



UNIVERSITY
OF SKÖVDE

**PERFORMANCE EVALUATION of
MILITARY TRAINING EXERCISES USING
DATA MINING**

**Master Degree Project in Informatics with a
Specialization in Data Science**

**One year Level 15 ECTS Spring
term 2016**

Rohini Dubey

Supervisor: Maria Riveiro

Examiner: Göran Falkman

DECLARATION

It is certified that the dissertation, **“PERFORMANCE EVALUATION OF MILITARY TRAINING EXERCISE USING DATA MINING”** carries my original research work and has not been submitted either in the part or full for any other degree or diploma in this or any other University. It is also certified that all the ideas and techniques of different researchers have been properly referenced and duly acknowledged in the dissertation.

Rohini Dubey

(a15rohdu)

Abstract

Attaining training objectives is the measure of a successful training as objectives defines the purpose of instructional events. Application of the training objectives is challenging in large and complex military trainings. The trainings in military domain not only focus on the completion of the trainings but effectively achieving the objectives of the training is the goal of the exercises. It has been realized that the performance to achieve the goal is strengthened by the instructional processes and materials which are crafted to address specific training objectives. Simulation is one of the effective and realistic learning tools which can be used in trainings. As it is known that simulation generates enormous data, analysis of this data which may contain hidden information is a challenging task. The use of data mining is a solution to this problem. The aim of this project is to propose a framework of a system for the instructors which can be followed for evaluating trainee's performance so that their fulfillment of the training objectives can be improved. A proposal which is studied in this project is learning from previous training experiences using data mining techniques to improve the effectiveness of the training by predicting the performance of the trainee. For selecting the good prediction model to estimate the learning outcome of the trainees, different classification techniques have been compared. CRISP-DM model is considered as a base for proposing the framework in this dissertation. Proposed framework is then applied on the dataset obtained from the Swedish Military for the exercises which involved shooting the target.

Keywords: Simulation, framework, Data Mining, Classification, CRISP- DM.

Table of Contents

1. Introduction.....	1
1.1 Organization of thesis.....	3
2. Background.....	4
2.1 Training Exercise.....	5
2.2 Data Mining.....	6
2.3 Visual Analytics.....	7
2.4 Related Work.....	8
3. Problem Formulation.....	11
3.1 Motivation.....	11
3.2 Objective of the study.....	12
4. Method.....	14
4.1 CRISP DM model.....	12
4.2 Proposed Framework.....	15
4.3 Tools Used for Implementation.....	20
4.3.1 WEKA.....	20
4.3.2 Tableau.....	20
4.4 Data Mining Techniques Used for Implementation.....	21
4.4.1 Feature Selection.....	21
4.4.2 Logistic Regression.....	22
4.4.3 Decision Tree.....	23
4.4.4 Artificial Neural Network.....	24
5. Implementation of the Framework.....	26
5.1 Business Understanding.....	26
5.2 Data Understanding.....	26
5.2.1 Analysis of Data.....	27
5.3 Data Preparation.....	30
5.3.1 Dimensionality Reduction.....	30
5.4 Modeling Technique Selection.....	32
5.5 Building the Model.....	33
5.5.1 Experimental Setup.....	33
5.5 Model Evaluation.....	36
5.6 Model Deployment.....	37
6. Discussion.....	38
7. Conclusion.....	41
7.1 Contribution.....	42
7.2 Future Work.....	43

8. References.....	44
9. Appendix.....	49

List of figures

1. Screen shot of simulation based shooting exercise.....	2
2. The Crisp DM process Model.....	19
3. Structure of single neuron.....	24
4. Trainees Performance in terms of hit counts.....	28
5. Ammunition and Hit Rate relation.....	28
6. Vehicle orientation per Player.....	29
7. Output of filter method of feature selection.....	31
8. Output of Logistic Regression technique.....	34
9. Output of Neural Network technique.....	35
10. Output of Decision Tree technique.....	35
11. System Framework.....	37

CHAPTER 1

1. Introduction

The environment in which the homeland and the coalition forces operate is comprised of stressful, complex and ambiguous environment [Laurence & Mathews, 2012; Salas, Priest, Wilson & Burke, 2006]. These kinds of situations help improve soldier's adaptability to tackle difficult situations and be decisive in resolving intricate situations in the war time [Andrews & Fitzgerald, 2010]. "The ability to augment, replace, create, and /or manage trainee's actual experience with the world on the provision of realistic content and embedded instructional features," is as Simulation Based training [CannonBowers & Bowers, 2009; Wang et al., 2015]. Simulated trainings are best achieved in military domain because these trainings provide safe, effective and efficient training, which has economical and practical advantage over traditional methods [Walcutt et al., 2013]. Simulation based trainings usually collect huge amount of data and thus applying data mining techniques can help the instructors to provide the immediate feedback based on the patterns observed and knowledge gained from the data. For the knowledge discovery and for the analysis of the enormous historical data, data mining has become a striking and innovative tool. Various organizations, which are in the field of business, science, engineering, sports etc., have accomplished a considerable success and lucrative results through implementing data mining. That is why, military has spent huge amount of resources for the development of replicable and generalizable training systems [Walcutt et al., 2013] and many companies have invested enormous amount of time and money to develop virtual reality programs and support systems [Walcutt et al., 2013]. Efficiency and effectiveness of the embedded instructional strategies and also the cognitive capabilities and the limitations of the trainees haven't been considered while designing many of these trainings [Walcutt et al., 2013] because of which they are conceptualized as practice platform rather than training devices [Nicholson et al., 2007].

The absence of the feedback and the presence of the inappropriate feedback are strong impediments to learning [Klueger & DeNisi, 1998]. Thus, feedback provides the knowledge of

results and serves as a source of motivation [Day et al., 2006]. There is overwhelming evidence that the direct instructional support is necessary for the optimal training in case of the novice trainees [Mayer, 2004; Walcutt et al., 2013]. The analysis of musicians, athletes and other professionals show that expertise is primarily a function of practice and feedback that are deliberately designed to strengthen weak skills [Ericsson & Lehmann, 1996; Stacy & Freeman, 2016; Riveiro et al., 2016].



Fig 1. Screen shot of simulation based shooting exercise

Data mining can contribute to the problem of finding knowledge from data. It can be used to extract, previously unknown, knowledge or hidden information from the data collected which can be potentially interesting and useful. This dissertation develops a data mining approach to enhance the evaluation of the training. A framework is proposed for improving the resulting outcome of the trainee for their exercise by applying data mining techniques. For this dissertation work, exercises which involve shooting in the simulated environment by means of ammunition are considered. Proposed framework is based on the CRISP DM model as it is broadly applicable for wide array of projects which involve data mining. Since the success of any project which makes use of data mining relies upon how good the model is, thus, this dissertation recommends that selection of a modeling techniques before it is exercised on data should be acknowledged as an important step. Therefore, CRSIP DM should be revised to include “selection of modeling technique” as a separate phase so that more focus can be given to this activity.

1.1 Organization of thesis

The chapters of the thesis are organized as follows:

CHAPTER 1: This chapter provides introduction to the problem and the basic information as to why it is important to implement data mining techniques on data collected from simulation based trainings and the importance of feedback in military trainings.

CHAPTER 2: This chapter begins by explaining the concept of simulation based trainings, data mining and visualization; and their importance. It also talks about the training exercise considered in this dissertation and the previous work done in the similar area.

CHAPTER 3: This chapter begins by discussing the purpose of the thesis work. It then clearly describes the objectives of the dissertation. It ends by explaining the motivation which led to work on this project.

CHAPTER 4: This chapter provides discusses the proposed framework which can be applied on the data, tools and techniques used in the research along with some important concepts.

CHAPTER 5: In this chapter the proposed methodology has been implemented. This chapter covers all the aspects of the thesis work like data pre-processing, implementation of the data mining techniques, their results and evaluation and how the model has been built.

CHAPTER 6: This chapter discusses the proposed model.

CHAPTER 7: This chapter concludes the thesis by summarizing the result, and proposes possible directions for future work.

CHAPTER 2

2. Background

The use of the simulation based trainings is increasing widespread because of the advancement in the computer technologies which allows more complex and realistic training scenarios that can be simulated. Simulation based training is potentially strong for designing a highly realistic training environment which allows trainees to participate more actively in the training process. Some of the researches have shown that simulation based trainings are not only capable of modifying trainee's behavioral pattern but can also improve their self- efficacy [McGaghie et al., 2010]. Trainees are supposed to act as if they are present in the real situations while performing the training. The evaluation of training is a systematic process of collecting and analyzing data in order to determine whether and to what degree objectives were or are being achieved [Boulmetis & Dutwin, 2000; Harshit Topno, 2012]. The effectiveness of evaluation is the determination of the extent to which a program has met its stated performance goals and objectives [Schalok, 2001; Harshit Topno, 2012]. As said by Spitzer in 1999, "training can be turned into a powerful force for improving the business, for organization as well as people in it". With the aid of error correction, timely feedbacks and examination, simulation based training allows a repeated practice and tends to move towards the excellence. Trainees are helped to achieve the expertise and the necessity of maintaining these skills and behavior patterns [Issenberg et al., 2005]. Once the instructor diagnoses the learning behavior of the trainee, useful feedback can be given to the trainee to improve the learning performance of the trainee. For improving the performance of the trainees, instructors can also rate the trainee's behavior and provide them the feedback based on their rating.

There are many techniques which are proposed for evaluating the performance of the trainees. Data mining is one of the popular techniques for analyzing the performance [A.M. Shahiri & W.Husain, 2005]. Data mining is a good approach to innovate the problem of predicting performance because of its ability to discover hidden relationships and patterns in large amount of data which are helpful in decision making. Data mining uses algorithms to find out the knowledge behind the data. Two approaches can be used to discover this knowledge: supervised

or unsupervised learning. If the available data already contains an output or target variable (which a model is supposed to predict) then supervised learning can be applied otherwise unsupervised learning should be opted. The useful information and patterns can be used to predict the trainee's performance. Therefore, it would assist trainers in providing the effective training approaches. Considered training approach tries to emulate the different shooting scenarios in which a trainee tries to fire a shot with the use of different ammunitions in order to strike the target. As it has been considered that the main goal of the training exercise for this dissertation is to hit the target with the use of the ammunition, thus, statistics of the trainee will serve as input and whether the trainee will hit or miss the target will be the output in case of the supervised learning approach. Thus, the problem of enhancing the effectiveness of the training is converted to a data mining problem which is predicting the outcome of the trainee based on his/her previous performance so that proper feedback can be given to him. CRISP DM methodology is a structured approach for providing a blueprint for conducting a data mining project. It was developed by industry leaders in late 1996 by taking inputs from many data mining users and data mining tool/service providers. It encourages best practices and provides structure which is needed for better and faster results from data mining. The framework which is proposed in this dissertation also has a foundation on CRISP DM methodology so that better and faster results can be obtained.

Visualizing the large amount of data using graphs and charts is easier because of the way the human brain processes information. Data visualization is easy and quick way for conveying the concepts and one can experiment with various scenarios. Visual Analysis is the study of visual representation of the data. The aim of the visual analysis is to communicate the information in graphical representation form. It helps decision makers in identifying new patterns by enabling them to view analytics presented visually. Reaching a good level of expertise is not an overnight job. It requires following strategies and building stamina over time Thus, trainers can see how the trainee's performance is evolving with time. This will help trainers to develop strategies for improved performance or see the effect of a training strategy on training personal. Visualization is very useful in analyzing the time series data which can be done in the form of trend lines. Visualization is very helpful in analyzing the patterns and lets decision makers to see how the things have changed over time.

2.1 Training Exercise

Military exercise can be considered as the employment of military personal in trainings for military operations. Military trainings can be either compulsory or voluntary. The primary training is the basic training colloquially called boot camp. It attempts to teach the basic training and information which is necessary to become an effective member. Therefore, service members are drilled technically, physically and psychologically. There are variety of military trainings exists these days. They include field exercises which is considered as the practice of warfare; command post exercise which has a focus on war readiness of staff; simulation based trainings which allows the imitation of a real world process, also called as virtual battlefields and joint exercises which involves training of the different armed forces together. The exercise considered in this dissertation is simulation based trainings for the military. Thus, the exercise has taken place in the virtually simulated environment. The exercise consists of shooting the target using the ammunitions in the computer controlled environment. The trainees were provided with gadgets for their trainings and were supposed to shoot the target in a simulated environment until he/she gets successful in hitting the target. The trainee can be either static or moving while shooting the target. The instructor can vary the conditions like ammunition type, target type, distance and other parameters. The performance is computed based on how many hits a trainee takes before he/she actually shoot the target and consistency in shooting the target. Thus, the trainee will exhibit low performance if he/she takes more number of shots to hit the target or if the performance is not steady.

2.2 Data Mining

The aim of the data mining is to derive meaningful patterns and rules from the dataset so that they can be transformed into an understandable structure which can be used further [Han & Kamber, 2006; Witten, Frank & Hall, 2011]. Data mining is also known as Knowledge discovery in databases. Data mining tools predict the future behaviors and trends which can be used to make proactive, knowledge driven decisions. Data mining is able to tell the important things like what is going to happen next in data through the technique known as Modeling. Modeling is a process of building a model on the data collected from different situations and then apply the

model to the other situations where the answer is not known. Data mining includes various techniques:

- Anomaly detection: the identification of the outliers.
- Association Rule mining: finding relationship between the variables.
- Classification: is the task of building a model to predict the classes of the data object for which the class labels are unknown.
- Clustering: grouping data that are similar in some way without taking help from the known structures in the data.
- Regression: process of finding a function that can be used to model the data with least error.

2.3 Visual Analytics

Visual analytics is the science of analytical reasoning supported by interactive visual interfaces [Thomas & Cook, 2005]. The aim of the visual analytics is to transform the information load problem into an opportunity of analyzing the data containing enormous information so that efficacious actions can be applied in the real life situations. However, on the contrary to the visualization, automated methods are better for analyzing big datasets but they fail sometimes because the user cannot intervene once the process has commenced. Visual Analytics however addresses the deficiencies of the visualization as well as automated methods by bringing them together and exploiting their strengths. Visualization has three main goals:

- a) Presenting the outcome of the analysis efficiently and effective.
- b) Investigating the hypothesis by having a goal oriented approach.
- c) Interactive exploratory data analysis and mostly search for trends is undirected [Keim, Mansmann, Stoffel & Ziegler, 2009].

2.4 Related Work

There is not much work done in implementing data mining techniques for improving the performance of the trainees involved in the simulation based trainings. But lot of work has been done in predicting the outcomes of the performance using data mining approaches and the evaluation of the data mining techniques separately.

Previous work on predicting the performance outcome using data mining

Morbitzer, Strachan & Simpson [2004] has described how the data mining techniques can be used in simulation exercises to improve the analysis of results obtained. They used clustering in improving the analysis of building simulation performance prediction and identified that it is one of the technique which can be implemented.

Fanhui [2013] has discussed the application of the data mining techniques in sports training. He applied neural networks approach to predict the performance of the athletes. The data used for forecasting was comprised of the student's sports score and physical health questionnaire. His experiment showed that the use of neural networks for predicting athlete's performance has good approximation ability.

Kabakchieva [2013] worked on predicting the outcome of the trainee/students based on their past learning behaviors. He has tried to develop model for predicting student's performance based on their personal and university performance characteristics.

Ola and Pallaniappan [2013] proposed a model for evaluating performance of instructors in the institutions of learning by using machine learning algorithms. The proposed framework is suitable for predicting performance and recommending necessary course of actions which can be taken to help school administrators for decision making.

Leung and Joseph [2014] tried to predict the result of the football bowl game based on the past results of the games and the statistics (attempts, time taken for passes) of the team using data mining techniques. The game results were extracted by comparing the results of the competing teams and were used for making the prediction of outcome of the bowl games.

Wang et al [2015] tried to develop a hybrid framework which comprises of data mining techniques and simulation based trainings to improve the effectiveness of evaluation of the training. Data mining techniques were applied to analyze the profiles of the trainees as well as the data generated from the simulation based trainings. Learning outcome of the trainees was based on the trainee's learning behavior, knowledge and confidence about the exercise.

Previous work on evaluating the performance of various classification techniques of data mining

Many works has been done which are related to the comparison of the classification methods. These works has compared the classification techniques with each other with regards to the evaluation criteria and the test data.

Shuhui, Wunsch, Hair and Giesselmann [2001] reported that for the wind farm data neural networks performance was better than the linear regression. They compared the techniques on the basis of the root mean square error and used them to estimate the wind turbine power.

Kumar [2005] also used RMSE criteria for comparing neural network and regression on the real and simulated data. The results of his experiment showed that the regression acts better when the data include errors and the real values of attribute are not available.

Ibrahim and Rusli [2007] compared artificial neural network, decision tree and regression techniques based on the square root of average squared error on the academic data of the students and reported that artificial neural network performs better.

Kim [2008] used RMSE criteria to compare decision tree, neural networks and logistic regression techniques of classification based on the kind of attribute and size of the dataset. His result stated that linear regression works best when the number of categorical variable is one for both continuous and categorical independent variables. Also, artificial neural networks perform best when the categorical variables are two or more.

Betul and Erkan [2013] evaluated classification algorithms using Mc Nemar's test for both nominal and numeric attributes and they found out that multi layer perceptron performed better

than the others. Furthermore, they observed that the evaluation results concurred with Mean Root Square Error.

Kaur, Singh and Jason [2015] used educational dataset to compare the performance of the various classifiers in order to predict the slow learners. They used WEKA to implement the algorithms and among all the classifiers used, multi layer perceptron performed the best..

All the articles mentioned above have compared the different classifiers based on the datasets which are related to a determined problem. Therefore, decision making based on the results provided by them is not general and make different judgments. None of the studies have compared classifiers based on the data coming from simulated shooting exercises.

CHAPTER 3

3. Problem Formulation

Training enhances the efficiency of the trainee and develops the systematic way of performing assigned duties and tasks. It involves imparting specific skills to the trainees for a particular purpose. It is an act of improving the skills of the trainee for doing a particular job. Training is important in order to eliminate the performance deficiencies. The most important aspect of the training is its evaluation. It ensures if the trainees are able to implement their learning in their respective work. Evaluation of the effectiveness of the training is the measurement of the improvement in the trainee's knowledge, skill and behavioral pattern as a result of training program. The army's main objective is to fight and win the war for the country. During the time of peace, their main role is to train for the wartime mission. So it becomes critically important that the training which is provided to them should be highly effective and it should increase the performance of the trainees to a great extent.

This dissertation aims to propose a framework which will be helpful for providing training effectively by considering the past learning behavior of the trainee. The overall objective of the project can be elucidated as the proposal of a framework which can be used to amplify the efficiency and effectiveness of training system by applying data mining techniques on the data generated from the simulation based shooting exercises. The exercise considered for this project involves shooting the target in a simulated environment with or without employing any vehicle which means that trainee can be either static or moving in some sort of a simulated vehicle. This dissertation also attempt to see if CRISP DM model can be amended to get better results as proposed framework has grounds on it.

3.1 Motivation

While going through the series of the research papers to understand the application of data mining techniques in performance prediction in depth, I came across the paper written by Juite

Wang, Y.L Lin and S.Y Hou namely “A data mining approach for training evaluation in simulation-based training” in which the authors have used data mining techniques on the simulation based trainings to improve the evaluation of the trainings. The authors have used knowledge level and confidence level of the trainees to assess the learning outcomes of the trainees. The work done by the authors was quite commendable but could be modified to make the task of evaluating the training more automatic by eliminating interaction like interviews and see if training can be evaluated. The work in this dissertation is kept limited to the training which involves shooting exercises. Also, the dataset used for this project is considerably different as this data consists of shooting exercises by Swedish military personal. Another area which is different is that for this dissertation, past learning behavior of the trainee has been used to predict their learning outcome rather than using skill and confidence level of the trainee.

3.2 Objective of the Study

The objective of this study is broken down into smaller categories so that they are easier to understand.

- The first goal of the study to screen the features of the data which are important in predicting the performance of the trainee. This can be achieved by first getting the insight of the data and then applying feature selection method to select the features which have high contribution in predicting the output.
- The second objective requires the evaluation of various classification techniques and then selecting the technique which gives highest accuracy in predicting the outcome.
- The third objective is to see how CRSIP DM helps in deriving the results in the current scenario when a new phase of selecting a modeling technique is inculcated into it which is also a proposed framework.

Summing up, the overall objective is to propose a framework for enhancing the trainings of the personal by applying data mining techniques on the data generated by the simulator involving shooting exercises. This framework can be used as steps which should be followed to meet the objective of the similar problem. CRISP DM model is taken as the basis for proposing the framework as the problem is treated as data mining problem and test if it fits the best for the current scenario or any amendment can be done to it in order to meet the objective.

CHAPTER 4

4. Method

A method is needed in order to solve the objective of the dissertation; this section illustrates the process of accomplishing the aim of the research. Classification algorithms are applied on the data to achieve the objective of the dissertation which is to propose a framework which can be useful for improving the performance of the trainees which involves predicting how the trainee would perform in the training so that proper guidance can be given. Due to the need of selection of important attributes which influence trainee's outcome and the availability of variety of classification algorithms, data is analyzed to select the attributes and then various classification algorithms are implemented to select the algorithm which gives good accuracy on the data available. This includes steps followed to devise the framework and tools and techniques used for carrying out the work.

4.1 CRISP DM Model

Data mining is relatively new to the field of performance evaluation of personal. However, the Cross Industry Standard Process for Data Mining (CRISP-DM) is a general framework designed for defining all the steps required in data mining projects and it is independent of both the industry sector as well as technology used [Wirth & Hipp, 2000]. In order to provide a framework general purpose for a data mining analysis, CRISP-DM is followed in this dissertation. It is comprised of six phases for applying the data mining techniques to build a model for evaluating the training exercises. The six phases includes:

- **Business Understanding:** The first stage of the model is to understand what needs to be accomplished from a business point of view. This phase has many steps which include, determination of business objectives, assessment of the situation, determination of data mining goals and creation of project plan. This step also involves discovering important factors which can influence the output of the project.

- **Data Understanding:** This phase begins with the collection of initial data. In order to identify the problems related to data quality or to have insight into the data, data analysts start to get familiarize with the data. This step involves tasks like collection of data, verification of data quality and exploration of data.
- **Data Preparation:** All the activities which are related to the construction of the final dataset from the initial data that will be used in building the model are covered in this step. It includes cleaning the data, attribute selection, formatting the data.
- **Model Building:** In this phase, various data mining techniques are applied on the data to construct models. Optimal values of the parameters for the model are calibrated in this step.
- **Evaluation:** The degree to which model built in the previous step meets the business objective is assessed in this step. A key is to ensure that obtained results can be used by the organization.
- **Deployment:** In this step, the model which has been selected in the previous steps, the model which exhibits the best performance is selected for the deployment so that it can be applied on the test data.

4.2 Proposed framework

The six phases of the CRISP DM model pretty much explains the important steps, from understanding the objective to the deployment of the model for new data, which needs to be followed while working on the data mining projects for the successful implementation of the project. It benefits analysts by being used as check list of each task by making sure that nothing important has been missed. One of the important phase of the CRISP DM model is when different data mining techniques are implemented on the data to build the model which can then be deployed if it meets the validation criteria. Data mining is a complex process which requires various tools and techniques. Therefore, before building a model, there is a dire need to select the

relevant techniques which can be implemented for the creation of the model. It is important to decide what techniques should be followed in order to meet the objective of the data mining problem. For example, which techniques in classification, clustering or association analysis is to be considered for getting the desired outcome with the good efficiency. Since selecting a model is one of the essential task for the success of the project, thus, should not be overlooked. Therefore, it should be kept in the checklist or should be a considered as a separate phase in the model. So, this dissertation suggests that one more step should be added in the CRISP DM model which emphasizes on selecting the data mining technique which can then be applied in the “Building Model” phase.

Phases of the new proposed framework are:

- **Business Understanding:** This phase focuses on understanding the objectives and requirement of the project from the business perspective. This knowledge is then converted into a data mining problem definition. It is the initial step which has emphasis on understanding training requirements and objectives. It is important to have knowledge of the tasks to be trained and the context of the training to have a good way to analyze the training. In order to enhance the data mining effectiveness and to comprehend the problem, it is important to gather domain knowledge. Also, help should be taken from domain experts to outline the objectives of the project and requirements from the business point of view so that the objectives of the project could be effectively translated into the problem definition from the data mining perspective. Therefore, in this step, problem and its constraints are defined. The objectives and which data mining technique has to be applied are determined in this step.
- **Data Understanding:** The main foundation of the data mining is to collect the right data. Thus, experts need to be consulted for their opinion to select the important feature which can influence the outcome of the trainee’s performance. Also, literature review related to the problem needs to be done. The data is collected keeping in mind the aspects which are important in knowing the performance of the trainee as this work is done for

evaluating the training. For this dissertation, data has been taken directly from the simulator used in the providing the training.

- **Data Preparation:** It is necessary to clean the data before it can be used for applying data mining techniques. Data which is incomplete or contains noise needs to be removed. In order to select the attributes which has high influence on predicting the outcome of the trainee, feature selection techniques can be applied. This also helps in reducing the dimensionality of the dataset which makes it easier to discard the attributes which are not of much relevance in forecasting the learning outcome of the trainee.

- **Modeling Technique Selection:** In this step, depending on the desired outcome, actual modeling technique is selected. Even though the tool to be used is selected in the business understanding phase, specific modeling techniques are selected at this stage like back propagation neural network or Cart in decision trees. Many modeling techniques form assumptions on the data like no missing values, uniform distribution of data etc. All these assumption should be recorded in this phase. Selecting right techniques is one of the major task in the accomplishment of the objective of the project. Since the problem identified in this dissertation is considered as a prediction problem and presence of the target variable directs the use of the supervised learning techniques thus, classification techniques which can be implemented on the data available are selected in this step.

- **Model building:** In this phase, data mining techniques are applied on the data for constructing models to evaluate the outcome of the trainee based on their learning behavior. The problem can be considered as classification problem for predicting the class of the data objects which has unknown class labels. Classification is a classical problem in data mining [Tsang et al., 2009]. For instance, the trainee's performance can be identified as if trainee will hit or miss the target using the classification model. Classification is divided into two categories as supervised classification and unsupervised classification. In supervised classification, dataset is analyzed for which the class for the target variable is known and the algorithm tries to find the relationship between the value of the target variable and the value of the predictors. All these relationships are then encapsulated into the model which is then applied on a dataset for which the class of the

target variable is unknown. On the other hand, in unsupervised classification, the value of class for the target variable is unknown, which provides the possibility to analyze the problem with pre-existing knowledge.

- **Evaluation:** In this step, the constructed model is evaluated. Evaluation of the model based on accuracy depends on the number of test records correctly and incorrectly identified by the model. These numbers are displayed in a tabular form known as confusion matrix. The table in table 1 shows the confusion matrix for the binary classifier. Entry f_{ij} in the matrix denotes the count of records belonging to class i but predicted to be part of class j . Therefore, sum of f_{00} and f_{11} denotes the number of correct prediction whereas sum of f_{01} and f_{10} denotes number of incorrect predictions made by the classifier. If the model has the satisfactory results then it can be used for training evaluation which is predicting the trainee’s performance.

		<i>Predicted Class</i>	
		<i>Class = 1</i>	<i>Class = 0</i>
<i>Actual Class</i>	<i>Class = 1</i>	f_{11}	f_{10}
	<i>Class = 0</i>	f_{01}	f_{00}

Table 1. Confusion Matrix for Binary Classifier

Thus, the accuracy can be calculated as

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total Number of predictions}} = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

The other method to check the performance of the model is to calculate the root mean square error (RMSE). Root mean square error is used to measure the difference between the expected or actual values and the values predicted by the model. The formula for calculating the root mean square error is:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

Where n is the number of instances, \hat{y}_j is the actual value and y_j is the predicted value.

- **Deployment:** In this phase, the model created in the previous step is implemented for improving the performance of simulation based shooting exercise. As an illustration, an organization may need to deploy a model to know which trainee tend to fail in the training or the one who is going to pass in the training so that the instructors can provide feedback to trainees according to their performance which can be helpful for improving their performance.

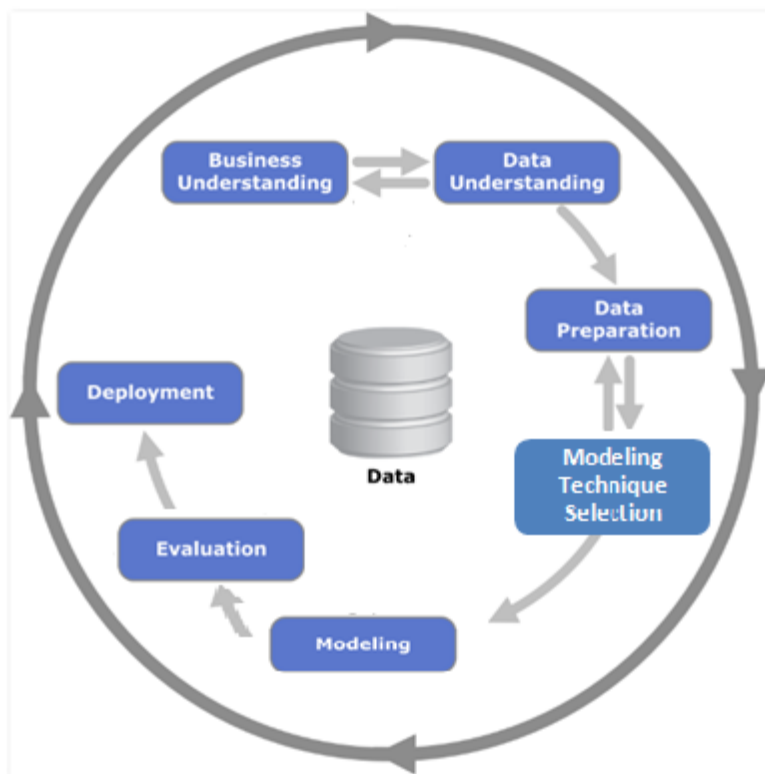


Fig 2. The Crisp DM process Model

4.3 Tools Used for the implementation

4.3.1 WEKA

Because of the need for the unified work bench to give researchers an easy access to the state of art machine learning techniques, WEKA (Waikato Environment for Knowledge Analysis) was developed at the University of Waikato, New Zealand and is implemented in Java. Weka is very well- known tool for data mining and machine learning. Weka is widely used in data mining research and has been accepted within academic and the business circles. It consists of visualization tools, data analysis and predictive modeling algorithms. It supports many data mining tasks, which includes data preprocessing, classification, clustering, feature selection, regression and visualization. Data can be loaded from various sources like databases, Urls, and files.

In this thesis work, Weka version 3.7 has been used for feature selection and then comparison of the performance of the classification algorithms has been done in order to select the best model which suits the data. Advantages of Weka include:

- Free availability under the GNU General Public License.
- Portability, since it is implemented in the Java programming language it is portable and thus runs on almost all modern computing platform.
- A complete collection of data preprocessing and modeling techniques.
- Provides graphical user interfaces for ease of use.

4.3.2 Tableau

The idea that visualization and analysis should not be different activities and should be integrated into visual analysis process gave birth to the Tableau software. Visual analysis products are built by Tableau in order to help analyst to ask and answer analytical queries which involves data that is stored in spreadsheets or databases. It lets the user analyze the data better by using their natural ability of thinking creatively when they analyze the data visually. The challenge for computer graphics pioneers to collaborate with database researchers at Stanford University by the US

Defense Department led to the development effort of Tableau. The tableau framework is based on five principles [Hanrahan, Stolte & Mackinlay, 2007]:

- Easy Interface: Users can analyze the data visually by just by dragging fields onto the worksheets.
- Data Exploration: It automatically creates the computations whenever a picture is composed.
- Expressiveness: It is based on declarative query language (VizQL) which is highly expressible because of which Tableau supports numerous class of visualizations.
- Database Independence: user can use same front end using different databases which saves training costs.
- Visualization Best Practices: graphic design principles are based on the best practices.

4.4 Data Mining Techniques Used for the implementation

4.4.1 Feature Selection

It has been widely recognized that working with large number of features adversely affects the performance of the inductive learning algorithms [Maaten, Postma & Herik, 200]. Since it is difficult by the machine learning algorithms to deal with high dimensional data, feature selection methods have become indispensable part of the leaning process [Kalousis, Prados & Hilario, 2007]. Feature selection is a process of selecting the relevant attributes and rejecting the irrelevant ones. The aim of the feature selection is three fold: improve the performance of the prediction by the predictors, provide more cost effective predictors, and provide the better understanding of the domain. There are three methods for finding the relevant features: wrapper method, filter method and embedded method [Guyon et al., 2006]. In the filter method, scoring is assigned to each feature based on the statistical properties. Then based on the ranking of the scores, features are either selected to be retained or discarded from the dataset. On the other hand, wrapper method assesses the subset of the attributes according to the efficiency of the given predictor and considers the selection of set of attributes as search problem. This method searches for the good subset by using the learning model as a part of evaluation function. The problem with this approach is the vastness of the feature space which makes it computationally

expensive to look at every possible combination. Embedded method considers which feature gives high accuracy to the model while the model is being created. It is computationally less expensive than wrapper method but it is specific to a learning model. Filter method has been applied on the military training dataset because it is computationally fast and simple and also independent of the learning algorithm to be applied on the dataset. Also, since many classification models will be applied on the dataset, application of filter method comes out to be a better choice.

4.4.2 Logistic Regression

Regression method - Logistic Regression is well suited classification method for describing and testing hypothesis for relationships between predictor variables and categorical outcome of the target variable [peng et al., 2002]. When the dependent variable is dichotomous in nature then logistic regression method is sufficient capable for classification [Abdolmaleki et al., 2004]. Since the result of the trainee is dichotomous, as he/she can hit the target or not, logistic regression method can be applied. The main goal of the logistic regression model is to predict the logit of dependent categorical variable 'Y' from the independent variables, $x_1, x_2, x_3 \dots x_n$. The logit function is the natural log of the odds that 'Y' is among one of the categories. Odd is the ratio of probability of 'Y' belonging to the category (p) to the 'Y' not belonging to that category (1-p).

The logistic regression method has formula:

$$\text{Logit}(Y) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

Where β is the regression coefficient.

According to the equation above, logistic regression method considers the logit as the dependent variable and converts it as a linear function of n predictors. Thus, the value of the regression coefficient determines the relationship between the predictors and the logit of Y. Predictor values are linked with large logit of Y when the value of the coefficient(β) is large or vice versa.

The main advantage of the logistic regression method is that the probability that certain instance belongs to a particular class and the confidence interval can be calculated directly by researchers

which makes this method easy to interpret. But the problem of overfitting may occur if the dataset has large number of attributes.

4.4.3 Decision Tree

Decision trees represent a sequence of rules in a tree like structure that leads to a class [Peng et al., 2009]. Decision tree approximates the discrete value target functions in which the function is represented with the use of decision tree. They are the most widely used and practical method for inductive learning [Tom M. Mitchell, 1997]. The decision tree consists of the three fundamental nodes known as root node, leaf node and internal node. Test condition on the attribute is denoted by the internal node, branch node denotes the result of the test and leaf node denotes the label of the class. Instances are classified by traversing from the root node to the leaf node. Attributes specified by each node are tested against the condition and then move forward to the branch node according to the value in the given set. Based on the attribute value test the decision tree algorithm recursively splits the training dataset into further subsets. For capturing the knowledge in expert systems, decision tree algorithms have been successfully implemented. The main task performed in expert systems is to apply inductive methods on the given attributes of unknown class to determine appropriate classification according to decision tree rules [Peng et al., 2009]. There are many algorithms present in the Decision tree learning like ID3, C4.5, CART etc. C4.5 algorithm for building decision tree is chosen for the application on the dataset because C4.5 method has good classification accuracy compared to other algorithms [Anyanwu & Shiva, 2009] and it deals with noise.

The main advantage of the decision trees is their robust nature towards the outliers because decision of divergence depends on the ordering of the attributes and not on the magnitude of the value [Leung, 2007]. Also, the presence of IF THEN condition makes them easier for understanding. But the decision trees have low performance than methods like artificial neural networks if the problem has non-linear structure because the decision trees follow linear structure during the tree generation.

4.4.4 Artificial Neural Network (ANN)

Forecasting is one of the main application areas of Artificial Neural Networks [Sharda, 1994]. These networks learn from the training set and find the functional relationship between the data attributes which have complex relationships or are hard to understand. Artificial Neural networks learn the data fed to them and even if the data is noisy they infer the dataset set very well. They are inspired from the biological neural network structure and their way of solving problems. A neural network comprises of simple processing units called as neurons, directed weighted connection between the neurons, and then mathematical function is applied to determine the activation of neuron. Neuron takes input data and performs operation on it. The outcome of these operations called as activation is passed onto the other neurons. The desired output can be obtained by adjusting the weights. More the weight of the neuron, stronger will be the input which has to be multiplied by it.

The back propagation algorithm is used in multi layered feed forward ANNs. The neurons are organized in layer form and the signals are sent in forward direction and then the errors are propagated backwards. The back propagation algorithm makes use of supervised learning mechanism. Therefore, the algorithm is fed with training examples and then the error is calculated. The aim of back propagation algorithm is to reduce the error until ANN learns the training data. Fig 3 shows the structure of neuron:

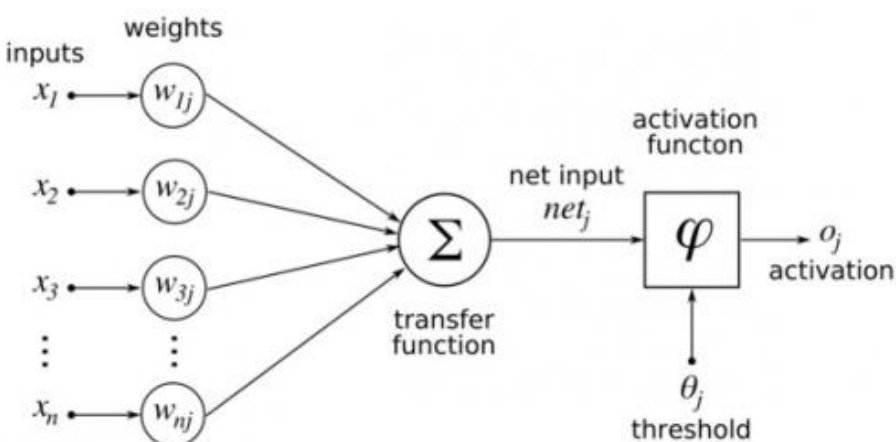


Fig 3. Structure of single neuron

Implementing back propagation algorithm, the activation function of the ANN is the sum of inputs (x) multiplied with their corresponding weights (w).

$$A_j(x, w) = \sum_{t=0}^x x_j W_{jt}$$

Sigmoid function is the most common output function. Thus, output will be

$$\text{Output}_j = \frac{1}{1 + e^{A_j(x, w)}}$$

The output and the corresponding target output of the node are compared to calculate the error. This error is then propagated to the network adjusting the weights to minimize the sum of square error.

Artificial neural networks are suitable for the problems in which the solution requires a knowledge which is difficult to specify but contains enough observations [Zhang, Patuwo & Hu, 1998]. But the size of the network depends on the degree of non-linearity and dimensionality of the dataset. Also, the generation of the weights is difficult to interpret because they are affected by the program that is used to generate them [Baxt, 1995]. The decision tree approach is better when a researcher needs to see how the particular conclusion has been made because neural network is more of a black box.

CHAPTER 5

5 Implementation of the framework

5.1 Business Understanding

At the beginning of project, meetings were held with the supervisor to discuss the expectations and the objectives of the project. With the aim to improve the learning outcomes of the trainee and improve their training performance, the Swedish army sought to enhance the evaluation of the training. Thus, the objective can be elucidated as the building a framework for analyzation of the data collected from the military shooting training exercises to examine the learning behavior of the trainees in order to predict their training outcomes so that the proper feedback can be provided to them to improve their performance. Data mining techniques are the most effective techniques to manage the data and discovering the useful information present in the data. That is why, data mining techniques are applied to build the model for the training evaluation. Data collected during the trainings can be used by the instructor to analyze the performance of the trainee so that feedbacks can be provided on time in order to improve the trainee's shooting performance. This will help in saving the time and efforts of the instructor in providing the timely feedbacks for improving trainee's behavior. Also, predictions can be made on how the trainee will perform in future based on the past learning behavior of the trainee. Based on the trend of how the trainee is performing, a personalized feedback and special training can be arranged for him/her.

5.2 Data Understanding

Once the understanding is made on what is expected to be the result of the project, it is a time to build an understanding on the data collected so that appropriate attributes and relationships between them can be selected for the further use. The data for this project has been collected by the Swedish Military following the quantitative approach as the data is used for performing statistical analysis. The data is then provided to the university through the Combitech company.

The data comes from the simulator which measures the factors like target coordinates, turret orientation point of contact etc. It has been collected keeping the important attributes in mind which should be considered for analyzing the performance of shooting. The data has been captured only for a day for various soldiers. So, in total there are 1193 sample data points are used. It comprises of the important attributes like coordinates, log index, orientation, ammunition type, projectile range, velocity, hit results etc which are helpful to see the pattern of the trainee's performance and can be used in designing model for predicting the outcome. The data is directly exported from simulator and saved in a comma separated format. The format is simple and could be used with different programs.

5.2.1 Analysis of Data

After achieving an appropriate understanding of the data, this step tries to point out the interesting patterns or facts present in the data which can play an important role in obtaining the output of the project. The right presentation of the data makes it easy to understand the hidden relationship between the attributes and information like outliers can be easily recognized. The dependency of one attribute on another helps to understand the trends or patterns which are present in the data. Understanding these patterns helps in knowing the data better which in turn helps in making decisions about factors influencing the dependant variable. In this dissertation, data visualization is important especially in understanding the main relationships present in the data and their characteristics. Tableau software has been used for analyzing the data for this dissertation. Some of the interesting aspects found while visualizing the data are:

- Fig 4 shows the performance of the trainees in terms of count of shots they made before they actually hit the target. Red color represents the bad performance of the trainee as it shows that the trainee has taken many trials to hit the target. On the other hand, green represents the good performance as the number of trials is low. Darker green color represents better performance.

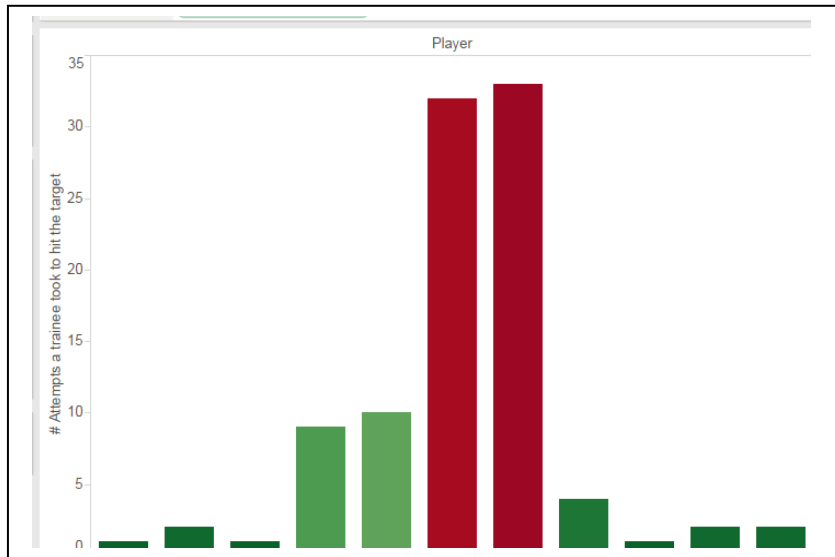


Fig 4. Trainees Performance in terms of hit counts

- Fig 5 shows the different types of ammunitions used in the training and their hit rates. Different ammunition types are represented by different colors of the bubble and size of the bubble represents the hit count. More the size of the bubble, high is the hit count of the ammunition. It can be clearly seen from the figure that ammunition type “HK 416” has high hit rate.

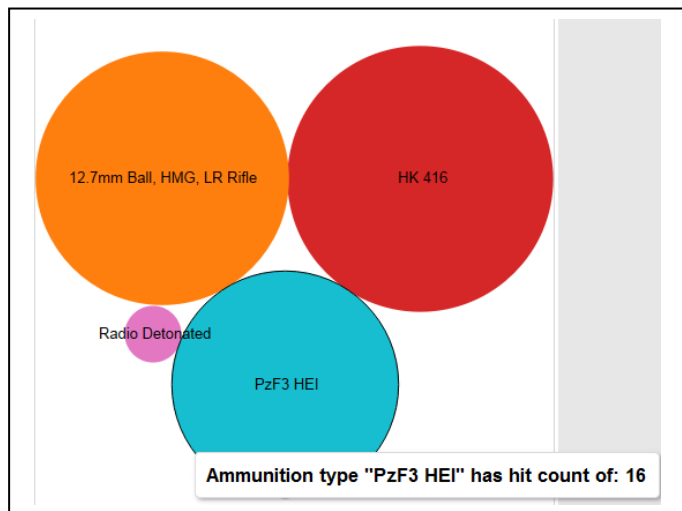


Fig 5. Ammunition and Hit Rate relation.

From the fig 4, it can be further analyzed the soldiers performing poor (represented by red bar) might belong the ammunition type "Radio Detonated" as the performance of the trainees handling this ammunition type is low (shown in fig 5). This case shows that the training which requires handling radio detonators has to be improved. But if the trainees (represented by red bar) belong to the other ammunition types then it can be inferred that the trainee is not performing well and should be given extra support.

- Fig 6 shows the performance of the players who are firing using some sort of vehicle. In this scenario, it is important to maintain the orientation of the vehicle balanced with the target of hitting. Graph starts form low value and reaches the peak value and then comes down which represents the start, peak and finishing/landing of the vehicle. All the players show the same behavior but for the second trainee, imbalance in the orientation can be seen before finishing the training. So, it can be concluded that second payer needs more guidance and training practice to fulfill the training objective more efficiently.

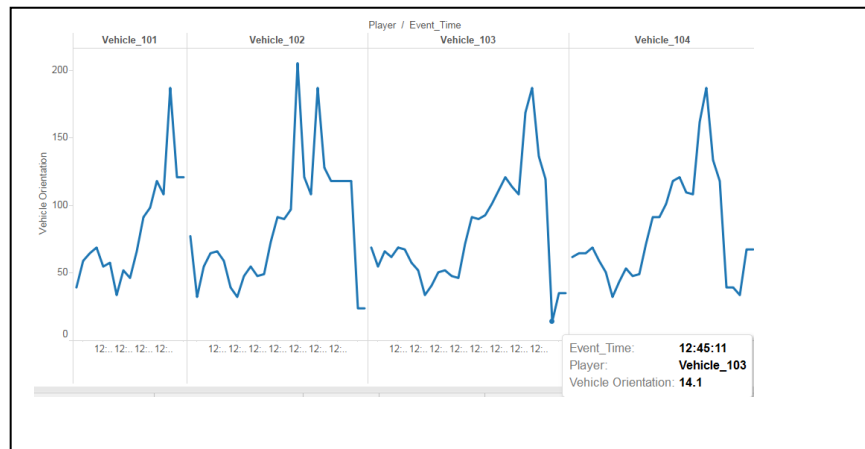


Fig 6. Vehicle orientation per Player

5.3 Data Preparation

Some prior knowledge can lead to the application of grouping or transformation on the input variables [see, e.g., Fan & Lv, 2008]. Simulated data consists of high dimensions or can be said it captures large number of features. Therefore, it is important to reduce the dimensionality of the data in order to avoid the complexity and the overfitting of the models. This step focuses on preparing the data so that it can be used for applying algorithms to build the model of predicting the outcome of trainee. This involves cleaning, selecting important attributes, filling missing values or discarding the rows which has missing values etc. Preprocessing of data has been done in this case. Firstly, the attributes which were not captured properly as they contained many missing values or had only one value has been discarded. Secondly, the attributes like “horizontal point of impact” or “velocity East” which doesn’t seem to be of much relevance in predicting the outcome are ignored. Initially, 37 attributes were present for each trainee but after cleaning the data only 15 attributes are left which are useful in assessing shooting performance for the trainees.

5.3.1 Dimensionality Reduction

Application of feature selection methods on the data containing many irrelevant features has become a necessity in many applications because many pattern recognition techniques cannot cope with high dimensional data [Guyon & Elisseeff, 2003; Liu & Motoda, 1998]. The performance of the model is highly dependent on the selection of the relevant variables in the dataset. There is a possibility that out of 15 variables in our dataset some variables might contain redundant and insignificant information for predicting outcome of the trainees. So, feature selection method has been applied with the help of Weka tool to select important attributes that contributes the most in the accuracy of the prediction. Dataset was loaded into the Weka and then in “Select attribute” section, filter method was applied on the dataset using the ranking function. The result of the filter method is shown in the fig 7. It shows the ranking of the attributes based on their relevance in predicting the dependent variable. First column tells the relevance percentage to the target variable and next column represents the number and name of the attribute from the dataset. Seeing the output, it can be concluded that Ammunition type has a

significant effect on the result that whether the trainee will hit the target or miss it. Since good number of attributes to predict the result are present, attributes which has influence of less than 30 percent on the dependent variable can be ignored. Therefore, for the classification method, the attributes which have relevance of more than 30 percent has been chosen as predictors and rest are discarded. Also, the attributes “Firing Player SimId” and “Ammunition Code Text” are the textual representation of their associated attributes “Firing Player” and “Ammunition Code”. Therefore, they represent the same information as their associated attribute and can be ignored. Finally, there are six attributes remaining which will be used in predicting the outcome of the classification model.

```
-  
  
Ranked attributes:  
0.6097 13 Ammunition Code Text  
0.6093 12 Ammunition Code  
0.4956 6 Firing Player  
0.4367 2 Player  
0.4232 11 Firing Player SimID  
0.3244 4 Event_Time  
0.3085 14 Associated RAD  
0.2091 9 Associated Player  
0.2037 10 Associated Player SimID  
0.1211 5 Log Index  
0.0616 7 Coordinates  
0.0308 8 Coordinates_Editted  
0 3 Event_Date  
0 1 Event Name  
  
Selected attributes: 13,12,6,2,11,4,14,9,10,5,7,8,3,1 : 14
```

Fig 7. Output of filter method of feature selection

After selecting the relevant features it is needed to choose the classification models which can be applied on the dataset to predict the outcome of the model. The following briefly describes the algorithms which will be applied on the training data.

5.4 Modeling Technique Selection

Model selection is a task of, given a dataset, selecting a statistical model from a set of available models. This phase demands investment of the time to discover and find models which suits best to the problem and to the data. Availability of the labeled data gives the ability to make use supervised learning techniques. Now, it is known that supervised learning techniques are going to be practiced; next step is to select the algorithms, available under supervised learning, which can be implemented on the data to train the model. These algorithms make use of set of available examples in order to make predictions. For instance, in the current scenario, historical performance of the personal can be used to predict his/her performance in the next training session. A Supervised learning algorithm tries to find patterns in the value labels and each algorithm aim to find different type of patterns. Once the algorithm observes the best pattern it can, then that pattern is used for making predictions for unlabeled testing data i.e. performance outcome of the trainee.

For selecting the model which will be implemented on the data, one should consider following below measures:

- **Accuracy:** It is a proportion of correctly classified instances. While evaluating the model, it is the first metric which is looked.
- **Training Time:** The time taken by the model to get trained varies a great deal between algorithms. It derives the choice of algorithm when the time is limited, especially when the size of the dataset is large.
- **Number of parameters:** Parameters works as knobs for setting up an algorithm. They affect the behavior of the algorithm like number of iterations or error tolerance. Sometimes accuracy and training time can be quite sensitive to getting right settings. Usually, algorithms having large number of parameters needs the most trial and error to find good combination. Having many parameters implies that algorithm has more flexibility. Provided the good combination of parameters can be achieved, they can achieve good amount of accuracy.

For this dissertation, because of the availability of the hit results of the trainees, supervised classification algorithms are applied on the dataset to assess the trainee outcome. After comparing the pros and cons of machine learning methods, three classifiers, logistic Regression [Khoshgoftaar et al., 1999], Decision tree [Quinlan, 1993], Artificial Neural Network techniques [Lippmann, 1987] of the supervised classification algorithms are selected for implementation on the dataset for carrying out the considered prediction task. Regression is intrinsically simple and it has low variance, thus, less prone to the overfitting. On the other hand, decision trees are like “white box” because the knowledge acquired by the model can be expressed in readable format and are less affected by outliers. It classifies the data without doing much calculation. Artificial Neural networks is a powerful self adaptive, data driven computational tool that has the ability to capture non linear and complex underlying aspects of the data with a good amount of accuracy. Therefore, simplicity of the Logistic regression, speed of the decision trees in building the model and the ability to detect possible interactions between dependent and independent variable by the Neural Networks has influenced the choice to consider these algorithms. The designed model will be evaluated and then based on the prediction accuracy and time taken to build the model, best model will be chosen for the deployment.

5.5 Building the Model

Based on the data collected from the training sessions, supervised classification techniques such as logistic regression, C4.5 method from decision trees and neural network learning methods have been applied using Weka to assess the performance of the trainee. The following section explains the model creation and the parameters that have been chosen for building the model.

5.5.1 Experimental Setup

Logistic regression method [Landwehr, Hall & Frank, 2005] is applied on the data with heuristicstop enabled as it gives a large speed up for small datasets. In this method, Logitboost with simple regression functions is used as base learner for fitting the logistic model. The maximum number of iterations for Logitboost is kept as 500 which is also the default value in Weka. Fig 8 shows the output screen of the result of the logistic regression model:

```

Time taken to build model: 1.31 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances      35      87.5 %
Incorrectly Classified Instances    5       12.5 %
Kappa statistic                    0.6552
Mean absolute error                 0.2542
Root mean squared error             0.3318
Relative absolute error             69.4808 %
Root relative squared error         76.5613 %
Total Number of Instances          40

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
      0.7      0.067   0.778     0.7     0.737     0.817    Hit
      0.933    0.3     0.903    0.933   0.918     0.817    Miss
Weighted Avg.   0.875   0.242   0.872    0.875   0.873     0.817

=== Confusion Matrix ===

 a  b  <-- classified as
 7  3 | a = Hit
 2 28 | b = Miss
    
```

Fig 8. Output of Logistic Regression technique

For the neural network learning model, back propagation method has been used for the classification of the instances. The nodes in the network are all sigmoid. The number of hidden layers is chosen to be 1 for the sake of the simplicity of the network and the number of neurons in the hidden layer is chosen based on the rule of thumb which says that number of neurons in hidden layer should be two third of the sum of the input neurons and the output neurons. Number of input neurons and the output neurons are kept as 5 and 2 respectively. The performance of the algorithm depends on the learning rate. If the learning rate is kept high then the algorithm will keep oscillating and become unstable whereas if the learning rate is very small then the algorithm will take lot of time to converge. Learning rate tells about how much adjustment to the weights has to be made. After analyzing different values for learning rate, it is kept as 0.2 for the model. Output screen of the neural network algorithm application on the dataset is shown in fig 9.

```

Time taken to build model: 1.77 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances      18          90 %
Incorrectly Classified Instances    2           10 %
Kappa statistic                    0.7333
Mean absolute error                 0.1534
Root mean squared error             0.2615
Relative absolute error             41.7456 %
Root relative squared error         60.3546 %
Total Number of Instances          20

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
          0.8      0.067    0.8        0.8     0.8        0.987    Hit
          0.933    0.2      0.933     0.933  0.933     0.987    Miss
Weighted Avg.    0.9      0.167    0.9        0.9     0.9        0.987

=== Confusion Matrix ===

 a b  <-- classified as
 4 1 | a = Hit
 1 14 | b = Miss
    
```

Fig 9. Output of Neural Network technique

The J48 method has been used to build C4.5 decision tree developed by Ross Quinlan [1993] in WEKA tool. C4.5 algorithm uses divide and conquer approach for the growth of the decision tree. The confidence factor has been kept as 0.25 for pruning the tree. Based on hit and trial method, number of folds was kept as 3 to determine the amount of data used for reduced error pruning. For improving the accuracy of the tree, 8 trials are used to grow the initial decision tree. For the subsequent validation phases in order to reduce the misclassification error, the final parameters have been determined. The fig 10 below is the screen shot from the WEKA tool for applying the J48 method on the data:

```

Time taken to build model: 0.01 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances      39          97.5 %
Incorrectly Classified Instances    1           2.5 %
Kappa statistic                    0.9355
Mean absolute error                 0.0813
Root mean squared error             0.1723
Relative absolute error             22.2122 %
Root relative squared error         39.7552 %
Total Number of Instances          40

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
          1      0.033    0.909     1       0.952     0.983    Hit
          0.967    0        1         0.967  0.983     0.983    Miss
Weighted Avg.    0.975    0.008    0.977     0.975  0.975     0.983

=== Confusion Matrix ===

 a b  <-- classified as
10 0 | a = Hit
 1 29 | b = Miss
    
```

Fig 10. Output of Decision Tree technique

5.6 Model Evaluation

Evaluation of the classifier is one of the key points in the process of data mining process. The main evaluation criterion for classifier is the overall accuracy attained by model validation and the root mean square error of the model. Same experiment procedure is used as suggested by the Weka to gauge and study the performance of the classification models namely Regression, Decision tree and artificial neural network. All the data in Weka is considered as instances and features in the data are called as attributes. To validate the models built above, 70 % of the data is used as training set and rest is used for the validation purpose. The results of the performance metrics of classification models are tabularized in table 2. Observing the results present in the table, it is seen that logistic regression has shown the least accuracy of 87.5% in predicting the trainee's performance outcome and has the maximum root mean square error of 0.33 as compared to other two models. Decision tree on the other hand has shown the maximum accuracy which is 97.5% and the least root mean square error 0.17 and time taken to build the model is 0.01 seconds. Artificial neural network has taken the maximum time to build the model with the accuracy of 90% and root mean square error of 0.26. Therefore, it can be agreed that based on our data, decision tree is the most accurate model for classification and has shown the maximum accuracy among other models. Therefore, it can be concluded that decision tree model fits the best on our data and is recommended for using it in predicting trainee's performance.

Parameter	Logistic Regression	Decision Tree	Artificial Neural Network
Accuracy Rate	87.5	97.5	90
Root Mean Square Error	0.33	0.17	0.26
Time Taken to build the model(sec)	1.43	0.01	1.77

Table 2. Output of the Classifiers for Performance Parameters

5.7 Model Deployment

The framework for the training evaluation is build over the algorithms used in the prior section. Since the decision tree method has shown the maximum accuracy and minimum root mean square error and output time, it is used for the deployment for training purposes. The system captures the data from the simulator used in the training exercise and then the data is processed to make use of the attributes which are highly significant in predicting the shooting outcome of the trainee. Then the selected attributes are used predict the performance of the trainee whether the trainee will be able to hit the target or not. If there exists a gap between the trainee’s performance and outcome of the performance evaluation model, then the trainee would be given appropriate feedback in order to improve his/her performance. Based on the feedback a trainee has got from the instructor and the scenarios provided by the simulator, he/she will perform the exercise again. The response data will again be provided by the simulator to the evaluation model. In this way, trainee’s learning behavior is captured over time and will be used further for improving the performance of the trainee and the accuracy of the evaluator. Instructor can also provide guidance according to his/her own experience integrated with the result of the model. Fig 11 below shows the system framework which can be deployed for assisting the instructor for enhancing the shooting performance of the trainees.

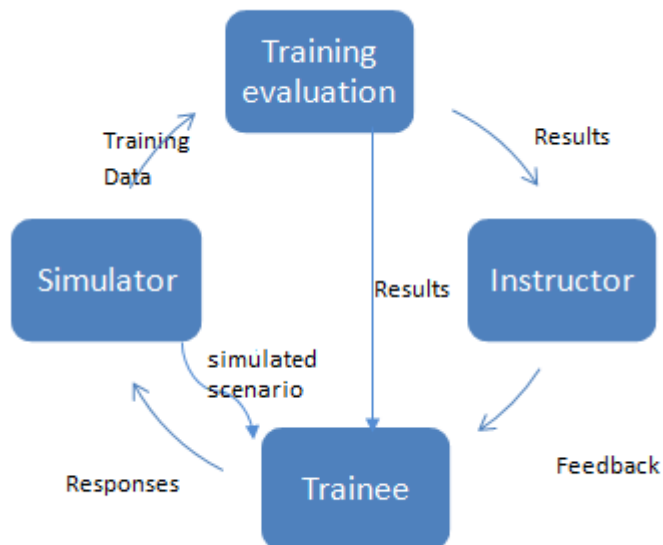


Fig 11. System Framework

CHAPTER 6

Discussion

Military training not only require trainees to finish the training but also finishing the training in the most efficient way as it involves the risk of lives when they have to work in reality. Also, data collected by simulators is usually high dimensional as they tend to record data from many aspects. Analysing such dataset is considered to be challenging for simulation studies. In conjunction, data mining is plausible in extracting knowledge, anomalies and relationships from large data. But the accuracy of classification algorithms deteriorates in high dimensions because of the phenomenon named as curse of dimensionality [Pappu & Pardalos, 2014]. Thus, it is very important to choose the good number of attributes which can be considered as independent variable while applying the classification algorithms. Also it is seen that classification techniques showed good amount of accuracy in predicting the outcome of the trainee, therefore, data mining is ought to be a good fit in building a model for evaluating trainee's performance. Since the problem is considered to be related to data mining, therefore, it is necessary to follow the steps which are standard for tackling data mining problems. Thus, CRISP DM model is taken as basis for the proposed framework. The proposed framework comprises of the seven steps.

In the first step which is business understanding, proper understanding of the objective of the project is taken care before initializing the actual work on the project. Meetings are conducted with the supervisor to have a clear picture of what is expected out of the project. In the second phase which is data understanding, purpose of each attribute is understood so that the meaning and importance of the values of the attribute can be apprehended in order to find the interesting patterns present in the data. To understand the relationship of the attributes, visual analytics approach has been considered and tableau has been used. Then in the third phase, data has been pre-processed to make it useful for implementing classification algorithms. Training data is screened using the feature selection method to determine the best attributes which can be used in classification algorithms to predict the performance of the trainee. In the fourth phase, suitable classification techniques are selected based on the functionality of the classifiers. In this phase, three classification techniques are chosen to be implemented on the data which are decision

trees, regression and artificial neural network. After selecting the classification techniques in the previous step, prediction model is built using those techniques and their performance measures are calculated. Training data for the military exercises is analysed by various classification algorithms for their accuracy and error rate in predicting the outcome of the trainee. In the sixth step, evaluation of the models built is done based on the accuracy and the time taken by the models. It is found that decision tree predicted outcomes are the most accurate. The decision on correctness of the algorithm has been taken based on how many instances an algorithm has classified correctly. Also, time taken by the algorithm to build the learning model has been examined while considering the best algorithm and is observed that it is also the least for decision tree algorithm. Therefore, the decision tree algorithm is chosen for building the model which will be used for improving the performance of the trainees. Then in the last step, since the decision tree showed the maximum accuracy, it is suggested that decision tree should be deployed on this kind of data.

The trainee's performance will be evaluated against the outcome predicted by the model and if there exists a gap then the instructor will help trainees by providing proper feedback to achieve the training goals. Following the proposed framework, trainees will be grouped according to their performance. The group of trainees who exhibit lower performance than expected by the model will be provided appropriate feedback and guidance to improve their performance so that they can meet their expected goals. On the other hand, for trainees who show better performance than predicted by the model, instructors can polish their skills towards the higher expert levels and will be moved to next cycle of the training to perform more comprehensive training as per their capabilities such as handling high end shooting gadgets or intensive shooting trainings. In this way, the trainees who are performing well in their training will keep moving forward their motivations will be maintained high. Also, they will not have the feeling that they are ossified. On the contrary, to make sure that trainees with low performance do not lose their confidence, they will be given extra support so that they can finish their training objectives effectively and will have the encouragement to accomplish the goals in order to move to the next levels. This will enable instructors to see how a certain strategy of training is working on a trainee individually or seeing trainee's learning curve new training strategies can be developed according to their calibre. Visual analysis is used as an aid for analysing the data for finding interesting relationships that exists in the data.

Result of feature selection and the fig 5 shows that the ammunition code has the most influence in predicting the outcome of the trainee and "HK416" has the highest hit rate. Therefore, it can be said that trainees using "HK416" in their trainings tend to show higher performance than others. So, it can be seen as the trainings involving the use of "HK416" has low difficulty level or the trainees have reached remarkable level in the trainings involving "HK416". In this case, more guidance is needed in the trainings which involve other ammunition types as trainees should perform well in handling all sorts of ammunitions or difficulty level of trainings involving "HK416" can be increased. They should be jack of all trades.

There are many classification methods available which can be used for predicting the trainee's performance but only three of the methods are considered in this dissertation. Even though the decision tree algorithms has given the maximum accuracy on the data present which is quite good but there might exist a case where some other classification algorithms would give better accuracy rate if the data increases in future.

It can be seen form the output of the different classification models that the time taken by models is very small (in seconds). This is because the data which was available for this dissertation was only for one day and for few soldiers. But if the dataset would have been larger, then the time taken by model will also be more. Also depending on the distribution of the data, classifiers will exhibit different performance.

Also, while evaluating the performance of the classifiers it is seen that the neural network algorithm has shown the lower accuracy than the decision tree algorithm. One of the possible reasons behind this could be overfitting of the model. Overfitting is caused when the algorithm is heavily swayed by the training set. This can be mitigated by adding more data for training the model.

CHAPTER 7

7.1 Conclusion

In this paper, a framework has been proposed, which is an extension of the CRISP DM model, on how the data mining techniques can be applied to improve the performance of the trainee. The framework has been implemented on the data which is collected from the simulator used for shooting exercise by the Swedish Military to analyse the performance of the trainee. In the first stage, dimensionality of the data has been reduced or can be said that important attributes have been selected, which has high participation in determining the outcome of the trainee's performance. Filter method of feature selection technique has been applied on the dataset using Weka tool. Result of this method has shown that the type of ammunition used in shooting exercise has substantial impact on the outcome of the trainee. Next, classification algorithms have been implemented on the dataset, produced by discarding the attributes which doesn't have much contribution in forecasting the performance, to predict the result of the trainee's performance. Out of the three techniques, Artificial Neural Network, Decision tree and Regression, which were applied on the data to predict the training outcome of the trainee, Decision tree algorithm has shown the maximum accuracy and has taken the least time for our dataset. The framework will be useful for instructors in improving the performance of the trainees as it considers only the past learning behaviour of the trainees which is the most important factor in physical trainings. It predicts the performance of the trainee from the data available and if there comes any discrepancy between the outcomes of the trainee and the result predicted by the model then instructor can pitch in and provide the feedback to the trainees to improve their trainings or learning behaviour of the trainee. Also, implementation of the proposed framework on the military shooting training exercise data has manifested that the selection of the modelling technique which will be used to build the model in the later steps should be an additional phase in the CRSIP DM model. This will help in saving the time and efforts in the next step of building the model.

7.2 Contribution

Wang et al. proposed the model for evaluating the trainee based on the knowledge a trainee has and his/her past behaviour. He collected the confidence knowledge of the trainees by interviewing them and by questionnaire. But, it is difficult to calculate the confidence level of the trainee as it is subjective in nature. It also makes the model applicable only for the trainees whose confidence level has been captured. Since the framework is proposed for military services, it is very difficult to interview all the trainees as the number of trainees seeking training is high. The proposed framework doesn't consider any sort of interaction with trainees to gather their performance data and is based only on the learning patterns of the trainees. The accuracy shown by the algorithms by implementing the proposed framework on the data taken from military shows that it can be used to evaluate the performance of the military trainees and their learning behaviour so that their learning capabilities can be improved without involving any sort of beforehand knowledge of their skills.

The proposed framework is an extension of the CRISP DM model. It suggests that since the data mining project has its success dependant on the performance of the data mining techniques, it is necessary to choose the suitable techniques among many available data mining techniques. Thus, selecting a modelling technique should be considered as one more step in addition to the six steps in CRISP DM model. Keeping selection of techniques as an additional step in the model accentuates that ample time should be spent in the selection criterion of the modelling techniques as building a model is an expensive task in terms of time and efforts. Therefore, applying appropriate techniques on the data and then checking their performance to select the most suitable technique is better than applying techniques randomly on the data and then discarding the techniques based on their performance as it will save lot of efforts in the model building phase.

This thesis work also evaluates the performance of the three classification methods used for predicting the outcome of the trainee by analysing their past learning behaviour. The classifiers are implemented on the dataset of the shooting exercise of the military and are compared based on root mean square error and time taken to build the model.

7.3 Future work

- For the future work, it will be helpful in getting more insight into the performance of the trainee if the coordinates at which trainee has made a shot can be captured into the data with the attributes which collect if the trainee has hit or missed the target so that more exhaustive guidance can be provided to the trainee. This will help in seeing how far the trainee from hitting the objective is.
- It has been anticipated that Neural Network algorithm's performance in predicting the result of the trainee has been impacted by the overfitting. Therefore, more data can be gathered to further investigate the performance of the algorithms.
- In this dissertation only three classification algorithms have been considered for building the model but if more data can be collected then other classification algorithms can also be applied to select the best algorithm that fits the dataset.
- The proposed framework is an extension of the CRSIP DM model which emphasizes on spending adequate effort on selecting modelling technique. The proposed framework has been applied to the current problem of building a framework for evaluating training of military personnel. But it will be good to see how it performs on the other data mining problems. Thus, thorough testing should be conducted on different kinds of data mining problems.
- Furthermore, expert systems can be integrated with the training evaluation system in future so that more comprehensive feedbacks can be provided to the trainees to improve the performance of the trainee.

REFERENCES

1. Amirah Mohamed Shahiri, Wahidah Husain Nur'aini Abdul Rashid (2005). A Review on Predicting Student's Performance Using Data Mining Techniques, Proceeding Computer Science, Volume 72, 2015, Pages 414-422
2. Andrews, D.H.,&Fitzgerald,P.C.(2010,May). Accelerating learning of competence and increasing long-term learning retention. Presented at the 2010 ITEC conference, London,UK.
3. Baxt WG (1995). Application of artificial neural networks to clinical medicine. *Lancet.*,346: 1135-8.
4. Bostanci, B., Bostanci, E. (2013). An evaluation of classification algorithms using Mc Nemar's test. *Advances in Intelligent Systems and Computing*, 201 AISC (VOL. 1), pp. 15-26.
5. C. Morbitzer, P. Strachan, and C. Simpson (2004). Data mining analysis of building simulation performance data,” *Building Services Engineers Res.Technologies*, vol. 25, no. 3, pp. 253–267.
6. Cannon-Bowers, J. A., & Bowers, C. A. (2009). Synthetic learning environments: On developing a science of simulation, games, and virtual worlds for training. In S. W. J. Kozlowski & E. Salas (Eds.). *Learning, training, and development in organizations* (pp. 229–261).
7. Carson K. Leung, Kyle W. Joseph (2014). Sports data mining: predicting results for the college football. 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES.
8. D. Kabakchieva (2013). Predicting Student Performance by Using Data Mining Methods for Classification. *Cybernetics and Information Technologies*, 13(1):61–72.
9. Daniel A. Keim and Florian Mansmann and Andreas Stoffel and Hartmut Ziegler (2009). Visual Analytics, *Encyclopedia of Database Systems*.

10. Day, E. A., Blair, C., Daniels, S., Kligyte, V., and Mumford, M.D. (2006). Linking instructional objectives to the design of instructional environments: The Integrative Training Design Matrix. *Human Resource Management Review*, 16(3), 376--395.
11. Ericsson, K.A., and A.C. Lehmann. (1996). Expert and Exceptional Performance: Evidence on Maximal Adaptations on Task Constraints. *Annual Review of Psychology* 47: 273_305.
12. Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. Roy. Statist. Soc. Ser. B* 70, 849-911.
13. Guoqiang Zhang, B. Eddy Patuwo, Michael Y. Hu (1998). Forecasting with artificial neural networks: The state of the art.
14. Guyon I, Gunn S, Nikravesh M, Zadeh L (2006) Feature extraction, foundations and applications. Springer, Heidelberg.
15. Guyon, I. and Elisseeff, A. (2003) An introduction to variable and feature selection. *J. Mach Learn Res.*, 3, 1157–1182.
16. Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed.). Morgan Kaufmann.
17. Hanrahan, Stolte, Mackinlay (2007). *Visual Analysis for everyone*.
18. HAO-YING JOANNE PENG, KUK LIDA LEE, GARY M. INGERSOLL (2002). An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research* 96(1):3-14
19. Harshit Topno (2012). Evaluation of Training and Development: An Analysis of Various Models. *IOSR Journal of Business and Management (IOSR-JBM)*, Volume 5, Issue 2, PP 16-22.
20. Issenberg, S. B., McGaghie, W. C., Petrusa, E. R., Lee, G. D., & Scalese, R. J. (2005). Features and uses of high-fidelity medical simulations that lead to effective learning: A BEME systematic review. *Medical Teacher*, 27(1), 10–28.
21. J Boulmetis and P. Dutin (2002). *The abc's of evaluation: Timeless techniques for program and project managers* (San Francisco, JosseyBass).
22. Juite Wang, Yung-I Lin b, Shi-You Hou (2015). A data mining approach for training evaluation in simulation-based training.
23. K. Ming Leung (2007). *Decision Trees and Decision Rules*.

24. Kalousis A, Prados J, Hilario M (2007) Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowl Inf Syst* 12(1):95–116.
25. Khoshgoftaar, T.M and Allen, E.B. (1999). Logistic regression modeling of software quality. *International Journal of Reliability, Quality and Safety Engineering*, vol. 6(4, pp. 303-317).
26. Klueger, A., and DeNisi, A. (1996). Effects of feedback intervention on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284.
27. Laurence, J.H., & Mathews, M.D. (Eds.). (2012). *The Oxford handbook of military psychology*. New York: Oxford University Press.
28. Laurens van der Maaten, Eric Postma, Jaap van den Herik (2009). *Dimensionality Reduction: A Comparative Review*.
29. Lippmann, R. (1987). An Introduction to computing with neural nets. *IEEE ASSP Magazine*, vol. (22).
30. Liu, H. and Motoda, H. (1998) *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, Norwell, MA.
31. Matthew N. Anyanwu, Sajjan G. Shiva (2009): Comparative Analysis of Serial Decision Tree Classification Algorithms. *International Journal of Computer Science and Security*, (IJCSS) Volume (3) : Issue (3)
32. Mayer, R.E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist*, 59(1), 14–19.
33. McGaghie, W. C., Issenberg, S. B., Petrusa, E. R., & Scalese, R. J. (2010). A critical review of simulation-based medical education research: 2003–2009. *Medical Education*, 44(1), 50–53.
34. Nicholson, D.M., Fidopiastis, C.M., Davis, L.D., Schmorow, D.D., & Stanney, K.M. (2007). An adaptive instructional architecture for training and education. *Foundations of Augmented Cognition*, 380–384.
35. Niels Landwehr, Mark Hall, Eibe Frank (2005). *Logistic Model Trees*.
36. Ola, A., and Pallaniappan, S. (2013): A data mining model for evaluation of instructors' performance in higher institutions of learning using machine learning

- algorithms, *International Journal of Conceptions on Computing and Information Technology* Vol. 1, sue 2; ISSN: 2345 - 9808 Methodology.
37. Parneet Kaur , Manpreet Singh and Gurpreet Singh Josan (2015): Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector. 3rd International Conference on Recent Trends in Computing 2015. vol. 57, 2015, Pages 500-508.
 38. Pappu, V.; Pardalos, P.M. (2014): High-Dimensional Data Classification. In *Clusters, Orders, and Trees: Methods and Applications* Springer: New York, NY, USA, 2014; pp. 119–150.
 39. P. Abdolmaleki , M. Yarmohammadi , M. Gity (2004).Comparison of logistic regression and neural network models in predicting the outcome of biopsy in breast cancer from MRI findings.
 40. Quinlan, J. R. (1993). *C45: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.
 41. R. Sharda (1994). Neural networks for the MS/OR analyst: An application bibliography. *Interfaces*, 24 (2) (1994), pp. 116–130.
 42. Riveiro, M., Gustavsson, P., Bengtsson, M., Blomqvist, P. and Wallinius, M. (2016 accepted). Enhanced Training through Interactive Visualization of Training Objectives and Models. 2016 NATO Modeling & Simulation Group (NMSG) Symposium, paper 1. Bucharest, Romania.
 43. Ross Quinlan (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.
 44. Salas, E., Priest, H.A., Wilson, K.A.,&Burke,C.S.(2006).Scenario-based training: Improving military mission performance and adaptability.InA.B. Adler, C.A.Castro,&T.W.Britt (Eds.). *The psychology of serving inpeace and combat operational stress* (Vol. 2).Westport,CT:Greenwood Publishing Group, Inc.
 45. Shuhui Li, Donald C. Wunsch, Edgar O’Hair and Michael G. Giesselmann (2001).Comparative analysis of regression and artificial neural network models for wind turbine power curve estimation. *Journal of Solar Energy Engineering*, 123, 327–332.
 46. Spitzer, D. R. (1999). Embracing evaluation. *Training*, 36(6), 42–47.

47. Stacy, W., & Freeman, J. (2016). Training objective packages: enhancing the effectiveness of experiential training. *Theoretical Issues in Ergonomics Science*, vol 17, no 2, pages 149--168.
48. Thomas, J. J., & Cook, K. (2005). Illuminating the path: the R&D agenda for visual analytics. IEEE.
49. Tom M. Mitchell, (1997). *Machine Learning*, Singapore, McGraw Hill.
50. Tsang S., Kao B., Yip K., Ho W. and Lee S (2009). Decision trees for uncertain data. In: *International Conference on Data Engineering (ICDE)*.
51. U.A. Kumar. (2005). Comparison of neural networks and regression analysis: A new insight. *Journal of Expert Systems with Applications*, Vol. 29, pp. 424-430.
52. U. F. Schalock (1998). *Outcome Based Evaluations* (Boston, Kluwer Academic/Plenum).
53. Wei Peng, Juhua Chen and Haiping Zhou (2009). An Implementation of ID3-Decision Tree Learning Algorithm. *Machine Learning University of New South Wales, School of Computer Science & Engineering, Sydney, NSW 2032, and Australia*.
54. Wirth R., Hipp J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Manchester, UK*, pp. 29–39.
55. Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Morgan Kaufmann.
56. Vogel---Walcutt, J.J.,Fiorella,L.,& Malone, N. (2013). Instructional strategies framework for military training systems. *Computers in Human Behavior*,29(4),1490--1498.
57. Y.S. Kim. (2008). Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size. *Journal of Expert Systems with Application*, Elsevier, pp. 1227-1234.
58. Zaidah Ibrahim, Daliela Rusli. (2007). Predicting student's academic performance: comparing artificial neural network, decision tree and linear regression. *21st Annual SAS Malaysia Forum, Shangri- La Hotel, Kuala Lumpur*. 21st Annual SAS Malaysia Forum, 5th September 2007, Shangri-La Hotel, Kuala Lumpur.

APPENDIX A

A.1 Setup, Feature Extraction, Model Creation (WEKA)

This appendix presents step-by-step instructions on how to apply data mining techniques to a set of captured data. The following experiment implements a filter method of feature selection technique and classification methods in WEKA.

A.1.1 Setup

- Download and install Weka from <http://www.cs.waikato.ac.nz/ml/weka/>

A.1.2 Feature Selection

1. Run Weka
2. From the Weka GUI Chooser, click on the Explorer button
3. From the Weka Explorer GUI, click on Open File option
4. Using explorer, open the filename.csv file
5. From the tabs above, click on “Select attributes”.
 - Attribute selection process selects the most relevant attributes according to the CSV file
6. Choose “FilteredAttributeEval” as an attribute evaluator.
 - Ranker method will automatically be chosen.
7. Use full training set as an attribute selection Mode.
8. Click on the Start button to start the method.

A.1.3 Model Creation

1. Run Weka
2. From the Weka GUI Chooser, click on the Explorer button
3. From the Weka Explorer GUI, click on Open File option

4. Using explorer, open the filename.csv file
5. Click on the Classify tab at the top of the Weka Explorer GUI
6. Click the “Choose” button to select a classifier
 - Expand the functions icon and select SimpleLogistic classifier to implement Regression method. or
 - Expand the functions icon and select MultilayerPerceptron classifier to implement Artificial Neural Network method. or
 - Expand the trees icon and select J48 Tree to implement Decision Tree method.
7. On the Classify GUI, click on the Start button to start the classifier.