CRC Construction Innovation

BUILDING OUR FUTURE

# Report
# Preliminary Report on Data Mining Techniques for Life Prediction

Research Project No: 2005-003-B-04


The research described in this report was carried out by:

| | |
|---|---|
| Project Leader | Penny Corrigan |
| Researchers | Angela Bradbury<br>Ivan Cole<br>Robin Drogemuller<br>Stephen Egan<br>Wayne Ganther<br>Tim Muster<br>David Paterson<br>Gerry Trinidad<br>Natalie Sherman<br>Andrew Martin<br>Wan Yee Chan<br>Richi Nayak<br>Esther Ge |
| Project Affiliates | Peter Hope<br>Michael Ball<br>Frank Turvey<br>Lee Wade<br>Wayne Muller<br>Lex Vanderstaay |



Research Program: B
Sustainable Built Assets

Project: 2005-2003-B
Learning System for Life Prediction of Infrastructure

Date: August 2006


Leaders in Construction and Property Research

## Distribution List

Cooperative Research Centre for Construction Innovation
Authors

## Disclaimer

The Client makes use of this Report or any information provided by the Cooperative Research Centre for Construction Innovation in relation to the Consultancy Services at its own risk. Construction Innovation will not be responsible for the results of any actions taken by the Client or third parties on the basis of the information in this Report or other information provided by Construction Innovation nor for any errors or omissions that may be contained in this Report. Construction Innovation expressly disclaims any liability or responsibility to any person in respect of any thing done or omitted to be done by any person in reliance on this Report or any information provided.



Please direct all enquiries to:

Chief Executive Officer
Cooperative Research Centre for Construction Innovation
9th Floor, L Block, QUT, 2 George St
Brisbane Qld 4000
AUSTRALIA
T: 61 7 3138 9291
F: 61 7 3138 9151
E: enquiries@construction-innovation.info
W: www.construction-innovation.info

# Table of Contents

Page
Table of Contents ..... iii
1 INTRODUCTION ..... 1
2 DATA MINING METHODS FOR PREDICTION ..... 1
3 DATA PRE-PROCESSING ..... 2
4 PREDICTIVE MODEL OVERVIEW ..... 2
5 CONCLUSIONS ..... 8

# 1 INTRODUCTION

This project is an extension of a previous CRC project (220-059-B) which developed a program for life prediction of gutters in Queensland schools. A number of sources of information on service life of metallic building components were formed into databases linked to a Case-Based Reasoning Engine which extracted relevant cases from each source. In the initial software, no attempt was made to choose between the results offered or construct a case for retention in the casebase.

In this phase of the project, alternative data mining techniques will be explored and evaluated. A process for selecting a unique service life prediction for each query will also be investigated. This report summarises the initial evaluation of several data mining techniques.

# 2 DATA MINING METHODS FOR PREDICTION

Data mining is a powerful technology to solve prediction problems. A literature review has been carried out and a number of data mining methods have been found that could be considered for the current problem. These include:

- Naïve Bayes
- K Nearest Neighbours (KNN)
- Decision Tree (DT)
- Neural Network (NN)
- Rough Set Theory (RS)
- Support Vector Machine (SVM)
- Bagging
- Boosting

Naïve Bayes is a statistical-based algorithm. KNN is very similar to case-based reasoning. It is based on the use of distance measures. Both Decision tree and Neural Network have been successfully applied into the prediction area and there are some similar applications using these methods. Rough set and support vector machine are relatively new methods. Rough set is very efficient in finding hidden patterns in data while SVM solves the problem of efficient learning from a limited training set.

Bagging and boosting are methods for improving the predictive performance. They generate multiple predictors and use these to get an aggregated predictor, which has better performance. All these methods have their advantages and disadvantages. The results usually depend on the data set on which they operate.

# 3 DATA PRE-PROCESSING

Data pre-processing is used to make data ready for mining. The data quality is a key aspect that influences the results of data mining. Raw data generally include many noisy, inconsistent and missing values and redundant information. Before building a prediction model, the data must be preprocessed.

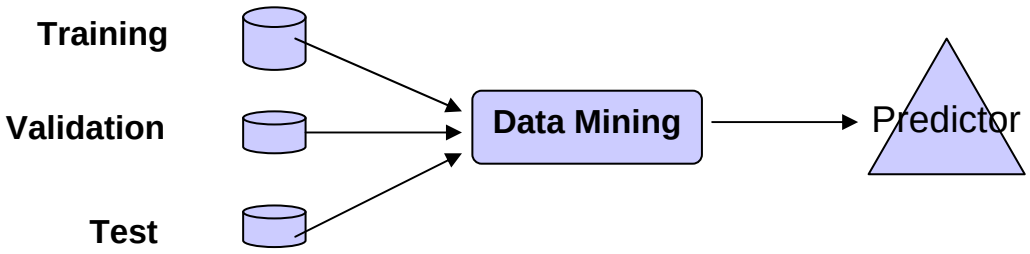
This involves three processes:

- Feature Selection - attributes that only provide identification information and only have one value are ignored (eg. Centre Code, Centre Name and LocID)
- Data Cleaning – this deals with inconsistent and missing values
    - the use of lowercases and capitals (eg. ‘Steel’ and ‘steel’)
    - different spellings/words but same meaning (eg. ‘Galvanised’ and ‘Galvanized’, ‘Steel in Hardwood’ and ‘Steel-Hardwood’)
    - More spaces are included in values (eg. ‘Residential ‘ and ‘Residential ’ )
- Data Discretization - all numeric attributes including target attributes are discretized to a nominal type by dividing them into ranges before applying Naïve Bayes and Decision Tree methodology
    - For example, ‘Mean’ contains values from 3 to 58. It is divided into 10 ranges: [3-13], (13-17], (17-21], (21-25], (25-29], (29-33], (33-37], (37-41], (41-45], (45-58].

# 4 PREDICTIVE MODEL OVERVIEW

The main objective in this research is to make an accurate prediction for the lifetime of metallic components. Therefore, the problem is a prediction data mining problem. After pre-processing, various data mining methods can be applied on new data sets to build prediction models. When performing a prediction task, a data set is usually separated into three parts: training set, validation set and test set. The training set is used to build a prediction model during a training phase. Usually when the target is a continuous value, the prediction model is called a predictor. The validation set is often used for improving the prediction accuracy. Finally the test set is used for evaluating the model. After the model is built, new data can be input into the predictor to produce the predicted result. This is illustrated in Figure 1.

**The data set is divided into three parts**




**Based on previous data, future events can be predicted:**

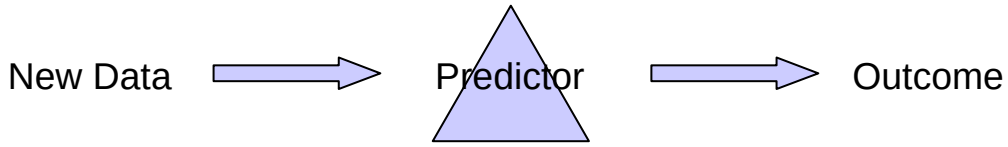



Figure 1: How data mining is used to predict outcomes

The workflow for the current problem is illustrated in Figure 2.

Data mining methods are applied to all three data sets to build three predictors first. After that, these three predictors can make predictions for the user's inputs. The final predicted life is either a multiple choice provided by three predictors or a value combined from the outputs of three predictors. This still needs to be determined.

In order to get accurate predicted results, various data mining methods including Naïve Bayes, K Nearest Neighbors (KNN), Decision Tree (DT), Neural Network (NN) and Support Vector Machine (SVM) have been applied on these data sets.

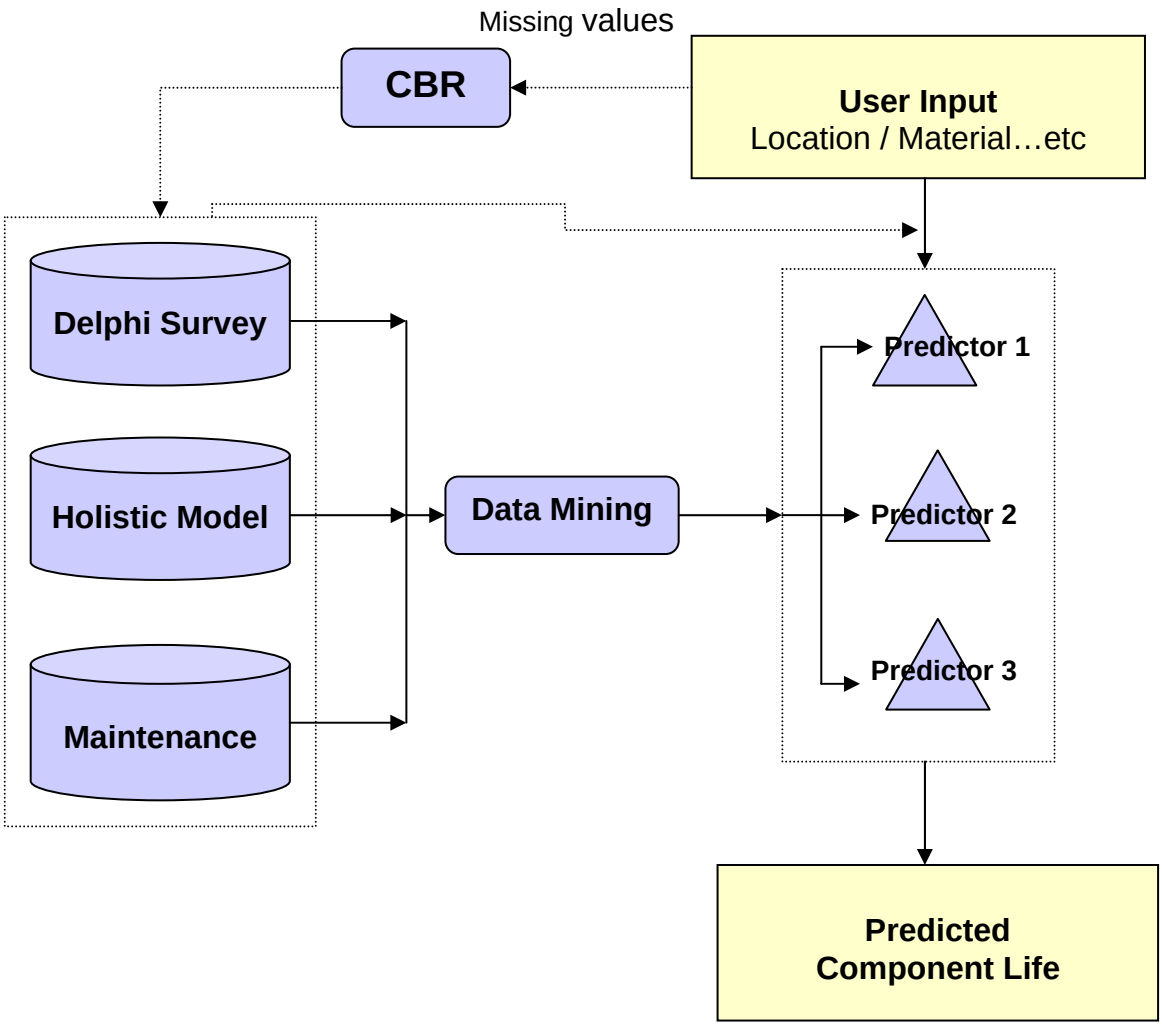



Figure 2: Diagram of life prediction application

Figure 3 shows an example of using the DT model for the Holistic data set. Each internal node is labeled with an attribute. Every path from a root node to a leaf node forms a rule. An example is:

IF 8.7320616733 <= SALannual AND
Material EQUALS ZINCALUME AND
GutterMaintenance EQUALS NOT CLEANED
THEN
NODE : 13
N : 30
AVE : 40.6522
SD : 6.83028

N: the number of non-missing observations
AVE: average of MLannual
SD: standard deviation

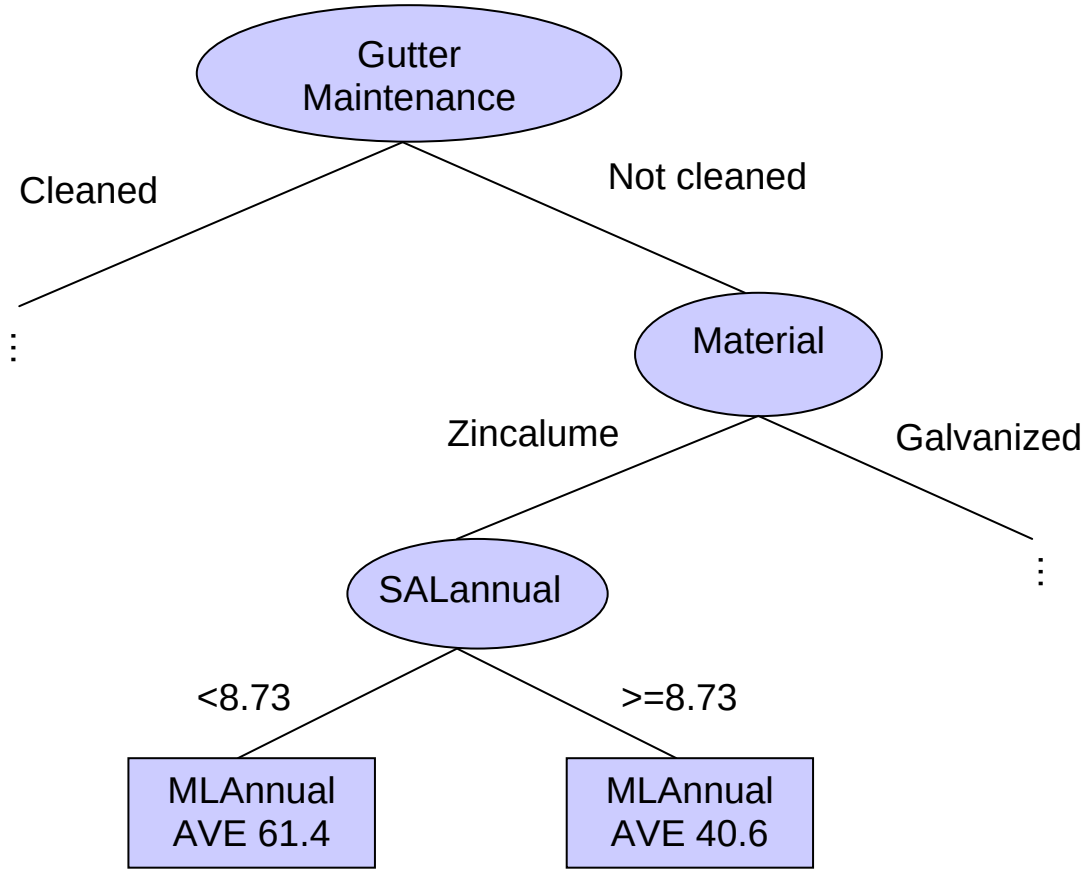



Figure 3. This is an example of using the DT theory on the Holistic Model dataset

Figure 4 shows an example of the Neural Network model, as applied to the Colorbond data. The input layer includes all attributes used for prediction. Every attribute is associated with a weight. The hidden layer includes three nodes. Only one target is in the output layer. NN is efficient for predicting a numerical target. It adjusts the weights to approximate the target value. But it is very difficult to extract rules from NN because of its complex structure.

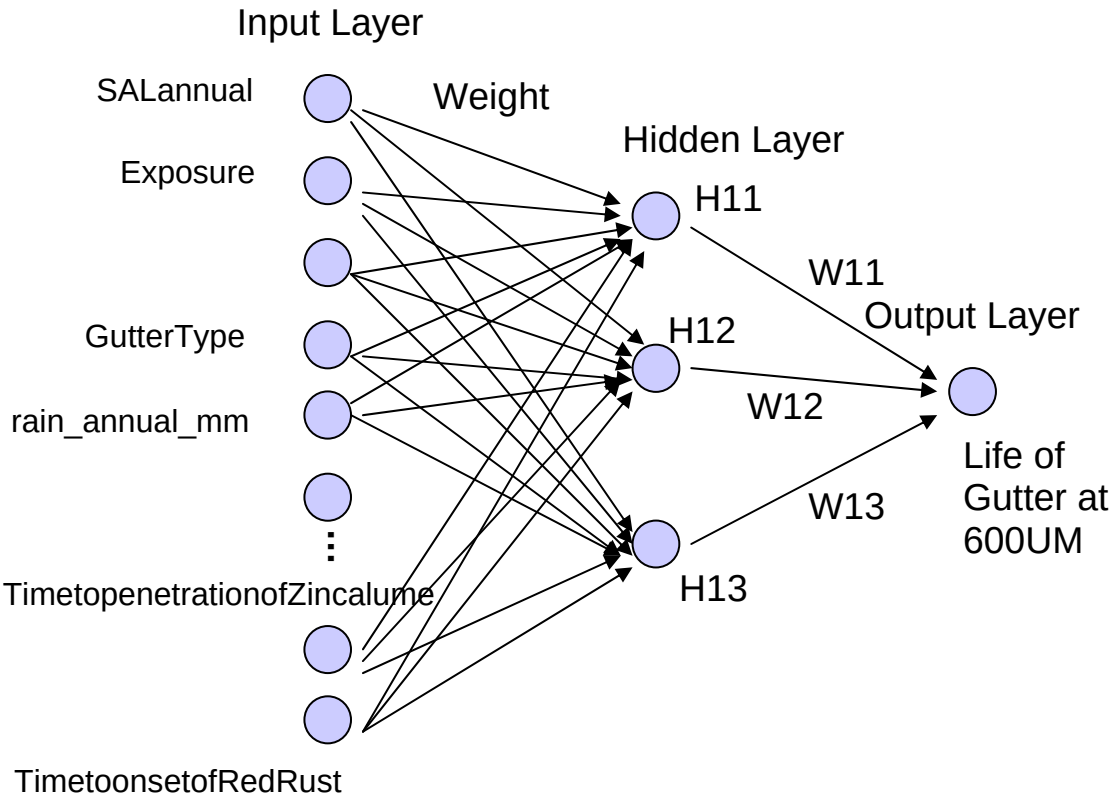



Figure 4. An example of the Neural Network model applied to the Colorbond data

## 4.1 Comparison of Results

Two different methods are used to compare the different data mining techniques depending on whether the numeric attributes are discretized or continuous.

Methods such as Naïve Bayes and Decision Tree data mining techniques have numeric attributes discretized to a nominal type including the target value. That means all continuous values are divided into ranges. So the targets became several predefined classes, our case became a supervised classification, and prediction accuracy can be used to measure performance, where:

Prediction accuracy = number of correctly classified instances / total number of instances

Table 1: Prediction accuracy for Naïve Bayes and Decision Tree techniques

| **Data Set** | **Prediction Accuracy** | |
|---|---|---|
| | **Naïve Bayes** | **Decision Tree** |
| Delphi | 45.748% | 54.985% |
| Holistic | 89.844% | 91.525% |
| ColorBond | 94.728% | 96.548% |
| Maintenance for Galvanized | 93.138% | 94.603% |
| Maintenance for Zincalume | 91.904% | 93.215% |

K Nearest Neighbours (KNN), Neural Network (NN) and Support Vector Machine (SVM) are three methods that can be used to predict continuous values. In this instance, a correlation coefficient is used to measure performance instead of prediction accuracy. Correlation coefficient measures the statistical correlation between the predicted and actual values. A prediction is good when the correlation coefficient is as large as possible. 1 is the best.

Table 2: Correlation Coefficients for KNN, NN and SVM techniques

| **Data Set** | **Correlation Coefficient** | | |
|---|---|---|---|
| | **KNN** | **NN** | **SVM** |
| Delphi | 0.8684 | 0.9539 | 0.9582 |
| Holistic | 0.9962 | 0.977 | 0.8408 |
| ColorBond | 0.9962 | 1 | 0.9999 |
| Maintenance for Galvanized | 0.9915 | 0.9994 | 0.9737 |
| Maintenance for Zincalume | 0.9886 | 0.999 | 0.9889 |

In this manner the best method for each data set is identified, which is SVM for Delphi Survey, KNN for Holistic Model and NN for Colorbond and Maintenance database with bagging. We find that a better correlation coefficient can be obtained using bagging for SVM, KNN and NN. It indicates that bagging is more accurate than the individual predictors.

Table 3: Comparison of effect of bagging

| **Data Set** | **Correlation Coefficient** | |
|---|---|---|
| | **SVM / KNN / NN** | **Bagged SVM / KNN / NN** |
| Delphi | 0.9582 | 0.9608 |
| Holistic | 0.9962 | 0.9967 |
| ColorBond | 1 | 1 |
| Maintenance for Galvanized | 0.9994 | 0.9997 |
| Maintenance for Zincalume | 0.999 | 0.9995 |

## 4.2 Existing Problems

Problems arising from this analysis of the data mining techniques are:

- No one data mining method is always best for all the data sets, and
- Large differences may exist among the predicted values from the three predictors, eg. one test case using Windsor State School / Roof / Zincalume / Maintenance = Yes as inputs gives:
    - Delphi: 52.877
    - Maintenance dataset: 29.928

  as the predicted service life.

## 4.3 Possible Solutions

A possible solution to this problem (Figure 5) is to build a knowledge base which contains all results we have got from past cases. When new data are presented, we first go to knowledge base to find a matching case. If yes, give the result, otherwise, go to the predictors to make the prediction. We may need to post-process the results using knowledge base if they are contradictory. The key of this solution is the construction of a knowledge base, which should be manual. Experience and

knowledge of domain experts can guide or assist with the construction of a knowledge base. The cases covered by rules in the knowledge base should be as many as possible. As a result, this solution can be human-cooperated mining

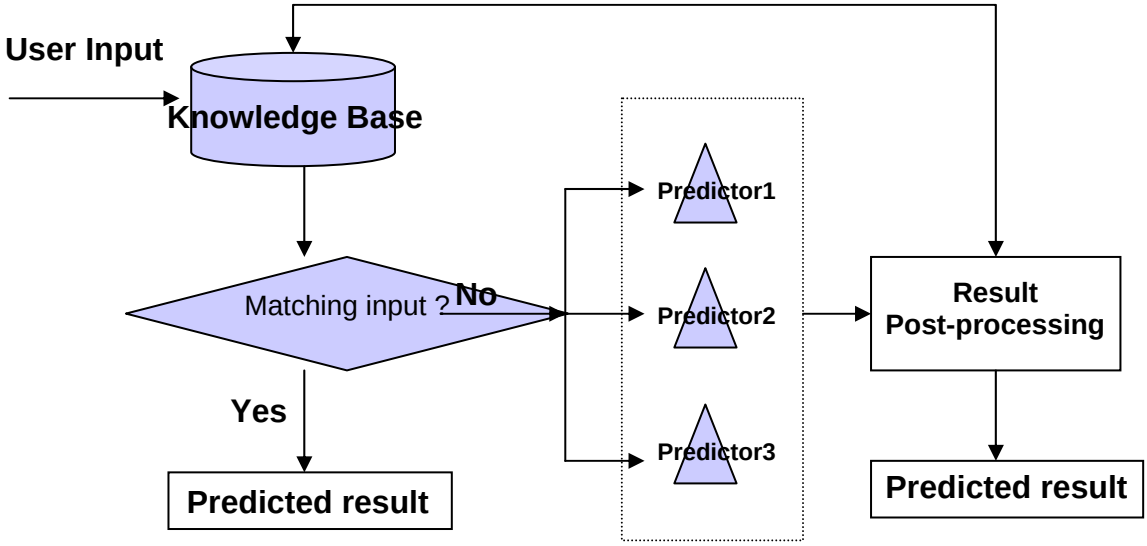



Figure 5. Proposed solution to conflicting results

# 5 CONCLUSIONS

The main objective of this project is to accurately predict the lifetime of metallic components. We intend to explore the various methods of data mining that can be considered to solve prediction problem. The data sets we are using include Delphi Survey, Holistic Model & Colorbond for Gutters and a Field data set for Roofs (' Maintenance') which is a combination of modeling and field measurement.

The work that has been done so far is as follows:

- Literature review on various predictive data mining methods
- Data pre-processing to make data ready for mining
- Experimental analysis of the data sets with traditional data mining methods such as Naïve Bayes, Decision Tree, K Nearest Neighbours, Neural Network and Support Vector Machine
- Selection of a best model for each data set (SVM model for Delphi Survey, KNN model for Holistic, NN for Colorbond & Field)
- Use of bagging on best models for improving performance
- Analysis of predicted results using some test cases
- A paper which has been submitted to AusDM (Australasian Data Mining Conference)

From the results we have obtained, we found that the prediction from Delphi Survey is not as good as we expected. Another main problem is that there is a confliction between the results from three data sets for a same test case. To solve this problem, we propose a possible solution. The basic idea of this solution is to build a

knowledge base for post-processing the results. The knowledge base will contain a set of rules which can be used to remove illogical results.

The main future tasks are summarized as follows:

- Analyse of Decision Tree Model for Delphi Survey to see what is the reason of poor results
- Implementation of Regression tree to see if the results for Delphi Survey can be better
- Implementation of Boosting on best models to see if the performance can be improved more
- Construction of the Knowledge base
- Pre-process the user input using matching
- Maintenance database input methods and ongoing update

Pre-processing user input is for dealing with missing values. Because the models are built on data sets with few missing values. The final users may not provide so many parameters, so we have to do pre-processing first.

**Cooperative Research Centre for Construction Innovation**

9th Floor, L Block
QUT Gardens Point
2 George Street
BRISBANE QLD 4001
AUSTRALIA

Tel: +61 7 3138 9291

Fax: +61 7 3138 9151

Email:
enquiries@construction-innovation.info

Web:
www.construction-innovation.info

Established and supported under the Australian Government's Cooperative Research Centres Program