# Review on Classification algorithms in healthcare

[1]A.Prettekha, [2]Dr.Pothula Sujatha, [3]S.Shanmugapriya
[1]M.Tech(CSE), [2]Assistant Professor, [3]M.Tech(CSE)
[1]Department of Computer Science,
[1]Pondicherry University, Pondicherry, India

***Abstract :*** Electronic health records (EHR) are rapidly becoming more casual among healthcare industries. Using data mining the healthcare suppliers enhance the proficiency of their organization by retrieving huge volumes of patients data from it. Data mining helps the healthcare industries to improve care and reduce costs. The tools used in the data mining examines the symptoms, causes, and treatment and determines which action will be effective for the patient. In this paper 5 classification algorithms are applied to patients dataset obtained from the UCI machine learning repository. These algorithms are applied to a data mining tool 'WEKA' and the results are compared to determine the best classifier using performance metrics.

***IndexTerms* – Data mining, electronic health records, classification algorithms.**

## I. INTRODUCTION

In olden days paper-based systems are used but in recent days it was changed to an electronic system. security of information, a backing of data and quick access of data can be achieved through electronic systems. New versions of algorithms were developed every day in the field of data mining which reduces the expenses of the organization. So a large amount of data is gathered every day.[1]

Electronic systems were used in most of the organizations and among that healthcare institutes plays a vital role of gathering electronic data by various methods including electronic health record(EHR), online insurance claim etc., A large amount of data is collected by hospitals and health institutes as an outcome of EHR systems.[1]

Data mining provides numerous applications in healthcare and its related disciplines. These can be analysed to discover new relationships and unique patterns. It makes the staffs take decisions based on data and knowledge. In proper use of data mining mostly there won't be any drawbacks, But when executed poorly or inefficient and inadequate data produces certain drawbacks. The combination of data mining techniques with healthcare system helps in increasing the efficiency of healthcare industries.[2]

In this paper, five algorithms have been considered, with the purpose of representing the most used algorithms in the healthcare sector. They are,

JRIP
K-Star
J48
NB Tree
RIDOR

## II. LITERATURE REVIEW

M.Sujatha et al., presented a study on numerous data mining classification techniques. It includes the genetic algorithm, KNN, SVM, C4.5, CART etc., the pros and cons of each algorithm was described in that paper. [15]

V. Krishnaiah et al., studied about various classification techniques and presented a survey on it. The detailed review of the techniques based on the decision tree, rule-based Algorithms, neural networks, support vector machines, Bayesian networks, and Genetic Algorithms and Fuzzy logic was presented in that paper. The advantage, disadvantage and the applications of these techniques are also described. [16]

Amit Gupta et al., presented a study on classification algorithms on crime and accident in a city of USA. They analyzed five classification algorithms such as JRIP, Naïve Bayes, J48, BayesNet, Decision Table and OneR. It is concluded that the JRIP and decision table provides the highest accuracy.[4]

Ms. S. Vijayarani and Ms. M. Muthulakshmi presented a Comparative Analysis of Bayes and Lazy Classification Algorithms. They compared algorithms such as Bayesian, BayesNet, Lazy, IBK, Naïve Bayes, IBL and K-star.. the results showed that IBK is best among all these algorithms.[8]

V.Veeralakshmi et al., studies about some classification algorithms in their paper on Ripple Down Rule learner (RIDOR) Classifier for IRIS Dataset. They compared algorithms such as JRIP, RIDOR and decision table. They concluded that RIDOR provides the best accuracy among the other algorithms.[6]

## III. ALGORITHMS

### 3.1 JRIP:

JRIP implements a propositional rule learner called as Repeated Incremental Pruning to Produce Error Reduction (RIPPER). It was proposed by William w. Cohen This algorithm undergoes 4 stages: initialization, building, Optimization and delete [4][5]

**Initialization**

Initialize RS = {}, and from each class from the less recurrent one to the most recurrent one.[5][6]

**Building stage**

Till there are no positive error rate >=50% repeat the grow phase and prune phase.
(i) Growing phase: Keep on growing by greedily adding clauses to the rule until the rule is perfect (100%)
(ii) Pruning phase: Each one of the rules is pruned by an incremental method. The sequence that is going to finish can also get pruned. The metrics for pruning is (p-n)/(p+n). [5][6]

**Optimization stage**

{Ri} is considered as the first rule set in this stage of the algorithm. Using random data and prune, create 2 variants from each rule Ri. One rule is created by adding greedily to the original rule and the other is created from an empty rule. (TP+TN)/(P+N) is the pruning metric used in this stage. Possible DL is calculated for the original rule. Finally, a variant with minimal DL is selected as the representative from {Ri}. If any residual positives persist from the {Ri} then using building stage more rules are generated depending on the residual positives.[5][6]

**Delete stage**

The rule that increases the DL from the rule set is deleted. The resultant rule set is considered as RS.[5][6]

### 3.2 K-STAR

K* comes under the category of an instance-based classifier. The area that k* tries to deal with are the missing values, real and symbolic attributes and problem in smoothness. An instance-based classifier Is that the test instance class is established upon the training instance class[7]. It compares the instance to a database of prior classified examples. The function of k* is computed as

$$K_{*}(y_i, x) = -\ln P_{*}(y_i, x) \qquad (3.1)$$

Here P* stands for the probability for all transformational path from x to y instance[8].

The distance function and the classification function are the two corresponding components of an instance-based learner. The first function regulates how two instances are similar and the second regulates how new instance is created based on the instance similarities. [9]

The entropic measure is used in the k-star algorithm. Instance distance is favourable when entropy is used as a meter. The distance between instances is calculated using information theory. The steps for measuring the distance is

i) Determine a fixed set of transformation
ii) Transform one instance to another with the help of transformation [9]

### 3.3 J48

An extension of an ID3 algorithm is J48. It is a widely used machine learning algorithm developed by Quinlan. It is an open source Java implementation of C4.5 algorithm in WEKA data mining tool. Some of the features of this algorithm is pruning decision trees, decreasing misclassification error, derivation of rules, accounting missing values etc.,[10][11]. The J48 algorithm accepts continuous and categorical attribute while building a decision tree. Using top down and bottom up approach decision trees are developed. Decision node and leaf node are the components of a decision tree. Based on the statistics of the theoretical attribute value of the present training data, the decision tree is built. Here the decision node is used to decide the test of attribute while the leaf node is used to denote the class values. This algorithm is also divided into datasets based on the attribute value of the present data. Many subsets are developed from a single training data which relates the values of chosen features, which are recurrent for every subset. Every attribute evaluates the gain value and the calculation until all the process ends[11]. The main goal of the J48 algorithm is the gradual generalization of decision tree until it gains stability in flexibility and accuracy[10].

Steps in the algorithm:

1) A leaf is represented in the tree and it returns labeling of the class when the instance belongs to the same class
2) The potential information and gain information are calculated for all attributes.
3) On the basis of present selection criterion and branching the best attribute is found [10].

### 3.4 NB TREE

Naïve Bayes classifier are probabilistic classifier constructed on applying Bayes theorem. By using Bayes theorem it considers the attributes to be independent of the class variable. The theorem of total probability and Bayes theorem are used in naïve Bayesian classifier[12]

In NB classifier the value of predictor on the class is independent of other predictors. Here the value of predictor is denoted as x and The class is denoted as c. This is an assumption and so it is called conditional independence[13].

$$\underbrace{P(c|x)}_{\text{Posterior Probability}} = \frac{\overbrace{P(X|C)}^{\text{Likelihood}}\,\overbrace{P(c)}^{\text{Class Prior Probability}}}{\underbrace{P(x)}_{\text{Predictor Prior Probability}}} \tag{3.2}$$

- $P(c|x)$ - posterior probability of class
- $P(c)$ - prior probability of class.
- $P(x|c)$ - likelihood which is the probability of predictor
- $P(x)$ - predictor prior probability.[13]

### 3.5 RIDOR

RIDOR Algorithm also known as Ripple Down Rule learner was developed by Brian R. Gaines and Paul Compton. [14]. It is a direct classification method which constructs default rules with error rate. The default rule is generated first followed by the error rate. The exceptions are found using incremental reduced pruning method. The set of rules that predict classes than the default are called as exceptions. These exceptions are created using IREP[6]

## IV. EXPERIMENT AND RESULTS

The five classification algorithms namely JRIP, J48, NB TREE, K-Star and RIDOR are trained and tested on a diabetes dataset on WEKA tool. A comparison such as correctly classified and incorrectly classified instance, mean absolute error and relative absolute error are provided in table 1 and TP rate, FP rate, precision, recall, F-measure and MCC is provided in table 2.

**Table 1**
The accuracy of classifiers.

| **Algorithms** | **Correctly classified instances** | **Incorrectly classified instances** | **Mean absolute error** | **Relative absolute error** |
|---|---|---|---|---|
| **JRIP** | 76.04 | 23.95 | 34.19 | 75.23 |
| **K-STAR** | 69.14 | 30.85 | 32.75 | 72.05 |
| **J48** | 73.82 | 26.17 | 31.58 | 69.48 |
| **NB-TREE** | 76.30 | 23.69 | 28.41 | 62.50 |
| **RIDOR** | 75 | 25 | 25 | 55 |

According to the comparison of the dataset in these algorithms it is known that NB tree and JRIP performs best among the classifiers with the accuracy of 76.30% and 76.04%. but when comparing the absolute error it is seen that JRIP posses more error compared to NB tree. So it is seen that NB tree is best among the classifiers.

**Table 2**
Performance Comparisons of algorithms.

| Algorithms | TP Rate | FP Rate | Precision | Recall | F Measure | MCC |
|---|---|---|---|---|---|---|
| **JRIP** | 0.760 | 0.322 | 0.755 | 0.760 | 0.755 | 0.457 |
| **K-STAR** | 0.691 | 0.415 | 0.680 | 0.691 | 0.683 | 0.293 |
| **J48** | 0.738 | 0.327 | 0.735 | 0.597 | 0.736 | 0.417 |
| **NB TREE** | 0.763 | 0.307 | 0.759 | 0.763 | 0.760 | 0.468 |
| **RIDOR** | 0.750 | 0.347 | 0.743 | 0.750 | 0.742 | 0.428 |

Comparing the performance measures of the algorithms, it is seen that the NB tree provides the best in all rates of performance. So it is concluded that NB tree is best among these classifiers.

### 4.1 Methodology

The main goal of this study is to find the best classification algorithm among these five algorithms. The architecture of this system is as follows:

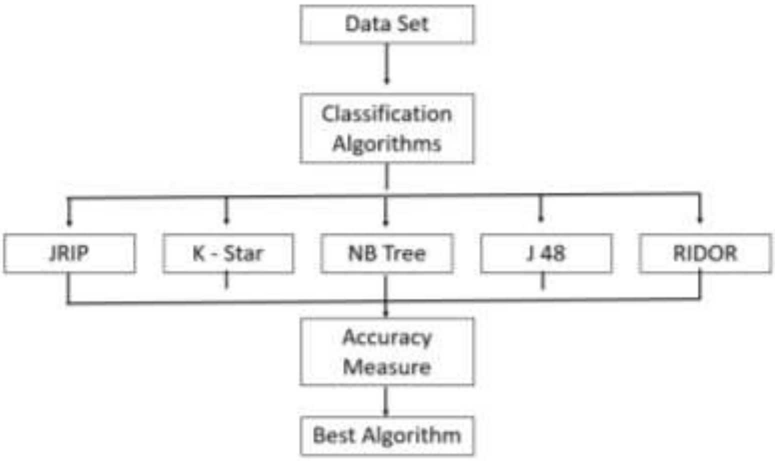


**Fig 1**: Methodology

#### 4.1.1 Dataset

The five data mining classification algorithms are applied on a dataset which produces the accuracy of each algorithm. The dataset used here was the Indian diabetes dataset which is used in some of the health-related research works. The dataset has 8 attributes with 768 instances. The attributes are a number of times pregnant, plasma glucose concentration, diastolic blood pressure, triceps skinfold thickness, 2-hour serum insulin, body mass index, diabetes pedigree function, and age.

#### 4.1.2 Classification

Classification is a function in data mining which is used to classify items in a large database to a set of classes. The classification process involves two steps – the training and testing of data. In this paper we have analyzed five classification algorithms, They are JRIP, K-Star, J48, NB tree, RIDOR.

#### 4.1.3 Weka Tool

Weka is an open source software which contains a collection of machine learning algorithms. It contains a set of tools for association, prediction, classification, clustering. In this paper, The accuracy of the five algorithms is analyzed using the weka tool.

In the field of health-related studies, classification plays a vital role. It precisely forecasts the target class from the data[3] current classification techniques provide more estimated results[3].

## 4.2 Performance Evaluation

**4.2.1 TP Rate:** TP stands for true positive. It denotes the number of predicted positive which is actually positive.

**4.2.2 FP Rate:** FP rate stands for False Positive. It denotes the number of predicted positive which is actually negative.

**4.2.3 Precision:** Precision is also called as positive predictive value(PPV). It is defined as the number of true positives which is divided by the total number of true and false positives.

$$Precision = \frac{tp}{tp+fp} \qquad (4.1)$$

**4.2.4 Recall:** Recall is also called as true positive rate. It is defined as the number of true positives which is divided by the total number of true positives and false negatives.

$$Recall = \frac{tp}{tp+fn} \qquad (4.2)$$

**4.2.5 F-Measure:** F-measure is the combination of precision and recall. F-measure is also known as balanced F-score. Here the weight of precision and recall are equal. F-measure is used to test the accuracy. The best value of this measure is 1 and the worst value is 0.

$$F - measure = \frac{2*Precision*recall}{tn+fp} \qquad (4.3)$$

**4.2.6 MCC:** MCC stands for Matthews correlation coefficient. It is performed between the observed and predicted classifications. This is used to calculate the values of the confusion matrix in data mining. It returns the value between -1 to +1.

The graph on comparison analysis of classifiers are provided in **Fig 2** and **Fig 3**

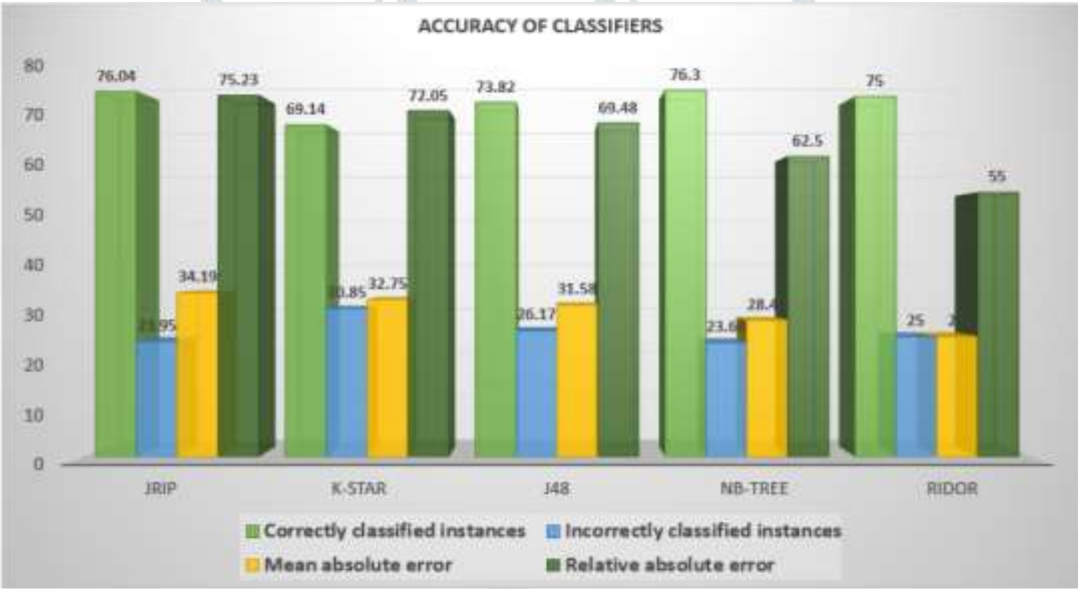
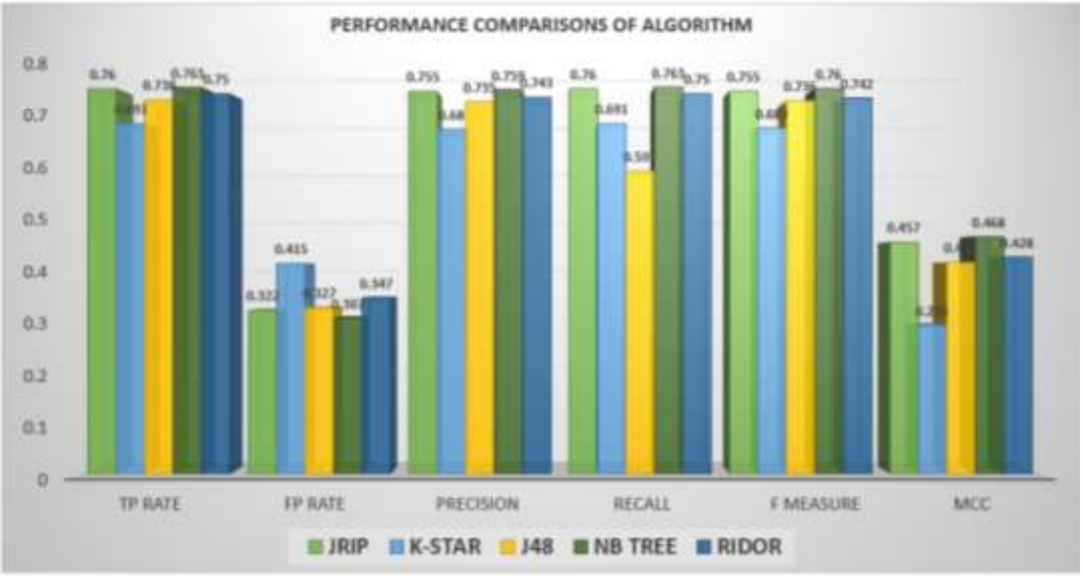


**Fig 2:** Accuracy of classifiers

The Accuracy of the classifiers is analyzed with the help of graph. It is proved that NB-Tree has the highest accuracy of 76.30% although the relative absolute error is little higher than RIDOR



.

**Fig 3:** Performance comparison of algorithms

According to the performance comparison of all algorithms, it is seen that NB-Tree has the highest accuracy among all the classifier in this dataset.

## V. CONCLUSION

In this paper, five classification algorithms in data mining were applied to the data set to predict the accuracy of each classifier. Based on the comparison of the accuracy it has been proved that the NB-Tree algorithm performs best on the Indian diabetes dataset with the accuracy of 76.30% and also a model graph also provided for reference. Therefore we conclude that NB Tree is potentially effective among all the five classifiers in this dataset.

## VI. ACKNOWLEDGMENT

We express our deep gratitude to Dr.Pothula Sujatha, our project guide for her enthusiastic encouragement and kind motivation throughout the period of study.

## REFERENCES

[1] Tekieh, M. H., & Raahemi, B. (2015). Importance of Data Mining in Healthcare. *International Conference on Advances in Social Networks Analysis and Mining*, 1057–1062. https://doi.org/10.1145/2808797.2809367

[2] Hassan, M. A., Shehab, M. E., & Hamed, E. M. R. (2016). A comparative study of classification algorithms in e-health environment. *2016 6th International Conference on Digital Information Processing and Communications, ICDIPC 2016*, 42–47. https://doi.org/10.1109/ICDIPC.2016.7470789

[3] Engineering, C. (2017). TranslatedcopyofTank_cultivation_of_Ulva_prolifera_in_deep_seawate, 13779–13786. https://doi.org/10.15680/IJIRCCE.2017.

[4] Engineer, C. (2016). a Comparative Study of Soil Classifications, *7*(7), 1–3. https://doi.org/10.14569/IJACSA.2016.070753

[5] Waseem, S., Salman, A., & Muhammad, A. K. (2013). Feature subset selection using association rule mining and JRip classifier. *International Journal of Physical Sciences*, *8*(18), 885–896. https://doi.org/10.5897/IJPS2013.3842

[6] Veeralakshmi, V., & Ramyachitra, D. (2015). Ripple Down Rule learner ( RIDOR ) Classifier for IRIS Dataset. *International Journal of Computer Science Engineering (IJCSE)*, *4*(03), 79–85.

[7] Tejera Hernández, D. C. (2015). An Experimental Study of K* Algorithm. *International Journal of Information Engineering and Electronic Business*, *7*(2), 14–19. https://doi.org/10.5815/ijieeb.2015.02.03

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

[9] Kasimzade, A. A., & Tuhta, S. (2013). Ambient vibration analysis of steel structure. *Test*, *80*(5), 483–492.

[10] Kaur, G., & Chhabra, A. (2014). Improved J48 Classification Algorithm for the Prediction of Diabetes. *International Journal of Computer Applications*, *98*(22), 13–17. https://doi.org/10.5120/17314-7433

[11] Aljawarneh, S., Yassein, M. B., & Aljundi, M. (2017). An enhanced J48 classification algorithm for the anomaly intrusion detection systems. *Cluster Computing*, 1–17. https://doi.org/10.1007/s10586-017-1109-8

[12] Patil, T. R., & Sherekar, S, S. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. *International Journal Of Computer Science And Applications, ISSN: 0974-1011*, *6*(2), 256–261. https://doi.org/ISSN: 0974-1011

[13] https://www.saedsayad.com/naive_bayesian.htm

[14] Mohamed, W. N. H. W., Salleh, M. N. M., & Omar, A. H. (2013). A comparative study of Reduced Error Pruning method in decision tree algorithms. *Proceedings - 2012 IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2012*, 392–397. https://doi.org/10.1109/ICCSCE.2012.6487177

[15] Sujatha, M., Prabhakar, S., & Devi, G. (2013). A Survey of Classification Techniques in Data Mining. *Ijiet.Com*, *2*(4), 86–92. Retrieved from http://ijiet.com/wp-content/uploads/2013/09/12.pdf

[16] Krishnaiah, V., Narsimha, G., & Chandra, N. S. (2014). International Journal of Computer Sciences Survey of Classification Techniques in Data Mining, (9).