International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 11 | Nov 2018 www.irjet.net p-ISSN: 2395-0072

# Feature Selection and Classifier Accuracy of Data Mining Algorithms

**Fifie Francis**[1]

[1]*Lecturer, Department of Humanities, ST PAULS COLLEGE, Bengaluru, Karnataka, India*

---

**Abstract –** *The combination of medical data and data mining algorithms gives a good amount of contribution in the field of medical diagnosis. In data mining classification algorithms are famous for prediction analysis. J48 and Navie Bayes are classification algorithms most commonly used in the area of prediction analysis, whose efficiency can be increased by preprocessing techniques. This paper aims at finding optimum features using Gain Ratio Attribute Evaluator and Ranker algorithm. Various combinations of the ranked attributes are made based on info gain value and the classifier accuracy is analyzed for each combination. The Chronic Kidney Disease dataset from UCI repository is used for the experimentation.*

***Key Words***: **Chronic Kidney Disease, Feature Selection, J48, Naive Bayes.**

## 1. INTRODUCTION

Data mining is the practice of categorization through large datasets to identify patterns and establish associations to find solutions using data analysis. Data mining tools allow predicting future trends using these association rules [1]. CRISP-DM is a comprehensive data mining methodology and process model that provides anyone from novices to data mining experts with a complete blueprint for conducting a data mining project. CRISP-DM breaks down the life cycle of a data mining project into six phases.

- Business Understanding
  - Understanding project objectives and requirements; Data mining problem definition
- Data Understanding
  - Initial data collection and familiarization; Identify data quality issues; Initial, obvious results
- Data Preparation
  - Record and attribute selection; Data cleansing
- Modeling
  - Run the data mining tools
- Evaluation
  - Determine if results meet business objectives; Identify business issues that should have been addressed earlier
- Deployment
  - Put the resulting models into practice; Set up for continuous mining of the data [2]

Chronic kidney disease is a human life condition that affects the normal functioning of kidney which leads to an unhealthy stage. When this disease get worse the after effects will be, high amount of waste in blood, high blood pressure, low blood count, weak bones, nerve damage etc. The major cause of this disease can be diabetes, high blood pressure. Initial stage detection and treatment keeps it from getting worse. If it progresses it will lead to kidney failure where the only option will be dialysis or kidney transplantation to maintain life [3].

## 2. LITERATURE SURVEY

Dr. S. Vijayarani et al., [4] proposed a paper titled "Data Mining Classification Algorithms For Kidney Disease Prediction". The aim of this research work is to predict kidney diseases using classification algorithms such as Naive Bayes and Support Vector Machine. The best algorithm is analyzed based on the accuracy and execution time, and the tool used for this research was MAT LAB. From the experimental results it is observed that the performance of the SVM is better than the Naive Bayes classifier algorithm. The algorithm which has the higher accuracy with the minimum execution time has chosen as the best algorithm.

Tabassum S et al., [5] proposed a paper titled "Analysis and Prediction of Chronic Kidney Disease using Data Mining Techniques". The proposed system in the research paper uses Chronic Kidney Disease dataset and Data Mining techniques like Classification and Clustering. The Clustering algorithm like Expectation Maximization [EM] algorithm, Classification Algorithms like Artificial Neural Network [ANN] and C4.5 Algorithm were used. Accuracy was calculated for each data mining algorithm. The accuracy results were the following EM: 70%, ANN: 75% and C4.5: 96.75% and the paper concluded with C4.5 is more accurate for prediction of Chronic Kidney Disease whether the patient is affected from disease or not.

Guneet Kaur et al., [6] proposed a paper titled "Predict Chronic Kidney Disease Using Data Mining Algorithms In Hadoop". The data mining algorithms taken for this study were KNN( K- nearest neighbor) and SVM (Support Vector Machine). The tool used for the stuy was MAT LAB by accessing Hadoop in itself. The accuracy rate observed is as follows, SVM: 99.29%, KNN: 97.83%, and the combination of KNN & SVM: 98.93%. The error rate observed is as the follows SVM: 0.5%, KNN: 2.1%, and the combination of KNN & SVM: 0.75%.

Pushpa M. Patil [7] proposed a paper titled "Review on Prediction of Chronic Kidney Disease Using Data Mining Techniques". This I a review paper on several research papers on the topic prediction of chronic kidney disease using data mining classifiers. Based on the review study the author has found some classifier algorithms with highest accuracy which were proved and the classifiers are

Multilayer Perceptron, Random Forest, Naïve Bayes, SVM, KNN and Radial Basis Function.

S.Dilli Arasu et al., [8] proposed a paper titled "Review of Chronic Kidney Disease based on Data Mining Techniques". In this paper, the various data mining techniques are surveyed to predict kidney diseases and major problems are briefly explained. From the survey the author comes to a conclusion that the results may vary for different stages of kidney disease diagnosis based on the tools and techniques used. Data mining provides enhanced outcome in disease diagnosis when suitable techniques used. Thus, data mining is the noteworthy field for healthcare predictions.

## 3. METHODOLOGY

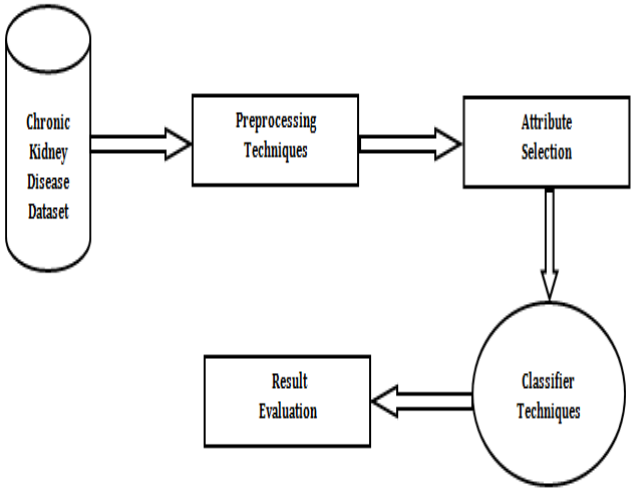


**Fig -1**: Methodology

Chronic Kidney Disease Dataset from UCI repository is used for this experiment on which different data mining preprocessing techniques are applied. Preprocessing techniques helps to understand remove all noisy data. The features are extracted from the preprocessed data using info gain attribute evaluator and ranker search method. Finally permutation combinations of attributes are formed with highest to lowest info gain value and the classifier techniques are executed for different combination of features. The results obtained for each combination are evaluated.

## 4. EXPERIMENTAL SETUP

A dataset is a pool of data. Most regularly a dataset corresponds to the contents of a single database table, where each and every column of the table speaks to a specific variable, and every tuple in the table relates to a given individual from the dataset being referred to. Chronic Kidney Disease Dataset from UCI repository is taken for the research work, contains 400 records with 250 chronic kidney disease (CKD) and 150 not chronic kidney disease and 24 + class = 25 attributes out of which 11 are numeric and 14 are nominal All instances in the dataset have 24 input attributes and 1 output attribute. The Table -1 below shows the attribute of this dataset. [9]

**Table -1:** Attribute Description

| Sr.No | Attriute Name | Description | Type |
|---|---|---|---|
| A1 | hemo | Hemoglobin in gms | Numerical |
| A2 | sc | Serum creatinine in mgs/dl | Numerical |
| A3 | sg | Specific Gravity | Nominal |
| A4 | pcv | Packed cell volume | Numerical |
| A5 | al | Albumin | Nominal |
| A6 | htn | Hypertension | Nominal |
| A7 | dm | Diabetes mellitus | Nominal |
| A8 | rbcc | Red blood cell count | Numerical |
| A9 | bu | Blood urea in mgs/dl | Numerical |
| A10 | bgr | Blood glucose random in mgs/dl | Numerical |
| A11 | sod | Sodium in mEq/L | Numerical |
| A12 | bp | Blood pressure in mm/Hg | Numerical |
| A13 | appet | Appetite | Nominal |
| A14 | pc | Pus cell | Nominal |
| A15 | pe | Pedal edema | Nominal |
| A16 | pot | Potassium in mEq/L | Numerical |
| A17 | rbc | Red blood cells | Nominal |
| A18 | su | Sugar | Nominal |
| A19 | ane | Anemia | Nominal |
| A20 | age | Age in years | Numerical |
| A21 | wbcc | White blood cell count | Numerical |
| A22 | pcc | Pus cell clumps | Nominal |
| A23 | cad | Coronary artery disease | Nominal |
| A24 | ba | Bacteria | Nominal |
| A25 | class | Class | Nominal |

The numbering scheme A1 till A25 defined in Table -1 is later referred in Table -2. Weka is a mainstream suite of machine learning software written in Java, created at the University of Waikato, New Zealand. The Weka work surface contains collection of visualization tools and algorithms for data analysis and predictive modeling. It is composed in Java and runs on any platform. The algorithms can either be connected straightforwardly to a dataset or called from your own Java code. [9]

## 5. RESULT AND DISCUSSION

There are plenty of attribute selecting algorithms available in the tool which we have selected for this research that is the Weka tool which helps to select specific number of attributes from large set of attributes. To choose those attributes the inbuilt algorithm used was Gain Ratio Attribute Evaluator and Ranker algorithm, which ranked the 24 attributes apart from the class attribute in the following order.

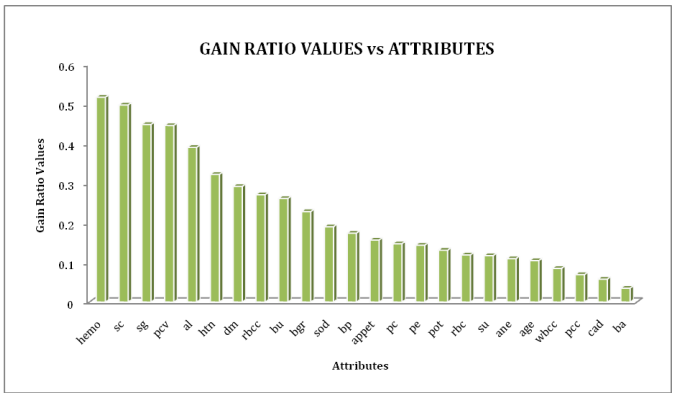


**Chart -1**: Gain Ratio Chart

Based on the above chart the 24 attributes of the chronic kidney disease dataset is taken in the descending order of their gain ratio values along with the class attributes, and both J48 and Navie Bayes algorithms are executed on each combination to analyze the performance of both algorithms on each combination of attributes and its hidden relation with the gain ratio values. The graphs below shows the accuracy of algorithm for different combination of attributes along with class attribute.

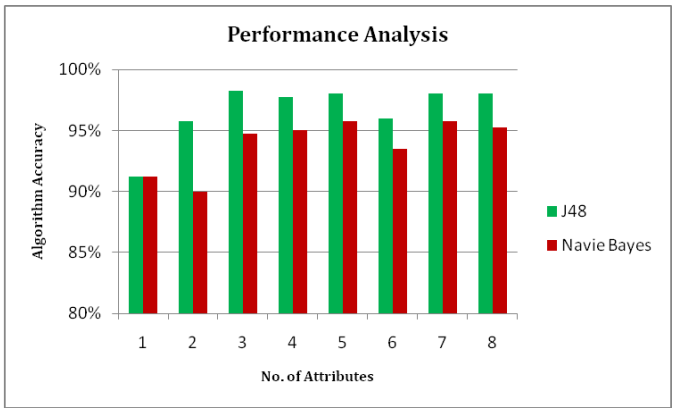


**Chart -2**: J48 and Navie Bayes performance for attribute count 1 to 8 along with class attribute

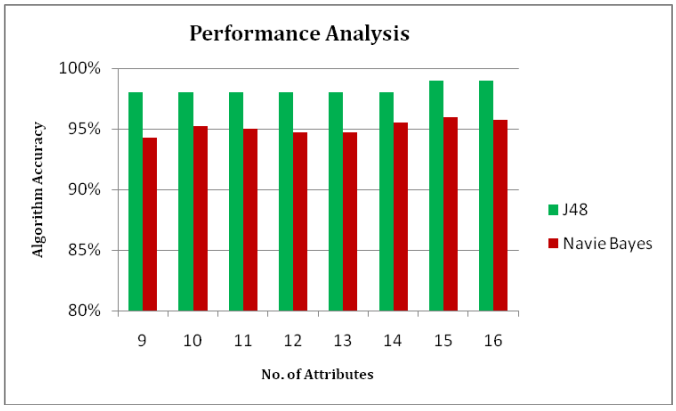


**Chart -3**: J48 and Navie Bayes performance for attribute count 9 to 16 along with class attribute

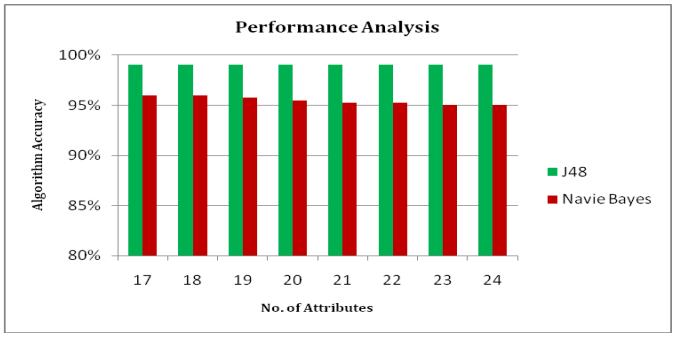


**Chart -4**:J48 and Navie Bayes performance for attribute count 17 to 24 along with class attribute

The Table- 2 below shows the accuracy obtained for J48 and Navie Bayes algorithms for the different attribute combination along with class attribute

**Table -2:** Algorithm Accuracy

| No. of Attributes | Name of Attribute | J48 | Navie Bayes |
|---|---|---|---|
| 1 | A1 | 91.25% | 91.25% |
| 2 | A1, A2 | 95.75% | 90% |
| 3 | A1- A3 | 98.25% | 94.75% |
| 4 | A1 – A4 | 97.75% | 95% |
| 5 | A1 – A5 | 98% | 95.75% |
| 6 | A1 – A6 | 96% | 93.50% |
| 7 | A1 – A7 | 98% | 95.75% |
| 8 | A1 – A8 | 98% | 95.25% |
| 9 | A1 – A9 | 98% | 94.25% |
| 10 | A1 – A10 | 98% | 95.25% |
| 11 | A1 – A11 | 98% | 95% |
| 12 | A1 – A12 | 98% | 94.75% |
| 13 | A1 – A13 | 98% | 94.75% |
| 14 | A1 – A14 | 98% | 95.50% |
| 15 | A1 – A15 | 99% | 96% |
| 16 | A1 – A16 | 99% | 95.75% |
| 17 | A1 – A17 | 99% | 96% |
| 18 | A1 – A18 | 99% | 96% |
| 19 | A1 – A19 | 99% | 95.75% |
| 20 | A1 – A20 | 99% | 95.50% |
| 21 | A1 – A21 | 99% | 95.25% |
| 22 | A1 – A22 | 99% | 95.25% |
| 23 | A1 – A23 | 99% | 95% |
| 24 | A1 – A24 | 99% | 95% |

The Table- 3 does an analysis about the performance obtained for both the algorithms for different attribute count.

**Table -3:** Performance Analysis

| No. of Attributes | J48 | No. of Attributes | Navie Bayes |
|---|---|---|---|
| 15 | 99% | 15 | 96% |

After analyzing the performance the maximum accuracy for J48 and Navie Bayes was given in an attribute count of 15 along with class label and the accuracy was 99% and 96 % respectively. The 15 attributes were

- Hemoglobin in gms
- Serum creatinine in mgs/dl
- Specific Gravity
- Packed cell volume
- Albumin
- Hypertension
- Diabetes mellitus
- Red blood cell count
- Blood urea in mgs/dl
- Blood glucose random in mgs/dl
- Sodium in mEq/L
- Blood pressure in mm/Hg
- Appetite
- Pus cell
- Pedal edema

## 6. CONCLUSION

The combination of medical data and data mining algorithms gives a good amount of contribution in the field of medical diagnosis. This study aims at analyzing the accuracy of J48 and Navie Bayes algorithms in different attribute numbers which are ranked using Gain Ratio Attribute Evaluator and Ranker algorithm. The highest accuracy obtained for J48 was 99% and for Navie Bayes was 96%. The ROC curve for the respective is given below.

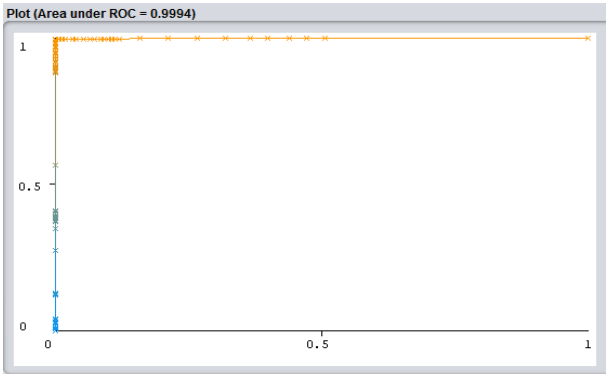


**Fig -2**: ROC Curve for J48

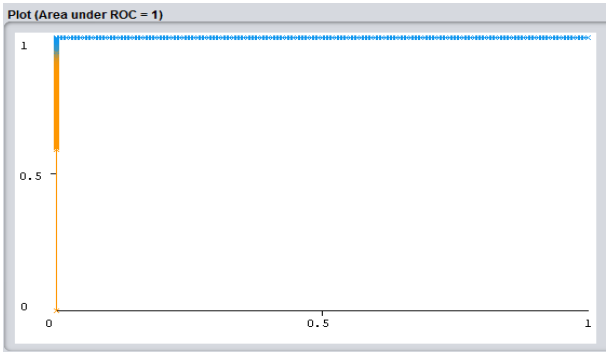


**Fig 3**: ROC Curve for Navie Bayes

## REFERENCES

[1] SearchSQLServer. (2018). What is data mining? - Definition from WhatIs.com. [online] Available at: https://searchsqlserver.techtarget.com/definition/data-mining [Accessed 26 Nov. 2018].

[2] Paginas.fe.up.pt. (2018). [online] Available at: https://paginas.fe.up.pt/~ec/files_0405/slides/02%20CRISP.pdf [Accessed 26 Nov. 2018].

[3] National Kidney Foundation. (2018). About Chronic Kidney Disease. [online] Available at: https://www.kidney.org/atoz/content/about-chronic-kidney-disease [Accessed 26 Nov. 2018].

[4] Dr. S. Vijayarani and Mr. S. Dhayanand "Data Mining Classification Algorithms for Kidney Disease Prediction" International Journal on Cybernetics & Informatics (IJCI) Vol. 4, No. 4, August 2015.

[5] Tabassum S, Mamatha Bai B G and Jharna Majumdar "Analysis and Prediction of Chronic Kidney Disease using Data Mining Techniques" International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 4, Issue 9, September 2017.

[6] Guneet Kaur and Ajay Sharma "Predict Chronic Kidney Disease Using Data Mining Algorithms In Hadoop" International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835 Volume-5, Issue-4, Apr.-2018

[7] Pushpa M. Patil "Review On Prediction of Chronic Kidney Disease Using Data Mining Techniques" International Journal of Computer Science and Mobile Computing, Vol.5 Issue.5, May- 2016.

[8] S.Dilli Arasu and Dr. R.Thirumalaiselvi "Review of Chronic Kidney Disease based on Data Mining Techniques" International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, November 23 (2017).

[9] Fifie Francis and Saleema J S "Feature Selection in Data Mining using Permutation Combination" International Journal of Advanced Research in Computer Science, Volume 8, No. 3, March – April 2017.

## AUTHORS

Lecturer
ST PAULS COLLEGE
Bengaluru, Karnataka
India.