5-2013

Determinants of Quality of Life of Sickle Cell Patients: A KDDM Process Model based Exploration

Gunjan Mansingh
The University of the West Indies, gunjan.mansingh@uwimona.edu.jm

Kweku-Muata Osei-Bryson
Virginia Commonwealth University, KMOsei@vcu.edu

Monika Asnani
The University of the West Indies, monika.parshadasnani@uwimona.edu.jm

Recommended Citation

Mansingh, Gunjan; Osei-Bryson, Kweku-Muata; and Asnani, Monika, "Determinants of Quality of Life of Sickle Cell Patients: A KDDM Process Model based Exploration" (2013). *CONF-IRM 2013 Proceedings*. 19.
http://aisel.aisnet.org/confirm2013/19

Determinants of Quality of Life of Sickle Cell Patients: A KDDM Process Model based Exploration

Gunjan Mansingh
The University of the West Indies,
gunjan.mansingh@uwimona.edu.jm

Kweku-Muata Osei-Bryson
Virginia Commonwealth University
KMOsei@VCU.edu

Monika Asnani
The University of the West Indies
monika.parshadasnani@uwimona.edu.jm

Abstract

Sickle Cell Disease (SCD) is the most common single-gene disorder worldwide and has multiple and variable manifestations. The many medical complications associated with it such as acute chest syndrome and painful crises, along with a lack of normal functioning, may lead to various psychosocial problems such as depression, loneliness and impaired quality of life. A few studies have sought to examine the relationships between demographics, disease severity, depression, loneliness and quality of life of such patients. In this paper we apply the knowledge discovery via data mining (KDDM) process to explore factors which impact the quality of life of sickle cell patients in Jamaica to explicate knowledge which can be used by medical professionals. We use multiple modeling techniques such as Decision Trees, Regression and Regression splines to generate multiple models on the dataset and then present a best set of models to the medical professionals. This allows the medical professionals to select models which will assist them in the decision making process. The benefits of using the process model are highlighted in this study.

Keywords

KDDM, Sickle Cell, Data Mining, Decision trees, Regression, Regression splines.

1. Introduction

Sickle Cell Disease (SCD) is a chronic disease which debilitates the patients suffering from it. In Jamaica, commonly occurring complications include painful crises, acute chest syndrome, leg ulcerations and priapism. Various studies have reported that complications of SCD may often lead to social isolation and depression (Ohaeri *et al.* 1995, Jenerette *et al.* 2005, Asnani *et al.* 2010). Recently Asnani et al. (2010) analyzed socio-demographic and clinical factors of a birth cohort of SCD patients in Jamaica in order to identify variables which affect depression and loneliness. They reported unemployment, having leg ulcerations, more frequent painful crises, and frequently visiting clinics as being factors that were all positively associated with depression.

There has also been an interest in understanding how depression and loneliness along with clinical and socioeconomic factors affect the *quality of life* (QOL) of persons with SCD.

Previous studies on the relationship between SCD and depression, loneliness, and QOL have typically involved the use of basic statistical techniques, including descriptive techniques and regression (Gil *et al.* 1989, Wilson Schaeffer *et al.* 1999, Asnani *et al.* 2010). However, it is known that each data analytic technique has limitations in terms of the answers that it can provide. For example, both regression and regression splines (Ko and Osei-Bryson 2004) can identify the order of importance of the independent variables in a predictive model, and estimate the value of the coefficient for each independent variable. However, if the impact of an independent variable on the dependent variable is conditional, then regression splines can identify such conditions while regression cannot. Thus, some questions cannot be answered using regression since it does not provide the means for exploring those questions. Previous studies have examined the impact of clinical and socioeconomic factors on QOL in terms of whether the impact is positive, negative, or non-existent, however they do not identify conditional relationships amongst and within the variables.

In order to improve the understanding of the factors that affect quality of life of sickle cell patient, in this study we propose the use of the knowledge discovery via data mining (KDDM) process model (Fayyad *et al.* 1996, Sharma and Osei-Bryson 2010) to explore impact of clinical and socioeconomic factors on QOL. The benefits of following the KDDM were evident in a study on an internet banking which performed multiple applicable techniques on a dataset and then presented a best set of models to the decision maker (Mansingh *et al.* 2013). Similarly, is this study we generate a set of best models which the medical professionals can use to assist in the decision making process.

2. Overview on the KDDM Process Model Methodology

The Knowledge discovery via data mining (KDDM) process is a multiple phase process that aims to semi-automatically extract new knowledge from a dataset (Kurgan and Musilek 2006, Sharma and Osei-Bryson 2010). It consists of the following phases: Business (or Application Domain) Understanding, Data Understanding, Data Preparation, Data Mining (or Modeling), Evaluation, and Deployment (see Figure 1 & Table 1). CRISP-DM (cross-industry standard procedure for data mining), the most popular of the KDDM process models, was developed by a multi-industry collective of practitioners after the practitioner community became aware of the need for formal data mining process models that describe the journey from data to discovering knowledge. The original model has been further extended by other researchers (Cios *et al.* 2000, Sharma and Osei-Bryson 2010, Osei-Bryson 2012).

One major difference between traditional confirmatory approaches to data analysis and a data mining (DM) based approach is that in the traditional approach a single model is generated, often based on default parameter settings, while with a DM-based approach multiple models would be generated using different data analysis techniques and different parameter settings. This DM-based approach allows for the generation and assessment of multiple competing models. A consequence of this is that, the Evaluation phase could be challenging for various reasons. For example, with regard to decision tree (DT) induction although the performance measures may be clear (e.g. Table

2), challenges include the need to evaluate a large number of DTs (Osei-Bryson 2004). Given this challenge Osei-Bryson (2004) proposed an approach for comparing and selecting the 'optimal' decision tree (DT) model given preference and value functions specified by the domain expert(s). In this paper we will use the approach of Osei-Bryson (2004).

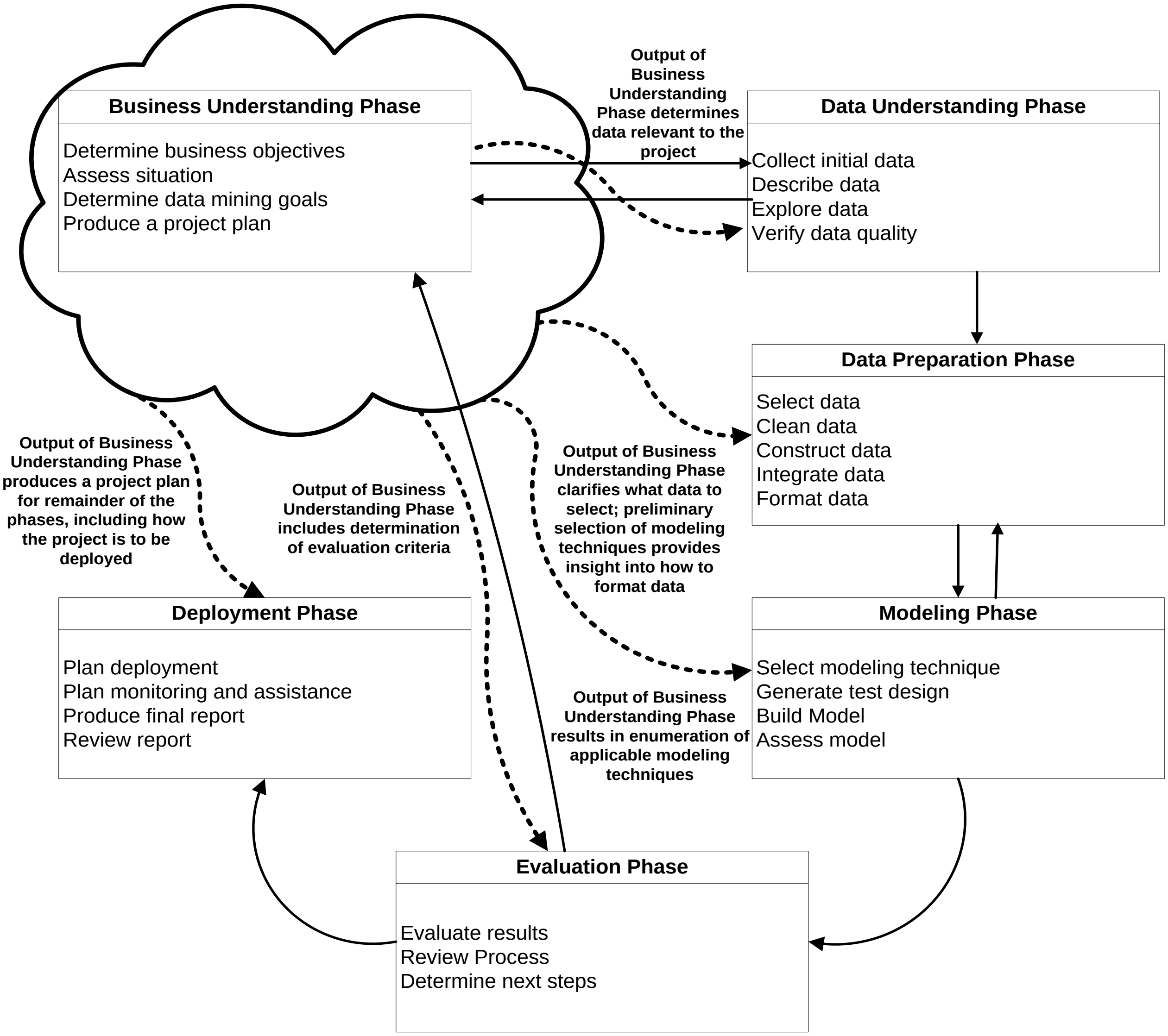
Business Understanding Phase
Determine business objectives
Assess situation
Determine data mining goals
Produce a project plan
Output of Business Understanding Phase determines data relevant to the project
Data Understanding Phase
Collect initial data
Describe data
Explore data
Verify data quality
Data Preparation Phase
Select data
Clean data
Construct data
Integrate data
Format data
Output of Business Understanding Phase produces a project plan for remainder of the phases, including how the project is to be deployed
Output of Business Understanding Phase includes determination of evaluation criteria
Output of Business Understanding Phase clarifies what data to select; preliminary selection of modeling techniques provides insight into how to format data
Deployment Phase
Plan deployment
Plan monitoring and assistance
Produce final report
Review report
Modeling Phase
Select modeling technique
Generate test design
Build Model
Assess model
Output of Business Understanding Phase results in enumeration of applicable modeling techniques
Evaluation Phase
Evaluate results
Review Process
Determine next steps

Figure 1: KDDM Process Model
Source: (Sharma & Osei-Bryson 2009)

Phase	Tasks
Business Understanding (BU)	a) Define Research Goals and Success Criteria. b) Use relevant existing theory to identify variables that are likely to be relevant to the phenomena of interest. c) Do preliminary identification of relevant data including sources of the data. d) If relevant: • Use existing theory & extant research to: provide guidance for the development of data collection instrument. • Develop, test & refine data collection instrument. e) Identify specific data analysis methods (e.g. regression, DT induction, Regression Splines, clustering, structural equation modeling, Data Envelopment Analysis) plus their parameter settings for use in the Modeling (i.e. Data Mining) phase. f) Determine whether available DM software offer adequate facilities for applying the selected Data Analysis methods. g) Elicit Preference Functions from Researcher (e.g. weights obtained using the AHP) that will be used in the Evaluation step for Comparing Causal Models. h) Elicit from Researcher Value Functions that may be relevant for some measures and DM methods (e.g. trapezoidal value functions for Simplicity)
Data Understanding (DU)	a) Collect Initial Data b) Describe Data (e.g. the format of the data, number of records and variables in each table, names of the variables) c) Explore Data (e.g. Determine data distributions using histograms, simple statistical analysis; Find outliers; Do Factor Analysis & Validity Tests; Determine if there are natural Groups) d) Explore relationships between pairs of variables using correlation analysis, etc e) Assess Data Quality
Data Preparation (DP)	a) Clean the Data b) Construct the data (i.e. create derived variables, discretize where relevant, integrate if necessary) c) Convert data to the format that the selected tool requires to satisfy the requirements of the given DM tool
Modeling or Data Mining (DM)	a) Apply to the prepared data, each DM method that was selected in the Business Understanding (BU) phase. b) Record the resulting data that corresponds to the DM performance measures elicited in the Business Understanding phase.
Evaluation (EV)	Evaluation of the generated knowledge from the business perspective a) Exclude models that do not satisfy the relevant threshold for any of the Performance Measures. b) For each model, use the Preference Function to generate a Composite Performance Score for that model. c) Rank models in descending sequence based on Composite Score.

Table 1: Phases of Adjusted KDDM Process Model

Measure	Description
Accuracy	The most commonly used accuracy measure for problems with discrete targets is the *proportion correctly classified*; for problems with interval targets the *R-Squared (R^2)*, or Average Squared Error (ASE) are often used.
Simplicity	In situations when the given model is to be applied by human beings rather than computers, there is the concern is that the model should be interpretable, thus facilitating ease of use.
Stability	This measure concerns our interest that there should not be much variation in this *accuracy rate* when a predictive model is applied to different datasets. Thus at a minimum one might expect that there should not be much variation in predictive accuracy of the given predictive model on the validation dataset when compared to that for the training dataset.
Descriptiveness	The user's subjective assessment of the descriptive power of the output provided by a particular technique (e.g. Regression, Decision Tree induction, Regression Splines) from the perspective of the end-user.

Table 2: Description of Measures for Evaluating Models

Table 3 presents sample outputs for some data analysis techniques. This output can be examined by the end user to determine a value for the measure *Descriptiveness*. The *Descriptiveness* scores would reflect the relative preferences of the user for the various types of output. For example, *Regression* output is in the form of an equation; *Decision Tree* output is in the form of a set of IF-THEN rules; *Regression Splines* output is in the form of an equation that incorporates conditions. The user could use a pairwise comparisons technique to generate *Descriptiveness* scores for the given set of data analysis techniques. It should be noted that like *Stability*, *Descriptiveness* is a subjective assessment of comprehensibility from the perspective of the user (e.g. the researcher), but while *Stability* is assessed at the level of the individual model, *Descriptiveness* is assessed at the level of the data analysis technique (e.g. Regression, Decision Tree (DT) induction). Thus, for example if multiple DT models were generated then they could have different *Simplicity* scores, they would have the same *Descriptiveness* score.

3. Application of KDDM to QOL data

In this study we apply the KDDM model to analyze the factors that affect the quality of life of Sickle cell patients. The emphasis in this paper is on the first four phases.

3.1 Business Understanding Phase

The objectives of the study were: "*Identifying the factors that impact the Quality of Life (QOL) for Sickle Cell patients?"* and *"How medical professionals can use this knowledge in their decision making process"*. Determining these objectives requires the use of classification and prediction modeling techniques that are explanatory. Regression, Decision Trees, Neural Networks, Memory based reasoning and Regression splines are examples of classification and prediction modeling. Since the target variable is interval valued, then relevant data analysis techniques that can be used to achieve this objective are Regression, Regression splines and Decision trees. Other techniques such as neural networks lack the explanatory power. Therefore in accordance with the business objective, the dataset and the available tools (i.e. Enterprise Miner from SAS, MARS from Salford Systems), regression (RG), decision trees (DT) and regression splines (RS) were the techniques that were selected to be used in this study.

Regression

```
                                 Standard
Parameter          DF   Estimate  Error   t-Value  Pr > |t|
Intercept           1   99.1398   4.0755    24.33    <.0001
Bone 0              1    4.9042   1.4069     3.49      0007
Dep_Score           1   -0.2465   0.0853    -2.89      0044
Hospital_visits 1 1     14.8514   4.7905     3.10    0.0023
Hospital_visits 2 1      3.1882   2.4451     1.30    0.1944
Hospital_visits 3 1     -8.4859   3.4137    -2.49    0.0141
Hospital_visits 4 1     -2.0608   5.5187    -0.37    0.7094
LONELY_score        1   -0.5816   0.1782    -3.26    0.0014
Surgery  0          1    3.0465   0.9721     3.13    0.0021
```

Decision Tree:

IF HOSPITAL_VISITS $\in$ {3, 4, 5}
THEN QOL : {AVE = 79.8658; SD = 5.77232}

IF LONELY_SCORE < 21.27 AND HOSPITAL_VISITS $\in$ {1, 2}
THEN QOL : {AVE = 99.2433; SD = 9.15824}

IF 21.27 <= LONELY_SCORE AND HOSPITAL_VISITS $\in$ {1, 2}
THEN QOL : {AVE = 87.9515; SD = 13.3212}

Regression Splines Model

QOL = 98.6731
- 1.33998 * max(0, LONELY_SCORE – 18)
- 0.368776 * max(0, DEP_SCORE - 0)
+ 5.01116 * (HOSPITAL_ADMISSIONS =1)
+ 0.0504462* (EDUCATION in ("2", "1"))*max(0, 21 - LONELY_SCORE) * max(0, DEP_SCORE - 0)
- 0.112928 * (HAVE_CHILDREN$ = "Y") * max(0, 28.9007 - AGE) * max(0, DEP_SCORE - 0)

Table 3: Sample Outputs from Various Data Analysis Techniques

Table 4 presents the relevant performance measures that will be used to compare models generated by these techniques and the corresponding weights of these performance measures. These weights could be determined by the user based using a pairwise comparison technique, and would reflect the user's understanding of the relative importance of the given measures. Tables 5a & 5b present value function for *Simplicity* and *Descriptiveness* that would also be determined by the user using a pairwise comparison technique. For example, the user (i.e.

researcher) may prefer RS & DT models mover the RG models because they allow for description of conditional relationships, and prefer RS models more than DT models because the latter also provide coefficients for the variables. The value function for *Simplicity* for regression models (RS and RG) is based on the number of predictors included in the model and for a DT it is determined by the number of rules that were generated (see Table 5a).

Measure	**Definition**	**Weight**
Accuracy	R^2	0.30
Stability	Min {Training R^2/ Validation R^2, Validation R^2/ Training R^2}	0.20
Simplicity	○ For Regression or Regression Splines models, this could be based on the Number of Predictors ○ For a Decision Tree model, this could be based in the Number of Rules	0.25
Descriptiveness	Descriptiveness of the output	0.25
Composite Score	0.30**Accuracy* + 0.20**Stability* + 0.25**Simplicity* + 0.25**Descriptiveness*	

Table 4: Performance Measures

IF Number_of_Rules > 2 and < 7	THEN $\text{Score}_{\text{Simplicity}}$ = 1.00
IF Number_of_Rules < 2 or >10	THEN $\text{Score}_{\text{Simplicity}}$ = 0.00
IF Number_of_Rules = 2 or (> 6 and < 9)	THEN $\text{Score}_{\text{Simplicity}}$ = 0.75
IF Number_of_Rules = 9	THEN $\text{Score}_{\text{Simplicity}}$ = 0.50
IF Number_of_Rules =10	THEN $\text{Score}_{\text{Simplicity}}$ = 0.25

Table 5a: Value Function for DT Simplicity

Data Analysis Technique	**Descriptiveness**
Regression (RG)	0.65
Decision Tree (DT)	0.80
Regression Splines (RS)	1.00

Table 5b: Descriptiveness Scores

3.2 Data Understanding Phase

The dataset consisted of 264 records. The distribution and the measurement levels of the variables in the dataset were examined. The first two columns in Table 6 show the variable names along with their corresponding measurement levels and column 4 shows the data values of the fields in the dataset.

3.3 Data Preparation Phase

In this phase the dataset is prepared to be consistent with the requirements of the chosen tool and the selected data mining techniques. Some of variables in the dataset were recoded and derived variables were created. Some nominal measurement levels were converted to ordinal variables (see Table 6). Derived variables were created using the depression and the loneliness scores (e.g. Lonely Score * Depression Score and Lonely Score/Depression Score). Previous studies have identified both Lonely Score and Depression Score as being important variables, therefore we created derived variables from these two variables to determine if interactions existed between these variables and whether their additives could become strong predictors.

3.4 Modeling Phase

In the modeling phase the three selected (see section 3.1) data mining techniques are applied and their results are compared.

- Linear Regression (SAS)
- Decision Trees (SAS)
- Regression Splines (MARS)

The dataset was partitioned into training and validation datasets. Both the datasets were created by using stratified sampling method. A new variable in lieu of *QOL Score* was created by binning the data values of *QOL Score* into 10 buckets. The new variable consisted of 10 possible values for *QOL Score* and its distribution was normal. This new *QOL Score* was then used for stratification. This sampling technique ensures that the original dataset is partitioned into complementary subsets in which both the training and validation datasets have adequate amounts of data belonging to each QOL Score value. This ensures that each of the partitions is representative of the complete dataset. The training dataset is used to build the model and the validation dataset is used to evaluate the model. For each technique we used different parameter setting and generated 45 models. 10 Decision Tree models, 34 Regression Models and 1 Regression Splines models were generated and compared based on the performance measures. Figure 2 displays a subset of the process flow diagram that was followed for performing Regression and Decision Trees techniques. For each of the modeling techniques different parameter setting were used. In Table 7 the results of a subset of the 45 models are presented. The first column gives the name of the model, the second and the third column inform us whether the model had used any derived variables or if variable selection was done. For each of the models the scores for performance measures along with composite score which multiples the score of a performance measure with the corresponding weights which were set in the BU phase (see Table 4) are given.

Variables	Original Measurement level	Measurement level	Values	Recoded Values
Sex	Binary	Binary	Male, Female	1, 0
Marital Status	Binary	Binary	Married, Not Married	1, 0
Lives with Family	Binary	Binary	Yes, No	1, 0
Have Children	Binary	Binary	Yes, No	1, 0
Education Level	Nominal	Ordinal	Primary Secondary Vocational Tertiary	1 2 3 4
Employed	Binary	Binary	Employed, Not employed	1,0
Hospital Visits	Nominal	Ordinal	Once a week 2-3 times a week Once a month 1-3 times a year Other	4 3 2 1 5
Pain	Nominal	Ordinal	Almost daily Once a week Once a month Rarely Never	5 4 3 2 1
Lung	Binary	Binary	Yes, Not Yes	1,0
Kidney	Binary	Binary	Yes, Not Yes	1,0
Stroke	Binary	Binary	Yes, Not Yes	1,0
Gall Bladder	Binary	Binary	Yes, Not Yes	1,0
Leg Ulcer	Binary	Binary	Yes, Not Yes	1,0
Eye	Binary	Binary	Yes, Not Yes	1,0
Bone	Binary	Binary	Yes, Not Yes	1,0
Hospital Admission	Nominal	Ordinal	More often than others Every few months Once a year Once every 3-5 years 0-2 times	5 4 3 2 1
Surgery	Binary	Binary	Yes, No	1,0
Lonely Score	Interval	Interval	8-31	
Depression Score	Interval	Interval	0-50	
QOL Score	Interval	Interval	60-117	

Table 6: List of Variables

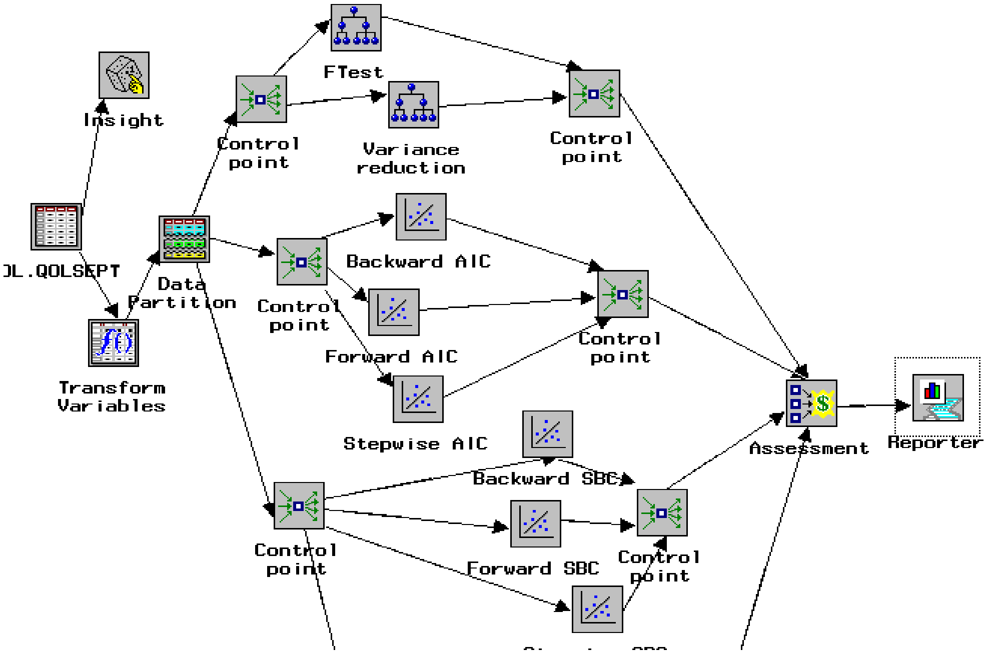

Figure 2: Process Flow Diagram – SAS Enterprise Miner

Model	Derived Variables	Variable Selection	Accuracy	Stability	Simplicity	Descriptiveness	Overall Score
DT-FTest	N	N	0.330	0.866	1.00	0.80	0.722
DT-FTest DV	Y	N	0.350	0.981	1.00	0.80	0.751
DT-Variance DV	Y	N	0.355	0.998	1.00	0.80	0.756
DT-FTest VS	N	Y	0.330	0.873	1.00	0.80	0.724
DT-Variance VS	N	Y	0.353	0.909	1.00	0.80	0.738
Reg-Backward AIC	N	N	0.254	0.495	0.50	0.65	0.463
Reg-Forward AIC	N	N	0.247	0.514	0.75	0.65	0.527
Reg-Stepwise AIC	N	N	0.338	0.839	1.00	0.65	0.682
Reg-Backward AIC VS	N	Y	0.320	0.682	0.75	0.65	0.583
Reg-Forward AIC VS	N	Y	0.315	0.680	0.75	0.65	0.581
Reg-Stepwise SBC VS	N	Y	0.344	0.951	1.00	0.65	0.706
Reg-Backward AIC DV	Y	N	0.134	0.269	0.75	0.65	0.444
Reg-Forward AIC DV	Y	N	0.277	0.685	1.00	0.65	0.633
Reg Splines	N	N	0.430	0.889	1.00	1.00	0.807

Table 7: Performance Scores of Models

The overall score of the models are examined in table 7 and the top models are selected to be examined by the decision makers. The multiple models allowed the decision maker to examine the dataset from multiple perspectives and provide alternate explanations for the phenomena under study. The results of decision trees DT-Variance–DV and DT-FTest are presented in tables 8 and 9. The previous studies had identified *Depression* and *Loneliness* as being important variables but in this study we are also able to identify the levels which affect the *QOL* of SCD patients. For example, the output in table 9 shows that the Lonely Score value of 21.27 is important in determining the *QOL* of SCD patients. The variable *Hospital Visits* was also identified as being an important determining factor for *QOL*. More frequent visits to hospital leads to lower *QOL* score. Identifying the variables and their values which affect the QOL score can help medical practitioners identify patients at risk and manage them better.

```
IF  LONELY_SCORE*DEP_SCORE < 37
THEN QOL : {AVE=101.145; SD=7.89344}

IF      LONELY_SCORE*DEP_SCORE > 355
THEN QOL : {AVE=80.5835;SD=10.9771}

IF  HOSPITAL_ADMISSIONS  ∈ {2,3,4,5}
     AND LONELY_SCORE*DEP_SCORE >= 37 and < 355
THEN QOL : {AVE=92.56; SD=10.355}

IF  EMPLOYED = 0
     AND HOSPITAL_ADMISSIONS = 1
     AND LONELY_SCORE*DEP_SCORE >= 37 and < 355
THEN QOL : {AVE=97.5807; SD=11.0052}

IF  AGE  <  27.81704
    AND EMPLOYED = 1
    AND HOSPITAL_ADMISSIONS = 1
    AND LONELY_SCORE*DEP_SCORE >=37 and < 355
THEN QOL : {AVE=103.723; SD=7.71358}

IF   AGE > 27.81704
        AND EMPLOYED = 1
         AND HOSPITAL_ADMISSIONS = 1
        AND LONELY_SCORE*DEP_SCORE >= 37 and < 355
 THEN  QOL : {AVE=101.505; SD=3.57751}
```

Table 8: Output of Decision Tree – DT-Variance- DV

```
IF  HOSPITAL_VISITS  IS ∈ {3,4,5}
    THEN QOL : {AVE=79.8658; SD=5.77232}

IF  LONELY_SCORE  <    21.27
AND HOSPITAL_VISITS  IS ∈   {1,2}
THEN QOL : {AVE=99.2433; SD=9.15824}

IF     LONELY_SCORE >=21.27
AND HOSPITAL_VISITS  IS ∈   {1,2}
THEN QOL : {AVE=87.9515; SD=13.3212}
```

Table 9: Output of Decision Tree - DT-FTest

4. Conclusions

This study highlights several benefits of using the KDDM process model. One of the benefits is that the process model ensures that the data mining goals are achieving the research objective. The process assists the data mining analyst to make the choices at each phase based on the knowledge garnered and the settings of the previous phases. KDDM recommends using multiple modeling techniques for a given research problem. Therefore, another benefit is that the data is analyzed from multiple perspectives. Also multiple performance measures are used therefore not only the analysis but also the assessment of the models is done from multiple perspectives. These performance measures are identified by the data mining analyst along with their value functions and corresponding weights. Adequate experimentation with the different modeling techniques and parameter setting ensure that we don’t just have 1 model but rather a set of best models. Therefore, it should be noted that the KDDM-based multi-criteria decision framework presented in this paper offers the researchers the opportunity to incorporate their informed subjective opinion about the relative importance of the different performance measures so that they can generate & investigate as many predictive, explanatory models as may be necessary to determine the most appropriate model without being overwhelmed by the information overload that might lead to the consideration of an insufficient number of models.

The results of the selected models can be examined by healthcare professionals to determine which models can be used to assist in the decision making process. The use of KDDM ensures that we are more likely to create models that can adequately identify patients at risk. This knowledge will assist healthcare workers in actively looking for problems and taking care of SCD patients to improve their quality of life.

References

Asnani, M.R., Fraser, R., Lewis, N.A. & Reid, M.E., 2010. Depression and loneliness in jamaicans with sickle cell disease. BioMed Central Psychiatry, 10 (40), pp. 1-7.

Cios, K., Teresinska, A., Konieczna, S., Potocka, J. & Sharma, S., 2000. Diagnosing myocardial perfusion from pect bull's-eye maps - a knowledge discovery approach. IEEE Engineering in Medicine and Biology Magazine 19 (4), pp. 17-25.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., 1996. From data mining to knowledge discovery in databases. AI Magazine, 17 (3), pp. 37-54.

Gil, K.M., Abrams, M.R., Phillips, G. & Keefe, F.J., 1989. Sickle cell disease pain: Relation of coping strategies to adjustment Journal of Consulting and Clinical Psychology, 57, pp. 725-731.

Jenerette, C., Funk, M. & Murdaugh, C., 2005. Sickle cell disease: A stigmatizing condition that may lead to depression. Issues Ment Health Nurs, 26, pp. 1081-1101.

Ko, M. & Osei-Bryson, K.-M., 2004. Using regression splines to assess the impact of information technology investments on productivity in the healthcare industry Information Systems Journal 14, pp. 43-63.

Kurgan, L.A. & Musilek, P., 2006. The survey of knowledge discovery and data mining process models. The Knowledge Engineering Review, 21 (1), pp. 1-24.

Mansingh, G., Rao, L., Osei-Bryson, K.-M. & Mills, A., 2013. Profiling internet banking users: A knowledge discovery in data mining process model based approach. Information Systems Frontiers, pp. 23, http://dx.doi.org/10.1007/s10796-012-9397-2.

Ohaeri, J.U., Shokunbi, W.A., Akinlade, K.S. & Dare, L.O., 1995. The psychosocial problems of sickle cell disease sufferers and their methods of coping. Soc Sci Med, 40 (7), pp. 955-960.

Osei-Bryson, K.-M., 2004. Evaluation of decision trees: A multi-criteria approach. Computer & Operations Research, 31, pp. 1933-1945.

Osei-Bryson, K.-M., 2012. A context-aware data mining process model based framework for supporting evaluation of data mining results. Expert Systems with Applications, 39 (1), pp. 1156-1164.

Sharma, S. & Osei-Bryson, K.-M., 2010. Toward an integrated knowledge discovery and data mining process model. The Knowledge Engineering Review, 25 (1), pp. 49-67.

Wilson Schaeffer, J.J., Gil, K.M., Burchinal, M., Kramer, K.D., Nash, K.B., Orringer, E. & Strayhorn, D., 1999. Depression, disease severity, and sickle cell disease. Journal of Behavioral Medicine, 22, pp. 115-126.