Mining Nutrition Survey Data

Lydia Manikonda*, Raghvendra Mall†, Vikram Pudi‡ and Raghunatha Rao§
*†‡*Center of Data Engineering*
Hyderabad, India
Email: {{lydia,raghvendra.mall}@research,vikram}@iiit.ac.in
§*National Institute of Nutrition*
Hyderabad, India
Email: drr_rao@yahoo.com

***Abstract*—In developing countries, many national development plans have included nutrition considerations for decades. This has led to a major thrust from the governments to promote nutrition education activities. Many public nutrition research institutes have stepped up their efforts to educate the public by conducting workshops which includes various techniques like giving lectures, providing questionnaires, etc., to train the participants in the most basic and important concepts of nutrition and dietetics. In this paper we focus on an application of mining questionnaires of such kind to determine the current knowledge of child participants and how this knowledge improves after the training session. This has been used to detect what kind of participant in the training session perform well/poorly and to determine if training has been successful and what improvements can be made in future training sessions.**

***Keywords*-Nutrition; Education; Questionnaires; Data Mining**

I. Introduction

Food systems in developing countries have been undergoing many changes as vast and rapid as any in history. Nutrition is not only the consequence of economic development but also it is one of the motive power which plays an important role in the economic development [6], [7]. Although most national development plans have included nutrition considerations for decades, they have traditionally been framed only as outcomes of economic growth. Fortunately, these countries are beginning to recognize the integral role nutrition plays in sustainable development, and the number of national development strategies that include explicit nutrition objectives is growing. As a part of this many government and voluntary organizations are striving hard to reach people and educate them with the basic nutrition concepts by conducting workshops to provide lectures, method demonstrations, questionnaires (surveys) in many rural and urban regions.

Conducting a survey in the form of a questionnaire is one of the easiest and useful ways to understand what people think. The Institute of Nutrition[1], here in referred to as the institute, helps the government by collecting the data is a premier research institute that has a vision to achieve optimal nutrition of vulnerable segments of population such as women of reproductive age, children, adolescent girls and elderly. This institute organizes many camps and surveys where they collect the real data and analyze the data. Because of this analysis, they are able to guide the government to take up some important projects for the benefit of the citizens. Also after critically analyzing some common problems which people are facing in the nutrition scenario, they come up with new nutritive products in their research laboratories which are beneficial for the society.

Currently for the assessment of knowledge and food preferences among the urban adolescent girls, the institute has taken an initiative to educate the adolescent girls from urban poor areas about functions of foods, anemia and folates, adolescence, family life education and infant feeding. The activity involves providing questionnaires to the girls who answer according to their knowledge without any training by the institute. After the training, they are again provided with the same questionnaires and were expected to answer the questions. The institute analyzes this data with the help of statisticians who provide the results being cross-checked by the scientists. The statisticians at this institute perform most of the work manually or by regression.

In this paper we focus on an application of mining questionnaires of such kind to determine the current knowledge of participants and how this knowledge improves after the training session. This paper is divided in to nine sections. Section II briefly explains the problem statement, Section III describes the problems we have faced while applying the data mining techniques on this kind of survey data provided by the institute. Section IV deals with the preprocessing stage followed by Section V which explains an overview of our approach and describes how it is beneficial to the institute to know the results. Section VI explains the statistical results obtained on the data followed by the results obtained by data mining techniques in Section VII. Section VIII shows the results obtained in the two different regions considered by the institute and the paper ends with the conclusions and the future work provided in section IX.

[1]Specific details have been removed to facilitate privacy issues

II. Problem Statement

Statistics provides tools for prediction using data and statistical models and hypothesis testing. For applying statistical models on data, the data should be clean. It relies on the fact that there should be a comprehensive collection of data to answer the given questions. On the other hand, data mining is the analysis of (often large) observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner [2]. Moreover, there are other advantages of data mining over statistics. An elaborated discussion on the relationship between data mining and statistics mentioned can be found in [13], [2], [4] and [3].

The statisticians at the institute use the naive methods of statistics and which might not provide some important information while trying to analyze the data which is collected during the surveys. The main goal of our paper is to apply various data mining techniques like classification, clustering and association rule mining on the data collected by the government research institute for a better analysis of the data.

The data we have considered contains the answers to the questionnaires from the girls of public schools which are located in two regions A and B[2]. For each region, the questionnaires are distributed to the girls twice - first, without training the girls and second, after the girls were given training in the basic concepts of nutrition and family life education. The analysis was done on young adolescent girls from the 8^{th} and 9^{th} grades. We performed analysis on records of 117 girls before the training period and 111 records after the training period. The questionnaire contained around 82 questions. The questions were majorly based on the concepts explained during the training by health educators have given to these girls along with some personal details like socio-economic/demographics details. The major areas on which training was given are:

1) **Adolescent Phase/Food Preference Details:** It is a critical phase of life between childhood and adulthood, characterized by rapid physical growth and development, physiological and emotional changes, social and psychological maturity, development of mental growth and search for adult identity and transition from total socio-economic dependence to relative independence. Food and nutrition play a very important role during this period. Adopting healthy eating life styles and following dietary guidelines are essential for the adolescent population to ensure optimal growth and development. During the training, more details about adolescent phase and tips to lead healthy adolescent life were emphasized.
2) **Breast Feeding and Complimentary Feeding:** Mother's milk contains all the essential nutrients including carbohydrates, proteins and micro-nutrients such as iron, vitamin A etc. Hence, it is often considered as wholesome food for the infants. Details about breast-feeding, complementary feeding, importance of green leafy vegetables (GLVs), fruits, amylase-rich foods (ARFs) and their preparation were more emphasized during the training.
3) **Food Groups and Balanced Diet:** Rapid physical growth demands more energy and nutrients. In order to follow dietary guidelines, classification of foods and the functions they perform should be understood. Foods are classified into three groups based on the functions they perform. They are energy yielding foods, body building foods, protective foods. A detailed emphasis was laid on the macro-nutrients, micro-nutrients and balanced diet during the training.
4) **Anemia and Folates:** In developing countries, Iron Deficiency Anemia(IDA) is a significant public health problem. It is particularly prevalent among pre-school children, adolescent girls and pregnant and lactating women. Recent survey conducted by the Family and Health department indicated that about 75% infants between 6 months and 3 years of age are anemic. About 50% of pregnant women continue to be anemic. As compared to the earlier survey, carried out 5 years ago, the proportion of anemia in fact had marginally increased today. Hence, causes of anemia, its consequences, symptoms and prevention are important and need to be explained. Also, folic acid, green leafy vegetables and how to cook them were also emphasized during the training.
5) **Family Life Education:** It is necessary to impart 'Family Life Education' to adolescent girls as most of them get ready to enter into a married life. They comprise a major portion (about one-fifths) of country's population. Important concepts like puberty, STD & HIV education, avoiding early marriages, other hygienic issues were emphasized.

[2]Specific details have been removed to facilitate privacy issues

III. Challenges faced

As we know that collecting data from the government institutes involves lots of paper-work and privacy has to be maintained. We have also faced several problems. There exists some practically challenging problems while we are attempting to apply data mining techniques to the data collected by the organization.

1) As was mentioned earlier, the biggest challenge was proper data collection. There were many concerns involved while collecting the data-be it from the organization's perspective to visit schools and collect the data or from our side to approach the organization or institute and perform analysis on the data.
2) Most of the government organizations maintain paper based records of data collected during the surveys.

The records which the organization collected were all paper based. This is one of the most basic and toughest challenge that we face in the field of health-care. Legibility and language issues existed. The documents were bulky and vulnerable to damage.

3) Conversion of records from paper to electronic form was a tedious, time consuming and difficult task. This process was done in parallel by all of us in the team. At the end of the task, it would have been very difficult if we had not followed certain format and standards to prevent errors and ambiguity. So it is necessary to chalk out a uniform, unambiguous template for storage of information.
4) For each question in the questionnaire four options were given and the students need to choose the correct choice. Only one correct answer is there for each question. Some of questions were left unanswered by the girls before the training period which resulted in missing values and uncertainty. In the other case, there were multiple answers for each question.
5) Applying data mining to real world applications requires data mining specialists to interact closely with domain experts. Firstly, we as knowledge engineers should have atleast some basic knowledge about nutrition, which could have helped us to better understand the terminology. The knowledge indirectly affects the kind of inferences we make from the data and able to judge whether the inferences are appropriate, interesting and significant.

IV. Pre-processing

Preprocessing of data is necessary to avoid any discrepancies. In our consideration, there is chance for errors to creep in because of issues such as language, hand writing, conversion from paper to electronic records, etc. Also, there were missing values in the answer sheets. For handling missing values, we have added another option to each question in the questionnaire which says $unknown$. To apply data mining techniques, identifying the distinct values for each attribute is helpful. To identify the distinct values of each attribute and the number of occurrence of each attribute value, a python language code has been implemented.

V. Approach

Initially for comparison purposes, we performed statistical analysis using WEKA [5], a popular data mining tool. This was followed by the classification of girls based on their knowledge, then clustering the girls, following which we try to find association rules showing the impact of the background on the knowledge of the girls and various such relations. We also tried to determine outliers which are extreme cases. These are the cases when a girl has exceptional knowledge as in comparison to her class mates or is lacking any general knowledge which other girls of the same age have. The patterns help the organization to identify the areas where they need to provide special training for improving the knowledge of the girls related to important issues of family life and nutrition which are the primary concerns in urban slum areas.

VI. Statistical Analysis

There are some important statistical results which we have identified. They include:

1) The age of girls varies from 11 to 16 years with the mean being 12.813 years and nearly 70% of the girls are aged between 12-13 years.
2) The number of girls from 8^{th} grade is 70% and from 9^{th} grade is 30%
3) The girls are classified in to three different communities C_1, C_2, C_3 based on their religion. Majority of the girls are from C_1 which occupied (82%), others are C_2 (7%) and C_3 (11%). Though the area where the survey was conducted has many families belonging to C_2 community, the number of girls belonging to this community who are getting educated in schools is particularly low.
4) Most of the girls had no knowledge about their height (98%) and weight (89%)
5) Around 30% of the girl's mothers are housewives while around 15% are laborers.
6) 30% of the girl's fathers are engaged in construction work or labour work while 10% of them were working as watchmen and 15% are drivers.
7) The maximum combined income of a family was \$277.78 and minimum was \$2.14 per month. About 50% of the families had a monthly income around \$64.10. From these numbers we can conclude the lower class girls were part of the survey. In some cases the father had expired and it was the mother who was the only source of income.
8) Around 55% of the girl's fathers were illiterate. And around 30% had primary education. So, lack of awareness about proper nutrition and family life planning is not a big surprise.
9) Around 75% of the girl's mothers are uneducated and 20% of them have received only primary education.
10) Looking at the background of the parents it is appropriate to suggest that the children too will lack proper knowledge about nutrition and family life unless educated properly. The lack of education is one of the main reasons behind the high mortality rate and large number of diseases which are prevalent in these urban backward areas.
11) Most of the girls had 2 or more siblings (>75%). This is another problem in the urban backward areas as the country's population is increasing alarmingly and there is no proper strategy and education to control it.

This kind of surveys will help to trace the alarming situations where action should be taken immediately. Similar statistical analysis was performed on the other categories as well. The analysis graph can be seen in Fig. 1.

VII. Analysis using Data Mining Techniques

We mainly focused on applying the data mining techniques like Classification, Clustering and Association Rule Mining on the survey data. The main algorithms which were used for this purpose were few basic algorithms from each of the techniques and are available in WEKA [5].

A. Classification Techniques

Three classification algorithms Naive Bayes [8], BayesNetwork [9], Decision Trees or C4.5 [10] were applied on the dataset. The purpose of choosing these three algorithms was that in some cases Naive Bayes is better than the other two, in some cases BayesNetwork is better and in some other cases C4.5 is the best. While performing analysis of data a major problem was that most of the girls had left some questions unanswered. Now we understood that there can be two reasons behind this:

1) The girls did not know about the question and so had left it unattempted. For example, "Do you know what anemia is". To this most of the girls have left the answer as blank. Here it means that they don't really know about the term "anemia".
2) The other option is that the girls knew the answer but for some reason they did not want to answer. For example, "Do you know how AIDS spreads". Some girls have left the answer as blank while most of the other girls have answered it. This may be due to the reason that the girls are shy.

So for this purpose we have assigned "NA" to the answers which were not attempted. It could mean either of the above two cases and its correct value is identified when we were performing classification.

Classification means that we build a classifier which is helpful to classify new girls on the basis of classifier training data which is the records about the girls which already have. Based on the similarities in the answers of the new girl to that with the ones in the classifier, the classifier predicts the class to which the girl should belong.

B. Clustering Techniques

Two clustering algorithms K-Means [12] and DBSCAN [1] were mainly used in our analysis. Clustering was performed to group the girls based on different categories or attributes and infer interesting results of the clusters. It helps in giving an overall scenario of a particular cluster. In this case-the life style, their knowledge about nutrition issues, etc. helps in taking the decisions quickly. Some critical situations like most of the people who work as labourers in construction companies for daily wages, don't want their children to go to school. Instead, most of them make their children to stop studying and start working with them. Clustering girls according to their parent's occupation and the grade they are in, can help us to come up with many such interesting conclusions. Also, if any new record comes, then based on the cluster it groups into, inferences can be made.

C. Association Rule Mining Techniques

Association rules identify the collection of data attributes that are statistically related in the data. As we know they form the attribute value pairs based on measures like confidence and support. Apriori [11] algorithm was used on this data to get interesting rules. Association rule mining techniques helped in a great way for this kind of data sets and encouraged the organization based on the inferences to take up the initiatives efficiently. The in-depth results with various examples are discussed in the next section.

VIII. Results

A. Region A Vs Region B

Application of data mining techniques have resulted in yielding few interesting results which were unknown using the statistical methods. Few of the interesting issues are:

1) Students belonging to Region A have performed very well compared to Region B candidates. When parent's role is considered, interestingly it is not playing a major role since most of them are illiterate. But age is playing a very interesting role. Most of the Region A students are older than Region B students. This clearly supports the statement that as we grow older, maturity levels will increase.
2) 25% of the Region B students don't include Green Leafy Vegetables (GLVs) in their diet whereas 92% of the Region A students include GLVs in their diet. The reason might be the income of their parents.
3) Association between mother's qualification and girl's knowledge about Anaemia was mined. Mothers of Region B students are less educated than the mothers of Region A's students. Region A students have more knowledge about Anemia than Region B students (this was inferred from the dataset obtained before training by the health educators).
4) Knowledge about balanced diet was also mined. In this case again, Region A students performed better than the students from Region B. The reason might be the parent's income. Graph showing the parents income and the number of students can be seen in Fig.2
5) Students from Region A outperformed Region B students in almost all sections of the questionnaire. The indirect reasons behind this may be the age of the students, their mother's educational background (since mother's play an important role in a girl's life especially in developing countries).

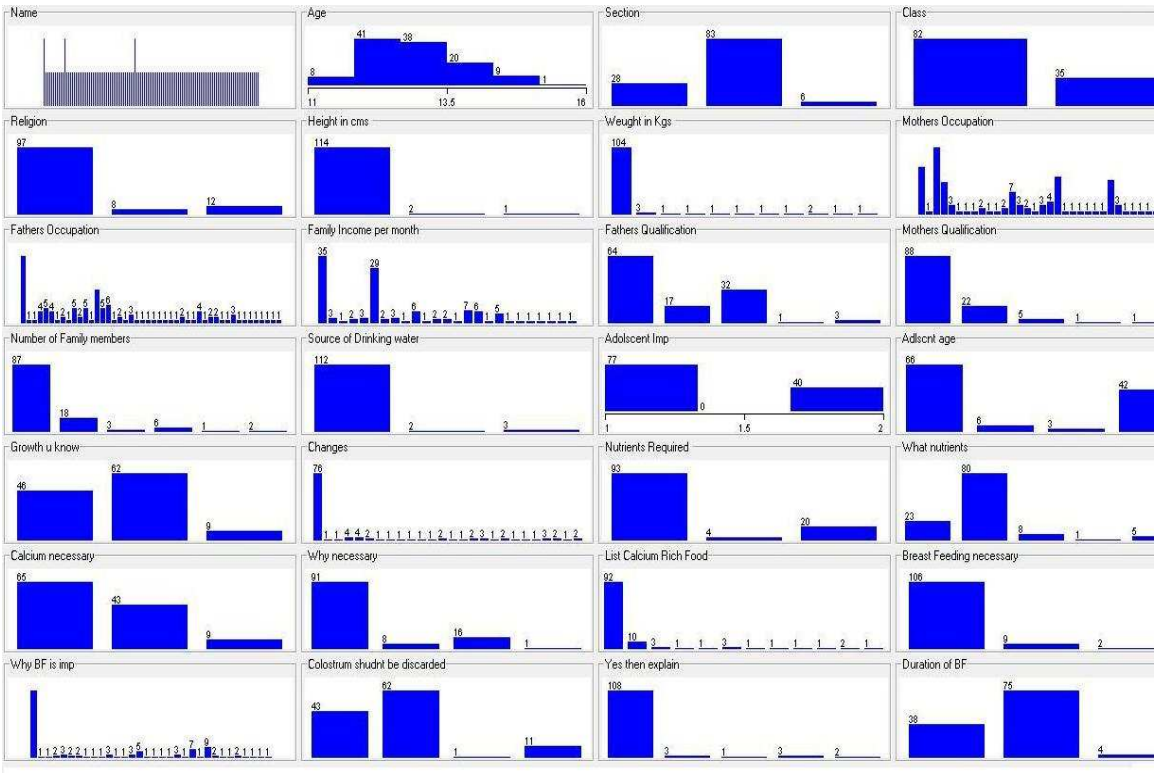

Figure 1. Statistical Analysis

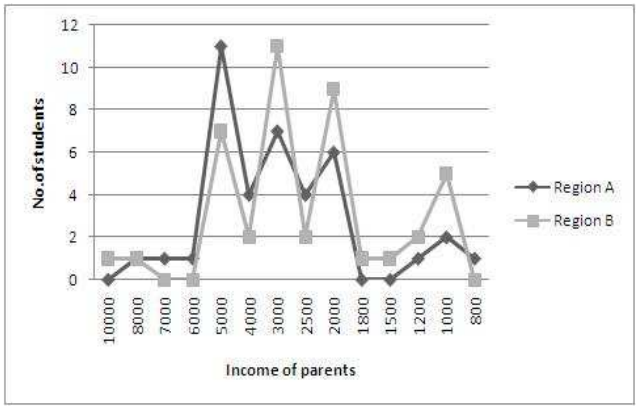

Figure 2. Income of parents Vs Number of students

B. Results from Region A data after the training by health educators

1) Knowledge about the adolescent phase or food preferences and practices was astonishingly improved from the earlier to after the training by health educators. Students were able to identify the answers comprehensively and that too without many mistakes. Classification methods prove that Girls are answering properly without any assumptions. For example, if they are mentioning that adolescent phase is important, then they are marking the correct age of this phase. Best rules identified are like:
   a) Adolescent age=2.0 47 → Adolescent Imp=1.0 46 conf:(0.98)
   b) Adolescent Imp=1.0 49 → Adolescent age=2.0 46 conf:(0.94)

   As we know that confidence measures the genuinity of the rule, so here we can say that if they are mentioning that Adolescent age (Adolescent Age) is 10-18yrs when they are saying that adolescent age is important. Other rules include knowledge about calcium, if it is necessary during the adolescent phase when considered after the training period:

   Why necessary=1.0 46 → Calcium necessary=1.0 45 conf:(0.98)

   When before training period was compared with this, there is lesser confidence than this:

   Why necessary=1.0 18 → Calcium necessary=1.0 17 conf:(0.94)
   So, these infer that knowledge about adolescence was improved very well.
2) Considering the Breast Feeding section, before training many students don't have any idea about

Colostrum or at which month complementary feeding should be started and the other related questions to this topic. But after the training, the answers showed us that girls have learned more about breast feeding. For example, when clustering was done, before training period, many clusters are being generated which suggests that the girls have varying opinions about breast feeding. But after training, very less number of clusters are generated about this question, which clearly shows us that learning took place.
Another example, if we consider the answer for complementary feeding, only 11% answered correctly while after the training by the health educators, 92% of the students answered correctly. This marks clearly that learning has taken place.

3) Considering the food groups and balanced diet, 82% of students don't know about Balanced diet and what constitutes a balanced diet. But after the training, more than 82% students have answered about balanced diet which can be considered as correct without any predictions. When association rule mining techniques were used and the rules were generated, the rules have shown that if the girls are able to identify the food groups and balanced diet, they are able to include the different nutritive food items in their diet which is extremely interesting.
4) Students are learning more knowledge about the importance of certain vitamins and minerals and other concepts like- if they lack certain vitamins and minerals, what are the side effects, etc. Here nearly 80.3% of learning took place in case of knowledge about Anemia.
Also in the case of the role of iodized salt, rules are occurring with good confidence which means that after training by health educators, students are having more awareness of goiter and are observing the fact that using iodized salt can be useful. The rule can be seen below.

aware of goiter=1.0 43 → iodized salt can help to get the required amount=1.0 41 conf:(0.95)

As per the results, even though much learning has taken place, there is only 13% increase in the use of GLVs. But this contributes to a large extent because around 92% of students already consume GLVs in their diet while the remaining might not afford it.
5) In our country the age when a girl reaches puberty has decreased recently. So it is also important for girls to have knowledge about the problems they might face and how to handle them effectively. Before training most of the girls have no idea about the menstrual cycle. From the total number of records, only a few of them have answered this question.

what age=NA 30 → Menstrual cycle=2.0 28 conf:(0.93)

But after training, 100% of the students came to know about the menstrual cycle and are able identify the correct age the cycle starts.
From the classification, it has been shown that before training many girls have predicted the right age of marriage, though initially stating that they don't know the appropriate age for marriage. But after classification those kind of predictions were gone and are clearly stating the fact if they know or not.

a) right age for marriage of girls=NA 15 → if yes mention it=1.0 15 conf:(1)
b) right age for marriage of girls=1.0 91 → if yes mention it=1.0 86 conf:(0.95)

In the above mentioned rules, for the first rule the girls are saying that they don't know the correct age and are answering correctly which is can be considered as a guess. And in contrary the second rule is stating that if they are telling that they know the right age, they are answering properly. But after training, most of the girls have understood facts about family life.

On the whole, there is a major improvement in learning about Health, Nutrition and Family Life Education. These kinds of trainings are very helpful and play a major role. And this kind of evaluation can help the organization to identify the major problems in this field. This helps them to conduct more number of calculated surveys in different regions and help improve the conditions of the society.

C. Results from Region B data after the training by health educators

1) Knowledge about the adolescent phase or food preferences and practices was improved from the earlier to the after training period. But the percentage of increase is not very satisfactory.
Classification methods prove that girls are answering properly without any assumptions. For example, if they are mentioning that adolescent phase is important, then they are selecting the correct age of this phase to a better extent than earlier. Best rules identified are like:

Adolescent Imp=1.0 40 → Adolescent age=2.0 32 conf:(0.8)

Here 40 is the number of rules having "Is Adolescent Imp" answered "Yes(1 is yes and 2 is no)" and "Adolescent age" occurring 32 time along with that answered "10-18 yrs" of age. "Conf" represents the confidence of the rule or how strong we can say

that if they are mentioning that Adolescent age is 10-18 years when they are saying that adolescent age is important. In case of knowledge about calcium necessary during the adolescent phase when considered after the training period, rules are like:

Calcium necessary=1.0 32 → Why necessary=1.0 13 conf:(0.41)

Before training period results was compared with this, there is lesser confidence than the above one:

Calcium necessary=1.0 12 → Why necessary=1.0 4 conf:(0.33)

So, these infer that knowledge about adolescence was improved well.

2) Considering the Breast Feeding section, before training many students don't have any idea about Colostrum or at which month complementary feeding should be started and related info. But after the training, the answers showed us that learning happened. But it was not to a great extent. When considered the use of Colostrum, whether to discard it or not, there is no much improvement in the answers. Training didn't improve their knowledge much. This indirectly shows us that there is need of further training on this topic. From 9.75% before training, it increased to 35% which is lesser compared to that of region A's School.
   Another example, consider the answer for complementary feeding. Only 9.75% answered correctly while after the training, 15.25% of the students answered correctly. This clearly shows us that much improvement was not seen over here.
3) Considering the food groups and balanced diet, 85.36% of students don't know about Balanced diet and what constitutes a balanced diet. But after the training, more than 54.23% students still have not answered about balanced diet which can be considered as correct without any predictions.
   When clusters were made, inference was made that after training the results show that learning has happened. Also when classifying food groups, there is an improvement of around 25% that students were able to correctly identify and classify the food groups. So we conclude that to some extent learning took place.
4) Students are learning more knowledge about the importance of certain vitamins and minerals and if they lack those, what are the side effects, etc. Since Anemia is much more prevalent in our country we've verified the amount of increase in the knowledge. Here nearly 33.36% of learning took place in case of knowledge about Anemia. This is comparatively better than many of the earlier inferences made about the girl's learning condition.
   Also in the case of the role of iodized salt, rules are occurring with good confidence which means that after training, students are getting more awareness about goiter and are observing the fact that using iodized salt can be useful. This was predicted when clustering was made. The post-training answers are showing a certain group of cluster, where as the pre-training results are showing them as unclustered instances. This shows that there is an improvement and the girls have similar opinion now about iodized salt.
   Since in the training period, there was much information explained on the importance of folates, iron, etc., there is no improvement in the inclusion of GLVs in their diet.
   Considering all these above cases, there is certainly an improvement but not as much as expected.
5) In contrary to the previous school, students here have more knowledge about the family life. Many of them have a better idea about Menstrual Cycle and there is 0% improvement in this case of knowledge about it. Apart from this, there is also a sufficient knowledge about marriage too. They are correctly identifying the right age of marriage.
   But when questioned about AIDS, there is a controversy to the above. 85% of the girls don't have an idea about that disease but after the training, only 37% of the students have not learned about it but the rest all identified well.
   On the whole, there is a major improvement in learning about Health, Nutrition and Family Life Education. But Region B's knowledge was not improved well compared to Region A. These issues when critically analyzed resulted that many factors as mentioned earlier come in to play. Parent's income, qualifications have more affect and on the other hand school environment also plays a major role as these children spend a lot of time at school. Thus, these kinds of training and analysis are very helpful and play a major role in taking some necessary actions and plans by the government to improve the living conditions of the under-privileged citizens of the nation.

IX. Conclusions and Future work

The above are some of the interesting issues which we were able to mine from the data and this process can be continued on as many other attributes as possible which if found are useful and necessary. We've shown that data mining is able to infer the results much better than the results obtained from traditional statistical analysis. Also, statistics cannot be applied on the original data since it cannot perform properly if data is not clean. But we can see practically that data mining can be performed easily after minimal reprocessing methods are applied.

We've clearly seen that conducting training sessions like this will be very useful especially for adolescent girls because they are the future mothers and they comprise one-fifth of the country's total population. In almost all the inferences, improvement was seen ranging from 0% to 100%. From the results we can also infer that there is some kind of variation in the school of Region B students compared to the Region A's students. There are many reasons which can lead to this inference. Some of the reasons might be:

1) Age of Region A students is more than Region B's students or
2) Mothers of Region A girls are well educated than Region B's girls
3) Social status, for example, income of parents of Region A students is better than Region B students as a whole
4) Perhaps some difficulties during the training period
5) Perhaps students are not able to follow what the health educators are explaining during the training phase
6) Perhaps the number of students in Region B who are absent is more than the number of students who are absent in Region A during the training provided by health educators

All of these reasons can play a significant role which lead to the unexpected level of learning in region B.

We performed our case study on the baseline data which is the data collected before training by the health educators and the post-training data. In the future, if large amount of data is collected, even more precise results can be obtained where measuring the difference of knowledge before training and after training can be easily observed from these methods instead of statistical approaches. This will enable us to provide information about all the important aspects for which training has to be given by the health educators. This can help the people of urban slum areas to live a better and healthy life.

REFERENCES

[1] Martin Ester and Hans-Peter Rigel and Jo erg Sander and Sower Xi, *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, In Second International Conference on Knowledge Discovery and Data Mining, 226–231, 1996.

[2] David J. Hand,*Data Mining: Statistics and More?* The American Statistician, May 1998 Vol. 52, No. 2

[3] Patricia Cherrita and John C. Cherrita, *Data and Text Mining the Electronic Medical Record to Improve Care and to Lower Costs*

[4] David J. Hand, *Statistics and Data Mining: Intersecting disciplines*

[5] Mark Hall and Elbe Frank and Geoffrey Holmes and Bernhard Porringer and Peter Reutemann and Ian H. Witten, *The WEKA Data Mining Software: An Update*, SIGKDD Explorations, 2009 Vol.11, Issue 1.

[6] Chen C and Wang Y, *Nutrition and economic development of poor areas*, Journal of hygiene research, 2000 Vol. 29, Issue 5, 305–307.

[7] John Strauss and Duncan Thomas, *Health, Nutrition, and Economic Development*, Journal of Economic Literature, 1998 Vol.36, no.2, 766–817.

[8] Duda, R.O. and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, 1973.

[9] N. Friedman, D. Geiger, and M. Goldszmidt, *Bayesian network classifiers*, Machine Learning, 1997 Vol.29, 131–163.

[10] Quinlan, J. R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.

[11] R. Agrawal, R. Srikant, *Fast Algorithms for Mining Association Rules in Large Databases*, In 20th International Conference on Very Large Data Bases, 478-499, 1994.

[12] J. A. Hartigan and M. A. Wong, *A K-Means Clustering Algorithm*, JSTOR: Journal of the Royal Statistical Society, Series C (Applied Statistics), Vol. 28, No. 1 (1979), 100–108.

[13] http://www.cs.csi.cuny.edu/∼imberman/DataMining/Statistics%20vs.pdf