An Efficient Sequential Clustering Based Classification Model for Diabetes Diagnosis and Prediction

K. Veerasekaran, P. Sudhakar

Abstract*: At last decade, the development of diverse models and the excessive data creation leads to an enormous production of dataset and source. The healthcare field offers rich in information and it needs to be analyzed to identify the patterns present in the data. The commonly available massive amount of healthcare data characterizes a rich data field. The way of extracting the medical design is difficult because of the characteristics of healthcare data like massive, real, and complicated details. Various machine learning (ML) algorithms has developed to predict the existence of the diabetes disease. Due to the massive quantity of diabetes disease dataset, clustering techniques can be applied to group the data before classifying it. A new automated clustering based classification model is applied for the identification of diabetes. To cluster the healthcare data, sequential clustering (SC) model is applied. Then, logistic regression (LR) model is applied for the effective categorization of the clustered data. The experimentations have been directed by the benchmark dataset. The simulation outcomes demonstrate that the efficiency of the SC-LR method beats the prevailing methods to predict the diabetes diseases.

Keywords*: Classification, disease diagnosis, diabetes, machine learning.

I. INTRODUCTION

At the previous days, developing of data is seen in all fields and almost under classes. Healthcare data is drastically improved in the earlier days because of the intense advancement of data in the therapeutic field. The volumes of data made by healthcare field contacts are extremely troublesome and enormous to be process and dissect by traditional methodologies [1]. This data conceals important information often. Research individuals in the healthcare field face an issue to discover huge data from this colossal volume of the asset. Along these lines, medicinal data is a rapidly rising strategy which fears to utilize software engineering to wellbeing data.

Social insurance data is the space of data science worried with the assessment, use and dissemination of therapeutic data utilize PCs to a few perspectives on medical field [2]. Wellbeing data is depicted as "each part of acknowledging and boosting the productive affiliation, examination, organization, and using data in medicinal field". It incorporates the use of data for finding and dealing with the new data connecting to wellbeing and disease. A reasonable PC based framework and viable precise methodology can support to take in critical concealed data from huge therapeutic dataset [3]. At present, data mining (DM) is turned out to be common in healthcare field.

DM offers advancements and approaches to change over enormous data into advantageous data for making decisions. DM is depicted as "a strategy for nontrivial mining of interpretation, prior unidentified and conceivably important data from the data put away in a database" [4]. A more extensive procedure is the center advance that is known as data disclosure in the dataset. This system contains the motivation behind some past preparing approaches unsurprising at streamlining the use of the DM strategy and post handling methods indicated separating and refining the acknowledged data.

Building up the prescient methodologies from a few therapeutic data assets is likely using learning recognition or DM strategy based on unique ML models and the assessed exactness of the result savvy framework could even arrive at higher accuracy. This strategy is recognizing relationships or structures between different territories in enormous medicinal databases. DM is the medicinal procedures can be utilized to investigate and disclosure concealed structures private patients' database or therapeutic dataset. Propelled by broadly developing of malignant growth, hepatitis, coronary illness and diabetes patients in yearly and the availability of tremendous number of patients' data is used by researchers for helping proficient restorative consideration in the disease the board.

AI is utilized for preparing the structure over the colossal database, where utilized ML models can be used to concentrate plans or build a model and use to made plans or models for making forecast in the forthcoming in unequivocal cases. The dataset used to ponder the models is called as the preparation dataset. The record creating the preparation set is meant as preparing models and coolly named the example occurrences. The strategies are made helpfully utilizing preparing dataset. As the qualities or class marks of each individual preparing case offered, this stage is called as regulated learning [5].

Revised Manuscript Received on January 15, 2020

K. Veerasekaran, Assistant Professor, Department of Computer Science, Thiru. A. Govindaswamy Government Arts College, India. Email: veeradevihar@hotmail.com

P. Sudhakar, Associate Professor, Dept. of Computer Science and Engineering, Annamalai University, Chidambaram. Email: kar.sudha@gmail.com

The dataset utilized to evaluate the model quality is called as test dataset. It is utilized to evaluate the analytical data. The dataset used to assess the model quality is called as test dataset. It is used to assess the systematic accuracy of the procedures. The test is haphazardly designated and self-deciding of preparing the test. The result of prediction methods on a predefined test set is % of test occasions which are appropriately estimated by the procedures. At the point when the accuracy of the example is estimated satisfactory, the systems can be used to sort or gauge up and coming data record or items for that the class marks or qualities are inconclusive [6].

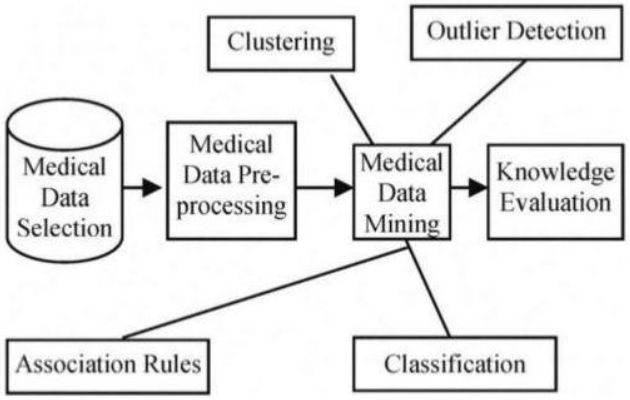

Fig. 1. Data mining process in healthcare data

Diabetes is a metabolic disease where a body is unable to control sugar classified by hyperglycaemia, for example higher blood glucose levels resultant to flaws while discharging insulin. The constant hyperglycaemia of diabetes is connected with clutters of starches, fats and protein digestion systems will prompts enduring damage, brokenness and different organ disappointments, particularly the nerves, kidneys, veins, eyes, and heart. The body needs insulin to use sugar, fats and proteins from the eating regimen for everyday exercises [7].

In the event of medicinal improvement, still, a considerable amount of diabetes mellitus are uninformed as different signs are indistinguishable from different diseases. In this way, it takes expanded interim for determination of such a diabetes disease [2]. Henceforth, there is a need to propose an expert framework that demonstrates the danger of disease and furthermore supports to find the arrangements utilizing information. The assessment paper we center around winning methodologies for diabetes recognition in order to know the ongoing advances in the field of diabetes under human services [8]. In the wake of examining every one of the strategies used to analyze diabetes disease, we make still some exertion in diagnosing the disease with high rightness. This procedure is performed on PIMA Indian dataset.

Various machine learning (ML) algorithms has developed to predict the existence of the diabetes disease. Due to the massive quantity of diabetes disease dataset, clustering techniques can be applied to group the data before classifying it. A new automated clustering based classification model is applied for the identification of diabetes. To cluster the healthcare data, sequential clustering (SC) model is applied. Then, logistic regression (LR) model is applied for the effective categorization of the clustered data. The experimentations have been directed by the benchmark dataset. The simulation outcomes demonstrate that the efficiency of the LR-SC method beats the prevailing methods to predict the diabetes diseases.

II. PROPOSED METHOD

The working process of the presented LR-SC is shown in Fig. 2. Initially, the data from the diabetes dataset is given as input to the prediction model. Then, data preprocessing takes place to remove the noise exist in the data. Next, the SC model gets executed to cluster the preprocessed data. Finally, the clustered data will undergo classification based on the existence or non-existence of diabetes dataset.

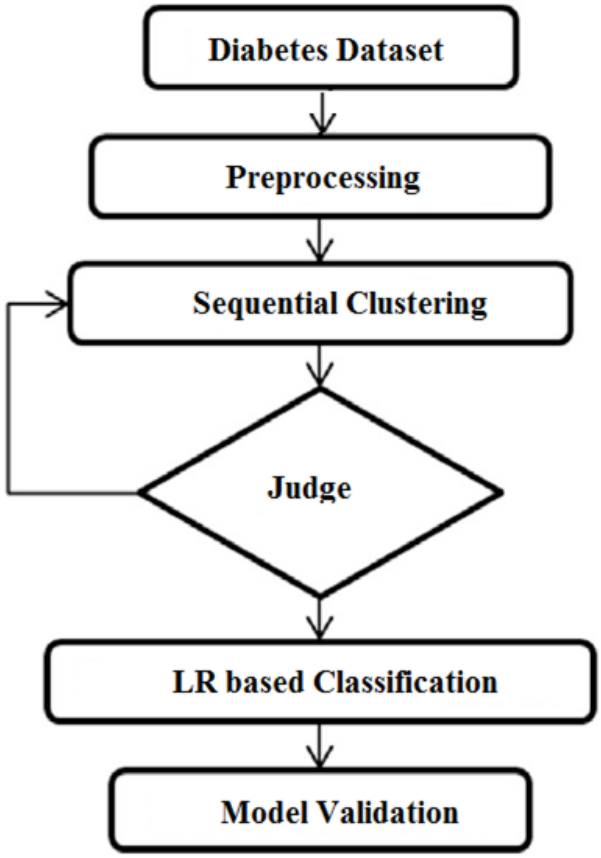

Fig. 2. Process involved in the LR-SC

A. Sequential clustering model

Let X indicates the set of objects and it is likely to identify a partition $T(X)$ that maximizes few score score function $F(T)$. A diverse score functions are considered and various models can be derived.

Here, the method begins with the division of objects X to singleton. In addition, in every step, a merging object is chosen in a greedy way for merging two clusters which performs score maximization. This greedy based merging step gets repeated until required cluster count is reached. Agglomerative clustering is mainly attractive in case the score function $\mathcal{F}$ can be decomposed.

Here, the modification in entire score through the merging of two clusters can be defined as

$$d_{\mathcal{F}}(t_i,t_j) = \mathcal{F}(\{t_i\} \cup \{t_j\}) - \mathcal{F}(\{t_i\}) - F(\{t_j\}) \quad (1)$$

A set of two major difficulties present in agglomerative clustering. Initially, the greedy method does not guarantee the identification of optimum partitions of X into a set of K clusters. Aurally, it is unsure to identify an optimum solution where every object belongs to a cluster which is almost identical to other. After that, an agglomeration process which holds the time complexity of $O(|X|^3|Y|)$ (where $|Y|$ is the dimension of the representation of each x) and a storage utilization of $O(|X|^2)$ that makes it not feasible for massive dataset. These issues can be resolved by the introduction of an efficient SC technique. Contrast to agglomerative clustering, this process manages a division which contains exact K clusters. It begins from a arbitrary initial partition $T = \{t_1, t_2, \ldots, t_K\}$ of X.

This SC model is significant of the default K-means technique. A major variation is that K-means execute parallel updates. Initially, a point is chosen for every new cluster, and then shits the components to the new clusters at one round. Consequently, the representation of clusters gets modified when every component moves to their recommended clusters. The algorithm of the SC method is shown in Fig. 3.

Input:
$|X|$ objects to be clustered
Parameters: K, n, $maxL$, ε

Output:
A partition T of X into K clusters

Main Loop:
For $i = 1, \ldots, n$
  $T_i \leftarrow$ random partition of X.
  $c \leftarrow 0$, $C \leftarrow 0$, $done = FALSE$
  While not $done$
    For $j = 1, \ldots, |X|$
      draw x_j out of $t(x_j)$
      $t^{new}(x_j) = \arg\min_{t'} d_{\mathcal{F}}(\{x_j\}, t')$
      If $t^{new}(x_j) \neq t(x_j)$ then $c \leftarrow c + 1$
      Merge x_j into $t^{new}(x_j)$
    $C \leftarrow C + 1$
    if $C \geq maxL$ or $c \leq \varepsilon \cdot |X|$ then
      $done \leftarrow TRUE$
$T \leftarrow \arg\max_{T_i} \mathcal{F}(T_i)$

Fig. 3. Algorithm of SC model

B. LR model

The classifier model intends to build a model which performs mapping process of the data items to a predefined type depending upon the available data. It is utilized for the extraction of important data items from the model or predicting the nature of the data. Under various instances, the dependency parameter of LR model is binary-classification. It indicates that the LR is applied for solving the 2-category problem.

The major aim of this simulation is the prediction of the presence or absence of diabetes which is considered as a classical binary-classification issue.

In addition, the LR model is mainly employed in the data mining process for the identification of diseases. In this study, LR is applied for the classification problem depending upon the linear regression model as defined below.

$$P = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m \quad (2)$$

The classifier issue is highly related to the linear regression issue that predicts the continuous value. It manages the reliable sensitivity over the field of real numbers. The prediction values of the classifier issue lies in the value of 0 or 1. The outcome will be 1 when the value exceeds the threshold value; else it is fixed to zero.

The final parameter range of LR lies in the range of 0 to 1. LR is a regression model which minimizes the predictive value and restricts the prediction value to [0, 1]. Depending upon the linear regression, LR includes and additional layer of non-linearity. The characteristics are totaled in a linear way and then the prediction process takes place by the use of sigmoid function. The major formula used in the LR classifier is given in Eqs. (3-5):

$$\Pr(Y = +1|X) \sim \beta \cdot X \; and \; \Pr(Y = -1|X) = 1 - \Pr(Y = +1|X) \quad (3)$$

$$\downarrow \sigma(x) : \frac{1}{1 + e^{-x}} \in [0;\ 1] \quad (4)$$

$$\Pr(Y = +1|X) \sim \sigma(\beta \cdot X) and \; \Pr(Y = -1|X) = 1 - \Pr(Y = +1|X) \quad (5)$$

Here, a set of two classes namely positive and negative is applied. The variable Y represents the patient suffers from diabetes. Then, X is an independent parameter indicating a set of 8 attributes present in the raw dataset. Each dependent variable X is allocated to a coefficient value known as β indicating the weights. Once the data is investigation using the LR model, the dataset exhibited each parameter's weight value. Diverse weights indicate distinctive overlapping among X and Y. When the regression model is applied, it is effective to provide new data and performs the prediction of the positive or negative dat. Generally, LR process takes place at the end stage as shown in Fig. 4.

III. PERFORMANCE VALIDATION

A. Dataset

Pima Indian dataset [9] is used to build the diabetes forecast model. The motivation behind this database is to dissect the diabetic patient dependent on explicit examination estimates incorporated into the dataset. While picking an example from the enormous dataset, numerous impediments were acknowledged. For this situation, each patient is female especially < 21 years of Pima Indian.

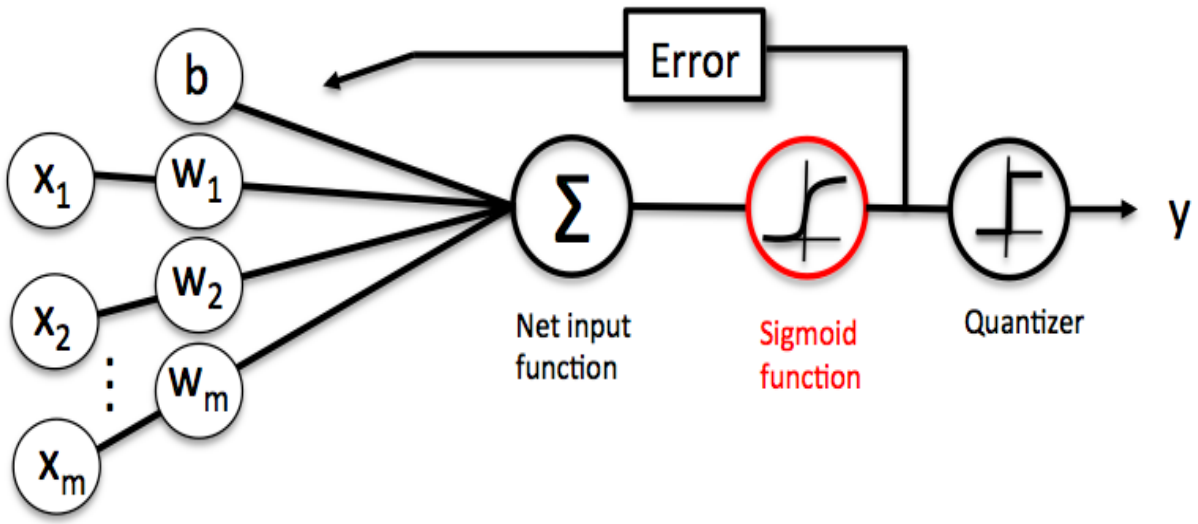

Fig. 4. LR process

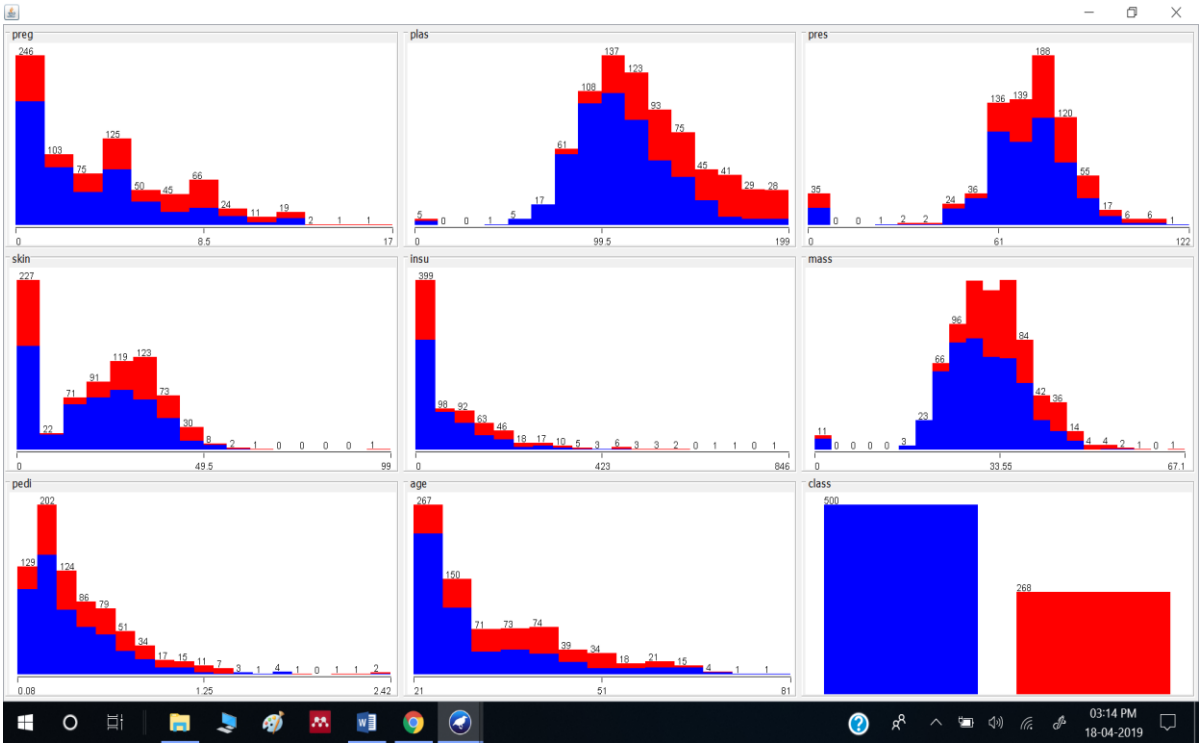

Fig. 5. Frequency Distribution of Pima Indian Diabetes Dataset for all Attributes

Table-I: Dataset Description

Description	Pima Indian Diabetes
Number of Instances	768
Number of Attributes	8
Number of Class	2
Percentage of Positive Samples	34.90%
Percentage of Negative Samples	65.10%

Table-II: Performance Evaluation of Different Classifiers with Proposed Method on Diabetes Dataset

Classifiers	Precision	Recall	Accuracy	F-score	Error Rate	MCC	Kappa
Logistic Regression	88.00	79.27	77.21	83.41	0.22	0.48	47.34
Voted Perceptron	92.40	68.04	66.79	78.37	0.33	0.17	13.52

Table-III: Confusion Matrix Before vs After sIB + Proposed

	LR-SC		LR	
Experts	**Negative**	**Positive**	**Negative**	**Positive**
Negative	213	32	440	115
Positive	51	106	60	153
Total	264	138	500	268
Total Instances	**402**		**768**	

The dataset incorporates numerous therapeutic indicator factors and one objective parameter as yield. These factors involved age, pregnancy check, BMI, level of insulin, etc. Table 1 reports the data about the connected dataset. The aggregate of 768 examples is available in dataset alongside the arrangement of 8 properties. Among them, around 34.90% and 65.10% of cases falls into the classes of positive and negative examples correspondingly. Moreover, a lot of two classes are available in dataset. This information are tabulated in Table 1.

Table 2 depicts the values attained by comparing the performance of the LR with VP classifiers. From the table, it is indicated that the presented VP model displays very lower outcome by achieving a precision of 88.00. However, the LR manages well and achieves maximum precision of 92.40. Nonetheless, the VP model showed poor performance by achieving a slightly lower recall of 68.04. Concurrently, the LR model exhibits maximum classification procedure by attaining a better recall value of 79.27. Similarly, the LR model improves the performance level is increase in the accuracy value of 77.21. On the other hand, the VP model showed poor performance over the LR by getting the accuracy value of 66.79. From the table, it is indicated that the presented VP model displays very lower outcome by achieving a F-score of 78.37.

However, the LR manages well and achieves maximum F-score of 83.41. Nonetheless, the VP model showed poor performance by achieving a minimum MCC of 0.17. Concurrently, the LR model exhibits maximum classification procedure by attaining a better MCC value of 0.48. Similarly, the LR model improves the performance level is increase in the kappa value of 47.34. On the other hand, the VP model showed poor performance over the LR by getting the least kappa value of 13.52. These values verified that the LR is a better model to calssify the diabetes dataset. Table 3 shows the confusion matrix produced by the LR with and without the use of SC model.

Table 4 shows the results attained by the LR model with and without the use of SC model. Fig. 6 also shows the results attained by various models on the applied dataset. From the figure, it is showed that the recall value of LR is increased from 79.27 to 86.94 by the inclusion of SC based clustering process. At the same time, it is also noted that the accuracy of LR is increased from 77.21 to 79.35 by the inclusion of SC based clustering process. Similarly, it is observed that the F-score of LR is increased from 83.41 to 83.69 by the inclusion of SC based clustering process. Likewise, interms of kappa value, it is noted that LR is increased from 47.34 to 55.66 by the inclusion of SC based clustering process.

On the other hand, the higher error rate of 0.22 by the LR is reduced to 0.21 by the application of the clustering process. In addition, the value of MCC by LR is increased from 0.48 to 0.56 by the inclusion of SC based clustering process. In overall, the presented LR-SC model exhibits maximum classification with the highest recall of 86.94, accuracy of 79.35, F-score of 83.69, MCC of 0.56 and kappa value of 55.66. These values verified that the nature of clustering mechanism greatly helps to increase the classification process.

Table-IV: Classifier results of LR model before and after SC

Classifiers	Precision	Recall	Accuracy	F-score	Error Rate	MCC	Kappa
LR+SC	80.68	86.94	79.35	83.69	0.21	0.56	55.66
LR	88.00	79.27	77.21	83.41	0.22	0.48	47.34

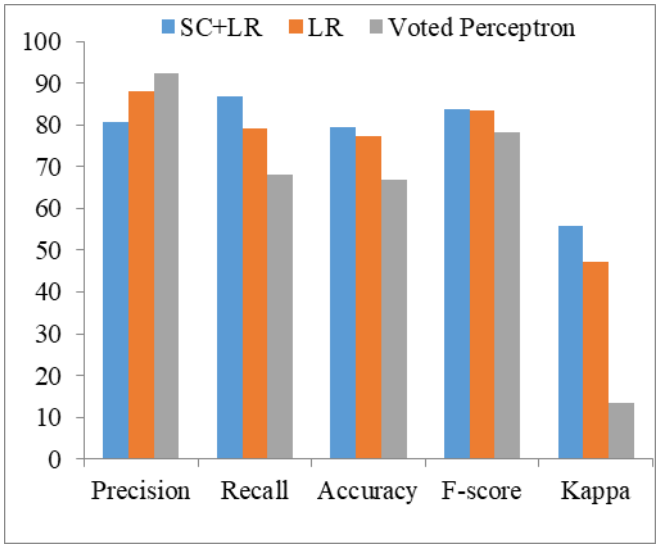

Fig. 6. Classifier results of LR model before and after SC

IV. CONCLUSION

DM offers advancements and approaches to change over enormous data into advantageous data for making decisions. Diabetes is a metabolic disease where a body is unable to control sugar classified by hyperglycaemia. ML algorithms have developed to predict the existence of the diabetes disease. Due to the massive quantity of diabetes disease dataset, clustering techniques can be applied to group the data before classifying it. A new automated clustering based classification model is applied for the identification of diabetes. To cluster the healthcare data, SC model is applied. Then, LR model is applied for the effective categorization of the clustered data. The experimentations have been directed by the benchmark dataset. The simulation outcomes demonstrate that the efficiency of the LR-SC method beats the prevailing methods to predict the diabetes diseases.

REFERENCES

1. Shardlow M. An analysis of feature selection techniques. The University of Manchester; 2016.
2. Dash M, Liu H. Feature selection for classification. Intell Data Anal 1997;1 (3):131–56.
3. Tang J, Alelyani S, Liu H. Feature selection for classification: a review. Data Classif: Algor Appl 2014;37
4. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011.
5. Tan PN. Introduction to data mining. Pearson Education India; 2006.
6. Muni Kumar N, Manjula R. Role of Big data analytics in rural health care – a step towards svasthbharath; 2014.
7. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3(Mar):1157–82.
8. Nalband S, Sundar A, Prince AA, Agarwal A. Feature selection and classification methodology for the detection of knee-joint disorders. Comput Methods Programs Biomed 2016;127:94–104.
9. https://www.kaggle.com/uciml/pima-indians-diabetes-database.