Child Development Prediction Model Using Data Mining

Mugdha Amonkar, Juhi Bansod, Ambica Naik, Tina Vaz

[1]mugdhaamonkar29@gmail.com
[2]juhibansod9@gmail.com
[3]naikambica1998@gmail.com
[4]tinavaz171@gmail.com

Information Technology,
Shree Rayeshwar Institute of engineering & Information Technology
Shiroda, India

ABSTRACT

Child development is the field that includes the scientific investigation of the patterns of growth, change and strength that occur from conception through adolescence. By better seeing how and why individuals change and grow, one can apply this knowledge to comprehend the requirements of a child and satisfying them and permit them to arrive at their maximum capacity. The aim of this study is to analyze the child growth based on the features such as age, height and weight. In order to understand how the physical growth of child changes with time, a data mining approach is applied to predict the child growth. In this approach comparison of two data mining techniques are carried out that is ID3 Decision Tree and Naïve bayes classifier on the premise of prediction accuracy, error rate and learning time.

Keywords—Child Development, Data Mining, Machine Learning, Prediction, Data Processing, Naïve Bayes Classifier, ID3 Decision Tree

ARTICLE INFO

I. INTRODUCTION

Early childhood is the period when most of transitions in children take place; this period is a significant influence of child development as children progress into adolescence and adulthood. This period plays a significant role in shaping other aspects of childhood development, such as growth. India has the second highest number of obese children in the world with 14.4 million and its occurrence is increasing rapidly. In 2015 over 2 billion children across the world were overweight. One in four children under age 5 worldwide is as short as to be classified as stunted. Child stunning is a key marker of child malnutrition. Classification is independent of each other, it is easy and fast to predict. The quality of a child's earliest environments and the availability of appropriate experiences at the right stages of development are crucial determinants of the way each child's growth develops. This is a big problem in many places such as remote villages where the chances of children getting affected by obesity and stunted growth is very high and availability of expert pediatrician is a problem. The problem of availability of experts can be solved by applying a data mining approach to predict the child's future physical growth.

II. RELATED WORK

In this paper [1] a comparative study is done on Decision tree and Naïve bayes to predict that child is developing normally according to its age. The data is collected from schools, anganwadis, and parents by generating questionnaire. Various data preprocessing techniques have been performed on the data. Finally decision tree gives more accurate result than naïve bayes algorithm. By combining both algorithms NB tree module is build for accurate prediction.

In this paper [2] eight machine learning techniques are applied to predict the basic mental health problems. The dataset is collected from clinical psychologists. A comparison is made on those eight machine learning techniques and looked out for the leading three which can be put into practice to help mental health experts in diagnosing mental health problems. Five basic mental health issues of children viz., Attention problem, Academic Problem, Anxiety Problem, Attention Deficit Hyperactivity Disorder (ADHD) and Pervasive Developmental Disorder (PDD) have been considered. The techniques used are AODEsr, Multi Layer Perceptron (MLP), RBF Network, IB1, KStar, Multi-Class Classifier (MCC), FT and LAD Tree. Interview was taken of clinical psychologists to identify the mental health problems that occur among children. This model assists the professionals to spot the

problem if the known evidences of the patient are given as input. All the attributes are of nominal type. Date set is pre-processed by eliminating irrelevant and redundant attributes using Best First Search technique. The WEKA tool is used to compare the accuracy level of classifiers based on three measures Kappa Statistic, Accuracy and ROC Area. This paper concludes that Multilayer Perceptron, Multiclass Classifier and LAD Tree produce more accurate results than the others.

In this paper [3] C5.0 decision tree and Association rules are applied to discover out what level of delays might occur from which types of illnesses in children. Dataset is obtained from Yunlin Developmental Delay Assessment Centre. Abnormal and incomplete data is deleted. This study uses data mining tool C5.0 of Clementine 7.2 software developed by SPSS in combination with the a priori model. This research has identified which type of illness items will cause certain types of delays by building a decision tree and association rule analysis to determine the correlations among cognitive, language, motor, social emotional developmental delays.

In this paper [4] various feature selection techniques have been used for the classification of childhood obesity. The data is collected from Standard Kecergasan Fizikal Kebangsaan untuk Murid Sekolah Malaysia (SEGAK) Assessment Program and the study questionnaire on sociodemographic, physical activity and dietary assessment. The chart of BMI-for-age is utilized for reference, in which BMI less than the 5th percentile is considered as "underweight", the BMI more noteworthy than the 5th but less than the 85th percentile is considered as "normal", a BMI prominent than the 85th percentile is considered as "overweight" whereas a BMI more noteworthy than 95th percentile is considered as "obese". The questionnaire is distributed to the students through and is divided into three sections; personal information, physical activity and dietary. The Classification techniques are used viz. Bayesian classifiers, decision tree, neural network and Support Vector Machine (SVM). Weka is used to collect optimal subset of attributes. It shows the comparison of performance between four classifiers viz. Bayes Net, J48, Naïve Bayes, MLP and SMO. Based on the result, J48 and SMO appear to be the best classifiers for predicting childhood obesity on these data sets.

In this paper [5] the working of ID3 decision tree learning algorithm examined against nominal attributes, continuous attributes and missing value attributes. Speculation is made that ID3 can indeed work well on datasets with missing attribute values to certain extent. In this paper algorithm has been implemented using java language. The experiments conducted conclude that ID3 works well on classification problems having datasets with nominal attribute values. It also works well in case of missing attribute values but the way missing attributes are handled actually governs the performance of the algorithm. In case of neglecting instances with missing values for the attribute leads to high error rate compared to selecting the missing value as a separate value.

In this paper[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 collected from Pima Indians Diabetes Database (PIDD) sourced from UCI machine learning repository. This research work focuses on pregnant women suffering from diabetes. WEKA tool is used for performing the experiment which includes a collection 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 higher accuracy in respective to other classification algorithms.

In this paper [7] the static and dynamic recording and evaluation of children's growth and development are performed. The dataset is gathered dependent on children's gender, date of birth, weight, and height data under the standard of World Health Organization (WHO). The framework likewise records the child's growth in the process of dribs and drabs, becoming a child's growth assistant and playmate. This paper principally presents the design and construction of the system architecture, technology selection, design and implementation of the database and the realization of the basic functions of the Android. Technologies that are utilized in research are Java, Xml ,Android Software Development kit (SDK),Application programming Interface (API), Apache Http-The network communication uses this protocol client, SQlite-The Android client stores the user's data in this local database, JDBC.

In this paper [8] Automated Menu Planning Algorithm have been developed for Children. Dataset on medical information of child such as underweight, normal or overweight and data on activity level of child is collected. In this dietary management system using ID3 decision tree algorithm has been proposed. The decision tree learning algorithm ID3 works admirably on any classification problems having dataset with the discrete values. The proposed framework will be very advantageous for mothers to take care of her child's health.

In this paper [9] comparison of logistic regression with six data mining techniques is done for predicting overweight and obesity. The data has been recorded of children of three years at birth, 6 weeks, 8 months and 2 years. It has been accounted that among two to four-year olds, obesity has doubled. Six data mining techniques 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 this paper [10], a model is designed that predicts the nutritional status of under five children using data mining techniques. Malnourished children experience ill effects of successive ailments, which adversely affects their growth. Six trials have been led utilizing three data mining classification algorithms i.e. J48 algorithm, Naïve Bayes

and PART rule induction classifier in order to build a model that predicts nutritional status of under-five year's children. Analyses have been implemented using Data understanding, Data preparation, Attribute selection, Selection of instances, Data transformation. As a result SVM and Bayesian algorithms appear to be the best two algorithms for predicting overweight and obesity.

In this paper [11], data mining techniques such as 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 analyse 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. The gathered data is handled, fed into the database and classified based on data mining algorithm finally a prediction system is developed to find out the cardiac risk levels.

III. 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.

A. 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 attributes age, height, weight and used for the prediction.

2) *Data Preprocessing:* The collected data is preprocessed   by using data preprocessing 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

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

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

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

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:
$$Mean= (X1+X2+\ldots+Xn) / n$$

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

Posterior probability is calculated.

Having better prediction rate, these two algorithms are chosen for this project. 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.
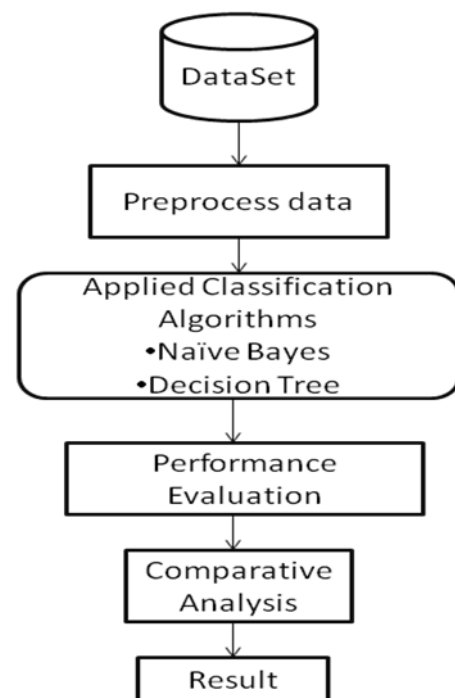
Model Diagram

IV. CONCLUSION

In this research study, analysis is done on various data mining classification and prediction algorithms for child growth development. Conclusion is made that compared to other algorithms Naive Bayes and ID3 Decision Tree technique gives the maximum accuracy for predicting child development analysis.

REFERENCES

[1]. Ambili K and 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. and 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 Saidah Abd Manan1, Aryati Ahmad1,Sharifah Wajihah Wafa1, Mohd Razif Shahril1, Nurzaime Zulaily1, Rahmah Mohd Amin1, and Amran Ahmed2 "Data Mining Techniques for Classification of Childhood Obesity Among Year 6 School Children" University Sultan Zainal Abidin, 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.Geetha1 and Dr.S.Rajalakshmi2 "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.