How to Cite:
Kumar, A. N., Vanaja, D. S., Reddy, A. M., Das, S., Dey, S., & Sucharitha, Y. (2022). Analysis of patient health condition based on hybrid machine learning algorithm. *International Journal of Health Sciences*, *6*(S3), 10532–10544. https://doi.org/10.53730/ijhs.v6nS3.8358

Analysis of patient health condition based on hybrid machine learning algorithm

Anil Kumar N.
Assistant Professor, Department of Electronics & Instrumentation Engineering, Sree Vidyanikethan Engineering College, Tirupati, Andhra Pradesh, India

D. S. Vanaja
Assistant Professor, Sri Venkatesa Perumal College of Engineering & Technology (Autonomous), Puttur, Andhra Pradesh, India

Alumuru Mahesh Reddy
Assistant Professor, Sri Venkatesa Perumal College of Engineering & Technology (Autonomous), Puttur, Andhra Pradesh, India

Susmita Das
Assistant Professor, Electronics and Instrumentation Engineering, Narula Institute of Technology, Kolkata, West Bengal, India

Sharmistha Dey
Assistant Professor, Department of UIC, Chandigarh University, Gharuan, Punjab, India

Yadala Sucharitha
Assistant Professor, Department of Computer Science and Engineering, CMR Institute of Technology, Hyderabad, TS state, India

***Abstract*---**In data mining, the classification methods are used to determine the relationships between the various objects of the interactions database. In this research, the objective is predominantly focused on the prediction of three types of functioning test level. The main purpose of the research work is to analyze human health condition, realised by fluctuation of specific ranges such as Bilirubin, Albumin, Prothrombin time (INR), Ascites, Encephalopathy , Bicarbonate, Calcium etc. It can improve the disease by taking medical diagnosis based on Apriori algorithm by generating rule for the most significant parameters of all three functioning test level. In this article, the experiment can be carried out by using a variety of different sizes of support count based on the association factor of attributes in the kidney and liver functional test data from a wide

range of patients. The goal of the experiment is to gain an understanding of the effect that the Apriori algorithm has on the amount of time it takes for the execution, the precision of the best rule discovered from the mining of frequent patterns, and the number of association rules that are generated. The efficiency of the algorithm is calculated based on different support counts and number of rule generation.

***Keywords*---**data mining, machine learning, health, apriori algorithm, liver, database.

Introduction

It is now possible for decision makers to make proactive, data-driven judgments by using a new powerful technique called data mining, which extracts predicted information from vast datasets. By using data mining to extract information from a data collection, it is possible to organise it in a way that will be useful in the future[1]. Analytical tools from data mining software are used to analyse data.. Use this programme to analyse data from a variety of categories or dimensions and to summarise the links that have been discovered. In order to find recurring patterns and connections in a huge amount of data, data mining is essential. Computational approaches for detecting patterns in big data sets using Data Mining (KDD) will be used in conjunction with artificial intelligence, machine learning, statistics, and database systems (KDD). Nontrivial extraction of implicit data from databases also refers to previously unknown and possibly relevant information. Data mining in databases is usually referred to as "knowledge discovery." However, data mining is essentially a component of the process of discovering new information[2]. As part of the Knowledge Discovery in Database (KDD) process, which is an iterative process, data cleaning, often referred to as data cleansing, is performed. Integration of data from several sources, typically heterogeneous, may be done via the use of a single data source. Analyzing, deciding, and retrieving data from a collection is all part of the data selection process. Data transformation, often known as data consolidation, is the process of transforming chosen data into forms suitable for mining [3]. Data mining is an application that uses a particular algorithm to identify patterns in data. Data preparation, data selection, data cleaning, effective interpretation of mining findings, and the inclusion of relevant previous knowledge are just a few of the tasks that the KDD process adds to the list. KDD is a term used to describe the process of extracting usable information from large amounts of data. It involves an interpretation of the pattern to make decision and evaluation and possibly the task of what qualifies as knowledge and process as shown in figure.1

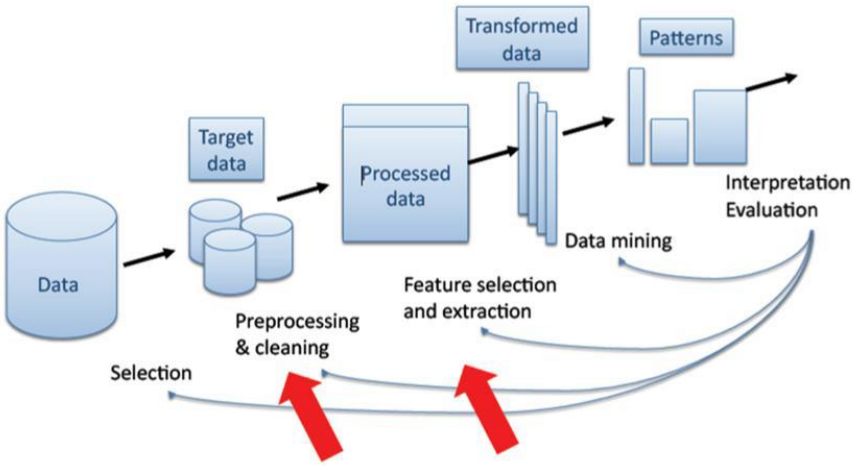

Figure 1. The steps in KDD process

Evaluation analysis

Time-series data analysis, pattern matching, prediction, or clustering of time-related data and similarity-based data analyses and model regularities or trends, for objects whose behaviour changes over time are only some of the distinct features of data evolution analysis.

Pattern discovery

Both descriptive and predictive data mining include describing the broad characteristics of the current data, while the latter involves making predictions based on inferences drawn from the available data. Functionalities and the diversity of information they uncover influence the patterns that may be identified.

Characterization

User-specified class labels are used in data characterisation, wherein generic traits are summarised with regard to characteristic rules under the selection target class. Typically, data retrieval requires a query to the database [4]. Subsequently, the module is condensed by deriving the core of the data at successive abstraction levels.

Discrimination

Discriminate rules are generated by the comparison of the general characteristics of objects in two classes, referred to as the target class and the contrasting class, in data discrimination.

Association analysis

Analysis of the frequency of items appearing together in transactional databases, and the threshold of support and confidence for recognising the frequent item sets, is the goal of the association analysis [5]. When one item is included in a transaction with another, the confidence factor is used to determine the conditional likelihood of that item appearing.

Classification

For a certain class, classification analysis is a set of data that may be analysed. Class labels may be used to arrange the data that is needed for categorization. Typically, this strategy is employed in training sets where all objects are already linked with established class labels and a model is constructed. Each piece of information in a dataset is given a predetermined class or group via classification, a data mining approach. There are two stages to classifying data. To begin, a model may be constructed from a collection of characteristics using the data tuples in training data. An method for classifying input data yields the class label value for each tuple in the training set. Check the correctness of the model of the test data in the second phase of the classification process.. The model may be used to categorise the unknown data tuples if the model's accuracy is satisfactory.

Clustering

Clustering, like classification, is the extraction and formation of groups of data in different classifications [6]. As a result, this kind of categorization is referred to as "unsupervised classification." Intra-class similarity is one of a number of methods for maximising the similarity between items inside a single class. The goal of anointer-class similarity is to minimise the similarity between objects of various classes.

Deviation analysis

The examination of time-related data that varies over time is known as deviation analysis. Analyzing data via the lens of evolutionary tendencies allows us to characterise, categorise, or cluster data through time. However, the discrepancy between observed and expected values is taken into account and the source of deviations from predicted values is sought out. [7] A flexible and inclusive data mining system is consequently necessary, allowing for the finding of a wide range of information and at various levels of abstraction.

Literature Survey

The main illness's risk factors enable healthcare practitioners identify people at high risk of developing the condition. All health care providers have access to a large quantity of patient information. It is critical to analyse these datasets in order to get relevant information. The data mining technique is a valuable tool for analysing data and extracting meaningful information from the information contained within. Cardiologists have benefited from the use of data mining tools

to aid in the identification of the condition. Here are a few instances of how advanced data mining methods have been used to aid with illness detection. Some of the data mining approaches used in the medical literature review are discussed here. As a result, data mining is a cross-discipline endeavour that draws on a variety of expertise and approaches from a variety of fields. Different approaches from other fields, such neural networks [8], fuzzy or rough set theory, knowledge representation, inductive logic programming or high-performance computing may also be utilised depending on the data mining strategy. The data mining system may also incorporate techniques from spatial data analysis, information retrieval, pattern recognition, image analysis, signal processing, computer graphics, Web technology, economics, business, bioinformatics, or psychology, depending on the types of data to be mined or on the specific data mining application. Figure 2. shows numerous criteria that may be used to classify data mining systems.

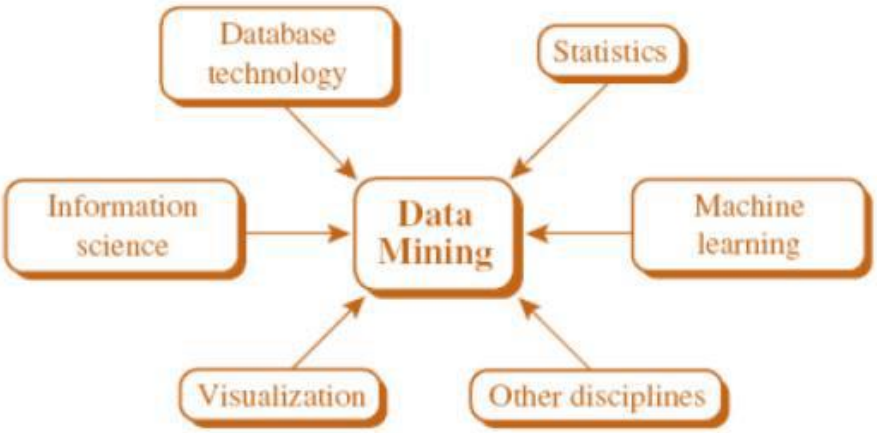

Figure 2. Data Mining and its Categories

Classification according to the kinds of databases mined

You may classify a data mining system by the types of databases it extracts data from. There are a variety of ways to classify database systems (such as data models, or the sorts of data or applications involved), each of which may need its own data mining approach.

Classification according to the kinds of knowledge mined

It is possible to classify data mining systems by the sorts of information they mine (e.g., characterisation and discrimination, association and correlation analysis, classification, prediction and clustering) [8] based on their data mining functions. In general, a complete data mining system has several data mining capabilities, some of which may be combined together [9].

Classification according to the kinds of techniques utilized

According on the underlying data mining methodologies, data mining systems may be divided into several subcategories. Systems can be classified as

autonomous, interactive, exploratory, or query-driven, depending on how much user interaction is required, or by the methods of data analysis they employ (such as database-oriented or data warehouse–oriented techniques, machine learning and statistical methods, visual representations, pattern recognition, and neural networks, among others). Multiple data mining methods may be used in a sophisticated data mining system or a successful, integrated strategy that combines the advantages of a few distinct approaches.

Classification according to the applications adapted

Applications may also be used to classify data mining systems. Depending on the application, data mining systems may be customised for a variety of different fields. Application-specific techniques are often required for integration with various applications. Consequently, a general data mining system may not be suitable for mining tasks that are specialised to a certain area. Data mining systems are plagued by several issues and snares. Advances in data mining have formed the current data mining applications to manage a wide range of issues, thanks to a variety of approaches and techniques. When applied to a bigger database, a system that performs well on short training sets may behave quite differently. It's possible that a data mining system to function well with clean training data, yet drastically degrade when the training data is contaminated with random noise. By giving it big datasets, machine learning can perform a wide range of jobs.

The general machine learning framework

Rather than being a single tool, machine learning is a collection of tools with varying strengths and shortcomings. It is through giving a model that acts as a learning framework that we enable computers to learn. To help you choose the right model for your machine learning application, scikit-learn (a collection of Python implementations of numerous machine learning algorithms and functions) has provided a guidance below. Knowledge of machine learning tools is an important part of becoming an expert [10]. The model should be built. Tune all important parameters for best performance after training the model on a dataset. Assess the model's usefulness and effectiveness. Supervised learning refers to the process of teaching a computer to do a certain job by giving it input data with predetermined results. Spam filters may be trained by supplying an algorithm with a large number of emails that have been classified as "spam" or not spam. Unsupervised learning is the process of letting a computer run wild with a dataset and seeing what it comes up with on its own, with no guidance from the user. Because it takes time to classify datasets for training supervised models, unsupervised machine learning approaches are appropriate. Similarly, there's semi-supervised learning, which uses a small amount of tagged observations together with a much larger set of unlabeled observations to train a model.

Regression and clustering methods may be used to categorise items in images, forecast the best price for the property to sell, and segment customers and markets. It is the science of making computers to behave without being explicitly programmed, which is called machine learning. Self-driving vehicles, realistic voice recognition, successful online search, and a much enhanced knowledge of

the human genome have all been made possible by machine learning in the last decade. Today, machine learning is so widely used that you probably don't even realise how many times you've been relying on it. Many AI researchers believe this is the greatest path to human-level AI advancement. Primary goal is for computers to learn on their own, with no human input, and then modify their behaviour appropriately. The following are some examples of machine learning: Unsupervised and supervised machine learning algorithms are typically referred to as such.

In order to make predictions about the future, supervised machine learning algorithms may apply what they've learned in the past to fresh data. With the use of an existing training dataset, the learning algorithm constructs an inferred function from which it may make educated guesses about the expected results. The material utilised to train unsupervised machine learning algorithms is unclassified and unlabeled. When data is unlabeled, unsupervised learning investigates how computers might infer functions that describe a hidden structure. Because they employ both labelled and unlabeled data, semi-supervised machine learning algorithms sit somewhere in the middle of supervised and unlabeled learning methods. Algorithms that use reinforcement learning to interact with their environment and find faults or rewards are called reinforcement machine learning algorithms. Reinforcement learning's most salient features are trial-and-error searching and delayed rewards.

In [11] author's, A-priori and k-means algorithms were used to construct a heart disease and renal failure prediction system. Her study employed A-prior and k-mean algorithm for 42 variables to identify renal failure patient. A-prior and k-means algorithms were used to analyse the data using machine learning techniques including distribution and attribute statistics. Calibrating and ROC plots were used to examine the data for accuracy. In [12] author's machine learning methods such as Support Vector Machine [SVM] and Random Forest [RF] were used in the presentation. These were used to examine, categorise, and compare various kernels and kernel parameters in relation to cancer, liver, and heart disease data sets. Random Forest and Support Vector Machine results were evaluated for several datasets, including breast cancer, liver, and heart illness. Results from several kernels were fine-tuned by carefully selecting and adjusting various parameter values. The findings were analysed more thoroughly in order to develop better methods for making predictions. As a consequence, various kernel functions of the SVM classification approach yielded diverse results.

In [13] author's classification techniques like Nave Bayes and Support Vector Machine may be used to predict renal illness. The primary goal of this study was to identify the classification algorithm with the optimum combination of classification accuracy and execution speed. The SVM outperforms the Naive Bayes classification method, according to the results of the experiments. In [14] A number of machine learning techniques, including Support Vector Machine (SVM), Decision Tree (C4.5), and a Bayesian Network (BN), were addressed by the authors in order to predict renal illness. There are 400 instances of Chronic Kidney Disease Dataset from the UCI Machine Learning Repository. SVM is placed second, but excels in classification time and accuracy. For classification and prediction in the medical arena, C4.5 has shown itself to be a strong classifier in

terms of accuracy and the shortest execution time. In [15], Evaluation of clustering techniques using Indian liver datasets. To begin, we used data from the UCI Machine Learning Repository (ILPR) [15], which comprises 583 medical records and 10 diagnostically essential factors for each patient. The second batch of data includes 500 records for liver patients and 13 different characteristics. We've combined the two datasets and utilised 10 similar features to create 1083 records for the testing. It was found that k-Means, AGNES, DBSCAN, OPTICS, and EM were the best clustering algorithms for the Indian liver dataset, with high accuracy, low entropy, high purity and high f-measure, as measured by Accuracy, Entropy, F-measure, and Purity.

Proposed System

A decision tree may be built from the provided collection of characteristics. greedy technique that generates a decision tree by a sequence of locally optimal decisions concerning the qualities to be used for data partitioning, Medical diagnostic data may be efficiently and accurately classified using standard machine learning techniques. Glomerular Filtration Rate (GFR), serum test, amount of blood urea cretonne, uric acid, as well as liver function test under Child-Turcotte-Pugh Score may all be used to identify and diagnose kidney function issues. Figure.3 depicts the medical diagnosis of kidney function test data and liver function test data as a set of parameters Pre-processing is the process of converting numerical data into nominal data, which is then used in the classification process. Bagging with C4.5 algorithm; Hybrid method- AdaBoost; Bagging with random forest; C4.5 with 10 fold-cross validation; Attribute selection with C45 algorithm; Attribute selection with C4.5 algorithm.

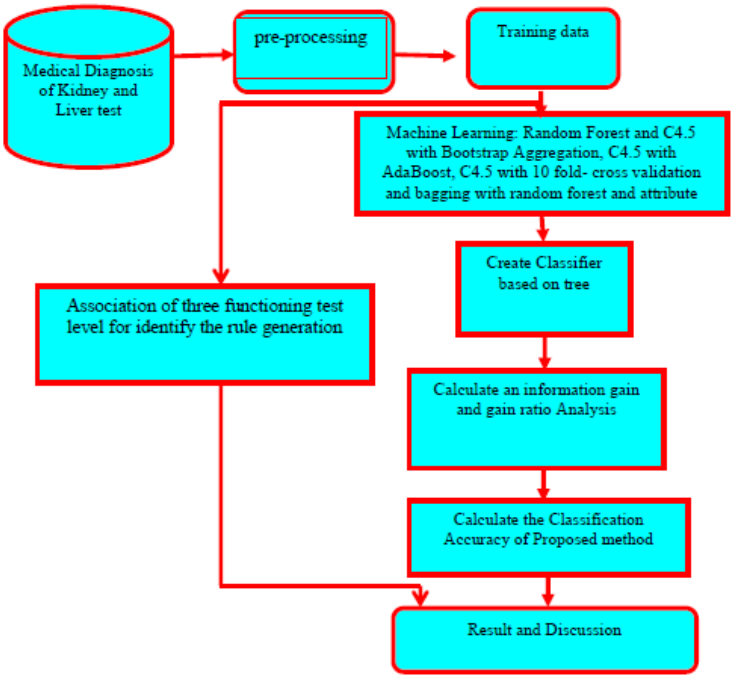

Figure 3. Frame work Architecture

Using an Apriori algorithm, generate a rule for the most important parameters of all functional test levels, including the heart pump test, in order to evaluate the classification process's correctness. With a support of 0.4 and a confidence of 0.5, the association rule is produced from frequently occurring item sets.

APRIORI Algorithm

To effectively count candidate item sets, Apriori makes advantage of breadth-first search. It builds k-item candidate sets from k-1-item sets. Then, it eliminates candidates who have a sub pattern that is seldom used. The candidate set comprises all common k-length item sets, according to the downward closure lemma. After that, it looks through the transaction database to see which of the candidates have the most common collections of items in their profile. An association rule generating process in the Apriori algorithm may be summarised as follows: one A little amount of assistance is provided in the first step in order to identify frequently satisfied item sets. Second: The requisite frequent item-sets and limitations are used to mine all association rules in the Second to determine the minimal level of confidence.

Apriori Algorithm Pseudo code procedure is give below:
Apriori (D, min_Sup) {//D: Database and min_Sup: minimum support F1= {frequent items};
for (i= 2; Fi-1 !=∅; i++)
{ Ck= candidates generated from Fk-1
//that is Cartesian product Fk-1 x Fk-1 and eliminating any F-1 size itemset that is not
frequent
for each transaction D in database do{
#increment the count of all candidates in Ck that are contained in D
Fk= candidates in Ck with min Support
}//end for each }//end for
returnUkFk
; }

As seen in the following pseudo code, Apriori's algorithm scans the database on a continuous basis. To effectively count candidate item sets, it makes use of a tree structure and a breadth-first search [16]. It produces candidate sets of items of length I from sets of items of length k 1 using this algorithm. Then, it eliminates candidates who have a sub pattern that is seldom used. Finally, every conceivable combination of frequently occurring item sets is identified until no candidate item is generated. For frequent item sets to be complete, it is necessary to have a downward closure property on the support. Prior candidate item sets are not filtered out by Apriori; rather, it aids in decreasing the number of candidate item sets that must be scanned. As a result, scanning a database takes longer[14]. It is not totally efficient to implement a procedure. An item set's support supp(X) is defined as the percentage of transactions in the data set that include the item set, as shown by the equation Supp (X) = (Number of transactions that contain the item set X)/. (total no. of transactions). It is defined as an estimate of the probability P(Y | X), the likelihood of finding the rule's right-hand side in transactions if they also include the rule's left-hand side.

Results and Discussion

The Apriori method may be used to generate the rules for the experiment analysis, and 500 training data can be used to generate 230 rule generations. In Figure.4, four candidate generation item-sets may provide 230 rule generations with a given support of 0.4. Figure 5 shows the rule generation as a percentage of confidence values ranging from 50% to 100%.

```
Number of cycles performed: 12

Generated sets of large itemsets:

Size of set of large itemsets L(1): 9

Large Itemsets L(1):
C=Normal 342
U=Normal 423
TP=Normal 423
UA=Normal 367
CAL=Low 411
BIC=Low 411
Sod=Normal 329
SC=Normal 309
Liver FunctionTest=Class C 408
```

```
Size of set of large itemsets L(2): 16

Large Itemsets L(2):
C=Normal U=Normal 342
C=Normal TP=Normal 342
C=Normal UA=Normal 302
U=Normal TP=Normal 423
U=Normal UA=Normal 364
U=Normal CAL=Low 334
U=Normal BIC=Low 334
U=Normal Liver FunctionTest=Class C 333
TP=Normal UA=Normal 364
TP=Normal CAL=Low 334
TP=Normal BIC=Low 334
TP=Normal Liver FunctionTest=Class C 333
CAL=Low BIC=Low 411
CAL=Low Liver FunctionTest=Class C 384
BIC=Low Liver FunctionTest=Class C 384
Sod=Normal SC=Normal 300
```

```
Size of set of large itemsets L(3): 14

Large Itemsets L(3):
C=Normal U=Normal TP=Normal 342
C=Normal U=Normal UA=Normal 302
C=Normal TP=Normal UA=Normal 302
U=Normal TP=Normal UA=Normal 364
U=Normal TP=Normal CAL=Low 334
U=Normal TP=Normal BIC=Low 334
U=Normal TP=Normal Liver FunctionTest=Class C 333
U=Normal CAL=Low BIC=Low 334
U=Normal CAL=Low Liver FunctionTest=Class C 309
U=Normal BIC=Low Liver FunctionTest=Class C 309
TP=Normal CAL=Low BIC=Low 334
TP=Normal CAL=Low Liver FunctionTest=Class C 309
TP=Normal BIC=Low Liver FunctionTest=Class C 309
CAL=Low BIC=Low Liver FunctionTest=Class C 384

Size of set of large itemsets L(4): 6

Large Itemsets L(4):
C=Normal U=Normal TP=Normal UA=Normal 302
U=Normal TP=Normal CAL=Low BIC=Low 334
U=Normal TP=Normal CAL=Low Liver FunctionTest=Class C 309
U=Normal TP=Normal BIC=Low Liver FunctionTest=Class C 309
U=Normal CAL=Low BIC=Low Liver FunctionTest=Class C 309
TP=Normal CAL=Low BIC=Low Liver FunctionTest=Class C 309
```

Figure 4. Candidate generation of item

This is an experimental 500 training data set that may be used to analyse how patients' functional test levels vary over time by varying support counts, as illustrated in Figure 5.

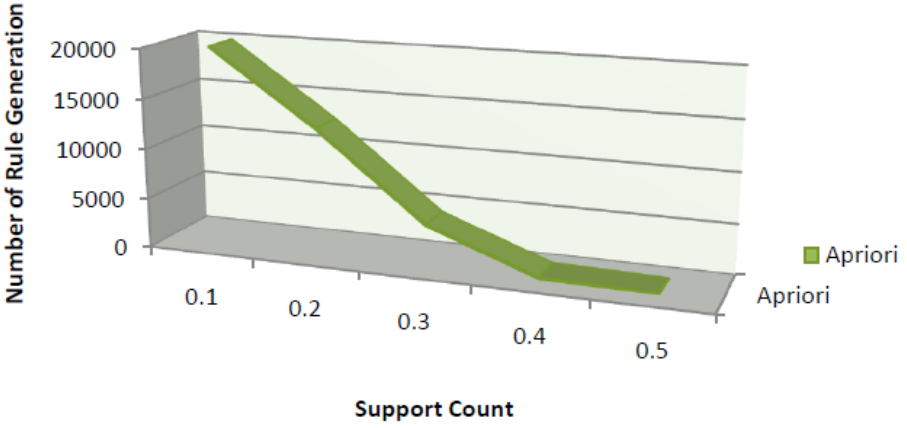

Figure 5. Rule generation found by different support count

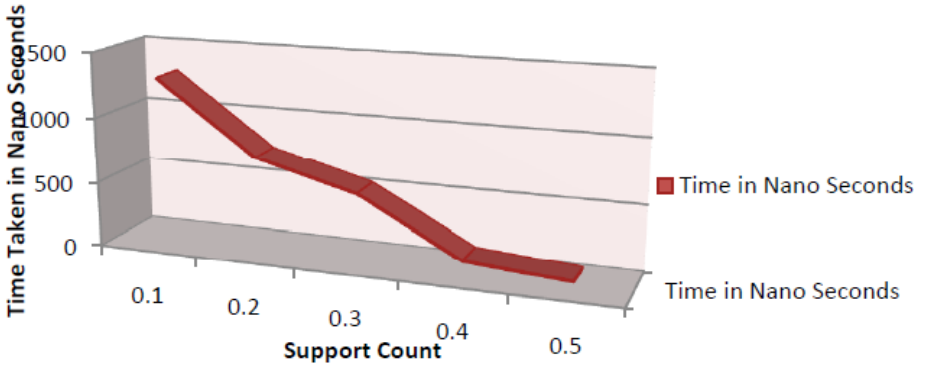

Figure 6. Time Taken in Nano Second

It is possible to experiment with the run duration and accuracy of rule generation in MATLAB, as shown in Figure 6. A 120 nanosecond execution time is required for a 0.4 support count.

Conclusion

In this research work, an attempt is made to review the basic concepts of Data Mining tasks, Data Mining methods, Clustering Techniques, Issues in Data Mining, Research Challenges in Data Mining and Applications, Recent research achievements and Data Mining tools. The main purpose of the research work is to analyse the human health condition that can be realised by fluctuation of specific ranges such as Bilirubin, Albumin, Prothrombin time (INR), Ascites, Encephalopathy, Bicarbonate, Calcium etc. It can improve the disease by taking medical diagnosis based on Apriori algorithm by generating rule for the most significant parameters of all three functioning test level. The experiment can be carried out with a variety of support count sizes in order to gain a better comprehension of the effect that the Apriori algorithm has on the amount of time required for its execution, the precision of the best rule discovered from the mining of frequent patterns, and the quantity of association rules that are produced.

References

1. Anu Chaudhary, Puneet Garg, Detecting and Diagnosing a Disease by Patient Monitoring System, International Journal of Mechanical Engineering And Information Technology, Vol. 2 Issue 6 June Page No: 493-499, 2014.
2. Ashfaq Ahmed K, Sultan Aljahdali and Syed Naimatullah Hussain,(2013) “Comparative Prediction Performance with Support Vector Machine and Random Forest Classification Techniques”, International Journal of Computer Applications Volume 69– No.11, May page no 12-16, 2013.
3. Basma Boukenze, et.al,” Performance of Data Mining Techniques to Predict in Healthcare Case Study: Chronic Kidney Failure Disease”, International Journal of Database Management Systems (IJDMS) Vol.8, No.3, pp: 1 to 4, 2016.

4. Endo, A, Shibata, T and Tanaka, H (2008) “Comparison of Seven Algorithms to Predict Breast Cancer Survival”, Biomedical Soft Computing and Human Sciences, 13(2), pp.11-16, 2008.
5. Dietterich, T. G., “An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting and randomization”. Machine learning, 40: 139-157, 2001.
6. Freund, Y. Schapire, R. (1996). “Experiments with a new boosting algorithm”, In Proceedings of the Thirteenth International Conference on Machine Learning, 148-156, 1996.
7. Lakshmi. K.R, Nagesh. Y and VeeraKrishna. M, (2014) Performance Comparison of Three Data Mining Techniques For Predicting Kidney Dialysis Survivability, International Journal of Advances in Engineering & Technology, Mar., Vol. 7, Issue 1, pg no. 242-254, 2014.
8. McLernon DJ, Donnan PT, Sullivan FM*, et a, “* Prediction of liver disease inpatients whose liver function tests have been checked in primary care: model development and validation using population-based observational cohorts”, *BM J*;4:e004837. doi:10.1136/bmjopen- 2014-004837, 2014.
9. Nazmun Nahar and Ferdous Ara, “Liver Disease Prediction by Using Different Decision Tree Techniques”, International Journal of Data Mining & Knowledge Management Process, Vol.8, No.2, PP-1-9, DOI: 10.5121/ijdkp.2018.8201, 2018.
10. J. Pradeep Kandhasamy, S. Balamurali, “Performance Analysis of Classifier Models to Predict Diabetes Mellitus”, Procedia Computer Science Issue 47 pp(45 – 51), doi: 10.1016/j.procs.2015.03.182, 2015.
11. Sajidaperveenaet. Al, “Performance Analysis of Data Mining Classification Techniques to Predict Diabetes”, Procedia Computer Science Elsevier, 82 (2016) 115 – 121.
12. K.Swapna and Prof. M.S. Prasad Babu,, “A Critical Study on Cluster Analysis Methods to Extract Liver Disease Patterns in Indian Liver Patient Data”, International Journal of Computational Intelligence Research, Volume 13, Number 10, pp. 2379- 2390, ISSN 0973-1873, 2017.
13. Tapas Ranjan Baitharua , Subhendu Kumar Panib, ”Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset”, Procedia Computer Science volume- 85, 862 – 870, doi: 10.1016/j.procs.2016.05.276, 2016.
14. J.Vijayalakshmi, Kidney Failure Due to Diabetics – “Detection using Classification Algorithm in Data Mining”, International Journal of Data Mining Techniques and Application, Volume: 06, Issue: 02, , Page No.62-64 ISSN: 2278-2419, 2017.
15. Dr. S. Vijayarani, Mr.S.Dhayanand, “Data Mining Classification Algorithms For Kidney Disease Prediction”, International Journal on Cybernetics & Informatics, Vol. 4, No. 4, pp: 13 to 25 DOI: 10.5121/ijci.2015.4402, 2015.
16. Pugh RN, Murray-Lyon IM, Dawson JL, et. al. Transection of the oesophagus for bleeding oesophageal varices. *Br J Surg.*; 60:646, 1973.