K-MEANS WITH APRIORI & ID3 FOR HEART STROKE RISK LEVEL PREDICTION

[1]Fetenech Meskele, [2]Dr.Mohd Abdul Hameed, [3] Dr.O.Jamsheela
[1]Research Scholar, [2]Assistant Professor, [3]Assistant Professor
[1]Dept. of Computer Sciences and Engineering
[1]University College of Engineering,
Osmania University, Hyderabad, INDIA

Abstract :The interpretation of problem is a significant and tedious task in medicine. The detection of heart problem from various factors or symptoms is an issue which is not free from false presumptions often accompanied by unpredictable effects. Thus the effort to utilize knowledge and experience of numerous specialists and clinical screening data of patients collected earlier to facilitate the interpretation process is considered a valuable option. This paper is presented as an efficient approach for the prediction of heart stroke risk levels from the heart patients' dataset (Cleveland dataset) by using Machine Learning techniques. The earlier researchers have used k-means mafia algorithm and the accuracy of the algorithm was 74%. Implementing the same algorithm with a modification of k-means with Apriori and ID3 algorithm achieved 85% of the accuracy, which is a growth of 15% in accuracy.

IndexTerms **- Machine Learning, Heart Problem, K-means, Apriori Algorithm, ID3 Algorithm, heart patients' risk level prediction. Stroke.**

I. INTRODUCTION

Heart problem or Cardiovascular problem is a kind of serious health imperiling and frequent happening problem. The world health organization has estimated that 12 million deaths occur worldwide every year due to cardiovascular problem. Advances in the field of medicine over the past few decades enabled the identification of risk factors that may contribute to cardiovascular problem. The most common cause of heart problem is narrowing or blockage of the coronary arteries, the blood vessels that supply blood to the heart itself. This is called coronary artery problem and happens slowly over time. It is the major reason of heart strokes. So automation of the risk level prediction would be very useful. All doctors are unfortunately not equally skilled in every subspecialty and they are in many places a scarce resource. An automated system for diagnosing health problem would enhance medical care and also can reduce costs. A physician must be experienced and highly skilled to diagnose heart strokes of a patient. Thus the effort to utilize knowledge and experience of numerous specialists and clinical screening data of patients to facilitate the problem process is considered a value system that is the integration of clinical decision support with computer. The patient records which are already collected could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. The main objective of this research is to develop an Intelligent Heart Problem Prediction System by using K-means clustering, Apriori algorithm and Decision tree [1] methods .

II. RELATED WORK.

Different supervised machine learning algorithms i.e. Naïve Bayes, Neural Network, along with weighted association Apriori algorithm, Decision algorithm have been used for analyzing the dataset in [2]. The Machine Learning tool Weka 3.6.6 is used for the experiment. Weka is a collection of Machine learning algorithms for Machine Learning tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, Classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Some of the classification algorithms in Machine Learning are: Chaurasia and Pal conducted a study on the prediction of heart stroke risk levels from the heart patients' data[16][17]18]. The prediction of heart problem significantly uses 11 important attributes, with basic Machine Learning technique like Naïve Bayes, J48 decision tree and Bagging approaches. The outcome shows that bagging techniques performance is more accurate than Bayesian classification and J48. The results show that the bagging prediction system is capable of predicting the heart stroke effectively [3]. A simple probabilistic, the naive Bayes classifier is used for classification earlier on Bayes' theorem. According to naïve Bayesian classifier the occurrence (or nonoccurrence of a particular feature of a class is considered independent to the presence (or absence) of any other feature. When the dimension of the input is high and more efficient result is expected, the chief Naïve Bayes Classifier technique [4,5,6,] is applicable. Naive Bayes model identifies the physical characteristics and features of patients suffering from heart problem. For each input it gives the possibility of an attribute for the expectable state.

III. K-MEANS CLUSTERING.

K-means is the simplest learning algorithm to solve the clustering problems[19]. The process is simple and easy, it classifies given data set into a certain number of clusters. It defines k centroids for each cluster. They must be placed as much as possible far away from each other. Then take each point belonging to a given data set and relate into the nearest centroid. If no point is pending then a group age is done. Then we re-calculate k new centroid for the cluster resulting from previous steps. When we get the k

centroid, a new binding is to be done between same data points and nearest centroid. A loop is being generated because of this loop key centroid change the location step by step until no more changes are done[4]. The advantages of k means clustering algorithms are simplicity and speed. This paper proposed the most popular distance measure, Euclidean distance, which is defined as

$$\text{dist}(i, j) = \| x_i - x_j \|$$

$$= \sqrt{(x_{i1}-x_{j1})^2 + (x_{i2}-x_{j2})^2 + \ldots (x_{ip}-x_{jp})^2}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p_dimensional data objects.

The steps involved in a K-means algorithm:

1. *x* points denoting the data to be clustered are placed into space. These points denote the primary group centroids.
2. The data are assigned to the group that is adjacent to the centroid.
3. The positions of all the *j x* centroids are recalculated as soon as all the data are assigned.

Steps 2 and 3 are reiterated until the centroids stop moving any further. This results in the separation of data into groups from which the metric to be minimized can be deliberated. The advantages of k-means clustering algorithm are its simplicity and speed.

In this system, we mainly used clustering for grouping the attributes. As we take almost 20 attributes such as age, obesity, gender, cholesterol, smoking, blood pressure, chest pain, blood sugar, Alcoholic intake etc. this attributes are grouped using K-Means clustering algorithm For example if we take an attribute such as age and considered the age of the person is between 0-100. After applying K-means algorithm on this dataset of age it will find the centroid and divide it into groups. It calculates the means. Here, age will be divided into 3 groups such as:-

0-30,31-60,61-100.
It will give values such as
0-30=0, 31-60=1, 61-100=2

For gender attribute, it will divide into groups such as Male=1 and Female=0.

K-means will be applied to each and every attribute. After that, the attributes and their values will be added in a dataset accordingly. Then the model is being ready for prediction.

IV. APRIORI ALGORITHM TO FIND THE FREQUENT PATTERN ON HEART STROKE DATASET.

Apriori algorithm is the most classical and important algorithm for mining frequent itemsets proposed by R.Agrawal and R.Srikant in 1994[20]. Apriori is an efficient algorithm to find frequent item sets from large datasets. The key idea of Apriori algorithm is to make multiple passes over the dataset. It applies an iterative breadth-first search (level-wise search) through the dataset, where frequent k-item sets are used to explore frequent (k+1)-item sets. The working of Apriori algorithm is based on the Apriori property which states that" All nonempty subsets of a frequent item set must be frequent". It also applies the anti-monotone property which proves that if an itemset cannot pass the minimum support test, all its supersets also will fail to pass the test. Therefore if any itemset is infrequent then all its supersets are also infrequent and vice versa. This property is used in Apriori to prune the infrequent candidate elements. As the first step, the set of frequent 1-itemsets are collected. The set of frequent 1-itemsets contains item names with support count, which satisfies the support threshold and is denoted by L. Each subsequent n^{th} pass starts with the set of itemsets which is collected in the previous $n\text{-}1^{th}$ pass and used to find larger frequent itemsets. At the end of each pass k, a set of frequent k-itemsets are collected and they become the inputs for the next pass k+1. Therefore, L is used to find L1, the set of frequent 2-item-sets, which is used to find L2, and so on, until no more frequent k- item sets can be found.

V. ITERATIVE DICHOTOMIZE 3 (ID3)

Itemized Dichotomized 3 algorithm or better known as ID3 algorithm [7] was first introduced by J.R Quinlan in the late 1970s.The concept of information theory is applied in the field of Machine Learning. As in the algorithms of Machine Learning, the classification is an essential step, using an information theoretic measure in ID3 algorithm, one of the key algorithms of decision tree algorithms, they have discussed the different steps of the development of a decision tree so that the best classification criteria can be developed which is helpful in making good decisions. From the data under consideration having a set of values, a property on the basis of calculation is selected as the root of the tree and the process is repeated to develop complete decision tree. ID3, Iterative Dichotomized 3 is a decision tree learning algorithm which is used for the classification of the objects with the iterative inductive approach. In this algorithm, the top to down approach is used. The top node is called the root node and others are the leaf nodes. Therefore, it's a traversing from the root node to leaf nodes. Each node requires some test on the attributes which decide the level of the leaf nodes. These decision trees are mostly used for the decision making purpose [8] [9]. Machine Learning techniques basically use the ID3 algorithm as it's the basic algorithm of classification. In the medical field, ID3 was mainly used for Machine Learning. The information gain, Gain (S, A) of an attribute A, relative to a collection of example S, is defined as,

Where pi is the proportion of S belonging to class i.

GAIN(S,A)=ENTROPY(S)-$\sum_{vEValues(A)}(\frac{Sv}{S})$ENTROPY(SV)…….(1)

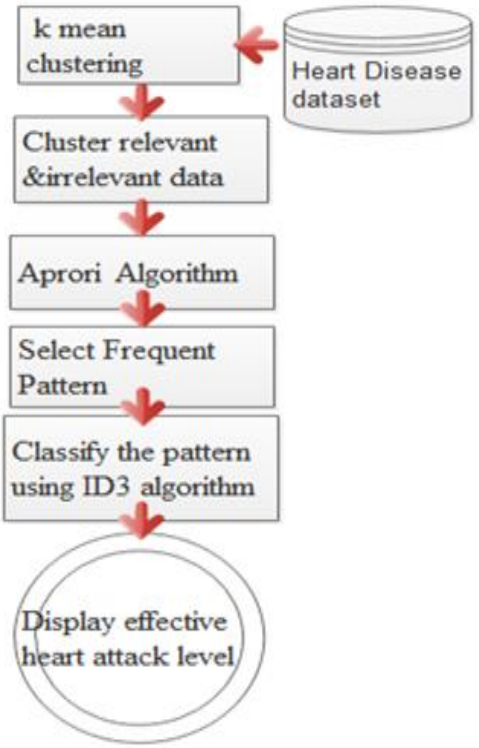
ENTROPY(S)=$\sum_{i=1}^{c} pilog2pi$ ……(2)

Where *pi* is the proportion of *S* belonging to class *i*.

VI. EXPERIMENTAL RESULTS.

The results of the experimental analysis of the new algorithm for finding significant patterns for heart stroke prediction are presented in this section. With the help of the dataset, the patterns significant to the heart stroke prediction are extracted using the approaches discussed above. The heart patients' dataset is preprocessed successfully by removing duplicate records and supplying missing values as shown in Table II. The final dataset, resultant from preprocessing, is then clustered using k-means algorithm with k -value as 2. Once the cluster consists of the data relevant to the heart problem are created then the frequent patterns are mined efficiently from the cluster using the Apriori Algorithm. The sample combinations of heart stroke parameters for normal and risk level along with their values and levels are shown in Table III and the other tables contains the remaining data. The sample combinations are shown in table such as if the value is less than or equal to (0.4) the system will predict it as normal level. If the value is greater than or equal to (0.4), then it shows that it is not normal and is in risk level. And if the value is (0.8) or (0.9) then it shows that it is in the higher risk level. Then the frequent patterns are mined efficiently from the cluster using the Apriori algorithm. Finally the example of training data to predict the heart stroke level and then the efficient heart stroke level with tree using the ID3 by information gain and performance measure measurement of accuracy, precision and recall are shown by using graphs.

VII. SYSTEM ARCHITECTURE



VIII. PERFORMANCE MEASURES

The performance measures like RECALL (SENSITIVITY), SPECIFICITY and F-measure are also used for calculating other aggregated performance measures. The goal is to have high accuracy, besides high precision and recall metrics. These metrics can be derived from the confusion matrix and can be easily converted to true-positive (TP) and false-positive (FP) metrics.

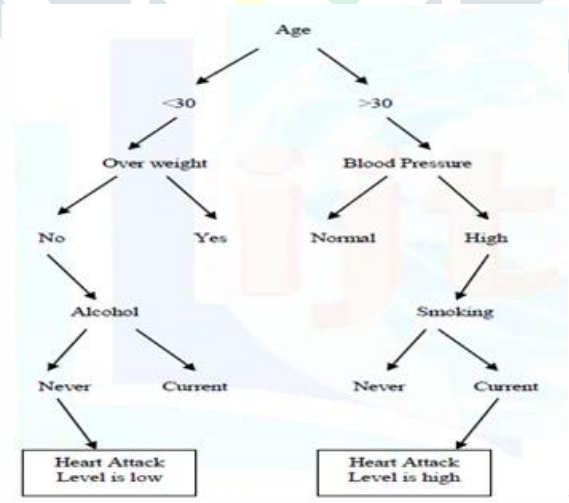
Precision =$\frac{TP}{TP+FP}$, recall= Recall =$\frac{TP}{TP+FN}$ and Accuracy = TP+TN/TP+FP+FN+TN

1. True Positive (TP): Total percentage of members classified as Class A belongs to Class A.
2. False Positive (FP): Total percentage of members of Class A but does not belong to Class A.
3. False Negative (FN): Total percentage of members of Class A incorrectly classified as not belonging to Class A.
4. True Negative (TN): Total percentage of members who do not belong to Class A are classified not a part of Class A. It can also be given as (100% - FP).

Table 1 Heart Patients' Dataset

id	Datasets
1	Age
2	Sex(value 1:Male; value 0: Female)
3	Slope: the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: downsloping).
4	famhist : family history of coronary artery problem (value 1: yes; value 0: no)
5	Fasting Blood Sugar (value 1: >120 mg/dl; value 0:
6	painloc: chest pain location (value 1:substernal; value 0: otherwise)
7	Thal (value 1: normal; value 2: fixed defect; value 3: reversible defect)
8	chol: serum cholesterol
9	trestbps: resting blood pressure
10	Exang: exercise induced angina (value 1:yes; value 0: no)

Fig-2 A decision tree for predicting heart stroke level



If

Age=<30 and Overweight=no and Alcohol Intake=never
Then
Heart stroke level is Low

(Or)

If

Age=>70 and Blood pressure=High and
Smoking=current
Then
Heart stroke level is High

Table-2 Clustered data of Heart Patients'' dataset

Id	Reference id	Attribute
1	#2	Age
2	#3	Sex
3	#5	chestpain
4	#13	trestbpd
5	#9	fbs
6	#8	smok
7	#11	Fms(familyhistory)
8	#10	Dm(History of diabetics)
9	#20	Alcohol intake

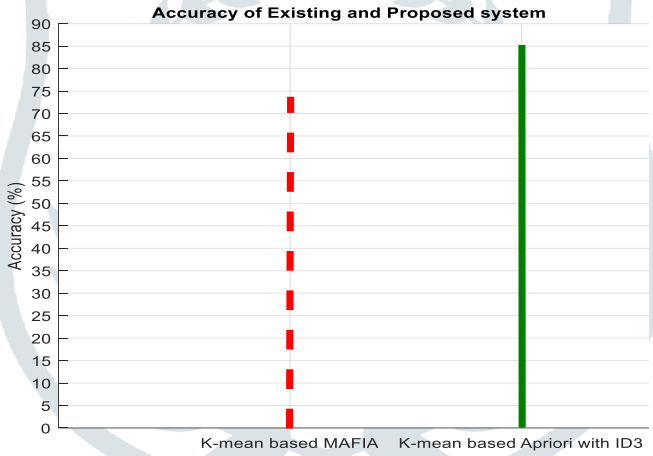
Table-3 heart stroke parameters with corresponding values and their weights

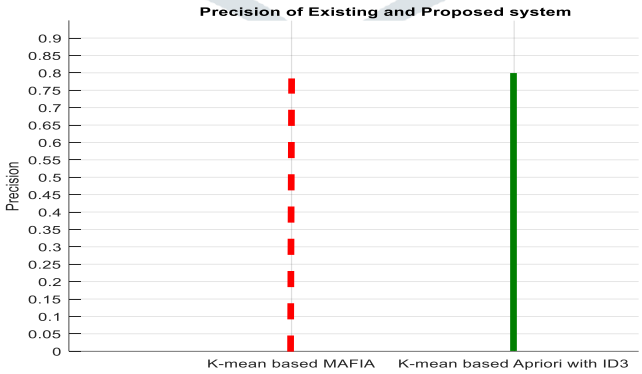
Parameter	Weight	Risk level
Male& Female	Age<30 Age>30	0.1 0.8
Smoking	Never Past Current	0.1 0.3 0.6
chol	High Normal	0.8 0.1
Overweight	Yes No	0.8 0.1
Alcohol Intake	Never Past Current	0.1 0.3 0.6
Sedentary Lifestyle/ inactivity	Yes No	0.7 0.1
Family history	Yes No	0.7 0.1
Blood Pressure	Normal (130/89) Low(< 119/79) High(>200/160)	0.1 0.8 0.9

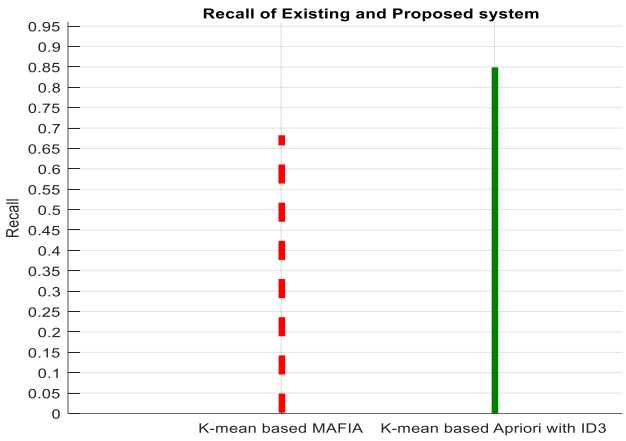
High Salt Diet	Yes No	0.9 0.1

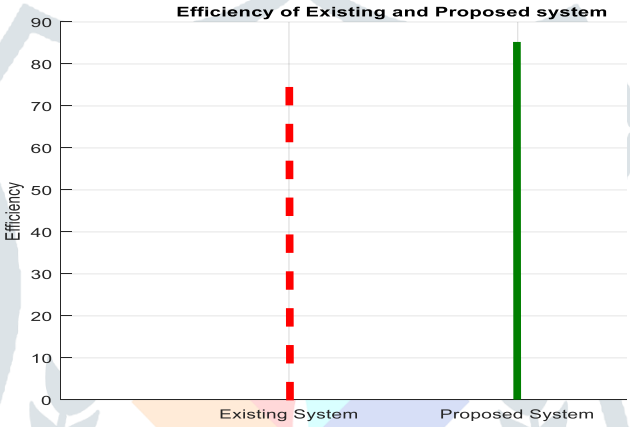
Table-4 Comparison between simple mafia with k-means earlier apriori & ID3 Algorithm

Technique	Precision	Recall	Accuracy%
K-means earlier MAFIA	0.78	0.67	74%
k-means earlier Apriori with ID3 (proposed algorithm)	0.80	0.85	85%









IX. CONCLUSION AND FUTURE WORK

The early prediction of heart problem is very important because heart problem is one of the leading causes of death in worldwide. The computer-aided heart problem prediction system helps the physician as a tool for detecting heart problems earlier and very easily. In this paper a Heart Problem Detection System using Machine Learning techniques is presented. Apriori algorithm is used to find the frequent item sets and the maximal frequent item set is generated. Clustering is performed using a k-means clustering algorithm. Lastly, the ID3 algorism is applied to show the classification. Defining the clusters earlier on maximal frequent item sets improves the accuracy of the algorithm. As a future work, we have planned to design and develop an efficient heart stroke prediction system with the aid of python programming language on an Oracle Dataset.

REFERENCES

[1] Awang, "Intelligent Heart Problem Prediction System Using Machine Learning Techniques" IEEE Conference, 2008, pp 108-115.

[2] Mark Hall; Eibe Frank; Georey Holmes; Bernhard Pfahringer; Peter Reutemann; Ian H. The weka Machine Learning software: An update. SIGKDD Explorations, 11, 2009.

[3] Shanta Kumar, B.Patil, Y.S.Kumaraswamy, "Predictive Machine Learning for medical problem of heart problem prediction"IJCSE Vol .17, 2011

[4] Liu X, Lu R, Ma J, Chen L. Privacy-preserving patient-centric clinical decision support system on naïve Bayesian classification. IEEE Journal of Biomedical and Health Informatics. 2016; 20(2):655–88.

[5] Patil RR. Heart problem prediction system using naïve Bayes and Jelinek-mercer smoothing. International Journal of Advanced Research in Computer Science and Communication Engineering. 2014; 3(5):6787–9.

[6] Pattekari SA, Parveen A. Prediction system for heart problem using naive Bayes. International Journal of Advanced Computer and Mathematical Sciences. 2012; 3(3):290–4.

[7] Komal G, Vekariya V, "Novel approach for heart problem prediction using a decision tree algorithm", International Journal of Innovative Research in Computer and Communication Engineering, 3(11):11544–1, 2015.

[8] Anand Bahety, "Extension and Evaluation of ID3 – Decision Tree Algorithm". University of Maryland, College Park.

[9] S. K. Yadav and Pal S., "Machine Learning: A Prediction for Performance Improvement of Engineering Students using Classification", World of Computer Science and Information Technology (WCSIT), 2(2), 51-56, 2012.

[10] Quinlan J. C4.5: programs for machine learning. San Mateo, CA: Morgan Kaufmann; 1993.

[11] Fayyad, P.-S. S. (1996). Advances in Knowledge Discovery and Machine Learning. AAAI Press / The MIT Press, 1-34.

[12] Srinivas, K., "Analysis of coronary heart problem and prediction of a heart stroke in coal mining regions using Machine Learning techniques", IEEE Transaction on Computer Science and Education (ICCSE), p(1344 - 1349), 2010.

[13] M. Anbarasi et. al. "Enhanced Prediction of Heart Problem with Feature Subset SelectionusingGeneticAlgorithm", International Journal of Engineering Science and Technology Vol. 2(10), 5370-5376, 2010.

[14] Shanta Kumar.Patil, Y.S.Kumaraswamy, _Intelligent and Effective Heart Stroke Prediction System Using Machine Learning and Artificial Neural Network_; European Journal of Scientific Research, ISSN 1450-216X Vol.31 No.4, 2009.

[15] Majali J, Niranjan R, Phatak V, Tanaka O. Machine Learning techniques for problem and prognosis of cancer. Inter¬national Journal of Advanced Research in Computer and Communication Engineering. 2015; 4(3):613–6.

[16] Chaurasia, V. and Pal, S., 2013. Early prediction of heart diseases using data mining techniques.

[17] Chaurasia, V. and Pal, S., 2014. Data mining approach to detect heart diseases. International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol, 2, pp.56-66.

[18] Chaurasia, V. and Pal, S., 2017. Data mining techniques: to predict and resolve breast cancer survivability.

[19] Hartigan, J.A. and Wong, M.A., 1979. Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), pp.100-108.

[20] Agrawal, R. and Srikant, R., 1994, September. Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).