A comparative analysis of the performance of multiple data mining classification approaches using the K_n fold validation

Abstract:

In Healthcare the data is very large and sensitive. The data is mandatory to be handled very carefully without any negligence. A variety of data mining categorization approaches have been employed in the healthcare industry to assess the quality of services. On the basis of 150 patients' records, this study provides and evaluates the experience of implementing various data mining methodologies and procedures. Using data mining techniques, a new method for determining a product's correctness has emerged. The evaluation of performance on data mining classification by using a different algorithms like Decision Tree, Naïve Bayes, KNN, Radom Tree Set and Rule Model. Finally we tend to aim to contemplate the performance analysis of accuracy, sensitivity and specificity proportion to produce a result.

Keywords: Healthcare, Data Mining, Classification, fibroid data set, Performance Analysis, Rapid Miner.

I. Introduction on Data mining

Two Cross Corporation (1999) defines data mining as "the act of discovering patterns and relationships in data that may be used to produce credible predictions" utilising a number of data analysis techniques. "The repetitive and participatory process of uncovering valid, unique, useful, and intelligible patterns or models in enormous databases," Kumer and Zaki (2000) define it as well. Data mining is a concept that has existed for quite some time. As Michael (2002) pointed out, the approaches have existed as algorithms used in academic research in the domains of statistics and machine learning for years or decades. Many of the techniques that are currently part of data mining are centred on pattern recognition and categorization.

For obtaining knowledge from big databases, a range of data mining approaches are available. "Descriptive" and "predictive" data are the two types (Hong and Weiss, 1999; (Han and Kamber, 2002). (Han and Kamber, 2002). Explanatory models that summarise data in order to draw conclusions are dealt with in the descriptive. The main applications of descriptive data mining are database summarization and visualisation. Using this idea, one may investigate a data set's general behaviour, which is difficult to derive from a huge database, using several degrees of abstraction. When it comes to predictive analytics, on the other hand, it's all about constructing models that can foretell future events. Cataloging and regression are dual of the most popular mining of data jobs.

Classification is an assignment in data mining. As previously stated, mining of data is a form of

learning machines, influenced by pattern recognition, an area of science tasked with the task of classifying things into a quantity of distinct groupings mentioned to as classes. Objects are little data units that are particular to a problematic and are generally referred to as patterns.

The creation of a classifier and its application are two distinct stages of classification prediction. The former focuses on developing an arrangement prototypical by specifying a fixed of predefined classes obtained from a training set using machine learning. Every sample in the working out set is presumed to fit to a previously defined class based on the session attribute label. The model is signified using arrangement rubrics, mathematical formulae or decision tree. The latter encompasses the use of a classifying forecast or categorized unidentified objects using designs discovered in the preparation set.

Data mining is already widely employed in the healthcare business, and as a result of its enormous success, interest in this subject is growing. As our dataset, we used the fibroid data set. We selected it since it is a sickness that affects the whole world. Female reproductive system fibroids are the maximum prevalent kind of feminine reproductive system tumour. Fibroids, alternatively referred to as fibromas/ leiomyomas, uterine myomas, are stiff, compact tumours that develop in the uterus and are comprised of smooth muscle cells and fibrous connective tissue [1]. This study compares the performance of numerous categorization methods.

We sought to compare the results of several classification algorithms in the Rapid Miner data mining programme using a few ROC parameters.

In this research, we examined the efficacy of multiple participant data mining categorization approaches utilising fibroid data. This work assisted in the discovery of the most accurate and least error-prone categorization algorithms in comparison to other techniques for a certain characteristic. The techniques of kNearest Neighbours (kNN), Random Tree Set (RTS), Naive Bayes (NB), Rule Mode (RM), Decision Tree (DT), and others were studied. The accuracy of these strategies is used to assess their training performance. To solve classification difficulties, machine learning tools like Rapid Miner are used. After determining the most appropriate classification technique for the fibroid dataset, this study will assist researchers in determining the best potential findings from the available data inside the datasets. As a result, the researchers will be able to perform data mining analysis more efficiently if they employ a proper classification technique on the fibroid dataset.

II. Introduction on Fibroids

- **Fibroids as a Health Problem:** Fibroids are benign tumours of the uterine muscle (womb). Fibroids are referred to in medicine as "myomas." In most cases, fibroid tumours are not dangerous (not cancerous). Fibroids can develop as a single tumour or as a collection of tumours throughout the womb. As little as an apple seed, they can grow into grapefruit-sized fruits. They may grow to enormous sizes in certain circumstances[1]. When fibroids develop in the uterus, they can impair fertility and pregnancy, as well as raise the possibility of complications after delivery.

- **Who is at risk of developing fibroids?**

  - **Age:** Women in their 30s and 40s through menopause are more likely to develop fibroids. Fistulae tend to decrease in size after menopause.
  - **Family History:** Fibroids run in our family, which raises our chances of developing them ourselves. Women who have fibroids in their mothers have a three-fold greater chance of

developing them themselves.

- **Ethnic Origin**: Fibroids are more common in African-American women than in white women.
- **Obesity**: Fibroids are more common in women who are obese. The risk is three to four times larger in obese women.
- **Eating Habits**: A diet high in red meat (e.g., beef)[1].

We want to utilise Orange Canvas to develop a Naive Bayesian model so that the fibroids' causes may be classified as mild or severe. Their symptoms were studied.

III. Literature Survey

There have been reports of a slew of scientists working on the detection of common ailments. Infertility, fibroids, and the evidence: a new comprehensive review A team of medical professionals led by Elizabeth A Pritts's MD, included William H Parker's MD, and Olive's MD[2]. Uterine fibroid identification utilising wavelet characteristics and a neural network classifier has been developed by Sriraam et al (2010). The vertical and horizontal coefficients are calculated using a three-level wavelet packet decomposition based on a user-defined ROI.[7].

IV. Description of The Data Set

The dataset used in our study is fibroid data set. This data set is collected by expert doctors and collected by some experts. There are 151 observations and it contains 10 attributes and one target class with no missing values. All the patients in this data set are females, living near Tumkur dist, Karnataka. Patients' names are hidden. The attributes are shown in Table 1 below,

Attribute no	Attribute name	Description
1	Age	Age of a patient
2	Status	Married or Single
3	HB	Heavy Bleeding
4	PP	Pelvic Pain
5	FT	Fibroid type
6	LBP	Lower back pain
7	PDS	Pain during sex
8	FU	Frequent Urination
9	NFP	No. of fibroid
10	SF	Size of a fibroid

Data Set	No of Examples	Input attributes	Output Classes	Total no of attributes	Missing attributes	Noise attributes
Fibroid	150	10	1	11	No	No

Table 1: Fibroid Data set.

a. Selection Step

This stage retrieves data from the database that is relevant to the analysis task. During the selection process, a dataset for analysis is generated. Data from medical fibroid databases were used to evaluate the efficacy of various data mining classification algorithms.[4].

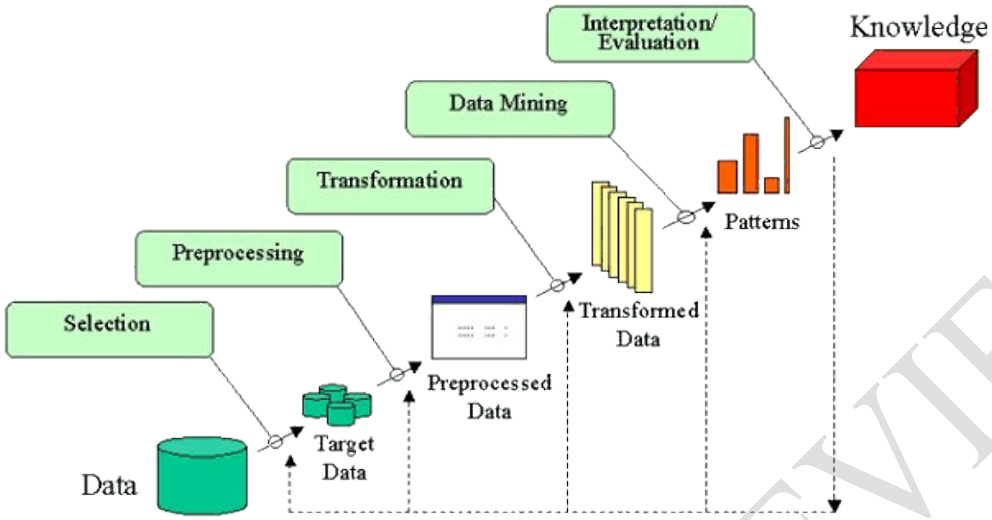


Figure 1: KDD Process.[14]

b. Preprocessing step

A database's ability to be harmed by noise, missing data, and inconsistency is heightened when it is large, complicated, and possibly derived from numerous different heterogeneous sources. Since the selection stage selected a target dataset, it's treated in this phase to address the aforementioned problem. Using the fast miner data mining application, the selected healthcare fibroid datasets are treated to remove missing values and noise before being presented to the classification algorithm.

c. Transformation step

In this stage, procedures like as smoothing, summary or aggregate, generalisation, normalisation, discretization, and feature development are used to change or combine data into forms suited for mining. The data mining programme Rapider Miner was used in this experiment to accomplish the aforementioned purpose. Rapider Miner modules such as derive, feature selection, and type are linked to the dataset to perform data transformations as needed for the research.[5].

d. Data Mining step

In the KDD procedure, data mining systems are employed to extract patterns from data. Intelligent techniques are utilised to extract data patterns at this stage of the KDD process. Analyses of the dataset are performed utilising the data mining job that was utilised. CHAID, ID3, Bayes Net, C4.5, LDA, KNN, FT, NB, LMT, SVM, C-RT, QUEST, NN-RBFN, J48, MLP, and Prototype-NN are used in this work to excerpt data designs from fibroid datasets by means of a machine learning tool.

e. Interpretation step

This stage comprises the assessment of patterns and the illustration of knowledge. This critical step makes use of imagining tools to aid consumers in grasping and analyzing data mining results. In this research, the cataloging strategy with the uppermost percentage of cataloging accuracy is measured to be the best cataloging approach for a precise dataset. The contrast is illustrated via charts. Healthcare researchers make judgments depending on the classification technique's produced classifier [6].

V. ASSESSMENT METHODOLOGIES

In this exertion, exactitude and recall, Receiver Operating Characteristic (ROC), and lift were utilised as enactment indicators for determining the accurateness of a cataloging typical . Each one was utilized as needed throughout the performance analysis. At a distance from the aforementioned critical enactment criteria, the inquiry will also evaluate the classifiers' swiftness and resilience.

- ❖ Accuracy.
- ❖ Exactness and Recall.

The exactitude of percentages as a routine statistic has remained demonstrated to be disingenuous (Provost et al. in 1998, Guo and Viktor in 2004). Given that 96 percent of the majority may reside in that place, a classifier labelling all sections as the mainstream class will achieve a 94 percent accuracy. Meanwhile, since to the dataset's bias, the pigeon-holing may have wrongly labelled specific occurrences of the smaller class as mainstream, despite their apparent accuracy. As a result, it is necessary to evaluate accuracy using another metric, such as precision and recall. A classifier's performance can be represented by the "confusion matrix" shown below [9], which is a two-class issue.

Table 2: Confusion Matrix.[15]

Predicted Class	True Outcome: Customers Default or Not	
	Positive (or Good)	Negative (Bad)
Positive (or Good)	True Positives	False Negatives (Type I Error)
Negative (Bad)	False Negatives (Type II Error)	True Negatives

Where TP, TN, FP, and FN correspond to the confusion matrix illustrated in the picture. Precision is used in this case to refer to the actual proportion of mail responses projected by the cataloging model, which interprets into revenues on sending expenses [10]. The recollection, on the additional hand, quantifies the proportion of consumers who were recognized and had to be besieged.

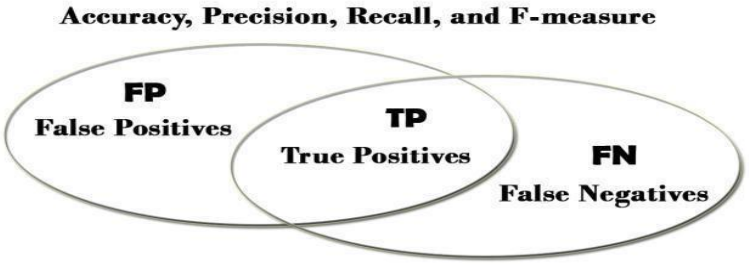


Fig. 2. Accuracy and precision [16]

VI. THE RESULTS OF EXPERIMENTS AND THE ANALYSIS

This segment contains the investigational data and examination from this study. Section 3 discussed the experimental design process. For the experimentations, the data set was submitted to a number of categorization systems. Table 1 summarizes the dataset used in this investigation.

To accomplish the study's objectives, the Rapid Miner machine learning tool for data mining was

Precision P=TP/T+P
Recall r=TP/TP+FN
Accuracy r=TP+Tn/TP+TN+FP+FN

Harmonic mean of F measure: precision and recall

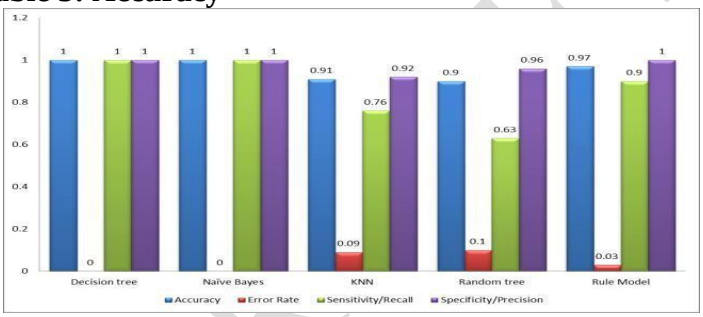
Accuracy F=1

1/2(1/P+1/r)

employed. The percentage of accurateness and the blunder rate for cataloging techniques are the examination's measuring metrics. These statistics suggest that when a classification strategy is applied to a dataset, a high correctness rate and a stumpy mistake rate imply that the dataset is properly classified by the classifier developed. On the other hand, a low accurateness ratio and a high mistake rate for a organization strategy pragmatic to a dataset suggest that the resulting classifier classified the dataset erroneously. For experiments, the data is initially partitioned into working out and testing sets. The classifier is developed on the basis of the working out set and verified on the basis of the test set. In this experiment, the percentages utilized for training and testing data were 66% and 34%, respectively. Then, using the machine learning procedures discussed before, the involved classification algorithms are implemented via tenfold cross validation using a linear sample type (alternative methods include automated, shuffled, and stratified). Finally, the results are summarized in percentages of accurateness and mistake rates.

Table 3: Accuracy

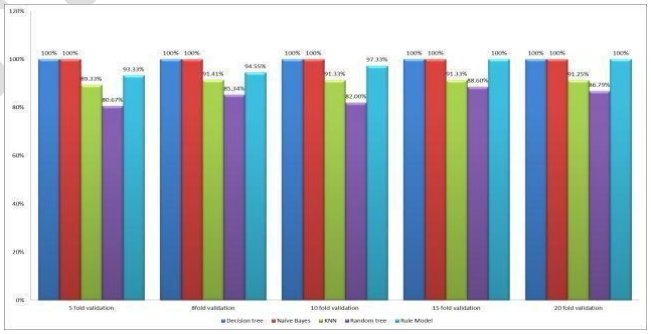
Algorithm	Accuracy	Error Rate	Sensitivity/ Recall	Specificity/ Precision
Decision tree	1	0	1	1
Naïve Bayes	1	0	1	1
KNN	0.91	0.09	0.76	0.92
Random tree	0.9	0.1	0.63	0.96
Rule Model	0.97	0.03	0.9	1



Let us focus on the accuracy factor in the table above. We may get the following results by utilising the Rapid Miner programme for different folds and algorithms:

Table 4: Different fold

Algorithm	5 fold validation	5 fold validation	10 fold validation	15 fold validation	20 fold validation
Decision tree	100%	100%	100%	100%	100%
Naïve Bayes	100%	100%	100%	100%	100%
KNN	89.33%	91.41%	91.33%	91.33%	91.25%
Random tree	80.67%	85.34%	82.00%	88.60%	86.79%
Rule Model	93.33%	94.55%	97.33%	100%	100%



VII. CONCLUSION

In this learning study, we have used different classification methods for data mining under the Rapid Miner tool. We used a different fold range of 5 to 20 for each classification method (the Rapid Miner tool defaults to 10 fold validation), so the results will be different. According to table 4, we can determine as follows: decision tree and naive Bayes: For both the algorithms, we analyse that for all the different folds, they will give the constant result of 100%.

KNN: In this algorithm, for 5 fold we will get a result of 89.33%. It will increase up to 10 folds, i.e., 91.33%, then it will decrease after the 17th fold up to the 20th fold validation.

Random Tree: In this algorithm, it is clear that alternative fold validation will give increasing and decreasing nature. To some extent, we get an increasing accuracy value, but after that, we will get a constant result of 100%.

Finally, by taking table 3, we can conclude that among the different classification methods of data mining, i.e., decision tree (DT) and Nave Bayes (NB) algorithms, the latter gives performance analysis

with the best results for all the different results.

VIII. References

[1]. Elizabeth A. Pritts, M.D.,a William H. Parker, M.D.,b and David L. Olive, M.D.a "Fibroids and infertility: an updated systematic review of the evidence"

[2]. [Neelamadhab Padhy, Dr. Pragnyaban Mishra and Rasmita Panigrahi, "The Survey of Data

i. Mining Applications and Feature Scope, *International Journal of Computer Science, Engineering and Information Technology*

[3]. (IJCSEIT)", vol.2, no.3, June

[4]. Heikki, Mannila, *Data mining: machine learning, statistics and databases*, IEEE, 1996.

[5]. Fayadd, U., Piatesky -Shapiro, G., and Smyth, P, From Data Mining To Knowledge Discovery in Databases", The MIT Press, ISBN 0–26256097– 6, Fayap, 1996.

[6]. Piatetsky-Shapiro, Gregory, *The Data-Mining Industry Coming of Age*,"IEEE Intelligent

[7]. Systems, 2000.

[8]. Jing He, Advances in Data Mining: History and Future, Third international Symposium on Information Technology Application, 978-0-7695- 3859-4, IEEE, 2009. http://shodhganga.inflibnet.ac.in/bitstream/10603 /22904/7/07_chapter2.pdf

[9]. Berry, M.and Linoff, G., Master Data Mining: The Art and Science of Customer Relationship Management, Wiley publisher, 2000.

[10]. Provost, F. and Fawcett, T. (1997). Analysis and

[11]. visualization of classifier performance: Comparison under imprecise class and cost distributions. In proceedings of the Third international conference on Knowledge discovery and data mining, Menlo park, CS.

[12]. Flach, P. and Gamberger D.(2001) Subgroup evaluation and decision support for a direct mailing marketing problem. Aspects of Data Mining, Decision Support and Meta- Learning,

[13]. [Online] available from: http://www.informatik.unifreiburg

[14]. http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html

[15]. https://glassboxmedicine.com/2019/02/17/measuring-performance-the-confusion-matrix/

[16]. https://www.researchgate.net/figure/A-Typical-precision-recall-diagram_fig1_225579773