Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 12, December 2015, pg.125 – 134

Comparative Analysis of Classification Techniques in Data Mining Using Different Datasets

Ritu Sharma[1], Mr. Shiv Kumar[2], Mr. Rohit Maheshwari[3]

[1]Department of CSE, Mewar University, India
[2]Department of CSE, Mewar University, India
[3]Department of CSE, Mewar University, India
[1] ritusharma1489@gmail.com, [2] shivkumar004@gmail.com, [3] rohit.maheshwari27@gmail.com

***Abstract:** Data mining is the invention of knowledge and useful information from the large amounts of data stored in databases. It is referred as an analysis study of the Knowledge discovery in database process or KDD. Data mining tools are used in forecasting future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. Classification is an important data mining technique with broad applications. It classifies data of different kinds. Classification is used in every field of our life. Classification is used to classify each item in a set of data into one of predefined group of classes. The present study aimed to do the comparative analysis of several data mining classification techniques on the basis of parameters accuracy, execution time, types of datasets and applications. Several major kinds of classification techniques are present, the main concerned of this work on decision tree based (M5P), nearest neighbour based (K star), rule-based (M5 Rule), neural network based (Multilayer Perceptron).*

***Keywords:** Data Mining, M5P, K star, M5Rule, Multilayer Perceptron*.

I. INTRODUCTION

The data mining is the process of extracting unknown and predictive information from huge amount of data. It is an innovative tool with great potential to help companies and mainly focus on the most essential information in their data warehouses. Most commonly data mining is also known as Knowledge Discovery in Databases (KDD). KDD is the important process of identifying valid, new, potentially useful, and finally understandable patterns in data. Knowledge discovery process has iterative sequential steps of processes and data mining is one of the KDD processes [6].

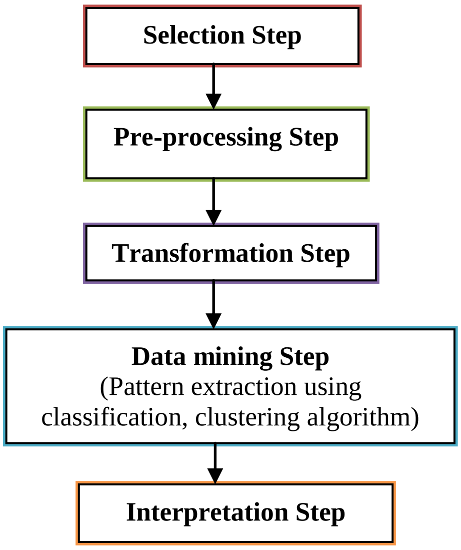


Figure 1: Sequential Steps of KDD process.

1.1 *Selection Step*

The first step in the KDD process is selection, in this step the data appropriate to the analysis task are retrieved from the database and objective dataset is formed. In this work, we have taken various datasets from the performance analysis of several data mining classification techniques.

1.2. *Pre-processing Step*

The second step is pre-processing step, in this existing databases are highly susceptible to noisy, missing, and inconsistent data due to huge size, complexity. So, in this step the dataset which is selected during the selection step is pre-processed to handle the above problem and transformed into a form that is appropriate for the classification techniques mainly with the help of Weka data mining tool.

1.3. *Transformation Step*

The third process is transformation step, in this step data are transformed or consolidated into forms appropriate for mining by performing smoothing, summary or aggregation, generalization, normalization, discretization and feature construction operations. In this work, Weka data mining tool is used for the above purpose.

1.4. *Data mining Step*

The KDD process in the data mining methods is used for extracting patterns from data. In this step of KDD process various methods are applied to extract data patterns. The data mining task are used for analyzed the dataset. In this work, data mining classification techniques like decision tree, artificial neural network, nearest neighbour and rule-based classification are used to extract the data patterns on the various datasets using WEKA machine learning tool.

1.5. *Interpretation Step*

This step involves pattern evaluation and knowledge representation. In this step visualization techniques are used to help users understand and interpret the data mining results correctly.

II. LITERATURE SURVEY

Megha Gupta, Naveen Aggarwal [2010][2], presented this paper on “Classification Techniques Analysis” to analyze advantages and disadvantages of various classification techniques when this techniques applied on XML data.

Dr. A. Padmapriya [2012][3],” Prediction of Higher Education Admissibility using Classification Algorithms”. This paper proposes to apply data mining techniques to predict higher education admissibility. Several well known data mining classification algorithms, including a decision tree classifier and Naive Bayesian classifier, are applied on the dataset. The performance of these algorithms is analyzed and compared.

S.Neelamegam, Dr.E.Ramaraj [2013][7], “Classification algorithm in Data mining: An Overview”. In this paper, an overview of several major kinds of classification method including decision tree, Bayesian networks, k-nearest neighbour classifier, Neural Network, Support vector machine are discussed.

S.Archana, Dr. K.Elangovan [2014][10],” Survey of Classification Techniques in Data Mining”. Several major kinds of classification algorithms including C4.5, k-nearest neighbour classifier, Naive Bayes, SVM, and IB3.This paper provide a general survey of different classification algorithms and their advantages and disadvantages.

Dr.A.Bharathi, E.Deepankumar [2014][11], presented paper on “Survey on Classification Techniques in Data Mining”. In this paper, different kinds of classification techniques are discussed such as Association Rule Mining, Bayesian Classification, and Decision Tree Classification, nearest neighbour classifier, neural Networks and Support Vector Machine.

III. PROBLEM STATEMENT

Data mining is a broad area that integrates techniques from various fields including machine learning, artificial intelligence, statistics and pattern recognition, for the analysis of large amount of data. Each of these methods can be used in various situations as needed where one tends to be useful while the other may not and vice-versa.

These classification algorithms can be implemented on several types of data sets like data of students, rainfall data according to performances. Therefore these classification techniques show how a data can be determined and grouped when a new set of data is available.

IV. OBJECTIVE

1. Study of following classification techniques in data mining:
   - Decision Tree Induction(M5P)
   - K-Nearest Neighbour(K-star)
   - Rule-Based Classifier(M5Rule)
   - Artificial Neural Network(MLP)

2. Comparative Analysis of classification techniques on the basis of following parameters:
   - Accuracy
   - Execution Time
   - Types of Dataset
   - Applications

V. CLASSIFICATION TECHNIQUES

Data mining consists of number of techniques which are used to mine appropriate and interesting knowledge from data. Data mining has some tasks such as association rule mining, classification and prediction, and clustering. Among all these classification techniques are supervised learning techniques to classify data item into predefined class label. It is one of the generally used techniques in data mining that construct classification models from an input data set and predict future data trends. The main part of this work is concerned with analysis of decision tree based (M5P), neural network based (Multilayer Perceptron), nearest neighbour based (K-Star) and rule-based (M5Rule) algorithms.

5.1 *Decision Tree Classification:*

A decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents a result of the test, and each leaf node holds a class label. The topmost node in a tree is the root node. We can simply obtain the rules related to the tree by traversing each leaf of the tree starting from the node. Decision tree are attractive in data mining as they stand for rules which can be expressed in natural language [1].

M5P Algorithm: It is most commonly used algorithm based on decision tree for numeric data prediction and at each leaf it stores a linear regression model that predicts the class value of instances that reach the leaf. It is the reconstruction of Quinlan's algorithm for inducing trees of regression models. It combines a predictable decision tree with the possibility of linear regression functions at the nodes [5].

Advantages of M5P Algorithm:

This algorithm does not require any domain knowledge or parameter setting, and therefore is appropriate for knowledge discovery. It can handle high dimensional data. The learning and classification steps of algorithm are simple, fast and have a good accuracy.

Limitations of M5P Algorithm

M5P algorithm does not easily handle non-numeric data, when training set is small there is high classification error rate in comparison with the number of classes. And it requires that the target attribute will have only discrete values.

5.2 *Artificial Neural Network:*

Artificial neural networks (ANNs) are stimulated by biological neural networks that correspond to brain image for information processing. Similar to human brains, neural networks are also consisting of processing units (artificial neurons) and connections (weights) between them. The processing units convey received information on their outgoing connections to other units. The most important feature of these networks is their adaptive nature where "learning by example" replaces "programming" in solving problems. This feature makes such computational models very attractive in application domains where one has little or incomplete understanding of the problem to be solved but where training data is readily available [1].

Multilayer Perceptron: MLP is one of the most common neural network models. Neural network of this type is known as a supervised network because it requires an output to learn. The main aim of this type of network is to create a model that correctly maps the input to the output using historical data so that the model can then be used to produce the output when the desired output is not identified. The graphical representation of MLP is shown in figure 2.

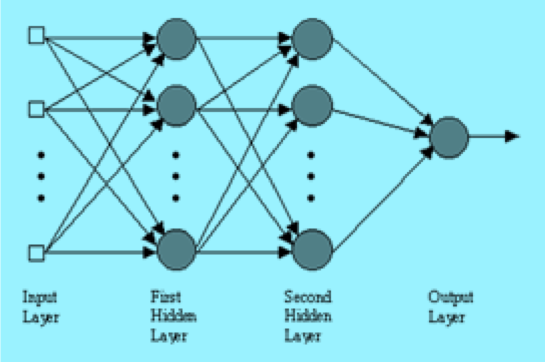


Figure 2: Graphical Representation of MLP

And with each input the output of the neural network is compared with the desired output and an error is computed. Then this error is fed back to the neural network [5].

Advantages of Multilayer Perceptron:

It is very flexible about incomplete, missing and noisy data. It can be updated with fresh data and implemented in parallel hardware. When an element of this algorithm is failed, it can continue without any problem by their parallel nature.

Limitations of Multilayer Perceptron:

There are no any methods to find out the best possible number of neurones necessary for solving any problem and it is very difficult to select a training data set which fully describes the problem to be solved.

5.3 *K-Nearest Neighbour Classification:*

K-nearest neighbour classification is based on learning by an evaluation, that is, by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by *n* attributes. Each tuple represents a point in an *n*-dimensional space. In this way, all of the training tuples are stored in an *n*-dimensional pattern space. When given an unknown tuple, a ***k***-nearest-neighbor classifier searches the pattern space for the *k* training tuples that are closest to the unknown tuple. These *k* training tuples are the *k* “nearest neighbors” of the unknown tuple. “Closeness” is defined in terms of a distance metric, such as Euclidean distance. The Euclidean distance between two points or tuples, say, $\boldsymbol{X}1 = (x_{11}, x_{12}, :::, x_{1n})$ and$\boldsymbol{X}2 = (x_{21}, x_{22}, :::, x_{2n})$, is

$$dist(\boldsymbol{X}_1, \boldsymbol{X}_2)=\sqrt{\sum_{i=1}^{n}(X1i - X2i)pow2} \text{ ....(i)[1]}$$

Therefore it can suffer from poor accuracy. Therefore, it has been modified to incorporate attribute weighting and the pruning of noisy data tuples [1].

K-Star: K star is one of the nearest neighbour lazy learning classification method with generalized based on transformations. It provides a reliable approach to handle symbolic attributes, real valued attributes and missing values. Space required for the storage is very large as compared to other algorithms. And it is generally slower in evaluating the result [12].

Advantages of K-star:
It is robust to noisy training data and it is more effective when applied on large data set.

Limitations of K-star:
In this algorithm, it is required to determine value of parameter k. The computation cost to calculate distance of each instance to all training sample is very high.

5.4 *Rule-Based Classification*

In rule-based classification a set of IF-THEN rules are used for classification. An IF-THEN rule is an expression of the form

IF condition THEN conclusion

In the above expression the "IF"-part of a rule is known as the rule antecedent or precondition. And the "THEN"-part is the rule consequent. In the rule antecedent, the condition consists of one or more attribute tests that are logically ANDed. The rule's consequent contains a class prediction. If the condition in a rule antecedent holds true for a given tuple, we say that the rule antecedent is satisfied and that the rule covers the tuple [1].For example if we are predicting that whether a student will get admission in ph.d or not, then R1 can be written as

R1: (age = 25) ^ (post graduate = yes)) (get admission in ph.d = yes).

M5Rule: M5Rule create a decision list for regression problems which is used to divide and to conquer. It builds a model tree and makes the "best" leaf into a rule in each iteration of progress. The approach for generating rules from model trees, called M5-Rules.In its work flow a tree learner is apply to the full training dataset and a pruned tree is learned then, the best branch is made into a rule and the tree is discarded. All instances covered by the rule are removed from the dataset. The process is applied recursively to the remaining instances and terminates when all instances are covered by one or more rules. This is a fundamental divide-and-conquer strategy for learning rules and its "best" branch into a rule [8].

VI. METHODOLOGY

Weka (Waikato Environment for Knowledge Analysis) is a popular set of machine learning algorithms developed at the University of Waikato, New Zealand, for solving real-world data mining problems [4]. It is written in Java and runs on almost any platform. It is an open source application which is freely available. Data pre-processing, classification, clustering, association, regression and feature selection these standard data mining tasks are supported by Weka. For classification purpose classify tab in Weka Explorer is used [9]. Advantages of Weka tool:

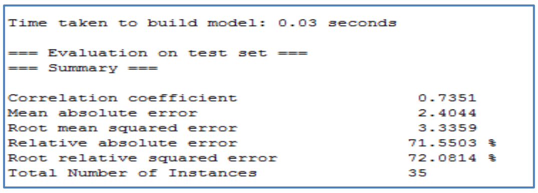
i. Available freely under the GNU General Public License.

ii. It is portable, as it is implemented in the Java programming language and thus runs on almost any platform.

iii. It is easy to use due to its graphical user interfaces.

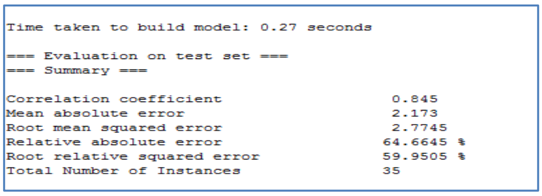
VII. RESULTS

A comparison of classifiers for different datasets has been done on the basis of accuracy, execution time, type of data sets by classifiers to analysis the performance of classification algorithm and its application domain is also discussed.



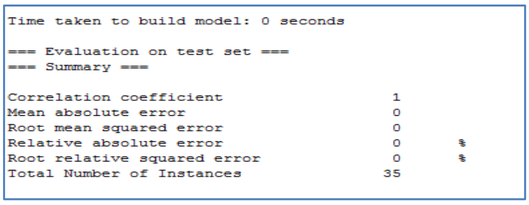
```
Time taken to build model: 0.03 seconds

=== Evaluation on test set ===
=== Summary ===

Correlation coefficient                  0.7351
Mean absolute error                      2.4044
Root mean squared error                  3.3359
Relative absolute error                 71.5503 %
Root relative squared error             72.0814 %
Total Number of Instances               35
```

M5P



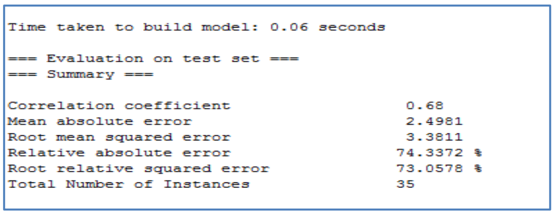
```
Time taken to build model: 0.27 seconds

=== Evaluation on test set ===
=== Summary ===

Correlation coefficient                  0.845
Mean absolute error                      2.173
Root mean squared error                  2.7745
Relative absolute error                 64.6645 %
Root relative squared error             59.9505 %
Total Number of Instances               35
```

MLP



```
Time taken to build model: 0 seconds

=== Evaluation on test set ===
=== Summary ===

Correlation coefficient                  1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances               35
```

K-Star



```
Time taken to build model: 0.06 seconds

=== Evaluation on test set ===
=== Summary ===

Correlation coefficient                  0.68
Mean absolute error                      2.4981
Root mean squared error                  3.3811
Relative absolute error                 74.3372 %
Root relative squared error             73.0578 %
Total Number of Instances               35
```

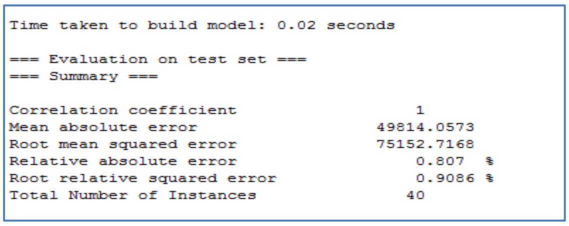
M5Rule

Figure 3: Performance Evaluation of Different Classification Algorithm for Rainfall Dataset

TABLE 1

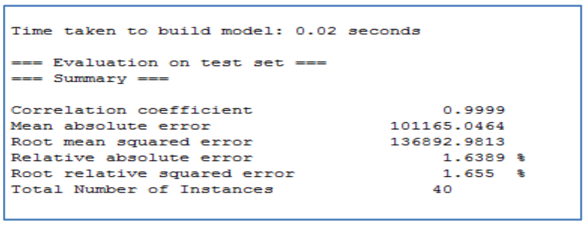
COMPARISON OF CLASSIFIER'S PERFORMANCE FOR RAINFALL STATISTICS IN CHITTORGARH

For Rainfall Statistics in Chittorgarh				
Classifier	**M5P**	**MLP**	**K-star**	**M5Rule**
Dataset Type	Sequential	Parallel	Parallel	Sequential
Applications	Medical ,Security manufacturing and production, financial analysis	Character recognition, image compression, stock market, Medical	Engineering , Medical, Bussiness	Medical, Financial
Execution Time	0.03 sec	0.27 sec	0.0 sec	0.06 sec
Accuracy	73.51%	84.50%	100%	68.00%



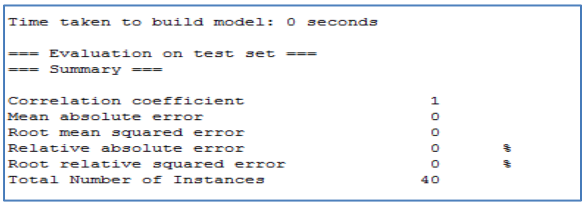
```
Time taken to build model: 0.02 seconds

=== Evaluation on test set ===
=== Summary ===

Correlation coefficient                  1
Mean absolute error                  49814.0573
Root mean squared error              75152.7168
Relative absolute error                  0.807  %
Root relative squared error              0.9086 %
Total Number of Instances               40
```

M5P



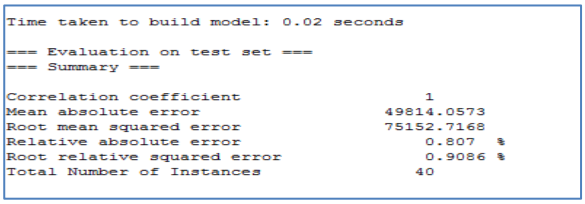
```
Time taken to build model: 0.02 seconds

=== Evaluation on test set ===
=== Summary ===

Correlation coefficient                  0.9999
Mean absolute error                 101165.0464
Root mean squared error             136892.9813
Relative absolute error                  1.6389 %
Root relative squared error              1.655  %
Total Number of Instances               40
```

MLP



```
Time taken to build model: 0 seconds

=== Evaluation on test set ===
=== Summary ===

Correlation coefficient                  1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances               40
```

K-Star



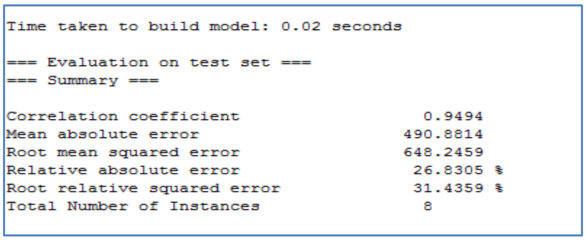
```
Time taken to build model: 0.02 seconds

=== Evaluation on test set ===
=== Summary ===

Correlation coefficient                  1
Mean absolute error                  49814.0573
Root mean squared error              75152.7168
Relative absolute error                  0.807  %
Root relative squared error              0.9086 %
Total Number of Instances               40
```

M5Rule

Figure 4: Performance evaluation of Different Classification Algorithm for Tourist Dataset

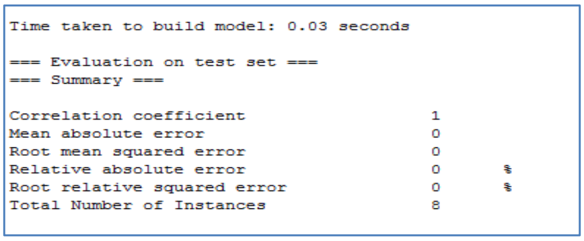
TABLE 2
COMPARISON OF CLASSIFIER'S PERFORMANCE FOR TOURISM STATISTICS OF RAJASTHAN.

For Rajasthan Tourism Statistics				
Classifier	**M5P**	**MLP**	**K-star**	**M5Rule**
Dataset Type	Sequential	Parallel	Parallel	Sequential
Applications	Medical ,Security manufacturing and production, financial analysis	Character recognition, image compression, stock market medical	Engineering , Medical, Bussiness	Medical, Financial
Execution Time	0.02 sec	0.02 sec	0.0 sec	0.02 sec
Accuracy	100%	99.99%	100%	100%



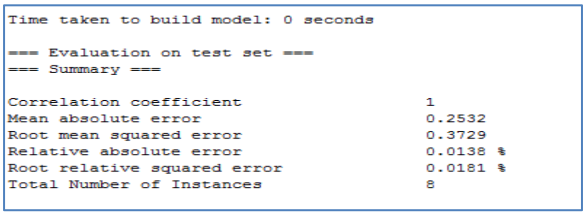
```
Time taken to build model: 0.02 seconds

=== Evaluation on test set ===
=== Summary ===

Correlation coefficient                  0.9494
Mean absolute error                    490.8814
Root mean squared error                648.2459
Relative absolute error                 26.8305 %
Root relative squared error             31.4359 %
Total Number of Instances                8
```

M5P



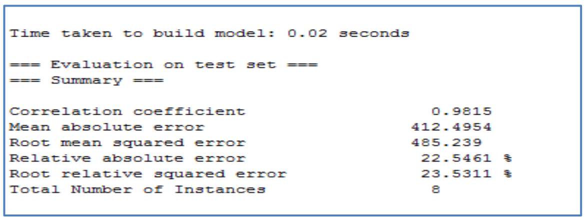
```
Time taken to build model: 0.03 seconds

=== Evaluation on test set ===
=== Summary ===

Correlation coefficient                  1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances                8
```

MLP



```
Time taken to build model: 0 seconds

=== Evaluation on test set ===
=== Summary ===

Correlation coefficient                  1
Mean absolute error                      0.2532
Root mean squared error                  0.3729
Relative absolute error                  0.0138 %
Root relative squared error              0.0181 %
Total Number of Instances                8
```

K-Star



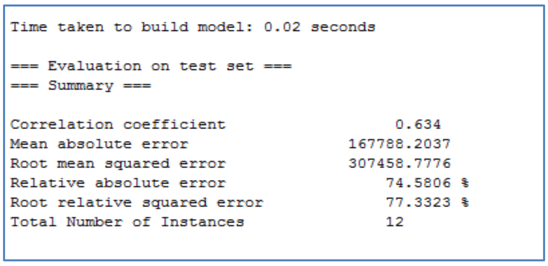
```
Time taken to build model: 0.02 seconds

=== Evaluation on test set ===
=== Summary ===

Correlation coefficient                    0.9815
Mean absolute error                      412.4954
Root mean squared error                  485.239
Relative absolute error                   22.5461 %
Root relative squared error               23.5311 %
Total Number of Instances                  8
```

M5Rule

Figure 5: Performance evaluation of Different Classification Algorithm for Admission Dataset

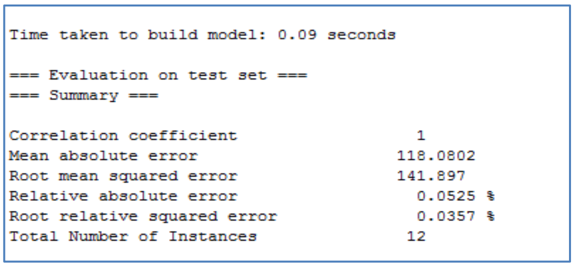
TABLE 3
COMPARISON OF CLASSIFIER'S PERFORMANCE FOR STUDENT'S ENROLMENTS IN PH.D.

For Student's Enrolment Statistics				
Classifier	**M5P**	**MLP**	**K-star**	**M5Rule**
Dataset Type	Sequential	Parallel	Parallel	Sequential
Applications	Medical, Security Manufacturing and production, financial analysis	Character recognition, image compression, stock market, medical	Engineering , Medical, Bussiness	Medical, Financial
Execution Time	0.02 sec	0.03 sec	0.0 sec	0.02 sec
Accuracy	94.94%	100%	100%	98.15%



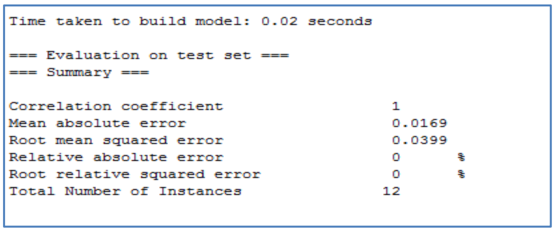
```
Time taken to build model: 0.02 seconds

=== Evaluation on test set ===
=== Summary ===

Correlation coefficient                    0.634
Mean absolute error                   167788.2037
Root mean squared error               307458.7776
Relative absolute error                   74.5806 %
Root relative squared error               77.3323 %
Total Number of Instances                 12
```

M5P



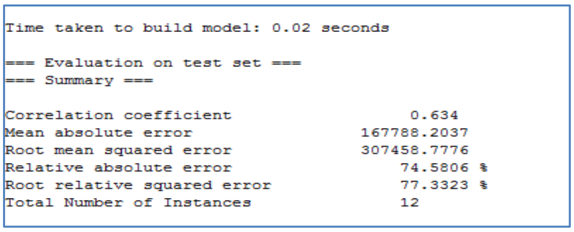
```
Time taken to build model: 0.09 seconds

=== Evaluation on test set ===
=== Summary ===

Correlation coefficient                    1
Mean absolute error                      118.0802
Root mean squared error                  141.897
Relative absolute error                    0.0525 %
Root relative squared error                0.0357 %
Total Number of Instances                 12
```

MLP



```
Time taken to build model: 0.02 seconds

=== Evaluation on test set ===
=== Summary ===

Correlation coefficient                    1
Mean absolute error                        0.0169
Root mean squared error                    0.0399
Relative absolute error                    0      %
Root relative squared error                0      %
Total Number of Instances                 12
```

K-Star



```
Time taken to build model: 0.02 seconds

=== Evaluation on test set ===
=== Summary ===

Correlation coefficient                    0.634
Mean absolute error                   167788.2037
Root mean squared error               307458.7776
Relative absolute error                   74.5806 %
Root relative squared error               77.3323 %
Total Number of Instances                 12
```

M5Rule

Figure 6: Performance evaluation of Different Classification Algorithm for Population Dataset

TABLE 4
COMPARISON OF CLASSIFIER'S PERFORMANCE FOR POPULATION DATA OF CHITTORGARH

For Population Statistics of Chittorgarh				
Classifier	**M5P**	**MLP**	**K-star**	**M5Rule**
Dataset Type	Sequential	Parallel	Parallel	Sequentia l
Applications	Medical, Security Manufacturing and production, financial analysis	Character recognition, image compression, stock market, medical	Engineering , Medical, Bussiness	Medical, Financial
Execution Time	0.02 sec	0.09 sec	0.02 sec	0.2 sec
Accuracy	63.40 %	100%	100 %	63.40%

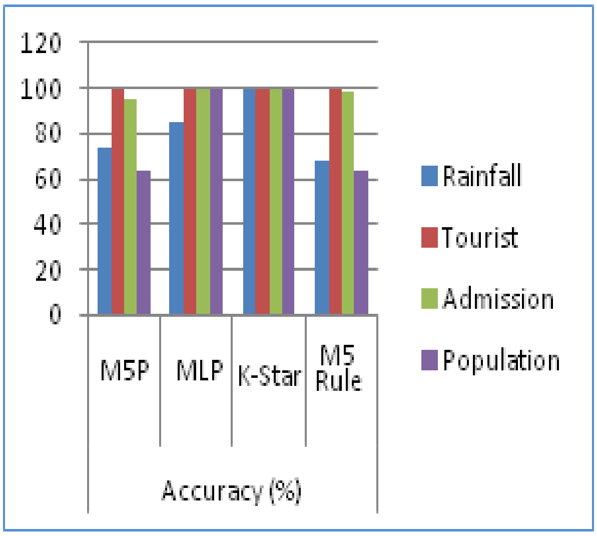


Figure 7: Graphical Representation of Accuracy of Various Classifiers on Different Datasets

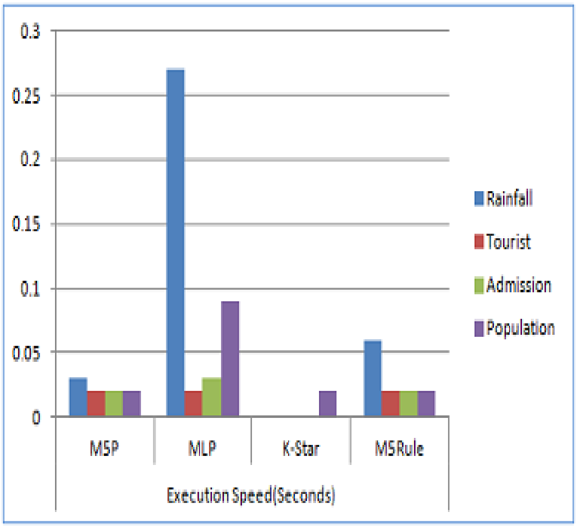


Figure 8: Graphical Representation of Execution Time of Various Classifiers on Different Datasets

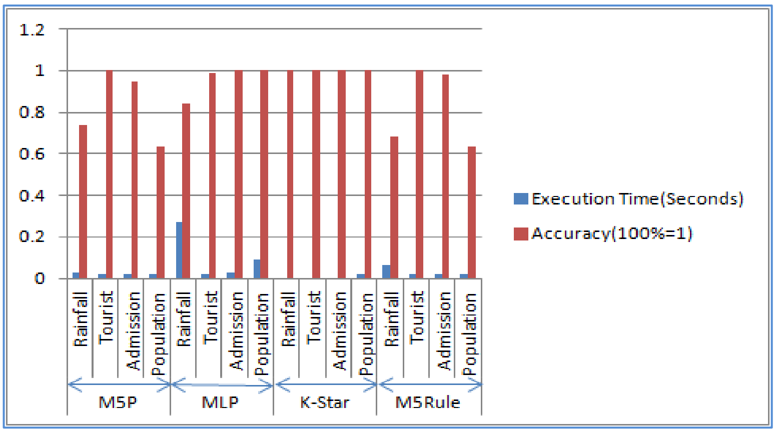


Figure 9: Comparative Analysis of Accuracy and Execution Time of Various classifiers using Different Datasets

VIII. CONCLUSION

Accordingly, in this work we have compared and analysis the performance of various classifiers on the basis of accuracy, execution time, type of dataset and domain. The analysis and comparison of these algorithms shows that k-star has highest accuracy for large dataset but other are not and for small dataset performance of algorithms are comparatively same. Therefore no particular algorithm is best suited for specific situation, the performance of the classification algorithms depends on the type and size of data sets, one algorithm is more appropriate for one data set while other algorithm is not appropriate for the same data set.

IX. FUTURE WORK

The future work will focus on the improvement of classifier's performance so that the efficiency of classification techniques would be improved in a decreased time. A combination of classification techniques will also be used to improve the performance.

X. ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my respected guide and co-guide and all faculty members of CSE Department of Mewar University. I also would like to thank my family and friends.

REFERENCES

[1] Jiawei Han,"*Data Mining:Concepts and Techniques*", Second Edition, Morgan Kaufmann,2006

[2] Megha Gupta, Naveen Aggarwal, "CLASSIFICATION TECHNIQUES ANALYSIS", National Conference on Computational Instrumentation,March 2010

[3] Dr. A. Padmapriya, "Prediction of Higher Education Admissibility using Classification Algorithms" , International Journal of Advanced Research in Computer Science and Software Engineering , Volume 2, Issue 11, November 2012

[4] Qasem A. Al-Radaideh , Eman Al Nagi, "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance", International Journal of Advanced Computer Science and Applications, Vol. 3, No. 2, 2012

[5] E. K. Onyari and F. M. Ilunga," Application of MLP Neural Network and M5P Model Tree in Predicting Streamflow: A Case Study of Luvuvhu Catchment", International Journal of Innovation, Management and Technology, Vol. 4, No. 1, February 2013

[6] N. Abirami, T. Kamalakannan, Dr. A. Muthukumaravel, "A Study on Analysis of Various Datamining Classification Techniques on Healthcare Data", International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 7, July 2013

[7] S.Neelamegam, Dr.E.Ramaraj, "Classification algorithm in Data mining: An Overview", International Journal of P2P Network Trends and Technology, Volume 4 Issue 8- Sep 2013

[8] Peyman Mohammadi, Abdolreza Hatamloun and Mohammad Masdari," A COMPARATIVE STUDY ON REMOTE TRACKING OF PARKINSON'S DISEASE PROGRESSION USING DATA MINING

METHODS", International Journal in Foundations of Computer Science & Technology , Vol. 3, No.6, November 2013

[9] Dr. Sudhir B. Jagtap, Dr. Kodge B. G, "Census Data Mining and Data Analysis using WEKA", International Conference in "Emerging Trends in Science, Technology and Management,2013

[10] S.Archana1, Dr. K.Elangovan, "Survey of Classification Techniques in Data Mining", International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 2, February- 2014

[11] Dr.A.Bharathi, E.Deepankumar ," Survey on Classification Techniques in Data Mining", International Journal on Recent and Innovation Trends in Computing and Communication, Volume. 2 Issue. 7,July 2014