IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-3, March 2019 E-ISSN: 2347-2693

Predicting the Birth of Healthy Babies with Gestation Period Observations using Machine Learning Algorithms

K. Menaka[1*], B. KeerthanaKani[2]

[1,2]Department of Computer Science, Shrimati Indira Gandhi College, Tiruchirappalli – 2

[*]*Corresponding Author: kmenakasigc@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i3.271275 | Available online at: www.ijcseonline.org

Accepted: 10/Mar/2019, Published: 31/Mar/2019

Abstract Machine learning is the most familiar division of Artificial Intelligence to perform exploratory data analysis tasks and to work out a variety of problems such as weather forecasting, drug discovery, encrypted image detection etc., This paper discusses about varieties of data mining classification algorithms that are commonly used to extract considerable knowledge from huge volumes of data. Identification of the healthiness of a baby with the observations during the gestation period of a mother requires various parameters to be taken into consideration during that period. Decision Tree (DT) algorithms could be very much helpful in predicting the healthiness of a baby. The numerical form of the data sets are taken and are fed to the DT algorithms to make calculations for the prediction of the healthiness of the baby. The data sets are taken and analyzed in the Waikato Environment for Knowledge Analysis (WEKA) platform.

Keywords: Data Mining, Knowledge Discovery, Classification Algorithms, WEKA.

I. INTRODUCTION

Data Mining refers to the extraction of hidden and possibly useful information from a database. Data Mining in general is a part of Knowledge Discovery process [1][2]. The basic idea here is to develop computer programs that sieve through the databases easily, looking for the retrieval of useful patterns. Powerful patterns, if generated, will simplify the process of accurate predictions in future. Data Mining involves various tasks like preprocessing, classification, clustering, outlier detection etc. The classification task accurately predicts the target class for each case in the data. Clustering involves the analysis of large amount of data to extract unknown useful and important patterns such as group of data records. This clustering stage in the data mining process would recognize multiple groups in the data that will be used to acquire more precise prediction results.

WEKA [3] is a computer program that was developed at the University of Waikato in New Zealand. WEKA tool is primarily used to identify information from raw data from unripe domains. The basic principle of this tool is to exploit a computer application that can be skilled to execute machine learning capabilities and obtain functional information in the form of trends and patterns. WEKA operates on the assumption that the data object supplied by the user is characterized by a fixed number of attributes which are a specific type like alpha-numeric or numeric values. With the help of WEKA, varieties of data mining tasks could be performed in medical databases. One among them and the one which is most uncommon is the prediction of the birth of healthy babies using the gestational period observations. The period of gestation is typically divided into three "trimesters" of approximately equal length. But, there are a wide variety of maternal and fetal conditions that can lead to complications of pregnancy. The motivation of this work is to take the above mentioned problem into account and it compares different classification algorithms for it and predicts the birth of healthy/unhealthy babies with the gestational period observations using Machine Learning algorithms.

This paper has been organized as follows: Section-I contains the Introduction for the problem, Section – II contains the Materials and Methods used for the work, Section-III briefs the Methodology adopted in this work, Section – IV states the Results and Discussion and Section-V presents the Conclusion and Future Scope of this work.

II. MATERIALS AND METHODS

The following lists the various algorithms which have generated the best results for the supplied set of data in this work though other algorithms also are available for the classification process.

Classification

Classification is the process which refers to **t**he categorization of objects or ideas in the form in which they have been identified or distinguished. A classifier algorithm

is one which implements the classification process with a set of supplemented data. In machine learning terminology, classification is thought of as an instance of the supervised learning process and the analogous unsupervised learning process is generally referred to as clustering or cluster analysis.

In the machine learning classification procedure, a model is generated with the training examples which learns and classifies the data samples into known classes. The following sequence of operations are involved in the classification process (i) Training data set creation (ii) Identification of classes attributes and classes (iii) Significance of useful attributes for classification (iv) Generation of learning models by tuning the to produce accurate results (v) Validating the generated model by supplying the known set of data and (vi) Testing the model with different set of data.

A brief description of the characteristic features of the algorithms employed in this study is outlined below

Bayes

The Naive Bayesian classification algorithm

Naïve Bayesian classifier is one of the most widely used classification algorithms. The Naïve Bayesian Classifier provides a simple approach with clear semantics, to represent, use and to learn probabilistic knowledge. The method is intended for use in supervised induction task in which the performance objective is to precisely predict the class of test instances and in which the training instances use class information [4].

Function

Multi Layer Perceptron

Multi Layer Perceptron is a classifier that uses the machine learning algorithm backpropagation to classify instances. WEKA uses a graphical interface that lets us to create a network structure with as many perceptrons and connections as desired. The network parameters can also be monitored and modified during training time. The network nodes are all sigmoid.

Rules

OneR

OneR (One Rule) is a simple and accurate classification algorithm which generates one rule for each predictor in the data. This algorithm then selects the rule with the smallest total error as its "one rule". To create a rule for a predictor, a frequency table is constructed for each predictor against the target.

ZeroR

ZeroR is the simplest classification method which relies on the target and ignores all predictors. It just determines the most common class−Or the median (in the case of numeric values)−Tests how well the class can be predicted without considering other attributes−Can be used as a Lower Bound on Performance.

Decision Tree algorithms

The solution to a given problem can be effectively found with the help of the Decision Tree (DT) algorithm. The internal nodes of the tree represent attributes and the leaf nodes represent class labels. The decision trees used in data mining are of two main types: those used for classification analysis where the predicted outcome is the class and those used for regression analysis where the predicted outcome is a real number. The difference between them resides in making the split points at the class-labeled training tuples. The principle components of the DT are the decision node, branches and the leaves and the first component is the decision node (root node) which indicates a test to be carried out. Based upon the results of the test, the tree will be splitted into branches and each branch will represent one of the possible answers.

Decision stump

A decision stump is a machine learning model consisting of a one-level decision tree [5]. It has one internal node connected to the terminal (leaf) nodes. It performs regression or classification.

J48

The J48 algorithm generates a pruned or unpruned decision tree. [6]. The decision here is built with labeled training data set using information gain and it is examined clearly which results from choosing an attribute for data split. The attribute with the highest normalized information gain is used to make decisions.

Random Forest

In this type of classifier, a forest of random trees is constructed. The random forest classifier adopts bootstrap aggregations or bagging of decision trees [7]. This builds multiple DTs by repeatedly reassembling training data with substituting and voting the trees for a consensus prediction.

Random Tree

Random tree is an ensemble learning algorithm and it is a supervised classifier [8]. It generates many individual learners. In this type of classifier, each node is split using the best split among all the variables. This constructs K randomly chosen attribute at each node. It has no pruning.

III. METHODOLOGY

Feature selection and test data

The datasets for the experiments were taken from [9]. It contains the records which are taken during the gestational period of a mother. The following parameters were found in the data base: *mcode, height, age, SGA (Small for gestational age : 0= no, 1 = yes), parity, smoker,*

birth_weight, sex of infant, gestational_age. During the preprocessing stage of the work, another parameter called *Category* has been added as another parameter for the data base for predicting the birth of healthy babies. There are varieties of maternal and fetal conditions that may lead to complications during pregnancy. For example, (i) *Low Birth Weight (LBW)* – The babies whose weights are less than 2500g at birth are considered to be unhealthy babies; (ii) *Pre-term delivery:* Babies born prior to 37 weeks of gestation will undergo lower birth weight complications and are also considered to be unhealthy babies. *(iii) Small for gestational age (SGA) :* Intra-uterine growth restrictions may also lead to low weight babies and are also considered to be unhealthy than normal babies and this is usually considered to be due to insufficient nutrition received by the fetus. There are some maternal factors like chronic medical conditions like diabetes and hypertension, poor maternal nutrition, maternal smoking which may also lead to the birth of unhealthy babies.[9].

The data sets taken from [9] consists of 755 records of pregnant women from South Africa At each clinic visit, the womens' weight and symphysis fundal height (SFH) were recorded along with the above mentioned characteristics. Based on the above observations, during the preprocessing stage of the work, a condition is formed which will predict whether the mother will yield a healthy baby or not during the gestational period itself. Using the above measurements, taken prior to 30 weeks of pregnancy, a model has been developed in this work using WEKA which has accurately distinguished whether a woman will deliver a healthy baby or an unhealthy baby.

Machine Learning Experiments

The inoculation of artificial intelligence, the DT algorithms or any other machine learning algorithms can be done with the proper set of preprocessed data. The learning experiments are performed using WEKA (Waikato Environment for Knowledge Analysis) (Version 3.8.3) [3] which is a popular suite of machine learning software written in JAVA. All the experiments are carried out with a 2.4 GHz, i5 processor in a 64-bit operating system with 8 GB RAM. Three phases of experiments were conducted viz. (i) Training Phase (ii) Validation Phase and (iii) Testing Phase. The 744 records in the dataset are divided into two sets, one for training and validation phase with 644 records (from which only 60% of records are taken for performing validation) and another one for the testing phase with 100 records.

IV. RESULTS AND DISCUSSION

Evaluation Criteria

The classifiers' performance of the machine learning algorithms can be calculated on the basis of some parameters. The most common among them are: TP rate, FP rate, Precision, Recall F-Measure and ROC area. They are all explained below. The classifier's accuracy for a given test set of data can be identified from the percentage of test set tuples that are correctly classified by it. The misclassification rate of a classifier is termed as the Error rate M which is (1-Acc(M)), where Acc(M) defines the accuracy value of M. The confusion matrix in the classifier is helpful to analyze how well the classifier recognizes the records of different classes. Sensitivity and Specificity measures are used to find the accuracy of a classifier. Sensitivity can also be referred to as true positive rate which is the proportion of positive tuples that are identified correctly. Specificity refers to the true negative rate which is the proportion of negative tuples that are identified correctly. The above measures can be defined as follows:

Accuracy = {(TP+TN)/(TP + FP+TN+FN)}

Sensitivity or Accuracy on Class-I (Healthy Baby)

= {TP/ (TP + FN)}

Recall or Precision Class-I (Healthy Baby)

= {TP/ (TP + FP)}

Specificity or Accuracy on Class-II (Unhealthy Baby)

= {TN/ (TN + FP)}

Recall or Precision Class-II (Unhealthy Baby) = {TN/(TN + FN)}where, True Positive (TP) and True Negative (TN) are correctly predicted Class-I and Class-II categories respectively. Similarly, False Positive (FP) and False Negative (FN) are wrongly predicted Class-II and Class-I categories respectively. To check whether the present classification scheme is better than a random prediction, a reliability factor (R) can be computed. It is given as

R = [((TP + FN) * (TP + FP)) + (TN + FN) * (TN + FP))] / (TP + TN + FP + FN)

A factor **S** which is independent of the total number of samples in the data set can also be computed as

S = ((TP + TN) - R) / [((TP + TN + FP + FN) - R) x 100]

and it gives the normalized percentage of correctly classified Class-II better than random classification. S = 100% signifies a perfect classification and S= 0% signifies a poor classification. The overall performance could also be exposed with a static parameter called F which is expressed as F = (2 x TP) / (2 x TP + FP + FN).

Performance of DT algorithms

Training

All the computational parameters are performed with the default parameters provided by the WEKA tool for the respective algorithms. An accurate classification of healthy (class I) and unhealthy (class II) babies are achieved in the training phase in almost all the algorithms except for the Rule based *ZeroR* algorithm which showed only 90.21% accuracy in which only 581 records are correctly classified and the remaining 63 records are incorrectly classified while the others showed more than 95% accuracy. The *RandomForest* algorithm of the decision tree showed the best result among all the other algorithms with 99.68% accuracy. The following screen shot shows the performance of the *RandomForest* algorithm during the training phase.

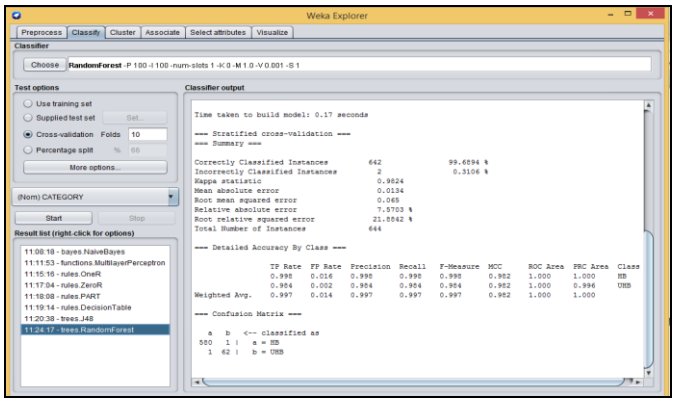
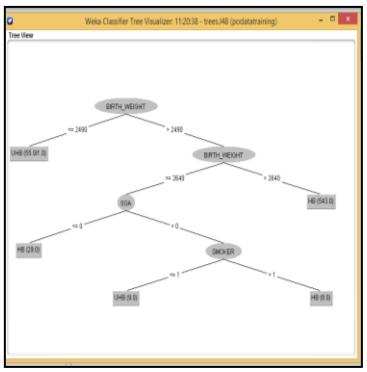


Figure 1: Performance of RandomForest algorithm during Training phase



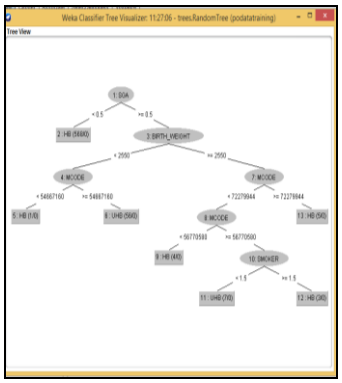


Figure2. Performance of J48 and Random Tree Algorithms

Validation

A classifier data model which is generated can be evaluated by splitting the data sets into two parts: training and testing parts. It is necessary to make the splitting either as a percentage split or as cross-validation procedure. The percentage split method has been adopted in this work by separating 60% of the data for learning and the remaining 40% for testing. In this phase also, the zeroR algorithm provided poor result of 89.53% of accuracy only. All the other algorithms showed more than 95% accuracies and the Multilayer Perceptron algorithm showed the best result of 98.83 % accuracy in this phase. The following screen shot shows the performance of the *Multilayer perceptron* algorithm during the validation phase.

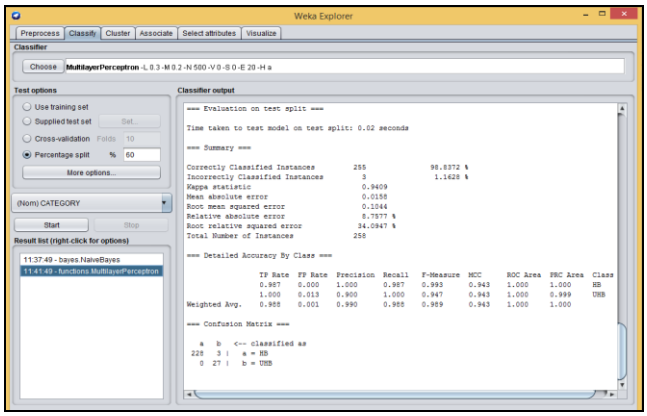


Figure 3: Performance of MultiLayer Perceptron algorithm during Validation Phase

Testing:

In this phase, it has been found that all the algorithms performed effectively produced good classification result except the *Randontree* algorithm and the *ZeroR* algorithm. They both produced only 80% accuracy in the testing phase. All the remaining algorithms showed above 99% accuracies in this phase also. The Function based *Multilayer Perceptron* algorithm produced 100% accuracy in the Testing Phase which proves that this algorithm is the best one for the problem under consideration.

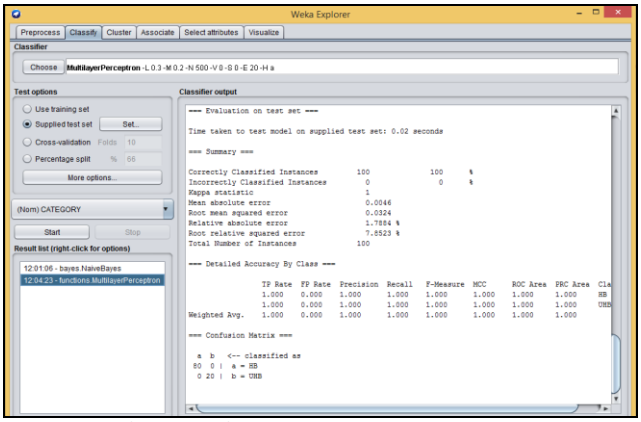


Figure 4: Performance of MultiLayer Perceptron algorithm during Testing Phase

V. CONCLUSION AND FUTURE SCOPE

In this work, the evaluation of the classifying power of some familiar DT algorithms on the given datasets. Among all the algorithms used for this study, the Multilayer Perceptron performed reliably good during training, validation and the testing phases. The other algorithms also produced more than 95% of accuracy for the prediction of healthy/unhealthy babies except the ZeroR algorithm. We conclude that the problem taken under consideration is a rare one and will be effectively useful as it helps for regular monitoring of the pregnant woman to look for symptoms or signs indicative of prediction of the birth of the babies during the gestational

period. The work could be extended by incorporating it by adding more parameters and finding out other complications that may occur during pregnancy period of a woman and providing better machine leaning algorithm outputs to avoid those complications prior to the delivery. This could also be extended by comparing the performance of these algorithms for their multilevel classification efficiencies.

REFERENCES

[1]J. Han and M. Kamber. "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2000.

[2]G. Piatetsky-Shapiro, U. M. Fayyad, and P. Smyth, "From data mining to knowledge discovery: An overview". In U.M. Fayyad, et al. eds.), Advances in Knowledge Discovery and Data Mining, 1-35. AAAI/MIT Press, 1996.

[3]Eibe Frank, Mark A. Hall, and Ian H. Witten (2016), " The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[4] George H. John and Pat Langley, "Estimating continuous distribution in Bayesian Classifier", In the proceedings of the Eleventh conference on uncertainty in Artificial Intelligence,Morgan Kaufmann Publishers, San Mateo, 1995.

[5] WayneIba; and Pat Langley,(1992); Induction of One-Level Decision Trees, in ML92: Proceedings of the Ninth International Conference on Machine Learning, Aberdeen, Scotland, 1–3 July 1992, San Francisco, CA: Morgan Kaufmann, P.233–240.

[6] J.Ross Quinlan, "Programs for Machine Learning", Edited by Morgan Kaufmann Publishers,Inc.,1993.

[7] Leo Breiman, "Random Forests Machine Learning" 45(1):5-32; (2001).

[8] Cutler.A, "Fast Classification Using Perfect Random Trees", TechnicalReport5/99/99, Department of Mathematics and Statistics, Utah State University, May 1999.

[9] Scott S. Emerson, M.D., Ph.D. Professor of Biostatistics Department of Biostatistics, University of Washington, Emerson Statistics: Datasets and Documentation.

Authors Profile

Dr. K. Menaka is working as an Assistant Professor in the Department of Computer Science of Shrimati Indira Gandhi College, Tiruchirappalli affiliated to Bharathidasan University, Tiruchirappalli. She has completed her Ph.D. in Computer Science in the year 2015. She has 14 years of Teaching Experience and 4 years of Research experience. Her area of specialization is Artificial Neural Networks. She has published many papers in her research area and has been guiding research scholars.

Ms. B. Keerthanakani is doing her M. Phil. in computer science at Shrimati Indira Gandhi College, Tiruchirappalli. She has finished her under graduate and her post graduate in Computer Science and presently she is in her M.Phil. She has put her effort in this work along with the first author of this paper for completing her project work during the course successfully. Her area of interests include Machine Learning Algorithms and Network Security.

