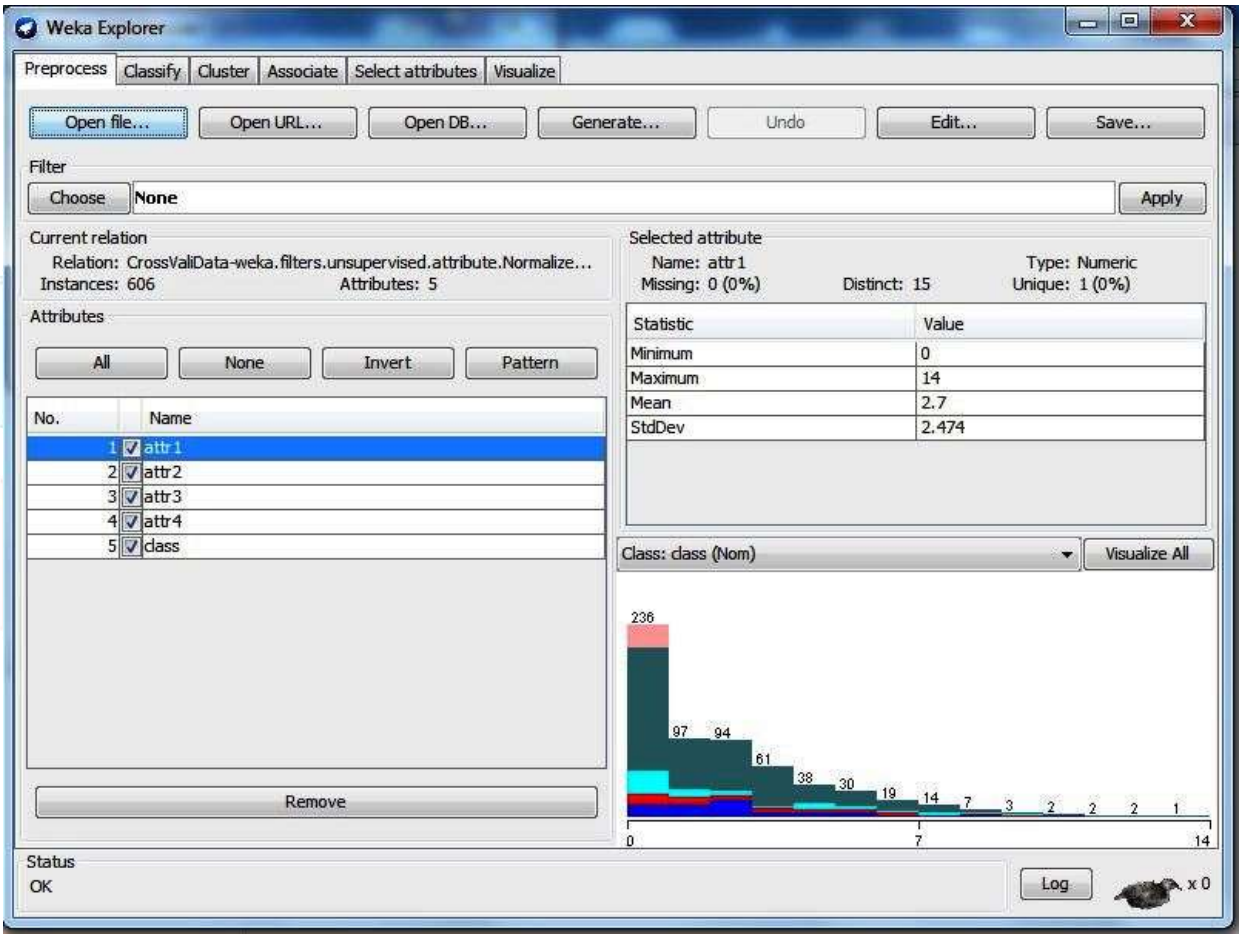

Figure 2: Opening of dataset of Car Reviews.arff file by Weka Explorer

1.2 Classification

Classification may refer to categorization, the process in which ideas and objects are recognized, differentiated, and understood. An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm that maps input data to a category. In the terminology of machine learning, classification is considered an instance of supervised learning, i.e. learning where a training set of correctly identified observations is available. The corresponding unsupervised procedure is known as clustering or cluster analysis, and involves grouping data into categories based on some measure of inherent similarity. Classification is a data mining algorithm that creates a step-by-step guide for how to determine the output of a new data instance. The tree it creates is exactly that: a tree whereby each node in the tree represents a spot where a decision must be made based on the input, and to move to the next node and the next until one reach a leaf that tells the predicted output. Sounds confusing, but it's really quite straightforward.

There is also some argument over whether classification methods that do not involve a statistical model can be considered "statistical". Other fields may use different terminology: e.g. in community ecology, the term "classification" normally refers to cluster analysis, i.e. a type of unsupervised learning, rather than the supervised learning [1],[2],[11].

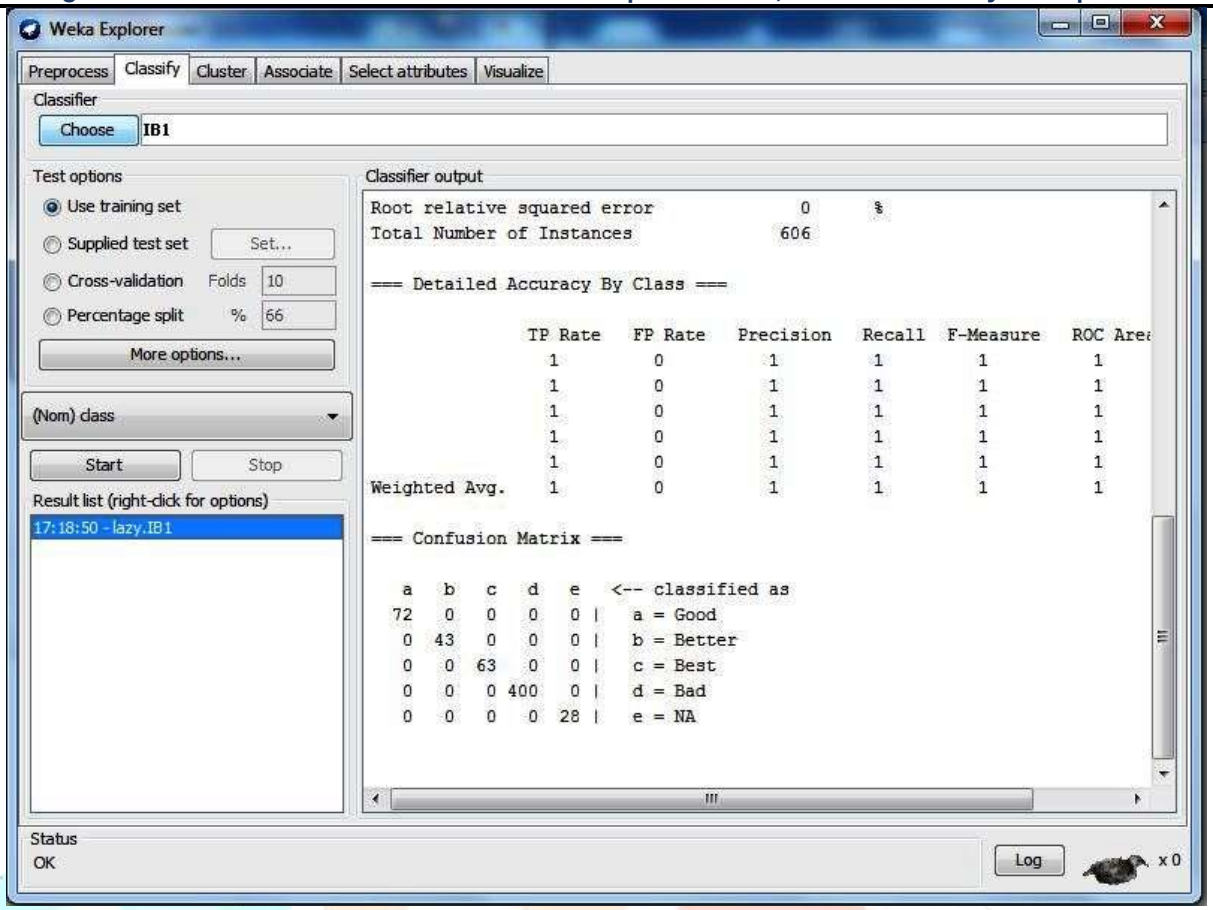

Figure 3 : Processing of arff file by IB1 Classifier on Test Mode Use Training Set

1.3 Lazy Classifiers

Lazy learners store the training instances and do no real work until classification time. Lazy is a classification technique, it includes IB1, IBk, KStar, LWL methods to classify the database. Lazy method doesn't do anything until last minute. The lazy learners use the same dataset as both training set and testing set.

The main benefit gained in employing a lazy learning method is that the target function will be approximated locally such as in the k-nearest neighbour algorithm. The disadvantages with lazy learning include (1) the large space requirement to store the complete training dataset. (2) lazy learning methods are usually slower to evaluate. Lazy learning solve multiple problems consecutively and deal the problem area in successful. In this paper comparative assessment has been done using IB1 and IBk Lazy Classifiers in test mode (i) evaluate on training data, 10-fold cross-validation in the context of dataset of Car Reviews. [1], [3], [7].

1.4 IB1

IB1 is a basic instance-based learner that finds the training instance closest in Euclidean distance to the given test instance and predicts the same class as this training instance. If multiple instances have the same (smallest) distance to the test instance, the first one found is used. IB1 is identical to the nearest neighbour algorithm except that it normalizes its attributes' ranges, processes instances incrementally, and has a simple policy for tolerating missing values.

Nearest neighbour is one of the simplest learning/classification algorithms, and has been successfully applied to a broad range of problems. The nearest neighbour classifier works based on the intuition that the classification of an instance is likely to be most similar to the classification of other instances that are nearby to it. [1],[3], [4], [5], [6].

1.5 IBk

IBk is an implementation of the k-nearest-neighbours classifier. In weka it's called IBk (instance based learning with parameter k) and it's in the lazy class folder. Fundamentally "IB" remains for Instance-Based and "k" determines number of neighbors that are analysed. It can select appropriate value of K based on cross-validation. IBk is an instance-based learning approach like the K-nearest neighbour method. The basic principle of this algorithm is that each unseen instance is always compared with existing ones using a distance metric, most commonly Euclidean distance and the closest existing instance is used to assign the class for the test sample weka's default setting is K = 1. Compared to other algorithms, it needs more time to predict the test samples' classes.

Opening of IBk classifier have following steps. The first step to choose weka Explorer initially, then choose dataset, and choose classify tap to get options from IBk implementation. It has the cross-validation option that can help by choosing the best value automatically. Weka uses cross-validation to select the best value for KNN (that is k-nearest neighbor algorithm).

It can also do distance weighting. A variety of different search algorithms can be used to speed up the task of finding the nearest neighbors. The default is the same as for IB1—that is, the Euclidean distance. The number of nearest neighbors (default k = 1) can be specified explicitly in the object editor or determined automatically using leave-one-out cross-validation, subject to an upper limit given by the specified value. Predictions from more than one neighbor can be weighted according to their distance from the test instance. [1], [3], [4], [5], [7], [8], [9], [10].

III. System Design

In order to co-relate Reviews with the categories, a model based on the machine learning was designed. As an input to the model, various quality car reviews are considered which are available online. Around 606 car reviews samples were collected on above repository using internet. In order to extract context from the car reviews, the car reviews was process with stop word removal, stemming and tokenization on the car reviews contents. The car reviews then separated into 5 categories GOOD, BETTER, BEST, BAD, NA (not applicable) and then converted into the term frequency matrix for further analysis purpose. Frequency matrix then converted to arff file using Java Programming. Finally classification is processed using WEKA Explorer; this can be seen in following Fig. 4. Due to classification in above 5 categories we are also able to find the GOOD, BETTER, BEST, BAD, NA count on every data set which help for market analysis, product rating and much more purposes. Based on this data, features (i.e. metadata) were extracted so that contextual assignment of the car reviews to the appropriate content can be done. This process is known as metadata processing [11].

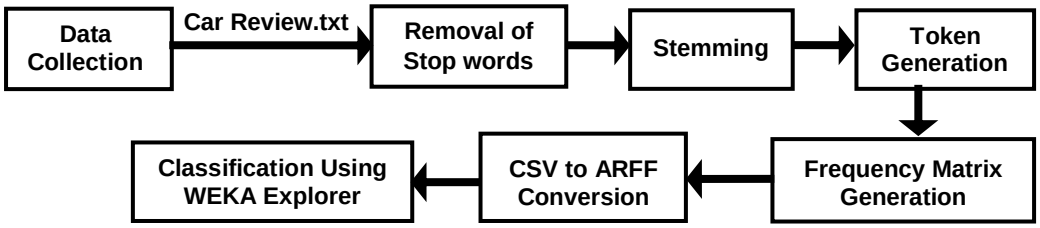

Figure 4. Flow Diagram of the Model

IV. Data Collection

Hence, it was proposed to generate car reviews data. Consequently the national and international resources were used for the research purpose. Data for the purpose of research has been collected from the various online resources using internet. They are downloaded and after reading the car reviews they are manually classified into 12 (Twelve) categories. There are 606 car reviews in total. The details are as shown in following table 1. The attributes consider for this classification is based on GOOD, BETTER, BEST, BAD, NA count each classification having their own data dictionary and based on this they are classified, the review are made by expert and user. Hence, there will be drastic enhancement in e-Contents when we refer to the latest material available in this regards [11].

Table I: Categorization of Car Review Dataset

Sr. No.	Car Companies	Numbers of Reviews
1	Chevrolet	38
2	Fiat	27
3	Ford	36
4	Honda	47
5	Hyundai	59
6	Mahindra & Mahindra	63
7	Maruti Suzuki	95
8	Renault	53
9	Skoda	23
10	Tata Motors	90
11	Toyota	41
12	Volkswagan	34
	Total	606

V. Performance Analysis

The Data so collected needed a processing. Hence as given in the system design phase, all the 606 data were processed for stop word removal, stemming, tokenization and ultimately generated the frequency matrix based on GOOD, BETTER, BEST, BAD and NA count. Stemming is used as many times when Car Review Data is printed, for a same there can be many variants depending on the tense used or whether it is singular or plural. Such words when processed for stemming, generates a unique word. Stop words needs to be removed as they do not contribute much in the decision making process. The dictionary of words is checked and removed such word from it. Frequency matrix thus generated can be processed for generating a model by converting CSV to ARFF file and the model so generated was used in further decision process. The two different test mode i.e. i) Use Training set and ii) 10-folds cross validation used for IB1 and IBk Lazy Classifiers. For processing WEKA APIs were used. The following tables shows the Confusion Matrix and True positive (TP) and False Positive (FP) rate of IB1 and IBk. In the given result the 1.0 represent the BEST, whereas the WORST is 0.0.

The following tables 2, 4, 6 and 8 show the result for the Confusion Matrix and the Tables 3, 5, 7, and 9 show True Positive and False Positive rate of IB1 and IBk for test mode: i) Use Training set and ii) 10-folds cross validation.

Table 2: Confusion Matrix for IB1 for Test Mode: Use Training Set

Classified as →	Good	Better	Best	Bad	NA
Good	72	0	0	0	0
Better	0	43	0	0	0
Best	0	0	63	0	0
Bad	0	0	0	400	0
NA	0	0	0	0	28

Table 3: TP and FP Rate of IB1 for Test Mode: Use Training Set

Class ↓	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Good	1	0	1	1	1	1
Better	1	0	1	1	1	1
Best	1	0	1	1	1	1
Bad	1	0	1	1	1	1
NA	1	0	1	1	1	1
Weighted Avg. →	1	0	1	1	1	1

Table 4: Confusion Matrix for IB1 for Test Mode: 10-fold Cross-Validation

Classified as →	Good	Better	Best	Bad	NA
Good	70	0	0	2	0
Better	0	37	6	0	0
Best	0	3	57	3	0
Bad	1	0	2	397	0
NA	0	0	0	1	27

Table 5: TP and FP Rate of IB1 for Test Mode: 10-fold Cross-Validation

Class ↓	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Good	0.972	0.002	0.986	0.972	0.979	0.985
Better	0.86	0.005	0.925	0.86	0.892	0.928
Best	0.905	0.015	0.877	0.905	0.891	0.945
Bad	0.993	0.029	0.985	0.993	0.989	0.982
NA	0.964	0	1	0.964	0.982	0.982
Weighted Avg. →	0.97	0.021	0.97	0.97	0.97	0.974

Table 6: Confusion Matrix for IBk for Test Mode: Use Training Set

Classified as →	Good	Better	Best	Bad	NA
Good	72	0	0	0	0
Better	0	43	0	0	0
Best	0	0	63	0	0
Bad	0	0	0	400	0
NA	0	0	0	0	28

Table 7: TP and FP Rate of IBk for Test Mode: Use Training Set

Class ↓	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Good	1	0	1	1	1	1
Better	1	0	1	1	1	1
Best	1	0	1	1	1	1
Bad	1	0	1	1	1	1
NA	1	0	1	1	1	1
Weighted Avg. →	1	0	1	1	1	1

Table 8: Confusion Matrix for IBk for Test Mode: 10-fold Cross-Validation

Classified as →	Good	Better	Best	Bad	NA
Good	70	0	0	2	0
Better	0	37	6	0	0
Best	0	4	56	3	0
Bad	1	0	2	397	0
NA	0	0	0	1	27

Table 9: TP and FP Rate of IBk for Test Mode: 10-fold Cross-Validation

Class ↓	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Good	0.972	0.002	0.986	0.972	0.979	0.986
Better	0.86	0.007	0.902	0.86	0.881	0.971
Best	0.889	0.015	0.875	0.889	0.882	0.977
Bad	0.993	0.029	0.985	0.993	0.989	0.988
NA	0.964	0	1	0.964	0.982	0.979
Weighted Avg. →	0.969	0.021	0.969	0.969	0.969	0.985

VI. CONCLISION

The following table 10 shows the summary of Classification.

Table 10: Summary of Classification

Classifier →	IB1		IBk	
Test Mode	Use Training Set	10-fold Cross-Validation	Use Training Set	10-fold Cross-Validation
Correctly Classified Instances	606 (100%)	588 (97.03%)	606 (100%)	587 (96.86%)
Incorrectly Classified Instances	0 (0%)	18 (2.97%)	0 (0%)	19 (3.14%)

In this paper as per the previous performance analysis, Table 10 Summary of Classification shows that both the Classifier IB1 & IBk has the accuracy for test mode evaluate on training data is 100%, This 100% accuracy for test mode evaluate on training data for both the Classifier IB1 & IBk is achieved due to the fact that both are Lazy Classifiers and further IB1 is Instance-Based learner with fixed neighbourhood, K sets the number of neighbors to use &whereas IB1 is equivalent to IBk for K = 1. WEKA's nearest neighbor implementations (IBk) has been used to generate a classifier based on one neighbor (IB1). Further both the Classifier IB1 & IBk has the accuracy for test mode 10-fold Cross-Validation is 97.03% and 96.86% respectively. Overall Performance of IB1 and IBk algorithm is acceptable for the test mode: 10-fold cross-validation, except some instances of Car Review Data are not correctly classified.

From all the above result in the Table 2 to Table 10, it is observed that performance of both the Classifier IB1 & IBk is excellent on test mode evaluate on training data as compared to the test mode 10-fold cross-validation.

VII, ACKNOWLEDGMENT

Author thanks to Dr. Sachin N. Deshmukh, Professor, Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad who initiated this line of research. Author also thanks University of Waikato for WEKA tool availability as an open source. Finally author thanks to the entire researcher, whose papers were used as a reference as listed in references section.

REFERENCES

[1] Ian H. Witten, Eibe Frank & Mark A. Hall. 2016, Data Mining Practical Machine Learning Tools and Techniques, (Third Edition, Morgan Kaufmann Publishers is an imprint of Elsevier).

[2] http://en.wikipedia.org/wiki/Classification.

[3] Ms S. Vijayarani and Ms M. Muthulakshmi, 2013, Comparative Analysis of Bayes and Lazy Classification Algorithms, International Journal of Advanced Research in Computer and Communication Engineering, 2(8), 3118-3124

[4] R. Preethi, G. M. SuriyaaKumar, N. G. Bhuvaneswari Amma and G Annapoorani, 2014 , Performance Analysis of Classifiers to Effieciently Predict Genetic Disorders Using Gene Data, International Journal of Innovative Research in Computer and Communication Engineering, 2(11), 6960-6966

[5] V. Prasathkumar, A. Deepalakshmi and N.Ramkumar, 2017. Performance Investigation of Lazy Classifiers for the Classification of Multivariate Data, Journal of Information technology and Its Applications, 2(2), 1-9

[6] Jasmina NOVAKOVIĆ, Perica STRBAC and Dusan BULATOVIĆ, 2011, TOWARD OPTIMAL FEATURE SELECTION USING RANKING METHODS AND CLASSIFICATION ALGORITHMS, Yugoslav Journal of Operations Research, 21(1), 119-135

[7] K.K.Revathi and K.K.Kavitha, 2017, COMPARISON OF CLASSIFICATION TECHNIQUES ON HEART DISEASE DATA SET, International Journal of Advanced Research in Computer Science, 8(9), 276-280

[8] Govinda.K and Narendra B,, 2018, Opinion mining using Classification Techniques, International Journal of Pure and Applied Mathematics, 118(9), 535-544

[9] S. Venkata Lakshmi and T. Edwin Prabakaran, 2015, Performance Analysis of Multiple Classifiers on KDD Cup Dataset using WEKA Tool, Indian Journal of Science and Technology, 8(17), 1-10

[10] Sonali Kadam, Rutuja Pawar, Manisha Kumari, Shweta Phule and Priyansha Kher, 2017, Performance Analysis of Pre-Processing Techniques with Ensemble of 5 Classifiers, International Journal on Recent and Innovation Trends in Computing and Communication, 5(5), 1250 – 1255

[11] Prof. Sushilkumar Rameshpant Kalmegh, 2022, COMPARATIVE ANALYSIS OF CATEGORIZATION OF CAR REVIEWS DATASET BY USING WEKA CLASSIFICATION ALGORITHM RULES NNGE AND JRIP, International Journal of Current Science (IJCSPUB), 12(1), 875-881