A PREDICTIVE SYSTEM FOR DETECTION OF BANKRUPTCY USING MACHINE LEARNING TECHNIQUES

Kalyan Nagaraj and Amulyashree Sridhar

PES Institute of Technology, India

ABSTRACT

Bankruptcy is a legal procedure that claims a person or organization as a debtor. It is essential to ascertain the risk of bankruptcy at initial stages to prevent financial losses. In this perspective, different soft computing techniques can be employed to ascertain bankruptcy. This study proposes a bankruptcy prediction system to categorize the companies based on extent of risk. The prediction system acts as a decision support tool for detection of bankruptcy

KEYWORDS

Bankruptcy, soft computing, decision support tool

1. INTRODUCTION

Bankruptcy is a situation in which a firm is incapable to resolve its monetary obligations leading to legal threat. The financial assets of companies are sold out to clear the debt which results in huge financial losses to the investors. Bankruptcy results in decreased liquidity of capital and minimized financial improvement. It is reported by World Bank data that Indian government resolves the insolvency in an average of 4.3 years [1]. There is a need to design effective strategies for prediction of bankruptcy at an earlier stage to avoid financial crisis. Bankruptcy can be predicted using mathematical techniques, hypothetical models as well as soft computing techniques [2]. Mathematical techniques are primary methods used for estimation of bankruptcy based on financial ratios. These methods are based on single or multi variable models. Hypothetical models are developed to support the theoretical principles. These models are statistically very complex based on their assumptions. Hence soft computing techniques are extensively used for developing predictive models in finance. Some of the popular soft computing techniques include Bayesian networks, logistic regression, decision tress, support vector machines and neural networks.

In this study, different machine learning techniques are employed to predict bankruptcy. Further based on the performance of the classifiers, the best model is chosen for development of a decision support system in R programming language. The support system can be utilized by stock holders and investors to predict the performance of a company based on the nature of risk associated.

DOI : 10.5121/ijdkp.2015.5103

2. BACKGROUND

Several studies have been conducted in the recent past reflecting the importance of machine learning techniques in predictive modelling. The studies and the technologies implemented are briefly discussed below.

2.1. MACHINE LEARNING

Machine learning techniques are employed to explore the hidden patterns in data by developing models. It is broadly referred as knowledge discovery in database (KDD). Different learning algorithms are implemented to extract patterns from data. These algorithms can either be supervised or unsupervised. Supervised learning is applied when the output of a function is previously known. Unsupervised learning is applied when the target function is unknown. The general layout for machine learning process is described below:

Data collection: The data related to domain of concern is extracted from public platforms and data warehouses. The data will be raw and unstructured format. Hence pre-processing measures must be adopted

Data pre-processing: The initial dataset is subjected for pre-processing. Pre-processing is performed to remove the outliers and redundant data. The missing values are replaced by normalization and transformation

Development of models: The pre-processed data is subjected to different machine learning algorithms for development of models. The models are constructed based on classification, clustering, pattern recognition and association rules

Knowledge Extraction: The models are evaluated to represent the knowledge captured. This knowledge attained can be used for better decision making process [3].

2.2. CLASSIFICATION ALGORITHMS

Several classification algorithms are implemented in recent past for financial applications. They are discussed briefly below:

Logistic Regression: It is a classifier that predicts the outcome based probabilities of logistic function. It estimates the relationship between different independent variables and the dependent outcome variable based on probabilistic value. It may be either binary or multinomial classifier. The logistic function is denoted as:

$$F(x) = \frac{1}{1+e^{-(\beta0+\beta1x)}}$$

β_0 and $\beta_1$are coefficients for input variable x. The value of F(x) ranges from zero to one. The logistic regression model generated is also called as generalized linear model [4].

Naïve Bayes classifier: It is a probabilistic classifier based on the assumptions of Bayes theorem [5]. It is based on independent dependency among all the features in the dataset. Each feature contributes independently to the total probability in model. The classifier is used for supervised learning. The Bayesian probabilistic model is defined as:

$$p(C_k \mid x) = \frac{p(C_k)p(x \mid C_k)}{p(x)}$$

$p(C_k|x)$ = posterior probability

$p(C_k)$=prior probability

p(x)= probability of estimate

$p(x|C_k)$=likelihood of occurrence of x

Random Forest: They are classifier which construct decision trees for building the model and outputs the mode value of individual trees as result of prediction. The algorithm was developed by Breiman [6]. Classification is performed by selecting a new input vector from training set. The vector is placed at the bottom of each of the trees in the forest. The proximity is computed for the tree. If the tree branches are at the same level, then proximity is incremented by one. The proximity evaluated is standardized as a function of the number of trees generated. Random forest algorithms compute the important features in a dataset based on the out of bag error estimate. The algorithm also reduces the rate of overfitting observed in decision tree models.

Neural networks: They are learning algorithms inspired from the neurons in human brain. The network comprises of interconnected neurons as a function of input data [7]. Based on the synapse received from input data, weights are generated to compute the output function. The networks can either be feed-forward or feed-back in nature depending upon the directed path of the output function. The error in input function is minimized by subjecting the network for back-propagation which optimizes the error rate. The network may also be computed from several layers of input called as multilayer perceptron. Neural networks have immense applications in pattern recognition, speech recognition and financial modeling.

Support vector machine: They are supervised learning algorithms based on non-probabilistic classification of dataset into categories in high dimensional space. The algorithm was proposed by Vapnik [8]. The training dataset is assumed as a p-dimensional vector which is to be classified using (p-1) dimensional hyperplane. The largest separation achieved between data points is considered optimal. Hyperplane function is represented as:

$$f(x, < w, b >) = sign(w.x + b)$$

w = normalized vector to the hyperplane
x = p-dimensional input vector
b = bias value

The marginal separator is defined as $2|k|/||w||$. 'k' represents the number of support vectors generated by the model. The data instances are classified based on the below criteria
If $(w \cdot x + b) = k$, indicates all the positive instances.
If $(w \cdot x + b) = -k$, indicates the set of negative instances.
If $(w \cdot x + b) = 0$, indicates the set of neutral instances.

3. RELATED WORK

Detection of bankruptcy is a typical classification problem in machine learning application. Development of mathematical and statistical models for bankruptcy prediction was initiated by Beaver in the year 1960 [9]. The study focused on the univariate analysis of different financial factors to detect bankruptcy. An important development in this arena was recognized by Altman who developed a multivariate Z-score model of five variables [10]. Z-score model is considered as a standard model for estimating the probability of default in bankruptcy. Logistic regression was also instigated to evaluate bankruptcy [11]. These techniques are considered as standard estimates for prediction of financial distress. But these models pose statistical restrictions leading to their limitations. To overcome these limitations probit [12] and logit models [13] were implemented for financial applications. In later years, neural networks were implemented for estimating the distress in financial organizations [14, 15, and 16]. Neural networks are often subjected to overfitting leading to false predictions. Decision trees were also applied for predicting financial distress [17, 18]. Support vector machines have also been used employed in predicting bankruptcy for financial companies [19, 20]. In recent years, several hybrid models have been adopted to improve the performance of individual classifiers for detection of bankruptcy [21, 22].

4. METHODOLOGY

4.1. Collection of Bankruptcy dataset

The qualitative bankruptcy dataset was retrieved from UCI Machine Learning Repository [23]. The dataset comprised of 250 instances based on 6 attributes. The output had two classes of nominal type describing the instance as 'Bankrupt' (107 cases) or 'Non-bankrupt' (143 cases).

4.2. Feature Selection

It is important to remove the redundant attributes from the dataset. Hence correlation based feature selection technique was employed. The feature based algorithm selects the significant attributes based on the class value. If the attribute is having high correlation with the class variable and minimal correlation with other attributes of the dataset it is presumed to be a good attribute. If the attribute have high correlation with the attributes then they are discarded from the study.

4.3. Implementing machine learning algorithms

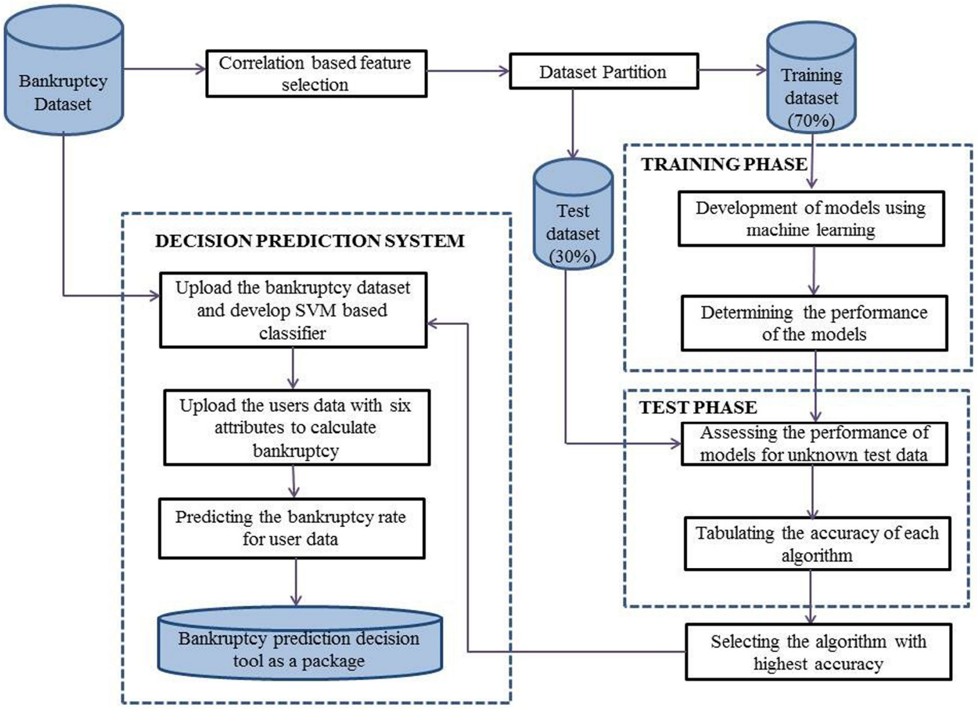
The features selected after correlational analysis are subjected to data partitioning followed by application of different machine learning algorithms. The dataset is split into training (2/3rd of the dataset) and test dataset (1/3rd of the dataset) respectively. In the training phase different classifiers are applied to build an optimal model. The model is validated using the test set in the testing phase. Once the dataset is segregated, different learning algorithms was employed on the training dataset. The algorithms include logistic regression, Bayesian classifier, random forest, neural network and support vector machines. Models generated from each of the classifier were assessed for their performance using the test dataset. A ten-fold cross validation strategy was adopted to test the accuracy. In this procedure, the test dataset is partitioned into ten subsamples and each subsample is used to test the performance of the model generated from training dataset. This step is performed to minimize the probability of overfitting. The accuracy of each algorithm was estimated from the cross validated outcomes.

4.4. Developing a predictive decision support system

From the previous step, the classifier with highest prediction accuracy is selected for developing a decision support system to predict the nature of bankruptcy. The prediction system was implemented in RStudio interface, a statistical programming toolkit. Different libraries were invoked for development of the predictive system including 'gWidgets 'and 'RGtk2'. The predictive tool develops a model for evaluating the outcome bankruptcy class for user input data. Predicted class is compared with the actual class value from the dataset to compute the percentage of error prediction from the system. The support system estimates the probability of bankruptcy among customers. It can be used as an initial screening tool to strengthen the default estimate of a customer based on his practises.

The methodology of this study is illustrated in Figure 1.

Figure 1: The flowchart for developing a decision support system to predict bankruptcy


Bankruptcy Dataset
Correlation based feature selection
Dataset Partition
Training dataset (70%)
Test dataset (30%)
TRAINING PHASE
Development of models using machine learning
Determining the performance of the models
TEST PHASE
Assessing the performance of models for unknown test data
Tabulating the accuracy of each algorithm
Selecting the algorithm with highest accuracy
DECISION PREDICTION SYSTEM
Upload the bankruptcy dataset and develop SVM based classifier
Upload the users data with six attributes to calculate bankruptcy
Predicting the bankruptcy rate for user data
Bankruptcy prediction decision tool as a package

5. Results And Discussion

5.1. Description of Qualitative Bankruptcy dataset

The bankruptcy dataset utilized for this study is available at UCI Machine Learning Repository. The dataset comprising of six different features is described in Table 1. The distribution of class outcome is shown in Figure-2.

Table 1: Qualitative Bankruptcy Dataset. (Here P=Positive, A=Average, N=Negative, NB=Non-Bankruptcy and B=Bankruptcy)

Sl. No	Attribute Name	Description of attribute
01.	IR (Industrial Risk)	Nominal {P, A, N}
02.	MR (Management Risk)	Nominal {P, A, N}
03.	FF (Financial Flexibility)	Nominal {P, A, N}
04.	CR (Credibility)	Nominal {P, A, N}
05.	CO (Competitiveness)	Nominal {P, A, N}
06.	OR (Operating Risk)	Nominal {P, A, N}
07.	Class	Nominal {NB, B}

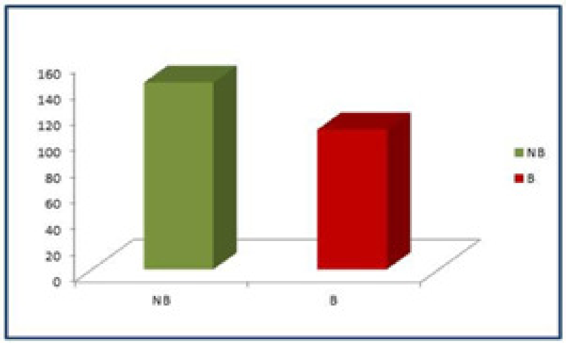
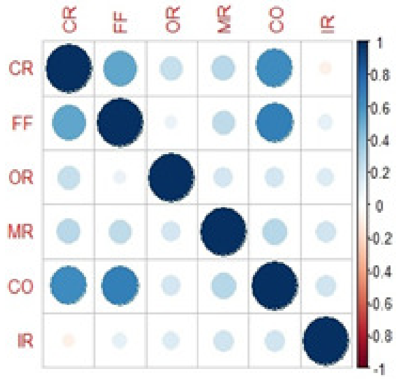


Figure 2: The distribution of class output representing Non Bankruptcy and Bankruptcy

5.2. Correlation based Attribute selection

The bankruptcy dataset was subjected for feature selection to extract the relevant attributes. The nominal values in bankruptcy dataset were converted to numeric values for performing feature selection. The values of each of the descriptors were scaled as P=1, A=0.5 and N=0 representing the range for positive, average and negative values. The procedure was repeated for all the six attributes in dataset. Pearson correlation filter was applied for the numerical dataset to remove the redundant features with a threshold value of 0.7. The analysis revealed that all the six attributes were highly correlated with the outcome variable. In order to confirm the results from correlation, another feature selection method was applied for the dataset. Information gain ranking filter method was applied to test the importance of features. The algorithm discovered similar results as that of correlation. Hence all the six attributes from the dataset were considered for the study. The correlational plot for the features is shown in Figure 3.

Figure 3: The correlational plot illustrating the importance of each feature



5.3. Machine learning algorithms

The features extracted from previous step were subjected for different machine learning algorithms in R. The algorithms were initially applied for the training set to develop predictive models. These models were further evaluated using the test set. Performance of each model was adjudged using different statistical parameters like confusion matrix and receiver operating characteristics (ROC) curve. Confusion matrix is a contingency table that represents the performance of machine learning algorithms [24]. It represents the relationship between actual class outcome and predicted class outcome based on the following four estimates:

a) True positive (TP): The actual negative class outcome is predicted as negative class from the model

b) False positive (FP): The actual negative class outcome is predicted as a positive class outcome. It leads to Type-1 error

c) False negative (FN): The actual positive class outcome is predicted as negative class from the model. It leads to Type-2 error

d) True negative (TN): The actual class outcome excluded is also predicted to be excluded from the model

Based on these four parameters the performance of algorithms can be adjudged by calculating the following ratios.

$$Accuracy(\%) = \frac{TP + TN}{TP + FP + TN + FN}$$

$$TPR(\%) = \frac{TP}{TP + FN}$$

$$FPR(\%) = \frac{FP}{FP + TN}$$

$$precision(\%) = \frac{TP}{TP + FP}$$

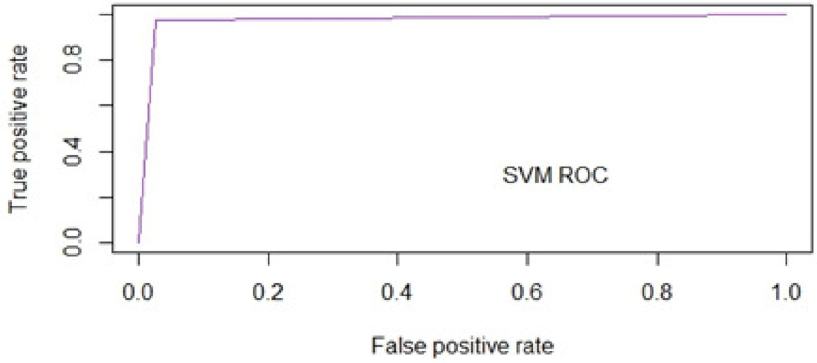
ROC curve is a plot of false positive rate (X-axis) versus true positive rate (Y-axis). It is represents the accuracy of a classifier [25].

The accuracy for all the models was computed and represented in Table 2. SVM classifier achieved better accuracy compared to other machine learning algorithms. Henceforth the ROC plot of RBF-based SVM classifier is represented in Figure 4.

Table 2: The accuracy of bankruptcy prediction of machine learning algorithms

Sl. No	**Algorithm**	**Library used in R**	**Accuracy of prediction (%)**	**True positive rate**	**False positive rate**	**Precision**
01.	Logistic regression	glmnet	97.2	0.972	0.028	0.97
02.	Rotation forest	randomForest	97.4	0.974	0.026	0.97
03.	Naïve Bayes	e1071	98.3	0.983	0.017	0.98
04.	Neural network	neuralnet	98.6	0.986	0.014	0.98
05.	RBF-based Support vector machine	e1071	99.6	0.996	0.004	0.99

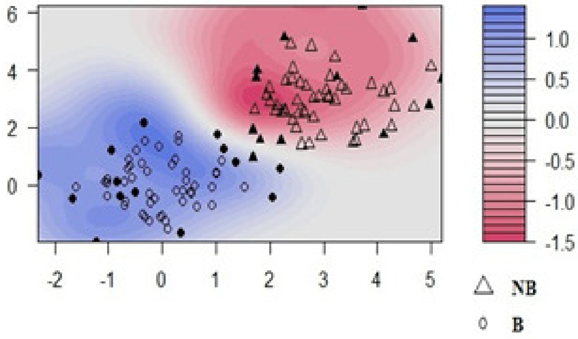
Figure 4: The ROC curve representing the accuracy of SVM classifier



5.4. SVM based decision supportive system in R

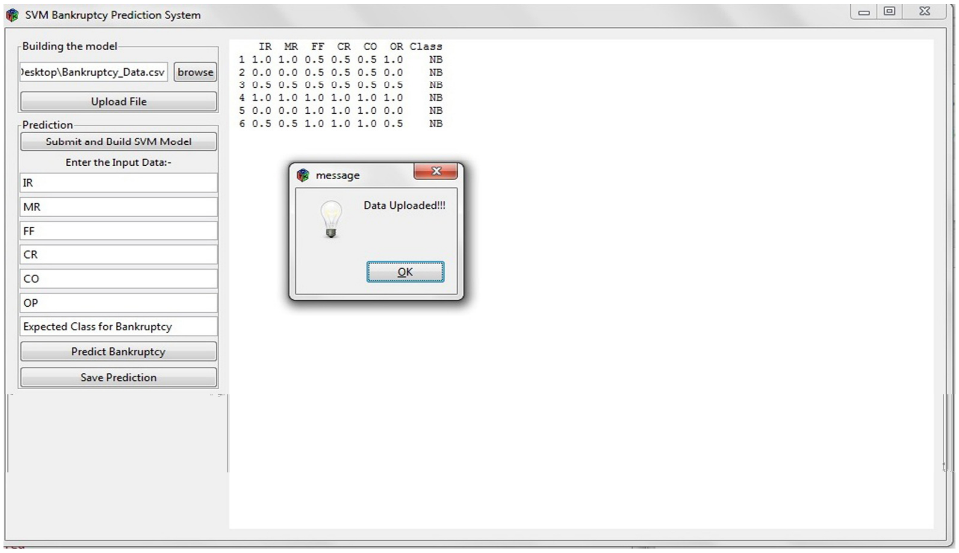
Based on the accuracy in previous step, it was seen that support vector based classifier outperformed other machine learning techniques. The classifier was implemented using radial basis function (RBF) kernel [26]. It is also referred as Gaussian RBF kernel. The kernel representation creates a decision boundary for the non-linear attributes in high dimensional space. The attributes are converted to linear form by mapping using this kernel function. An optimal hyperplane is constructed in feature space by considering the inner product of the kernel. Hyperplane is considered as optimal if it creates a widest gap from the input attributes to the target class. Furthermore, to achieve optimization C and gamma parameters are used. C is used to minimize the misclassification in training dataset. If the value of C is smaller it is soft margin creating a wider hyperplane, whereas the value of C being larger leads to overfitting called as hard margin. Hence the value of C must be selected by balancing between the soft and hard margin. Gamma is used for non-linear classifiers for constructing the hyperplane. It is used to control the shape of the classes to be separated. If the value of gamma is small it results in high variance and minimal bias producing a pointed thrust. While a bigger gamma value leads to minimal variance and maximum bias producing a broader and soft thrust. The values of C and gamma were optimized and selected for classification. The classified instances from RBF kernel is observed in Figure 5.

Figure 5: RBF classifier based classification for bankruptcy dataset as either NB or B



Based on the RBF classifier the prediction system was constructed in R. The bankruptcy dataset is initially loaded into the predictive system as a .csv file. The home page of predictive tool loaded with bankruptcy dataset is shown in Figure 6.

Figure 6: SVM based predictive tool with the bankruptcy dataset



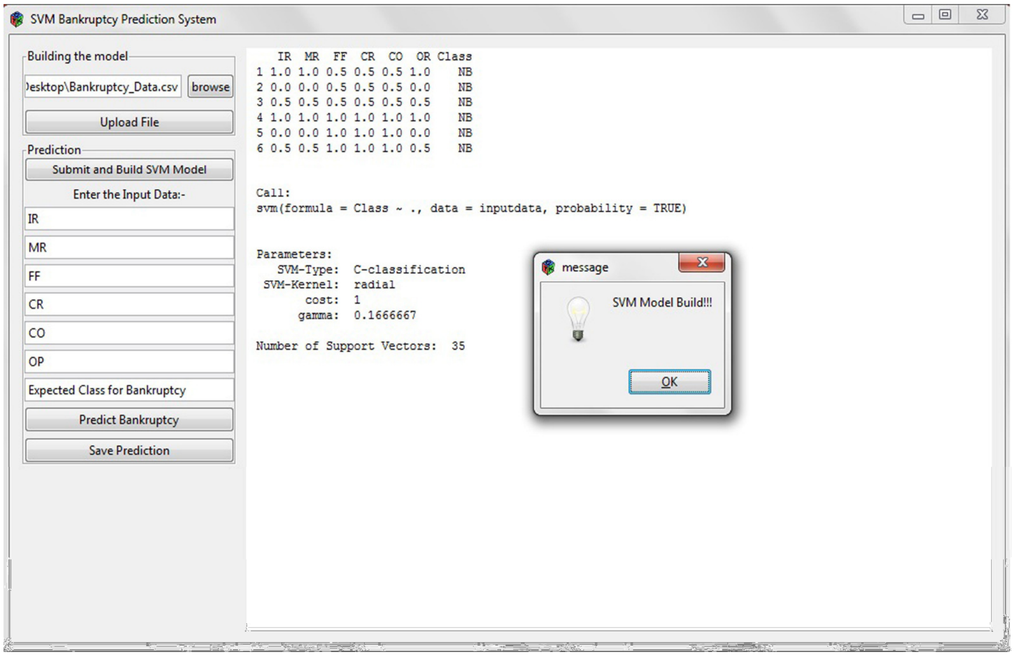
)esktop\Bankruptcy_Data.csv

```
    IR  MR  FF  CR  CO  OR Class
1 1.0 1.0 0.5 0.5 0.5 1.0    NB
2 0.0 0.0 0.5 0.5 0.5 0.0    NB
3 0.5 0.5 0.5 0.5 0.5 0.5    NB
4 1.0 1.0 1.0 1.0 1.0 1.0    NB
5 0.0 0.0 1.0 1.0 1.0 0.0    NB
6 0.5 0.5 1.0 1.0 1.0 0.5    NB
```

Data Uploaded!!!

The system fetches the dataset and stores as a dataframe. Dataframe is a vector list used to store the data as a table in R. RBF-kernel SVM model is developed for the bankruptcy dataset. It is displayed in Figure 7.

Figure 7: Radial based SVM model developed for bankruptcy dataset



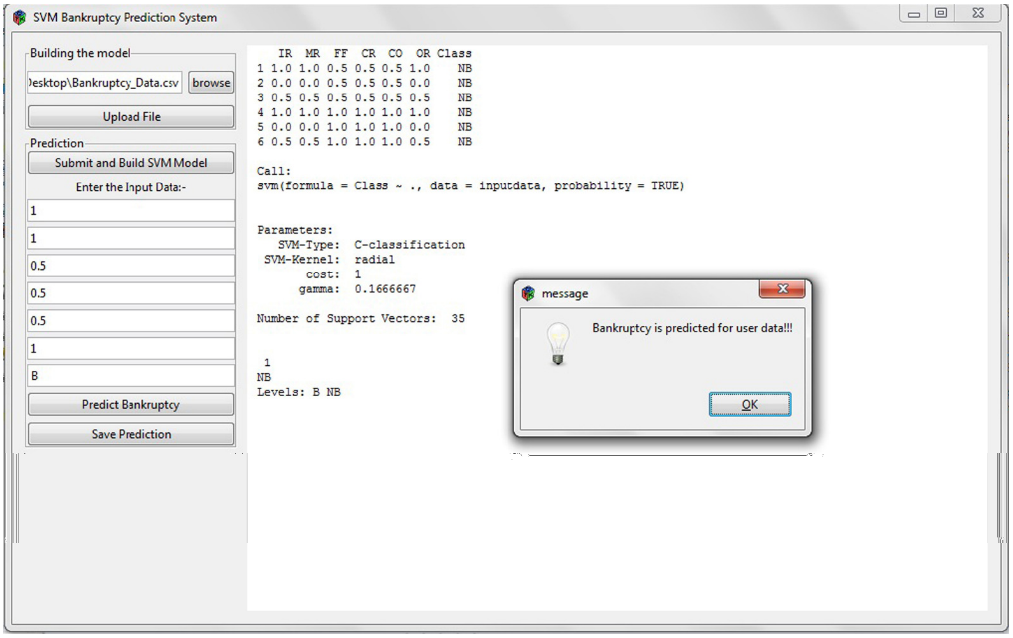
)esktop\Bankruptcy_Data.csv

```
    IR  MR  FF  CR  CO  OR Class
1 1.0 1.0 0.5 0.5 0.5 1.0    NB
2 0.0 0.0 0.5 0.5 0.5 0.0    NB
3 0.5 0.5 0.5 0.5 0.5 0.5    NB
4 1.0 1.0 1.0 1.0 1.0 1.0    NB
5 0.0 0.0 1.0 1.0 1.0 0.0    NB
6 0.5 0.5 1.0 1.0 1.0 0.5    NB


Call:
svm(formula = Class ~ ., data = inputdata, probability = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.1666667

Number of Support Vectors:  35
```

SVM Model Build!!!

After the model is developed, users can enter their data in the text input boxes for predicting bankruptcy. Each of the six input parameters have values as 1, 0.5 or 0 (positive, average and negative) respectively. Based on SVM model built for the initial dataset, the predictive system estimates the probability of bankruptcy as either B (Bankruptcy) or NB (Non Bankruptcy). The predictive tool was tested for both non-bankruptcy and bankruptcy conditions. The results from prediction are shown in Figure 8 and 9 respectively.

Figure 8: The predicted result as NB (Non-Bankruptcy) for user input data based on RBF-kernel.



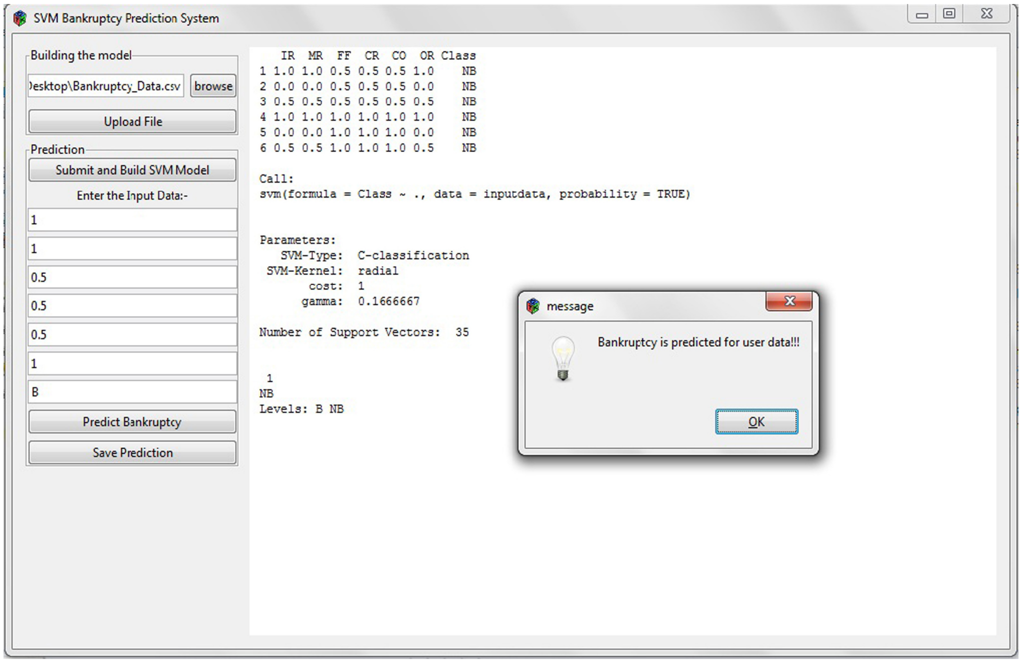
)esktop\Bankruptcy_Data.csv

1
1
0.5
0.5
0.5
1
B

```
   IR  MR  FF  CR  CO  OR Class
1 1.0 1.0 0.5 0.5 0.5 1.0    NB
2 0.0 0.0 0.5 0.5 0.5 0.0    NB
3 0.5 0.5 0.5 0.5 0.5 0.5    NB
4 1.0 1.0 1.0 1.0 1.0 1.0    NB
5 0.0 0.0 1.0 1.0 1.0 0.0    NB
6 0.5 0.5 1.0 1.0 1.0 0.5    NB

Call:
svm(formula = Class ~ ., data = inputdata, probability = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.1666667

Number of Support Vectors:  35


 1
NB
Levels: B NB
```

Bankruptcy is predicted for user data!!!

Figure 9: The predicted result as B (Bankruptcy) for user input data based on RBF-kernel.



)esktop\Bankruptcy_Data.csv

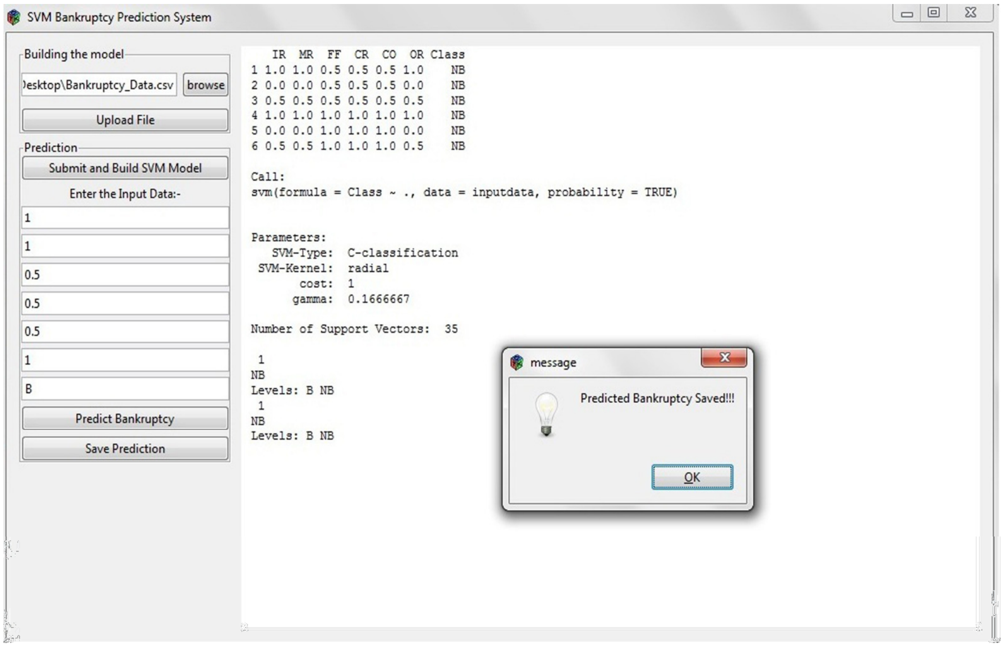
1
1
0.5
0.5
0.5
1
B

```
   IR  MR  FF  CR  CO  OR Class
1 1.0 1.0 0.5 0.5 0.5 1.0    NB
2 0.0 0.0 0.5 0.5 0.5 0.0    NB
3 0.5 0.5 0.5 0.5 0.5 0.5    NB
4 1.0 1.0 1.0 1.0 1.0 1.0    NB
5 0.0 0.0 1.0 1.0 1.0 0.0    NB
6 0.5 0.5 1.0 1.0 1.0 0.5    NB

Call:
svm(formula = Class ~ ., data = inputdata, probability = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.1666667

Number of Support Vectors:  35


 1
NB
Levels: B NB
```

Bankruptcy is predicted for user data!!!

The performance of the tool was computed by comparing the predicted value with the actual bankruptcy value for user data. It was found that the predictions were in par with the actual outcomes for both the cases (NB and B). The predicted outcome is saved as a .csv file in the local directory of user's system to view the results represented in Figure 10.

Figure 10: The predicted results for bankruptcy from the tool are saved



```
    IR  MR  FF  CR  CO  OR Class
1 1.0 1.0 0.5 0.5 0.5 1.0    NB
2 0.0 0.0 0.5 0.5 0.5 0.0    NB
3 0.5 0.5 0.5 0.5 0.5 0.5    NB
4 1.0 1.0 1.0 1.0 1.0 1.0    NB
5 0.0 0.0 1.0 1.0 1.0 0.0    NB
6 0.5 0.5 1.0 1.0 1.0 0.5    NB

Call:
svm(formula = Class ~ ., data = inputdata, probability = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.1666667

Number of Support Vectors:  35

 1
NB
Levels: B NB
 1
NB
Levels: B NB
```

Predicted Bankruptcy Saved!!!

5.5. Developing the prediction system as a package

The predictive system for detecting bankruptcy was encoded as a package in RStudio. The package was developed using interdependent libraries 'devtools' and 'roxygen2'. The package can be downloaded by users in their local machines followed by installing and running the package in RStudio. Once the package is installed users can run the predictive system for detection of bankruptcy using their input data.

6. CONCLUSIONS

The results suggest that machine learning techniques can be implemented for prediction of bankruptcy. To serve the financial organizations for identifying risk oriented customers a prediction system was implemented. The predictive system helps to predict bankruptcy for a customer dataset based on the SVM model.

REFERENCES

[1] Personal Bankruptcy or Insolvency laws in India. (http://blog.ipleaders.in/personal-bankruptcy-or-insolvency-laws-in-india/). Retrieved on Nov, 2014.

[2] Hossein Rezaie Doolatabadi, Seyed Mohsen Hoseini, Rasoul Tahmasebi. "Using Decision Tree Model and Logistic Regression to Predict Companies Financial Bankruptcy in Tehran Stock Exchanges." International Journal of Emerging Research in Management &Technology, 2(9): pp. 7-16.

[3] M. Kantardzic (2003). "Data mining: Concepts, models, methods, and algorithms." John Wiley & Sons.

[4] X. Wu et. al (2008). "Top 10 algorithms in data mining". Knowl Inf Syst. 14: pp 1-37.

[4] Hosmer David W, Lemeshow, Stanley (2000). "Applied Logistic Regression". Wiley.

[5] Rish, Irina (2001). "An empirical study of the naive Bayes classifier". IJCAI Workshop on Empirical Methods in AI.

[6] Breiman, Leo (2001). "Random Forests." Machine Learning 45 (1):pp. 5–32.

[7] McCulloch, Warren; Walter Pitts (1943). "A Logical Calculus of Ideas Immanent in Nervous Activity". Bulletin of Mathematical Biophysics 5 (4): pp. 115–133.

[8] Cortes, C.; Vapnik, V. (1995). "Support-vector networks". Machine Learning 20 (3): 273.

[9] Altman E (1968). Financial ratios, discriminant analysis and prediction of corporate bankruptcy. The Journal of Finance. 23(4), 589-609.

[10] Martin, D (1977). Early warning of bank failures: A logit regression approach. Journal of Banking and Finance. 1, 249-276.

[11] Lo, A (1984). Essays in financial and quantitative economics. Ph.D. dissertation, Harvard University.

[12] J. A. Ohlson (1980). “Financial ratios and the probabilistic prediction of bankruptcy,” Journal of Accounting Research, 18(1), pp. 109-131.

[13] B. Back et. al (1996). “Choosing Bankruptcy Predictors using Discriminant Analysis, Logit Analysis and Genetic Algorithms,” Turku School of Economics and Business Administration.

[14] J. E. Boritz, D. B. Kennedy (1995) “Effectiveness of neural network types for prediction of business failure,” Expert Systems with Applications. 9(4): pp. 95-112.

[15] P. Coats, L. Fant (1992). “A neural network approach to forecasting financial distress,” Journal of Business Forecasting, 10(4), pp. 9-12.

[16] M. D.Odom, R. Sharda (1990), “A neural network model for bankruptcy prediction,” IJC NN International Joint Conference on Neural Networks, 2, p. 163-168.

[17] J. Sun, H. Li (2008). “Data mining method for listed companies’ financial distress prediction,” Knowledge-Based Systems. 21(1): pp. 1-5. 2008.

[18] İ. H. Ekşi (2011). “Classification of firm failure with classification and regression trees,” International Research Journal of Finance and Economics. 76, pp. 113-120.

[19] V. Fan, v. Palaniswami (2000).“Selecting bankruptcy bredictors using a support vector machine approach,” Proceedings of the Internal Joint Conference on Neural Networks. pp. 354-359.

[20] S. Falahpour, R. Raie (2005). “Application of support vector machine to predict financial distress using financial ratios”. Journal of Accounting and Auditing Studies. 53, pp. 17-34.

[21] Hyunchul Ahn, Kyoung-jae Kim (2009). “Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach”. Applied Soft Computing. 9(2): pp. 599-607.

[22] Ming-Yuan Leon Li, Peter Miu (2010). “A hybrid bankruptcy prediction model with dynamic loadings on accounting-ratio-based and market-based information: A binary quantile regression approach”. Journal of Empirical Finance. 17(4): pp. 818-833.

[23] A. Martin, T Miranda Lakshmi, Prasanna Venkateshan (2014). “An Analysis on Qualitative Bankruptcy Prediction rules using Ant-miner”. IJ Intelligent System and Applications. 01: pp. 36-44.

[24] Stehman, Stephen V. (1997). "Selecting and interpreting measures of thematic classification accuracy". Remote Sensing of Environment. 62 (1): pp. 77–89.

[25] Fawcett, Tom (2006). "An Introduction to ROC Analysis". Pattern Recognition Letters. 27 (8): pp. 861 – 874.

[26] Vert, Jean-Philippe, Koji Tsuda, and Bernhard Schölkopf (2004). "A primer on kernel methods". Kernal Methods in Computational Biology.

Audit Sistem Informasi Menggunakan Framework COBIT 4.1 (Dengan Domain *Monitor and Evaluate*) Pada PT. Samudera Indonesia Tbk

Gerald Vidisa Jourdano [1], Alexander Setiawan [2], Agustinus Noertjahyana [3]
Program Studi Teknik Informatika, Fakultas Teknologi Industri, Universitas Kristen Petra
Jl. Siwalankerto 121 – 131 Surabaya 60236
Telp. (031) – 2983455, Fax. (031) – 8417658
E-mail: vjgerald9@gmail.com [1], alexander@petra.ac.id [2], agust@petra.ac.id [3]

ABSTRAK

PT. Samudera Indonesia Tbk merupakan perusahaan yang bergerak di bidang jasa logistik meliputi seluruh wilayah Indonesia dan mengjangkau ranah internasional. Perusahaan ini membagi portfolio bisnisnya menjadi 4 lini bagian, yaitu Samudera Shipping (bisnis pelayaran), Samudera Agencies (bisnis keagenan), Samudera Logistics (bisnis logistik), dan Samudera Terminal (bisnis pelabuhan) guna menghadirkan layanan jasa transportasi dan logistik terpadu, salah satu kantor cabang di Surabaya melayani di bidang keagenan.

Bagi sebuah perusahaan yang besar dengan proses bisnis dengan kompleksitas tinggi dan dibantu dengan teknologi informasi, perusahaan harus mampu memberikan pelayanan yang sesuai dengan tujuan bisnis yang ingin dicapai. Investasi terhadap teknologi informasi yang sudah diterapkan, belum memberikan jaminan yang pasti bagi perusahaan. Ada faktor lain yang harus diperhatikan, tidak hanya berfokus pada penggunaan aset teknologi informasi, melainkan perusahaan harus berfokus pada pemantauan, pemeliharaan, pengelolaan dan adanya jaminan bahwa perusahaan mematuhi peraturan terkait teknologi informasi yang berlaku.

Oleh sebab itu, peneliti menggunakan domain Monitor and Evaluate, untuk mengukur tingkat kedewasaan dari teknologi informasi perusahaan berdasarkan proses pemeliharaan, pengelolaan dan sejauh mana PT. Samudera Indonesia Tbk mematuhi peraturan hukum dalam teknologi informasi yang berlaku dan persyaratan eksternal lainnya.

Kata Kunci: Audit Sistem Informasi, COBIT 4.1, Monitor and Evaluate

ABSTRACT

PT. Samudera Indonesia Tbk is a company engaged in the field of logistics services throughout the territory of Indonesia and reach the international sphere. The company divides its business portfolio into 4 lines, namely Samudera Shipping (shipping business), Samudera Agencies (agency business), Samudera Logistics (logistics business), and Samudera Terminal (port business) to deliver integrated transportation and logistics services, and one of branch office located in Surabaya which serving in the field of agency.

For a large company with high-complexity business processes and assisted with information technology, the company must be able to provide services that match the business objectives to be achieved. Investment on information technology that has been applied, has not provided a firm guarantee for the company. There are other factors to consider, not just focusing on the use of information technology assets, but companies should focus on monitoring, maintaining, managing and ensuring that companies comply with relevant information technology regulations.

Therefore, the researchers using domain of Monitor and Evaluate, to measure the maturity level of enterprise information technology based on monitoring, managing process and how far PT. Samudera Indonesia Tbk comply with applicable information technology laws and external requirements.

Keywords: *Audit Information System, COBIT 4.1, Monitor and Evaluate*

1. PENDAHULUAN

Dewasa ini, perkembangan ilmu pengetahuan dan teknologi yang terus meningkat, membuat kedua hal tersebut saling berpadu menjadikan sebuah teknologi informasi yang membantu pekerjaan manusia. Dalam bidang teknologi informasi, khususnya dalam dunia bisnis, hal ini sangat diandalkan untuk menjalankan proses bisnisnya. Salah satunya, perusahaan memanfaatkan teknologi informasi (TI) untuk menjadikan sistem informasi yang terintegrasi dan dapat diakses secara langsung oleh karyawan, sehingga dengan adanya sistem informasi tersebut pengerjaan operasional bisa terbantu.

Selain penerapan TI, perusahaan memerlukan proses pengontrolan terhadap sistem informasi yang diterapkan, yaitu dengan memanfaatkan audit sistem informasi sesuai standar *Control Objective for Information and related Technology* (COBIT), yang dikeluarkan oleh organisasi bernama *Information System Audit and Control Association* (ISACA) pada tahun 1992. COBIT juga memiliki model kematangan (*Maturity Model*) yang digunakan untuk mengetahui posisi kematangannya saat ini dan secara terus menerus serta berkesinambungan harus berusaha untuk meningkatkan levelnya sampai tingkat tertinggi agar aspek pengelolaan (*governance*) terhadap TI dapat berjalan secara efektif.

PT. Samudera Indonesia Tbk., yang memiliki kantor cabang di Kota Surabaya merupakan perusahaan yang bergerak di bidang jasa yaitu, *shipping agency*.

Selama ini, perusahaan telah mengimplementasikan sistem *Open Ticket Request System* (OTRS) untuk mengetahui permasalahan TI yang terjadi. Cara kerja sistem tersebut, yaitu staf TI maupun staf biasa yang menemui kendala yang berkaitan langsung dengan sistem informasi atau infrastruktur jaringan (semua bagian dari critical assets) yang dikelola oleh perusahaan, dapat melaporkan masalah tersebut melalui sistem OTRS. Sistem tersebut memiliki daftar kendala yang biasa dialami oleh staf, beserta dengan besaran biaya yang dicantumkan sesuai masalah yang ditangani oleh staf TI. Namun, sistem tersebut belum memiliki standar yang jelas atau ukuran yang formal mengenai pelaporan kinerja TI, salah satunya, kinerja staf TI dinilai ketika ada setiap laporan permasalahan dari staf terkait penggunaan TI dalam perusahaan.

Oleh sebab itu, PT. Samudera Indonesia Tbk. memerlukan adanya pengontrolan atau audit sistem informasi, untuk mengawasi dan mengevaluasi seluruh aset TI yang digunakan oleh staf perusahaan, sehingga penulis menggunakan metode audit sistem informasi dengan framework COBIT 4.1 yang berfokus pada domain Monitor and Evaluate. Domain ini menitikberatkan pada proses pengawasan dan evaluasi yang ditujukan untuk solusi TI pada perusahaan dan seluruh proses bisnis yang diterapkan, sehingga ada standar penilaian secara umum bagi perusahaan.

2. TINJAUAN PUSTAKA

2.1 COBIT 4.1

COBIT (*Control Objective for Information and Related Technology*) merupakan kerangka dari best of practices manajemen teknologi informasi (TI) yang membantu organisasi untuk memaksimalkan keuntungan bisnis, serta dapat membantu auditor, user dan manajemen mengelola resiko bisnis dan masalah-masalah teknis dalam organisasi. *Framework* COBIT disusun oleh *Information System Audit and Control Association* (ISACA) dan IT *Governance Institute* (ITGI) [3]. COBIT dapat digunakan untuk mengukur level kedewasaan (*maturity level*) dalam proses TI dan mengukur kesesuaian antara kebutuhan bisnis dan tujuan TI dalam organisasi [9].

Domain COBIT yang digunakan oleh peneliti, yaitu:

1. ME 1 (*Monitor and Evaluate IT Performance*)

Domain ME 1 menjelaskan mengenai kebutuhan proses pemantauan terhadap kinerja manajemen TI dalam perusahaan yang efektif. Proses ini mencakup indikator kinerja, mendefinisikan hal-hal yang relevan, pelaporan yang sistematis dan tepat waktu, dan cepat bertindak atas penyimpangan. Pemantauan diperlukan untuk memastikan bahwa hal yang benar dilakukan dan sejalan dengan arah dan kebijakan yang telah ditetapkan.

2. ME 2 (*Monitor and Evaluate Internal Controls*)

Domain ME 2 menentukan program pengendalian internal dan proses monitoring untuk TI perusahaan. Proses ini meliputi pemantauan dan pelaporan kontrol, ulsan dari hasil penilaian diri dan pihak ketiga. Manfaat utama dari pemantauan pengendalian internal adalah untuk memberikan keyakinan yang berkaitan dengan operasi yang efektif dan efisien serta kepatuhan terhadap hukum dan peraturan yang berlaku.

3. ME 3 (*Ensure Compliance with External Requirements*)

Domain ME 3 menjelaskan mengenai kepatuhan perusahaan terhadap undang-undang dan peraturan persyaratan kontrak. Proses ini meliputi identifikasi persyaratan kepatuhan, mengoptimalkan dan mengevaluasi respon, memperoleh jaminan bahwa persyaratan telah dipenuhi dan pada akhirnya, mengintegrasikan pelaporan kepatuhan TI dengan bisnis

4. ME 4 (*Provide IT Governance*)

Domain ME 4 mempunyai tujuan memberi kepastian pada perusahaan apakah investasi kebutuhan TI sesuai dengan strategi bisnis yang sudah diterapkan. Selain itu, domain ini menjelaskan mengenai pembentukan kerangka kerja pengelolaan teknologi informasi yang efektif untuk mencapai strategi tersebut.

2.2 *Control Practices*

COBIT memiliki standar *control practices* yang digunakan sebagai pengukur bagi sebuah organisasi. Dari domain *monitor and evaluate* yang digunakan oleh peneliti, memiliki *control practices*, antara lain:

1. ME 1 (*Monitor and Evaluate IT Performance*)
   a. ME 1.1 *Monitoring Approach*
   b. ME 1.2 *Definition and Collection of Monitoring Data*
   c. ME 1.3 *Monitoring Method*
   d. ME 1.4 *Performance Assessment*
   e. ME 1.5 *Board and Executive Reporting*
   f. ME 1.6 *Remedial Actions*
2. ME 2 (*Monitor and Evaluate Internal Controls*)
   a. ME 2.1 *Monitoring of Internal Control Framework*
   b. ME 2.2 *Supervisory Review*
   c. ME 2.3 *Control Exception*
   d. ME 2.4 *Control Self-assesment*
   e. ME 2.5 *Assurance of Internal Control*
   f. ME 2.6 *Internal Control at Third Parties*
   g. ME 2.7 *Remedial Actions*
3. ME 3 (*Ensure Compliance with External Requirements*)
   a. ME 3.1 *Identification of External Legal, Regulatory and Contractual Compliance Requirements*
   b. ME 3.2 *Optimisation of Response to External Requirements*
   c. ME 3.3 *Evaluation of Compliance With External Requirements*
   d. ME 3.4 *Positive Assurance of Compliance*
   e. ME 3.5 *Integrated Reporting*
4. ME 4 (*Provide IT Governance*)
   a. ME 4.1 *Establishment of an IT Governance Framework*
   b. ME 4.2 *Strategic Alignment*
   c. ME 4.3 *Value Delivery*
   d. ME 4.4 *Resource Management*
   e. ME 4.5 *Risk Management*
   f. ME 4.6 *Performance Measurement*
   g. ME 4.7 *Independent Assurance*

2.3 *Maturity Model*

Tingkat kedewasaan memiliki peranan sebagai pengukur seberapa matang proses TI yang sudah diterapkan oleh perusahaan. Penerapan yang tepat pada tata kelola TI di lingkungan perusahaan, tergantung pada pencapaian tiga aspek kedewasaan (*maturity*), yaitu kemampuan, jangkauan dan kontrol. Dampak dari peningkatan maturity akan mengurangi risiko dan meningkatkan efisiensi, mendorong berkurangnya kesalahan dan meningkatkan kuantitas proses yang dapat diperkirakan kualitasnya, serta mendorong efisiensi biaya terkait dengan penggunaan sumber daya TI [5]. Tingkat kemampuan pengelolaan TI berdasarkan kerangka kerja COBIT 4.1., memiliki level kedewasaan dengan skala dari level 0 sampai level 5, antara lain:

Level 0 *Non-existent* perusahaan tidak mengetahui dan tidak memahami proses teknologi informasi yang harus dilakukan.

Level 1 *Initial*: pada level ini, secara keseluruhan manajemen TI belum diatur dengan baik. Terdapat bukti bahwa perusahaan telah mengetahui proses-proses pengendalian sistem. Walaupun tidak ada proses yang sesuai standar, ada pendekatan secara ad hoc dan hanya diterapkan pada case tertentu saja.

Level 2 *Repeatable*: proses yang telah dilakukan sampai tahap, yaitu untuk prosedur yang sama dilakukan oleh orang yang berbeda didalam melakukan tindakan yang sama. Tidak terdapat pelatihan resmi atau koordinasi mengenai prosedur standar dan tanggung jawab standar yang diberikan kepada setiap personel.

Level 3 *Defined*: proses yang dilakukan telah memiliki standar dan perusahaan mendokumentasi yang telah dikomunikasikan melalui pelatihan, sehingga memaksa setiap personel atau staf untuk mengikuti prosedur dan meminimalkan kejadian penyimpangan. Prosedur yang ada, masih dinilai tidak memuaskan.

Level 4 *Managed*: proses yang dilakukan berada dalam peningkatan yang konsisten dan mengarah pada tujuan, serta memungkinkan untuk melakukan pengawasan dan mengukur tingkat kesesuaian dengan prosedur. Apabila proses yang dijalankan tidak berjalan efektif, akan diambil tindakan.

Level 5 *Optimised*: secara keseluruhan dari proses yang ada, telah mencapai tingkat best practices, dan didasarkan pada hasil pengembangan secara kontinu, serta membandingkan pemodelan maturity dengan organisasi atau perusahaan lain. Proses TI yang diterapkan dapat digunakan terintegrasi untuk membuat sistem alur kerja yang lebih otomatis. Penyediaan atau pengadaan perangkat untuk meningkatkan efektivitas dan kualitas yang membuat perusahaan lebih cepat beradaptasi.

3. METODE PENELITIAN

3.1 PT. Samudera Indonesia Tbk

PT. Samudera Indonesia merupakan perusahaan yang bergerak dibidang jasa atau pelayanan transportasi kargo dan logistik terpadu bagi pelanggan domestik dan internasional. Awalnya, perusahaan didirikan oleh Soedarpo Sastrosatomo, yang mulai merintis usaha di bidang keagenan pelayaran pada tahun 1953. Perusahaan kemudian berkembang dan resmi berstatus sebagai sebuah perusahaan pelayaran dengan nama PT. Samudera Indonesia pada tahun 1964, dan mampu bertahan hingga kini, serta berpusat di Ibu Kota Jakarta.

Perusahaan membagi portofolio bisnisnya ke dalam 4 lini bisnis. Keempat lini bisnis tersebut antara lain, Samudera Shipping (bisnis pelayaran), Samudera Agencies (bisnis keagenan), Samudera Logistics (bisnis logistik), dan Samudera Terminal (bisnis pelabuhan) guna menghadirkan layanan jasa transportasi dan logistik terpadu [1]. Salah satu anak perusahaannya yang bergerak di bidang keagenan terdapat di Surabaya, berlokasi di Perak Barat No. 400, Surabaya, Jawa Timur.

Proses bisnis yang dijalankan pada kantor di Surabaya, yaitu customer yang akan mengirim suatu barang dalam jumlah besar, akan dilayani oleh staf *Sales*. Setelah pendataan barang dan informasi dari *customer* didapatkan, maka staf *sales* akan mengkonfirmasikan semua data yang didapat kepada staf *customer services* dan staf *financial*. Staf *customer services* akan mendata ulang secara lengkap jenis pengiriman barang, seberapa banyak barang yang akan dikirim, berapa lama perkiraan barang akan sampai ke tempat tujuan dan memberikan informasi bagaimana customer dapat melakukan *tracking* terhadap barangnya yang dikirim. Sedangkan untuk staf *financial*, akan mengecek apakah customer tersebut memiliki riwayat pengiriman sebelumnya atau tidak. Jika belum ada, akan dibuatkan akun baru sesuai dengan identitas *customer*. Staf *financial* juga akan menetapkan harga pengiriman barang berdasarkan jarak lokasi pengiriman, total berat barang yang akan dikirim, dan setelah data total pembayaran terhitung nominalnya, akan diserahkan kepada staf *sales* untuk disampaikan kepada *customer*.

3.2 Visi, Misi, Target, dan Strategi PT. Samudera

PT. Samudera Indonesia memiliki visi, yaitu Global Connectivity to meet people's need, yang berarti perusahaan memiliki tujuan menghubungkan secara global untuk memenuhi kebutuhan konsumen [1].

Selain visi yang telah disebutkan diatas, PT. Samudera Indonesia memiliki misi, yaitu:

- Menyediakan layanan jasa transportasi untuk memenuhi kebutuhan distribusi barang dari dan ke seluruh Indonesia maupun Internasional.
- Senantiasa memastikan pertumbuhan bisnis yang berkelanjutan seraya memberikan nilai tambah bagi pemegang saham.
- Berkontribusi positif terhadap pertumbuhan ekonomi Indonesia dengan memberikan solusi logistik yang efisien.
- Turut berperan serta dalam menciptakan lapangan kerja dan membangun kompetensi sumber daya manusia di Indonesia.

Sesuai dengan visi dan misi yang dimiliki oleh perusahaan, maka target perusahaan untuk ke depannya, yaitu berusaha menyediakan layanan berkualitas tinggi dalam transportasi barang dan logistik untuk konsumen, serta untuk mencapai target tersebut, perusahaan memiliki strategi, antara lain:

- Memberikan pelayanan yang terbaik untuk pelanggan di setiap lini bisnis perusahaan.
- Meningkatkan investasi di sarana pendukung kegiatan logistik seperti pusat layanan peti kemas, gudang, truk dan heavy-lift equipment.

3.3 Metode Audit Sistem Informasi

Dalam melaksanakan audit sistem informasi, penulis menerapkan metodologi yang diterapkan sesuai dengan metodologi yang dianjurkan oleh *IT Assurance Guide: Using COBIT*. Dasar untuk melaksanakan metodologi pengumpulan data dalam audit sistem informasi, meliputi observasi dan wawancara dengan pihak perusahaan. Berikut penjabaran metode audit yang dilakukan oleh penulis antara lain:

1. Menentukan Audit Resource

Tahap ini bertujuan untuk mengumpulkan sumber data atau dokumen yang diperlukan untuk proses audit sistem informasi, menghubungi narasumber yang berkaitan dengan pengelolaan atau monitor aset-aset TI dalam perusahaan, untuk meminta kesediaan mengisi kuisioner, melakukan wawancara mengenai kondisi TI perusahaan, dan menyamakan pendapat berdasarkan data yang sudah diperoleh.

2. Evaluasi Kontrol

Tahap selanjutnya ditujukan untuk mengetahui apakah seluruh kontrol (seluruh peraturan, standar prosedur, dan strukur organisasi) yang sudah diterapkan dapat memenuhi standar berdasarkan COBIT 4.1. Apabila kontrol yang sudah ada, dapat dipenuhi secara efektif sesuai dengan standar COBIT tersebut, maka kontrol dapat digunakan sebagai standar untuk pengukuran tahap berikutnya, yaitu evaluasi kesesuaian antara proses terhadap kontrol. Jika kontrol tersebut tidak memenuhi standar pengukuran, maka proses berikutnya melalui evaluasi substansi signifikan.

3. Evaluasi Kesesuaian Proses terhadap Kontrol

Tahap ini dilakukan apabila kontrol dinyatakan secara efektif mencapai standar ideal yang sesuai dengan COBIT 4.1. Untuk proses selanjutnya, membandingkan realita yang terjadi di perusahaan dengan kontrol (peraturan dan standar prosedur), untuk memeriksa kesesuaian proses sesungguhnya yang telah diterapkan dengan konsisten.

4. Evaluasi Substansi Terbatas

Tahap ini dilakukan apabila tahap ketiga memiliki kesimpulan yang tidak absolut. Pada dasarnya, tahap ketiga biasanya sudah dapat diketahui apakah sebuah proses dapat mencapai target yang sesuai terhadap kontrol. Untuk mengatasi hal tersebut, maka proses yang tidak dapat di evaluasi secara absolut memerlukan uji substansi dengan memanfaatkan dokumen proses, kuesioner, dan wawancara terhadap pelaku proses agar dapat mengambil keputusan. Tahap ini tidak dilakukan dalam seluruh aspek audit, hanya untuk proses-proses yang memerlukan pengujian lebih lanjut.

5. Evaluasi Substansi Signifikan

Tahapan ini dilakukan apabila pada tahap kedua, hasil yang diperoleh membuktikan bahwa kontrol tidak dapat mencapai gambaran ideal secara efektif sesuai COBIT. Aspek yang akan diuji merupakan aspek keseluruhan mengenai pengelolaan dan pemeliharaan aplikasi, infrastruktur atau aset TI tanpa melihat apakah kontrol untuk aspek tersebut mencapai nilai yang efektif atau tidak.

6. Pengukuran Level Kedewasaan

Tahap ini memberikan informasi mengenai level perbandingan antara kondisi ideal yang selalu dinyatakan dengan level 5 dengan kondisi yang ada di lingkungan perusahaan.

7. Menentukan Kesimpulan dan Rekomendasi

Setelah melalui tahap kedua hingga tahap keenam, rekomendasi yang dibuat didasarkan pada hasil evaluasi sesuai *control objectives* dan memanfaatkan penilaian dari auditor atau staf ahli professional di bidang ini.

3.4 Langkah Audit Sistem Informasi

Langkah-langkah yang dilakukan dalam proses audit sistem informasi adalah:

1. Membuat kuisioner untuk mengetahui tujuan teknologi informasi (TI) dan survei ke perusahaan untuk menentukan domain audit yang sesuai dengan permasalahan.
2. Melakukan pengumpulan data dengan metode judgemental sampling, wawancara, observasi, dan kuisioner.
3. Memahami strategi dan proses bisnis yang dimiliki oleh PT. Samudera Indonesia Tbk.
4. Melakukan mapping antara tujuan TI dan proses TI.
5. Penentuan aspek yang perlu diukur tingkat kedewasaan (maturity level) sesuai dengan domain audit (dalam skripsi ini domain yang digunakan adalah monitor and evaluate).
6. Mengukur tingkat kedewasaan (maturity level) sesuai dengan domain audit.
7. Validasi data yang diperoleh secara internal dan eksternal dengan narasumber.
8. Mengambil kesimpulan dan rekomendasi berdasarkan hasil analisa audit sistem informasi, serta saran maupun harapan yang diberikan oleh staf IT perusahaan.

3.5 Kondisi Teknologi Informasi di PT. Samudera Indonesia Tbk.

Jumlah komputer yang dimiliki oleh PT. Samudera Indonesia Tbk., mencapai 50 komputer yang berbasis Windows 7. Semua komputer sudah terkoneksi dengan seluruh aplikasi yang dimiliki oleh perusahaan. Berikut spesifikasi komputer yang dimiliki oleh perusahaan:

- *Operating System*: Microsoft Windows 7 Professional 32-bit
- *Processor*: Intel® Core (TM) i3 CPU
- RAM : 4 Gigabyte (Gb)
- VGA : Intel® High Definition (HD) Graphics
- HDD : 500 Gigabyte (Gb)
- *Monitor*: HP
- *Mouse*: HP
- *Keyboard*: HP

4. HASIL

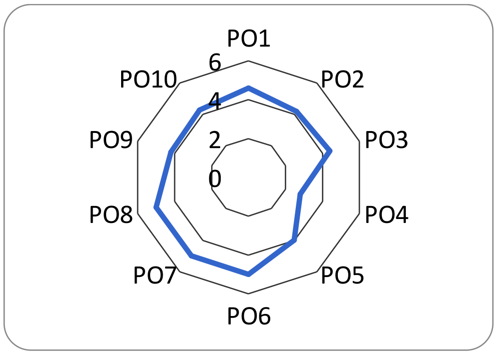
4.1 Audit Awal Teknologi Informasi

Proses audit awal ini dilakukan dengan melakukan pengisian kuisioner keseluruhan domain yang terdapat di COBIT 4.1, hal ini dilakukan untuk mengetahui sejauh mana kondisi TI yang diketahui oleh staf. Hasil dari kuisioner audit awal (yang berfokus pada domain *Monitor and Evaluate*) digunakan sebagai nilai pembanding dengan hasil analisa kontrol dan analisa evaluasi oleh penulis.

Setelah melakukan penyebaran kuisioner kepada staf IT di perusahaan, berikut hasil olah data kuisioner dari PT. Samudera Indonesia Tbk., mengenai keseluruhan domain yang terdapat pada COBIT:

Tabel 1. ***Planning and Organization (PO)***

Domain	Nilai rata-rata	Nilai Maks.
PO1	4.6	5
PO2	4.2	5
PO3	4.4	5
PO4	2.8	5
PO5	4	5
PO6	5	5
PO7	5	5
PO8	5	5
PO9	4.2	5
PO10	4.28571429	5

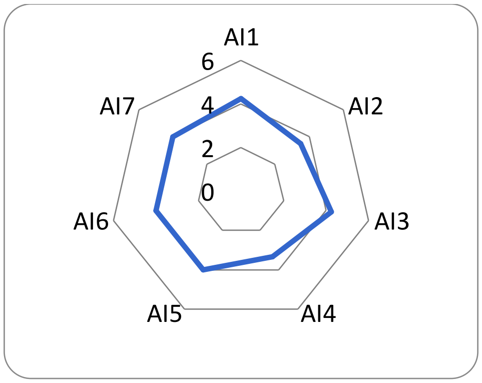


Gambar 1. Diagram Hasil Kuisioner PO

Perhitungan nilai rata-rata pada Tabel 1 berasal dari nilai yang didapat dari hasil kuisioner yang diisi oleh pihak perusahaan. Pada kuisioner tersebut, setiap domain mempunyai poin kontrol yang menjelaskan mengenai domain tersebut dengan nilai maksimal 5 dan dapat dilihat pada grafik Gambar 1.

Tabel 2. ***Acquire and Implementation*** **(AI)**

Domain	Nilai rata-rata	Nilai Maks.
AI1	4.25	5
AI2	3.5	5
AI3	4.25	5
AI4	3.33333333	5
AI5	4	5
AI6	4	5
AI7	4	5

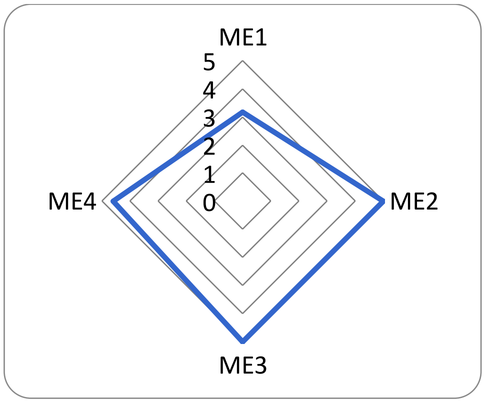


Gambar 2. Diagram Hasil Kuisioner AI

Perhitungan nilai rata-rata pada Tabel 2 berasal dari nilai yang didapat dari hasil kuisioner yang diisi oleh pihak perusahaan. Pada kuisioner tersebut, setiap domain mempunyai poin kontrol yang menjelaskan mengenai domain tersebut dengan nilai maksimal 5 dan dapat dilihat pada grafik Gambar 2.

Tabel 3. ***Monitor and Evaluate*** **(ME)**

Domain	Nilai rata-rata	Nilai Maks.
ME1	3.16666667	5
ME2	5	5
ME3	5	5
ME4	4.6	5

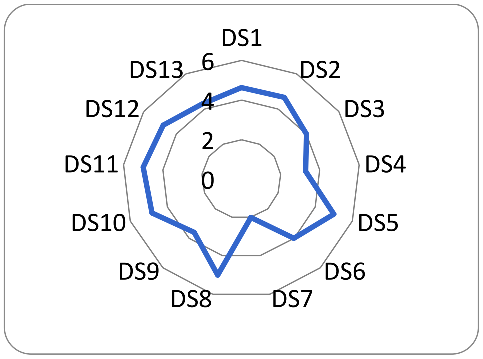


Gambar 3. Diagram Hasil Kuisioner ME

Perhitungan nilai rata-rata pada Tabel 3 berasal dari nilai yang didapat dari hasil kuisioner yang diisi oleh pihak perusahaan. Pada kuisioner tersebut, setiap domain mempunyai poin kontrol yang menjelaskan mengenai domain tersebut dengan nilai maksimal 5 dan dapat dilihat pada grafik Gambar 3.

Tabel 4. *Deliver and Support* (DS)

Domain	Nilai rata-rata	Nilai Maks.
DS1	4.625	5
DS2	4.66666667	5
DS3	4	5
DS4	3.27272727	5
DS5	5	5
DS6	4	5
DS7	2	5
DS8	5	5
DS9	3.6	5
DS10	4.83333333	5
DS11	5	5
DS12	4.8	5
DS13	4.28571429	5



Gambar 4. Diagram Hasil Kuisioner DS

Perhitungan nilai rata-rata pada Tabel 4 berasal dari nilai yang didapat dari hasil kuisioner yang diisi oleh pihak perusahaan. Pada kuisioner tersebut, setiap domain mempunyai poin kontrol yang menjelaskan mengenai domain tersebut dengan nilai maksimal 5 dan dapat dilihat pada grafik Gambar 4.

4.2 Perbandingan Tingkat Kedewasaan PT. Samudera Indonesia dan Hasil Observasi Lapangan

Tabel 5. Perbandingan Tingkat Kedewasaan

Domain	Tingkat Kedewasaan PT. Samudera Indonesia	Tingkat Kedewasaan Hasil Observasi Lapangan
ME1	3.16666667	4,04
ME2	5	3,87
ME3	5	3,74
ME4	4.6	3,91

Dalam uraian tabel 5, perhitungan nilai tingkat kedewasaan PT. Samudera Indonesia merupakan total nilai yang didapat pada domain ME saat melakukan audit awal. Sedangkan untuk kolom selanjutnya, tingkat kedewasaan berdasarkan kegiatan observasi dan wawancara dari penulis dengan narasumber mengenai evaluasi kontrol, evaluasi proses, sehingga hasil akhir yang didapat merupakan analisa yang dilakukan oleh penulis beserta pemberian saran maupun rekomendasi.

5. KESIMPULAN DAN SARAN

5.1. Kesimpulan

Secara keseluruhan proses TI yang ada di PT. Samudera Indonesia sudah dilakukan kegiatan pemantauan dan evaluasi sesuai dengan prosedur yang ada, namun belum dilakukan dengan sangat baik dan belum mencapai control practices dari COBIT 4.1.

1. Hasil penilaian tingkat kedewasaan mengenai domain ME1, ME2, ME3, dan ME4, sudah tergolong sesuai prosedur yang dimiliki, dan dengan hasil perolehan nilai ME1 *Monitoring and Evaluate IT Performance* sebesar 4,04; ME2 *Monitoring and Evaluate Internal Control* sebesar 3,87; ME3 *Ensure Compliance with External Requirements* sebesar 3,74: ME4 *Provide IT Governance* sebesar 3,91.
2. Penilaian tingkat kedewasaan TI pada PT. Samudera Indonesia yang tergolong kriteria *defined*, disebabkan karena perusahaan sudah menerapkan proses TI sesuai dengan prosedur dan standar yang telah ditentukan, dan sudah disesuaikan dengan model bisnis, strategi dan tujuan bisnis dari perusahaan.

5.2. Saran

1. Kegiatan pemantauan sangat penting bagi perusahaan untuk mencegah risiko bisnis yang timbul akibat investasi TI yang tidak sesuai dengan tujuan bisnis dan perlu memperhatikan kerangka kerja, yang sebaiknya dibuat oleh divisi IT perusahaan sesuai dengan kebutuhan dan memberikan hasil evaluasi yang dapat menekan pengeluaran biaya investasi TI, sehingga bersifat realistis serta mematuhi peraturan, hukum dan regulasi yang berlaku.
2. Pertanyaan wawancara dibuat lebih detail dengan mengarahkan narasumber sesuai topik agar dapat memahami pertanyaan lebih baik.

6. DAFTAR PUSTAKA

[1] *Annual Report* PT. Samudera Indonesia. 2016. Laporan Tahunan.

[2] Fauzan., & Latifah, Rani. 2015. *Audit Tata Kelola Teknologi Informasi Untuk Mengontrol Manajemen Kualitas Menggunakan COBIT 4.1 (Studi Kasus: PT Nikkatsu Electric Works).* Jurnal Teknik Informatika dan Sistem Informasi Vol. 1. No. 3.

[3] Gondodiyoto, Sanyoto. 2007. *Audit Sistem Informasi + Pendekatan COBIT*. Jakarta: Mitra Wacana Media.

[4] Heryudo, Stenly. 2013. *Analisis Tata Kelola Teknologi Informasi Dengan Menggunakan Framework Cobit 4.1 Dalam Mendukung Layanan Teknologi Informasi Studi Kasus : PT. Pupuk Sriwidjaja Palembang*. Skripsi Fakultas Informatika. Institut Teknologi Telkom.

[5] IT Governance Institute. 2007. *COBIT 4.1 Framework Control Objectives, Management Guidelines, Maturity Models*. IT Governance Institute.

[6] Kesumawardhani, Dwi R. 2012. *Evaluasi IT Governance Berdasarkan COBIT 4.1 (Studi Kasus di PT TIMAH (PERSERO) Tbk)*. Skripsi Program Ekstensi Akuntansi. Universitas Indonesia.

[7] Madarina, Thalita. 2013. P*enyusunan Mekanisme pemantauan dan Evaluasi Sistem E-learning IT Telkom dengan COBIT 4.1*. Skripsi

[8] Maghfiroh, Inayatul. 2016. *Analisis dan Perancangan Tata Kelola TI Menggunakan Framework COBIT 4.1 Domain Deliver and Support (DS) dan Monitor and Evaluate : Studi Kasus PT. Bio Farma (PERSERO)*. Skripsi Program Studi Sistem Informasi. Universitas Telkom.

[9] Sarno, R. & Tanuwijaya, H. 2010. *Comparation of CobiT Maturity Model and Structural Equation Model for Measuring the Alignment between Universit Academic Regulations and Information Technology Goals*. International Journal of Computer Science and Network Security, vol.10 (no. 6), p. 80.

[10] Weber, Ron. 2000. *Information Control and Audit*. Printice Hall, Inc. New Jersey.

[11] Wella, Johan S. 2015. *Audit Sistem Informasi Menggunakan Cobit 4.1 pada PT. Erajaya Swasembada Tbk*. Skripsi Sistem Informasi. Universitas Multimedia Nusantara.

[12] Wella., & Tjhin, V. U. 2013. *Pengukuran Tingkat Kematangan Implementasi Teknologi Informasi Pada Domain Monitor and Evaluate Dengan Menggunakan COBIT 4.1 Pada PT ERAJAYA SWASEMBADA, TBK*. Jurnal Seminar Sistem Informasi Program Studi Manajemen, STIE Wiyatamandala.

[13] Wijaya, Alief F. 2014. *Audit Tata Kelola Teknologi Informasi Bagian Pengelolaan Data Menggunakan Framework COBIT 4.1 pada Bank JATENG*. Skripsi Sistem Informasi. Universitas Dian Nuswantoro.

DATA, TEXT, AND WEB MINING FOR BUSINESS INTELLIGENCE: A SURVEY

Abdul-Aziz Rashid Al-Azmi

Department of Computer Engineering, Kuwait University, Kuwait
Fortinbras222@hotmail.com

ABSTRACT

The Information and Communication Technologies revolution brought a digital world with huge amounts of data available. Enterprises use mining technologies to search vast amounts of data for vital insight and knowledge. Mining tools such as data mining, text mining, and web mining are used to find hidden knowledge in large databases or the Internet. Mining tools are automated software tools used to achieve business intelligence by finding hidden relations, and predicting future events from vast amounts of data. This uncovered knowledge helps in gaining completive advantages, better customers' relationships, and even fraud detection. In this survey, we'll describe how these techniques work, how they are implemented. Furthermore, we shall discuss how business intelligence is achieved using these mining tools. Then look into some case studies of success stories using mining tools. Finally, we shall demonstrate some of the main challenges to the mining technologies that limit their potential.

KEYWORDS

business intelligence, competitive advantage, data mining, information systems, knowledge discovery

1. INTRODUCTION

We live in a data driven world, the direct result of advents in information and communication technologies. Millions of resources for knowledge are made possible thanks to the Internet and Web 2.0 collaboration technologies. No longer do we live in isolation from vast amounts of data. The Information and Communication Technologies revolution provided us with convenience and ease of access to information, mobile communications and even possible contribution to this amount of information. Moreover, the need of information from these vast amounts of data is even more pressing for enterprises. Mining information from raw data is an extremely vital and tedious process in today's information driven world. Enterprises today rely on a set of automated tools for knowledge discovery to gain business insight and intelligence. Many branches of knowledge discovery tools were developed to help today's competitive business markets thrive in the age of information. World's electronic economy has also increased the pressure on enterprises to adapt to such new business environment. Main tools for getting information from these vast amounts are automated mining tools, specifically speaking data mining, text mining, and web mining.

Data Mining (DM) is defined as the process of analysing large databases, usually data warehouses or internet, to discover new information, hidden patterns and behaviours. It's an automated process of analysing huge amounts of data to discover hidden traits, patterns and to predict future trends and forecast possible opportunities. DM analyse datasets of rational

DOI : 10.5121/ijdkp.2013.3201

databases, in multiple dimensions and angles, producing a summary of the general trends found in the dataset, relationships and models that fits the dataset. DM is a relatively new interdisciplinary field involving computer science, statistical modelling, artificial intelligence, information science, and machine learning [1]. One of the main uses of DM is business intelligence and risk management [2]. Enterprises must make business critical decisions based on large datasets stored in their databases, DM directly affect decision-making. DM is relied on in retail, telecommunication, investment, insurance, education, and healthcare industries they are data-driven. Other uses of DM includes biological research such as DNA and the human genome project, geospatial and weather research for analysing raw data used to analyse geological phenomenon.

A related field is Text Mining (TM), which deals with textual data rather than records. TM is defined as automatic discovery of hidden patterns, traits, or unknown information from textual data [7]. Textual data makes up huge amounts of data found on World Wide Web WWW, aside from multimedia. TM is related field to DM, but differs in its techniques and methodologies used. TM is also an interdisciplinary field encompassing computational linguistics, statistics, and machine learning. TM uses complex Natural Language Processing (NLP) techniques. It involves a training period for the TM tool to comprehend patterns and hidden relations. The process of mining text documents involve linguistically and semantically analysis of the plain text, thus structuring the text. Finally relates and induces some hidden traits found in the text, like frequency of use for some words, entity extractions, and documents summarizations. TM is used, aside from business applications, for scientific research, specifically medical and biological [22]. TM is very useful in finding and matching proteins' names and acronyms, and finding hidden relations between millions of documents.

The other mining technique is Web Mining (WM). WM is defined as automatic crawling and extraction of relevant information from the artefacts, activities, and hidden patterns found in WWW. WM is used for tracking customers' online behaviour, most importantly cookies tracking and hyperlinks correlations. Unlike search engines, which send agents to crawl the web searching for keywords, WM agents are far more intelligent. WM work by sending intelligent agents to certain targets, like competitors sites' [8]. These agents collect information from the host web server and collect as much information from analysing the web page itself. Mainly they look for the hyperlinks, cookies, and the traffic patterns. Using this collected knowledge enterprises can establish better customer relationships, offers and target potential buyers with exclusive deals. The WWW is very dynamic, and web crawling is repetitive process where contentious iteration will achieve effective results. WM is used for business, stochastic, and for criminal and juridical purposes mainly in network forensics.

In this survey paper, we shall look at the main mining technologies used through information systems for business applications to gain new levels of business intelligence. Furthermore, we shall look at how these techniques can help in achieving both business leadership and risk management by illustrating real enterprises' own experience using mining techniques. In addition, we shall look at the main challenges facing data, web, and text mining today.

2. HISTORY AND BACKGROUND

Many developments were made leading to mining technologies we have today. These developments date back to early days of mathematical models and statically analysis using regression and Bayesian methods in mid-1700s. With the advent of commercial electronic computers after World War II, large data sets were stored into magnetic tapes to automate the work. In the 1960s were data stored in computers helped analysers to answer simple predictive questions. With the development of programming languages, specifically COmmon Business Oriented Language or COBOL, and Rational Database Management Systems RDBMS, querying databases were possible. Meaning more complex information and knowledge can be extracted. Development of advanced object oriented languages such as C++, Java, multi-dimensional databases, data warehousing, and Online Analytical Processing OLAP made way for an automated algorithmic way of extracting patterns, knowledge from such large data sets. DM tools today are more advanced and provide more than reporting capabilities, they can discover hidden patterns and knowledge. These DM tools were developed in the 1990s.

After the Internet and the WWW revolution in the early 1990s, many research and developments were made to automate the search and exploration of the net, especially text, found in the URLs. Developments in NLP, neural networks and text processing led ultimately to search engines development. The need for better search algorithms led to textual exploration of web pages. These developments greatly enhanced the search engines and opened the door for text mining to be applied in several other applications. Search engines' technologies were centred on agents that could map the vast WWW and correlate keywords and similar other possible keywords. These developments will lead to the more intelligent agents that search the WWW for not only keywords but also site visitors' patterns. Ultimately, the developments in both DM and TM lead to the notion of WM, were the WWW is used as a source for looking for new knowledge, hidden away somewhere. WM agents are small standalone software, that crawl the WWW, acquiring logging data, cookies, and site visits behaviour found on the servers and other machines attached to the WWW.

The tremendous advancements made in the mining technologies have shifted thought from data collection to knowledge discovery and collection [9]. With today's powerful and relatively inexpensive hardware and network infrastructure, matched with advanced software for mining, enterprises are adapting mining technologies as essential business processes. In addition, the Internet has an integral role as network and communications are ubiquitous today, mining is carried over the world through the network of databases. The vast amount of knowledge is not only consumed at the top senior management level but at all the other levels of an enterprise as well.

Today mining software utilizes complex algorithms for searching, pattern recognition, and forecasting complex stock market changes. IBM and Microsoft are on an epic race to produce best DM software to date; this is also influenced by security and intelligence agencies such as FBI and CIA. Multi-linguistic and semantic TM is a hot new research topic. As modern as it is today, WM has become an increasingly adopted business process as well. WM is suited more for e-commerce than DM and TM. The nature of e-commerce suggests the direct exploitation of customers' online behaviours. Many surveyors, such as Gartner Group, predict that over 5 billion dollars of business will be net worth of e-commerce in the coming years [10]. WM is heavily used for e-education and e-business, as the WWW is again their main platform. As developments were huge in the 1990's in terms of hardware support for mining techniques and the further leaps achieved by modern software, mining techniques are more of a must than a commonplace for modern business today. Relatively new and emerging mining techniques are what are known collectively as Reality Mining [65]. Reality mining is the collection of transactions made daily by individuals to realize how they live and react. Reality mining is aimed at developing our understanding of our modern societies, economies and politics. This is technology is made possible by the ICT world we live in today. Reality mining which is very controversial as it infiltrate individuals privacy, is catching the intention of governments and corporate, as it can be used for potential business benefits. Reality mining really mines what is known as reality traces, these include all patterns of human life in digital form. Traces include banking transactions, travel tickets, mobile telecommunications calls, blogs, and every possible digital transaction. The aim of such emerging technology is to better understand societies as well as individual and to further develop solutions aimed at them. The main problem facing such new mining technology is privacy concerns from individual, and governments, as data spread on the Internet is not really owned by any legislative body.

3. Related Work

Much work was done in surveying business applications of the aforementioned mining techniques. However, most work considers each mining technique separate from one another. In [4] the authors have provided an overview of Knowledge Discovery in Databases (KDD) approaches. They also classified the approaches depending on software characteristics. In [5] the authors demonstrated how modern technologies shifted the process of decision-making, from manual data analysis using modelling and stochastic to an automated computer driven process. The authors also stated that knowledge discovery tools have benefits such as increased profitability. In addition, risk management and market segmentation is another advantage. A survey of visual data mining techniques is found in [11]. The authors have stated that large data sets with complex dimensions need a better way for representation. In their paper, the authors have reviewed previous work done in data visualization [12]. The authors classified data visualization techniques into six different classes, based on the parameters of the data.

In [13], the authors have surveyed the relatively young and interdisciplinary field of TM. Since most information found in computerized form are textual, the need to extracts this unstructured text into informative knowledge demands new tools. TM tools are machine tools that analyse written text with a certain context [15]. A case study of TM is found in [16], the paper discusses the use of TM for patent analysis. The authored discussed how professional patent information business is sceptical in using TM tools. The paper discussed showcased PackMOLE (Mining Online Expert on Packaging Patents), a TM tool, designed for mining patent information in the packaging field. The authors showed that PackMOLE tool has advantages over the manual patent portfolio analysis techniques. However, the tool calibration of its internal clustering processes is difficult, and consumes time. This leads to the use of a hybrid use of text mining techniques and manual patent classifications in conjunction. In [18], the authors presented a review of TM techniques. The authors clearly stated that TM faces challenges as natural language processing NLP techniques are not readily made for mining activities. The paper illustrated several TM technique that included information extraction, topic tracking, summarization, categorization, clustering, concept linkage, information visualization and question answering to name a few. Finally, the author stated that TM is used in media, banking, politics, and even in insurance.

How business intelligence is derived from web mining is found in [17]. WM or web usage mining is described as an intelligence tool to aid enterprises in the intense competition found in e-commerce. The paper presented a review of current WM techniques used as well as introducing a novel approach called intelligent miner. Intelligent miner (called i-Miner) is a hybrid framework for WM; it uses a combination of algorithms for finding and processing log files from web servers. Its then applies rules and structures to find hidden patterns found in the log files. In [20], the authors stated how difficulties arise in WB from all the fuzziness and unstructured nature of the data found in the WWW. The paper also illustrates the evolution of DM that lead to WM. The authors stated that WM has these main task, associations, classification, and sequential analysis. The paper included a WM study on two online courses. Using WM to improve the two online courses experience, based on the results from the WM tool that used the logging files. An excellent discussion on the characteristics of WM is found in [26]. The authors relate the development of the soft computing, which is a set of methodologies to achieve flexible and natural information processing capabilities. The paper discusses how it is difficult to mine the WWW with its unstructured, time varying, and fuzzy data. The paper also specifies four phases that include information retrieval IR, information extraction, generalization, and finally analysis of the gathered data. The authors also classified WM into three main categories, Web Content Mining WCM, Web Structured Mining WSM, and Web Usage Mining WUM. WCM is about retrieving and mining content found in the WWW like multimedia, metadata, hyperlinks, and text. WSM the mining of the structure of WWW, it finds all the relations regarding the hyperlinks structure, thus we can construct a map of how certain sites are formed, and the reason why some documents have more links than others. Finally, WUM, which is the mining of log files of web servers, browser generated logs, cookies, bookmarks and scrolls. WUM helps to find the surfing habits customers and provides insights on traffic of certain sites.

4. MINING FOR INFORMATION AND KNOWLEDGE

How does mining really work? Let’s look on how DM works. It’s regarded as the analysis step of the Knowledge Discovery in Databases KDD process [5]. KDD usually has three steps, pre-processing, then DM, then finally data verification [6]. For DM, it uses data stored in data warehouses for analysis. DM technologies use Artificial Intelligence AI and neural networks; a good review of neural networks applications in business is found in [45]. AI and neural networks are nonlinear predictive models [3] that learn through experience and training. Furthermore, AI techniques are highly used in business models and predictions [50]. AI led to the more advanced technique of machine learning. Machine learning is the ability of the machine to adapt and learns from previous trials and errors, and then it tries to find out how to be more effective. Other techniques are decision trees and genetic algorithms. Decision trees [21] are top down induction rationale thinking tools, they support classifying the decisions into different branches. Starting from their roots and stemming to the leaves, each decision branching out has risks, possibilities, and outcomes. Certain decision trees type used are Classification and Regression Trees CART.

DM tasks are several and depend on the different fields where the DM is applied. Classifying data stored in multi-dimensional databases is a prominent task. Classifying involves identifying all groups that can be found in the data, like grouping fraudulent transactions in a separate group from legitimate transactions. Associations and rule inductions, intelligently inferring if-then relations from patterns found hidden in the data. This leads to finding hidden correlations, like market baskets, which are the products bought mostly together. Another task is regression modelling or predictive modelling [47], which helps in predicting future trends. Usually regression is used for extrapolating data in mathematics, in DM; it helps to find a model that fits dataset. Data visualization, visually aids in linking multi-dimensional data together, such as Exploratory Data Analysis EDA and model visualization. Data visualization tasks are emerging task of newer interactive DM tools. Other tasks include Anomaly detection, were anomalies are caught. Summarizations that express information extracted in a compact form; and aggregation of data, where sums of data are compacted to single figures or graphs [4].

DM requires huge computational resources. DM requires as a perquisite a data source, usually a data warehouse, or a database. Data warehouses [19] are large databases used for data analysis and aggregation. It extracts and transforms the data from the DBMS the enterprise use for daily activates and through Extraction, Transformation, and Loading ETL processes [20]. ETL extracts the data from the DB, then it pre-process the data and finally load it into the data warehouse for further processing of this cleaned up data. Aside from this, DM requires substantial processing power, usually top class server level computing prowess. A growing interest in combing data mining with Cloud Computing technologies is emerging such as providing DM as a service, such as found in [40]. DM applies number crunching algorithms, parallel processing, and neural networks with AI techniques, thus requiring this huge computational resources. For standardizing DM, the CRoss Industry Standard Process for Data Mining or CRISP-DM for short, was developed [14]. CRISP-DM is standard model developed in 1997 by the ESPRIT funding initiative, as part of a Euro Union project, and lead by leading industry companies such as SPSS, now part of IBM, and NCR Corporation. Until data, CRISP-DM is the leading standard for DM, as most leading DM software implements it.

The next question is how does TM work? TM essentially is based on how do we read and comprehend text. This process of reading, then understanding what is read is to some extent imitated. However, this is not a very easy process for computers. For TM techniques, it first has to retrieve relevant documents, data, or text found on the WWW. For this step Information Retrieval IR systems are used [23], Google search engine is an example of such systems. The second process is NLP; this is the most difficult part of TM. In NLP, AI and neural networks are again used to parse the text the same way humans do to comprehend the text. The text is parsed; nouns and verbs are used to understand grammatically the meaning on each sentence. The final step is information extraction; here linguistic tools are used to get information from the comprehended text. Entities, characters, verbs, and places are correlated and hidden unknown new information is generated. This is where DM techniques are applied at the final stages to extract the information. This is how most TM tools work.

TM techniques try to emulate human comprehension of textual data. TM highly used in many applications to replace manual search in textual documents by humans. TM techniques, unlike DM techniques, deal much with unstructured data sets, thus more complicated. Healthcare and medical usage includes linking several hundred of medical records together, finding relations between symptoms and prescriptions [47]. Media applications also use TM, especially in the political aspects of certain controversial issues or controversial characters. TM is also used for clustering archived documents into several clusters according to predefined semantic categories [24]. TM is newly finding new uses in the legal and jurisdictional fields, as TM is applied to patents, and criminal profiling. TM is used in text summarizations, where it effectively identifies the main names, characters, verbs, and the most used words or referenced subjects in large documents. TM is also used in TM OLAP, found in [59], as a textual search tool rather than numerical. Furthermore, TM is an essential part of modern IR engines, such as Google's search engine and Yahoo's search engine, as they apply sophisticated TM techniques to correlate search queries together.

The final question is how does WM work? WM is based on Internet and agent technologies [25], that utilize soft computing and fuzzy logic techniques. Again, WM technologies rely on IR tools to find the data it needs. IR systems provide means and ways in which these intelligent agents can scour the WWW. In addition, it is worth noting that such IR systems are greatly supported by the developments of Semantic Web technologies, SW technologies [51]. SW technologies were developed to build semantics over web published content and information on the WWW, in the aim of easing the retrieval process of the content by humans and machines. As SW technologies are still a work in progress, as more new web technologies are implementing SW techniques to aid in the search process of such contents. The next step is the agents programming, these agents are programmed after finding their targets to apply their mining techniques. These include analysing the HTML document, parsing, and extracting all the hyperlinks, multimedia among other things. Users' preferences and online accounts are specifically tracked to identify their sessions and transactions logged in that system. Server data, like site traffic, activates and even possible proxies are retrieved for further analysis. Final step for these agents is to analyse the gathered data, using DM techniques, to understand the habits, patterns found in the WWW.

WM tasks depend on the mining purpose intended. WM is divided into three main categories as in [26]; web content mining WCM, web structure mining WSM, and web usage mining WUM. WCM is intended for retrieval can fetch and locate context sensitive text, multimedia, and hyperlinks depending on the giving context. Other WM tasks intended for WSM include finding the chain of links or site maps of certain sites, mainly to find were most of the traffic is headed. Finally, WM intended for WUM can collect logs, cookies, bookmarks, and even browsers history and metadata of targeted users. WUM is also used for mining social networks, namely online blogs [33]. These tasks are all WWW oriented, but WM also plays a key role in mining social media networks for investigations, intranets, and to lesser extent Virtual Private Networks VPNs. A new use of WM is multimedia mining, where WM tools will try to mine out multimedia as pictures, movies, audio files, and applications.

4.1. Mining Tools Software

Mining tools of sometimes called siftware, as they sift through the data. Mining tools varies depending on their sophistication, as state of the art tools are expensive. In 2008, the world market for business intelligence software reached over 7.8 billion USD according to [27]. IBM SPSS is an example of business intelligence software, offering mining capabilities. Clementine was a graphical and widely used DM tool of the late 1990's; this case study utilizes that software [43], it is precursor to SPSS. IBM also provides online services for WM, called Surfaid Analytics; they provide sophisticated tools for WM [32]. IBM is one of main providers of solution-oriented packages such as IBM's Cognos 8 solutions [44]. Other types of mining or business intelligence tools are integrated tools, like Oracle Data Mining, a part of the company flagship RDBMS software. SAS offers its SAS Enterprise Miner [47], a part of its enterprise solutions. Another major player in the enterprise and business information systems is SAP, it offers world known ERP solutions along with providing other mining tools software that can be integrated into their ERP solutions. Atos is an international information technology services provider that utilizes SAP software [63]. Microsoft offers SQL Server Analysis Services, a platform dependant solution integrated in Microsoft SQL for Microsoft Windows Server. Other Microsoft products include PowerPivot, a mining tool for small and middle size enterprises. Open source mining tools include the Waikato Environment for Knowledge Analysis or WEKA [28]. Other open source tools include RapidMiner and KNIME. Situation with open source mining tools however is not as other open source software, as their use is quite limited [49]. Furthermore, with huge decrease of in costs of storing and acquiring data from WWW, data acquisition tools such as RFID tag readers and imaging devices, e-commerce, and telecommunications, mining software costs of procuring dropped considerably [42].

A new trend in utilizing such mining tools for middle-sized and small enterprises is Cloud Computing based Mining tools [58]. As small and middle-sized enterprises, lack the infrastructure and budget available for large enterprises. Cloud Computing helps in providing such mining tools at relatively lower costs. Cloud Computing provides web data warehousing, were the actual data warehouse is outsourced and accessed through the web. They also provide sophisticated mining analysis of the data as the enterprise specifies and demands. Aside from lowering the costs of the mining tools infrastructure, Cloud based mining also provides expertise that are not available in such middle-sized and small enterprises. Usually start-up entrepreneur level enterprises lack not only the financial resources but also the human resources and expertise in the IT field. The Infrastructure-As-Service IAS provides middle-sized enterprises comfort from the burden of software, hardware, and human resources managements as well. The main downfalls of Cloud Computing are the dependency and privacy issues that occur from the fact that another party that the enterprise agrees to store its data on its machines and data warehouses. Such issues are limit and turn off large capable enterprises from going with Cloud based solution. These enterprises can set up their own mining solutions instead is much less risky. Dependency means that the whole service depends on the other party, not the enterprise itself, meaning that the enterprise is pretty much tied up with what the service provider has to offer. The privacy concerns arise from the fact that the enterprise's data is technically not under its control or even possession, the other party has it, it utilize its resources to give results and analysis. The privacy concern entails the misuse of the data, mostly causing confidentiality risks

.

5. Business Intelligence through Data, Text, and Web Mining

Business applications rely on mining software solutions. Mining tools are now an integral part of enterprise decision-making and risk management. Acquiring information through mining is referred to as Business Intelligence BI. Enterprise datasets are growing rapidly, thanks to use of Information Systems IS, and data warehousing. On average, credit card companies usually have millions of transactions logged per year [29]. Largest data sets are usually generated by large telecommunications and mobile operators as they mount up to 100 million user accounts generating thousands of millions of data per year [30]. As these number mount up, analytical processing such as OLAP and manual comprehension seems ineffective. With BI such tasks are within reach. According to Gartner Group "Data mining and artificial intelligence are at the top five key technology areas that will clearly have a major impact across a wide range of industries within the next three to five years," this was back in a 1997 report [31]. Also according to Gartner Group reports in 2008 [53], it was found that 80% of data found in enterprises information systems is unstructured, and would very likely to double in size nearly every three months. BI has become the prevalent decision support systems in organizations. BI has dominated many industries including retail, banking, and insurance [41]. The First American Corporation FAC is an example of a success story in implementing BI to improve its customers' loyalty and better investment. It's worth mentioning that BI software is aimed at knowledge workers, mainly executives, analysts, middle management, and to a lesser extent operational management. Fig.1, illustrates how data today is transformed to acquire BI.

BI is implemented through mining tools; these tools generate findings that are ultimately used to gain competitive advantage over rivals, better and efficient business operations, and better survivability and risk management. Mining tools provide better customers' relationship management CMR, through mining real habits, patterns, and even customers churn. Customers churn is defined as the per cent of customers that have left the enterprise, most likely to other rivals or due to the inability to keep your competitive advantage and customer satisfaction levels. Habits and trends in customers' data help in discovering the customers' segmentations, what customers to target, especially alpha customers. Alpha customers are those that play a key role in a product success, thus finding what they want is essential. This means that, mining tools are essential for catalogue marketing industries and advertising agencies. In addition, mining tools, especially DM, provides market basket analysis that helps the discovery of products that are bought usually together. As modern economies around the world today are driven by information, becoming information and knowledge based economies [66]; BI tools are from the top reasons of development information technologies in business today. BI tools in business today are integrated in most enterprises tools such as Enterprise Resource Planning ERP tools, Customer Relationship Management CRM tools, supply chain management tools, data warehouses, and even RDBMS. BI is also the main tools for decision support in modern enterprises. BI tools provide competitive advantages, better customer relationships managements, and better management of risk in investments.

Mining tools provides predictive profiling; this means that using current and historical behaviours of your customers, possible future behaviours of purchase are predicted. The insurance industry is most interested in their customers' medical records and its history. Stock market predictions are mostly done using mining, NASDAQ, is a major DM user. NASDAQ had spent over 450 million dollars [64], estimated costs, on implementing a full BI solution for its stock market exchange. Their solution consists of smart BI agents that buy and sell stock autonomously, with human monitoring for erroneous errors. Google, the technology giant, uses BI on its Google Finance service [59]. The Google Finance web page contains dynamic charts of international stock markets, with references to critical points in the graph directing to web pages that are the service got its information, providing assurance to end users. Mining tools are also used for automatic spam detection, and the defence against fraud, through fraud detection techniques utilizing mining tools [30]. Most major banking and telecommunication companies apply automated fraud detection systems through mining techniques, AT&T, bank of America are examples of such users of fraud detection. In the next subsections we will look at the main aspects were BI through mining tools is used to gain business proficiency. BI and mining tools are used exchangeable in the following text.

5.1. Achieving Competitive Advantages

Competitive advantage is driven by competitive pressure. According to [41], competitive pressure is degree of pressure that companies feel from rivals and possible new entrants. This pressure is lessened by gaining a competitive advantage. For gaining competitive advantages, enterprises develop market research groups that analyse the large data sets to acquire knowledge. Market research, through mining, try and find what products dominate the market, why this is and what hidden elements that set such products leading in sales. For example, media networks use mining in their market research to set the common factors between audience and the program's scheduled slot. BBC, used to hire human experts to schedule its programs slots, now its uses fully automated mining tools for scheduling, the results were equivalent or better than the human manual scheduling [36]. Marketing use-mining tools to get the market's baskets, as mentioned before, market basket are associations of certain products that are highly likely bought together. No competitive retail enterprise is without its set of market baskets, leading in this segment are Wal-Mart, Costco, and K-Mart.

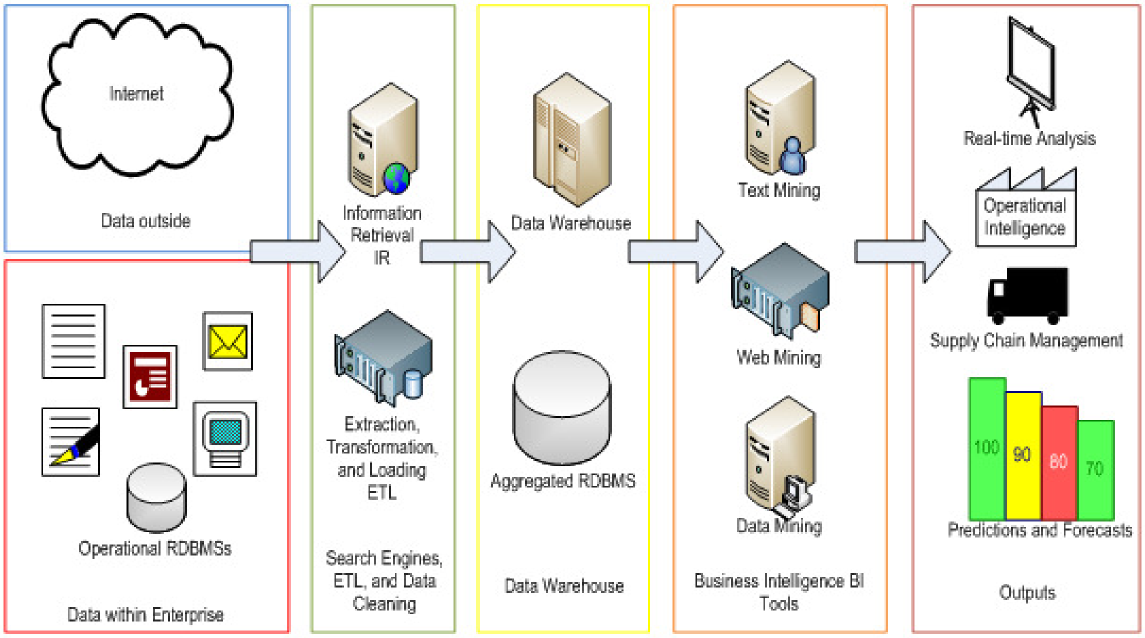


Figure 1. Data Transforming into Business Intelligence

For better risk management, in banking for example mining tools are used to automate risk of bankruptcy. An example is bankruptcy predictions [38]. Bankruptcy predictions are essential risk assessment processes, nearly all large banks such automated tools in their bankruptcy predictions. Most of these tools use neural networks in assessing the counterpart's liability to pay back the loan given. Bankruptcy prediction tools help the bank in reducing the risk of lending money to trouble or would be troubled customers. These tools also depending on the risk the bank is willing to take, will estimate the fair interest rates depending on the credibility of the other party. In addition, bankruptcy prediction tools are also used to assess the credit loss of current and on-going bank's loan portfolio. Credit loss is the prediction of possibility that depreciation or loss levels of loan will be more than the expected 1% in a predefined time frame. These tools came to the foreground after strict regulations especially after the 2008 market crash and following economic crisis that led many banks to file for bankruptcy. Banking sectors today utilize BI with bankruptcy tools together to create sophisticated profiles on their potential customers before any lending. Usually banks have setup return/risk parameters depending on the other party's market worth, market share, and current debt. According to [38], bankruptcy tools are also used by accounting companies. Auditing now takes into account the possibility of going bankrupt in the near future. Lawsuits are filed against accounting companies if they did not warn the contracting party of its possible bankruptcy. J.P. Morgan utilize its' own bankruptcy tool called CreditMetrics [39], it model the change in the credit quality ratings, it can obtain a possible estimate of the risk incurred in the loan.

In manufacturing industries, sophisticated Management Information Systems MIS are used to automate the processes. Such MIS tools generate reports that can be feed to BI tools to better optimize the manufacturing processes [61]. MIS tools are used to automate the work, but such tools were not meant to monitor the performance nor smartly detect how to use the operation data to further develop and optimize the manufacturing process. MIS software for the manufacturing industry includes Material requirements planning MRP systems, which detects when the finished products will be needed or shipped out of the factory. Other example of MIS software for manufacturing is Just in time JIT inventory systems, precise and immediate delivery of materials before usage. These tools also generate huge amount of raw data, reports, and schedules, which could be used for mining. According to [61], BI tools using such data can generate adaptive manufacturing schedules to match the demands. BI tools help the top management further to lessen the wastage and unnecessary expenditures. BI tools not only improve the manufacturing processes in the physical factory, but also optimize pricing analysis, by examining sales, production, and quarterly reports, the manufacturing enterprise can adjust its pricings on its items. Other optimization of other processes include warranty analysis, by examining the warranty claims and lawsuits occurring from manufacturing errors, failing parts, and possible risk factors. As such, the manufacturing enterprise can reduce such errors, forecast possible risks and avert flawed warranty strategies.

5.2. Improved Customers' Relationships

Customers today are well informed, and demanding. Meaning customers are getting even more difficult to please. For these reasons, Customer Relationship Management (CRM) systems were developed to better handle the customers. CRM is defined as the process of managing the relationships of a company's customers through technology focusing on the customers' satisfaction and further development of these relationships [46]. With information systems availability since the mid 1980's, companies have used information technologies further to develop their customers' relationships. CRM software automates how the company deals with its customers' information, but with the introduction of BI tools, further and newer exploitations of such information are possible. Questions like, how to keep your customers satisfied? What makes a customer happy? How can we attract potential customers? Why did some customers have our company left for other rival companies? All these questions are answered by BI tools. BI tools are used with existing CRM systems to help in customers oriented decisions. These decisions include customers' service, market targeting, product evaluation, and even in manufacturing. The main users of BI tools with CRM systems are retailers, electronic retailers or e-tailers, services providers, and telecommunications.

Though BI helps in gaining a competitive advantage, some businesses are heavily dependent on its customers' satisfaction [48]. This leads to extensive BI exploitation of the CRM data. Companies can achieve competitive advantages in providing value propositions to its customers; these value propositions include lowest cost, bundles and offers, special discounts, and loyalty and partnership privileges. Lowest cost however is not easily achieved, as the WWW had brought new ways for the enterprise customers to search for other rivals possibly offering better prices. Transparency in pricing is another factor that customers would appreciate and look for. BI tools help in avoiding price transparency problems by offering the enterprise the choice of price discrimination. Price discrimination [48] means that the company offers the same product with different prices for different people and regions; this of course depends on each individual's willingness to pay for how much. This dynamic trait allows the enterprise to appeal to its diverse customers with offers that are identical to the lowest cost offers. This ability is not possible without prior knowledge of how much each individual is willing to pay of course. Here is where BI tools help in evaluating your customers buying and spending habits. For example, some customers are willing to pay more for certain products, depending on personal preferences such as hobbies, age, or culture. These customers the company offers the price it thinks they will gladly pay, most of the time they will pay. Other customers, would never pay for that price, instead they may be willing to pay for less. Such customers the company can offer limited and special discount prices these customers will surely be happy with such pricings. It worth noting that sophisticated BI techniques must be employed to achieve effective price discrimination. A tool that utilizes DM in CRM software is found in [67], through a data mining engine DME, which mines the CRM database.

BI helps in assessing possible switching costs of customers. As stated in [48], high switching costs leads to lower customers' satisfaction. As customers are not satisfied with being stuck with a company because of such costs, it reduces their overall satisfaction. In addition, BI finds the market baskets, as stated earlier; this is from the customers' side though. As market basket, analysis can be considered from the customer side and the company market side. Through CRM data, BI tool can construct specific market baskets for certain groups of users that is more specific than the ordinary market baskets. Amazon, Google are among the best utilizes of customers market baskets. Market baskets are seen in the form or recommendations, special discounts on certain products, and the usually advertisements regarding specific products. Furthermore, market baskets implemented for e-tailers, another name for e-commerce retailers, are more effective because it is more easily exploited than traditional brick and mortar retailers, where certain products could not physically be with others.

Services industries have another use of BI tools; it is in their Customers Experience Management CEM systems. Such systems keep track of the experience of customers. BI tools help find evidence of what make customers satisfied, happy, and what do they expect from the enterprise. As rivals are always trying to take the enterprise's customers and recruiting them. BI is also used to find reasons for customers churns. Customers churn, as defined earlier, are enterprise health indicators. As high customers churn ratings are an alarming indicator that customers are unsatisfied. Churn analysis can also show reasons and causes for these turn over, like turning to rival companies, or found substitutes to the enterprise products. BI tools utilizing CEM systems can also help identify the alpha customers, as stated earlier, those highly valued customers that the company must take special care in dealing with them. The emerging Reality Mining RM techniques can also be used to acquire further insight of the costumers. As it is an emerging field, most of the methodologies are experimental. As mentioned before RM faces tremendous pressure as it deliberately violates individuals' privacy, anonymity, and even to a lesser extend security.

5.3. Better Logistics and Inventory Management

Supply chain management systems such as Warehouse Management Systems (WMS) [34] and Inventory Management Systems IMS; provide automated control and management of the supply chains. These systems achieve logistics and production efficiencies, such as high distributors and customers satisfactions, on demand supplies, and quick response to inventory levels. Logistics industries make profitability out of efficiency of their supply chain managements. Using BI tools, help supports supply chain management; enterprises have now better management over its supplies. BI tools, integrated in WMS or IMS, help find patterns, shortages, possible overproduction, and underproduction, but mostly quick response to demand spikes that are extremely important to catch.

Traditionally, complex mathematical models were used for supply chain managements and logistical problems. Logistics can affects inventories (under-flows or over-flows), specifically from slow response to spikes in demands and lack of precise forecasting. Order batching methods were used to batch many items to as many as possible different locations [35], also routing batched orders was another problem. Using association rules and clustering, BI can help in effectively in routing, and batch these orders. BI tools mine markets current situations, providing forecasting predictions. These predictions help in managing inventories, keeping under-flows and over-flows at bay. BI tools have complete control over forecasting, as demands and surges are estimated through reading market history, political events, or even rivals retributions. Modelling, beside prediction, is another BI facility used for warehouse management, as certain industries fall in certain demand and supply cycles. Models also help automate replenishment processes. It is worth noting that sophisticated BI tools can predict certain market segments demands. Such market segments are used in niche marketing.

BI help logistics in providing strict and timely decision supports. The utilization of using such BI tools results in delivering on time decisions and faster feedbacks, such benefits are of BI in food supply chain networks are found in [60]. In the food industry, time of delivering the goods is critical as the stock of food supplies or livestock can rot or die during anyone of the many supply chain stages. As poor management and slow reactions in such supply chains can give low yields, meaning that slow reactions can lead to massive loss that could be avoided. The problem of Death-On-Arrival DOA occurs when many of the livestock die during transport due to reasons unknown or known much later in the supply chain. BI tools can effective identify such signs of dying livestock at an early stage, that some of the livestock are dying; this information can help in reducing the casualties or identifying the problem and rescuing the remaining livestock. Aside from the livestock management and preservation in the supply chain, food products like milk, sugar, or fruits, and other raw food materials, spoil if not handled in time. BI help in dynamically routing the supplies before it rot during the storage, or distribution. Many BI tools provide timing managements of such supply chains.

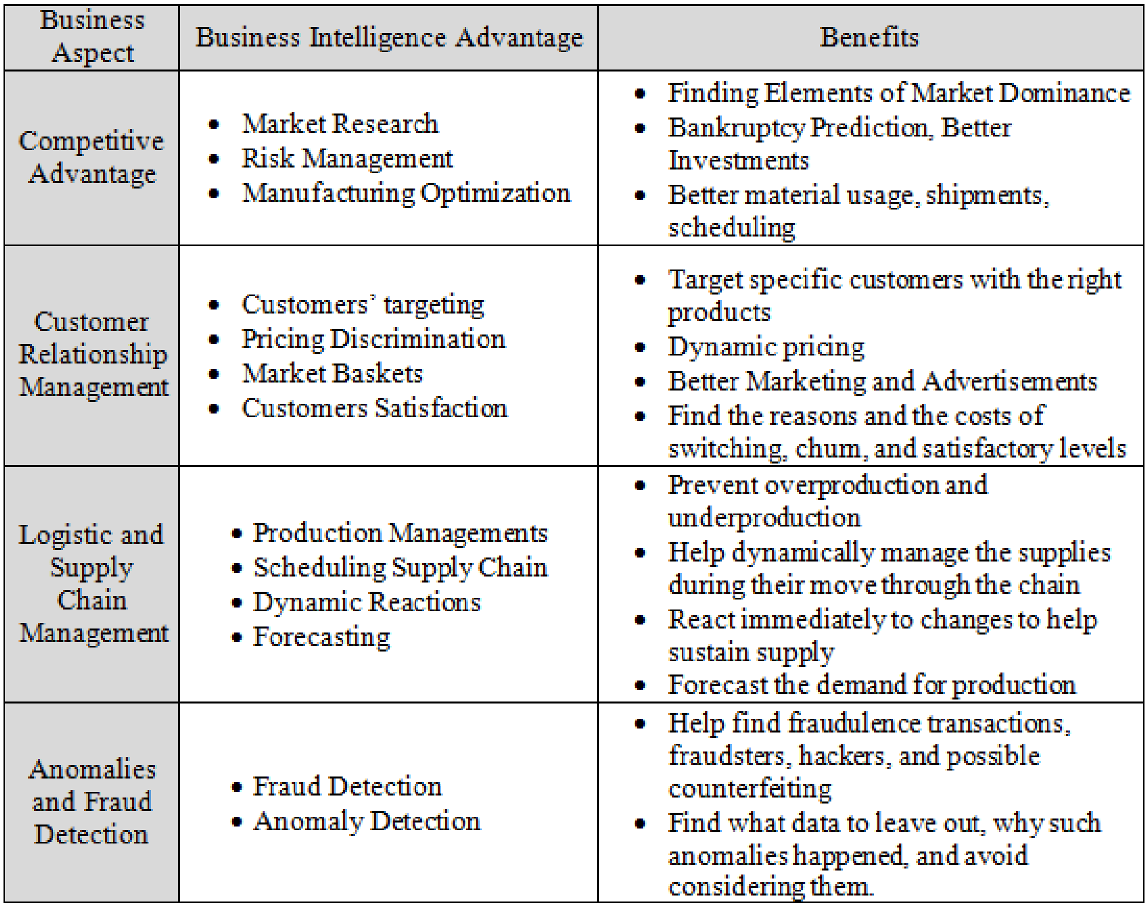
One of the main users of BI tools in supply chain managements are the Third Party Logistics (TLP) or 3PL. TLPs are an outsourced party that provides logistics services to the contracting first party. TLPs using BI can provide much more services to their customers [62], in addition to the traditional services; BI enables new services such as detailed reports and forecasts about their delivered goods. Another additional service is the cost-benefit analysis, this service helps their customers to analyse their current supplies and evaluate their suppliers. Another service powered by BI tools is the supply chain visibility, allowing the customers to track their supplies online dynamically and in real time.

5.4. Anomalies and Fraud Detection

BI tools through its sifting capabilities can locate certain hidden patterns found in daily transactions. These patterns are used for possible fraud and anomaly detection. Fraud and anomaly detection is defined as detection of deceptive transactions or strange and unusual transactions that need inspection. BI tools are one of the main tools used in forensics to detect fraudulence. Most insurance and telecommunication companies use fraud detection daily. As it’s the norm to investigate suspicious transaction for possible fraudulence or anomalies. As such, some businesses are plagued by fraudulence and possible anomalies in their transactions [30]. Telecommunications companies and credit industries are some of the most plagued industry, since it is hard to detect such fraudulence in these businesses. On the other hand, anomalies are not deliberate actions like fraudulence. They are unusual behaviours that may manifest in normal data for unknown reasons, corrupted data, glitches, or network errors in transmissions. If not identified correctly, anomalies can lead to a shift and possible divergence in finding real frauds. Anomalies must be detected and left outside the data set, as to leave out the surge or upheavals these anomalies bring in the data. Fraudulence is however deliberate actions are, usually caused for monetary and financial gains, and is usually carried out by white-collar criminals, insiders, high profile technology criminals, and expert computer hackers.

Usually credit companies try to find certain patterns of charging for its customers. For example, usually stolen credit cards will result in erroneous behaviours and transactions. These transactions usually happen in short periods, within few hours; huge sums of figures in thousands to millions are spent in numerous transactions. The fraud detection used to be manual, through reviewing record manually. As BI fraud detection can easily spot such anomalies. Fraud detection is not only limited to fraudulent customers, but to also fraud within the company, as in fraudulent reports and predictions. BI tools detect suspected behaviours over periods of time, providing what accounts or individuals that require special intention. BI is also used in financial crimes like money laundering, through sifting records of individuals, and insider trading, trading upon secret inside information, Intrusion detection, by outsiders into the system, and spam detection, a major problem facing many enterprises email systems. Finally, we can say that the main problem with such tools would be the false positives that can occur with high percentages in some cases, and miss predictions that have legal and monetary consequences.

Table 1. Summary of BI Advantages



Business Aspect	Business Intelligence Advantage	Benefits
Competitive Advantage	• Market Research • Risk Management • Manufacturing Optimization	• Finding Elements of Market Dominance • Bankruptcy Prediction, Better Investments • Better material usage, shipments, scheduling
Customer Relationship Management	• Customers' targeting • Pricing Discrimination • Market Baskets • Customers Satisfaction	• Target specific customers with the right products • Dynamic pricing • Better Marketing and Advertisements • Find the reasons and the costs of switching, churn, and satisfactory levels
Logistic and Supply Chain Management	• Production Managements • Scheduling Supply Chain • Dynamic Reactions • Forecasting	• Prevent overproduction and underproduction • Help dynamically manage the supplies during their move through the chain • React immediately to changes to help sustain supply • Forecast the demand for production
Anomalies and Fraud Detection	• Fraud Detection • Anomaly Detection	• Help find fraudulence transactions, fraudsters, hackers, and possible counterfeiting • Find what data to leave out, why such anomalies happened, and avoid considering them.

6. UTILIZING MINING TO GAIN BUSINESS ADVANTAGES

6.1. Jaeger Uses Data Mining to Locate Losses

Jaeger is midsized privately owned British chain cloth retailer in the UK [56] with hundreds of outlets through the different regions of the UK. During the 2008 recession in the UK, shoplifting was on the rise. Retailers like Mark and Spencer, Tesco were accumulating losses in a rate that is at least two times higher than the previous years. According to The Centre for Retail Research, customers stole more than £ 1.6 billion, while employees stole £ 1.3 billion, and finally suppliers took more than £ 209 million fraudulently. The centre has also a Global Retail Theft Barometer, which was estimated at 1.3% of the total loses in sales in the UK in 2007. Current technologies to defend against such theft are the common Closed-Circuit Televisions CCTV surveillance systems, also widely used in retail shops are Electronic Article Surveillance EAS. However, even with such technologies implemented; many retailers still can’t estimate losses. The new and innovative technologies made possible that data mining tools, to help in locating the lose source.

Jaeger, in 2008, has implemented a DM application to identify its losses. Their application was centralized, used to process data collected from all the other outlets to try to make sense of it. The DM application for Jaeger called LossManager, provided by IDM software. The software uses feed from Jaeger’s other information systems such as the Electronic Point of Sales EPOS system; as it was meant to monitor the employees’ behaviour as well as an EPOS. As it is quite common for fraud transactions to be made by some employees in the form of unauthorized discounts. Jaeger did had another DM software before its current solution, but it was far too cumbersome and it did not integrate with the Jaegers’ other information systems very well. The current LossManager however was developed with the current systems in mind to provide possible integration.

Jaeger’s main aim with LossManager was not to detect shoplifters, as CCTV and the EAS systems were aimed at that purpose. LossManager’s aim was to spot theft coming from employees, through fraudulent transactions, loss of inventories, and marginal losses from unnecessary working practices. The audit team at Jaeger was in charge of reading the reports generated from the DM tool. The team analysed the reports to identify the possible dishonest employees and the sources of lost money. However, the report generate by LossManager, led to some false positives. To overcome this, audit team used these reports to investigate further through the data using the DM tool. These questions helped avoid false positives the system generated. Questioning process helped the team to understand how the tool generated reports and what reasons for the patterns did it conceive.

LossManager reduced losses to less than losses accumulated from theft and losses in 2007. Jaeger is already expecting a Return on Investment ROI in its first year of using the system. Although the managers were concerned with the double checking process that the audit team takes in every step. Major findings were that the losses from employees theft consists a very small portion of the losses. In addition, the DM tools found out many erroneous transactions, not all were identified as fraud though. The system also helped Jaeger to better manage its inventory stocks, reducing lose gained from the stock going off-season, or missing items from the other outlets’ inventories.

Jaeger experience with LossManager proved that the DM tool, which was meant for fraud and theft detection, did manage to be useful in inventory stock management as well. The DM tool helped Jaeger in 2009, when the recession worsened. The DM tool helped in keeping a low profile on its losses generated from lose and theft. It worth noting that the main difficulty faced was mainly came from the integration with the current systems at Jaeger. The DM tool, LossManager, was implemented in C++, using Microsoft Development Environment, to interface with most other systems. The feeds from the EPOSs, EASs were challenging, as they were the source of info into the DM tool. In addition, the system being centralized made the data collection time consuming as well. As LossManager was integrated well with the EAS feeds, it showed a significant relation between stores with high lose figures and frequent EAS breakdowns.

6.2. KFC/Pizza Hut Find a Better BI Tool

KFC/Pizza Hut in Singapore [68], have more than 120 outlets, with a workforce of 5000 employees. As an international fast food franchise, they deliver food and beverages to customers through outlets, drive-through, and by home delivery. To deal with such workload, KFC/Pizza Hut have used a BI tool; the tool was growing increasingly inefficient with each month. Tool didn’t meet with time requirements to deliver business reports. It was also had problems with performance benchmarking, plus daily reports across multiple systems was tedious. KFC/Pizza Hut most important daily operation was to calculate payments needed for daily paid workers, such as deliver staff. The used BI tool, managers would take hours and had to work for extra hours to sum up pay correctly. Finally, the old system reporting was slowing down KFC/Pizza Hut ability to match and adapt to current and rapid changes.

Solution was to find another BI tool that was modern. KFC/Pizza Hut contracted with Zap, a BI vendor, using their product Zap Business Intelligence. New solution was web based; it was also linked to other external sources. Corporate data warehouse was remodelled as to include the point of sales POS, marketing, human resources HR, and the corporate very own supply chain. In September 2009, after two month of testing, KFC/Pizza Hut went live with the new BI tool. The employees and managers were generally happy with the new tool. As it was web based, and it offered modern BI capabilities like dashboards, instant report generation, KPI benchmarking, scoreboards, and a very user-friendly interface.

The benefits of the new tool were significant. The improvement included optimized market spending, through live updates; KFC/Pizza Hut immediately responded and adjusted its marketing campaigns and offers. Restaurant planning and outlet location managements were based on reports given for the tool, to cope with KFC/Pizza Hut strategy of being close to its customers. Customer service was highly improved, especially the home delivery service, as the tool accurately capture the parameters of such deliveries to optimize the delivery process. In addition, the POS integration into the data warehouse allowed KFC/Pizza Hut to manage its deals and offers per outlet; different customers at different locations had very varied demands.

Finally, the new BI tool, Zap Business Intelligence, had an expected ROI within 12 months of deployment, as the staff members were reduced, and daily time wasted on reporting was cut to minutes. Workforce efficiency increased and managers do not have to work those extra hours every day. The major cost reduction came from the lessened reliance on the IT staff. The new tool was web based and very intuitive and friendly to use, as most of the operations were carried over to the servers situated at KFC/Pizza Hut’s central IT centre. Little IT provision was needed on the different outlets, and fewer staff could manage the new BI tool.

7. Challenges Facing Mining Technologies

Many challenges hinder mining tools in business today. They come in three main categories, technological, ethical, and legislative challenges. Mining tools are classified as highly specialized software. This means that they cost in millions, and require extensive infrastructure. The need for this infrastructure stems from their sophisticated and specialized nature, as these tools are considered business professional tools. Aside from this, the human resource, to manage such tools are very scarce, as most business graduates are not trained to use such tools, as most of them are trained on using manual techniques [37].

Technical challenges include huge and elaborate infrastructural needs, and software limitations. Most mining tools need data warehouses as a perquisite with its hardware infrastructure, in case of DM. TM and WM require dedicated hardware and a set of software, the hardware include high-end application server and web servers, distributed computational grids are the main platform for such mining tools. The set of software include, network software tools, NLP packages, even supporting DM packages, and a IR engine to help in search the WWW. These perquisites not only have high costs, but also need expert IT staff as well. Finally, the limitations of such tools today may hinder their usage, as most software package bought from vendors are highly specialized in one single area, like data visualization tools, market analysis tools, for example. In addition, these packages are limited in the sense that they are not extendable; they use certain models, certain techniques that may not be suited for the business models or environment of the procuring enterprise. To a lesser extent, the software interface is limited in some tools or cumbersome, reducing they usage by employees, managers, and executives. Finally, software limitation includes scalability, not all solution scale as well or adaptable to the business environment.

Ethical challenges raised from public concerns about the data found in the Internet. Customers' profiles include private data. Aim of mining tools is not to identify such individuals however, as most data is anonym-zed before use. Still, concerns are raised around how enterprises use individuals' data. Beside the mining application for business, governments are utilizing such mining tools for its national security purposes. Such governmental security agencies try to locate individuals, possible terrorists. These uses, along with business uses of mining tools made public awareness of their legal and privacy rights more evident in programs like Total Information Awareness program [55]. The Total Information Awareness Program was a secret program for the Pentagon, it was aimed at national security and the identification of terrorists, and it used mining tools to sift private individuals' records. Public awareness against the exploitation of individuals' privacy and private data forced the congress to stop funding for this program in 2003. Legal regulations were issued to address these concerns, acts like Health Insurance Portability and Accountability Acts HIPAA, in the United States stated by the congress. The HIPAA act requires a prior consent from individuals regarding the use of their information and the notification of the purpose will their information will be used. Another ethical issue in the mining tools is that they made Globalization far easier. Globalization has dire consequences on emerging economies, as emerging businesses that can never compete with international top companies.

Legislatively BI has resulted in new levels of transparency, due to the vast data decimation across the net willingly or unwillingly, Wikileaks for example. The term data quality [54], is a relatively recent term, refers to authentic, complete, and accurate data and that the source of this data is legally liable for its authenticity. International legislative and professional organizations have made standards and regulations regarding the quality of published data from companies and other agencies. According to [57], more than 25% of critical data in top companies' databases are inaccurate and incomplete. Since BI relies on the data its fed, the quality plays a crucial role, as the quality of the BI is as good as the fed data quality, better and accurate data yields better BI decisions. Open and free market standards today have regulated and insisted that public companies must give accurate fiscal reports, with actual numbers and figures. Such quality of these reports is yet a problem facing BI tools. As it is a problem with taxes collection from such free market enterprises, tax evasion. The transparency of such reports and facts are especially crucial in the energy market. Whereas these energy companies, incredibly and heavily dependent on fiscal reports, cannot only shift market predictions, but also shift other enterprises' plans as well as governments' plans and budget estimates. High quality data is needed for BI in energy markets [52], as demand predictions and market forecasting is the only way to schedule operations and supplies.

8. CONCLUSION

Competitiveness today is driven through BI. Companies achieving high competitiveness are the companies utilizing BI tools. Mining technologies had come a long way. The software development, along with hardware developments made possible of more commercially available mining tools. Revolution of information brought in by the Internet and the telecommunications technologies, made them a huge source of information, sometimes for free even.

BI utilizing this vast amounts of data can help in achieve competitive advantages, better customers' relationships, effective resource planning, and fraudulence detection. As BI tools implement AI techniques, decision trees, NLP, and SM technologies, they are considered as sophisticated and highly specialized tools. Many challenges hinder the further developments of such tools. The challenges are technological, ethical, and legislative. As more enterprises and governments are more dependent on such tools, we think that some obstacles have to go in order to progress. As more developments and innovation is coming everyday into the free market, we see a very bright future for BI tools. A long side the information systems that companies are dependent on these days, future companies will depend on BI tools just as much as their information systems.

ACKNOWLEDGEMENTS

The author would like to thank Professor Hameed Al-Qaheri, College of Business Administration, Kuwait University for his support and help during this research.

REFERENCES

[1] Bill Palace, (1996) "Technology Note prepared for Management 274A" Anderson Graduate School of Management at UCLA.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, (2008) "The Elements of Statistical Learning: Data Mining, Inference and Prediction," New York, Springer-Verlag, ISBN 0 387 95284-5

[3] Doug Alexander, (2011) "Data Mining", dea@tracor.com

[4] Michael Goebel, Le Gruenwald, (1999) "A Survey Of Data Mining And Knowledge Discovery Software Tools," SIGKDD Explorations, Vol. 1, Issue 1. Pg 20, ACM SIGKDD.

[5] Chidanand Apte, Bing Liu, Edwin P.D. Pednault, Padhraic Smyth, (2002) "Business Applications of Data Mining," Communications of the ACM, Vol. 45, No. 8.

[6] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, (1996) "From Data Mining to Knowledge Discovery in Databases," AI Magazine, American Association for Artificial Intelligence AAAI, Vol. 17 No. 3.

[7] Marti Hearst, (2003) "What Is Text Mining?" SIMS, UC Berkeley.

[8] Prof. Anita Wasilewska, (2011) “Web Mining Presentation 1” CSE 590 Data Mining, Stony Brook.

[9] Prasanna Desikan, Colin DeLong, Sandeep Mane, Kalyan Beemanapalli, Kuo-Wei Hsu, Prasad Sriram, Jaideep Srivastava, Vamsee Venuturumilli, (2009) “Web Mining for Business Computing” Handbooks in Information Systems v.3, Emerald Group Publishing Limited.

[10] MineIT (2010) “Web Mining, The E-Tailers’ Holy Grail?” www.mineit.com

[11] Maria C. Ferreira de Oliveira and H. Levkowitz, (2003) “From Visual Data Exploration to Visual Data Mining: A Survey” IEEE Transactions on Visualization and Computer Graphics, Vol. 9, No. 3.

[12] E.H. Chi, (2000) “A Taxonomy of Visualization Techniques Using the Data State Reference Model,” In the Proceedings of the Information Visualization Symposium InfoVis 2000, pp. 69-75.

[13] A. Hotho, A. Nu¨rnberger, G. Paaß, (2005) “A Brief Survey of Text Mining” GLDV-Journal for Computational Linguistics and Language Technologies.

[14] The Cross Industry Standard Process for Data Mining Blog (2008).

[15] Feldman, R. & Dagan, I. (1995) “Knowledge discovery in texts” In Proceeding of the First International Conference on Knowledge Discovery (KDD), pp. 112–117.

[16] Michele Fattori, Giorgio Pedrazzi, Roberta Turra, (2003) “Text mining applied to patent mapping: a practical business case” World Patent Information, Volume 25, Issue 4.

[17] Ajith Abraham, (2003) “Business Intelligence from Web Usage Mining” Journal of Information & Knowledge Management, Vol. 2, No. 4, iKMS & World Scientific Publishing Co.

[18] Vishal Gupta, Gurpreet S. Lehal, (2009) “A Survey of Text Mining Techniques and Applications” Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1.

[19] W. H. Inmon, (1996) “The Data Warehouse and Data Mining” Communications of the ACM, Vol. 39, No. 11, ACM.

[20] Rajender Singh Chhillar, (2008) “Extraction Transformation Loading, A Road to Data Warehouse,” Second National Conference Mathematical Techniques: Emerging Paradigms for Electronics and IT Industries, India, pp. 384-388.

[20] Samia Jones, Omprakash K. Gupta, 2006) “Web Data Mining: A Case Study” Communications of the IIMA, Vol. 6, Issue 4.

[21] J.R. Quinlan, (1986) “Induction of Decision Trees”, Machine Learning, Kluwer Academic Publishers, Boston.

[22] Cohen KB, Hunter L, (2008) “Getting Started in Text Mining” PLoS Comput Biol.

[23] Judy Redfearn and the JISC Communications team, (2006) “What Text Mining can do” Briefing paper, ‘Joint Information Systems Committee’ JISC.

[24] Neto, J., Santos, A., Kaestner, C., Freitas, A. 2000) “Document Clustering and Text Summarization” In the Proceeding of the 4th International Conference Practical Applications of Knowledge Discovery and Data Mining PADD-2000, London, UK.

[25] R. Kosla and H. Blockeel, (2000)“Web mining research a survey,” SIGKDD Explorations, vol. 2, pp. 1–15.

[26] Sankar K. Pal, Varun Talwar, Pabitra Mitra, (2002) “Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions” IEEE Transactions on Neural Networks, Vol. 13, No. 5.

[27] Ralf Mikut, and Markus Reischl, (2011) “Data mining tools” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 1, Issue 5.

[28] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten (2009) “The WEKA data mining software: an update” SIGKDD Explorer News.

[29] Dorronsoro, J., Ginel, F., Sanchez, C. & Cruz, C. (1997) “Neural Fraud Detection in Credit Card Operations” IEEE Transactions on Neural Networks.

[30] Clifton Phua, Vincent Lee, Kate Smith, Ross Gayler, (2010) “A Comprehensive Survey of Data Mining-based Fraud Detection Research” Cornell University library, CoRR.

[31] Sang Jun Lee, Keng Siau, (2001) “A Review of Data Mining Techniques” Industrial Management and Data Systems, 101/1, MCB University Press.

[32] IBM, SurfAid Analytics (2003).

[33] Federico Michele Facca, Pier Luca Lanzi, (2005) “Mining interesting knowledge from weblogs: a survey” Data & Knowledge Engineering, 53, Elsevier.

[34] Mu-Chen Chen, Cheng-Lung Huang, Kai-Ying Chen, Hsiao-Pin Wu, (2005) “Aggregation of Orders in Distribution Centers using Data Mining” Expert Systems with Applications, Volume 28, Issue 3, Pages 453-460, Elsevier.

[35] Van den Berg, J. P. (1999) “A literature survey on planning and control of warehousing systems” IIE Transactions, 31, PP.751–762.

[36] Fitzsimons, M., Khabaza, T., and Shearer, C. (1993) “The Application of Rule Induction and Neural Networks for Television Audience Prediction” In Proceedings of ESOMAR/EMAC/AFM Symposium on Information Based Decision Making in Marketing, Paris, pp 69-82.

[37] Gregory Piatetsky-Shapiro, Ron Brachman, Tom Khabaza (1996) “An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications” KDD-96 Proceedings.

[38] Amir F. Atiya, (2001) “Bankruptcy Prediction for Credit Risk Using Neural Networks: A Survey and New Results” IEEE Transactions on Neural Networks, vol. 12, no. 4.

[39] M. Crouhy, D. Galai, and R. Mark, (2000) “A comparative analysis of current credit risk models,” J. Banking & Finance, vol. 24, pp. 59–117.

[40] Marinela Mircea, Bogdan Ghilic-Micu, Marian Stoica, (2007) “Combining Business Intelligence with Cloud Computing to Delivery Agility in Actual Economy” Department of Economic Informatics The Bucharest Academy of Economic Studies.

[41] Thiagarajan Ramakrishnan, Mary C. Jones, Anna Sidorova, (2011) “Factors Influencing Business Intelligence and Data Collection Strategies: An empirical investigation”, Decision Support Systems.

[42] Surajit Chaudhuri, Vivek Narasayya, (2011) “New Frontiers in Business Intelligence” The 37th International Conference on Very Large Data Bases, Seattle, Washington, Vol. 4, No. 12, VLDB.

[43] Consumer Packaged Goods Company Multi-Model Study, (1998) “Data Mining Case Study: Retail”.

[44] IBM Software Group Case Study. (2010) “Great Canadian Gaming Corporation Leverages IBM Cognos 8: Solutions for Financial Consolidation and Reporting Standardization”.

[45] A. Vellidoa, P.J.G. Lisboaa, J. Vaughan, (1999) “Neural Networks in Business: a Survey of Applications (1992–1998)” Expert Systems with Applications 17, pp. 51–70, Elsevier Science.

[46] Injazz J. Chen, K. Popovich, (2003) "Understanding Customer Relationship Management (CRM): People, process and technology", Business Process Management Journal, Vol. 9, pp.672 – 688.

[47] Dave Smith (2010) “Using Data and Text Mining to Drive Innovation” PhUSE 2010, UK.

[48] Dien D. Phan, Douglas R. Vogel, (2010) “A Model of Customer Relationship Management and Business Intelligence Systems for Catalogue and Online Retailers”, Information & Management, Vol. 47, Issue 2, Pages 69-77.

[49] Christian Thomsen, Torben Bach Pedersen (2009)“A Survey of Open Source Tools for Business Intelligence” International Journal of Data Warehousing and Mining, Vol. 5, Issue 3, IGI Global.

[50] Meryem Duygun Fethi, Fotios Pasiouras (2010) “Assessing Bank Efficiency and Performance with Operational Research and Artificial Intelligence Techniques: A survey” European Journal of Operational Research, pp. 189–198, Elsevier.

[51] Rafael Berlanga, Oscar Romero, Alkis Simitsis, Victoria Nebot, Torben Bach Pedersen, Alberto Abelló, María José Aramburu (2012) “Semantic Web Technologies for Business Intelligence” IGI.

[52] Manuel Mejía-Lavalle, Ricardo Sosa R., Nemorio González M., and Liliana Argotte R. (2009) “Survey of Business Intelligence for Energy Markets” E. Corchado et al. (Eds.): HAIS, LNAI 5572, pp. 235–243, Springer-Verlag Berlin Heidelberg.

[53] Shantanu Godbole, Shourya Roy, (2008) “Text Classification, Business Intelligence, and Interactivity: Automating C-Sat Analysis for Services Industry” KDD’08, ACM Las Vegas, USA.

[54] Carlos Rodríguez, Florian Daniel, F. Casati, Cinzia Cappiello (2010) “Toward Uncertain Business Intelligence: The Case of Key Indicators” Internet Computing, IEEE, vol.14, no.4, pp.32-40.

[55] K.A. Taipale (2003) "Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data" Columbia Science and Technology Law Review 5.

[56] Will Hedfield (2009) “Case study: Jaeger uses data mining to reduce losses from crime and waste”.

[57] K. Laundon and J. Laundon (2011) “Foundations of Business Intelligence: Databases and Information Management” Managing Information Systems: Managing the Digital Firm, Pearson Education Inc.

[58] Oksana Grabova, Jerome Darmont, Jean-Hugues Chauchat, Iryna Zolotaryova (2010) “Business Intelligence for Small and Middle-Sized Enterprises” SIGMOD Rec. 39.

[59] Byung-Kwon Park and Il-Yeol Song (2011) “Toward total business intelligence incorporating structured and unstructured data” In Proceedings of the 2nd International Workshop on Business intelligencE and the WEB (BEWEB '11), ACM, NY, USA.

[60] Y. Li, M.R. Kramer, A.J.M. Beulens, J.G.A.J. van der Vorst (2010) “A Framework for Early Warning and Proactive Control Systems in Food Supply Chain Networks” Computers in Industry, Vol. 61, Issue 9, pp. 852-862.

[61] MAIA Intelligence (2009) “Business Intelligence in Manufacturing”.

[62] Srinivasa Rao P, Saurabh Swarup (2001) “Business Intelligence and Logistics” Wipro Technologies.

[63] Atos, (2011) “Business Intelligence solutions: Decisions that are Better-Informed Leading to Long-Term Competitive Advantage”.

[64] K. Laundon and J. Laundon (2012) “Enhancing Decision Making” Managing Information Systems: Managing the Digital Firm, Pearson Education, Pearson Hall.

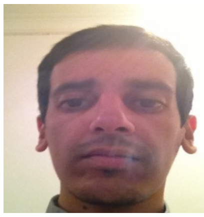
[65] INSEAD, World Economic Forum (2009) “The Global Information Technology Report 2008–2009: Mobility in a Networked World”, Geneva.

[66] Aura-Mihaela Mocanu, Daniela Litan, Stefan Olaru, A. Munteanu (2010) “Information Systems in the Knowledge Based Economy” WSEAS Transactions on Business and Economics, Issue 1, Vol. 7

[67] A. S. Al- Mudimigh, F. Saleem, Z. Ullah, F. N. Al-Aboud (2009) “Implementation of Data Mining Engine on CRM -Improve Customer Satisfaction” International Conference on Information and Communication Technologies ICICT '09, vol., no., pp.193-197.

[68] Case study by Zap Technology, (2010) “KFC/Pizza Hut makes efficiency gains with Zap Business Intelligence: Businesses become more agile, responsive and performance-focused”.

Author : Abdulaziz R. Alazemi



A. R. Alazemi is a graduate researcher interested in the fields of personal information, privacy, networks, and data mining. He graduated from Kuwait University in 2009. He published conference and journal papers regarding PII, data mining and visualization, and distributed systems algorithms.

PREDICTING STUDENTS' PERFORMANCE USING ID3 AND C4.5 CLASSIFICATION ALGORITHMS

Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao

Department of Computer Engineering,
Fr. C.R.I.T., Navi Mumbai, Maharashtra, India

ABSTRACT

An educational institution needs to have an approximate prior knowledge of enrolled students to predict their performance in future academics. This helps them to identify promising students and also provides them an opportunity to pay attention to and improve those who would probably get lower grades. As a solution, we have developed a system which can predict the performance of students from their previous performances using concepts of data mining techniques under Classification. We have analyzed the data set containing information about students, such as gender, marks scored in the board examinations of classes X and XII, marks and rank in entrance examinations and results in first year of the previous batch of students. By applying the ID3 (Iterative Dichotomiser 3) and C4.5 classification algorithms on this data, we have predicted the general and individual performance of freshly admitted students in future examinations.

KEYWORDS

Classification, C4.5, Data Mining, Educational Research, ID3, Predicting Performance

1. INTRODUCTION

Every year, educational institutes admit students under various courses from different locations, educational background and with varying merit scores in entrance examinations. Moreover, schools and junior colleges may be affiliated to different boards, each board having different subjects in their curricula and also different level of depths in their subjects. Analyzing the past performance of admitted students would provide a better perspective of the probable academic performance of students in the future. This can very well be achieved using the concepts of data mining.

For this purpose, we have analysed the data of students enrolled in first year of engineering. This data was obtained from the information provided by the admitted students to the institute. It includes their full name, gender, application ID, scores in board examinations of classes X and XII, scores in entrance examinations, category and admission type. We then applied the ID3 and C4.5 algorithms after pruning the dataset to predict the results of these students in their first semester as precisely as possible.

DOI : 10.5121/ijdkp.2013.3504

2. LITERATURE SURVEY

2.1. Data Mining

Data mining is the process of discovering interesting knowledge, such as associations, patterns, changes, significant structures and anomalies, from large amounts of data stored in databases or data warehouses or other information repositories [1]. It has been widely used in recent years due to the availability of huge amounts of data in electronic form, and there is a need for turning such data into useful information and knowledge for large applications. These applications are found in fields such as Artificial Intelligence, Machine Learning, Market Analysis, Statistics and Database Systems, Business Management and Decision Support [2].

2.1.1. Classification

Classification is a data mining technique that maps data into predefined groups or classes. It is a supervised learning method which requires labelled training data to generate rules for classifying test data into predetermined groups or classes [2]. It is a two-phase process. The first phase is the learning phase, where the training data is analyzed and classification rules are generated. The next phase is the classification, where test data is classified into classes according to the generated rules. Since classification algorithms require that classes be defined based on data attribute values, we had created an attribute "class" for every student, which can have a value of either "Pass" or "Fail".

2.1.2. Clustering

Clustering is the process of grouping a set of elements in such a way that the elements in the same group or cluster are more similar to each other than to those in other groups or clusters [1]. It is a common technique for statistical data analysis used in the fields of pattern recognition, information retrieval, bioinformatics, machine learning and image analysis. Clustering can be achieved by various algorithms that differ about the similarities required between elements of a cluster and how to find the elements of the clusters efficiently. Most algorithms used for clustering try to create clusters with small distances among the cluster elements, intervals, dense areas of the data space or particular statistical distributions.

2.2. Selecting Classification over Clustering

In clustering, classes are unknown apriori and are discovered from the data. Since our goal is to predict students' performance into either of the predefined classes - "Pass" and "Fail", clustering is not a suitable choice and so we have used classification algorithms instead of clustering algorithms.

2.3. Issues Regarding Classification

2.3.1. Missing Data

Missing data values cause problems during both the training phase and to the classification process itself. For example, the reason for non-availability of data may be due to [2]:

- Equipment malfunction
- Deletion due to inconsistency with other recorded data

- Non-entry of data due to misunderstanding
- Certain data considered unimportant at the time of entry
- No registration of data or its change

This missing data can be handled using following approaches [3]:

- Data miners can ignore the missing data
- Data miners can replace all missing values with a single global constant
- Data miners can replace a missing value with its feature mean for the given class
- Data miners and domain experts, together, can manually examine samples with missing values and enter a reasonable, probable or expected value

In our case, the chances of getting missing values in the training data are very less. The training data is to be retrieved from the admission records of a particular institute and the attributes considered for the input of classification process are mandatory for each student. The tuple which is found to have missing value for any attribute will be ignored from training set as the missing values cannot be predicted or set to some default value. Considering low chances of the occurrence of missing data, ignoring missing data will not affect the accuracy adversely.

2.3.2. Measuring Accuracy

Determining which data mining technique is best depends on the interpretation of the problem by users. Usually, the performance of algorithms is examined by evaluating the accuracy of the result. Classification accuracy is calculated by determining the percentage of tuples placed in the correct class. At the same time there may be a cost associated with an incorrect assignment to the wrong class which can be ignored.

2.4. ID3 Algorithm

In decision tree learning, ID3 (Iterative Dichotomiser 3) is an algorithm invented by Ross Quinlan used to generate a decision tree from the dataset. ID3 is typically used in the machine learning and natural language processing domains. The decision tree technique involves constructing a tree to model the classification process. Once a tree is built, it is applied to each tuple in the database and results in classification for that tuple. The following issues are faced by most decision tree algorithms [2]:

- Choosing splitting attributes
- Ordering of splitting attributes
- Number of splits to take
- Balance of tree structure and pruning
- Stopping criteria

The ID3 algorithm is a classification algorithm based on Information Entropy, its basic idea is that all examples are mapped to different categories according to different values of the condition attribute set; its core is to determine the best classification attribute form condition attribute sets. The algorithm chooses information gain as attribute selection criteria; usually the attribute that has the highest information gain is selected as the splitting attribute of current node, in order to make information entropy that the divided subsets need smallest [4]. According to the different values of the attribute, branches can be established, and the process above is recursively called on each branch to create other nodes and branches until all the samples in a branch belong to the same category. To select the splitting attributes, the concepts of Entropy and Information Gain are used.

2.4.1. Entropy

Given probabilities $p_1, p_2, \ldots, p_s$, where $\sum p_i = 1$, Entropy is defined as

$$H(p_1, p_2, \ldots, p_s) = \sum - (p_i \log p_i)$$

Entropy finds the amount of order in a given database state. A value of H = 0 identifies a perfectly classified set. In other words, the higher the entropy, the higher the potential to improve the classification process.

2.4.2. Information Gain

ID3 chooses the splitting attribute with the highest gain in information, where gain is defined as difference between how much information is needed after the split. This is calculated by determining the differences between the entropies of the original dataset and the weighted sum of the entropies from each of the subdivided datasets. The formula used for this purpose is:

$$G(D, S) = H(D) - \sum P(D_i)H(D_i)$$

2.5. C4.5

C4.5 is a well-known algorithm used to generate a decision trees. It is an extension of the ID3 algorithm used to overcome its disadvantages. The decision trees generated by the C4.5 algorithm can be used for classification, and for this reason, C4.5 is also referred to as a statistical classifier. The C4.5 algorithm made a number of changes to improve ID3 algorithm [2]. Some of these are:

- Handling training data with missing values of attributes
- Handling differing cost attributes
- Pruning the decision tree after its creation
- Handling attributes with discrete and continuous values

Let the training data be a set $S = s_1, s_2 \ldots$ of already classified samples. Each sample $S_i = x_l, x_2 \ldots$ is a vector where $x_l, x_2 \ldots$ represent attributes or features of the sample. The training data is a vector $C = c_1, c_2 \ldots$, where $c_1, c_2 \ldots$ represent the class to which each sample belongs to.

At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits data set of samples S into subsets that can be one class or the other [5]. It is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute factor with the highest normalized information gain is considered to make the decision. The C4.5 algorithm then continues on the smaller sub-lists having next highest normalized information gain.

3. TECHNOLOGIES USED

3.1. HTML and CSS

HyperText Markup Language (HTML) is a markup language for creating web pages or other information to display in a web browser. HTML allows images and objects to be included and that can be used to create interactive forms. From this, structured documents are created by using structural semantics for text such as headings, links, lists, paragraphs, quotes etc.

CSS (Cascading Style Sheets) is designed to enable the separation between document content (in HTML or similar markup languages) and document presentation. This technique is used to improve content accessibility also to provide more flexibility and control in the specification of content and presentation characteristics. This enables multiple pages to share formatting and reduce redundancies.

3.2. PHP and the CodeIgniter Framework

PHP (recursive acronym for PHP: Hypertext Preprocessor) is a widely-used open source general-purpose server side scripting language that is especially suited for web development and can be embedded into HTML.

CodeIgniter is a well-known open source web application framework used for building dynamic web applications in PHP [6]. Its goal is to enable developers to develop projects quickly by providing a rich set of libraries and functionalities for commonly used tasks with a simple interface and logical structure for accessing these libraries. CodeIgniter is loosely based on the Model-View-Controller (MVC) pattern and we have used it to build the front end of our implementation.

3.3. MySQL

MySQL is the most popular open source RDBMS which is supported, distributed and developed by Oracle [8]. In the implementation of our web application, we have used it to store user information and students' data.

3.4. RapidMiner

RapidMiner is an open source data mining tool that provides data mining and machine learning procedures including data loading and transformation, data preprocessing and visualization, modelling, evaluation, and deployment [7]. It is written in the Java programming language and makes use of learning schemes and attribute evaluators from the WEKA machine learning environment and statistical modelling schemes for the R-Project. We have used RapidMiner to generate decision trees of ID3 and C4.5 algorithms.

4. IMPLEMENTATION

We had divided the entire implementation into five stages. In the first stage, information about students who have been admitted to the second year was collected. This included the details submitted to the college at the time of enrolment. In the second stage, extraneous information was removed from the collected data and the relevant information was fed into a database. The third stage involved applying the ID3 and C4.5 algorithms on the training data to obtain decision trees of both the algorithms. In the next stage, the test data, i.e. information about students currently enrolled in the first year, was applied to the decision trees. The final stage consisted of developing the front end in the form of a web application.

These stages of implementation are depicted in Figure 1.

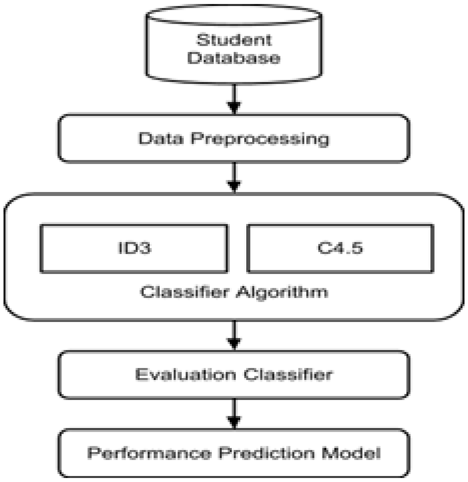


Figure 1. Processing model

4.1. Student Database

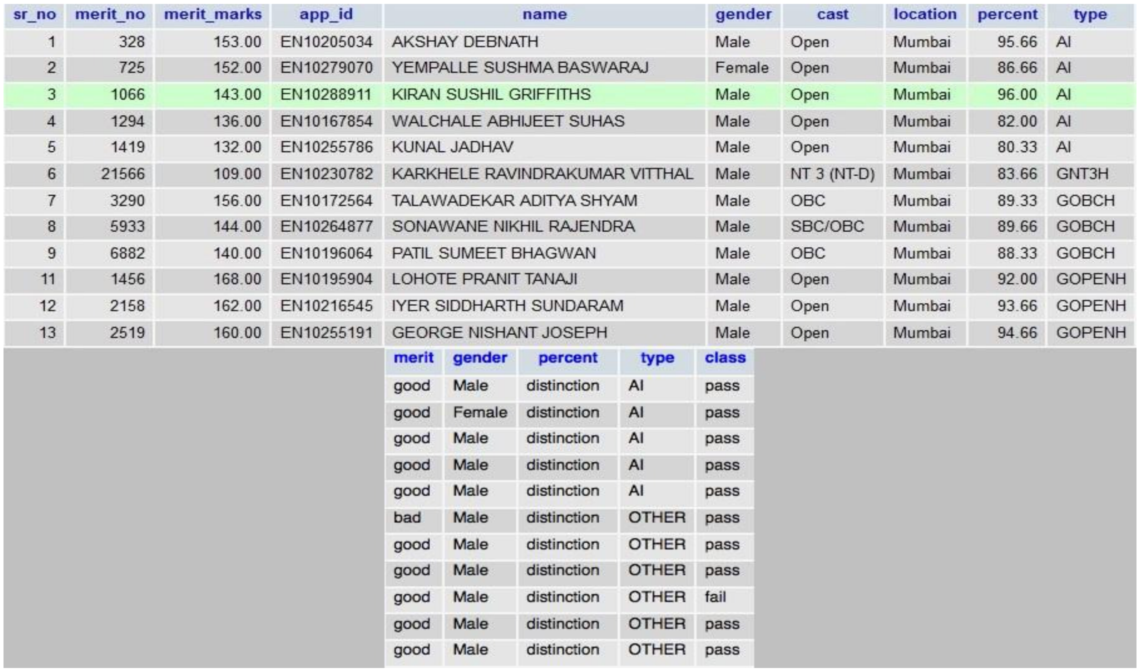
We were provided with a training dataset consisting of information about students admitted to the first year. This data was in the form of a Microsoft Excel 2003 spreadsheet and had details of each student such as full name, application ID, gender, caste, percentage of marks obtained in board examinations of classes X and XII, percentage of marks obtained in Physics, Chemistry and Mathematics in class XII, marks obtained in the entrance examination, admission type, etc. For ease of performing data mining operations, the data was filled into a MySQL database.

4.2. Data Preprocessing

Once we had details of all the students, we then segmented the training dataset further, considering various feasible splitting attributes, i.e. the attributes which would have a higher impact on the performance of a student. For instance, we had considered ‘location’ as a splitting attribute, and then segmented the data according to students’ locality.

A snapshot of the student database is shown in Figure 2. Here, irrelevant attributes such as students residential address, name, application ID, etc. had been removed. For example, the admission date of the student was irrelevant in predicting the future performance of the student. The attributes that had been retained are those for merit score or marks scored in entrance examination, gender, percentage of marks scored in Physics, Chemistry and Mathematics in the board examination of class XII and admission type. Finally, the “class” attribute was added and it held the predicted result, which can be either “Pass” or “Fail”.

Since the attributes for marks would have discrete values, to produce better results, specific classes were defined. Thus, the "merit" attribute had a value "good" if the merit score of the student was 120 or above out of a maximum score of 200, and was classified as "bad" if the merit score was below 120. Also, the value that can be held by the "percentage" attribute of the student are three - "distinction" if the percentage of marks scored by the student in the subjects of Physics, Chemistry and Mathematics was 70 or above, "first_class" if the percentage was less than 70 and greater than or equal to 60, then it was classified as "second_class" if the percentage was less than 60. The attribute for admission type is labelled "type" and the value held by a student for it can be either "AI" (short for All-India), if the student was admitted to a seat available for All-India candidates, or "OTHER" if the student was admitted to another seat.



sr_no	merit_no	merit_marks	app_id	name	gender	cast	location	percent	type
1	328	153.00	EN10205034	AKSHAY DEBNATH	Male	Open	Mumbai	95.66	AI
2	725	152.00	EN10279070	YEMPALLE SUSHMA BASWARAJ	Female	Open	Mumbai	86.66	AI
3	1066	143.00	EN10288911	KIRAN SUSHIL GRIFFITHS	Male	Open	Mumbai	96.00	AI
4	1294	136.00	EN10167854	WALCHALE ABHIJEET SUHAS	Male	Open	Mumbai	82.00	AI
5	1419	132.00	EN10255786	KUNAL JADHAV	Male	Open	Mumbai	80.33	AI
6	21566	109.00	EN10230782	KARKHELE RAVINDRAKUMAR VITTHAL	Male	NT 3 (NT-D)	Mumbai	83.66	GNT3H
7	3290	156.00	EN10172564	TALAWADEKAR ADITYA SHYAM	Male	OBC	Mumbai	89.33	GOBCH
8	5933	144.00	EN10264877	SONAWANE NIKHIL RAJENDRA	Male	SBC/OBC	Mumbai	89.66	GOBCH
9	6882	140.00	EN10196064	PATIL SUMEET BHAGWAN	Male	OBC	Mumbai	88.33	GOBCH
11	1456	168.00	EN10195904	LOHOTE PRANIT TANAJI	Male	Open	Mumbai	92.00	GOPENH
12	2158	162.00	EN10216545	IYER SIDDHARTH SUNDARAM	Male	Open	Mumbai	93.66	GOPENH
13	2519	160.00	EN10255191	GEORGE NISHANT JOSEPH	Male	Open	Mumbai	94.66	GOPENH

merit	gender	percent	type	class
good	Male	distinction	AI	pass
good	Female	distinction	AI	pass
good	Male	distinction	AI	pass
good	Male	distinction	AI	pass
good	Male	distinction	AI	pass
bad	Male	distinction	OTHER	pass
good	Male	distinction	OTHER	pass
good	Male	distinction	OTHER	pass
good	Male	distinction	OTHER	fail
good	Male	distinction	OTHER	pass
good	Male	distinction	OTHER	pass

Figure 2. Preprocessed student database

4.3. Data Processing Using RapidMiner

The next step was to feed the pruned student database as input to RapidMiner. This helped us in evaluating interesting results by applying classification algorithms on the student training dataset. The results obtained are shown in the following subsections:

4.3.1. ID3 Algorithm

Since ID3 is a decision tree algorithm, we obtained a decision tree as the final result with all the splitting attributes and it is shown in Figure 3.

4.3.2. C4.5 Algorithm

The C4.5 algorithm too generates a decision tree, and we obtained one from RapidMiner in the same way as ID3. This tree, shown in Figure 4, has fewer decision nodes as compared to the tree for improved ID3, which is shown in Figure 3.

4.4. Implementing the Performance Prediction Web Application

RapidMiner helped significantly in finding hidden information from the training dataset. These newly learnt predictive patterns for predicting students' performance were then implemented in a working web application for staff members to use to get the predicted results of admitted students.

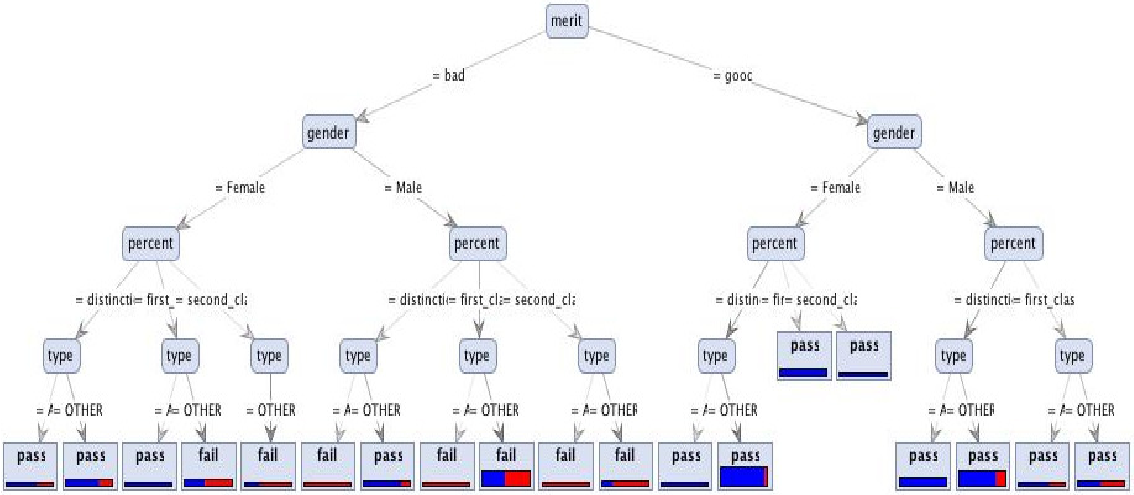


Figure 3. Decision tree for ID3

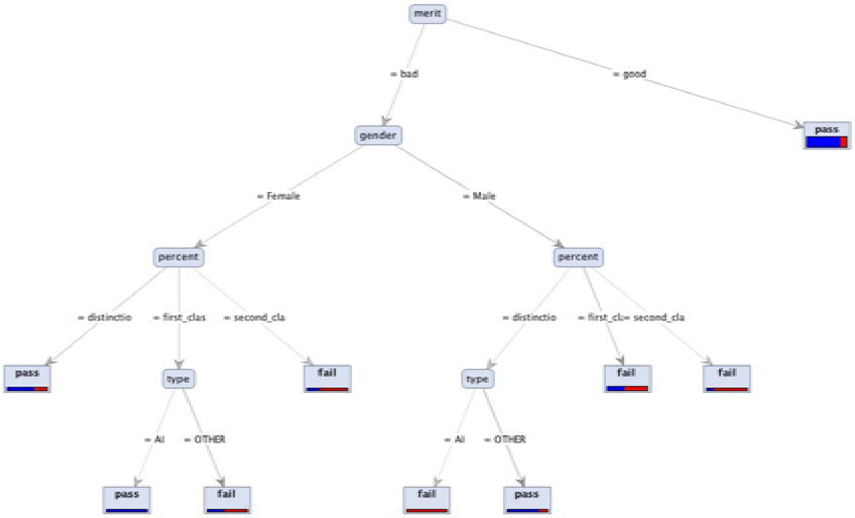


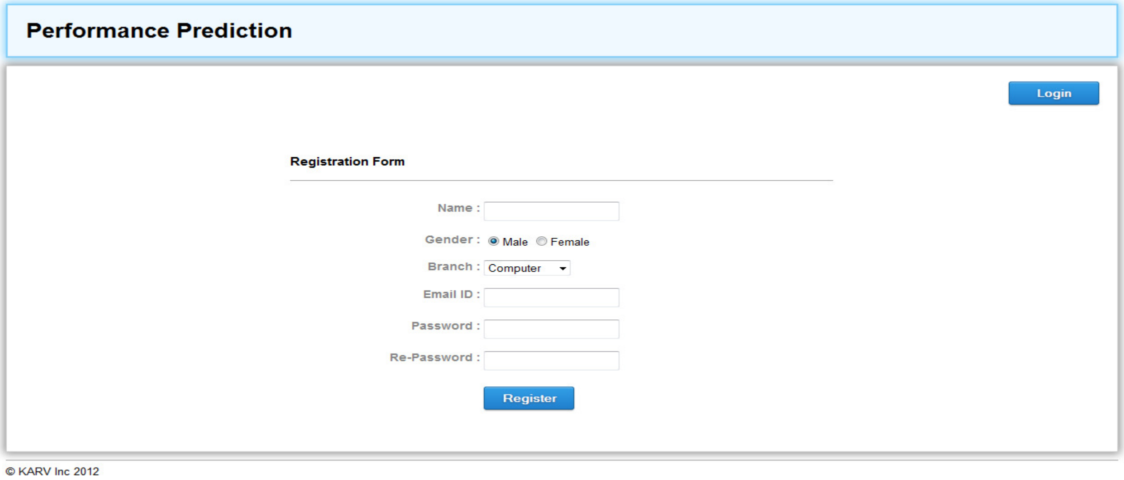
Figure 4. Decision tree for C4.5

4.4.1. CodeIgniter

The web application was developed using a popular PHP framework named CodeIgniter. The application has provisions for multiple simultaneous staff registrations and staff logins. This ensures that the work of no two staff members is interrupted during performance evaluation. Figure 5 and Figure 6 depict the staff registration and staff login pages respectively.

4.4.2. Mapping Decision Trees to PHP

The essence of the web application was to map the results achieved after data processing to code. This was done in form of class methods in PHP. The result of the improved ID3 and C4.5 algorithms were in the form of trees and these were translated to code in the form of if-else ladders. We then placed these ladders into PHP class methods that accept only the splitting attributes - PCM percentage, merit marks, admission type and gender as method parameters. The class methods return the final result of that particular evaluation, indicating whether that student would pass or fail in the first semester examination. Figure 7 shows a class method with the if-else ladder.



Performance Prediction

Registration Form

Name :

Gender : ● Male ○ Female

Branch : Computer

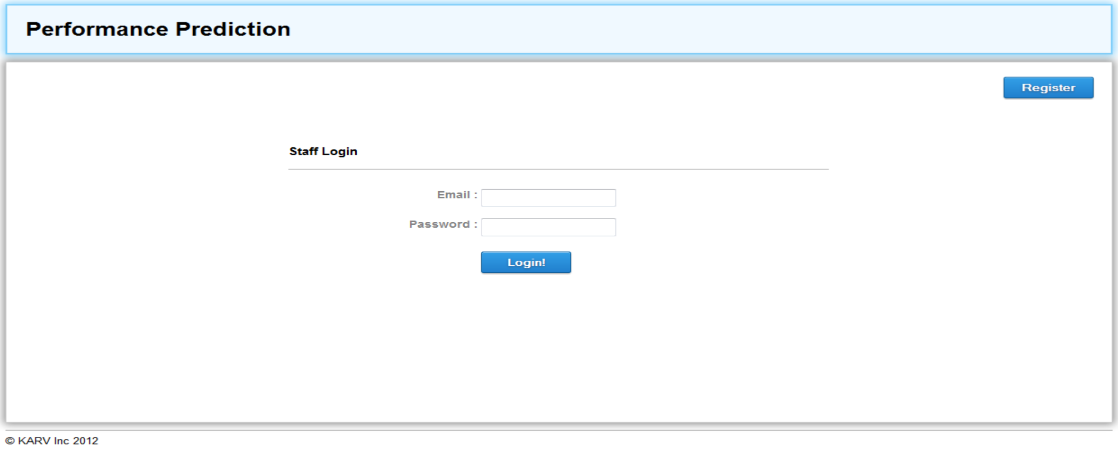
Email ID :

Password :

Re-Password :

© KARV Inc 2012

Figure 5. Registration page for staff members



Performance Prediction

Staff Login

Email :

Password :

© KARV Inc 2012

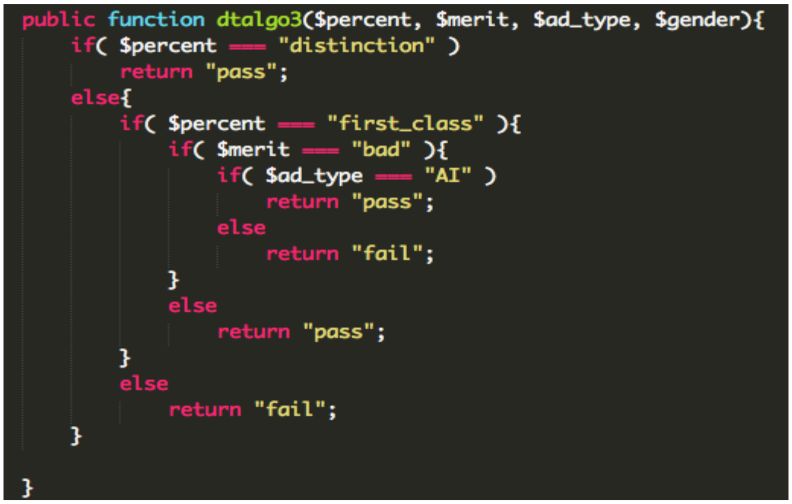
Figure 6. Login page for staff members

4.4.3. Singular Evaluation

Once the decision trees were mapped as class methods, we built a web page for staff members to feed values for the name, application ID and splitting attributes of a student, as can be seen in Figure 8. These values were then used to predict the result of that student as either “Pass” or “Fail”.

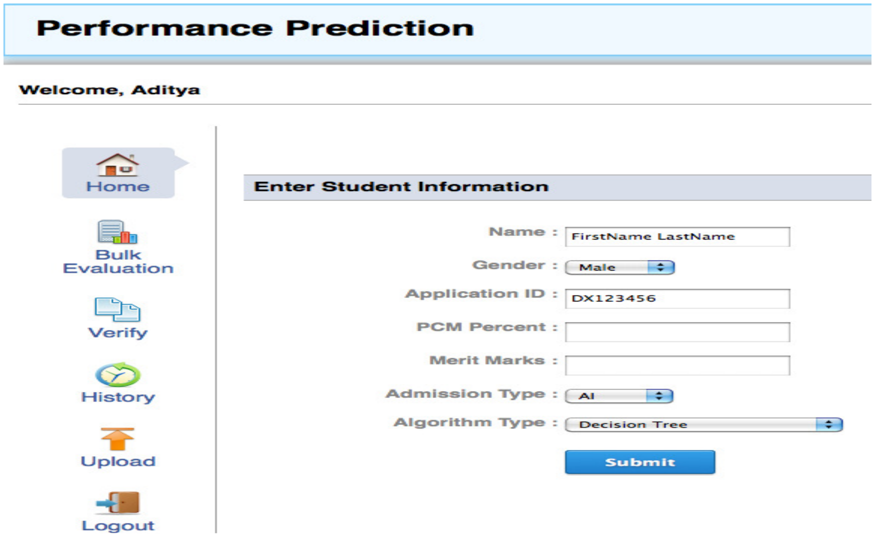
4.4.4. Upload Excel Sheet

Singular Evaluation is beneficial when the results of a small number of students are to be predicted, one at a time. But in case of large testing datasets, it is feasible to upload a data file in a format such as that of a Microsoft Excel spreadsheet, and evaluate each student's record. For this, staff members can upload a spreadsheet containing records of students with attributes in a predetermined order. Figure 9 shows the upload page for Excel spreadsheets.



```
public function dtalgo3($percent, $merit, $ad_type, $gender){
    if( $percent === "distinction" )
        return "pass";
    else{
        if( $percent === "first_class" ){
            if( $merit === "bad" ){
                if( $ad_type === "AI" )
                    return "pass";
                else
                    return "fail";
            }
            else
                return "pass";
        }
        else
            return "fail";
    }

}
```

Figure 7. PHP class method mapping a decision tree



Name : FirstName LastName

Gender : Male

Application ID : DX123456

PCM Percent :

Merit Marks :

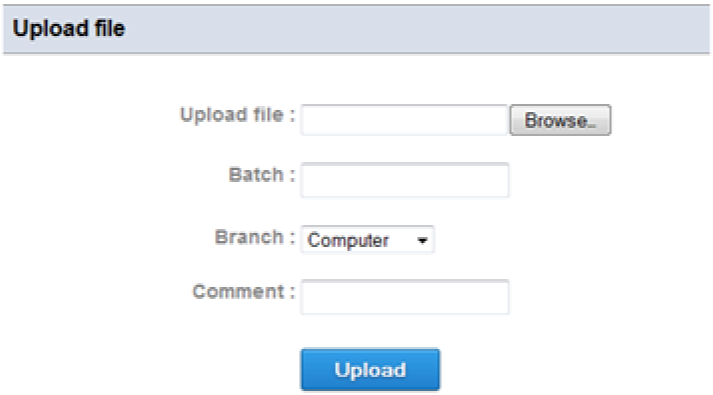
Admission Type : AI

Algorithm Type : Decision Tree

Figure 8. Web page for Singular Evaluation

4.4.5. Bulk Evaluation

Under the Bulk Evaluation tab, a staff member can choose an uploaded dataset to evaluate the results, along with the algorithm to be applied over it. After submitting the dataset and algorithm, the predicted result of each student is displayed in a table as the value of the attribute "class". A sample result of Bulk Evaluation can be seen in Figure 10.



Upload file

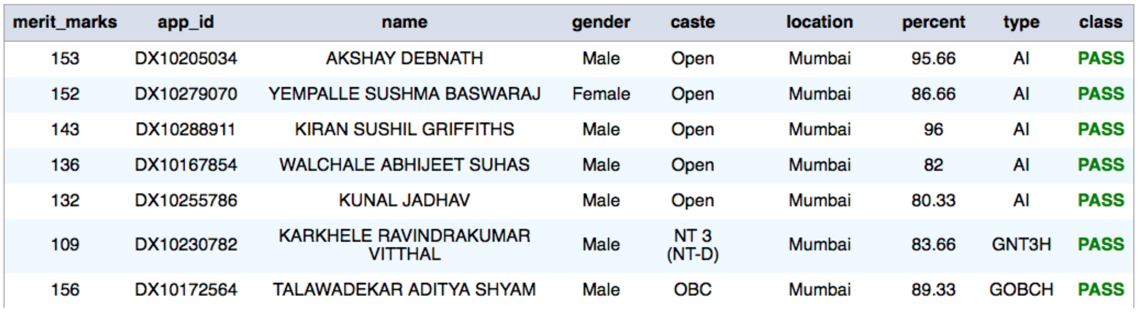
Upload file :

Batch :

Branch : Computer

Comment :

Figure 9. Page to upload Excel spreadsheet

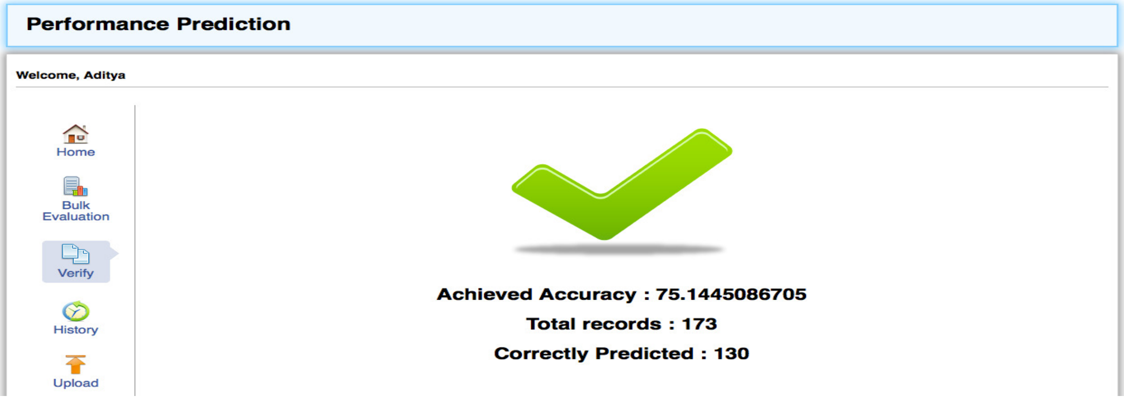


merit_marks	app_id	name	gender	caste	location	percent	type	class
153	DX10205034	AKSHAY DEBNATH	Male	Open	Mumbai	95.66	AI	PASS
152	DX10279070	YEMPALLE SUSHMA BASWARAJ	Female	Open	Mumbai	86.66	AI	PASS
143	DX10288911	KIRAN SUSHIL GRIFFITHS	Male	Open	Mumbai	96	AI	PASS
136	DX10167854	WALCHALE ABHIJEET SUHAS	Male	Open	Mumbai	82	AI	PASS
132	DX10255786	KUNAL JADHAV	Male	Open	Mumbai	80.33	AI	PASS
109	DX10230782	KARKHELE RAVINDRAKUMAR VITTHAL	Male	NT 3 (NT-D)	Mumbai	83.66	GNT3H	PASS
156	DX10172564	TALAWADEKAR ADITYA SHYAM	Male	OBC	Mumbai	89.33	GOBCH	PASS

Figure 10. Page showing results after Bulk Evaluation

4.4.6. Verifying Accuracy of Predicted Results

The accuracy of the algorithm results can be tested under the Verify tab. A staff member has to select the uploaded verification file which already has the actual results and the algorithm that has to be tested for accuracy. After submission the predicted result of evaluation is compared with actual results obtained and the accuracy is calculated. Figure 11 shows that the accuracy achieved is 75.145% for both ID3 and C4.5 algorithms. Figure 12 shows the mismatched tuples, i.e. the tuples which were predicted wrongly by the application for the current test data.



Performance Prediction

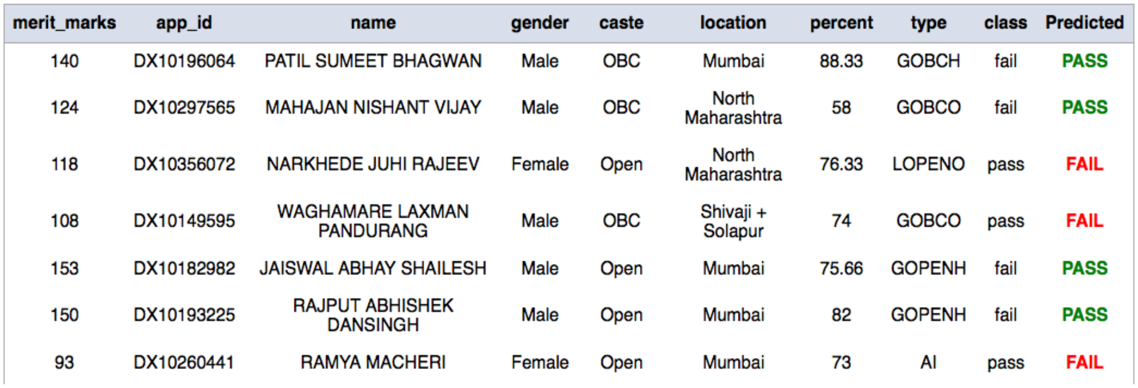
Welcome, Aditya

Achieved Accuracy : 75.1445086705

Total records : 173

Correctly Predicted : 130

Figure 11. Accuracy achieved after evaluation

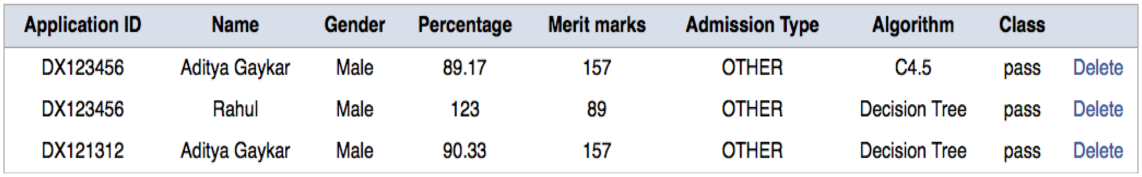


merit_marks	app_id	name	gender	caste	location	percent	type	class	Predicted
140	DX10196064	PATIL SUMEET BHAGWAN	Male	OBC	Mumbai	88.33	GOBCH	fail	PASS
124	DX10297565	MAHAJAN NISHANT VIJAY	Male	OBC	North Maharashtra	58	GOBCO	fail	PASS
118	DX10356072	NARKHEDE JUHI RAJEEV	Female	Open	North Maharashtra	76.33	LOPENO	pass	FAIL
108	DX10149595	WAGHAMARE LAXMAN PANDURANG	Male	OBC	Shivaji + Solapur	74	GOBCO	pass	FAIL
153	DX10182982	JAISWAL ABHAY SHAILESH	Male	Open	Mumbai	75.66	GOPENH	fail	PASS
150	DX10193225	RAJPUT ABHISHEK DANSINGH	Male	Open	Mumbai	82	GOPENH	fail	PASS
93	DX10260441	RAMYA MACHERI	Female	Open	Mumbai	73	AI	pass	FAIL

Figure 12. Mismatched tuples shown during verification

4.4.7. Singular Evaluation History

Using the web interface, staff members can view all Singular Evaluations they had conducted in the past. This is displayed in the form of a table, containing attributes of the student and the predicted result. If required, a record from this table may be deleted by a staff member. A snapshot of this table is shown in Figure 13.



Application ID	Name	Gender	Percentage	Merit marks	Admission Type	Algorithm	Class	
DX123456	Aditya Gaykar	Male	89.17	157	OTHER	C4.5	pass	Delete
DX123456	Rahul	Male	123	89	OTHER	Decision Tree	pass	Delete
DX121312	Aditya Gaykar	Male	90.33	157	OTHER	Decision Tree	pass	Delete

Figure 13. History of Singular Evaluations performed by staff members

5. FUTURE WORK

In this project, prediction parameters such as the decision trees generated using RapidMiner are not updated dynamically within the source code. In the future, we plan to make the entire implementation dynamic to train the prediction parameters itself when new training sets are fed into the web application. Also, in the current implementation, we have not considered extra-curricular activities and other vocational courses completed by students, which we believe may have a significant impact on the overall performance of the students. Considering such parameters would result in better accuracy of prediction.

6. CONCLUSIONS

In this paper, we have explained the system we have used to predict the results of students currently in the first year of engineering, based on the results obtained by students currently in the second year of engineering during their first year.

The results of Bulk Evaluation are shown in Table 1. Random test cases considered during individual testing resulted in approximately equal accuracy, as indicated in Table 2.

Table 1. Results of Bulk Evaluation

Algorithm	Total Students	Students whose results are correctly predicted	Accuracy (%)	Execution Time (in milliseconds)
ID3	173	130	75.145	47.6
C4.5	173	130	75.145	39.1

Table 2. Results of Singular Evaluation.

Algorithm	Total Students	Students whose results are correctly predicted	Accuracy (%)
ID3	9	7	77.778
C4.5	9	7	77.778

Thus, for a total of 182 students, the average percentage of accuracy achieved in Bulk and Singular Evaluations is approximately 75.275.

ACKNOWLEDGEMENTS

We express sincere gratitude to our project guides, Ms M. Kiruthika, and Ms Smita Dange for their guidance and support.

REFERENCES

[1] Han, J. and Kamber, M., (2006) *Data Mining: Concepts and Techniques*, Elsevier.

[2] Dunham, M.H., (2003) *Data Mining: Introductory and Advanced Topics*, Pearson Education Inc.

[3] Kantardzic, M., (2011) *Data Mining: Concepts, Models, Methods and Algorithms*, Wiley-IEEE Press.

[4] Ming, H., Wenying, N. and Xu, L., (2009) "An improved decision tree classification algorithm based on ID3 and the application in score analysis", Chinese Control and Decision Conference (CCDC), pp1876-1879.

[5] Xiaoliang, Z., Jian, W., Hongcan Y., and Shangzhuo, W., (2009) "Research and Application of the improved Algorithm C4.5 on Decision Tree", International Conference on Test and Measurement (ICTM), Vol. 2, pp184-187.

[6] CodeIgnitor User Guide Version 2.14, http://ellislab.com/codeigniter/user-guide/toc.html

[7] RapidMiner, http://rapid-i.com/content/view/181/190/

[8] MySQL – The world's most popular open source database, http://www.mysql.com/