Random Forest for Salary Prediction System to Improve Students' Motivation

Pornthep Khongchai
Department of Computer Science
Faculty of Science and Technology
Thammasat University
Pathumthani, Thailand
p.khongchai.22@gmail.com

Pokpong Songmuang
Department of Computer Science
Faculty of Science and Technology
Thammasat University
Pathumthani, Thailand
pokpong@cs.tu.ac.th

Abstract— **A salary prediction model was generated for graduate students using a data mining technique to generate for individuals with similar training attributes. An experiment was also conducted to compare the two data mining techniques Decision Trees ID3, C4.5 and Random Forest to determine the most suitable technique for salary prediction, tuned with key important parameters to improve the accuracy of the results. Random Forest gave the best accuracy at 90.50%, while Decision Trees ID3 and C4.5 returned lower accuracies at 61.37% and 73.96%, respectively for 13,541 records of graduate students using a 10-fold cross-validation method. Random Forest generated the best efficiency model for salary prediction. A questionnaire survey was conducted to determine usage evaluation with 50 samples. Results indicated that the system was effective in boosting students' motivation for studying, and also gave them a positive future viewpoint. The results also suggested that the students were satisfied with the implemented system since it was easy to use, and the prediction results were simple to understand without any previous background statistical knowledge.**

Keywords- Motivation; Salary prediction system; Educational data mining; Classification technique;Decision trees; Random Forest

I. INTRODUCTION

Most university students selected their degree courses without any future career goals, choosing a field of study following their friends or current trends. This caused boredom while studying which often led to failure through bad performance in the examinations [1]. Some students dropped out from university willingly while others were told to leave.

To increase motivation, a salary prediction system was developed based on a decision tree model with two features. The first identified examples of successful graduate students with a good career and salary. They were selected as role models to motivate current students to study hard and achieve their career plans and goals.

The second feature was the salary prediction system, where salary was the main factor for improving the quality of life, and also a tool for managing and setting stable goals. Salary was used as a means to motivate students and enhance their study performance. There are many factors for students to consider such as their GPA, enrollment history, English skills, and job training experience. These factors all affected the salary prediction model. Using the model, a salary prediction system [2,3] was implemented as a tool to show examples of students who followed successful careers, along with activities that they were involved with during student life.

Previous researchers [4] encountered problems with low accuracy of the Decision Tree. Therefore, Random Forest was applied instead of Decision Tree to enhance both the performance and accuracy of salary prediction and achieve higher efficiency. Random Forest was applied from Decision Tree as it is easy to understand and also more effective and accurate. The Random Forest algorithm is popular and widely used in research. Since the majority of the required learning processes from Random Forest were influenced by its own parameters, the model could be adjusted to determine individual results.

This research, therefore, compared algorithms for applied models to predict students' future salaries based on historical results by inputting details of individual student's activities and grades. Data mining techniques were exploited to build the salary prediction model. Several data mining techniques were tested and compared to find the best suited for the task.

The literature review is described in Section II and the proposed system is explained in detail in Section III. Section IV presents the experimental accuracy and efficiency results. Finally, Section V provides the conclusions with a summary of the findings.

II. LITERATURE REVIEW

Previous researchers used Decision Trees as a standard data mining technique to compare prediction efficacy [5,6].

The first algorithm as Decision Trees was the most popular. Some early examples include ID3 and C4.5 which were proposed by J. Ross Quinlan. Decision Trees are the most well-known and used machine learning methods [7-9]. A wide range of issues can be related to Decision Trees, from the core

978-1-5090-5698-9/16 $31.00 © 2016 IEEE
DOI 10.1109/SITIS.2016.106

algorithm for building an initial tree to methods for pruning, converting trees to rules, and handling various problems such as missing attribute values. Quinlan presented clear descriptions of the problems, usually accompanied by examples, and described how C4.5 handled them. The detailed examples were usually drawn from real data sets and greatly helped to illustrate each problem. Decision Tree creates a structure with a set of instances which can be used to classify new instances. Each instance is described by a set of attributes or features which can have numeric or symbolic values.

Over the past decade, the basic Decision Tree model proposed by Breiman has been improved and developed into the highly efficient Random Forest, a data mining techniques that has gained increasing popularity. An ensemble prediction method by aggregating the result as individual decision trees is a property of Random Forest. The general property of Random Forest is simulating a black box itself, with a few levers whose values can be altered and each of these levers has some effect on either the efficiency or the resource and time of the model [5].

To improve the accuracy of Random Forest, input parameter values were tuned to create a Random Forest model that gave out predictions [5]. A Random Forest for salary prediction was created based on Breiman's method.

The two main parameters for tuning Random Forest accuracy are the number of K trees to form a random forest and the number of F randomly sampled features for building a decision tree. Parameter K was set to 25 and parameter F was computed by $F = [\log_2 M + 1]$. For large and high dimensional data, where M is the number of inputs, a large K and F should be used [5,6]. Moreover, a technique called 'Ensemble' is used in many classification models to find answers. This technique has high efficacy for many competitors. The winner normally uses the Ensemble method in Random Forest. This creates various models from randomly selecting training data using the same classification models which result in many subsets. For example: even though the Decision Tree was used for all three models, different data were input which yielded different results. Similarly, after creating a bagging model with the Decision Tree technique [10], the next step used the created model to form a prediction as shown in Figures 1 and 2. This step outlined the training data procedures and used the model for the new data prediction.

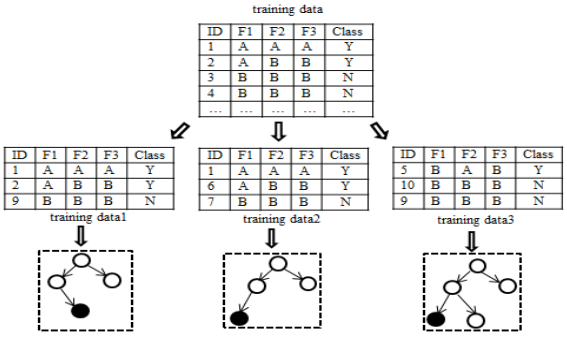


Figure 1. The process used the Random Forest model for individual tree classifiers

Then, step of the used model.

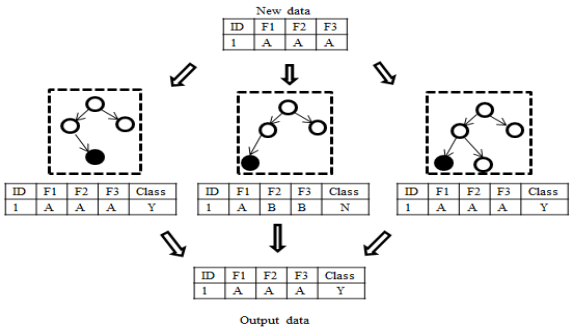


Figure 2. The process used the Random Forest model for new data classifiers

Figure 2 shows new data with one example of unknown classes of results:

a) Decision Tree (Model 1): The predictions of this model yielded the result Y from example1.
b) Decision Tree (Model 2): The predictions of this model yielded the result N from both examples.
c) Decision Tree (Model 3): The predictions of this model yielded the result Y from both examples.

From the predictions of all three models, the output data was Y, since two models had the same result as Y, even though one result yielded N.

However, data mining techniques have never been applied for student salary prediction [4]. Therefore, tuning parameter values were conducted and the performances of the data mining techniques for salary prediction were compared to determine the most suitable for creating a salary prediction system.

III. RANDOM FOREST FOR SALARY PREDICTION BASED ON DATA MINING TECHNIQUE USING TUNED PARAMETERS

Parameters in Random Forest either increase the predictive effectiveness of the model or make model training easier. The parameters have three primary features which can be tuned to improve the predictive effectiveness of the model. The maximum number of features Random Forest is allowed to try in individual trees.

1) The number of random trees. This parameter specifies the number of random trees to generate. Increasing the maximum number of features generally improves the performance of the model at each node, however, the speed of the algorithm decreases. To generate the optimal features and the right balance 25 random trees were specified.

2) Criterion. The criterion on which attributes are selected for splitting was chosen as gain ratio. This variant adjusts the information gain for each attribute to allow for breadth and uniformity of the attribute values.

3) Maximal depth. The depth of a tree varies depending on the size and nature of the instance set. This parameter is used to restrict the size of the Decision Tree. The tree generation process is not continued when the tree depth is equal to the

maximal depth. If its value is set to '-1', the maximal depth parameter puts no bounds on the depth of the tree and a tree of maximum depth is generated. If its value is set to '1', a tree with a single node is generated. The value of maximal depth was specified as 20 to fit the criterion and data set.

To create the Random Forest, Breiman used the bagging method to design training data subsets as trees which were then extended to forests. An accurate figure was used to assess the importance of trees which did not cause over-fitting problems and gave an unbiased result. Figure 2 shows the process using the Random Forest model for new data classifiers.

Similar to Breiman's method, the proposed method had one more key parameter as the number of P-trees selected to form an improved Random Forest. For large and high dimensional data more trees need to be selected to create an accurate salary prediction system.

IV. SALARY PREDICTION SYSTEM

The aim was to determine the salary of students based on their activities while at university by comparison with students who had graduated. Hopefully, knowing their predicted salary will urge them to pay more attention to studying with a clearer future direction and outcome. The data mining technique was a core component of the prediction system. Data for the graduates and their salaries were used to train a salary-activity model to predict a salary for the current students.

A. System Design

The system required two data inputs as follows:

1) Student profiles, and

2) Profiles of former student graduates and their salaries.

The three profiles of former students with the greatest similarity to those with the highest salary were represented as exemplars. Feature selection was applied to scale down the number of features to only those which significantly affected the prediction. The selected features for comparison were:

- Gender (Male, Female)
- Faculty (Engineering, Business, Arts, etc.)
- Program (Engineering Civil, Computer Science, Marketing, etc.)
- Job Training (Yes, No): did the student apply for job training?
- Certificate (Yes, No): did the student receive a certification in their study related field?
- GPA (>2.79, <2.8)
- Salary (four levels as classes: less than 13,500, 13,501 - 15,300, 15,301 - 18,000, and more than 18,000 baht): prediction result for current students and labels from those who had graduated.

B. Model Training

Many data mining techniques can be applied for salary prediction. However, five models were selected for comparison as follows:

Random Forest (RF) [5,9,10] is an ensemble classification method by voting the result of individual decision trees.

Decision trees (ID3,C4.5) [6-8] represents the hierarchical nature of a structure by node graphs using WEKA. This algorithm was applied to compare all experimental techniques.

The efficiencies of these two data mining techniques were compared for predicting salary, and the best result was installed in the system.

C. Apply rules for application

Trees were created from each data set. Each set of trees contained many rules such as the tree of A, B, and C as shown in Figure 3.

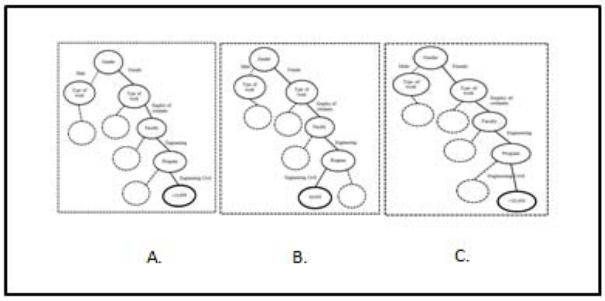


Figure 3. The process of creating the tree models and rules from the data set

The features from user tree A were then input shows the root node is Gender and has two branches that are left node as Male and right node as Female. Once Gender is focused on Female, the next node becomes node of Type of works. Once Type of works is focused on Employ of company, the next node becomes node of Faculty. Once Faculty is focused on Engineering, the next node becomes node of Program. Once Program is focused on Engineering Civil. The next node becomes leaf node of > 18,000 and compared to the rule set of tree A. The rules used were similar to a one rule system. The system acquired that particular rule, collected it, and then input the features from the user to compare with the rule set of tree B shows the root node is Gender and has two branches that are left node as Male and right node as Female. Once Gender is focused on Female, the next node becomes node of Type of works. Once Type of works is focused on Employ of company, the next node becomes node of Faculty. Once Faculty is focused on Engineering, the next node becomes node of Program. Once Program is focused on Engineering Civil. The next node becomes leaf node of 18,000 and tree C shows the root node is Gender and has two branches that are left node as Male and right node as Female. Once Gender is focused on Female, the next node becomes node of Type of works. Once Type of works is focused on Employ of company, the next node becomes node of Faculty. Once Faculty is focused on Engineering, the next node becomes node of Program. Once Program is focused on Engineering Civil. The next node becomes leaf node of > 18,000. The overall rules had similar sets of trees, but only one rule was chosen from each set. A set of rules was then created from the collected rules and voted on, with the majority then selected for application. Rules are shown in Figure 4.

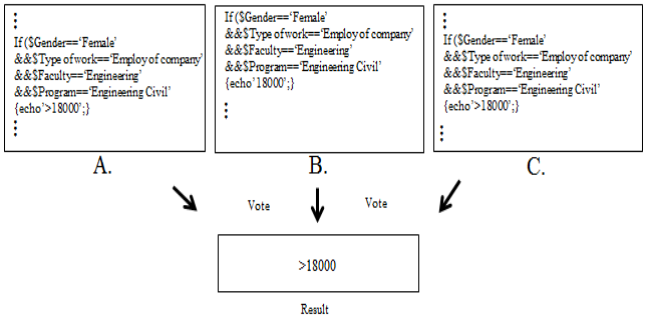


Figure 4. The process for the vote result

Many rules were complex in creating the Salary Prediction System.

D. *Usage of the Salary Prediction System*

The goal was to predict salaries to increase students' motivation. The system motivated the students through 1) salary prediction from student profiles, and 2) examples of three graduates with similar profiles and top-ranked salaries. Accordingly, users were able to understand the predicted results without the necessity of a statistical background.

For system input data, students were asked to submit their information as a profile. The user interface for data input is shown in Figure 5.

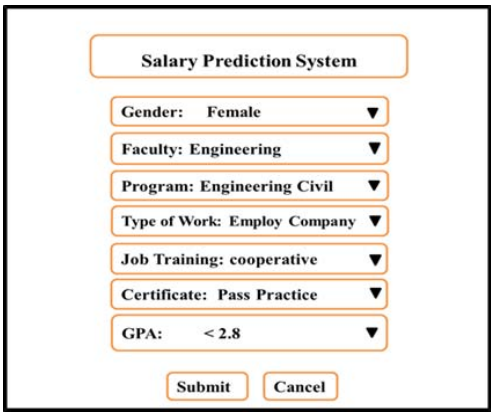
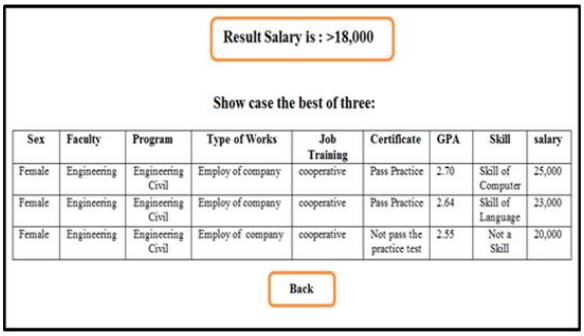


Figure 5. Interface of the Salary Prediction System

Figure 5 illustrates the interface for data input to the salary prediction system. The interface contained seven requested attributes from each student as gender, faculty, program, type of work, job training, certificate, and GPA. After data input and submission, the system compared the inputs of attributes with the rules and displayed the predicted salary of the three graduates whose attributes were the most similar to the input with the highest salary rates as shown in Figure 6.



Sex	Faculty	Program	Type of Works	Job Training	Certificate	GPA	Skill	salary
Female	Engineering	Engineering Civil	Employ of company	cooperative	Pass Practice	2.70	Skill of Computer	25,000
Female	Engineering	Engineering Civil	Employ of company	cooperative	Pass Practice	2.64	Skill of Language	23,000
Female	Engineering	Engineering Civil	Employ of company	cooperative	Not pass the practice test	2.55	Not a Skill	20,000

Figure 6. Salary Prediction Result

Figure 6 shows predicted salary as more than 18,000 baht. The salary prediction model compared the inputs of attributes with the profile in the database and selected three graduates with similar attributes and the highest salaries. The system then presented profiles of these three graduates as follows: gender, faculty, program, type of work, job training, certificate, GPA, skill and salary.The proposed salary prediction system showed potential in overcoming the problems of previous systems as 1) the salary prediction system predicted salaries of individual students, and 2) users easily understood the results from the salary prediction system.

V. EXPERIMENT AND RESULTS

To select the best data mining technique for the salary prediction system, an experiment was conducted to compare the six techniques mentioned in Section III A. Test data were used for students who had graduated between 2006 and 2015 from Rajamangala University of Technology Thanyaburi, Thailand. The historical data included gender, student faculty, student program, type of work, job experience training, certification, total Grade Point Average (GPA), salary, address, telephone number, and e-mail. In total, 13,541 profile records were used for training data.

A. *Data Preparation*

Records for graduates were accumulated over 10 years, and the base salaries increased relatively each year in economic value. Therefore, former salaries could not be used directly to create a salary prediction model for recent students. Linear equating techniques were applied to adjust previous salaries to current values, based on the hypothesis that the distribution of salaries was similar in every year. Firstly, step-wise data preparation was conducted, and salary prediction models were then created and used in the salary prediction system.

However, the data for graduates contained several missing values and excessive numbers of attributes for creating the salary prediction model. In addition, the data format required reformation since it was apparently unacceptable for the data mining tool. Data were therefore prepared following three steps:

1) Data Selection: Relevant attributes were selected for salary prediction, and forward selection [11] analysis was used to select 7 out of 108 total attributes. These 7 attributes were as follows: gender, faculty, program, type of work, job training, certificate, and GPA.

2) Data Cleaning: Outlier data (e.g. noticeably high salary) and missing values (e.g. no salary records) were manually removed. After data cleaning, the remaining data comprised 13,541 rows.

3) Data Transformation: This step prepared the data format as usable for the data mining tool. The salary data was changed from continuous to an interval format. User specific discretization [12] was applied, and the salaries were divided into four levels as classes: less than 13,500, 13,501 - 15,300, 15,301 - 18,000, and more than 18,000 baht.

B. Comparison of Data Mining Techniques

Data mining techniques were compared for predicting salary using the data mining tool WEKA (Waikato Environment for Knowledge Analysis) version 3.6 [13]. A 10-fold cross-validation technique was employed to evaluate the efficiency of the salary prediction model created using data described in the previous section. Table I presents the results of tuned Random Forest models and Table II presents the results of the salary prediction models in terms of Recall, Precision, F-measure, and Accuracy [14,15].

TABLE I. SUMMARY OF TUNE RANDOM FOREST MODELS

Number of trees	Maximal depth	Recall (%)	Precision (%)	F-measure (%)	Accuracy (%)
20	15	72.70	72.80	72.80	72.74
20	20	73	73	73	73
20	25	73.10	73.10	73.10	73.09
25	15	83.02	83.02	83.02	83
25	**20**	**90.50**	**90.50**	**90.50**	**90.50**
25	25	79	79	79	79
30	15	73.40	73.40	73.40	73.40
30	20	75	75	75	75
30	25	84.90	84.90	84.90	84.90

*BASED ON THE NUMBER OF ATTRIBUTES TO BE USED: KVALUE= 4 .

TABLE II. SUMMARY OF SALARY PREDICTION MODELS

Recall(%)		
ID3	C4.5	Random Forest (RF)
61.40	*74*	***90.50***
Precision (%)		
ID3	C4.5	Random Forest (RF)
61.40	*74*	***90.50***
F-measure (%)		
ID3	C4.5	Random Forest (RF)
61.40	*74*	***90.50***
Accuracy (%)		
ID3	C4.5	Random Forest (RF)
61.37	*73.96*	***90.50***

The results for the tuned Random Forest models are shown in Table I. Number of trees, Maximal depth, Recall, Precision, F-measure and Accuracy were based on a K value of 4 which was the most appropriate for 25 trees, the Maximal depth was 20, and the highest Recall result was 90.50%. Precision F-measure and Accuracy also showed high results at 90.50%. The prediction for each data mining technique is shown in Table II. For Recall, Precision, F-measure and Accuracy the highest overall prediction was 90.50% for Random Forest, with the lowest at 61.37% as Accuracy for Decision Tree (ID3).

For a small number of features, Random Forest showed higher efficiency than the other techniques and was also easy to implement. The result from the tuned parameters on the other hand, with a larger number of features required more complex techniques. Random Forest showed the best overall accuracy; hence, a model from Random Forest was selected for application in the salary prediction system.

C. Evaluation of Student Motivation

A questionnaire was designed in three parts to evaluate student motivation. The samples were 50 students from Rajamangala University of Technology Thanyaburi, Thailand. The same source used for the salary prediction system.

a) Part 1: General student information

Basic student information was collected including gender, age, faculty, and program. There were 26 males and 24 females with ages ranging from 19 to 22. The students came from 7 different faculties, and numbered 7, 15, 5, and 23 from years 1 to 4, respectively.

b) Part 2: Questions to examine student motivation before using the system.

The questions and results are shown in Table III.

TABLE III. QUESTIONS BEFORE USING THE SALARY PREDICTION SYSTEM AND RESULTS

Question No.	Question	Scores / Answers	Mean(SD)/Full scores
Q1	Which level of satisfaction do you have for your faculty/ program?	1. Least, 2. Little, 3.Moderate, 4.Much, 5.Most	3.8 (0.57) / 5
Q2	Does a high salary affect your study motivation?	1. Least, 2. Little, 3.Moderate, 4.Much, 5.Most	4.16 (0.47) / 5

c) Part 3: Questions to examine student motivation after using the system

The questions and results are shown in Table IV.

TABLE IV. QUESTIONS AFTER USING THE SALARY PREDICTION SYSTEM AND RESULTS

Question No.	Question	Scores / Answers	Mean(SD)/Full scores
Q3	How much satisfaction do you have with your current faculty/program?	1. Least, 2. Little, 3.Moderate, 4.Much, 5.Most	4.16 (0.55) / 5

Q4	How much does the system motivate you to study for your expected salary?	1. Least, 2. Little, 3.Moderate, 4.Much, 5.Most	4.18 (0.40) / 5

The mean score results in Q1 showed that students were entirely satisfied with their current faculty and program. Moreover, combining Q1 with Q3, the predicted salary increased students' satisfaction with their current faculty and program. Q4 results inferred that knowledge of predicted salary gave the students a much clearer goal and motivated them to study hard to achieve their expected salary.

VI. SUMMARY

This paper presented a salary prediction system using data mining techniques. The system was designed to support individual salary predictions by comparing the profiles of current students with graduates. Data mining techniques were compared for performance. An experiment was conducted comparing 13,541 data records of graduates by 10-fold cross-validation. Results indicated that Random Forest gave the best accuracy at 90.50% while Decision Tree (ID3) returned the lowest at 61.37%. Usage evaluation from 50 samples indicated that the system was helpful and increased student motivation to study, realize their plans and achieve their goals. Moreover, students mentioned the ease of system usage, and the prediction result was simple and comprehensible.

ACKNOWLEDGMENT

The authors express their thanks to Office of Academic Resource and Information Technology, Rajamangala University of Technology Thanyaburi, Thailand for providing the dataset of graduated students.

REFERENCES

[1] Lumsden and L.S., "Student Motivation To Learn". ERIC Digest, Number 92, 1994.

[2] Y.Lee and M.Sabharwal, "Education–Job Match, Salary, and Job Satisfaction Across the Public, Non-Profit, and For-Profit Sectors: Survey of recent college graduates", Public Management Review, 18:1,p 40-64, 2014.

[3] J.Jerrim, "Do college students make better predictions of their future income than young adults in the labor force?", Education Economics, 23:2, p 162-179, 2013.

[4] P.khongchai and P.Songmuang, "Improving Students'Motivation to Study using Salary Prediction System". 13th International Joint Conference on Computer Science and Software Engineering (Preprint), Khon Kaen,Thailand, 2016.

[5] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone.,"Classification and Regression Trees". Wadswart,Belmont, 1984.

[6] L. Breiman, "Random forests". Machine Learning, vol. 45, no. 1, pp. 5-32.2001.

[7] J.Ross Quinlan,"Induction of Decision Trees". Machine Learning 1:1, 81-106., 1986

[8] J.Ross Quinlan, "Unknown attribute values in induction". Proceedings of the Sixth International Machine Learning Workshop (pp. 164-168). San Mateo, CA: Morgan Kaufmann. ,1989.

[9] J.Ross Quinlan, "C4.5 Programs for Machine Learning". San Mateo, CA: Morgan Kaufmann.,1992.

[10] B. Xu, J.Li, Q.Wang, X.Chen, "A Tree Selection Model for Improved Random Forest", Bulletin of Advanced technology research, vol.6(2), 2012.

[10] T.G.Dietterich,"An Experimental Comparison of Three Methods for Constructing Ensombles of Decision Trees: Bagging,Boosting, and Randomization", Machine Learning, vol. 40, no. 2, pp. 139-157,2001.

[11] M.A.Hall and G.Holmes, "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining", IEEE Transactions on Knowledge and Data Engineering, vol.15,no. 3, 2003.

[12] P. Berka. and I. Bruha., "Discretization and grouping: preprocessing steps for Data Mining", Principles of Data Mining and Knowledge Discovery, Second European Symposium, PKDD '98, Nantes, France, 1998.

[13] Machine Learning Group at the University of Waikato, http://www.cs.waikato.ac.nz/ml/weka/downloading.html

[14] O.Villacampa., "Feature Selection and Classification Methods for Decision Making: A Comparative Analysis", College of Engineering and Computing Nova Southeastern University, 2015.

[15] G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification", Journal of Machine Learning Research 3, p 1289-1305, 2009.