Exploration of Co-Relation between Depression and Anaemia in Pregnant Women using Knowledge Discovery and Data Mining Algorithms and Tools

Fayyaz Ahmed[1], *Adibah Sitara*[2]
[1]Tata Consultancy Services, Toronto, ON, Canada
[2]Pakistan Army, Rawalpindi, Punjab, Pakistan

Abstract— Data Mining (DM) and Knowledge Discovery (KD) requires certain steps or rules that determine the data access mechanism and facilitate the Decision Support System (DSS). This survey has reviewed standard algorithms that are well known in research community and has discussed the criterion for these algorithms, which are classification, regression, segmentation, association, and sequence analysis. These data mining classifications are subsets of standard algorithms and are used by data miner software vendors for their database server data analysis services. Depressive symptoms are common amongst pregnant women following anaemia and predict subsequent maternal mortality and morbidity and fetal abnormalities. A data set for South East Asia (Pakistan) low income group, second and third trimester pregnancy, is collected to form a test bed for data miner tools and algorithms. The analysis shows that both Cross Industry Standard Process for Data Mining (CRISP-DM) and Sample, Explore, Modify, Model, Assess (SEMMA) are practical e-health management tools. To the best of the researchers' knowledge, the classification data mining algorithm is better due to its operational simplicity and ease of handling of the statistical data. The research discovered a correlation between depressions related anaemia during pregnancy, and it enables researchers to better understand pregnancy complications related to depression.

***Index Terms*— Data mining, Knowledge discovery, Algorithms**

I. Introduction

Well known algorithms in the research community are C4.5 [1], k-Means [2], SVM [3], Apriori [4], EM [5], PageRank [6], AdaBoost [7], kNN [8], Naïve Bayes [9], and CART [10]. The C4.5 algorithm is based on classification method and takes collection of cases as input; each collection belongs to classes and described by its values for a fixed set of attributes. The output is a classifier that predicts the class to which a new case belongs.

K-means is a user defined k-clustered iterative method to partition a given dataset. Support Vector Machines (SVM) is a classification method to distinguish between members of two classes in the training data. In apriori algorithms frequent item sets can be sorted using candidate generation. Expectation maximum (EM) algorithm is an iterative algorithm that iterates between maximum likelihood and maximization step. It is used to fit the mixture models by maximum likelihood. The expectation (E) step computes the expectation of the log-likelihood evaluated, using the current estimate for the latent variables, while the maximization (M) step computes parameters maximizing the expected log-likelihood found on the *E* step. The distribution of the latent variables can be predicted in the next E step. PageRank is patented by Stanford University and used by Google for static ranking of web pages, with the page value computed for each page offline. PageRank is a graph theory concept for ranking of internet pages by computing the limit stationary probability distribution of a random walk on the internet graph, where the nodes are pages, and the edges are links among the pages. AdaBoost extracts one classifier from the pool in each of M iteration. The elements in the data set are weighted according to their current relevance at each of the iteration. All elements are assigned the same weight at the beginning e.g. just 1, or 1=N if a total sum of 1 for all weights is required. As the drafting progresses, the elements are assigned larger and larger weights. AdaBoost is one of the most important ensemble methods since it has solid theoretical foundation, very accurate prediction, and greater simplicity. The KNN classification is the nearest neighbor finding algorithm, often termed as regression analysis. The K algorithm holds good for points, if regressions are at uniform interval. It finds a group of “k” objects in the training set that are closest to the test objects “x”. The value of “k” often depends upon the measure matrix applied for finding the value of “k”. Euclidian, hamming distance and large margin nearest neighbor are known K matrices for continuous variables, text classification and optimized classification respectively. The Naive Bayes classifier is a conditional independence statistical classifier, based on the Bayesian theorem and can predict class membership probability that a given sample belongs to a particular class. The CART decision tree is a binary recursive

- *Fayyaz Ahmed is Software engineer at Tata Consultancy, Toronto, Ontario, Canada.*
- *Adibah Sitara served as Lieutenant Colonel in Army Medical Corps, Rawalpindi, Pakistan.*

partitioning procedure that is constructed by splitting a node into two child nodes repeatedly, beginning with the root node that contains the whole learning sample. Table 1 shows the classical algorithms, their techniques and data processing description.

Algorithm	**Technique**	**Data Processing**
Naive Bayes	Classification	Supervised binning of attributes
GLM (General Linear Model)	Classification and Regression	Normalized attributes
SVM (Support Vector Machine)	Classification and Regression	Normalized attributes.
k-Means	Clustering	Outlier-sensitive normalization.
O-Cluster	Clustering	Binned attributes (equi-width binning). Columns with all nulls or a single value are removed.
MDL (Minimum Description Length)	Attribute Importance	Supervised binning
Apriori	Association Rules	Simple association
NMF (Non-negative Matrix Factorization)	Feature Extraction	Normalized attributes
EM (Expectation Minimization)	Clustered	Maximum likelihood/ maximum a posteriori (MAP) estimates of data sets
Page Rank	Link Analysis	Numerical weighting to each element of a data set
Adaptive Boosting (AdaBoost)	Machine learning	Large pool of classifier (two-class pattern recognition) Strong out of weak classifier
KNN (K Nearest Neighbor)	Classification	Classifying objects based on nearest neighbor method
CART (Classification and Regression Techniques)	Decision Tree	Combination of Classification and Regression

Table 1 Data Mining Techniques and Algorithms

The supervised binning method, which is transform numerical variables into categorical counterparts, is in fact grouping the values of variable into a relatively small set of discrete values (bins), each of which represents a range of values on the original variable. The discretization allows an analysis that is restricted to categorical variables. Bin normalization is classified as min-max, scale, and z-score normalization. Shift value for min-max is minimum value and scale is the min-max difference. Shift is zero in scale normalization and scale is the maximization of absolute value of min-max values.

Scale = max {abs (max), abs (min)}

The z-score normalization uses mean and standard deviation with shift as mean and scale as the standard deviation. In order to avoid information loss during supervised binning, outliers should be treated appropriately. Outliers are the inappropriate tail and head values that can cause information loss. Outliers use trimming of data, which is assigning null values data sets as missing; while the winsorizing outlier technique sets a percentile value for the minimum and maximum value. Outliers filter the unwanted information and set a cut-off point that can minimize computation time and optimize results.

II. Data Mining Techniques

Data warehouse technologies rely on standard data mining techniques and algorithms. This section has discussed these techniques in detail.

A. Classification:

This technique predicts the outcome of a product, business strategy and policy. The standard algorithms used in this technique are regression, C4.5, SVM, Naïve Bayes, and CART. The Generalized Linear Model (GLM) is a statistical technique for linear modeling for logistic regression. GLM consists of three elements: a probability distribution from the exponential family, a linear predictor and a link function. The GLM has a dependent variable, an independent variable, a link function, and linear predictor to derive a mathematical function to generate an expected value for the dependent variable. Taking an independent variable A, and dependent variable B for a mean value of combination of probability distributions ranging from Poisson, normal and binomial, the expression can be derived as

$$E\ (A) = \mu$$

$$\mu = \frac{B\rho}{G}$$

$$E\ (A) = \frac{B\rho}{G}$$

The expected value of dependent variable *E(A)* is equivalent to the mean of the distribution function µ, which is the arithmetic relation of link function *G*, and the linear predictor *Bρ*. The variance of the dependent variable and the mean of distribution function follows the same relationship.

B. Regression Technique:

This technique predicts one or more continuous variables, such as profit or loss, based on other attributes in the dataset. Time series algorithm provides regression techniques that are optimized for the forecasting of continuous values, such as product sales, over time. A time series model can predict trends based only on the original dataset that is used to create the model. New data can be added to the model to predict and automatically incorporate the new data in the trend analysis. The Regression technique is supported by SVM and multiple regression algorithms. Support Vector Machines (SVM) is based on the Vapnik-Chervonenkis theory. SVM can model problems such as text and image classification, character recognition, and bioinformatics and bio-sequence analysis. SVM performs well on data sets that have many attributes, with no upper limit on the number of attributes, even if there are very few cases on which to train the model.

C. Attribute Importance Technique:

This technique is an automated solution for improving the speed and possibly the accuracy of classification models built on data tables with a large number of attributes. We can term it as a filtering technique based on the patterns of the data set. Inductive inference for model selection bears the insight that any irregularity in data can result in data reduction. The minimum description length (MDL) algorithm used for the attribute importance technique is a machine learning process which selects a cut off attribute strategy for filtering data sets thereby reducing the computational time and complexity.

D. Anomaly Detection Technique:

This method involves comparing data in the referenced data set block to some baseline model data set from historical data. A way to accomplish this is to monitor statistical measures computed for combinations of categorical attributes in the database. Considering such combinations gives rise to a multidimensional array at each time interval. Each dimension of such an array corresponds to the levels of a categorical variable. Anomaly detection objective is to identify unusual or suspicious activity based on deviation from the normal historical data. Cases of interest are fraud in banking, health care and tax compliance. The data mining tools available use SVM one-class classifier for anomaly detection. The SVM case analysis produces a prediction and a probability for each case in the scoring data. For a typical case, the prediction is 1 and 0 for an anomalous case.

E. Clustering Technique:

Clustering is forming a subset of data from a selected data set. It is different from classification as it classifies data into newer groups, and forms a homogenous group that can serve as supervised data model for prediction. Anomaly detection in clustering is for data sets that are misfit in a clustered environment, termed as outliers and needing to be normalized for optimal data processing. A clustered data model can be hierarchical, partitioned, locality based or gird based. The data sets generated in clustering are defined by their position in the cluster hierarchy, their rules for positioning in the hierarchy, distribution of values of an attribute in a cluster, and by typical attributes existing within a cluster. The clusters are hierarchical within a cluster through a parent/child relation.

F. Association Technique:

The Association technique is useful for developing a decision strategy for large databases. Association relations between different attributes possess well defined patterns that are helpful to develop business rules [11]. The association technique can be sub divided into co-occurrence of data sets and frequently occurring data sets. The standard Point Of Sale method (POS) [12] is based on co-occurrence of data sets of different groups. The most common applications of association are web semantic analysis of online database entities, and data compression based on certain patterns. The Association technique is only suitable for frequently occurring data sets. A minimum threshold value may increase computational time and generate results which are non-predictive. The anomaly detection technique is better for less frequent attributes. Association rules require various analyses such as in breadth [13], in depth [14], frequent pattern compression [15], and conjunction and disjunction association [16].

G. Feature Extraction Technique:

The Feature extraction technique is based on time series data extraction. The problem associated with time series is high dimensionality. Correlated data sets extracted through technique should be filtered to get reduction down features. Discrete Wavelet Transform (DWT) and Discrete Fourier Transform (DFT) are most commonly used feature extraction techniques. Indexing time series data can reduce computational overhead and optimize feature extraction technique for multi-dimension data set.

III. Data Mining Model Methodologies

A. CRISP-DM:

CRISP-DM [17] stands for Cross-Industry Standard Process for Data-Mining and this research has reviewed six steps for CRISP-DM methodology. The data mining is modeling or exploring of the data based on its historical or predictive analysis. The classification in the modeling phase would require a regression analysis, classification analysis, association analysis or clustering analysis. Historical data can be divided into categorical or numerical analysis.

A.1: Problem definition:

Understanding the business problem is the first phase of a data mining project. For business project objectives and requirements the data mining experts, business experts, and domain experts work together. The project objective is formulated as data mining problem definition. In this phase no data mining tools are required.

A.2 Data exploration:

Domain experts collect, describe, and explore the data, and also identify quality problems. Interaction of data mining and business experts from the first phase is vital in this phase. Conventional data analysis tools such as statistical and visualization techniques are used to explore the data. The exploration would essentially require uni-variate or bi-variate analysis. Uni-variate performs step wise variable analysis, either categorically or numerically. Bivariate is association analysis between two variables. It can be numerical and numerical, numerical and categorical, and categorical and categorical.

A.3 Data preparation:

Domain experts collect, cleanse, and format the data and also create new derived attributes, such as, an average value. Data is prepared for the modeling tool by selecting tables, records, and attributes. Data is collected through counting (categorical) or measurement (numerical). Data set is a combination of rows (records), columns (variables) and values (data). Database is a system that enables users to retrieve, add, update or remove the data. The system itself is named as a Database Management System (DBMS). Data mining toolboxes connect to databases through Open Database Connectivity (ODBC) or Java Database Connectivity (JDBC) interfaces.

A.4 Modeling:

In this phase data mining experts select and apply various mining functions, as some of the mining functions require specific data types, therefore data mining experts must assess each model. In this phase, domain experts from the data preparation phase are exchanged to make the model feasible. Modeling and evaluating phase are coupled and are repeated several times for optimal values. The summarized model reports can form different patterns that can be classified, sequenced, associated, clustered etc.

A.5 Evaluation:

The evaluation model is evaluated by data mining experts for an optimal and/or a sub optimal model; the model is rebuilt by changing its parameters until optimal values are achieved. Finally, data mining experts decide how to use the data mining results. There are two methods of evaluating models in data mining; hold-out and cross-validation. The first one deals with large data sets and the second with smaller data sets. The hold-out is further divided into a training set for predictive models, validation set to assess the performance of the model and a test set to assess the likely future performance of a model

A.6 Deployment:

Data mining results are exported into database tables or into other applications. The Deployment phase can be simple e.g. generating a report or complex e.g. implementing a repeatable data mining process. Most commonly, the customer carries out the deployment steps. There are four way of deploying the models in data mining:

- Through cloud
- Through programming language
- Through Database and SQL script
- Through PMML (Predictive Model Mark up Language)

B. SEMMA Model:

SEMMA [18] methodology stands for Sample, Explore, Modify, Model, and Assess. This definition defines the five steps involved in this methodology. This model uses various statistical and visualization techniques: this model selects and transforms the most significant predictive variables, models the variables to predict outcomes, and confirms a model's accuracy.

B.1 Sample:

One or more data tables are created by data mining experts that contain significant information. The size of the tables is optimal so they can be processed quickly.

B.2 Explore:

Data mining experts explore data by searching for relationships, trends, and anomalies in order to gain understanding and ideas.

B.3 Modify:

Data mining experts modify the data by creating, selecting, and transforming the variables to focus the model selection process.

B.4 Model:

This phase is also similar to the CRISP-DM model phase. The software performs an auto search for a combination of data that reliably predicts a desired outcome.

B.5 Assess:

Similar to the evaluation phase of CRISP-DM, the data is evaluated for usefulness and reliability of the findings from the data mining process.

The SEMMA model is an iterative process that starts with sampling the data, and then we use visualization or clustering techniques to explore the data. The data explored is modified according to pre-determined criteria. Various statistical model techniques are used during the modeling phase such as neural networks, and decision tree. Finally, the model is evaluated for the business objectives and can be re-modeled for optimized results.

CRISP-DM and SEMMA analysis indicates that CRISP-DM methodology is more comprehensive and covers all the practical implementation of the DM model. SEMMA seems to lack the end user knowledge and business objective phase but an analytical review shows that without end user requirements, knowledge data cannot be sampled and summarized. Sampling is merging the phases of business understanding and data understanding. CRISP–DM business and the data understanding correspond to sample and explore of SEMMA. The modification phase of SEMMA is in line with data preparation of CRISP-DM. Model and modeling is similar for both CRISP-DM and SEMMA. Assessment of SEMMA is a combination of evaluation and deployment of CRISP-DM. Both CRISP-DM and SEMMA are practical DM tool and are widely used in industry for formulating business solutions.

IV. Data Mining Mathematical Model

There are various implementations of mathematical models which are applied in vendor specific DM tools. These models can be categorized into supervised and unsupervised learning processes.

A. Apriori Algorithm:

Apriori algorithm considers an X-space problem containing input Variables from X_1 to X_n, with Y as the class variable termed as output. The association property using Apriori algorithm for large values of inputs and simplified domain can be accomplished by setting a minimum value for a threshold. The pseudo code can be expressed as:

```
Find all frequent item sets
  Get frequent items
        Items whose occurrence in database is greater than or
        equal to the min.support threshold
  Get frequent item sets
        Generate candidates from frequent items
            Prune the results to find the frequent item sets
        Generate strong association rules from frequent item
        sets
Define Rules which satisfy the min.support and
min.confidence threshold
```

The support of a rule indicates how frequently the items in the rule occur together, and the confidence of a rule indicates the probability of both the antecedent and the consequent appearing in the same transaction. The IF component of an association rule is known as the antecedent; the THEN component is known as the consequent. The antecedent and the consequent are disjoint; they have no items in common. Confidence is the conditional probability of the consequent given the antecedent.

For large values of data sets, the probability of happening of a particular item in a data set is

$$P(x_1, x_2, \dots \dots \dots x_p)$$

The domain simplification process for 1 to j values are given as

$$P\left(\bigwedge_{j=1}^{p} \bigvee_{xj\,\varepsilon\,Sj} (X_j = x_j)\right)$$

With

$$S_j \cong \text{Domain } (X_j)$$

Assuming absolute value of S_j is equal to one for sub set values of X:

$$\bigvee_{xj\,\varepsilon\,Sj} (X_j = x_j) \equiv (X_j = x_j) \equiv \top$$

Where T is the item set and T (J ε j) is the support variable with X_j as the binary variable, such that

$$(P) \begin{cases} 1 & if\ P = T \\ 0 & otherwise \end{cases}$$

For the threshold values t, considering output item sets J with the condition that T(J) > t, now for the association rule we take another item set ζ, such that ζ ≡ J, then T(ζ) ≥ T(J), which implies that any item set deleted for summarization in J could also be deleted from ζ.

B. Classification Algorithm:

The classification algorithm works on the items' similarity or dissimilarity contained in a data set. Choice of appropriate similarity or dissimilarity is more important than choice of the algorithm itself. This choice is dependent on the problem domain, mixture of qualitative and quantitative variables and missing variables would bring in complication for the classification algorithm.

For data set D, containing the set of variables X = {X_1 ……X_p}, for variable X_j attains value $x_{i,\ j}$ within instance x_i ε D. Now the dissimilarity d ($x_{i,\ j}$, $x_{k,j}$) for values between $x_{i,\ j}$ and $x_{k,j}$ of variable X_j.

$$\blacktriangle\ (x_i, x_k) = \sum_{j=1}^{p} \omega_j\, d(x_{i,j}\ , x_{k,j})$$

Where ω_j is sample weight such that $\sum_{j=1}^{p} \omega_j = 1$ and average dissimilarity for data set D, for N = $|D|$

$$\blacktriangle\ = \frac{1}{N^2} \sum_{i=1}^{N}\ \sum_{k=1}^{N}\ (x_i\ , x_k) = \sum_{j=1}^{p} \omega_j\, \overline{d_J}$$

$\overline{d_\iota} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{k=1}^{N} d(x_{i,j}\ ,\ x_{k,j})$ which is reciprocal to the associated weights.

C. Clustering Algorithm:

Clustering algorithm also separates data set items based on their similarity but the separation is group based. Unlike classification where there is a predefined target, the clustering does not have a specified target. Rounding or omitting certain attribute values might generate some errors that develop anomalies in the specified cluster. The clusters are formed hierarchically. Final results are leaf clusters whereas intermediate clusters are higher in order within a hierarchy. The difference between two instances of a data set x_i and x_j for a cluster can be expressed as

$$\text{T} = \frac{1}{2} \blacktriangle\ (x_i\ , x_k) \sum_{i=1}^{N} \sum_{k=1}^{\text{N}}$$

where T is a constant cluster value for a particular data set, and the components of cluster hierarchy is within cluster (wc) and between cluster (bc) groups scattering, whereas for bc values, k ≠ 1. The clustering algorithm is based on the assumption as how can the wc be minimized that is the difference between T and bc. K means is a further variation of a clustering algorithm with an embedded Euclidian distance between instances in a data set, for wc scatter and dissimilarity.

V. Simulation Test Bed

There are various data mining mathematical tools such as KNIME, Rapid Miner, Weka, TANGRA, Orange, and Teradata that facilitate statistical simulations. All of these data mining tools follow both CRISP-DM and SEMMA methodology. Due to scarcity of space, this research has only used Orange and Weka data miner.

A. Background Information:

Malnutrition is widely recognized as a public health problem of critical importance for developing countries. Women and children are most affected by malnutrition, especially during pregnancy due to increased nutritional requirement. Malnutrition is not merely the result of ignorance of what constitutes a balanced diet. It is also a reflection of a wide range of socio-economic factors such as poverty, lack of education, poor social and health services, high fertility rate, unhygienic living conditions and food shortage.

Anaemia is one of the most prevalent nutritional deficiency problems affecting pregnant women [15]. For many developing countries, prevalence rates of up to 75% are reported. In industrialized countries, anaemia in pregnancy occurs in less than 20% of women. This does not, however, still reach the level of public health significance (≥10%). Published rates of prevalence for developing countries range from 35% to 72% for Africa, 37–75% for Asia and 37–52% for Latin America [16]. High prevalence of anaemia is expected to contribute significantly to maternal mortality and morbidity and fetal abnormalities.

Over the last few decades, Pakistan has made significant efforts for economic development and provision of social services for its population but unfortunately, it has not improved the nutrition situation in the country. Malnutrition continues to be a major problem in Pakistan, and inadequate food intake contributes substantially to increased maternal morbidity and mortality and neonatal deaths.

Anaemia in pregnancy is an important preventable cause of maternal and prenatal morbidity and mortality [17]. In published studies, the mean minimum normal hemoglobin in healthy pregnant women living at sea level is 11.0–12.0 g/dl. The mean minimum by World Health Organization (WHO) criteria is 11.0 g/dl in the first half of pregnancy and 10.5 g/dl in the second half of pregnancy [18]**.** Anaemia in pregnancy is further divided into mild anaemia (Haemoglobin 10.0–10.9 g/dl), moderate anaemia (Haemoglobin 7.0–9.9 g/dl) and severe anaemia (Hemoglobin >7.0 g/dl) [19]. Each year more than 500,000 women die from pregnancy-related causes, 99% of these in developing countries [20-21]. Estimates of maternal mortality resulting from anaemia range from 34/100,000 live births in Nigeria to as high as 194/100,000 in Pakistan [16]. Complete blood count (CBC), including hemoglobin concentration and red blood cell indices, mean corpuscular volume (MCV), mean corpuscular hemoglobin concentration (MCHC), and reticulocyte count, are the first steps for the positive and differential diagnosis of anaemia [22].

Anaemia in child bearing age and pregnant women is related to iron deficiency [23]. This research explored that Serum folate or vitamin B12 deficiency is crucial for pregnant and lactating mothers and their infants. For research study group in Northern Pakistan, Serum folate or vitamin B12, has found to be considerably lower than prescribed standards. Reproductive aged women and growing children are the principle groups at risk of anaemia. Nutritional anaemia is more widespread among pregnant and lactating women because of the increased needs for iron during those periods. Due to frequent pregnancies most women enter pregnancy with inadequate iron to meet the increased iron needs, required for red blood cell mass expansion in the mother, as well as for the development of the fetus and the placenta. Approximately 1000 mg of iron are needed during pregnancy [24] of which 500 mg are used to support the expanding maternal hemoglobin mass and 300 mg for the development of the fetus and placenta.

Iron is obtained in the form of non-haem iron from vegetables and as haem iron from meat. Haem iron is absorbed about two to three times and is better than non-haem iron. A small amount of haem iron in the diet improves absorption of non-haem iron and thus the diet composition is an important determinant of the amount of iron actually absorbed. Iron is stored in the reticulo-endothelial system as ferritin and haemosiderin. Iron is a component of hemoglobin and iron deficiency ultimately leads to defective erythropoietin and anaemia**.** In many developing countries, it is difficult to meet daily nutrient requirements with diet alone especially for pregnant women. Animal products and fats are often relatively expensive and in addition, there may be food taboos which influence dietary intake in pregnancy.

In mild cases the pregnant women may not have any symptom as the body system adjusts to reduced hemoglobin mass. However, patient can still come up complaints of ill health fatigue, dyspnoea, palpitations and tachycardia, vertigo, loss of appetite and cravings for soil. In severe cases the deficiency may cause oedema.

Anaemia during pregnancy has a detrimental effect both on mother and fetus. The risks of prematurity and low weight are increased for infants of anemic women. Iron deficiency anaemia leads to abnormalities in host defense and neurological dysfunction increased risks of premature labor and low birth weight.

B. Eperimental Study:

In order to better understand the pregnancy complications in developing countries, there is a need of analysis for these countries facing scarcity of food and poverty. Prevalence of nutritional anaemia and its effects are widespread especially in high risk group, which is pregnant women and children; therefore it is worth to conduct a study for the group. This study focuses on assessing the prevalence of anaemia in second and third trimester of pregnancy in low socioeconomic group. It also compares the prevalence of anaemia in second and third trimester among expectant mothers. The study's biochemical tests for assessment of anaemia such as serum iron profile and serum folate levels were not performed due to financial limitations.

The data set analyzed through Orange/Weka data miner has a poulation set of age group ranging from twenty (20) to forty (40). The group is moduled into four sub-groups ranging from twenty to twenty five, twenty six to thirty, thirty one to thirty five and thirty six to forty. The research also scale the iron-rich food intake from one to ten, with one being the worst case scenario while ten being the ideal condition. The food intake is a dependent entity based on the demographic and income background. Although the data set collected belongs to a group of Pakistan Army house wives, yet the army ranks and income group was not taken into account for computer simulations. The research also set some criteria for data collection. The inclusion criteria was to include all pregnant women in second and third trimester of pregnancy, belonging to low socio-economic class, visiting a medical facility for antenatal check up (irrespective of their parity and age). Pregnant women suffering from any major illness such as Diabetes Mellitus, Tuberculosis, Renal, and Cardiovascular diseases and Hematological disorder were excluded from the study. The clinical examinations were conducted for Conjunctival pallor, Palmer pallor, Angular stomatitis, Koilonychia, Glossitis, Palpitation and Oedema. The biochemical finding were based on the readings of Hemoglobin concentration, RBCs count, Packed Cell Volume (PCV), MCH (Methylcyclohexenon), MCHC (Microcrystalline hydroxyapatite) and Peripheral picture.

AGE	*Hemoglobin concentration (g/dl)*	*RBCs Count (million/cu mm)*	*PCV (ft)*	*MCH (pg)*	*MCHC (g/dl)*
23	10	4	74	26	30
34	11.6	4.2	76	27	30
30	9.8	3.8	72	26	29
---	---	---	---	---	---
39	11.9	4.3	87	29	31
40	10.2	3.8	74	24	28

Table 1 Study Participants Data

Cluster based analysis is sensitive to initialization, and is prone to anomoalies (if data is not well described). For clustered data attributes, the cost of optimal solution decreases with increase in k cluster until zero (when k = n). As the initialization criteria are changed there is an apparent shift in the clusters that are regrouped in different parts of the graph.

The data set collected were attributed according to anaemic and non-anaemic classification, educated and non-educated classification and urban and rural classification. The data were run on both weka and orange data miner.

The data sets which are collected over a period of time without any boundary lines are often difficult to understand and simulate. The results generated are often confusing as the data sets are not properly initialized. The values which are extended during data mining often do not match the actual/original values. Based on the data collected, attributes can be determined by the prediction model. The attribute value can be either true or false. This prediction model can classify data sets into two categories: Category one matches one attribute, category two matches another (can be taken as true or false). The results are analyzed through observers' expertise, historical data and simulation results. Tools supporting data prediction model are widely used by research community. Complex data sets can define more than two attributes which can refine the prediction model. For example if we are dealing with patient data: a patient can be anaemic or non anaemic. The correlation model can be anaemic patient with depression or without depression and vise versa.

Depressive symptoms are common following acute anaemia and predict subsequent morbidity. Depression in patients appears to be independent of clinical disease severity. This study assessed the relationship between anaemia on admission and depressive symptoms.

Design: Longitudinal clinical observational study.

Setting: Mother care unit.

Patients: 223 patients with documented anaemic and non anaemic pregnancy.

Main outcome measures: Depressive symptoms measured with the Beck Depression Inventory BDI [25].

The cause of depression can be endogenous or exogenous. The social factors such as girl child, low income, unwanted pregnancy, are exogenous depression factors. Endogenous depression factors are dependent on age and medical conditions. The phenomenon of depression during pregnancy is complex.

C. *Results Discussion:*

Anaemia was defined with WHO criteria, anaemia predicted raised depression scores 3 weeks later independently of age, gender, educational attainment, smoking, The odds of a Beck Depression Inventory score >= 10 among anaemic patients were 4.03 (95% CIs 1.48 to 11.00), adjusted for covariates. Sensitivity analyses indicated that effects were also present when hemoglobin was analyzed as a continuous measure. Classification algorithm using Orange data miner simulations are presented in Fig. 1 and 2a – 2c. Simulations show a co-relation between the anaemia and haemoglobin. There is an apparent sign of depression association with haemoglobin (anaemia) and this phenomenon can better help pregnant patient for controlling depression by administering drugs to improve haemoglobin (control anemia) and thereby decreasing chances of depression.

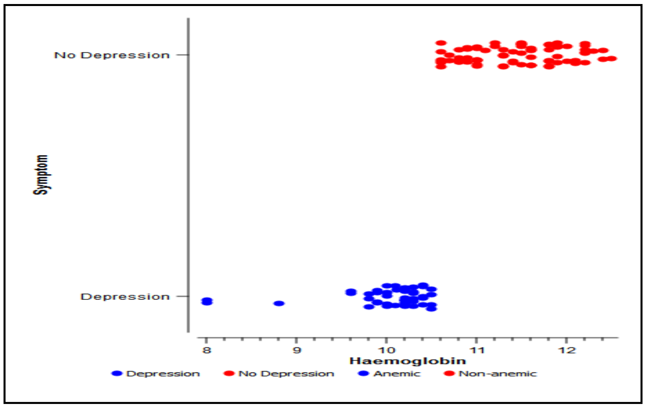

Fig. 1 Co-Relation of Anaemia and Depression

The same relation is observed in Fig. 2a – 2c.

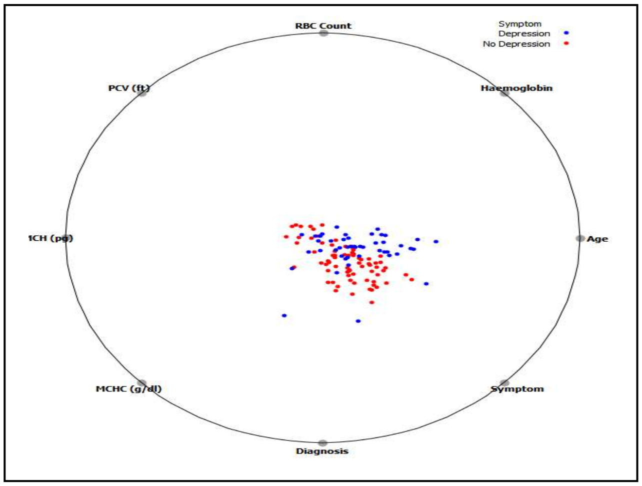

Fig. 2a Attributed Co-Relation for Pregnant Women

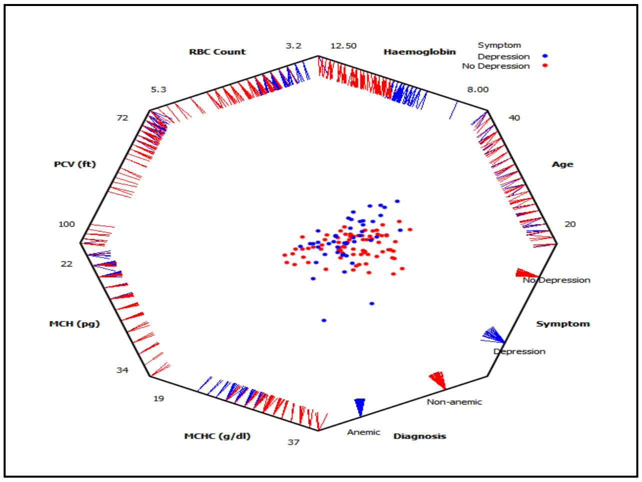

Fig. 2b Attributed Co-Relation for Pregnant Women

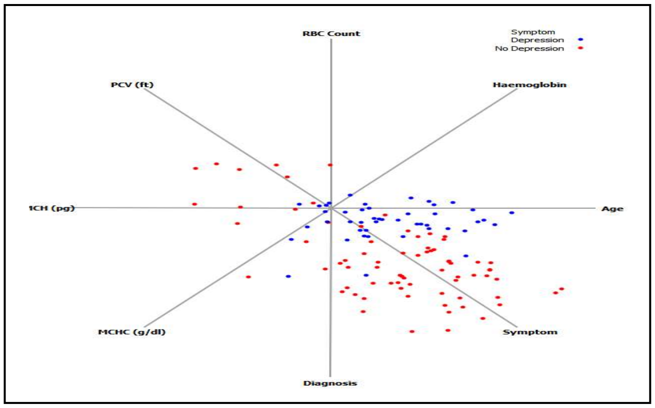

Fig. 2c Attributed Co-Relation for Pregnant Women

VI. Conclusion

Tool selection is an important step for knowledge discovery. For this research, data mining Weka and Orange tools were selected, and it is explored that Orange data miner is more flexible than Weka for its large statistical spectrum and flexible GUI interface. It has built-in data sets and data upload is simpler as compared to the weka data miner that requires complex programming skills. For data miner algorithms, to the best of researcher's knowledge, the classification method can be regarded as the best data mining algorithm because of its simpler data preparation, handling of numerical and categorical data with the statistical tools for large data sets.

Previously, multiple studies performed for determining anaemia and depression in pregnant women have revealed contradictory results, advocating both relation and negation. All those studies were conducted, using one statistical model and without any counter re-run, to verify the results. This research is the first attempt (to the best of researchers' knowledge) to present results which were counter verified by using all the known statistical models embedded in data miner tools. Study of pregnant women data simulations concludes that anaemia appears to contribute to depression and is associated with future child morbidity. Studies evaluating the effects of anaemia management will help delineate the role of this pathway more precisely.

For future research, it is suggested that a resource comparative study can be conducted for CPU usage time versus the data miner input-output operation delays for various data mining engines. Virtual cloud computing and index coherence cloud are recently explored topics that also requires researcher's attention.

VII. Acknowledgment

The authors appreciate Mr. Manir Islam, Master Engineering student at University of Ontario, Institute of Technology, Ontario, Oshawa, Canada, for simulation help.

VIII. Appendix

A Query using Pearson coefficient [26] to determine correlation between two attributes (Anaemia and Depression):

```
SELECT
Group1, Group2,
((psum - (sum1 * sum2 / n)) / sqrt((sum1sq - pow(sum1, 2.0) / n)
* (sum2sq - pow(sum2, 2.0) / n)))
AS
r, n
FROM
(SELECT
n1.Group AS Group1,
n2.Group AS Group2,
SUM(n1.depression) AS sum1,
SUM(n2.depression AS sum2,
SUM(n1.depression * n1.depression) AS sum1sq,
SUM(n2.depression * n2.depression) AS sum2sq,
SUM(n1.depression * n2.depression) AS psum,
COUNT(*) AS n
FROM
testdata AS n1
LEFT JOIN
testdata AS n2
ON
n1.anaemic = n2.nonanaemic
WHERE
n1.Group > n2.Group
GROUP BY
n1.Group, n2.Group) AS step1
ORDER BY
r DESC,
n DESC
```

References

[1] Fails, J. A. and Olsen, D. R. Interactive machine learning. In Proceedings of the International Conference on Intelligent User Interfaces (IUI) Miami, Florida, January 12-15, p.39-45, 2003.

[2] Chris Ding and Xiaofeng He., "K-means Clustering via Principal Component Analysis", In Proceedings of International Conference of Machine Learning, pp 225-232. July 2004.

[3] Ovidiu Ivanciuc "Applications of Support Vector Machines in Chemistry", Reviews in Computational Chemistry, Volume 23, pp. 291–400, 2007.

[4] Farah Hanna AL-Zawaidah and Yosef Hasan Jbara "An Improved Algorithm for Mining Association Rules in Large Databases," World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 1, No. 7, 311-316, 2011.

[5] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome. "8.5 The EM algorithm", The Elements of Statistical Learning. New York: Springer. pp. 236–243. 2001.

[6] Andrea Esuli and Fabrizio Sebastiani. "Page Ranking WordNet synsets: An Application to Opinion-Related Properties", In Proceedings of the 35th Meeting of the Association for Computational Linguistics, Prague, CZ, pp. 424–431. Retrieved June 30, 2007.

[7] Saehoon Kim, Yoonseop Kang, and Seungjin Choi, "Sequential Spectral Learning to Hash with Multiple Representations," Lecture Notes in Computer Science Volume 7576, pp 538-55 1, 2012.

[8] Bremner D, Demaine E, Erickson J, Iacono J, Langerman S, Morin P, and Toussaint G. "Output-sensitive algorithms for computing nearest-neighbor decision boundaries", Discrete and Computational Geometry Journal 33 (4): 593–604, 2005.

[9] Webb, G. I., J. Boughton, and Z. Wang, "Not So Naive Bayes: Aggregating One-Dependence Estimators", Journal of Machine Learning 58(1). Netherlands, Springer, pages 5-24, 2005.

[10] G. Bellala, S. K. Bhavnani and C. Scott, "Extensions of Generalized Binary Search to Group Identification and Exponential costs," Advances in Neural Information Processing Systems, 2010.

[11] Jan Rauch, "Logical Aspects of the Measures of Interestingness of Association Rules," Advances in Machine Learning II, Studies in Computational Intelligence Volume 263, pp 175-203, 2010.

[12] J. Burez, D. Van den Poel, "Handling class imbalance in customer churn prediction," Elsevier Expert Systems with Applications 36, 4626–4636, 2009..

[13] Colin Shearer, "The CRISP-DM model: the new blueprint for data mining," Journal of Data Warehousing, Vol. 5, No. 4, 2000.

[14] Gonzalo Mariscal, Óscar Marbán and Covadonga Fernández, "A survey of data mining and knowledge discovery process models and methodologies," The Knowledge Engineering Review, Volume25, Issue02, pp 137-166, 2010.

[15] Yarlini Balarajan, Usha Ramakrishnan, Emre Özaltin, Anuraj H Shankar, Dr SV Subramanian, "Anaemia in low-income and middle-income countries," The Lancet, Volume 378, Issue 9809, Pages 2123 - 2135, 2011.

[16] Andrew F Goddard, Martin W James, Alistair S McIntyre, Brian B Scott, Guidelines for the management of iron deficiency anaemia," GUT An International Journal of Gastroenterology and Hematology, pp. 1309-1316, 2010.

[17] Shulman CE, Levene M, Morison L, Dorman E, Peshu N, Marsh K, "Screening for severe anaemia in pregnancy in Kenya, using pallor examination and self reported mortality," Trans R Soc Trop Med Hyg ; 95: 250-5, 2001.

[18] Henna Hämäläinen, Katja Hakkarainen, and Seppo Heinonen,"Anaemia in the first but not in the second or third trimester is a risk factor for low birth weight," Clinical Nutrition, An International Journal devoted to Clinical Nutrition and Metabolism, 2002. Volume 22, Issue 3, Pages 271-275, 2003.

[19] A. Demmouche, A. Lazrag, S. Moulessehoul, "Prevalence of anaemia in pregnant women during the last trimester: consequences for birth weight," European Review for Medical and Pharmacological Sciences; 15 (4): 436-445, 2011.

[20] Karine Tolentino, Jennifer F. Friedman, "An Update on Anemia in Less Developed Countries," American Journal on Trop. Medical Hygiene; 77(1): 44-51, 2007.

[21] World Health Organization, "Iron Deficiency Anaemia: Assessment, Prevention and Control. Geneva: World Health Organization," 2001.

[22] Bentley M E; Griffiths P L., "The burden of anemia among women in India," European journal of clinical nutrition; 57(1):52-60, 2003.

[23] Baig-Ansari N, Badruddin SH, Karmaliani R, Harris H, Jehan I, Pasha O, et al., "Anaemia preva-lence and risk factors in pregnant women in an urban area of Pakistan," Food Nutrition Bulletin;29 (2):132-9, 2008.

[24] Hyder AA, Wali SA, McGuckin J. "The burden of disease from neonatal mortality: a review of South Asia and Sub-Saharan Africa". BJOG: an international journal of obstetrics and gynecology, 110:894-901, 2003.

[25] J. L. Beard, M. K. Hendricks, E. M. Perez et al., "Maternal iron deficiency anemia affects postpartum emotions and cognition," *Journal of Nutrition*, vol. 135, no. 2, pp. 267–272, 2005.

[26] Toby Segaran, "Programming Collective Intelligence Building Smart Web 2.0 Applications," O'Reilly Media, Pages: 362, 2007.

Fayyaz Ahmed received his MASc degree in Electrical and Computer Engineering from University of Ontario Institute of Technology, Oshawa, Canada and MSc Degree in Computational Engineering from McMaster University, Hamilton, Canada in 2011 and 2009 respectively. His area of research is Communication Networked Control Systems. Currently he is working at Tata Consultancy as Software Engineer.

The author is certified with Microsoft, IBM, Sun Microsystems, Oracle, and Research in Motion with computer programming, networking, database-system and network administration since 2000. He is IBM academic initiative member, and Microsoft TechNet club member and active in blogging volunteer activities for Microsoft SQL server technologies.

Previously, he completed his BA degree in Computer Applications from AIOU, Islamabad, B-Tech Honours degree in Electrical Engineering from UET Lahore and MSc in Computer sciences from Al-Khair University, AJK and MS in Information Technology from Hamdard University, Karachi, Pakistan. He has more than twenty years of experience for IT operations and Management at CANDU systems.