LITERATURE SURVEY

1. Bellaachia and Guven (2005) proposed predicting breast cancer lastingness using data mining method.

The authors have examined three data mining methods such as Naïve Bayes, propagated neural networks, and c4.5 decision tree algorithms. Naïve Bayes method is the first method that uses the Bayesian method, because of its simple, clear, and fast predictive nature. The second method is artificial neural networks (ANNs) that uses multilayer network with transmission utilisation. Finally, they used c4.5 decisiontree algorithms. On the whole, the authors' work shows that the preliminary results are challenging prediction problem in medical data sets.

Data mining is the apt technology to predict patterns in the health sector data set. Though it is tedious to make the prediction of few diseases such as heart attack, due to its complexity, such tasks need more skill. Masethe and Masethe (2014) discussed to determine heart disease using classification algorithms. Few data mining algorithms such as j48, Naïve Bayes, REPTREE, and classification and regression trees (CART) are applied to predict heart attacks.

The author's research work result shows that prediction accuracy is 99%, and j48, REPTREE, and CART gave a prediction model of 89 cases with a risk factor positive for heart attacks. From these techniques, it was identified that prediction of diagnoses can be done by data-mining algorithms.

2. A medical data of large size need powerful data analysis tools for processing

Data mining techniques can also be used for the diagnosis and predictive analysis. Ramaraj and Thanamani (2013) proposed predictive analytics methods to identify heart diseases. The authors' aim was to design a predictive method for heart disease detection.

Classification accuracy report among various data mining techniques with the difference in error rates is provided in analysis part. The authors' final result shows that CN2Rule performs classification more accurately than the other methods.

Nasridinov et al. (2014) discussed a study on crime pattern prediction using data mining techniques. The authors analysed many data mining techniques with generated test data to determine the best method to perform crime pattern prediction task.

Specifically, the authors did an extensive performance analysis of various data mining prediction algorithms such as support vector machine (SVM), decision tree, neural network, k-nearest neighbour, and Naïve Bayes.

The authors assumed that wearable sensor devices are attached to the clothes of the user of the proposed method. It captures the inner temperature and heartbeat of a user and sends these data to the server to perform emotion mining.

3. Chandra Shekar et al. (2012) make up a better algorithm for prediction of heart disease using case-based machine learning-based methods technique on non-binary data sets.

Mining frequent item-sets in non-binary search space presented fascinating challenges over conventional mining in binary search space. Initially, the non-binary search space needs innovative tactics to calculate support and must be active.

As there is a chance of removal of candidate item-set from the non-binary data set due to pruning, applying it at a higher level may become frequent. Support calculation and candidate generation at each level are carried out using separate mechanism. The author's final result was a prototype for generating frequent item sets for non-binary data set that was developed

Mining frequent item-sets in non-binary search space presented fascinating challenges over conventional mining in binary search space. Initially, the non-binary search space needs innovative tactics to calculate support and must be active. As there is a chance of removal of candidate item-set from the non-binary data set due to pruning, applying it at a higher level may become frequent.

Support calculation and candidate generation at each level are carried out using separate mechanism. The author's final result was a

prototype for generating frequent item sets for non-binary data set that was developed.

4.Kone and Karwan (2011) predicted the expense incurred in delivering bulk (liquefied) gas to new customers making use of a multifactor linear regression model.

Development of a single model, i.e. evaluating all the observations one time, leads to poor prediction outcomes. Hence, before regression analysis, a novel supervised learning method is utilised for grouping the customers who have similarity in some or the other perception.

Hyperboxes are used to denote classes on customers, and subsequently, a linear regression model is developed within every class. To increase with the combination of data classification and regression, the accuracy of the prediction is indicated.

Bhat et al. (2011) presented a new preprocessing phase along with imputation of missing value for numerical and also categorical data.

A hybrid combination consisting of classification and regression trees (CART), genetic algorithms for imputing the missing sequential values and self-organising feature maps (SOFM) for imputing the categorical values are used in the work.

A linear regression model is developed within every class. To increase with the combination of data classification and regression, the accuracy of the prediction is indicated.

5.Various data mining methodologies were used for predicting the heart disease by Soni et al. (2011).

The accuracy of those algorithms are verified, in which the accuracies of Naïve Bayes, ANN and decision tree are said to have accomplished a respective 86.53, 85.53, and 89%.

The data mining algorithms such as ANN, decision trees, and C4.5 apply ECG signals to analyse the heart disease. Decision tree algorithm is found to be the best and obtains 97.5% accuracy. The C4.5 algorithm yields an accuracy of 99.20%, whereas Naïve Bayes algorithm produces 89.60% of accuracy (Aneeshkumar and Venkateswaran, 2012). Therefore, these algorithms are employed for estimating the supervision over liver disorder.

C5.0 is a classification algorithm that is applied on huge data sets. It overcomes C4.5 in terms of the memory and speed along with the performance. This technique divides the sample depending on the field which provides the high information gain.

Later, the obtained sample subset received earlier will be divided. The action will persist till the sample subset cannot be further divided. At last, the lowest level split in the sample subsets that have less than acceptable level contribution for the model will be eliminated. C5.0 methodology easily deals with the missing attribute and the multivalued attribute from data set (Patil et al., 2012).