Prediction of Child Development using Data Mining Approach

Juhi Bansod
SRIEIT, Goa University
Goa, India

Mugdha Amonkar
SRIEIT, Goa University
Goa, India

Ambica Naik
SRIEIT, Goa University
Goa, India

Tina Vaz
SRIEIT, Goa University
Goa, India

Manjusha Sanke
SRIEIT, Goa University
Goa, India

Shailendra Aswale
SRIEIT, Goa University
Goa, India

ABSTRACT

Child development is the sector that consists of scientific research of the patterns of growth, power and change that arise from conception through adolescence. One can apply this knowledge to realize the necessities of a child by viewing how and why individuals alternate and grow. Thus, pleasing them and allowing them to arrive at their maximum capacity. The intention of this study is to research the child growth based on the features consisting of age, height and weight. In order to understand how the physical growth of child switches with time, data is gathered from various sources such as Anganwadis, Primary Schools and Primary health Centres and a data mining method is implemented to expect the child growth. In this method, assessment of two data mining approaches ID3 Decision Tree and Naïve Bayes classifier is carried out on the basis of factors such as prediction accuracy, error rate and learning time.

Keywords

Child Development, Data Mining, Prediction, Naïve Bayesian Classifier, ID3 Decision Tree, Machine Learning

1. INTRODUCTION

Early childhood is the duration when most of transitions in children take place; this period is a significant have an effect on of child development as child progress into formative years and adulthood. This length plays a large position in shaping other components of formative year's development, including growth. India has the second one maximum variety of obese children within the global with 14.4 million and its incidence is growing rapidly. In 2015 over 2 billion children internationally were overweight. One in four children beneath age 5 worldwide is as brief as to be classified as stunted. Child stunning is a key marker of child malnutrition. Classification is independent of each other, it is straightforward and fast to predict. The excellent of a baby's earliest environments and the provision of appropriate experiences at the proper tiers of improvement are critical determinants of the manner each baby's growth develops. This is a large hassle in many locations which includes remote villages in which the possibilities of child getting suffering by weight problems and stunted growth is very excessive and availability of expert pediatrician is a trouble. The hassle of availability of experts may be solved by using applying a data mining technique to predict the child's future physical growth.

2. RELATED WORK

A survey is conducted on related work as illustrated in table 1.

In [1], a comparative study is performed on Decision tree and Naïve bayes to predict common development of child in step with its age. The data is gathered from schools, Anganwadis, and parents by generating a questionnaire. Naïve Bayes and Decision tree algorithms are used. It is concluded that decision tree gives more correct result than naïve bayes algorithm.

In [2], various machine learning strategies are implemented to predict the primary mental fitness issues. The dataset is accumulated from medical psychologists. An evaluation is made on the ones 8 machine learning techniques and appeared out for the leading three which may be put into exercise to help mental health experts in diagnosing mental fitness problems. Interview was taken of clinical psychologists to pick out the mental fitness issues that occur among children. This model assists the specialists to identify the problem if the recognized evidences of the affected person are given as input. All the attributes are of nominal type. Date set is pre-processed by getting rid of irrelevant and redundant attributes using Best First Search technique. The WEKA tool is used to compare the accuracy level of classifiers based totally on 3 measures Kappa Statistic, Accuracy and ROC Area. This paper concludes that Multilayer Perceptron, Multiclass Classifier and LAD Tree produce more accurate outcome than the others.

In [3], C5.0 decision tree and Association rules are implemented to find out what stage of delays might occur from which varieties of ailments in children. Dataset is acquired from Yunlin Developmental Delay Assessment Centre. Abnormal and incomplete information is deleted. This study has identified which kind of illness items will cause certain sorts of delays by means of building a decision tree and association rule evaluation to decide the correlations amongst cognitive, language, motor, social emotional developmental delays.

In [4], various feature selection strategies have been used for the classification of childhood obesity. The data is accrued from Standard Kecergasan Fizikal Kebangs aanuntuk Murid Sekolah Malaysia (SEGAK) Assessment Program and the study questionnaire on socio demographic, physical activity and nutritional assessment. The chart of BMI-for-age is utilized for reference, wherein BMI much less than the 5th percentile is taken into consideration as "underweight", the BMI more noteworthy than the fifth however less than the

85th percentile is taken into consideration as “normal”, a BMI prominent than the 85th percentile is taken into consideration as “overweight” whereas a BMI greater noteworthy than 95th percentile is considered as “obese”. The questionnaire is sent to the students through and is divided into three sections; personal information, physical interest and dietary. The Classification strategies are used viz. Bayesian classifiers, decision tree, neural network and Support Vector Machine (SVM). Weka is used to accumulate most fulfilling subset of attributes. It indicates the assessment of performance between four classifiers viz. Bayes Net, J48, Naïve Bayes, MLP and SMO. Based at the result, J48 and SMO appears to be the quality classifiers for predicting childhood obesity on these data sets.

In [5], the working of ID3 decision tree learning algorithm examined towards nominal attributes, continuous attributes and missing value attributes. The dataset is gathered from UCI Machine Learning Repository. Speculation is made that ID3 can indeed work properly on datasets with lacking attribute values to sure extent. In this paper set of rules has been implemented using java language. The experiments conducted conclude that ID3 works well on classification issues having datasets with nominal characteristic values.

In [6] three machine learning classification algorithms namely Decision Tree, SVM and Naive Bayes are used to detect diabetes at an early stage. The dataset is accrued from Pima Indians Diabetes Database (PIDD) sourced from UCI machine learning repository. This research work makes a speciality of pregnant women grief from diabetes. WEKA tool is used for performing the experiment which incorporates a set of various machine learning methods for data classification, clustering, regression, visualization etc. Accuracy, F-Measure, Recall, Precision and ROC (Receiver Operating Curve) measures are used for the classification of this work. As a result Naive Bayes gives better accuracy in respective to different classification algorithms.

In [7] the static and dynamic recording and assessment of children's growth and development are performed. The dataset is accrued dependent on children's gender, date of birth, weight, and height data under the usual of World Health Organization (WHO). The framework likewise records the child's growth within the manner of dribs and drabs, becoming a child's growth assistant and playmate. This application in the market can meet the parent’s desires to view child growth. This paper principally presents the design and construction of the system architecture, technology selection, layout and implementation of the database and the realization of the basic features of the Android

In [8], Automated Menu Planning Algorithm have been developed for Children. Dataset on medical statistics of child inclusive of underweight, normal or obese and records on activity stage of child is collected. In this nutritional management system the use of ID3 decision tree algorithm has been proposed. The decision tree learning algorithm ID3 works admirably on any classification issues having dataset with the discrete values. The proposed framework will be very effective for mothers to attend to her child’s health.

In [9], evaluation of logistic regression with six data mining strategies is achieved for predicting obese and obesity. The data has been recorded of child of 3 years at birth, 6 weeks, 8 months and a pair of years from UCI records. It has been accounted that among two to four-year olds, obesity has doubled. Six data mining strategies have been compared viz. Decision tree (C4.5), association rules, Neural Networks (NNs), Naïve Bayes, Bayesian networks and Support Vector Machines (SVMs). This paper concludes that SVM and Bayesian algorithms appear to be the best two algorithms for predicting overweight and obesity.

In [10], a model is designed that predicts the nutritional status of fewer than five children by using data mining techniques. The dataset is collected from Wirral infant Database. Malnourished children experience ill effects of successive ailments, which adversely impact their growth. Six trials have been led utilizing three data mining classification algorithms i.e. J48 algorithm, Naïve Bayes and PART rule induction classifier with a purpose to construct a model that predicts nutritional repute of under-five year’s children. Analyses have been implemented the use of Data understanding, Data preparation, Attribute selection, Selection of instances, Data transformation. As a result SVM and Bayesian algorithms appear to be the high-quality two algorithms for predicting overweight and obesity.

In [11], data mining strategies which include classification and prediction are used to predict the cardiac problems of patients based on the analysis from their symptoms as well as based on the risk factors. The goal is to analyze various data mining tools and techniques in health care domain that can be used in prediction of cardiac problems of patients. A system is built to gather data using Naive Bayes algorithm by developing a web based application.

Table 1. Survey on various related research

Paper	Purpose	Algorithm used	Data set used	Accuracy
[1]	To predict that the child is developing normally	Naïve Bayes ID3 Decision tree	Anganwadis, Parents, Health centres	High
[2]	To predict basic mental health problems using machine learning techniques.	AODEsr, Multi Layer Perceptron (MLP), RBF Network, IB1, KStar, Multi-Class Classifier (MCC), FT and LADTree.	Clinical Psychologist	High
[3]	This study conducts data mining focusing specifically on developmentally-delayed children.	C5.0 decision tree	Yunlin Developmental Delay Assessment Center.	Moderate

[4]	This study was to identify the factors that influence the childhood obesity using various feature selection techniques.	Bayesian classier, decision tree , neural network and Support Vector Machine (SVM), SMO MLP, J48.	(SEGAK) Assessment Program & the questionnaire.	Moderate
[5]	To analyze the working of ID3 algorithm with different attribute values.	ID3 Decision Tree	UCI Machine Learning Repository	Moderate
[6]	The motive of this study is to design a model which can predict the likelihood of diabetes in patients with maximum accuracy.	1. Decision Tree 2. SVM 3. Naive Bayes	Pima Indians Diabetes Database (PIDD) sourced from UCI machine learning repository	Moderate
[7]	The static and dynamic recording and evaluation of children's growth and development were performed.	—	Children's gender, date of birth, weight, and height data was taken from the standard of World Health Organization (WHO).	–
[8]	To give proper diet to children as per their profile, Dietary Management System using ID3 is proposed.	ID3 decision tree	Medical information, under-weight, normal, or overweight, Activity level of child,The food preference, Indian food database	Moderate
[9]	Many companies like credit card, insurance, bank, retail industry require direct marketing. Data mining can help those institutes to set marketing goal.	ID3 Decision tree Naïve Bayes classifier	UCI data	High
[10]	There is a growing epidemic of obesity affecting all age groups. Being fat as a child causes many diseases.	Decision trees (C4.5), association rules, Neural Networks (NNs), naïve Bayes, Bayesian networks and Support Vector Machines (SVMs)	Wirral child database	High

3. PROPOSED METHODOLOGY

The aim of the project is to study the child's growth data from 0-5 years and to model a classifier to predict whether a child's height and weight is accurate or not according to its age.

3.1 Tools and Technology

1) Datasets: The data is collected from the three locations:

a) Health centres

b) Pre-primary schools

c) Anganwadis

The dataset has 3 feature attributes: age, height, weight and used for the prediction.

2) Data Preprocessing: The collected data is pre-processed by using data pre-processing algorithms in python. First the data is categorized into dependent and independent variables. After this, missing values are replaced by its mean value. Then the data is divided into training and testing dataset.

3) Algorithms: ID3 Decision Tree supports decision making process and risk analysis. Algorithm iteratively divides attributes into two groups which are the most dominant attributes and to construct a tree.

ID3 Decision Tree formulae:

Calculate the Entropy of every attribute using the dataset S

$$\text{Entropy}(S) = \sum -p\ (I).\log_2 p\ (I)$$

Split the set S into subsets using the attribute for which the resulting entropy (after splitting) is minimum (Or equivalently information gain is maximum)

$\text{Gain}(S, A) = \text{Entropy}(S) - \sum [p\ (S|A).\text{Entropy}\ (S|A)]$

Make a decision tree node containing that attribute.

Recurse on subsets using remaining attributes.

Naïve bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. Every pair of features being classified is independent of each other. A naïve bayes classifier considers each of features to contribute independently to the probability.

Naïve bayes formulae:

$$\text{Mean} = (X1+X2+\ldots+Xn) / n$$

$$\text{Variance} = \sigma^2 = \sum_{i=1} = (X_i - \mu)^2 / n$$

Posterior probability is calculated.

Having better prediction rate, these two algorithms are chosen. By comparing both these techniques, more prominent algorithm can be detected for child growth analysis. These are evaluated on the basis of three criteria that is prediction accuracy, learning time and error rate.

The accuracy for ID3 Decision tree is 85% and for Naïve Bayes is 57%. From this experiment we can say that ID3 decision tree works well and gives better prediction than Naïve Bayes. It can work in any environment and gives better prediction rate.

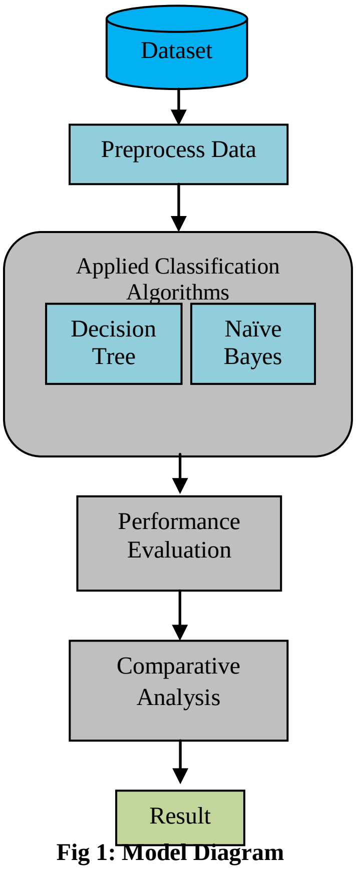

Fig 1: Model Diagram

4. CONCLUSION

In this research study, analysis is done on various data mining classification and prediction algorithms for child growth development. By better understanding how and why people change and grow, one can apply this knowledge to understand the needs of a child and fulfilling them and allow them to reach their full potential.

As compared to other algorithms, Naive Bayes and ID3 Decision Tree technique gives the maximum accuracy for predicting child development analysis. This research can be used in women and child department. It can be used in hospitals by doctors to advise parents how to take care of babies.

5. REFERENCES

[1] Ambili K, Afsar P, “A Prediction Model for Child Development Analysis using Naive Bayes and Decision Tree Fusion Technique – NB Tree” International Research Journal of Engineering and Technology (IRJET) July 2016 .

[2] Sumathi M.R, Dr. B. Poorna, “Prediction of Mental Health Problems Among Children Using Machine Learning Techniques”(IJACSA) International Journal of Advanced Computer Science and Applications, 2016.

[3] Taiwan ROC “A study of applying data mining to early intervention for developmentally-delayed children”.

[4] Fadzli Syed Abdullah1, Nor SaidahAbd Manan1, Aryati Ahmad1, Sharifah Wajihah Wafa1, MohdRazif Shahril1, Nurzaime Zulaily1, RahmahMohd Amin1, and Amran Ahmed2 “Data Mining Techniques for Classification of Childhood Obesity among Year 6 School Children” University Sultan ZainalAbidin, 21300 Kuala Terengganu.

[5] Anand Bahety “Extension and Evaluation of ID3 – Decision Tree Algorithm”.

[6] Deepti Sisodia and Dilip Singh Sisodia, “Prediction of Diabetes using Classification Algorithms.”

[7] “The Research and Development of Growth Curve for Children's Height and Weight on Android Platform.” 10th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics (CISP-BMEI 2017).

[8] “Automated Menu Planning Algorithm for Children: Food Recommendation by Dietary Management System using ID3 for Indian Food Database.”

[9] Masud Karim and Rashedur M. Rahman , “Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing” Journal of Software Engineering and Applications, 2013, 6, 196-206.

[10] Shaoyan Zhang, Christos Tjortjis, Xiaojun Zeng, Hong Qiao , Iain Buchan and John Keane “Comparing data mining methods with logistic regression in childhood obesity prediction” Published online: 24 February 2009 # Springer Science + Business Media, LLC 2009.

[11] V.Geetha and Dr.S.Rajalakshmi “A detailed analysis and comparison of decision tree vs naïve bayes algorithm in cardio vascular datasets” International Journal of Pure and Applied Mathematics Volume 119 No. 15 2018, 437-444.

[12] Saurabh Shastri, Paramji tKaur, Ankush Gupta, ShakshiS ambyal, Arun Singh Bhadwal, Amardeep Sharma, Professor Vibhakar Mansotra, Dr. Anand Sharma “Development of a Data Mining for Classification of Child Immunization Data” International Journal of Computational Engineering Research(IJCER)

[13] Ahmad Ansari, Iman Paryudi, A Min Tjao “Performance comparison between Naïve Bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool” International Journal of Advanced Computer Science and Applications(IJACSA)4,2013.

[14] William A Altemeier III, Peter M Vietze, Kathryn B Sherrod, Howard M Sandler, Susan Falsey, Susan O’Connor “Prediction of child maltreatment during pregnancy” Journal of the American Academy of Child Psychiatry 18(2)

[15] Muhamad Hariz B Muhamad Adnan, Wahidah Hussain, Faten Damanhoori “A survey on utilization of data mining for childhood obesity prediction” 8th Asia-

Pacific Symposium on Information and Telecommunication Technologies, 1-6

[16] T.M. Dugan, S. Mukhopadhyay, A. Carroll, S. Downs "Machine Learning Techniques for Prediction of Early Childhood Obesity"

[17] Leslie C Philipsen, Margaret R Burchinal, Carollee Howes, Debby Cryer "The prediction of process quality from structural features of child care" Early childhood research quarterly 12(3)

[18] Chung-Lang Chang "A study of applying data mining to early intervention for developmentally-delayed children" Expert Systems with Applications 33(2)

[19] Zenebe Markos "Predicting Under nutrition status of under-five children using data mining techniques: The Case of 2011 Ethiopian Demographic and Health Survey".

[20] Las Johansen B. Caluza "Machine Learning Algorithm Application in Predicting Children Mortality: A Model Development" International Journal of Information Science and Application.