International Journal of Computing and Artificial Intelligence

E-ISSN: 2707-658X
P-ISSN: 2707-6571
IJCAI 2021; 2(2): 23-28
Received: 13-05-2021
Accepted: 15-06-2021

Baduru Sneha
Department of Computer Science, SDHR College, Tirupati, Andhra Pradesh, India

B Pavan Kumar Reddy
Assistant Professor, Department of Computer Science, SDHR College, Tirupati, Andhra Pradesh, India

Corresponding Author:
Baduru Sneha
Department of Computer Science, SDHR College, Tirupati, Andhra Pradesh, India

An efficient prediction of diabetes using machine learning approaches

Baduru Sneha and B Pavan Kumar Reddy

DOI: https://doi.org/10.33545/27076571.2021.v2.i2a.34

Abstract
Diabetes is a one of the fundamental wellsprings of visual disability, kidney frustration, expulsions, cardiovascular breakdown and stroke. Diabetes is conceivably the most appalling disease that humanity is defying at the present time. The sickness occurs considering body's unseemly response to insulin: which is a huge substance in our body that converts sugar into energy needed for authentic working of standard life. The diabetic disorder has genuine complexities on our body as it assembles the peril of making kidney ailment, coronary ailment, eye retinal contamination, nerve damage and vein hurt. This paper bases on progressing developments in AI which have had colossal impacts in the ID and finish of diabetes. In this paper we developed a conjecture model for diabetes representation using Decision tree and KNN classifier. The introduction of these applied methods is settled using the components precision, exactness and review. The results gotten shown that choice tree beats KNN and choice tree with most raised exactness of 100%. Execution assessment of these request strategies helps us with picking which fitting strategy to pick in future for separating the given dataset.

Keywords: machine learning, decision tree, KNN and classification

1. Introduction
Diabetes is genuinely not an inborn issue in any case heterogeneous get-together of strife which could ultimately achieve an impact of glucose inside the blood and nonappearance of glucose inside the pee. Diabetes is ordinarily coming about in light of genetic characteristics, technique forever and ecological elements. Eating a hazardous weight decrease plan, being overweight expect part in developing the diabetes. High glucose levels can moreover achieve kidney ailments, coronary heart ailments [10]. The excess of sugar in the blood can hurt the tiny veins in your packaging. Signs of diabetes are foggy creative and farsighted, unbelievable yearning, remarkable weight decline, ordinary pee and dried. In this paper, limits used inside the real factors set to discover the diabetes are Glucose, Blood pressure, pores what's more, skin thickness, Insulin, Age. Tremendous volumes of bits of knowledge units are made by clinical consideration adventures. Those real factors sets are a collection of patient information about the diabetes from the crisis facilities. Enormous records examination is the getting ready which it reviews the information units and shows the restricted information. Pima Indians Diabetes Database, this dataset is taken from the public Institute of Diabetes and Digestive disorders. The objective of the dataset is to anticipate whether the patient has diabetes or on the other hand not, basically established on insightful assessments in the dataset. A couple of objectives were taken from the immense informational index.

1.1 Signs or Symptoms of Diabetes: Frequent Urination, expanded thirst, Increased longing, Tired/Sleepiness, Weight mishap, Blurred vision. Mental scenes, Confusion and inconvenience concentrating, ordinary infections/vulnerable recovering. That is the explanation it's so basic to both realize what advised signs to look for and to see a clinical consideration provider reliably for routine wellbeing screenings [10]. PC Supported Diagnosis is a rapidly creating novel space of examination in clinical industry. The new experts in AI ensure the improved precision of understanding and assurance of disorder. Here the PCs are enabled to think by making information by learning. There are various sorts of Machine Learning Strategies and which are used to arrange the educational assortments [1].

2. Characterization

AI computations are generally requested as being directed or solo. An oversaw learning computation uses the previous experience to create assumptions on new or subtle data while independent estimations can draw surmising from datasets. The directed learning is moreover called portrayal. This examination businesses portrayal system to make a more exact perceptive model as it is perhaps the most regularly applied AI technique that takes a gander at the getting ready data and makes an assembled limit, which can be used for arranging new or covered models [4, 5]. The critical goal of the plan strategy is to appraise the target class absolutely for each case in the data.

Request is a framework to characterizations data into an ideal and specific number of classes where we can consign imprint to each class. Additionally, conjecture models expect steady regarded limits. Effectively diagnosing these issues are used to data mining and AI strategies [6, 7]. Portrayal is one of the immense procedures for disorder gauge. Request is the most outstanding data mining assignments [2, 3]. Enormous proportion of business and clinical instructive assortments commonly incorporates request. Portrayal is a data mining work that can administers the things in a collection to target classes. Request is maybe the most basic unique systems in various authentic world issue [6, 8]. In proposed, using AI estimations. In this work, the key objective is to mastermind the data as diabetic or non-diabetic and improve the course of action precision. For a couple of request issue, the higher proportion of tests picked at this point it doesn't prompt higher request accuracy. In various perspectives, the execution of computation is high with respect to speed anyway the precision of data portrayal is low. The essential objective of our model is to achieve high precision.

3. Choice Tree

Choice tree is a gathering technique. This system is generally use for assumption and gathering. A tree contains ways, branches and leave centers. Grouping of branches is considered way and addresses the attribute regard. Leaves tended to Class regard. Each route in decision tree addresses a standard which is used for game plan or assumption. In this computation whole truths are tended to inside the condition of tree in which each leaf is identifies with the class name likewise, trademark are identifies with inner center of the tree. Choice tree segments the data into subsets or center points. Root center tends to the all-out dataset. Tree pruning is preformed after tree is built completely. Pruning is started from the lead center [6].

Choice tree uses tree structure and the tree begins with a single center tending to the readiness test. If the models are all in a comparable class, the center transforms into the leaf and the class marks it. Something different, the computation picks the one-sided quality as the current center point of the choice tree. As shown by the value of the current choice center point trademark, the arrangement tests are isolated into serval subsets, all of which shapes a branch, and there are serval regards that structure serval branches. It is a controlled learning strategy, which is used for handling plan issues. Choice tree [6, 8] is a strategy which iteratively breaks the given dataset into at any rate two model data. The goal of the technique is to predict the class worth of the goal variable. The choice tree will help with confining the educational assortment and builds the decision model to predict the dark class names. A decision tree can be worked to both twofold and constant components. Choice tree in a perfect world finds the root center point subject to the most imperative entropy regard. This gives decision tree an advantage of picking the most unsurprising theory among the readiness dataset. A commitment to the choice tree is a dataset, including a couple attributes and models regards and yield will be the choice model.

3.1 Steps for Decision Tree Algorithm

- Build tree with centers as data incorporate.
- Select part to expect the yield from input include whose information secure is generally raised.
- The most vital information secure is resolved for every property in each center point of tree.
- Rehash stage 2 to shape a sub tree using the part which isn't used in above center.

3.2 K-Nearest Neighbor Algorithm

K-nearest neighbor is essential portrayal and backslide computation that used non parametric procedure proposed by Aha *et al.* [7]. In plan affirmation, the k-nearest neighbors' computation is a non-parametric procedure used for gathering and backslide. In the two cases, the information contains the k closest getting ready models in the segment space. The yield depends upon whether KNN is used for plan or backslide. The computation records each and every considerable characteristic and describes new credits subject to their comparability measure. To choose the partition from point of convergence to centers in planning data set it uses tree like data structure. The property is orchestrated by its neighbors. In a gathering system, the value of k is reliably a positive number of nearest neighbors. The nearest neighbors are perused a lot of class or article property esteem [8].

4. Test Results

The assessments have been coordinated by using Python programming language. It is an open-source programming language give astonishing utilization of different data examination and Visualization methodologies. It is a weighty library that gives numerous AI gathering computations, capable devices for data mining and data assessment. The Python Scikit-learn is a pack for data request, backslide, clustering and portrayal. We have thought about the Pima Indian Diabetes Dataset (PIDD) information from UCI Machine Learning Repository datasets [9]. This Data set has 768 lines and 9 segments. So, in this information there are two class marks i.e., the tried negative class has 500 and tried positive class has 268. The characteristic data information is consolidated in Table-1. The standard dataset is parceled into two sets (70% and 30%), one for getting ready containing 537 examples and another set for testing contains 231 cases.

Table 1: Attributes Information

S. No	Name of the Attribute	Description
1	preg	Number of times pregnant
2	plas	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3	pres	Diastolic blood pressure (mm Hg)
4	skin	Triceps skin fold thickness (mm)
5	test	2-Hour serum insulin (mu U/ml)
6	mass	Body mass index (weight in kg/(height in m)^2)
7	pedi	Diabetes pedigree function
8	age	Age (years)
9	class	Class variable (1:tested positive for diabetes, 0: tested negative for diabetes)

4.1 Measures for execution assessment

There exist various measures that can be utilized to assess the presentation of a classifier, like Accuracy, affectability, explicitness, exactness and review, and so forth Every one of these assessment measures have their own impediments and, therefore, a proper assessment measure which best suits the issue ought to be chosen. Because of the elements referenced, and to have a solid exhibition assessment, the evaluation ought to be viewed as dependent on the Cross approval.

To limit the inclination related with the arbitrary inspecting of the preparation and holdout information tests in contrasting the prescient precision of at least two techniques, scientists will in general utilize k-overlap cross-approval

Execution of every classifier is measure regarding disarray framework, affectability, particularity, exactness, review and precision. These measurements are customarily characterized for a twofold arrangement task with positive and negative classes. That is:

Accuracy: Accuracy is a measure which determines the probability that how much results are accurately classified.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision: Precision represents how precise the classifier predictions are since it shows the amount of true positives that were predicted out of all positive labels assigned to the instances by the classifier. Precision is the proportion of positive predictions that are correct

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Recall: Recall is the proportion of positive samples that are correctly predicted positive. It shows the amount of truly predicted positive classes out of the amount of total actual positive classes.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

Where,

- True positive (TP) = number of positive samples correctly predicted.
- False negative (FN) = number of positive samples wrongly predicted.
- False positive (FP) = number of negative samples wrongly predicted as positive.
- True negative (TN) = number of negative samples correctly predicted.

Table 2: Confusion Matrix of Prediction cases of classification

		Predicted	
		Positive	**Negative**
Actual Class	Positive	TP	FN
	Negative	FP	TN

Confusion matrix is a visualization tool which is commonly used to present the accuracy of the classifiers in classification that assist with performance evaluation purposes which consist of the concepts defined above measurements. This is illustrated in table-2. It is used to show the relationships between outcomes and predicted classes. The level of effectiveness of the classification model is calculated with the number of correct and incorrect classification in each possible value of the variable being classified in the confusion matrix.

4.2. Results

To approve the expectation consequences of the Naïve Bayes arrangement and the 10-overlap hybrid approval is utilized. The k-overlap hybrid approval is normally used to lessen the mistake came about because of irregular examining in the correlation of the exactness's of various forecast models. The current investigation partitioned the information into 10-folds where 1-crease was for trying and 9-folds were for preparing for the 10-overlap hybrid approval.

The disarray framework of Naïve Bayes arrangement technique is introduced in the table-3. The qualities to quantify the exhibition of the strategies (for example exactness, accuracy, review, and f1-score) are gotten from the disarray framework and appeared in table-4 and same appeared in graphical portrayal in figure-1

Table 3: Confusion Matrix of Pima Diabetes data classification

Algorithm	Training Data (231)		
	Desired Result	Output Result	
		Negative	Positive
Decision Tree	Negative	155	0
	Positive	0	76
KNN	Negative	123	31
	Positive	26	51

Table 4: Results of Pima Diabetes Classification

Algorithm	Accuracy	Precision	Recall	f1-score
Decision Tree	100	100	100	100
KNN	75.3	76	75	76

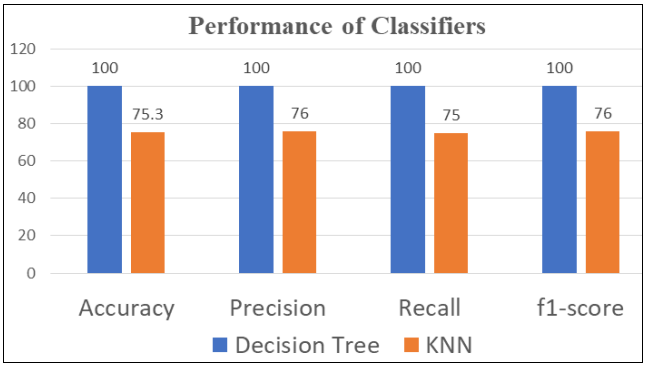
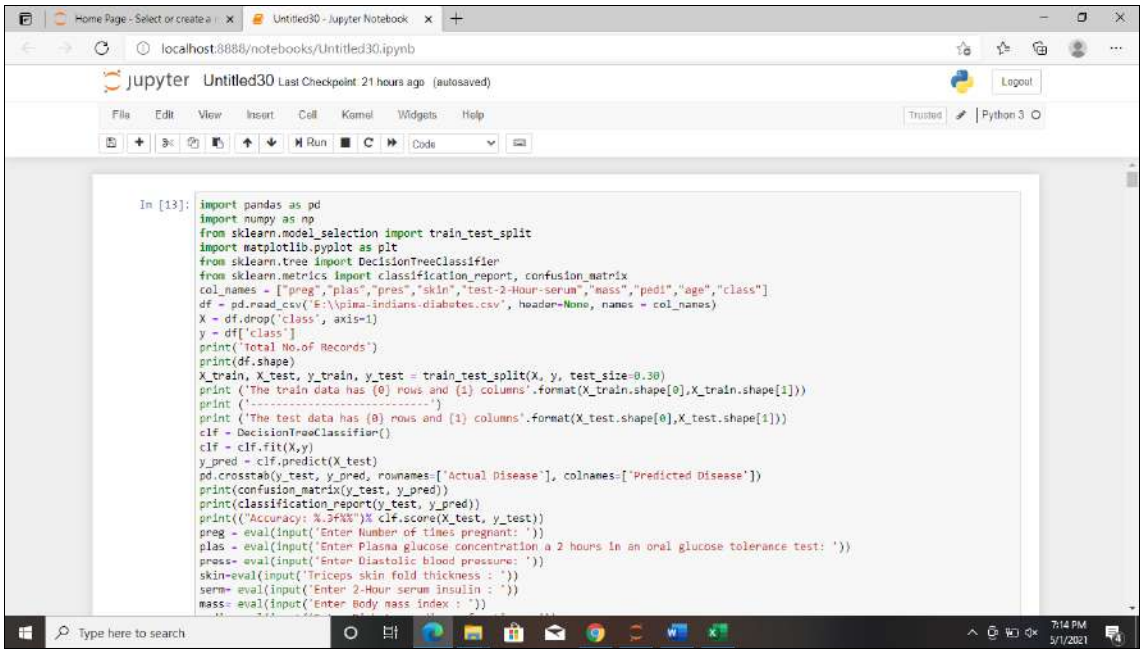
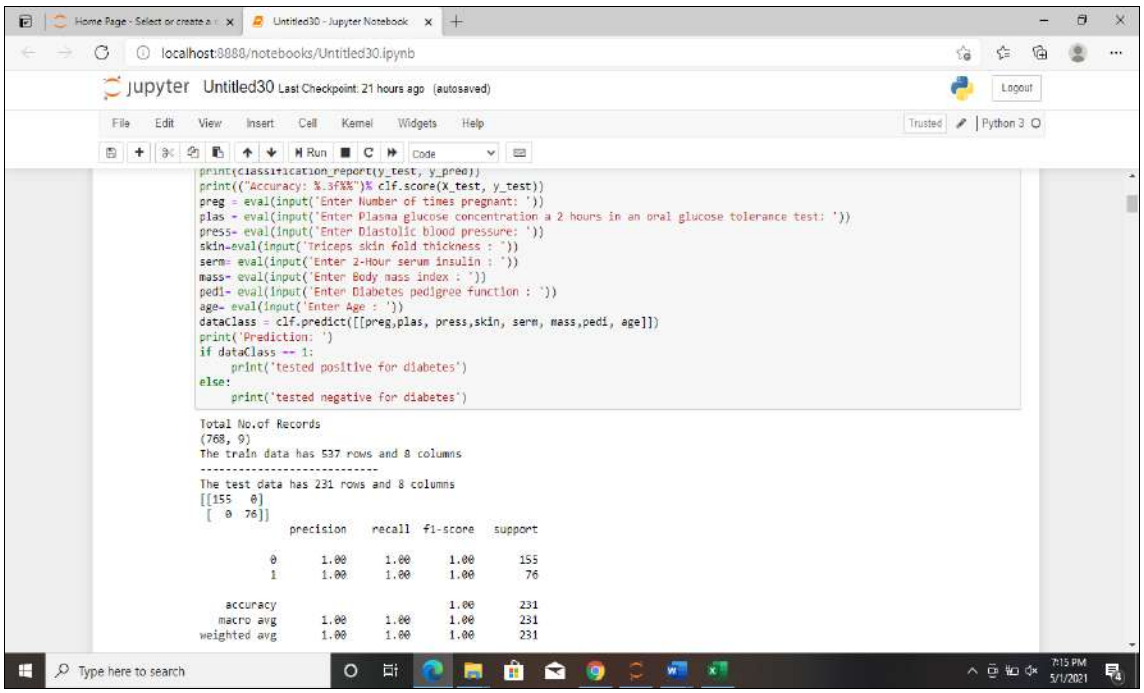


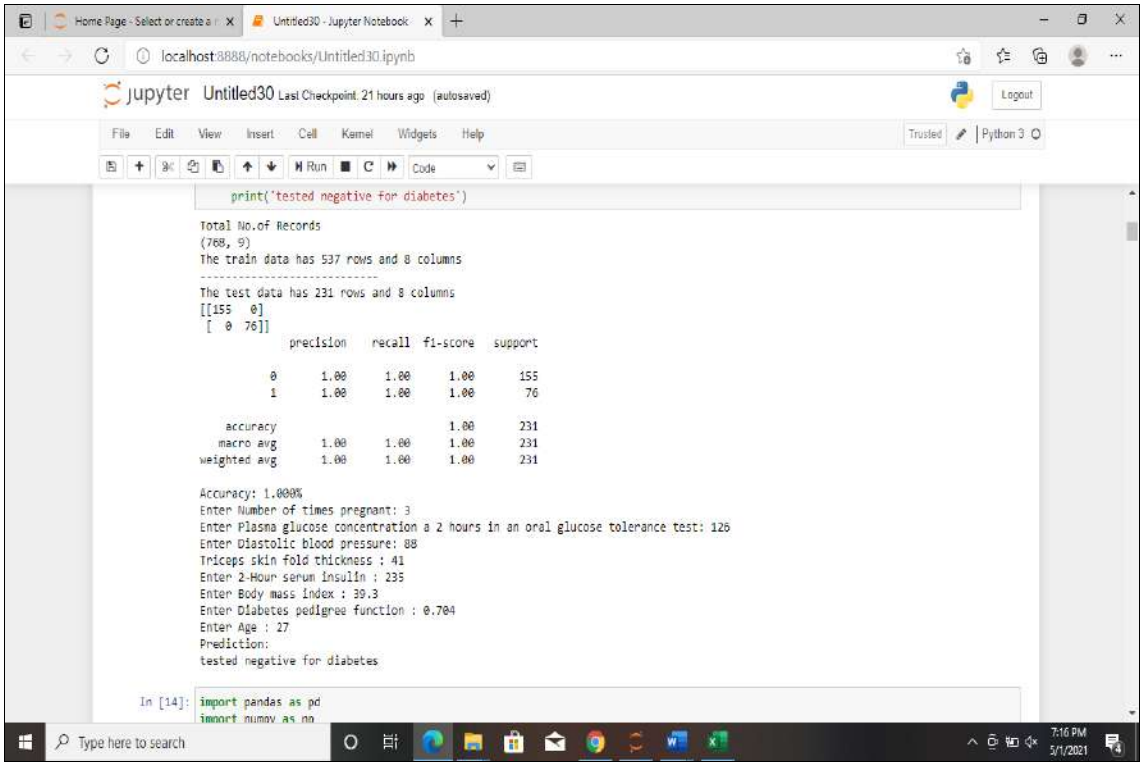
Fig 1: performance of Classifiers

We notice execution of Decision Tree and KNN as demonstrated in the figure-1, the precision of choice tree has accomplished 100%, while the KNN has 75.3% exactness. The accuracy of choice tree has 100% and KNN has 76% accomplished. So, in all the exhibition measurements, the choice has accomplished most elevated when contrasted with KNN characterization.

4.3 Screen shots







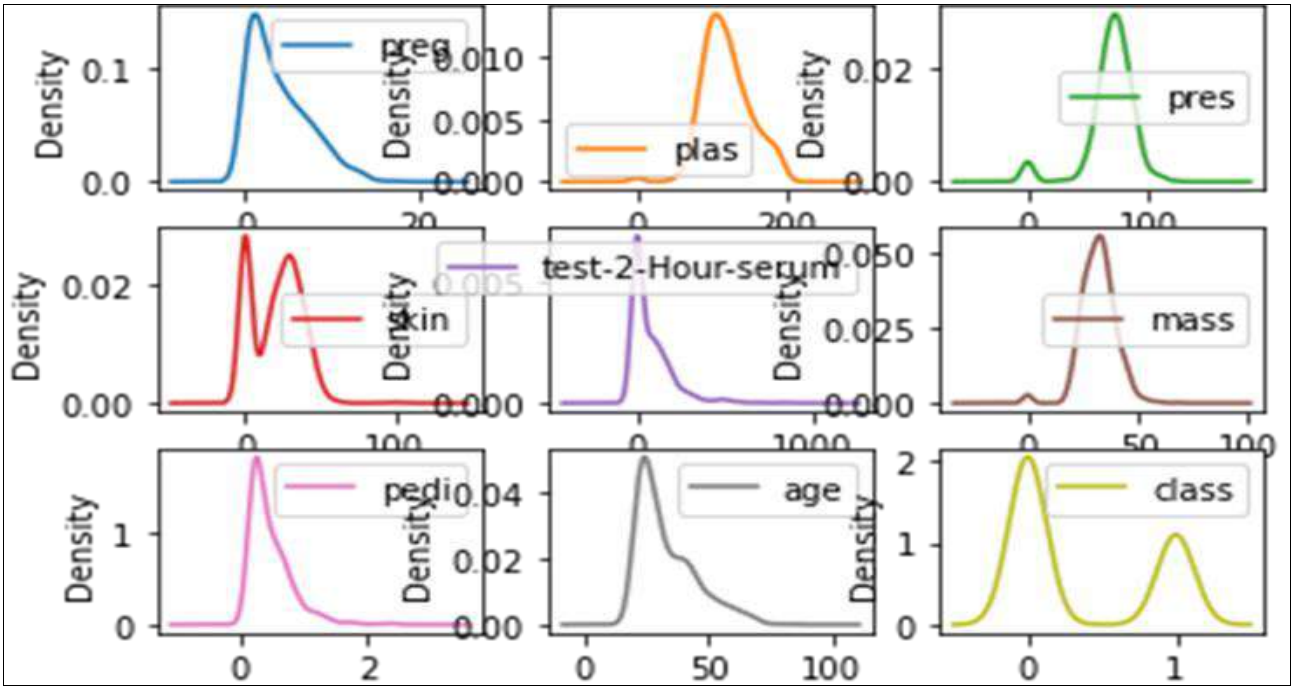


Fig 2: Density Plot of Diabetes data Statistical summary

5. Conclusion

This assessment focused in on AI request estimations for expecting diabetes infection with more accuracy. The customized assurance of diabetes is a critical genuine clinical issue. Area of diabetes in its starting stages is the key for treatment. This paper shows how Decision Trees and KNN are used to show genuine finding of diabetes for close by and conscious treatment, close by presenting related work in the field. The unmistakable display potential gains of course of action estimations are resolved on various measures. Train and test the data like Pima Indians Diabetes Dataset. The gathering computation achieved most limit testing precision. Preliminary outcomes show the feasibility of the choice tree model. The show of the techniques was investigated for the diabetes end issue. Preliminary outcomes display the adequacy of the proposed model.

6. References

1. Yu D, Deng L. Deep learning and its applications to signal and information processing, IEEE Signal Process. Mag 2011;28(1):145-154.
2. Ravi Kumar G, Nagamani K, Anjan Babu G. A Framework of Dimensionality Reduction Utilizing PCA for Neural Network Prediction", Lecture Notes on Data Engineering and Communications Technologies, ISBN 978-981-15-0977-3. Springer Nature Singapore Pte Ltd 2020;37:173-180.

3. Ravi Kumar G, Venkata Sheshanna Kongara, Dr Ramachandra GA. An Efficient Ensemble Based Classification Techniques for Medical Diagnosis, International Journal of Latest Technology in Engineering, Management and Applied Sciences 2013;2(8):5-9. ISSN-2278-2540.
4. Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann 2005.
5. Ho TJ. Data Mining and Data Warehousing, Prentice Hall 2005.
6. Han J, Kamber M. Data Mining concepts and Techniques, the Morgan Kaufmann series in Data Management Systems, 2nd ed. San Mateo, CA; Morgan Kaufmann 2006.
7. Michael N. Artificial Intelligence – A Guide to Intelligent Systems, 2nd Edition, Addison Wesley 2005.
8. Tan PN, Steinbach M, Kumar V. Introduction to Data Mining, A: Addision-Wesley 2005.
9. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/.
10. www.diabetesresearch.org/document.doc?id=284