Data Analytics Using Data Mining
Arpit Garg¹, Asouv Bahl², Amit Rai³, Rohan Bhattacharjee⁴, Swarnalatha. P⁵
VIT University, Vellore, Tamil Nadu, India

Abstract:
The purpose of this project to acquire data sets from varying fields such as agriculture, banking, health care, government sectors, education etc. and then process the data before it can be mined. The core aim of this project is to mine the processed data for decision support, classification and prediction. Doing so we can also identify interesting patterns and trends from the data.

I. INTRODUCTION

The objectives of this project are as follows:
➢ To acquire data sets from reliable sources and determine the feasibility of mining. Once it has been established that the data can be mined for useful pattern/predictions further processing can be done
➢ Once the data has been processed mining would be done to provide a reliable model for prediction and decision support
➢ This project can be applied to virtually any field ranging from healthcare to environmental data
➢ Once reliable model has been developed it would drastically reduce the cost involved in marketing, man power, promote and early identification of events. Overall when data mining is precisely used it will lead to the growth and prosperity of any organization/sector/public. Project stakeholders: Project stakeholders:

II. PROJECT STAKEHOLDERS

The people and organization that are going to be affected by this work are called the stakeholders of the project. They are as follows:-
➢ Data repository
➢ Project team leader
➢ Project team members (Data Analysts)
➢ Organization interested in decision support and analysis.
➢ Open Source Communities/ Sectors in need of data analysis.

A. Characteristics

Functional requirement:
➢ To create a reliable data mining system that can provide accurate analysis of information locally.
➢ To provide option for various operation associated with data mining. This include frequent item data set mining, regression analysis, classification model, cluster analysis etc.
➢ Provide various techniques which are associated with the above operation such as apriori technique, decision tree, classification, k-means clustering, SVM (support vector machines) etc.

➢ To provide the mining results in a form i.e. easy to interpret and understand such as histograms, cluster plots, graphs. These would be in the form of pdf’s/jpeg.

Non-functional requirement
➢ 24 X 7 availability for decision support.
➢ This project is a learning project hence it efficiency cannot be guaranteed when it is deployed in a real time scenario.
➢ Automation of the 3 modules involved in the project. This is because each module has to be executed manually.

III. THREE MAJOR MODULE

The 3 main activities/modules are:

Data Acquisition: The data used for this project will be acquired through online data sets available from government data sets, private surveys companies. In addition to this we may also acquire the data by conducting online surveys.

Data Preprocessing: To get quality mining results we need to process the data to remove the inconsistencies, noise and incompleteness. This preprocessing would involve Data Cleaning, Data Integration, Data Reduction, Data Transformation and Data Discretization.

Data Mining: This is the extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data. This involves
a. Frequent Item Set mining
b. Classification & Prediction
c. Cluster Analysis and Outlier Analysis

A. Abbreviations and Acronyms

NPV-Net Present Value
PERT-Program Evaluation Review Technique
RAD- Rapid Application Approach
LOC - Lines of Code
COCOMO-Constructive Cost Model
TDEV-Transport Level Deviation

B. Units
• KDSI
• PM( Programer Months)
• PD( Programer Days)
C. Equations

The following equations were used in the work:

Earned Value (EV) = BCWP/BAC (1)

Here BCWP stands for Budgeted Cost of Work Scheduled and stands for Budget at completion.

Schedule Performance Index (SPI) = BCWP / BCWS (2)

BCWS stands for Budgeted Cost of Work Scheduled

Schedule Variance (SV) = BCWP – BCWS (3)

Cost Performance Index (CPI) = BCWP/ACWP (4)

Cost Variance (CV) = BCWP-ACWP (5)

Utility programs: PM=3.0* (KDSI)^1.12 (6)

COCOMO = 2.5* (PM)^0.35 (7)

TDEV = 2.5*(2.666)^0.35 (8)

D. WORK BREAKDOWN STRUCTURE

We have followed an activity based approach for the work breakdown structures of our project. This is because our project is strictly a service based project not a product based one. As shown in Fig 1.

IV. CALCULATION

Budgeted Cost of Work Performed (BCWP) = 44 programmer days

Budgeted Cost of Work Scheduled (BCWS) = 7 programmer days

Budget at Completion (BAC) = 85 programmer days

Earned Value (EV) = BCWP/BAC => 44/85 = 0.517 = 51.76 %

Schedule Performance Index (SPI) = BCWP/BCWS => 44/7 = 628%

Schedule Variance (SV) = BCWP – BCWS => 44 – 7 = 37

Cost Performance Index (CPI) = BCWP/ACWP => 44/62 = 70.96% ~71%

Cost Variance (CV) = BCWP – ACWP => 44 – 62= -18 Days (here “-” indicates the no. of days behind schedule)

A. RISK MANAGEMENT

The following risks are involved in the project:

- **Outdated Data**: If the data set used contains data that is severely outdated then the mined results may be irrelevant in the current time.

- **Irregularities in the recorded data**: This may be due to equipment malfunction, recording inconsistencies, missing data.

- **Complex Data**: If the data sets are too huge to process then the project can exceed its deadlines.

- **Ambiguous data**: Some real time data are too random to identify a confident mining sequence. This means that we can face a situation where we can’t provide a significant confidence percentage for the mined results.
TABLE 2. Index

<table>
<thead>
<tr>
<th>Index</th>
<th>Activity Description</th>
<th>Required Predecessor</th>
<th>Duration (weeks)</th>
<th>Critical path</th>
<th>EST</th>
<th>ECT</th>
<th>LST</th>
<th>LCT</th>
<th>Slack time</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Problem identification (here)</td>
<td>3</td>
<td>*</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>Abstract</td>
<td>A</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>12</td>
<td>13</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>Project planning</td>
<td>A</td>
<td>10</td>
<td>*</td>
<td>3</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>Survey</td>
<td>A</td>
<td>7</td>
<td>*</td>
<td>3</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>Data acquisition</td>
<td>ABC</td>
<td>7</td>
<td>*</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>Data preprocessing</td>
<td>E</td>
<td>5</td>
<td>*</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>Data mining</td>
<td>F</td>
<td>15</td>
<td>*</td>
<td>15</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>H</td>
<td>Pattern/Validation</td>
<td>G</td>
<td>1</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>I</td>
<td>Real-time validation</td>
<td>G</td>
<td>15</td>
<td>*</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>J</td>
<td>Documentation</td>
<td>C</td>
<td>15</td>
<td>*</td>
<td>13</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>7</td>
</tr>
</tbody>
</table>

V. REFERENCES


