International Journal of Allied Practice, Research and Review
Website: www.ijaprr.com (ISSN 2350-1294)

A Comparative Study of Data Mining Techniques used for Determining Child Nutrition Status

Pooja Kadam[1] and Dr. Emmanuel M.[2]
[1]ME Student, Department of Information Technology, PICT,Pune, India
[2]Professor, Department of Information Technology, PICT, Pune, India

***Abstract*—Nutrition is a process by which organisms use food they consume normally through digestion, absorption, transportation, storage, metabolism and excretion to maintain life, growth and normal functions of the organs, as well as producing energy. So, nutrition plays a vital role in sustenance of life. Hence, early diagnosis of nutrition status will help enable us to improve the health levels of children. For this purpose, several researches have been carried out to classify nutrition status using different data mining techniques. In most of these studies, nutrition status was classified using anthropometric indices, Height-for-Age, Weight-for-Age and Weight-for-Height. Dataset used in such studies is mainly Demographic and Health Survey dataset. In this paper we are going to analyze the models built using techniques like Naïve Bayes, K-means clustering and Artificial Neural Networks for comparing child nutrition status. Finally, algorithms like Naïve Bayes, J48 Decision Tree and Multilayer Perceptron were applied to the Maharashtra data which was extracted from the Indian Demographic and Health Survey Dataset (IDHS) in order to determine their performance. Performance was determined with the help of accuracy parameters like Mean Absolute Error, Root Mean Squared Error and Time Taken to build the model.**

Keywords—Data mining, Demographic, Naïve Bayes, Decision tree, Anthropometry

I. Introduction

Nutrition can be measured with the help of indicators like anthropometry, biochemical tests, clinical signs and assessment of dietary intake. The anthropometric measurements method is used to compare various physical measurements like height, weight, mid-upper arm circumference, and body mass index (BMI) to those of a reference population. The World Health Organization had updated its international child growth standards in 2006. These

standards were first published in 1970 after a six-year study on 8,000 children in Ghana, Brazil, India, Norway, Oman, and the United States. The standards included indicators such as weight-for-age, height-for-age, weight-for-height, BMI, and key motor development milestones such as sitting, standing, and walking [1]. India adopted the World Health Organization's child growth measures in 2008. Before these measures were adopted in 2006, almost half of India's under-5children were stunted, or too short for their age; 20% were wasted, or too thin for height; and 43% were underweight, with some states, including Madhya Pradesh, Jharkhand, and Bihar, faring worse than others in terms of underweight children. Children in rural areas were more likely to be undernourished; yet even in urban areas, one-third of children were underweight [2].

According to the Swasth Report of Maharashtra, 2019 the state has been recognised for its improved healthcare facilities which is evident from its reduced infant and child mortality rates, but it has not been able to improve the status of malnutrition which is a major contributor to death among children under five in India [3]. This is keeping the state from achieving the Sustainable Development Goal (SDG-2) which aims to put an end to hunger and malnutrition [3]. According to experts, children can reach their full growth and development potential only when they are given appropriate nutrition and adequate care during the first five years. Children become vulnerable to poor growth and development if there is lack of proper nutrition. According to National Family Health Survey (NFHS 4), the prevalence of stunting (i.e. low height for age) among children under the age of five has been reduced. However, the state has not been able to fight wasting (i.e. low weight for height) and underweight problem among children below five years of age who were covered in the survey [3]. This is an indication of the failure in catering to the nutritional needs of the children in the state who were covered in the survey. The Department of Women and Child Development has been implementing Integrated Child Development Services (ICDS) for more than three decades and has been providing mid-day meals to about 82 per cent of the total number of children enrolled in public schools, to tackle malnutrition. For creating awareness about the importance of nutrition, programs like Village Health and Nutrition Day (VHND) is observed once a month in every village. ASHA (Accredited Social Health Activists) is being deployed for this. Despite of all these measures, the goal to eradicate malnutrition remains unfulfilled [3]. Therefore, there is a need to create an infrastructure that will help in providing better nutrition for children.

Machine learning techniques can be used to determine the nutrition status of children using different sources of data. With the help of Machine Learning, we can build computer systems having the ability to adapt, learn and simultaneously improve their performance through experience [4]. Data mining uses these machine learning techniques, mathematical functions, and statistical analysis to extract potentially useful knowledge that was previously unknown [4]. Classification is a data mining technique used to classify each item in a set of data into one of a predefined set of classes or groups. Classification makes use of mathematical techniques such as linear programming, decision trees, statistics and neural networks [4]. In classification, we build the software that is able to learn how to classify the data items into groups. Knowledge Discovery in Database (KDD) is a field encompassing theories, methods, and techniques that maps low-level data which is often too voluminous to be understood into high-level knowledge that is more compact and useful [4]. Data mining is a key step in the KDD process. It is defined as an iterative multistep process consisting of selection, preprocessing, transformation, data mining, interpretation and evaluation, and knowledge representation [4].

Determining the nutrition status among children below 5 years is a very challenging task. This study comes within the framework of nutrition evaluation among children and its main goal is to apply classification algorithms to predict nutrition status for children below five years and also compare these techniques. The tool employed in this study was the Waikato Environment for Knowledge Analysis (WEKA) Workbench since it allows to compare different machine learning solutions. The IDHS 2015-16 children data was used for the analysis. Classification is

done mainly into three categories i.e. stunted, underweight and wasted using anthropometric indices height-for-age, weight-for-age and weight-for-height respectively. The tasks involved in the study included data selection, anthropometric analysis, feature extraction, data preprocessing, data transformation and finally application of several classifiers and their evaluation through performance measures like the number of correctly classified instances, mean absolute error, root mean squared error and time taken to build the model. The main goal of this study was to apply different data mining techniques in WEKA environment to extract useful information from data and identify a suitable algorithm for generating an accurate predictive model to predict nutrition status of children below five years.

II. Related Work

The machine learning approach has been extensively used in the studies related to nutrition. Studies involve determining nutrition status, finding out the factors explaining malnutrition in children and comparing algorithms used for classification. The different techniques explored in the past studies include K-means clustering, Naïve Bayes, Hierarchical Agglomerative Clustering and Artificial Neural Networks. Some of the papers that used the above techniques are discussed below. Sri Winiarti et.al used k-means clustering to identify malnutrition status in toddlers based on the context data from community health center (PUSKESMAS) in Jogjakarta, Indonesia [5]. The study done by Riris Aulya Putri et.al in [6] classified toddlers's nutrition status based on 3 anthropometry indices with the help of Naïve Bayes algorithm. Reyhan Gupta et.al in [7] compared results of two clustering algorithms to better understand the malnutrition factors in Bihar. In the study unstructured data from Government reports, pertaining to malnutrition indices, demographic, social and medical factors that cause malnutrition was broken into clusters with the help of Rapid Miner Studio using k-means and Hierarchical agglomerative clustering. Rizky Ade Putranto et.al in [8] built a faster and more efficient system using the combination of Cloud Computing (CC) technology with the Certainty Factor (CF) and Forward Chaining expert system to solve the problem of monitoring infant nutritional status based on the anthropometric index in the postpartum period. The study done by Cynthia Hayat et.al in [9] involved the development of Artificial Neural Network (ANN) model to identify the types of malnutrition which gave 96% accuracy rate with MSE of 0.000997 during 5 seconds training period. Lastly an ANN approach was applied by Md Mehrab Shahriar et.al in [10] to Bangladesh Demographic and Health Survey 2014 (BDHS) children data to build a predictive model that classifies the malnutrition condition. The resulting model gave best accuracy for wasting, underweight, and stunting. In this paper we will discuss the techniques that were used in the above studies and then compare three algorithms i.e. Naïve Bayes, J48 Decision Tree and Multilayer Perceptron with the help of WEKA data mining tool.

III. Algorithms and Techniques used in Related Literature

3.1 Identification of Toddler's Nutritional Status using Data Mining Approach [1]

In this study, the patient's data that cannot directly map into appropriate groups of malnutrition status was mapped into several malnutrition status categories by using the k-means clustering data mining concept. K-means is the most commonly used algorithm for grouping objects by attributes into k number of clusters, where k is a positive integer defined by the user [5]. The nutrition report data from PUSKESMAS Umbulharjo Yogyakarta in 2016 was used for the study. The data included 6 months to 72 months old infant's data. Parameters like height, weight and age were used for grouping. Grouping is done by minimizing the sum of squares distance between the data and the appropriate cluster centroid. The k-means algorithm works by using eq (1) below:

$$arg_s min \sum_{i=1}^{k} \sum_{Xj \ni Si} \left\| X_j - \mu_i \right\|^2 \qquad (1)$$

Where,

- (X1, X2 …Xn): the observation results represent a cluster element with a real d dimensional vector.
- n: Number of observations where the observed value belongs to k set (k <= n) S = (S1, S2, … Sk).
- μi: the mean value of the point at Si

The system classifies nutrition status into 5 categories i.e. good nutrition, moderate nutrition, malnutrition, more nutrition and obesity based on 5 cluster centroids. The system is tested using cross validation method by comparing the manually calculated results with the results of the developed system. The system could determine nutritional status of toddlers with 90% accuracy.

3.2 Classification of Toddler Nutrition Status with Anthropometry Calculation using Naive Bayes Algorithm [2]

In this study, Naïve Bayes classification technique was used which is considered as one of the most popular classification algorithms and is among the 10 best algorithms for data mining. The purpose of this research was to classify the toddler's nutrition status into categories based on 3 anthropometric indices and they are: WFA for the malnutrition status, lack of nutrition, having decent nutrition and having excessive nutrition, HFA for very short nutrition, short, normal, and high, WFH for very thin status, thin, normal and fat [6].Survey data from Desa Tunjungtirto's Posyandu (Pos Pelayanan Terpadu/Integrated Service Post) was used in the study. The classification was done in two processes, they are training and testing process. The training process involves calculation of probability value for non-numeric data and mean & standard deviation calculation for numeric data. With the Naïve Bayes technique, probabilities and statistic method are used to predict a chance in the future based on the past experience. We get value of the probability from the frequency calculations and the combination of value based on the existing data. The equation used is:

$$P(C|X_1, \ldots., X_n) = P(C) \prod_{i=1}^{n} P(Xi|C) \qquad (2)$$

According to the Naïve Bayes theory, if the i-attribute is discrete, $P(X_i|C)$estimate is relative frequent from the sample having X_i as the i-attribute in a C class. On the other hand, if the i-attribute is continuous or numeric, $P(X_i|C)$ could be estimated using the function of Densitas Gauss in eq(3).

$$f(x) = \frac{1}{\sqrt{2\Pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad (3)$$

After we count based on Densitas Gauss, we can search the maximum probability constitute from the result of classification class. The classification was tested with k-fold cross validation method and results of the test was evaluated for Accuracy and Error Rate.WFA index gave the best accuracy value of 88%, the WFH index gave an accuracy value of 68% and finally HFA gave the lowest accuracy value of 64%. Also, WFA index gave lowest error rate of 12%, then WFH gave 32%, and HFA gave the highest error rate 36%.

3.3 Comparative Study of Clustering Algorithms by Conducting a District Level Analysis of Malnutrition [3]

Here, a comparison of K-Means and Hierarchical Agglomerative Clustering was done to better understand the malnutrition factors in Bihar. The NFHS 4 data released by the Government of India in 2017 which gives the district level details of many determinants of malnutrition was used for the study. With k-means algorithm we associate each data point with a cluster, where k different clusters are created from a data set that has n total objects such that k<=n. To find the optimal k value, the algorithms were run several times and the silhouette index was noted for each run. The value of k was selected when the silhouette index showed the maximum value corresponding to a particular k value. The silhouette index can be found out by considering the mean of the intra-cluster distance (a1) and the mean of the distance between the nearest clusters (b1) [7] given by the eq (4) below:

$$\text{Silhoutte Index} = (b1-a1)/\max(a1,b1) \qquad (4)$$

The hierarchical agglomerative algorithm takes the help of a bottom up procedure that creates sub clusters and creates more till termination point where only a single cluster remains.

In the first stage of the study, variables from non-clustered data were correlated with malnutrition indices. Further, Rapid Miner Studio was used to cluster data using k-means and hierarchical agglomerative clustering. Subsequently, each cluster was again analysed using the software and the correlation results were compared. Correlation is found using Pearson's correlation coefficient to determine how strongly they are correlated with dependent variables (Underweight, Wasting, Severe Wasting and Stunting). Significant variation was observed in most of the correlations in the data sets obtained by executing the two algorithms. A comparative study between k-means and hierarchical agglomerative clustering showed that the relevancy of hierarchical method was more than k-means when it came to deducing clusters pertaining to malnutrition in Bihar.

3.4 Cloud Computing Medical Record Related Baby Nutrition Status Anthropometry Index During Postpartum [4]

This system used Cloud Computing (CC) technology for quick classification of nutritional status of infants during postpartum. The combination of CC technology with the Certainty Factor (CF) and Forward Chaining (FC) expert system were used in monitoring infant nutritional status based on the anthropometric index. The CF shows the level of trust and the level of distrust in the facts of the data. In this system, the CF value ranges from -1 to 1. The -1 value indicates absolute distrust while the value 1 indicates absolute confidence. The cloud system consists of a Service App medical record data system, cloud database, and the App Service plan in a cloud server location. Data from Bhakti Bunda clinic healthcare data was used as a sample of the population of health services for all cities in Central Java Province for the period of 2017-2018. This data included maternal biodata data, baby biodata, and midwife biodata. The anthropometric indices, body weight according to age (Weight / Age), body height according to age (Height / Age) and body weight according to body height (Weight / Age) were used for classification. The baby's nutrition is determined using the Z-Score value for each anthropometric index which is calculated using eq(5) below:

$$\text{Z-Score} = \frac{Individual\ Value\ Subject - Median\ Standard\ Reference}{Standard\ Reference\ Deviation\ Value} \qquad (5)$$

The functions in the anthropometry index were translated into the PHP programming language. Rule based approach was used in modeling the baby nutritional status category. Each rule is defined as follows:

IF Weight_Age is X AND Height_Age is Y

THEN Weight_Height is Z

Data parameters were entered into the database which are then processed by the system which consists of cloud computing in which there are criteria and rule-based indicators to produce output in the form of a web dashboard consisting of baby nutrition status, baby medical record data, analysis of variables that has the most influence on the nutritional status of infants [8]. Infant nutrition classification for 108 infant patients, showed that babies experienced good nutrition status of 89% based on Weight / Age indicator, normal nutrition status of 53.2% based on Height / Age indicator and very thin nutrition status of 51.4% based on Weight / Height indicator.

3.5 The Modelling of Artificial Neural Network of Early Diagnosis for Malnutrition with Backpropagation Method [5]

The study involved the development of ANN model to identify the types of malnutrition. In the training phase ANN weight was generated using feed-forward activation function, and in the testing phase the result of the previous stage was tested to obtain output. Primary data from 14 Community Health Centers located in Nusaniwe and Sirimau Regencies, Ambon, Maluku was used here. The secondary data was obtained by conducting interviews with nutritionists, pediatricians, and parents and through observation conducted in 14 Community Health Centers. Initial stage involves preprocessing the data for the model followed by a rule-based system which was designed using if-then rules as existed in the expert system. Then, parameters and the activation function (binary sigmoid 'logsig') for the BPNN model were determined. The trial and error method were used to determine the number of hidden layer neurons. Finally, the weight of ANN generated from the training phase was implemented into the testing phase by inputting the symptoms experienced [9]. The accuracy rate of the resulting architectural model was 96% with MSE of 0.000997 during 5 seconds training period. Regression results show that the resulting model had a high degree of accuracy to produce output for malnutrition types such as marasmus, kwashiorkor, and marasmus-kwashiorkor.

3.6 A Deep Learning Approach to Predict Malnutrition Status of 0-59 Month's Older Children in Bangladesh [6]

In this study an ANN approach was applied to Bangladesh Demographic and Health Survey 2014 (BDHS) children data. Classification of nutrition level was done on the basis of Z-scores for 3 anthropometric indices which were calculated based on the 2006 Child Growth Standard defined by WHO. With reference to Primary Sampling Unit (PSU), 16 most significant features were extracted by literature review among the 1,742 features in BDHS 2014. Data was cleaned using mean imputation and changed to numeric values from string values using python's 'numpy' library. The pre-processed data was now ready to fit the deep learning model. The popular library, '*Keras'*, which runs *TensorFlow* at its backend was used to build the model to train the factors. The model used 16 neurons as features in the input layer, 8 in the hidden layer with three hidden layers which were tested one by one. A rectifier activation function was applied in hidden layers and sigmoid activation function in the output layer [10]. The probability of being malnourished was 1 and nourished was 0. 90% of the data was taken as training data to fit a model which runs 100-500 epochs. The best result was shown by ANN with accuracy close to 86.0% for wasting, 70.0% for underweight, and 67.30% for stunting. These tests have been done with 10% of the data using k-fold(k=10) cross-validation.

IV.COMPARISON OF TECHNIQUES

A brief comparison of all the above discussed techniques is done in Table 4.1. The table mentions the characteristics and limitations of the different techniques used so far.

S. No.	Name of Paper	Author	Year of Publication	Algorithm Used	Dataset	Advantages	Disadvantages
1	Identification of Toddlers' Nutritional Status using Data Mining Approach [1]	Sri Winiarti, Herman Yuliansyah, Aprial Andi Purnama	International Journal of Advanced Computer Science and Applications, 2018	K-means Clustering	Data from community health center (PUSKESMAS) in Jogjakarta, Indonesia	System gave accuracy of 90% when compared to the manually calculated results.	The initial centroids have a strong impact on the final results.
2	Classification of Toddler Nutrition Status with Anthropometry Calculation using Naive Bayes Algorithm [2]	Riris Aulya Putri, Siti Sendari &Triyanna Widiyaningtyas	IEEE 2018 International Conference on Sustainable Information Engineering & Technology (SIET)	Naive Bayes Algorithm	Survey data from Desa Tunjungtirto's Posyandu (Pos Pelayanan Terpadu/Integrated Service Post)	Naive Bayes algorithm has both high accuracy and speed on various classification problems.	Naive Bayes classifier needs to make a strong assumption on the shape of your data distribution.
3	Comparative Study of Clustering Algorithms by Conducting a District Level Analysis of Malnutrition [3]	Reyhan Gupta, Abhishek Singhal &A. Sai Sabitha	2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)	K-Means &Hierarchical Agglomerative Clustering	2017 NFHS 4 dataset	The clustered data provided better correlation between parameters as compared to the non-clustered data.	It is important to segregate the dataset into clusters having relatively similar attributes becauseextreme values present as outliers do not give correct interpretations.
4	Cloud Computing Medical Record Related Baby Nutrition Status Anthropometry Index During Postpartum [4]	Rizky Ade Putranto, Suryono Suryono & Jatmiko Endro Suseno	IEEE 2018 2nd Intenational Conference on Informatics and Computational Sciences (ICICoS)	Cloud Computing & Rule Based Indicators	Sample data of the population of health services for all cities in Central Java Province (2017-2018) obtained from Bhakti Bunda clinichealthcare	This system is faster & more efficient in reading the monitoring area for classification of infant nutritional status.	The time taken to formulate and implement new rules requires substantial time.
5	The Modeling of Artificial Neural Network of Early Diagnosis for Malnutrition with Backpropagation Method [5]	Cynthia Hayat & Barens Abian	2018 Third International Conference on Informatics and Computing (ICIC)	Backpropagation neural network (BPNN).	Primary data from 14 Community Health Centers located in Nusaniwe & Sirimau Regencies, Ambon, Maluku	Resulting architecture is simple, easy to build with good accuracy and error rates.	1.It has poor generalization i.e. Lack of program design that is rigorous to the theory. 2. It is difficult to control the training process.
6	A Deep Learning Approach to Predict Malnutrition Status of 0-59 Month's Older Children in Bangladesh [6]	Md Mehrab Shahriar, Mirza Shaheen Iqubal, Samrat Mitra & Amit Kumar Das	The 2019 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)	Artificial Neural Network using Tensorflow	2014 Bangladesh-DHS dataset	This approach is the most scientific way with best accuracy both for policymakers and clinicians	Need lot of data, especially for architectures with many layers.

Table 4.1: Comparison of Techniques

V. Proposed Methodology

In this study we are using the Knowledge Discovery Process (KDP) to build a predictive model using data mining techniques. KDP is a process of finding knowledge in data with the help of data mining methods (algorithms). For the study purpose, the WEKA data mining tool was used. The complete flow of the steps involved in our study is shown in Figure5.1.

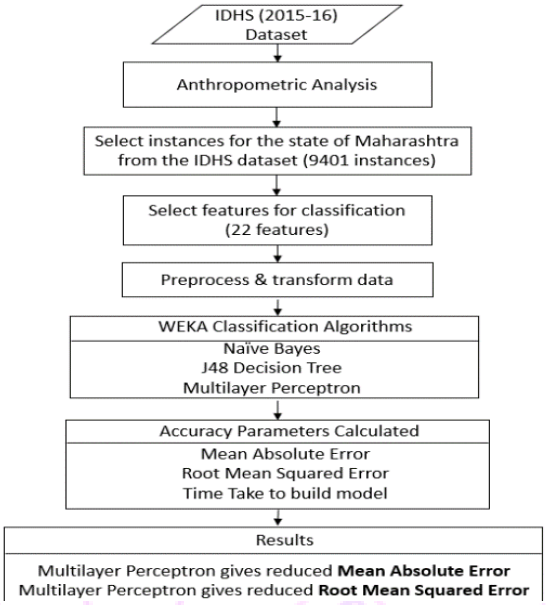

Figure5.1: Proposed methodology

Algorithms like Naive Bayes, J48 classifiers and Multilayer perceptron were used for the classification of data set. The performance of these classifiers was analysed with the help of accuracy parameters like Correctly Classified Instances, Mean Absolute Error, Root Mean-Squared Error and Time Taken to build the model and the results are shown statistical and graphically.

5.1 Data Source

The source data employed for the study is 2015-16 IDHS dataset. This census is conducted in every five years interval.The DHS dataset consists of data related to birth record, children record, couples record, HIV test record, household record, household member record and individual record across all states. In this project we will be using data related to children record.This Children's Data dataset (Children's Recode - KR) has one record for every child of interviewed women, born in the five years preceding the survey. The dataset consisted of a total of 245513 instances together with 1340 features (factors). In the study we would be using 9401 instances which are the instances for the state of Maharashtra and 13 attributes which have been selected after feature extraction and thorough literature review.

5.2 Anthropometric Analysis

Height- and weight-based anthropometric indicators are used worldwide to characterize the nutrition status of populations. The WHO has indicated that the standard deviation (SD) of *Z*-scores of these anthropometric indicators is relatively constant across populations, irrespective of nutritional status according to the 1978 WHO/National Center for Health

Statistics growth reference. Therefore, the SD of *Z*-scores can be used as an indicator for anthropometric data. In 2006, WHO published new growth standards [11]. The aim of this study is to classify nutrition status based on SD of height- and weight- based *Z*-score indicators from the 2006 WHO growth standards. The classification of data is done into three classes i.e. stunted, underweight and wasted using parameters HW70, HW71 and HW72 (HAZ, WAZ and WHZ) respectively from the IDHS dataset. Classification is done on the basis of below categories mentioned in the 'Guide to DHS Statistics DHS-7' [12]. Finally, we classify data into four labels, i.e. HAZ below -200 as Stunted, WAZ below -200 as Underweight, WHZ below -200 as Wasted. Table 5.1 below gives classification based in Anthropometry.

FEATURE VALUE	CATEGORY
HW70 < -300	Severely stunted
HW70 < -200	Moderately or severely stunted
HW71 < -300	Severely underweight
HW71 < -200	Moderately underweight
HW71 > 200 & HW71 < 9990	Overweight
HW72 < -300	Severely wasted
HW72 < -200	Moderately or severely wasted
HW72 > 200 & HW72 < 9990	Overweight

Table 5.1: Classification based on Anthropometry

The equation of z-score (Z) calculation is :

$$Z = \{(X/M)^*L - 1\}/(L^*S) \qquad (6)$$

[Where, X= weight/height/BMI, M, L and S are the age-specific values of appropriate table corresponding reference populations]. The WHO has developed a software Anthro to determine the nutritional status of children, which is very convenient to the users. Using, this tool you will able to calculate your desired anthropometric indices of height-for-age, weight-for-age and weight-for-height very easily among children 0-5 years.

5.3 Feature Extraction

Generally, as some of the features do not have any impact on malnutrition, labelling of data and feature extraction was a very challenging task. Out of a total of 1340 features, 13 most significant features were selected on the basis of thorough literature review. We also made use of the WEKA 3.9.3, Ranker method and GainRationAttributeEval attribute evaluator in order to rank the attributes for their significant impact on malnutrition. The attributes that were selected are given in the Table 5.2 below.

SELECTED ATTRIBUTES
1. Child Sex
2. Child Age In Months
3. No of Living Children
4. Wealth Index
5. Type Of Residence
6. Ever Had Vaccination
7. Mother's Current Age
8. Mother's BMI
9. Mother's Highest Educational Level
10. Mother's Occupation
11. Height Age Standard Deviation
12. Weight Age Standard Deviation

Table 5.2: Selected attributes

5.4 Data Preprocessing

At this stage, the inconsistent data is handled. Here issues such as missing and duplicate data are resolved, and the data is then moved to transformation process.

5.5 Data Transformation

In this stage, discretization is used for converting continuous valued variables to discrete values where limited numbers of labels are used to represent the original variables. Discretization is done on values of standard deviation of three anthropometric indices (HAZ, WAZ and WHZ) based on 2006 WHO Growth Multicenter standards.

Discrediting the values of HAZ attribute: HAZ attribute discredited categories are <-200 (Stunted), -200 to 200 (Normal) and >200 (NoClass)

Discrediting the values of WAZ attribute: WAZ attribute discredited categories are <-200 (Underweight), -200 to 200 (Normal) and >200 (NoClass).

Discrediting the values of WHZ attribute: WHZ attribute discretized categories are <-200 (Wasted), -200 to 200 (Normal) and >200 (NoClass).

5.6 Data Mining using WEKA

The WEKA 3.9.3 toolkit was used to analyse the dataset with the data mining algorithms. WEKA is a collection of tools of data classification, regression, clustering, association rules and visualization. The toolkit is developed in Java and is open source software issued under the GNU General public License [13].

5.7 Experimentation

Experimentation with Naïve Bayes Algorithm: Naïve Bayes is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the existence of a particular feature in a class is not related to the existence of any other feature. Naive Bayes model is easy to build and

especially useful for huge data sets. It is simple and outperforms other highly sophisticated classification methods. With this theorem, we can calculate posterior probability P(c|x) from P(c), P(x) and P(x|c). The equation is given below:

$$P(c|x) = \frac{P(c|x)P(c)}{P(x)} \qquad (7)$$

Where $P(C|X) = P(x_1|c) * P(x_2|c) * ... * P(x_3|c) * P(c)$

P(c|x) = posterior probability of class (c, target) for agiven predictor (x, attributes).

P(c) = prior probability of class.

P(x|c) = likelihood which is the probability of thepredictor given class.

P(x) = the prior probability of predictor.

Experimentation with J48 Algorithm: J48 is known as the optimized implementation of the C4.5. Its output is a decision tree. A decision tree is used to divide the input space of a data set into mutually exclusive areas, with each area having a label, a value or an action that describe its data points. Splitting criterion is used to calculate which attribute is the best to split that portion of the tree of the training data that reaches a particular node. One of the Decision trees for the parameters Child Sex, Child Age in Months, Mother's Highest Educational Level for Class Stunted from the data set is shown in Figure5.2.

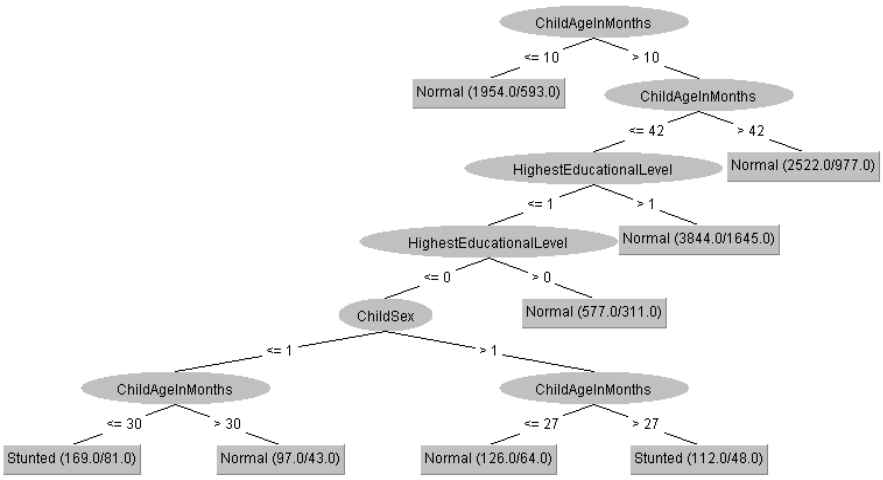

Figure5.2: Decision tree built by J48 algorithm in WEKA

Experimentation with Multi-Layer Perceptron Algorithm:A Multi-Layer perceptron (MLP) is a feedforward neural network with one or more hidden layers between input and output layer. Each neuron in each layer is connected to every neuron in the adjacent layer. The training or testing vectors are fed to the input layer and then processed by the hidden and output layers. The final hidden layer is connected to the output layer. Figure5.3. shows the multilayer perceptron model built using WEKA. It has 10 neurons at the input layer for 10 input parameters. The single hidden layer has 6 neurons each and the output layer has 3 neurons for the class labels (Stunted, Normal and NoClass).

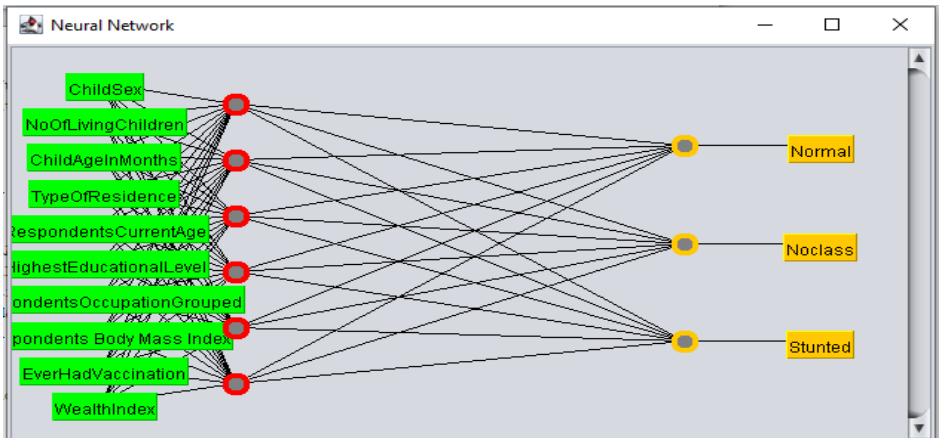

Figure5.3: Neural Network built using WEKA

We ran each algorithm for the dataset three times each for classes stunted, underweight and wasted respectively. The results of running the algorithms is given in Table 5.3, Table 5.4 and Table 5.5 below:

Algorithm (Total Instances: 9401)	Correctly Classified Instances % (value)	Incorrectly Classified Instances % (value)	Time Taken (seconds)	Kappa Statistic	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Squared Error (%)
Naïve Bayes	59.0469 %	40.9531 %	0.05	0.1033	0.3429	0.422	94.6117 %	99.1399 %
J48	55.3133 %	44.6867 %	1.98	0.0865	0.3405	0.4746	93.9322 %	111.4792 %
Multilayer Perceptron	59.2916 %	40.7084 %	11.03	0.0576	0.3372	0.4178	93.03 %	98.142 %

Table 5.3: Results of classifiers for class Stunted (HAZ)

Algorithm (Total Instances: 9401)	Correctly Classified Instances % (value)	Incorrectly Classified Instances % (value)	Time Taken (seconds)	Kappa Statistic	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Squared Error (%)
Naïve Bayes	61.3764 %	38.6236 %	0.02	0.1503	0.3204	0.4096	92.9641 %	98.6663 %
J48	59.7171 %	40.2829 %	0.62	0.1271	0.3184	0.4508	92.37 %	108.5936 %
Multilayer Perceptron	61.7488 %	38.2512 %	10.41	0.0773	0.3171	0.405	91.987 %	97.5669 %

Table 5.4: Results of classifiers for class Underweight (WAZ)

Algorithm (Total Instances: 9401)	Correctly Classified Instances % (value)	Incorrectly Classified Instances % (value)	Time Taken (seconds)	Kappa Statistic	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Squared Error (%)
Naïve Bayes	70.301 %	29.699 %	0.02	0.001	0.3	0.3872	99.9553 %	99.9469 %
J48	68.865 %	31.135 %	0.55	0.043	0.2921	0.4001	97.3152 %	103.2867 %
Multilayer Perceptron	70.4287 %	29.5713 %	10.22	0.0005	0.291	0.3869	96.9417 %	99.8699 %

Table 5.5: Results of classifiers for class Wasted (WHZ)

VI. Result Analysis and Comparison

In this paper, the following parameters were used to evaluate the performance of the above-mentioned classification techniques:

Correctly Classified Instances: This parameter gives the percentage of the correctly classified instances by the classifier.

Mean Absolute Error (MAE): Mean Absolute Error is a basic accuracy parameter which calculates the average magnitude of the errors of the forecasting results. In statistics, the MAE determines how close the forecasts are to the eventual outcomes. The MAE is given by eq(8) below:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|fi - yi| = \frac{1}{n}\sum_{i=1}^{n}|ei| \qquad (8)$$

Where f_i refers to the prediction and y_i refers to the true value

Root Mean-Squared Error (RMSE): The RMSE is the difference between forecast and corresponding observed values which are squared and then averaged over the sample. Then, we take the square root of the calculated average. The RMSE for a set of n values $\{x_1,x_2,\ldots,x_n\}$ is given by eq(9) below:

$$\text{Xrms} = \sqrt{\frac{1}{n}(x_1^2 + x_2^2 + \cdots + x_n^2)} \qquad (9)$$

Time: This parameter is the amount of time required to build the model.

These accuracy parameters are measured for the above algorithms for classes stunted (HAZ), underweight (WAZ) and wasted (WHZ). The comparison of the results is shown in Figure 6.1, Figure 6.2, Figure 6.3 and Figure 6.4 below:

Algorithm (Total Instances: 9401)	Stunted	Underweight	Wasted
Naïve Bayes	59.05%	61.38%	70.30%
J48	55.31%	59.72%	68.87%
Multilayer Perceptron	59.29%	61.75%	70.43%

Table 6.1: Correctly Classified Instances

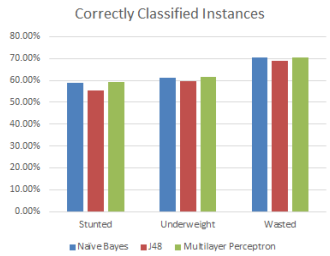

Figure6.1: Comparison of algorithms for Correctly Classified Instances

Algorithm (Total Instances: 9401)	Stunted	Underweight	Wasted
Naïve Bayes	0.3429	0.3204	0.3
J48	0.3405	0.3184	0.2921
Multilayer Perceptron	0.3372	0.3171	0.291

Table 6.2: Mean Absolute Error

Figure 6.2: Comparison of algorithms for Mean Absolute Error

Algorithm (Total Instances: 9401)	Stunted	Underweight	Wasted
Naïve Bayes	0.422	0.4096	0.3872
J48	0.4746	0.4508	0.4001
Multilayer Perceptron	0.4178	0.405	0.3869

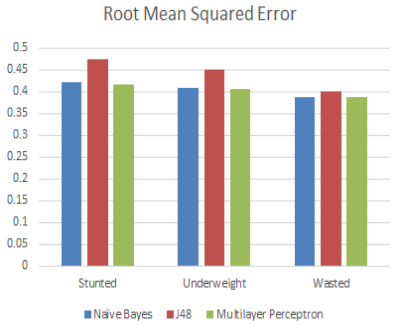

Table 6.3: Root Mean Squared ErrorFigure 6.3: Comparison of algorithms for Root Mean Squared Error

Algorithm (Total Instances: 9401)	Stunted	Underweight	Wasted
Naïve Bayes	0.05	0.02	0.02
J48	1.98	0.62	0.55
Multilayer Perceptron	11.03	10.41	10.22

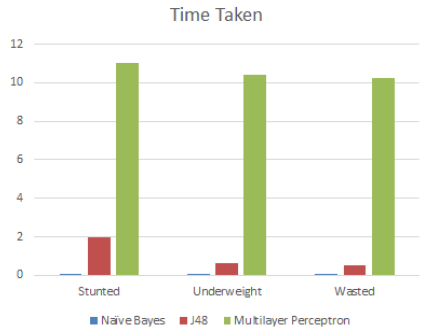

Table 6.4: Time Taken **Figure 6.4: Comparison of algorithms for Time Taken**

From the Accuracy parameters it can be observed that Multilayer Perceptron model performance is better than Naïve Bayes and J48 Decision Tree. The Multilayer Perceptron gives the highest correctly classified instances of 59.29 % for class Stunted (HAZ), 61.75 % for class Underweight (WAZ) and 70.43 % for class Wasted (WHZ). It also gives reduced values for Mean Absolute Error i.e. 0.3372, 0.3171 and 0.291 and Root Mean Squared Error i.e. 0.4178, 0.405 and 0.3869 for classes Stunted, Underweight and Wasted respectively. Among the three algorithms, the model built using Naïve Bayes takes the least time. Based on these results we can classify child nutrition status by selecting the algorithm that gives the most appropriate results in terms of the above accuracy parameters that will be relevant to our study.

VII. Conclusion

In this paper, three different classifiers (Naïve Bayes, Multilayer Perceptron and J48 Decision Tree) were used for the classification of Maharashtra data from 2015-16 IDHS dataset. Each technique was applied three times on the dataset for classification as the parameters for the three classes i.e. stunted, underweight and wasted are different. The parameter HAZ is used for stunted, WAZ is used for underweight and WHZ is used for wasted. The performance of the algorithms was tested based on four parameters i.e. Correctly Classified Instances, Mean Absolute Error (MAE), Root Mean-Squared Error (RMSE) and Time taken to build the model. After analysis and comparison of the above techniques, it can be concluded that the Multilayer Perceptron model performs better then Naïve Bayes and J48 Decision tree in terms of Correctly Classified Instances, Mean Absolute Error and Root Mean Squared Error. But in terms of the time taken to build the model, Naïve Bayes algorithm turns out to be the fastest.

Future work can be done to analyze data for different states and different set of parameters. Also, other algorithms can be applied on the dataset to analyze their performance.

Thus, choosing the most appropriate algorithm for classification of child nutrition status based on a combination of different accuracy parameter will help clinicians and policy makers to tackle the malnutrition problem.

VIII. References

[1] World Health Organization, "World Health Organization releases new Child Growth Standards," 2006. [Online]. Available: http://www.who.int/mediacentre/news/releases/2006/pr21/en/. [Accessed 20 January 2015].

[2] Keri Wachter, Julie Rosenberg, Robbie Singal, and Rebecca Weintraub, https://www.globalhealthdelivery.org/files/ghd/files/ghd-031_reducing_child_malnutrition_in_maharashtra.pdf, [Accessed October 2015]

[3] Swasth Report Of Maharashtra, "Swasth Report Of Maharashtra: Malnutrition Remains Leading Health Problem Among Children Under Five," 2019. [Online]. Available: https://swachhindia.ndtv.com/swasth-report-card-malnutrition-remains-leading-health-problem-among-children-under-five-in-maharashtra-40471/

[4] Diana Ferreira, Hugo Peixoto, José Machado and António Abelha, "Predictive Data Mining in Nutrition Therapy", 13th APCA International Conference on Automatic Control and Soft Computing, (CONTROLO) June 4-6, 2018

[5] Sri Winiarti, Herman Yuliansyah, Aprial Andi Purnama, "Identification of Toddlers' Nutritional Status using Data Mining Approach", International Journal of Advanced Computer Science and Applications, Vol. 9, No. 1, 2018.

[6] Riris Aulya Putri , Siti Sendari, Triyanna Widiyaningtyas, "Classification of Toddler Nutrition Status with Anthropometry Calculation using Naïve Bayes Algorithm", IEEE 2018 International Conference on Sustainable Information Engineering and Technology (SIET), 2018.

[7] Reyhan Gupta, Abhishek Singhal, A. Sai Sabitha, "Comparative Study of Clustering Algorithms by Conducting a District Level Analysis of Malnutrition", 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2018.

[8] Rizky Ade Putranto, Suryono Suryono, Jatmiko Endro Suseno, "Cloud Computing Medical Record Related Baby Nutrition Status Anthropometry Index During Postpartum", IEEE 2018 2nd International Conference on Informatics and Computational Sciences (ICICoS), 2018.

[9] Cynthia Hayat, Barens Abian, "The Modeling of Artificial Neural Network of Early Diagnosis for Malnutrition with Backpropagation Method", Third International Conference on Informatics and Computing (ICIC), 2018.

[10] Md Mehrab Shahriar, Mirza Shaheen Iqubal, Samrat Mitra, Amit Kumar Das, "A Deep Learning Approach to Predict Malnutrition Status of 0-59 Month's Older Children in Bangladesh", The 2019 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), 2019.

[11] Zuguo Mei and Laurence M Grummer-Strawn, "Standard deviation of anthropometic Z-scores and a data quality assessment tool using 2006 WHO growth standards: a cross country analysis", Bull World Health Organ, v.85(6), June 2007

[12] The DHS Program, "Guide to DHS Statistics. DHS-7," 2018. [Online]. Available: https://dhsprogram.com/pubs/pdf/DHSG1/Guide_to_DHS_Statistics_DHS-7.pdf

[13] WEKA: Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/WEKA/