

Proceedings of

2014 International Conference on Data and Software Engineering (ICODSE)

November 26th - 27th, 2014
Institut Teknologi Bandung
Bandung, Indonesia



IEEE Catalog Number : CFP14AWL-USB

ISBN : 978-1-4799-8177-9



Data and Software Engineering Research Group
School of Electrical Engineering and Informatics
Institut Teknologi Bandung



ISBN: 978-1-4799-8177-9

Proceedings of

**2014 International Conference
on Data and Software Engineering (ICODSE)**

Institut Teknologi Bandung, Indonesia

November 26th - 27th, 2014

2014 International Conference on Data and Software Engineering (ICODSE)

Copyright © 2014 by the Institute of Electrical and Electronics Engineers, Inc, All rights reserved.

Copyright and Reprint Permission

Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

Other copying, reprint or reproduction requests should be address to IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, P.O. Box 1331 Piscataway, NJ 08855-1331.

IEEE Catalog Number CFP14AWL-USB

ISBN 978-1-4799-8177-9

Additional copies of this publication are available from

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
+1 845 758 0400
+1 845 758 2633 (FAX)
email: curran@proceedings.com

General Chair's Message

Welcome to 2014 ICODSE.

It is a pleasure for us to host the 2014 International Conference on Data and Software Engineering (ICODSE) at Institut Teknologi Bandung (ITB) campus, Bandung, Indonesia. The conference is organized by Data and Software Engineering Research Group, School of Electrical Engineering and Informatics, Institut Teknologi Bandung. 2014 ICODSE is co-organized by Vienna University of Technology, Austria, and ASEA Uninet. 2014 ICODSE is also technically co-sponsored by IEEE Indonesia Section.

The 2014 ICODSE conference aims at uniting researchers and professionals in the domains of data and software engineering, presenting and discussing high-quality research results and outcomes in their fields. This year, the conference welcomes contributions for 2 tracks: Data Engineering and Software Engineering.

In total, we received 130 submissions of authors from 10 countries around the world. All submissions were peer-reviewed (blind) by at least three reviewers drawn from external reviewers and the committees, and as the result, 68 papers were accepted to be presented in this conference. These papers are in the proceedings of 2014 ICODSE.

Finally, as the General Chair of the Conference, I would like to express my deep appreciation to all members of the Steering Committee, Technical Programme Committee, Organizing Committee and Reviewers who have devoted their time and energy for the success of the event.

For all participants, I wish you an enjoyable conference in this beautiful city of Bandung.

Bayu Hendradjaya

General Chair of the ICoDSE2014

Committee

General Chair

Bayu Hendradjaya Institut Teknologi Bandung - Indonesia

Steering Committee

Benhard Sitohang Institut Teknologi Bandung - Indonesia
A Min Tjoa Vienna University of Technology - Austria
Ford Lumban Gaol IEEE - Indonesia
Iping Supriana Institut Teknologi Bandung - Indonesia
Richard Lai La Trobe Univesity - Australia

Technical Program Committee

Chair:

Wikan Damar Sunindyo Institut Teknologi Bandung - Indonesia

Members:

Adila Alfa Krisnadhi Wright State University - USA
Agung Trisetyarso Universitas Telkom - Indonesia
Ahmad Ashari Gadjah Mada University - Indonesia
Amin Anjomshoaa Vienna University of Technology - Austria
Ary Setijadi Prihatmanto IEEE Indonesia Computer Society Chapter
Arya Ardiansyah Eindhoven University of Technology - the Netherlands
Ayu Purwarianti IEEE Indonesia Education Activity Committee
Bayu Hendradjaya Institut Teknologi Bandung - Indonesia
Bernardo Nugroho Yahya Ulsan National Institute of Science & Technology - South Korea
Chitra Hapsari Ayuningtyas Alpen-Adria-Universitat Klagenfurt - Austria
Dade Nurjanah Telkom University - Indonesia
Dang Tran Khanh Ho Chi Minh City University of Technology - Vietnam
David Taniar Monash University - Australia
Dwi Hendratmo Widiyantoro Institut Teknologi Bandung - Indonesia
Erich Neuhold IFIP Technical Committee on Information Technology Application
Estefania Serral Asensio KU Leuven, Belgium
Fathul Wahid Universitas Islam Indonesia - Indonesia
Fazat Nur Azizah Institut Teknologi Bandung - Indonesia
GA Putri Saptawati Institut Teknologi Bandung - Indonesia
Gerald Quirchmayr University of Vienna - Austria
Josef Küng Johannes Kepler University of Linz - Austria

Khabib Mustofa	Gadjah Mada University - Indonesia
Ladjel Bellatreche	Laboratoire d'Informatique Scientifique et Industrielle - France
Maguelonne Teisseire	University of Montpellier 2 - France
Michel Hassenforder	UHA - France
MM Inggriani	Institut Teknologi Bandung - Indonesia
Muhammad Asfand-e-yar	Masaryk University - Czech Republic
Muhammad Ilyas	University of Sargodha - Pakistan
Nguyen Duy Thanh	Ho Chi Minh City University of Technology - Vietnam
Robert P. Biuk-Aghai	University of Macau - PR China
RV Hari Ginardi	Institut Teknologi Sepuluh Nopember - Indonesia
Saiful Akbar	Institut Teknologi Bandung - Indonesia
Siti Rochimah	Institut Teknologi Sepuluh Nopember - Indonesia
Soleh Udin Al Ayubi	Boston Children's Hospital - USA
Stephane Bressan	National University of Singapore - Singapore
Vladimír Mařík	Czech Technical University in Prague - Czech Republic
Wenny Rahayu	La Trobe University - Australia
Wichian Chutimaskul	King Mongkut's University of Technology Thonburi - Thailand
Wikan Dinar Sunindyo	Institut Teknologi Bandung - Indonesia
Yan Tang	European Space Agency - Netherland
Yudho Giri Sucahyo	University of Indonesia - Indonesia
Yudistira D. W. Asnar	Institut Teknologi Bandung - Indonesia
Zainal Arifin Hasibuan	University of Indonesia - Indonesia

Organizing Committee

Adi Mulyanto	Institut Teknologi Bandung - Indonesia
Christine Suryadi	Institut Teknologi Bandung - Indonesia
Hira Laksmiwati Soemitro	Institut Teknologi Bandung - Indonesia
Riza Satria Perdana	Institut Teknologi Bandung - Indonesia
Tricya E. Widagdo	Institut Teknologi Bandung - Indonesia
Yani Widayani	Institut Teknologi Bandung - Indonesia

Reviewers

A. Imam Kistijantoro	Institut Teknologi Bandung - Indonesia
Adila Krisnadhi	Wright State University - USA
Aditya Dakur	Hubvents VIT University - India
Agung Trisetyarso	Telkom University - Indonesia
Amin Anjomshoaa	Vienna University of Technology -Austria
Anto Nugroho	BPPT - Indonesia
Artem Lenskiy	Korea University of Technology and Education - South Korea
Ary Prihatmanto	Institut Teknologi Bandung - Indonesia
Arya Adriansyah	Eindhoven University of Technology - the Netherlands
Ayu Purwarianti	Institut Teknologi Bandung - Indonesia
Bayu Hendradjaya	Institut Teknologi Bandung - Indonesia
Benhard Sitohang	Institut Teknologi Bandung - Indonesia
Bernaridho Hutabarat	UPI YAI - Indonesia
Dade Nurjanah	Telkom University - Indonesia
David Taniar	Monash University - Australia
Dessi Puji Lestari	Institut Teknologi Bandung - Indonesia
Dicky Prima Satya	Institut Teknologi Bandung - Indonesia
Dwi Hendratmo Widyantoro	Institut Teknologi Bandung - Indonesia
Estefania Serral Asensio	KU Leuven - Belgium
Fathul Wahid	Universitas Islam Indonesia - Indonesia
Fazat Nur Azizah	Institut Teknologi Bandung - Indonesia
G. A. Putri Saptawati	Institut Teknologi Bandung - Indonesia
Hira Laksmiwati	Institut Teknologi Bandung - Indonesia
Indriana Hidayah	Gadjah Mada University - Indonesia
Iping Supriana Suwardi	Institut Teknologi Bandung - Indonesia
Jamaiah Yahaya	The National University of Malaysia - Malaysia
Ketut Wikantika	Institut Teknologi Bandung - Indonesia
Khabib Mustofa	Gadjah Mada University - Indonesia
Lukman Heryawan	Gadjah Mada University - Indonesia
Maguelonne Teisseire	University of Montpellier 2 - France
Mahmoud Neji	Sfax University - Tunisia
Maman Fathurrohman	Universitas Sultan Ageng Tirtayasa - Indonesia
Masayu Leylia Khodra	Institut Teknologi Bandung - Indonesia
Mewati Ayub	Maranatha Christian University - Indonesia
MM Inggriani Liem	Institut Teknologi Bandung - Indonesia
Muhammad Asfand-e-yar	Masaryk University - Czech Republic
Noor Maizura Mohamad Noor	Universiti Malaysia Terengganu - Malaysia
Nur Ulfa Maulidevi	Institut Teknologi Bandung - Indonesia

Oerip Santoso	Institut Teknologi Bandung - Indonesia
Ponmagal Rajendran	M G R Educational and Research Institute University - India
Rajesri Govindaraju	Institut Teknologi Bandung - Indonesia
Restya Astari	Institut Teknologi Bandung - Indonesia
Richard Lai	La Trobe University - Australia
Richki Hardi	University of Ahmad Dahlan - Indonesia
Rila Mandala	Institut Teknologi Bandung - Indonesia
Rinaldi Munir	Institut Teknologi Bandung - Indonesia
Robert P. Biuk-Aghai	University of Macau - PR China
Saiful Akbar	Institut Teknologi Bandung - Indonesia
Siti Rochimah	Institut Teknologi Sepuluh Nopember - Indonesia
Soleh Udin Al Ayubi	Boston Children's Hospital - USA
Soon Chung	Wright State University - USA
Suhardi Suhardi	Institut Teknologi Bandung - Indonesia
Thanh D. Nguyen	Ho Chi Minh City University of Technology - Vietnam
Thomas Basuki	Universitas Parahyangan - Indonesia
Tran Khanh Dang	Ho Chi Minh City University of Technology - Vietnam
Tricya Widagdo	Institut Teknologi Bandung - Indonesia
Wikan Dinar Sunindyo	Institut Teknologi Bandung - Indonesia
Yani Widyani	Institut Teknologi Bandung - Indonesia
Yudistira Dwi Wardhana Asnar	Institut Teknologi Bandung - Indonesia

Contents

Data Engineering Track

A Method for Automated Document Classification Using Wikipedia-Derived Weighted Keywords	1
<i>Robert P. Biuk-Aghai, Ka Kit Ng</i>	
A New Direct Access Framework for Speaker Identification System	7
<i>Hery Heryanto, Saiful Akbar, Benhard Sitohang</i>	
A New Scheme to Hide the Data Integrity Marker on Vector Maps Using A Feature-Based Fragile Watermarking Algorithm	12
<i>Shelvie Nidya Neyman, Yudhy Haryanto Wijaya, Benhard Sitohang</i>	
A Semantic Framework for Data Integration and Communication in Project Consortia	18
<i>Fajar J. Ekaputra, Estefania Serral, Dietmar Winkler, Stefan Biffi</i>	
Analysis and Modeling of Sequential Pattern as Multimedia Data Representation	24
<i>Dini Nurmalasari, Gusti Ayu Putri Saptawati</i>	
Application Program Interface to Build Executive Information System using Data Warehouse	30
<i>Mario Orlando Teng, Tricya Esterina Widagdo</i>	
Computing Preset Dictionaries from Text Corpora for the Compression of Messages	35
<i>Marc W. Abel, Soon M. Chung</i>	
CORRELATION ANALYSIS OF USER INFLUENCE AND SENTIMENT ON TWITTER DATA	40
<i>Fadhli Mubarak bin Naifa Hanif, G. A. Putri Saptawati</i>	
Data Migration Helper Using Domain Information	46
<i>Irfan Kamil, M. M. Inggriani, Yudhistira Dwi Wardhana Asnar</i>	
Detection of Potential Traffic Jam Based on Traffic Characteristic Data Analysis	52
<i>Anasthasia Amelia, G. A. Putri Saptawati</i>	
Exploration of Classification Using NBTree for Predicting Students' Performance	57
<i>Tjioe Marvin Christian, Mewati Ayub</i>	
Fostering Government Transparency and Public Participation through Linked Open Government Data	63
Case Study: Indonesia Public Information Service	
<i>Peb R. Aryan, Fajar J. Ekaputra, Wikan D. Sunindyo, Saiful Akbar</i>	
Handwriting Recognition Using a Combination of Structural Elements Similarity	69
<i>Mastur Jaelani, Iping Supriana</i>	

Improving Classification Performance by Extending Documents Terms <i>Widodo, Wahyu Catur Wibowo</i>	75
Information Extraction of Public Complaints on Twitter Text For Bandung Government <i>Dekha Anggareska, Ayu Purwarianti</i>	80
Information Extractor for Small Medium Enterprise Aggregator <i>Fabrian Oktavino H., Nur Ulfa Maulidevi</i>	86
Integration of Search Engine and Recommender System for the Holy Qur'an Personalization <i>Lukman Heryawan</i>	91
Java Archives Search Engine Using Byte Code as Information Source <i>Oscar Karnalim, Rila Mandala</i>	97
Modeling of Coastal Upwelling Using Spatiogram and Structuring Elements <i>Yus Sholva, Benhard Sitohang, Ketut Wikantika</i>	103
Modeling Unpredictable Data and Moving Object in Disaster Management Information System based on Spatio-Temporal Data Model <i>Hira Laksmiwati, Yani Widyani, Nisa'ul Hafidhoh, Atika Yusuf</i>	108
Natural Language Interfaces to Database (NLIDB): Question Handling and Unit Conversion <i>Filbert Reinaldha, Tricya E. Widagdo</i>	114
On Analyzing of Fingerprint Direct-Access Strategies <i>G. Indrawan, S. Akbar, B. Sitohang</i>	120
Predicting Information Cascade on Twitter Using Support Vector Regression <i>Irfan Aris Nur Hakim, Masayu Leylia Khodra</i>	126
Predicting Latent Attributes by Extracting Lexical and Sociolinguistics Features from User Tweets <i>Muhammad Afif Al hawari, Masayu Leylia Khodra</i>	132
Predictions based on Twitter - A Critical View on the Research Process <i>Lisa Madlberger, Amal Almansour</i>	137
Rule based Approach for Text Segmentation on Indonesian News Article using Named Entity Distribution <i>Saniati, Ayu Purwarianti</i>	143
Semantic Web-based Aggregation of Indonesian Open Development Data <i>Tubagus Andhika Nugraha, Windy Gambetta</i>	148
Sentence Extraction in Recognition Textual Entailment Task <i>Yudi Wibisono, Dwi H. Widyantoro, Nur Ulfa Maulidevi</i>	154
Smart Buildings: Semantic Web Technology for Building Information Model and Building Management System	159

Muhammad Asfand-e-yar, Adam Kucera, Tomáš Pitner

Spatial Data Model for Corporate Based on Google Maps Platform 165
Iping Supriana Suwardi, Dessi Puji Lestari, Dicky Prima Satya

System Information Log Visualization to Monitor Anomalous User Activity Based on Time 171
Jeremy Joseph Hanniel, Tricya E. Widagdo, Yudistira D. W. Asnar

The Application of Rough Set and Fuzzy Rough Set Based Algorithm to Classify Incomplete Meteorological Data 177
Winda Aprianti, Imam Mukhlash

Total Information Quality Management-Capability Maturity Model (TIQM-CMM): An Information Quality Management Maturity Model 183
Suhardi, I Gusti Ngurah Rama Gunawan, Ardani Yustriana Dewi

Two-stage feature extraction to identify Plasmodium ovale from thin blood smear microphotograph 189
Anto Satriyo Nugroho, Made Gunawan, Vitria Pragesjvara, Miranti Jatnia Riski, Desiani, Inas Ashilah, Isma Hariani, Ismail Ekoprayitno Rozi, Puji Budi Setia Asih, Umi Salamah, Esti Suryani

Unpredictable Data and Moving Object Handling Prototype Architecture Using Spatio-Temporal DBMS 193
Nisa'ul Hafidhoh, Atika Yusuf, Hira Laksmiwati, Yani Widyani

Using Dictionary in a Knowledge Based Algorithm for Clustering Short Texts in Bahasa Indonesia 198
Husni Thamrin, Atiqa Sabardila

Using Twitter Data to Improve News Results on Search Engine 202
Abraham Krisnanda Santoso, G. A. Putri Saptawati

Software Engineering Track

A New Proposal for The Integration of Key Performance Indicators to Requirements Elicitation Process Originating from Organization Goals 207
Fransiskus Adikara, Bayu Hendradjaya, Benhard Sitohang

A Quality Model for Mobile Thick Client that Utilizes Web API 213
Hanny Fauzia, Hira Laksmiwati, Bayu Hendradjaya

A Vulnerability Scanning Tool for Session Management Vulnerabilities 219
Raymond Lukanta, Yudistira Asnar, A. Imam Kistijantoro

A/B Test Tools of Native Mobile Application 225
Muhammad Adinata, Inggriani Liem

Academic Information System Quality Measurement Using Quality Instrument: A Proposed Model 231
Umi Laili Yuhana, Agus Budi Raharjo, Siti Rochimah

An Application Framework for Evaluating Methods in Biometrics Systems <i>Satria Ardhe Kautsar, Saiful Akbar, Fazat Nur Azizah</i>	237
Android Security Assessment Based on Reported Vulnerability <i>Eko Sugiono, Yudistira Asnar, Inggriani Liem</i>	243
Automatic Grader for Programming Assignment Using Source Code Analyzer <i>Susilo Veri Yulianto, Inggriani Liem</i>	249
Business Process Extraction for Information Technology/Business Alignment <i>Azmat Ullah, Richard Lai</i>	253
Case Study on Semantic Clone Detection Based on Code Behavior <i>Bayu Priyambadha, Siti Rochimah</i>	259
Case-based Reasoning Approach for Form Interface Design <i>Dwi H. Widyantoro, U. Ungkawa, Bayu Hendradjaya</i>	265
Component Design of Business Process Web Content Management System for Online Shop Website <i>Rizal Panji Islami, Adi Mulyanto</i>	271
Design of a Tool for Generating Test Cases from BPMN <i>Prat Yotyawilai, Taratip Suwannasart</i>	278
Developing Translation Rules of Java-JML Source Code to Event-B <i>Faisal Hadiputra, Yudistira D. W. Asnar, Bayu Hendradjaya</i>	284
Development of Game Testing Method for Measuring Game Quality <i>Rido Ramadan, Bayu Hendradjaya</i>	290
Educational platform for learning programming via controlling mobile robots <i>Artem Lenskiy, Heo Junho, Kim Dongyun, Park Junsu</i>	296
Efficiency Measurement of Java Android Code <i>Nugroho Satrijandi, Yani Widayani</i>	300
Improvement of Adaptable Model Versioning (AMOR) Framework for Software Model Versioning Using Critical Pair Analysis <i>Ni Made Satvika Iswari, Fazat Nur Azizah</i>	305
Input Injection Detection in Java Code <i>Edward Samuel Pasaribu, Yudistira Asnar, M. M. Inggriani Liem</i>	310
Integrating Cognitively-oriented and Pedagogically-oriented Methods in Adaptive Educational Hypermedia <i>Reza Akhmad Gandara, Dade Nurjanah, Rimba Whidina Ciptasari</i>	316
Managing Requirements Change in Global Software Development <i>Naveed Ali, Richard Lai</i>	322
Measuring the Constraint Complexity of Automotive Embedded Software Systems <i>Mohit Garg, Richard Lai</i>	327

Modeling The Requirements for Big Data Application Using Goal Oriented Approach	333
<i>Hanif Eridaputra, Bayu Hendradjaya, Wikan Danar Sunindyo</i>	
Multi-agent Sentiment Analysis using Abstraction-based Methodology	339
<i>Timotius Kevin Levandi, M. M. Inggriani, Nur Ulfa Maulidevi</i>	
OBD-II Standard Car Engine Diagnostic Software Development	345
<i>Alex Xandra Albert Sim, Benhard Sitohang</i>	
Software Modularization in Global Software Development	350
<i>Dilani Wickramaarachchi, Richard Lai</i>	
Software Reliability Model Selection for Component-Based Real-Time Systems	356
<i>Mohit Garg, Richard Lai</i>	
Software with Service Oriented Architecture Quality Assessment	362
<i>Aminah Nuraini, Yani Widyani</i>	
Vendor Capability for Offshore IT Projects: Analysing a case in Indonesian Context	368
<i>Rajesri Govindaraju, Leksananto Gondodiwiryo, Reza Andhika Zairul</i>	
Verifying UML-based Interaction Using Coloured Petri Nets	373
<i>Aditya Bagoes Saputra, Thomas Anung Basuki, Jimmy Tirtawangsa</i>	
Visually Scripting Portable BPMN Script Tasks	379
<i>Jessada Wiriyakul, Twittie Senivongse</i>	

Exploration of Classification Using NBTree for Predicting Students' Performance

Tjioe Marvin Christian

Department of Informatics Engineering
Faculty of Information Technology, Maranatha Christian
University
Bandung, Indonesia
e-mail: t.marvin.christian@gmail.com

Mewati Ayub

Department of Informatics Engineering
Faculty of Information Technology, Maranatha Christian
University
Bandung, Indonesia
e-mail: mewati.ayub@it.maranatha.edu

Abstract— The growth of academic data size in higher education institutions increases rapidly. This huge volume of data collection from many years contains hidden knowledge, which can assist the improvement of education quality and students performance. Students' performance is affected by many factors. In this study, the data used for data mining were students' personal data, education data, admission data, and academic data. NBTree classification technique, one of data mining methods, was adopted to predict the performance of students. Several experiments were performed to discover a prediction model for students' performance. The class labels of students' performance were students' status in study, graduates predicates, and length of study. The experiments were conducted with two-level classification, the university level and faculty level. The resulted model indicated that some attributes had significant influence over students' performance.

Keywords—classification; NBTree; student performance.

I. INTRODUCTION

In conjunction with the increase of huge volume of daily data collection, data mining has served as the tool for analyzing large amounts of data. Data mining has been expanded from not only to analyze financial data, retail industries data, recommender system data, and intrusion detection data, but also to analyze data in higher education [1][2].

Data mining tools have been the subjects of research in analyzing higher education data. Ranjan [3] proposed a framework to help academic institution to utilize hidden knowledge in historical academic data to improve education quality. In [4], Bhardwaj and Pal use Naïve Bayes classification to divide students based on their academic performance. Kasih, Ayub, and Susanto [5] utilize Apriori algorithms to predict students final passing results based on their performance in several subjects.

Classification as a supervised learning technique has been used in predicting new data to be classified based on training dataset. Model resulted from classification can be utilized to predict future data trends. The commonly used classification algorithms are decision tree, and Bayesian classifier. Other than those, there is an NBTree classification, one of the classification algorithms that combines decision tree classifiers and Naïve Bayes classifiers [6].

This paper describes exploration of data mining concepts, especially a classification using NBTree. The objective of this study is to build a model using NBTree to predict students' performance. Dataset used to predict students' performance during their study consisted of personal data, education data, admission data, and academic data.

This research is the extension from previous studies [7][8], the aim of which is to design data warehouse schema for academic data. The academic data schema consists of data mart schema for student and data mart schema for lecturer. The schema is able to become a basis for data mining analysis of academic data to obtain meaningful knowledge that can be used to improve education quality.

II. LITERATURE STUDY

A. Data Classification

Data classification is defined as a predictive methods in data mining that is used to classify unseen data [1][9]. There are two main steps in data classification, namely learning step and classification step. In learning step, a classification model is built using an algorithm on a training set. Training set used for learning step must have class labels for given data. After a classifier model is built, it is utilized for predicting class labels for unseen data.

NBTree is one of classification methods that was introduced by Ron Kohavi [6]. It is a hybrid algorithm from Naïve Bayes and decision tree combined. This algorithm is similar with decision tree except in its leaf. The decision tree has a branch from recursive process, and the leaves are from the Naïve Bayes classifier, not a node that contains final result from a class.

Fig. 1 shows the pseudo code from NBTree algorithm. When data reach a node, five-fold cross-validation using Naïve Bayes will be performed on each attribute. The split of attributes will be considered, based on the error rate that resulted from Naïve Bayes classifier in that node.

NBTree algorithm generates a decision tree that looks like in Fig. 2. The node in ellipsis is an attribute that will split dataset into two or more groups. The node in square is a leaf that classified by Naïve Bayes classifier [10].

Input: a set T of labelled instances

Output: a decision-tree with Naïve-Bayes categorizes at the leaf

Algorithm:

1. Foreach attribute X_i , evaluate the utility $u(X_i)$, of a split on attribute X_i . For continuous attributes, a threshold is also found at this stage.
2. Let $j = \arg \max_i$, *i.e.*, the attribute with the highest utility.
3. If u_j is not significantly better than the utility of the current node, create a Naïve-bayes Classifier for the current node and return.
4. Partition T according to the test on X_j . If X_j is continuous, a threshold split is used; if X_j is discrete, a multi-way split is made for all possible values.
5. Foreach child, call the algorithm recursively on the portion of T that matches the test leading to the child.

Fig. 1. NBTree pseudo code by Kohavi [6]

For each leaf in NBTree, there is a Naïve Bayes classifier described in Table I. The first column shows attributes and nominal values for each attribute. The other columns show class values and frequency counts (FC) of nominal values or parameters of normal distributions for numeric attributes [10].

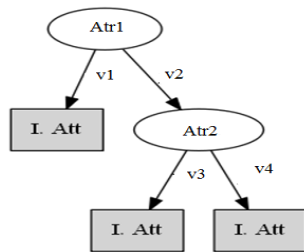


Fig. 2. Decision Tree Example for NBTree

To derive rules from a decision tree, one rule is generated for each path from the root to a leaf node. The rule antecedent is formed from each splitting condition along a given path, and the rule consequent is the class assigned by the leaf [1][10].

TABLE I. NAÏVE BAYES CLASSIFIER

Atr - i	Class = 1	Class = 2	...	Class = m
val-1	FC11	FC21		FCm1
val-2	FC12	FC22		FCm2
...				
val-n	FC1n	FC2n		FCmn

B. Evaluation

The evaluation was performed by building model in the classification with training data and test data. Training data was processed using the classification algorithm to build a classifier. To measure the performance of resulted classifier, test data was used to calculate the error rates. Training data and test data must be disjointed to ensure credibility of classifier evaluation [10].

The problem in the employed evaluation method is the availability of sufficient data set to be divided as training data and test data. This research employed the stratified tenfold

cross validation for the evaluation to overcome the limitation of data availability [10]. To implement this, the data were divided into ten parts randomly in nearly the same size. Each part was held out in turn and the nine-tenths were trained using classification algorithms. Afterward, the error rates were calculated. Thereby, the algorithm was executed ten times on different training sets. At last, the overall error rate was resulted from the average of ten error estimates.

III. METHODOLOGY

The research methodology used in this study consists of three steps as follow:

A. Data Preparation

This data set was built from student data mart schema that resulted from prior studies [7][8] as a data resource. The data set contains students' data from two faculties of a university. This study used two kinds of data set, the first was active students' data set and the second was graduates data set. Preprocessing was performed on raw data by removing outliers, resolving inconsistencies, and transforming data to obtain qualify data that ready for classification.

B. Data selection and transformation

As described in Table II, a group of attributes has been selected for active students' classification. These attributes consist of (a) personal information such as: gender and students home town, (b) education information such as: high school major, (c) admission information such as: type of admission phase and admission test score, (d) academic information such as: faculty, department and GPA. All attributes were utilized to predict students' status to be active (1) or drop out (2).

TABLE II. ACTIVE STUDENTS DATA SET

Attribute name	Description	Possible Values
Gender	Students Gender	[M=Male, F=Female]
Faculty	Students Faculty	[E=Faculty E, T= Faculty T]
Department	Students Department	[T=Department T, S= Department S]
Admission	Type of admission phase	[A1= Admission Phase 1, A2 = Admission Phase 2]
Test Score	Admission test score	[1 : excellent, 2 : good, 3 : fair, 4 : acceptable]
City	Students home town	[A = Bandung, B = outside Bandung]
Major	High school major	[A = major A, S = major S]
GPA	Grade point average	[1 : GPA = below, 2: GPA = good, 3 : GPA = excellent]
Status	Status of student	[1 = active, 2 = drop out]

For graduates' classification, a group of attributes has been selected as described in Table III. These attributes consist of (a) personal information such as: gender and students home town, (b) education information such as: high school major, (c) admission information such as: type of admission phase and admission test score, (d) academic information such as: faculty, department, and total of credit. These attributes were used for two kind of classification. The first classification is predicting

students GPA to be satisfactory (1), excellent (2), or cum laude(3). The second classification is predicting length of students study to be on time (1), or late (2).

TABLE III. GRADUATES DATA SET

Attribute name	Description	Possible Values
Gender	Gender	[M=Male, F=Female]
Faculty	Faculty	[E=Faculty E, T= Faculty T]
Department	Students Department	[T=Department T, S= Department S]
Admission	Type of admission phase	[A1= Admission Phase 1, A2 = Admission Phase 2]
Test Score	Admission test score	[1 : excellent, 2 : good, 3 : fair, 4 : acceptable]
City	Students home town	[A = Bandung, B = outside Bandung]
Major	High school major	[A = major A, S = major S]
Credit	Credit total	[C1 : Credit = 144, C2 : Credit > 144]
GPA	Grade point average	[1 : GPA = satisfactory, 2 : GPA = excellent, 3 : GPA = cum laude]
LST	Length of study	[1 : on time (LST ≤ 4), 2 : late (LST > 4)]

C. Classification using NBTree

This study performed NBTree classification on active students' data set and graduates data set using WEKA as a data mining toolkit [10]. For each data set, the classification was done at university level and faculty level. Data set at the university level was represented by data samples from two faculties. Data set at the faculty level was represented by data samples from faculty T.

In this study, the proposed model for prediction students' performance consisted of two parts. The first predicted students' performance through students' status, active or drop out. The second utilized graduates' GPA and LST to predict students' performance.

Classification for active students' data set used students' status attribute as the class. The attributes used for students' status prediction at university level consisted of Gender, Faculty, Admission, Test Score, City, Major, and GPA. At faculty level, we used Gender, Department, Admission, Test Score, City, Major, and GPA for prediction.

For graduates' data set, this study executed two kinds of classification. The first used GPA as the class attribute, and the second used LST as the class attribute. For both prediction, we used the same dataset, which at university level consisted of Gender, Faculty, Admission, Test Score, City, Major, and Credit. At faculty level, we used Gender, Department, Admission, Test Score, City, Major, and Credit.

IV. EXPERIMENT : THE RESULT AND INTERPRETATION

After the data had been prepared, the classification model construction was performed. In each of these experiments explained below, a tree was built using NBTree technique. In classification for university level, attribute faculty was used instead of attribute department. Attribute department was used in classification for faculty level. Data set for faculty T was used in classification for faculty level.

To validate the performance of each experiment, this study utilized ten fold cross validation because of the limitation of data availability.

A. Experiment-1(E1) : Active Students Data set

Active students' data set in Table II with Status as the class label for experiment-1 in university level (E1-U) consisted of 2687 instances. Fig. 3 shows the tree with ten leaves resulted from NBTree classification for the data set. The accuracy percentage for predicting performance in E1-U is 81.46%.

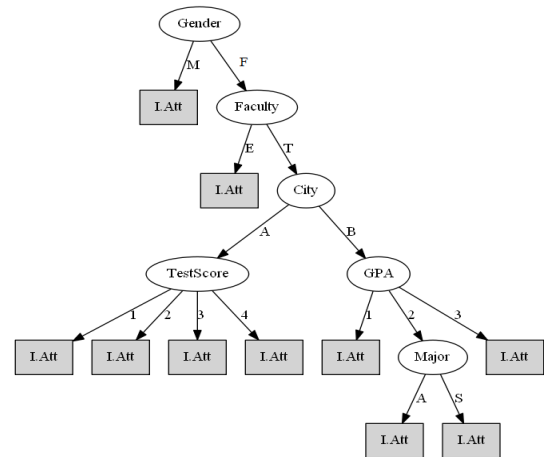


Fig. 3. NBTree for E1-U

The correct classified data from E1-U consisted of 2145 instances classified as active and 44 instances classified as drop out. Based on the correct classified data in the NBTree in Fig. 3, the meaningful rules generated from the tree are shown in Table IV.

Table IV indicates that Gender attribute was the most affective attribute in determining active students at university level. Generally, Male students have better performance than Female. But Female students in faculty T and came from Bandung have better performance than Male. So also the students in Faculty T, came from outside Bandung, and have good GPA or excellent GPA.

Active students' data set in Table II with Status as the class label for experiment-1 in faculty level (E1-F) consists of 875 instances. Fig. 4 shows the tree with seven leaves resulted from NBTree classification for the data set. The accuracy percentage for predicting performance in E1-F is 77.14%.

The correct classified data from E1-F consisted of 649 instances classified as active and 26 instances classified as drop out. Based on the correct classified data in the tree in Fig. 4,

the meaningful rules generated from the tree are shown in Table V.

TABLE IV. RULES FOR ACTIVE STUDENTS (E1-U)

Rule #	Rule Premise	Percentages of Instances	
		Active	Drop Out
1	IF Gender = M	97.70%	2.30%
2	IF Gender = F and Faculty = E	98.47%	1.53%
3	IF Gender = F and Faculty = T and City = B and GPA = 1	75.00%	25.00%
4	IF Gender = F and Faculty = T and City = A	100%	0%
5	IF Gender = F and Faculty = T and City = B and (GPA = 2 or GPA = 3)	100%	0%

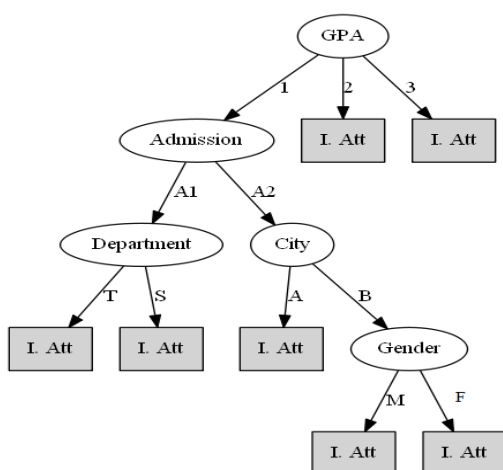


Fig. 4. NBTree for E1-F

Table V indicates that GPA attribute was the most affective attribute in determining active students in faculty T. Students with good or excellent GPA have better performance than below GPA. For students with below GPA, Admission, City, and Gender attributes determined whether they will survive to finish their study.

B. Experiment-2(E2) : Graduates Data set with GPA as the Class

Graduates data set in Table III with GPA as the class label for experiment-2 in university level (E2-U) consisted of 1209 instances. Fig. 5 shows the tree with three leaves resulted from NBTree classification for the data set. The accuracy percentage for predicting performance in E2-U is 68.74%.

The correct classified data from E2-U consisted of 498 instances classified as satisfactory GPA, 263 instances classified as excellent GPA, and 70 instances classified as cum laude GPA. Based on the correct classified data in the tree in Fig. 5, the meaningful rules generated from the tree are shown in Table VI.

Table VI shows that when the Credit was equal 144, there was 66.80% satisfactory GPA. However, when the Credit was

above 144, the satisfactory GPA increased to 93.62%, especially in faculty E. In faculty T, when the Credit was above 144, the GPA was dominated by excellent GPAs.

TABLE V. RULES FOR ACTIVE STUDENTS (E1-F)

Rule #	Rule Premise	Percentages of Instances	
		Active	Drop Out
1	IF GPA = 1 and Admission = E1 and Department = T	73.85%	26.15%
2	IF GPA = 1 and Admission = E1 and Department = S	64.71%	35.29%
3	IF GPA = 1 and Admission = E2 and City = B and Gender = M	94.34%	5.66%
4	IF GPA = 1 and Admission = E2 and City = A	100%	0%
5	IF GPA = 1 and Admission = E2 and City = B and Gender = F	100%	0%
6	IF GPA = 2 or GPA = 3	100%	0%

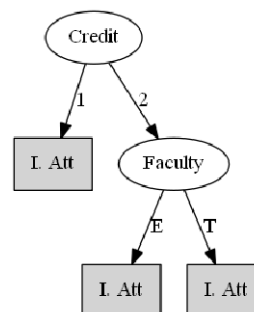


Fig. 5. NBTree for E2-U

Graduates data set in Table III with GPA as the class for experiment-2 in faculty level (E2-F) consisted of 437 instances. NBTree classification for the data set resulted in no tree. The accuracy percentage for predicting performance in E2 in faculty level is 63.84 %.

TABLE VI. RULES FOR GRADUATES – CLASS GPA (E2-U)

Rule #	Rule Premise	Percentage of Instances		
		Satisfactory	Excellent	Cum Laude
1	IF Credit = C1	66.80%	29.46%	3.74%
2	IF Credit = C2 and Faculty = E	93.62%	6.38%	0%
3	IF Credit = C2 and Faculty = T	0%	67.70%	32.30%

C. Experiment-3(E3) : Graduates Data set with LST as the Class

Graduates data set in Table III with LST as the class for experiment-3 in university level (E3-U) consists of 1209 instances. Fig. 6 shows tree with 23 leaves resulted from NBTree classification for the data set. The accuracy percentage for predicting performance in E3-U is 69.70%.

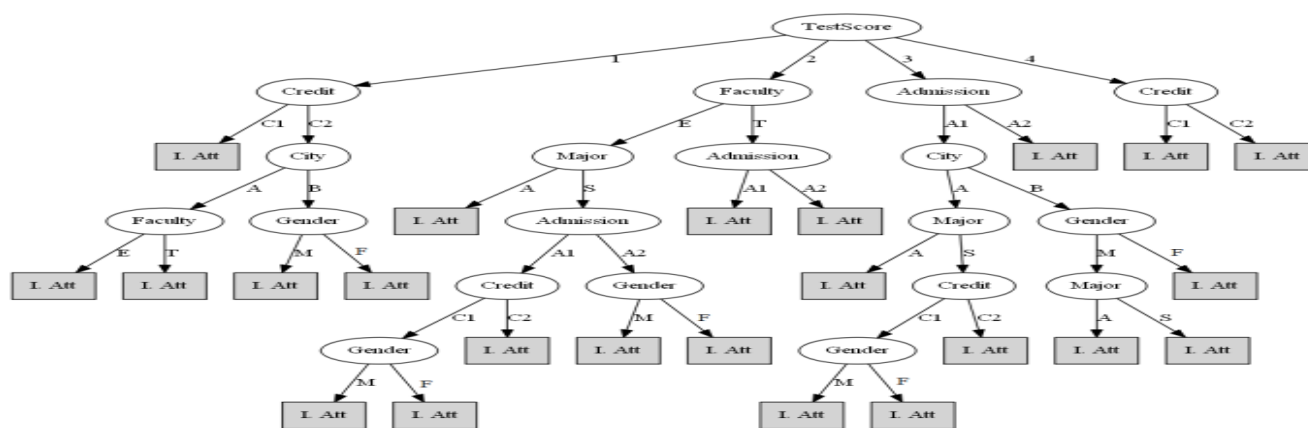


Fig. 6. NBTree for E3-U

The correct classified data in E3-U consisted of 315 instances classified as on time, and 520 instances classified as late. Based on the correct classified data in the tree in Fig. 6, the meaningful rules generated from the tree are shown in Table VII.

Table VII indicates that students will be graduated on time if they have excellent Test Score. Based on length of study, graduates in faculty E have higher performance than graduates in faculty T, especially if they have good Test Score.

Graduates data set in Table III with LST as the class for experiment-3 in faculty level (E3-F) consists of 437 instances. Fig. 7 shows tree with 28 leaves resulted from NBTree classification for the data set. The accuracy percentage for predicting performance in E3-F is 61.56%.

The correct classified data in E3-F consisted of 192 instances classified as on time, and 77 instances classified as late. Based on the correct classified data in the tree in Fig. 7, the meaningful rules generated from the tree are shown in Table VIII.

Table VIII indicates that Major attribute was the most affective attribute in determining length of study in faculty T. Students from Major A have higher performance than students from Major S. Whenever Department is taken into consideration, graduates from Department T have better performance than from Department S.

V. CONCLUSION

This paper focused on building a classification model to predict students' performance. To achieve the objective, many attributes has been tested, and some of them were found as influential attributes to performance prediction.

It can be concluded based on the classification model resulted from the experiments that:

- 1) Prediction at the university level for active students indicated that Gender attribute had significant influence to determine whether students will be survive to finish their study.

TABLE VII. RULES FOR GRADUATES – CLASS LST (E3-U)

Rule #	Rule Premise	Percentage of Instances	
		On Time	Late
1	IF Test Score = 1	100%	0%
2	IF Test Score = 2 and Faculty = E and Major = A	100%	0%
3	IF Test Score = 2 and Faculty = E and Admission = A2 and Major = S	100%	0%
4	IF Test Score = 2 and Faculty = T and Admission = E1	70.73%	29.27%
5	IF Test Score = 2 and Faculty = T and Admission = E2	66.67%	33.33%
6	IF Test Score = 3 and Admission = A1 and City = B and Gender = M and Major = A	66.67%	33.33%
7	IF Test Score = 3 and Admission = A1 and City = A and Major = S and Credit = C1	100%	0%

- 2) Prediction at the university level for graduates indicated that :
 - a. Credit attribute had significant effect to identify graduates GPA.
 - b. Test Score attribute had significant effect to determine graduates length of study.
- 3) Prediction at the faculty level for active students indicated that GPA attribute had significant influence to determine whether students will be survive to finish their study.
- 4) Prediction at the faculty level for graduates indicated that Test Score attribute had significant effect to determine graduates length of study in faculty T.

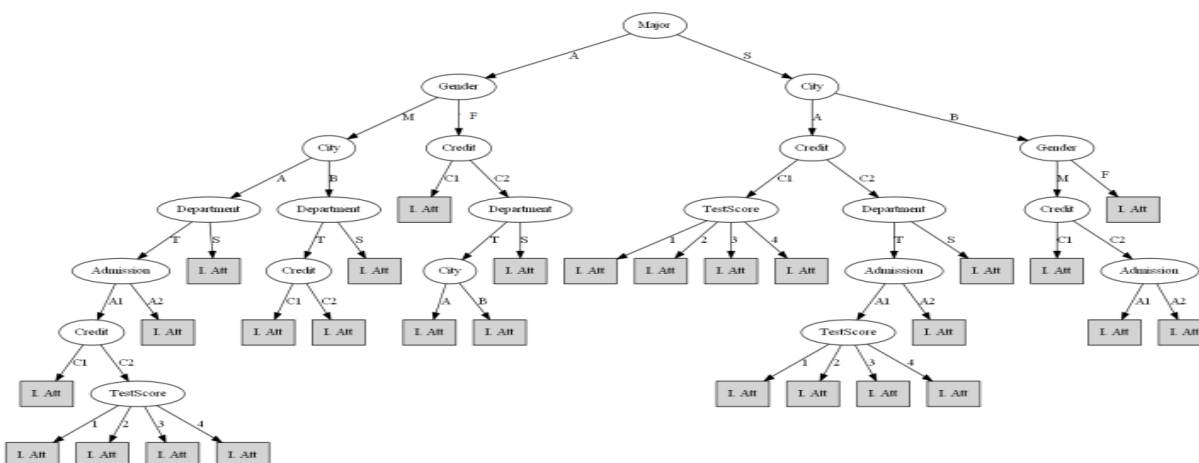


Fig. 7. NBTree for E3-F

TABLE VIII. RULES FOR GRADUATE – CLASS LST (E3-F)

Rule #	Rule Premise	Percentage of instances	
		On Time	Late
1	IF Major = A and Gender = F and Credit=C1	100%	0%
2	IF Major = A and Gender = F and Department = S and Credit=C2	100%	0%
3	IF Major = A and Gender = F and City = A and Department = T and Credit=C2	83.33%	16.67%
4	IF Major = A and Gender = M and City=A and Department=S	71.43%	28.57%
5	IF Major = A and Gender = M and City=B and Department=S	80%	20%
6	IF Major = A and Gender = M and City=B and Department=T	100%	0%
7	IF Major = A and Gender = M and City=A and Department=T and Admission = A2	100%	0%
8	IF Major = A and Gender = M and City=A and Department=T and Admission = A1 and Credit = C1	100%	0%
9	IF Major = A and Gender = M and City=A and Department=T and Admission = A1 and Credit = C2 and Test Score=1	100%	0%
10	IF Major = S and City=A and Department=S and Credit=C2	62.50%	37.50%
11	IF Major = S and Gender = M and City=B and Credit=C1	100%	0%
12	IF Major = S and City=A and Credit=C1 and Test Score = 1 or 2 or 3	100%	0%

5) Classification model to predict students' performance was resulted as a set of rules, which can be used to predict the new student performance.

Further research may collect more proper data from several higher education institutions to generate a correct model for students' performance.

ACKNOWLEDGMENT

The authors would like to thank DIPa Kopertis Wilayah IV, Ministry of National Education and Culture of Republic of Indonesia, as the funding sponsor of this research.

REFERENCES

- [1] J. Han, M. Kamber, and J. Pei. Data Mining Concepts and Techniques, 3rd ed. Waltham: Elsevier Inc, 2012.
- [2] C. Vialardi, J. Bravo, L. Shafiq, A. Ortigosa. "Recommendation in higher education using data mining techniques." Available: eric.ed.gov/?id=ED539088, Retrieved September 7, 2014.
- [3] J. Ranjan, S. Khalil, "Conceptual framework of data mining process in management education in India : an institutional perspective." Information Technology Journal, vol.7, pp.16-23, 2008
- [4] B.K. Bhardwaj, S. Pal. "Data mining : a prediction for performance improvement using classification." Available: arxiv.org/pdf/1201.3418, Retrieved March 6, 2014.
- [5] J. Kasih, M. Ayub, S. Susanto. "Predicting students' final passing results using the apriori algorithm." World Transaction on Engineering and Technology Education, Vol. 11, pp. 517-520, 2013.
- [6] R. Kohavi. "Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid." Available: robotics.stanford.edu/~ronnyk/nbtree.pdf, Retrieved November 1, 2013.
- [7] M. Ayub, T. Kristanti, M. Caroline. "Data warehouse sebagai basis analisis data akademik perguruan tinggi." Proceeding of Seminar Nasional Teknologi Informasi, Faculty of Information Technology Tarumanegara University, pp. 18 – 25, 2013.
- [8] M. Ayub, T. Kristanti. Model Analisis Classification dan Clustering untuk Data Mahasiswa dan Dosen di Perguruan Tinggi. Bandung: DIKTI Competitive Grant Report 2013 at Maranatha Christian University, 2013.
- [9] P.N. Tan, M. Steinbach, M., and V. Kumar. Introduction to Data Mining. Boston: Pearson, 2006.
- [10] I.H. Witten, E. Frank, and M.A. Hall. Data Mining Practical Machine Learning Tools and Techniques, 3rd ed. Burlington: Elsevier, 2011.