MATICS : Jurnal Ilmu Komputer dan Teknologi Informasi ISSN : 1978-161X(p); 2477-2550(e)
Volume 13, No. 1(2021), pp 33-41 DOI : 10.18860/mat.v13i1.9726
Received: December 26th 2021; **Accepted**: January 29th 2021; **Avalaible Online**: March 25th 2021

Identification of Student Academic Performance in Computer Science Based on *Naïve Bayes*

Kevin Elmy Aziz, Cahyo Crysdian, M. Imamudin

***Abstract*— Jurusan Teknik Informatika is one of the study programs at UIN Maulana Malik Ibrahim. Based on the current curriculum in Jurusan Teknik Informatika, the curriculum refers to the IEEE/ACM Computer Science Curricula 2013. The IEEE/ACM Computer Science Curricula 2013 has a knowledge area classification, which is mentioned in the curriculum as having 18 knowledge areas. The curriculum used in the current technical study program is formulated and determined from the entire content or collection of knowledge in the IEEE/ACM Computer Science Curricula 2013. In the Jurusan Teknik Informatika curriculum at UIN Malik Ibrahim Malang currently there are 76 subjects, 58 of which are Teknik Informatika subjects and 18 others are general subjects. To identify the academic performance of students it is necessary to classify the curriculum in the Department of Informatics Engineering to the knowledge area in the IEEE / ACM Computer Science Curricula 2013. Classification is done using the Naïve Bayes method by calculating the probability of each course of the knowledge area, after it is done classification, data will appear in the form of subject distribution to the knowledge area. After classification, it is necessary to determine the level of contribution of each course that has spread to the knowledge area. This contribution level is entered into the Joint formula with the value of the student transcript to calculate the student's academic performance. Testing is done by comparing the output in the form of knowledge area with the highest performance produced by the program with input in the form of knowledge area from the expert for each student. This research resulted in an accuracy of 78.95% from the results of twenty times experiment.**

***Index Terms*— Academic Performance; Classification; Naïve Bayes; Preprocessing; Zeno Dichotomy Paradox;**

I. INTRODUCTION

Jurusan Teknik Informatika is one of the study programs at UIN Maulana Malik Ibrahim. Based on the current curriculum in Jurusan Teknik Informatika, the curriculum refers to the IEEE/ACM Computer Science Curricula 2013. The IEEE/ACM Computer Science Curricula 2013 has a knowledge area classification, which is mentioned in the curriculum as having 18 knowledge areas. The curriculum used in the current technical study program is formulated and determined from the entire content or collection of knowledge in the IEEE/ACM Computer Science Curricula 2013.

In the Jurusan Teknik Informatika curriculum at UIN Malik Ibrahim Malang currently there are 76 subjects, 58 of which are Teknik Informatika subjects and 18 others are general subjects.

Meanwhile, in the IEEE/ACM Computer Science Curricula 2013, 18 knowledge areas are covering the whole topic of the IEEE/ACM Computer Science Curricula 2013, although in fact in IEEE/ACM Computer Science Curricula 2013 it does not confirm or propose a series of subjects or specific curriculum structures. Knowledge areas are not intended to be in a single relationship with certain subjects in the curriculum.

UIN Maulana Malik Ibrahim Malang Teknik Informatika curriculum covers material sourced from the IEEE/ACM Computer Science Curricula 2013 which are packaged into a variety of subjects. At present there is no research and determination raised by the Jurusan Teknik Informatika that discusses the relationship between the subject and the knowledge area, considering that in the IEEE/ACM Computer Science Curricula 2013 there are 18 knowledge areas that can structurally assist in the preparation and grouping of materials for further use as study material in each subject.

this study aims to determine the academic performance of students based on the existing knowledge area in the IEEE/ACM Computer Science Curricula 2013, by taking transcript scores during their lectures. But before that, it is necessary to classify to determine the relationship between the subject and knowledge area, so that the distribution of subjects will be known to the knowledge area. At present, there is no grouping or classification of subjects in UIN Maulana

Kevin Elmy Aziz is with the Informatic Engineering Departement of UIN Maulana Malik Ibrahim Malang, Indonesia (email 16650087@student.uin-malang.ac.id)

Cahyo Crysdian., is with the Informatic Engineering Departement of UIN Maulana Malik Ibrahim Malang, Indonesia.

M.Imamudin is the Informatic Engineering Departement of UIN Maulana Malik Ibrahim Malang, Indonesia

Malik Ibrahim's Teknik Informatika curriculum against the knowledge area in IEEE/ACM Computer Science Curricula 2013. To do this classification the Naive Bayes classifier method will be used to determine the distribution of each subject to the knowledge area, this method was chosen because in this classification process the data is in the form of text. As one of these successful methods, Naïve Bayes is popular in text classification due to its computational efficiency and relatively good predictive performance [1]. In each knowledge area, some topics are set as the focus of the material, while in the subject there are study materials as listed in the semester learning plan or RPS where the study material is the contents of the material in each subject. Then to determine the class or knowledge area that is most suitable for each subject will be classified in the form of text by taking topics from the knowledge area and study material from each subject.

In determining student academic performance, student value transcripts will be used as a reference in determining academic performance by calculating the value of each subject that has been classified into the knowledge area. After going through the calculation process, the output of this application is the value of each knowledge area, to determine the academic performance can be known by seeing the knowledge area with the highest value.

II. LITERATURE REVIEW

Related research conducted by Jiang take the first step by review the existing weighting approach for Naive Bayes and find that all of them only include the weight of the features studied in the classification of Naïve Bayes formulas and in no way include the weight of the features learned into estimation of conditional probabilities at all. Then, the researcher proposed a simple, efficient, and effective feature weight approach, called deep feature weighting (DFW), which estimates the conditional probabilities of Naive Bayes by calculating the frequency of feature weighted frequencies from the training data [2].

Alkubaisi et. al did a research and showed that their research has achieved high accuracy equal to 90.38% for HNBC with all classes (positive, negative and neutral). This result will enable decision-makers and investors in the domain of the stock market exchange to make safe, low-risk decisions because these results depend on facts about the stock market domain. Facts such as spatial and temporal features are needed in addition to the role of stock market experts in achieving real sentiment analysis. High classification accuracy with real sentiment analysis will produce reports and indicators that are accurate and reliable on company shares. from these results, it can be seen that machine learning methods that use sentiment analysis on Twitter such as the NB classifier produce high, real and reliable accuracy by simulating domain features and preparing datasets using the NLP method [3].

In other research, Naive Bayes with the Query Expansion Ranking feature selection to reduce the number of features in the classification process. The process of sentiment analysis consists of preprocessing, feature selection using the Expansion Ranking Query method, and classification with Naive Bayes. The test in this study is an accuracy test using variations in the ratio of feature selection, the result is feature selection 75% has the best accuracy of 86.6% [4].

Indrayuni researched and proved that based on testing the model using the Naive Bayes algorithm in experiments that have been carried out it is proven that the Naive Bayes algorithm is the simplest algorithm which is proven to produce high accuracy values up to 90.50% with an AUC value of 0.715 **[5]**.

Other studies regarding the comparison of methods in classification are carried out for personality classification. Testing was conducted using 10-fold cross-validations. In the crossvalidation testing, MNB got the best accuracy in three methods tested with average accuracy 60%. SVM and KNN performed similarly. SVM method performs worse than MNB due to difficulties separating a class of a word as dataset are not quite accurate. KNN method also performs worse than MNB. The alleged cause of the low accuracy of the KNN method because of the difficulty in determining the optimal value of K. Total value of K is crucial because the KNN's probability result will be calculated from the K samples. This is different from MNB that uses pure probability calculations on existing features. Based on macro-averaged scores in 59%-60%, this experiment fails to improve accuracy, as it is only equal to the best score from previous research (61%) [6].

In contrast to the above research, this study does not use initial data that already labeled like other research that already have data with labels that have been determined, in this study the amount of data will be the same as the number of courses and the data will be formed from terms that arise from each course with a class contains 18 knowledge areas. Then the level of contribution of each course will be determined to the knowledge area using the concept of the Zeno dichotomy paradox. After that, the last stage is the process of determining student academic performance will be used a transcript of student scores as input and the proposed formula regarding the distribution of courses in each knowledge area.

III. SYSTEM DESIGN

A. System Design

The System design is shown in Figure 1 which consists of some steps, namely dataset, preprocess, Naïve Bayes, Zeno Dichotomy Paradox, transcript score input, subject and score distribution, knowledge area performance and highest score of knowledge area. The next steps will be discussed in the next session.

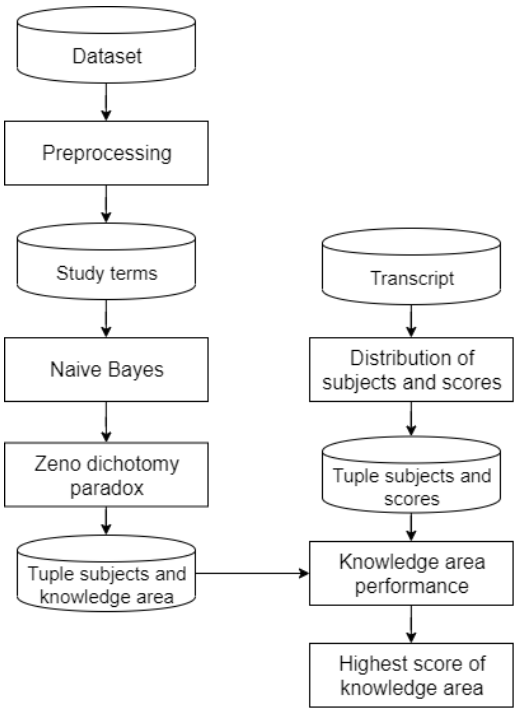


Figure 1. System Design

B. Dataset

Data is in form of text and numeric, taken from the RPS document which is a semester learning plan in the jurusan Teknik informatika, knowledge area in the IEEE/ACM Computer Science Curricula 2013, and student score transcripts as in Figure 2.

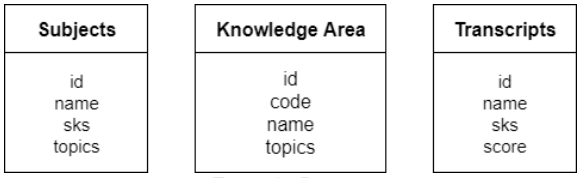
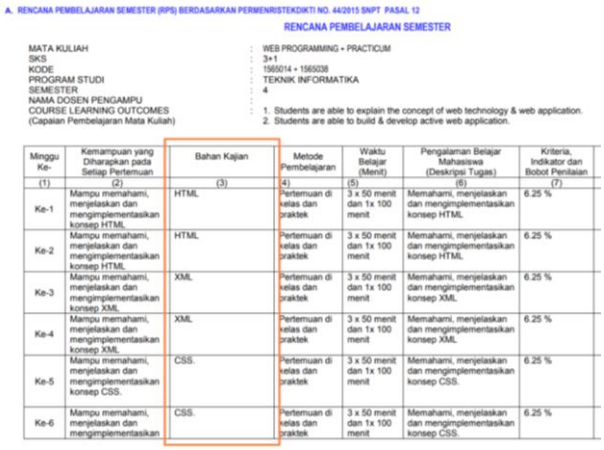


Figure 2. Dataset

One of the data to be used comes from the study material contained in the RPS document. Study material contains a collection of terms about what is to be taught in the subject. The study material in the RPS document as shown in Figure 3 below.



A. RENCANA PEMBELAJARAN SEMESTER (RPS) BERDASARKAN PERMENRISTEKDIKTI NO. 44/2015 SNPT PASAL 12

RENCANA PEMBELAJARAN SEMESTER

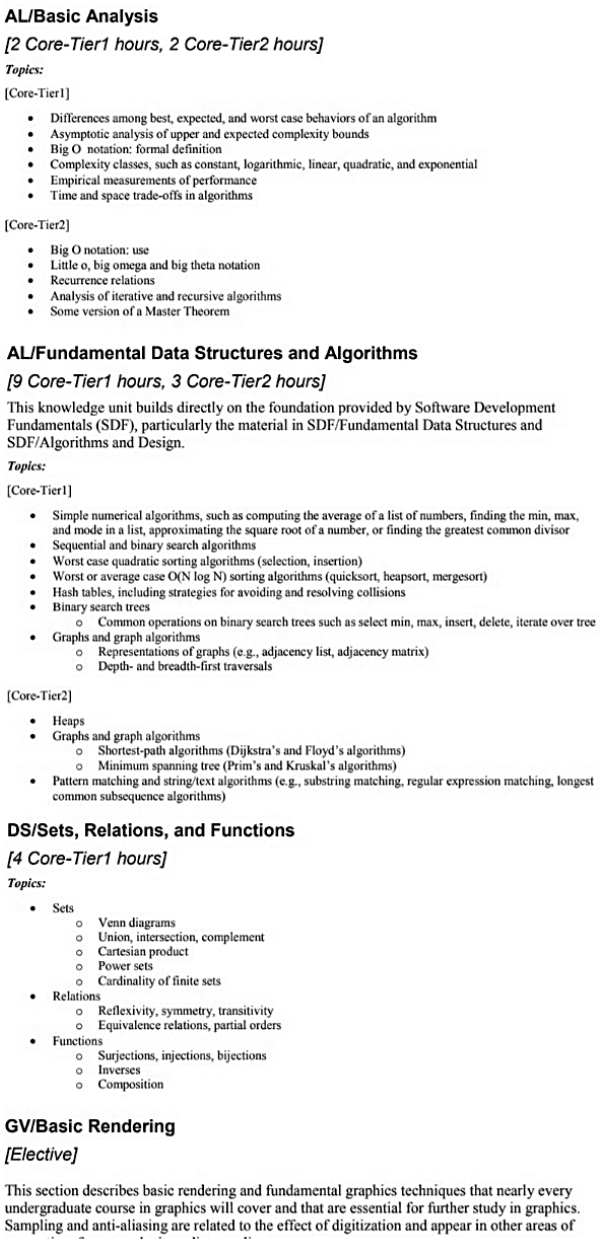
MATA KULIAH	:	WEB PROGRAMMING + PRACTICUM
SKS	:	3+1
KODE	:	1565014 + 1565038
PROGRAM STUDI	:	TEKNIK INFORMATIKA
SEMESTER	:	4
NAMA DOSEN PENGAMPU	:	
COURSE LEARNING OUTCOMES (Capaian Pembelajaran Mata Kuliah)	:	1. Students are able to explain the concept of web technology & web application. 2. Students are able to build & develop active web application.

Minggu Ke-	Kemampuan yang Diharapkan pada Setiap Pertemuan	Bahan Kajian	Metode Pembelajaran	Waktu Belajar (Menit)	Pengalaman Belajar Mahasiswa (Deskripsi Tugas)	Kriteria, Indikator dan Bobot Penilaian
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ke-1	Mampu memahami, menjelaskan dan mengimplementasikan konsep HTML	HTML	Pertemuan di kelas dan praktek	3 x 50 menit dan 1x 100 menit	Memahami, menjelaskan dan mengimplementasikan konsep HTML	6.25 %
Ke-2	Mampu memahami, menjelaskan dan mengimplementasikan konsep HTML	HTML	Pertemuan di kelas dan praktek	3 x 50 menit dan 1x 100 menit	Memahami, menjelaskan dan mengimplementasikan konsep HTML	6.25 %
Ke-3	Mampu memahami, menjelaskan dan mengimplementasikan konsep XML	XML	Pertemuan di kelas dan praktek	3 x 50 menit dan 1x 100 menit	Memahami, menjelaskan dan mengimplementasikan konsep XML	6.25 %
Ke-4	Mampu memahami, menjelaskan dan mengimplementasikan konsep XML	XML	Pertemuan di kelas dan praktek	3 x 50 menit dan 1x 100 menit	Memahami, menjelaskan dan mengimplementasikan konsep XML	6.25 %
Ke-5	Mampu memahami, menjelaskan dan mengimplementasikan konsep CSS.	CSS.	Pertemuan di kelas dan praktek	3 x 50 menit dan 1x 100 menit	Memahami, menjelaskan dan mengimplementasikan konsep CSS.	6.25 %
Ke-6	Mampu memahami, menjelaskan dan mengimplementasikan	CSS.	Pertemuan di kelas dan praktek	3 x 50 menit dan 1x 100 menit	Memahami, menjelaskan dan mengimplementasikan konsep CSS.	6.25 %

Figure 3. Study material contained in the RPS document

The next data that will be used is the topic of the existing knowledge area in the IEEE/ACM Computer Science Curricula 2013. The topic in this knowledge area contains a collection of terms forming a study material. The topic of the knowledge area is shown in Figure 4.

In Figure 4, you can see that each knowledge area has chapters and each chapter has topics. In each topic there are study materials which in the study material contain various terms that describe teaching materials or material in the knowledge area. If we look at the study material in the RPS document, the study material also contains terms that represent the material in the course. These terms will be used for the classification process by looking for the probabilities of each term in a course of the terms in each knowledge area that were previously pre-processed from the two data sources before they will be processed using the naïve bayes method.



AL/Basic Analysis

[2 Core-Tier1 hours, 2 Core-Tier2 hours]

Topics:

[Core-Tier1]

- Differences among best, expected, and worst case behaviors of an algorithm
- Asymptotic analysis of upper and expected complexity bounds
- Big O notation: formal definition
- Complexity classes, such as constant, logarithmic, linear, quadratic, and exponential
- Empirical measurements of performance
- Time and space trade-offs in algorithms

[Core-Tier2]

- Big O notation: use
- Little o, big omega and big theta notation
- Recurrence relations
- Analysis of iterative and recursive algorithms
- Some version of a Master Theorem

AL/Fundamental Data Structures and Algorithms

[9 Core-Tier1 hours, 3 Core-Tier2 hours]

This knowledge unit builds directly on the foundation provided by Software Development Fundamentals (SDF), particularly the material in SDF/Fundamental Data Structures and SDF/Algorithms and Design.

Topics:

[Core-Tier1]

- Simple numerical algorithms, such as computing the average of a list of numbers, finding the min, max, and mode in a list, approximating the square root of a number, or finding the greatest common divisor
- Sequential and binary search algorithms
- Worst case quadratic sorting algorithms (selection, insertion)
- Worst or average case O(N log N) sorting algorithms (quicksort, heapsort, mergesort)
- Hash tables, including strategies for avoiding and resolving collisions
- Binary search trees
  - Common operations on binary search trees such as select min, max, insert, delete, iterate over tree
- Graphs and graph algorithms
  - Representations of graphs (e.g., adjacency list, adjacency matrix)
  - Depth- and breadth-first traversals

[Core-Tier2]

- Heaps
- Graphs and graph algorithms
  - Shortest-path algorithms (Dijkstra's and Floyd's algorithms)
  - Minimum spanning tree (Prim's and Kruskal's algorithms)
- Pattern matching and string/text algorithms (e.g., substring matching, regular expression matching, longest common subsequence algorithms)

DS/Sets, Relations, and Functions

[4 Core-Tier1 hours]

Topics:

- Sets
  - Venn diagrams
  - Union, intersection, complement
  - Cartesian product
  - Power sets
  - Cardinality of finite sets
- Relations
  - Reflexivity, symmetry, transitivity
  - Equivalence relations, partial orders
- Functions
  - Surjections, injections, bijections
  - Inverses
  - Composition

GV/Basic Rendering

[Elective]

This section describes basic rendering and fundamental graphics techniques that nearly every undergraduate course in graphics will cover and that are essential for further study in graphics. Sampling and anti-aliasing are related to the effect of digitization and appear in other areas of computing, for example, in audio sampling.

Topics:

- Rendering in nature, e.g., the emission and scattering of light and its relation to numerical integration
- Forward and backward rendering (i.e., ray-casting and rasterization)
- Polygonal representation
- Basic radiometry, similar triangles, and projection model
- Affine and coordinate system transformations
- Ray tracing
- Visibility and occlusion, including solutions to this problem such as depth buffering, Painter's algorithm, and ray tracing
- The forward and backward rendering equation
- Simple triangle rasterization
- Rendering with a shader-based API
- Texture mapping, including minification and magnification (e.g., trilinear MIP-mapping)
- Application of spatial data structures to rendering
- Sampling and anti-aliasing
- Scene graphs and the graphics pipeline

IS/Fundamental Issues

[1 Core-Tier2 hours]

Topics:

- Overview of AI problems, examples of successful recent AI applications
- What is intelligent behavior?
  - The Turing test
  - Rational versus non-rational reasoning
- Problem characteristics
  - Fully versus partially observable
  - Single versus multi-agent
  - Deterministic versus stochastic
  - Static versus dynamic
  - Discrete versus continuous
- Nature of agents
  - Autonomous versus semi-autonomous
  - Reflexive, goal-based, and utility-based
  - The importance of perception and environmental interactions
- Philosophical and ethical issues. [elective]

IAS/Cryptography

[1 Core-Tier2 hour]

Topics:

[Core-Tier2]

- Basic Cryptography Terminology covering notions pertaining to the different (communication) partners, secure/unsecure channel, attackers and their capabilities, encryption, decryption, keys and their characteristics, signatures
- Cipher types (e.g., Caesar cipher, affine cipher) together with typical attack methods such as frequency analysis
- Public Key Infrastructure support for digital signature and encryption and its challenges

[Elective]

- Mathematical Preliminaries essential for cryptography, including topics in linear algebra, number theory, probability theory, and statistics
- Cryptographic primitives:
  - pseudo-random generators and stream ciphers
  - block ciphers (pseudo-random permutations), e.g., AES
  - pseudo-random functions
  - hash functions, e.g., SHA2, collision resistance
  - message authentication codes
  - key derivations functions
- Symmetric key cryptography
  - Perfect secrecy and the one time pad
  - Modes of operation for semantic security and authenticated encryption (e.g., encrypt-then-MAC, OCB, GCM)
  - Message integrity (e.g., CMAC, HMAC)
- Public key cryptography:
  - Trapdoor permutation, e.g., RSA
  - Public key encryption, e.g., RSA encryption, EI Gamal encryption
  - Digital signatures
  - Public-key infrastructure (PKI) and certificates
  - Hardness assumptions, e.g., Diffie-Hellman, integer factoring
- Authenticated key exchange protocols, e.g., TLS
- Cryptographic protocols: challenge-response authentication, zero-knowledge protocols, commitment, oblivious transfer, secure 2-party or multi-party computation, secret sharing, and applications
- Motivate concepts using real-world applications, e.g., electronic cash, secure channels between clients and servers, secure electronic mail, entity authentication, device pairing, voting systems.
- Security definitions and attacks on cryptographic primitives:
  - Goals: indistinguishability, unforgeability, collision-resistance
  - Attacker capabilities: chosen-message attack (for signatures), birthday attacks, side channel attacks, fault injection attacks.
- Cryptographic standards and references implementations
- Quantum cryptography

GV/Visualization

[Elective]

Visualization has strong ties to the Human-Computer Interaction (HCI) knowledge area as well as Computational Science (CN). Readers should refer to the HCI and CN KAs for additional opics related to user population and interface evaluations.

Topics:

- Visualization of 2D/3D scalar fields: color mapping, isosurfaces
- Direct volume data rendering: ray-casting, transfer functions, segmentation
- Visualization of:
  - Vector fields and flow data
  - Time-varying data
  - High-dimensional data: dimension reduction, parallel coordinates,
  - Non-spatial data: multi-variate, tree/graph structured, text
- Perceptual and cognitive foundations that drive visual abstractions
- Visualization design
- Evaluation of visualization methods
- Applications of visualization

CN/Processing

[Elective]

The processing topic area includes numerous topics from other knowledge areas. Specifically, coverage of processing should include a discussion of hardware architectures, including parallel systems, memory hierarchies, and interconnections among processors. These are covered in AR/Interfacing and Communication, AR/Multiprocessing and Alternative Architectures, AR/Performance Enhancements.

Topics:

- Fundamental programming concepts:
  - The concept of an algorithm consisting of a finite number of well-defined steps, each of which completes in a finite amount of time, as does the entire process.
  - Examples of well-known algorithms such as sorting and searching.
  - The concept of analysis as understanding what the problem is really asking, how a problem can be approached using an algorithm, and how information is represented so that a machine can process it.
  - The development or identification of a workflow.
  - The process of converting an algorithm to machine-executable code.
  - Software processes including lifecycle models, requirements, design, implementation, verification and maintenance.
  - Machine representation of data computer arithmetic.
- Numerical methods
  - Algorithms for numerically fitting data (e.g., Newton's method)
  - Architectures for numerical computation, including parallel architectures
- Fundamental properties of parallel and distributed computation:
  - Bandwidth.
  - Latency.
  - Scalability.
  - Granularity.
  - Parallelism including task, data, and event parallelism.
  - Parallel architectures including processor architectures, memory and caching.
  - Parallel programming paradigms including threading, message passing, event driven techniques, parallel software architectures, and MapReduce.
  - Grid computing.
  - The impact of architecture on computational time.
  - Total time to science curve for parallelism: continuum of things.
- Computing costs, e.g., the cost of re-computing a value vs. the cost of storing and lookup.

IS/Basic Knowledge Representation and Reasoning

[3 Core-Tier2 hours]

Topics:

- Review of propositional and predicate logic (cross-reference DS/Basic Logic)
- Resolution and theorem proving (propositional logic only)
- Forward chaining, backward chaining
- Review of probabilistic reasoning, Bayes theorem (cross-reference with DS/Discrete Probability)

Figure 4. Topics contained in the knowledge area of IEEE/ACM Computer Science Curricula 2013

C. Preprocessing

Preprocessing has three stages including, case folding, tokenizing, and stopword removal. In this system, the case folding and stopwords removal process is carried out in the middle of the tokenizing process. This is done to reduce the number of loops in the program because looping to tokenize is done to the level per word making it possible to do the case folding process as well as the stopwords removal process, ie checking whether the word entered into the list of words that must be removed. Here is the source code for the Knowledge Area data preprocessing process.

Each subject and knowledge area will be preprocessed, the preprocessing stage covers some stages including, case folding, tokenizing and stopword removal. Case folding is the process of turning all characters into standard shapes, in this study all characters will be converted to lowercase letters. For example, the word "binary" if the word is at the beginning of a sentence from the topic it will have capital letters as its first character, and when compared with the same word but do not have capital letters the system will be considered a different string because the string has case sensitive properties. So it is necessary to do a folding case so that all words that are visibly the same but have different character structures will be considered the same again because the form has been changed to a similar form.

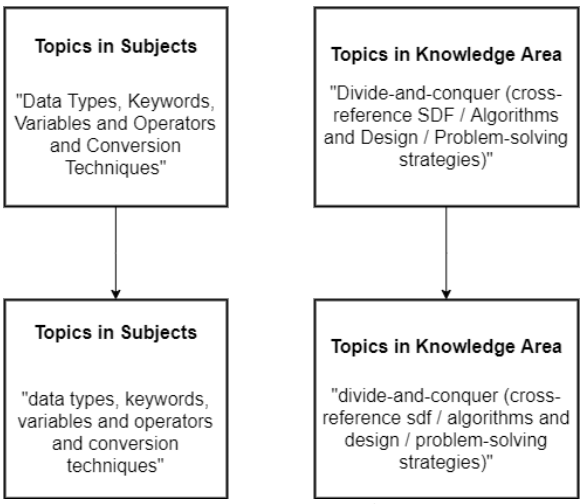


Figure 5. Output Casefolding

Each subject and knowledge area will be preprocessed, the preprocessing stage covers some stages including, case folding, tokenizing and stopword removal. Case folding is the process of turning all characters into standard shapes, in this study all characters will be converted to lowercase letters. For example, the word "binary" if the word is at the beginning of a sentence from the topic it will have capital letters as its first character, and when compared with the same word but do not have capital letters the system will be considered a different string because the string has case sensitive properties. So it is necessary to do a folding case so that all words that are visibly the same but have different character structures will be considered the same again because the form has been changed to a similar form.

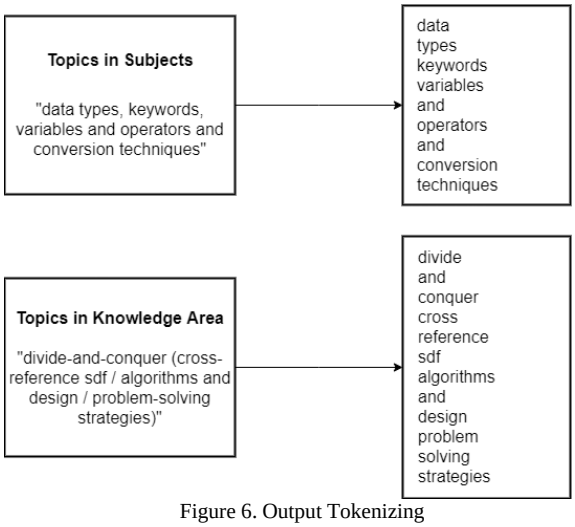


Figure 6. Output Tokenizing

The final step of the preprocessing stage is stopword removal, which is to discard words that are considered to have no meaning, usually conjunctions, or common words that have no meaningful value.

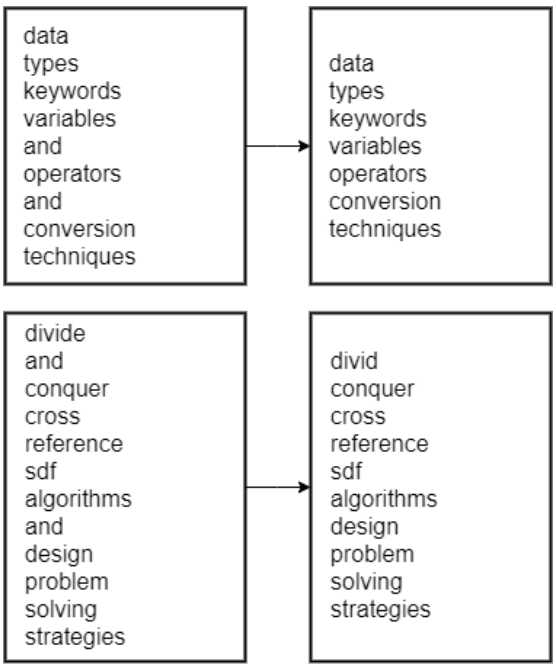


Figure 7. Output Stopwords Removal

D. Naïve Bayes

Then by using study material from subjects and topics from the knowledge area, it will form data such as the Table 1 below, the example table below shows the data generated from Web Programming subjects on the knowledge area.

Table 1. Subject terms – Web Programming

Knowledge Area	xml	css	php	java	script	framework
AL - Algorithms and Complexity	0	0	0	0	0	0
AR - Architecture and Organization	0	0	0	0	0	0
CN - Computational Science	0	0	0	0	0	0
DS - Discrete Structures	0	0	0	0	0	0
GV - Graphics and Visualization	0	0	0	0	0	0
HCI - Human-Computer Interaction	0	0	0	0	0	0
IAS - Information Assurance and Security	0	0	0	0	0	0
IM - Information Management	1	0	0	0	0	0
IS - Intelligent Systems	0	0	0	0	0	0
NC - Networking and Communications	0	0	0	0	0	0
OS - Operating Systems	0	0	0	0	0	0

PBD - Platform-based Development	0	1	1	1	1	0
PD - Parallel and Distributed Computing	0	0	0	0	0	0
PL - Programming Languages	0	0	0	0	0	1
SDF - Software Development Fundamental	0	0	0	0	0	0
SE - Software Engineering	0	0	0	0	0	0
SF - Systems Fundamentals	0	0	0	0	0	0
SP - Social Issues and Professional Practice	0	0	0	0	0	1

E. Naïve Bayes

Naive Bayes Classifier is a popular algorithm used for data mining purposes because of its ease of use and fast processing time, easy to implement with a fairly simple structure and high level of effectiveness [7].

The difference between Naïve Bayes classifiers and other learning methods lies in the process of developing hypotheses. In the Naïve Bayes classifier, a hypothesis is formed directly without a search process, only by calculating the frequency of occurrence of a word in the training data, whereas in other learning methods a hypothesized search is usually performed from the hypothesis space.

$$P(C|X) = \frac{P(x|c)P(c)}{P(x)}$$

Naive Bayes equation, in general, can Further elaboration of the Bayes formula is done by describing (c | x1, ..., xn), then the following equation applies written as follows.

$$P(c|X) = P(x_1|c)P(x_2|c)P(x_3|c) \dots P(x_n|c)P(c)$$

In this stage the results will be obtained is the subject classification of the knowledge area by taking the highest probability of each knowledge area. In the calculation phase using the Naive Bayes method, the probability sought is the probability of each knowledge area for a subject to find out which subject belongs to the knowledge area, by ranking the knowledge area based on the highest probability using the following formula.

$$P(KA|Subject(terms) = P(term1|KA)P(term2|KA) \dots P(termn|KA)$$

Naive Bayes formula above can bring up the possibility there will be a probability of 0 because there may be a knowledge area that has absolutely no term in the subject matter of a course. Therefore, efforts should be made to avoid the 0 probability. The method that can be used to avoid this is to use laplace correction.

$$\boldsymbol{\rho_i = \frac{m_i + 1}{n + k}}$$

The probability of each subject to the knowledge area will be calculated by calculating the probability of the term or token in the subject for each knowledge area, meaning that the results of this calculation will show the probability of each subject for each knowledge area.

After the probability value of each knowledge area is obtained, then the next step is the sorting stage, that is, the knowledge area will be sorted based on the probability value from highest to lowest and then proceed to the next stage, namely determining the value of the subject contribution to each knowledge area.

F. Zeno Dichotomy Paradox

Zeno Dichotomy Paradox is used to determine the weights of the size of features that are not fixed [8]. The weights assigned to each feature size follow the famous paradoxical Zeno Dichotomy series.

In this paradox, it is explained that in order to achieve a goal, a person must take a segment halfway, and after that to get through the next segment a person must still go through more segments including a quarter, eighth, sixteen and so on.

each subject has its knowledge area partner, certainly not only one knowledge area but it has a partnership with another knowledge area with the highest to lowest probability sequence. This shows that at this stage the knowledge area already has a collection of subjects that are related to the knowledge area, then based on the calculation of the probability that not all subjects enter the knowledge area of each subject will determine the level of contribution to a knowledge area based on the order of its probability of knowledge area.

Figure 8. Zeno Dichotomy Paradox

From Figure 8 it can be seen that in traveling a distance, each trip will take one segment in advance which segment is half of the journey of one segment. Following the concept of Zeno Dichotomy Paradox, this research will apply the concept of this paradox in determining the value of the contribution of each subject to a knowledge area.

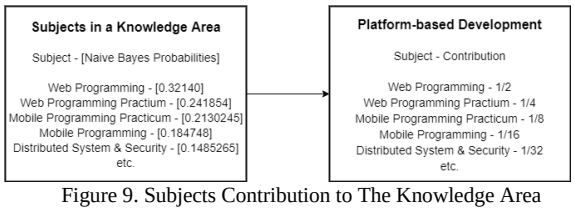


Figure 9. Subjects Contribution to The Knowledge Area

G. Input Transcript

The input transcript value of students is done by entering the transcript value file with excel format into the system to then be converted to an array of objects. Then from the value data will be entered into the structure of the object array that contains the contribution of subjects to each knowledge area. The

object generated from the excel input file is as shown in Figure 3.10 below.

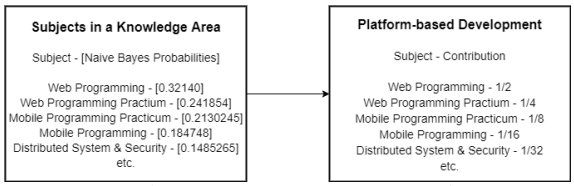


Figure 9. List of Subject inside Knowledge Area after Input Scores

H. *Knowledge Area Performance*

Knowledge area performance calculation is done by the proposed formula used to get the final value of each knowledge area, where each knowledge area has many subjects and each level contributes.

$$KA_i = \frac{\sum_{j=1}^{n}(score_j * sks_j * contribution_j)}{\sum_{j=1}^{n}(sks_j)}$$

Figure 9. Subjects Contribution to The Knowledge Area

In the equation above, to calculate the final value of each knowledge area, 18-times repetition or several existing knowledge areas will be repeated. Then in each iteration, there will be repeated to get the total value or score and the total of credits (sks). Repetition in each knowledge area will be carried out several times according to the number of subjects in the knowledge area.

IV. EXPERIMENT AND DISCUSSION

A. *Subject Classification*

Classification is done using the naïve Bayes method with a total of 18 classes, namely the number of knowledge areas. Each subject will have a probability for each knowledge area, which the probability will be a reference to determine the level of contribution of the subject to each knowledge area.

Data from the subject and knowledge area are in the form of text that has been packaged in JSON format. The text needs to be preprocessed inserting case folding, tokenizing, and stopword removal. Next will begin to be calculated for the probability of each term in the subject to the terms in the knowledge area to find out what is the final probability of the subject to the knowledge area.

After the subject probability of each knowledge area is obtained, the next step is to enter the contribution level by sorting the knowledge area with the highest to lowest probability and then each knowledge area is given a contribution level based on the Zeno dichotomy paradox concept in which each level is half the previous level starting from 0.5.

B. *Student Academic Performance Calculation*

In this section the subject has been classified, next is how to calculate student academic performance. In this calculation, the 7th-semester student transcript will be used which has been packaged in Excel format, which will then be read by the system and converted to the JSON format. After becoming a JSON format, the data can be presented and calculated values and other parameters using previous formula.

C. *Input Transcript*

At this stage the student transcript value will be inputted, using the transcript data from the excel file, the data must be extracted first and to facilitate the transcript input process, then the extracted data will be presented in the form of JSON. Here is the source code for extracting data in an excel file and presenting it in JSON format.


```
[
  {
    "id": 1565003,
    "mkName": "ALGORITMA & PEMROGRAMAN 1",
    "sks": 3,
    "scoreDis": "A"
  },
  {
    "id": 1565006,
    "mkName": "STRUKTUR DATA",
    "sks": 3,
    "scoreDis": "A"
  },
  {
    "id": 1565008,
    "mkName": "ELEKTRONIKA DIGITAL",
    "sks": 3,
    "scoreDis": "A"
  },
  {
    "id": 1565012,
    "mkName": "SISTEM KOMPUTER",
    "sks": 3,
    "scoreDis": "A"
  },
  {
    "id": 1565017,
    "mkName": "JARINGAN KOMPUTER",
    "sks": 3,
    "scoreDis": "A"
  },
  {
    "id": 1565018,
    "mkName": "KECERDASAN BUATAN",
    "sks": 3,
    "scoreDis": "A"
  },
```


Figure 8. Zeno Dichotomy Paradox

D. *Accuracy*

The output of this application is the value of student performance in each knowledge area, there are 18 knowledge areas, each of which already has a performance value. So to determine the knowledge area which student has the highest performance then has been sorted in the previous stage to display the knowledge area with the highest to lowest performance.

This experiment is carried out by calculating the accuracy of the output of this application compiled with input from the expert. At this stage, 20 transcript data for 7th-semester student grades will be taken and input from experts will also be taken, each student's transcript will be input by five experts. The expert will select three of the 18 knowledge areas based on the value of the transcript to determine the three selected knowledge

areas to be used as actual conditions or calculation material to calculate the accuracy of this application. After we get input from experts, the formula for the calculation below can be implemented.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \; x\; 100\%$$

Based on the formula above, For one student transcript, there will be 18 knowledge areas, the top three of the 18 will be a positive output, and the remaining 15 will be a negative output. TP is the amount of positive knowledge area that is also detected correctly or positively by the system, TN is a negative knowledge area that is also detected wrongly or negatively by the system, FP is a positive knowledge area that is detected wrongly or negatively by the system, and FN is negative data that is detected wrongly by the system.

E. Experimental Result

The first part of the results of this experiment is in the form of the probability results of the subject for each knowledge area. These results will determine the level of contribution of the subject to each knowledge area. Following is a table containing subject probabilities in each knowledge area. Below are the probabilities that result from naïve bayes calculations for foundation of computing subject.

Table 1. Probability of the Foundation of Computing subject

Knowledge Area	Probability
AL - Algorithms and Complexity	1.60419270661286e-32
AR - Architecture and Organization	1.60419270661286e-32
CN - Computational Science	8.213466657857843e-30
DS - Discrete Structures	1.60419270661286e-32
GV - Graphics and Visualization	2.566708330580576e-31
HCI - Human-Computer Interaction	8.213466657857843e-30
IAS - Information Assurance and Security	8.213466657857843e-30
IM - Information Management	2.566708330580576e-31
IS - Intelligent Systems	2.0533666644644607e-30
NC - Networking and Communications	1.283354165290288e-31
OS - Operating Systems	6.41677082645144e-32
PBD - Platform-based Development	1.60419270661286e-32
PD - Parallel and Distributed Computing	5.133416661161152e-31
PL - Programming Languages	2.566708330580576e-31
SDF - Software Development Fundamental	6.41677082645144e-32
SE - Software Engineering	2.0533666644644607e-30

Example of an accuracy calculation, based on previous Formula, One student transcript with Student ID 16650012 based on the output system has the top three knowledge namely IAS, HCI, and SE. Meanwhile, based on expert input, the transcript has the top three knowledge, namely SE, CN, and SP, the calculation is as follows.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \; x\; 100\%$$

$$= \frac{1+13}{1+13+2+2} \; x\; 100\%$$

Where:
TP = 1 (SE)
TN = 13 (18 knowledge area except IAS, HCI, CN, SP, SE)
FP = 2 (CN and SP)
FN = 2 (IAS and HCI)

From the example above, the accuracy of the student transcript with Student ID 16650012 is 77.78%. That is still from one transcript and there will still be 19 more trials, then final accuracy will be obtained from an average of 20 trials. The remaining 19 students have more transcripts whose accuracy will be shown in Table 2 below.

Table 2. Probability of the Foundation of Computing subject

NIM	Output System	Expert Input
16650012	IAS, HCI and SE	SE, SP, and CN
16650013	IAS, HCI, and CN	SP, SE and GV
16650015	IAS, CN and SE	CN, SE and SP
16650016	IAS, CN and SE	IAS, IM and SP
16650020	IAS, HCI and CN	CN, GV and SE
16650021	IAS, HCI and CN	IAS, SE and CN
16650029	IAS, HCI and CN	SE, PL and SP
16650031	IAS, SE and HCI	IAS, SP and IM
16650035	IAS, SE and CN	SE, IAS and GV
16650037	IAS, HCI and SE	PBD, SE, and GV
16650039	IAS, SE and CN	SE, CN and PBD
16650040	IAS, HCI and CN	GV, CN and IM
16650056	IAS, HCI and CN	SE, CN and IM
16650084	IAS, CN and SE	PBD, IAS and GV
16650085	IAS, HCI and CN	SE, SP and IM
16650086	IAS, HCI and CN	PD, CN, and SDF
16650087	IAS, HCI and CN	CN, GV and SDF
16650097	IAS, HCI and CN	CN, IAS and SDF
16650102	IAS, CN and HCI	SP, CN and GV
16650115	IAS, HCI and CN	CN, IM and SP

Data in the form of expert input in each of the students above were obtained from the input of several experts. To explain how input from the expert was obtained, it will be explained with an example of one of the students listed in Table 2. Students with NIM 16650012 get input from experts namely SE, SP, and CN. The knowledge area is obtained from the input of five experts in detail in Table 3 below.

Table 3. Expert Input for Student with ID 16650012

Expert Input	Knowledge Area
Expert 1	SE, SP, and CN
Expert 2	CN, SE and SP
Expert 3	IM, SE and SP
Expert 4	GV, SE and SP
Expert 5	CN, SF and GV

From Table 4.3 above, it can be seen that the input of the five experts varies, so to determine the three selected

knowledge areas is to calculate the frequency of each knowledge area based on the input of some of these experts. Therefore based on input from the five experts, it can be determined that the top three knowledge areas for students with ID 16650012 are SE, SP, and CN.

Table 4. Result of Calculation – Accuracy

NIM	TP	FP	TN	FN	Accuracy
16650012	1	2	13	2	77. 77777778%.
16650013	0	3	12	3	66.66666667%
16650015	2	1	14	1	88.88888889%
16650016	1	2	13	2	77.77777778%
16650020	1	2	13	2	77.77777778%
16650021	2	1	14	1	88.88888889%
16650029	0	3	12	3	66.66666667%
16650031	1	2	13	2	77.77777778%
16650035	2	1	14	1	88.88888889%
16650037	1	2	13	2	77.77777778%
16650039	2	1	14	1	88.88888889%
16650040	1	2	13	2	77.77777778%
16650056	1	2	13	2	77.77777778%
16650084	1	2	13	2	77.77777778%
16650085	0	3	12	3	66.66666667%
16650086	1	2	13	2	77.77777778%
16650087	1	2	13	2	77.77777778%
16650097	2	1	14	1	88.88888889%
16650102	1	2	13	2	77.77777778%
16650115	1	2	13	2	77.77777778%

The table above shows the details of the accuracy variables in the form of TP, FP, TN, and FN which were obtained from two classes, namely positive and negative, positively represented as the top three knowledge areas, and negative classes represented as 15 unselected knowledge areas.

From the results of the 20 test results above, it can be seen that the accuracy of the academic performance identification system of this student is 78.95%. These results are results that are purely based on existing data, namely terms that are in the subject and terms that are in the knowledge area. And the value that is on each student transcript that also gets input from the results of identification by experts.

F. Discussion

In this discussion section, it will explain the analysis of the subject classification results, student academic performance, and accuracy. The first is the result of the classification, it has been mentioned that there is a distribution of subjects to knowledge areas and their contribution rates. The results of the classification are the result of calculations using the Naïve Bayes method which depends on the probability term. In this study, all results are based on data, meaning the data conditions that determine the results of the classification, therefore the role of the RPS document is very important because it determines the classification, there are many differences of opinion from experts regarding the classification of subjects in the knowledge area, but again that the results in the study This is based on the terms contained in the RPS document.

Next is the academic performance of students, academic performance is also based on data which involves scores and subject or credit load. The results of this academic performance calculation are based on the contribution of each subject to the knowledge area. If we look back at the top, the output results of this system tend to be similar because indeed the distribution of the contribution of subjects to the knowledge area tends to be higher to the knowledge area and it is based on the dataset.

Next the third is accuracy, it is mentioned in the experimental results that the accuracy of this system is 78.95%. It can be seen in Table 4.3 above that the accuracy has only a few kinds because the first input from experts is the top three knowledge areas which cause the highest likelihood for TP is 3, as well as for others. And the error in the system when compared with expert input is the classification generated by the system based on the dataset and the expert version classification has a difference that also affects the calculation in the final output of the application, which is the academic performance of students through a formula involving credits (sks) and contribution rates too.

V. CONCLUSION

From the result of the implementation and experiments that have been carried out by researchers, it can be concluded that the accuracy from this study is 78.95%, obtained from the output system test result and input from the experts. This accuracy result is influenced by several factors, where the biggest factor is the result of the classification of subjects in the system based on pure datasets of RPS documents in Jurusan Teknik Informatika UIN Maulana Malik Ibrahim Malang and IEEE/ACM Computer Science Curricula 2013.

REFERENCES

[1] Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. Expert Systems with Applications, 36(3 PART 1), 5432–5435. https://doi.org/10.1016/j.eswa.2008.06.054.

[2] Jiang, L., Li, C., Wang, S., & Zhang, L. (2016). Deep feature weighting for naive Bayes and its application to text classification. Engineering Applications of Artificial Intelligence, 52, 26–39. https://doi.org/10.1016/j.engappai.2016.02.002.

[3] Alkubaisi, G. A. A. J, Kamaruddin, S. S., & Husni, H. (2018). Stock Market Classification Model Using Sentiment Analysis on Twitter Based on Hybrid Naive Bayes Classifiers. Computer and Information Science, 11(1), 52. https://doi.org/10.5539/cis.v11n1p52.

[4] Fanissa, S., Fauzi, M. A., & Adinugroho, S. (2018). Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naïve Bayes dan Seleksi Fitur Query Expansion Ranking. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Vol 2, 8, 2766-2770.

[5] Indrayuni, E. (2019). Klasifikasi Text Mining Review Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes. Jurnal Khatulistiwa Informatika, 7(1), 29–36. https://doi.org/10.31294/jki.v7i1.1.

[6] Pratama, B. Y., & Sarno, R. (2016). Personality classification based on Twitter text using Naive Bayes, KNN and SVM. Proceedings of 2015 International Conference on Data and Software Engineering, ICODSE 2015, 170–174. https://doi.org/10.1109/ICODSE.2015.7436992.
[7] Taheri, S., Mammadov, M., & Bagirov, A. M. (2010). Improving Naive Bayes classifier using conditional probabilities. Conferences in Research and Practice in Information Technology Series, 121(December 2011), 63–68.
[8] Crysdian, C. (2017). Performance measurement without ground truth to achieve optimal edge. International Journal of Image and Data Fusion, 9(2), 170–193. https://doi.org/10.1080/19479832.2017.1384764.