International Transactions on Artificial Intelligence (ITALIC)
Vol. 2 No. 1, November 2023, pp. 60~75
P-ISSN: 2963-6086, E-ISSN: 2963-1939, DOI: 10.33050

Playing Smart with Numbers: Predicting Student Graduation Using the Magic of Naive Bayes

Shilpa Mehta[1]
[1] Department of Electrical and Electronic Engineering, Auckland University of Technology, New Zealand
e-mail: shilpa.mehta@aut.ac.nz

Article Info

Article history:

Received August 25, 2023
Revised August 29, 2023
Accepted November 23, 2023

Keywords:

Predictive Modeling,
Student Graduation Prediction,
Naive Bayes Algorithm,
Educational Data Analysis,
Smart Analytics in Education

ABSTRACT

The quality of a higher education institution is often measured by the accreditation granted by the National Accreditation Agency for Higher Education (BAN-PT). In this context, one of the primary assessment criteria is the graduation rate of students. An intriguing study employs the Naive Bayes algorithm to forecast whether students will graduate on time or face delays. The resulting predictive outcomes offer valuable insights and input to universities for enhancing their educational standards. The Naive Bayes method brings its unique advantages, particularly in predicting graduation rates based on real-world data. This ensures that the generated predictions can be relied upon and utilized as guidelines for future projections. This predictive mechanism encompasses 14 pivotal factors. These factors include gender, student status, age, marital status, performance across semesters 1 through 8, as well as cumulative performance, culminating with the information of whether a student passed or not. Within this study, data from 302 students of the 2018 cohort were involved. Data processing was carried out using the Python programming language within the Jupyter Notebook environment. The results unveil an impressive accuracy rate, reaching 85%. In terms of precision, the prediction for delays achieved a value of 0.42, while timely graduation prediction scored 0.95. Furthermore, the accuracy in identifying delay cases reached 0.65, compared to 0.88 accuracy for timely predictions. The f1 score for delay predictions stood at 0.51, while timely graduation predictions reached 0.91. These results illustrate that this algorithmic approach is capable of providing accurate and well-balanced insights into student graduation predictions.

Corresponding Author:

Shilpa Mehta
Department of Electrical and Electronic Engineering, Auckland University of Technology, New Zealand
Email: shilpa.mehta@aut.ac.nz

1. INTRODUCTION

In the realm of education, higher education institutions are challenged to orchestrate a realm of quality education that births students who are not only competent, but also creative and brimming with competitiveness. The quality of Indonesian higher education institutions is reflected in the accreditation bestowed by the National Accreditation Agency for Higher Education, commonly known as BAN-PT. Expanding the horizon, the rate of graduation success stands as a pivotal gauge, shaping the evaluation that determines the quest for recognition [1], following the footsteps of regulations etched in the Appendix of National Accreditation Agency for Higher Education Regulation No. 23 of 2022.

The evaluation of graduation rates conducted so far has mostly relied on graduation registration data, often overlooking students who might be facing academic or administrative challenges. On the other hand, the university's response to students who do not graduate on time can be carried out through methods of persuasion, guidance, and mentoring, encouraging students to promptly complete their studies.

The challenge faced by universities is the absence of an integrated system for predicting student graduation. The consequences of this situation include, among others: the Academic and Student Administration Bureau cannot ensure that an entire cohort graduates on time, leading to a scenario where students find themselves without a solution to the issue of delayed graduation.

Drawing inspiration from the aforementioned issues, there arises a need for a student graduation prediction application within the university environment [2]. This application encompasses comprehensive data about students grouped within a single cohort, spanning various study programs. With the advent of this application, we embark on a new chapter where student quality and university accreditation are not only monitored but also continually enhanced on the journey towards excellence.

2. LITERATURE REVIEW

2.1. Literature Review

Regarding the matter of predicting student graduations, numerous studies have been conducted across various universities, employing a variety of methodologies. Previous research endeavors were undertaken by Armansyah and Rakhmat Kurniawan Ramli in a study titled "A Naive Bayes Approach to Predicting Timely Graduation of Students" [3]. They grappled with the challenge of declining graduation rates stemming from the disparity between incoming students and those who graduate, leading to detrimental effects on study programs across multiple aspects. They adopted an experimental approach and employed the naïve bayes method. The outcomes of this research yielded predictions of student graduation rates that showcased exceptional performance of this prediction model, reaching an accuracy of 100%.

Moving forward, let's delve into the study by Lydia Yohana Lumban Gaol, M. Safii, and Dedi Suhendro titled "Anticipating Successful Student Graduations in Stikom Tunas Bangsa's Information Systems Program Through the Implementation of the C4.5 Algorithm" [4]. The puzzle they tackle stems from the crucial role that graduation plays as a vital yardstick in evaluating the accreditation of higher education institutions. As a result, when an increasing number of students graduate within the designated time frame, it directly contributes to the institution's accreditation assessment climbing higher. Their chosen methodology revolves around the deployment of the C4.5 classification algorithm, deftly sifting through both numeric and categorical attributes within the dataset. The culmination of their research opens up a treasure trove of predictive data on student graduations, underlining that the most influential factor determining student success is the GPA attribute [5].

Stepping into the next research expedition, we enter the realm of exploration crafted by Ray Mondow Sagala with the title "Unveiling Student Graduation Forecasts Through the

K-means Algorithm in the World of Data Mining" [6]. The heart of the issue arises from the significant urgency surrounding student graduation, as the common thread interweaving various courses forms an inevitable connection. K-means, used as a tool, unlocks the door to interpreting research data into a series of numbers. The final hue of this research journey is a canvas of predictive data, proving that through the stride of k = 3 out of a total of 118 manipulated data points, 13 students were found not to conclude their journey, followed by 36 students opting for the path with satisfactory grades, and finally, 69 students anchoring within the realm of stellar scores.

In the ensuing research endeavor, the ladder of knowledge is scaled by Nursetia Wati with the title "Envisioning Student Graduations by Applying the K-Nearest Neighbor Approach Based on Particle Swarm Optimization" [7]. The foundation of this challenge is obstructed graduation rates, especially in the realm of the Faculty of Engineering. As this situation arises, like the rustling of leaves driven by the wind, every study program tirelessly delves into the journey to enhance the graduation rates, aiming to reach the pinnacle of desired quality. Employing the method of classifying data based on the distance from new to existing data, curiosity answers the call, and the experimental method named K-Nearest Neighbor (KNN) is chosen as the complement. The result of this research ripple, as it turns out, takes the form of notes, recording that the conducted testing yielded the best value when the K-Nearest Neighbor algorithm was applied.

2.2. Theoretical Framework

Unveiling the curtain of knowledge in the realm of data, we come across the term "data mining" as a tool to excavate intellectual treasures within the database vault. In this process, mathematics, statistical techniques, machine learning, and even artificial intelligence play pivotal roles. They collaborate, comb through data, and formulate the identification and extraction of various valuable pieces of information, as well as weighted knowledge from an array of expansive databases [8]. Maulana and Fajrin also share an intriguing perspective that data mining in the realm of research is fundamentally not a novel topic. It emerges as an added-value agent capable of enhancing the effectiveness of various previously employed techniques, thus addressing an array of challenges we commonly encounter [9].

Like a ready-to-eat dish, an application is a program menu waiting to be served to execute various commands of its users. In alignment with the given commands, it artfully crafts detailed outcomes, just as desired when preparing a meal. However, an application's role doesn't merely stop at being a digital cook; it also becomes a catalyst for solving puzzles through the application of one of the various data processing recipes available. Aligned with specific hopes or goals, it transforms into a data-grinding machine that functions harmoniously according to its capabilities. Delving deeper, we encounter another perspective that defines an application as a pre-assembled machine component ready for operation by its enthusiasts [10].

Once upon a time, Al Khawarizmi, a scholar from Persia, breathed life into algorithms for the first time. Like a seed sown, initially, algorithms were used to formulate solutions for arithmetic problems. However, algorithms underwent transformations over time, assuming the role of cracking various mathematical puzzles. Delving deeper, algorithms also weave an inseparable thread with mathematics, anchoring themselves at the heart of the world of knowledge [11]. Through another lens, T.S. Alasi mentioned that algorithms are a sequence of logical steps, speaking the language of order. They lead us on a journey through the forest of problems with a well-organized and systematic route [12].

Prediction is a mystical endeavor that leads us to peek through the door of the future, estimating the array of possibilities that might unfold there. It's like unearthing a magical chest of past data, sculpting forecasts guided by the stars of indicators. Various challenges require the enchantment of prediction, including peeling back layers in the tale of prices, unveiling the production veil, or dissecting the secrets of graduation rates—and many more [13].

Classification is like assembling puzzle pieces of data, carefully putting them together to predict the characteristics of new data. Just like a detective grouping evidence based on the clues left behind, classification uses existing data as a foundation to guess the nature of unfamiliar data. In the realm of classification, there are two main ingredients: test data, like gems whose light is being examined, and training data, the stepping stones that guide the learning process [14].

Imagine Jupyter Notebook as an enchanting laboratory holding three magical languages: Julia, the clever wizard; Python, the versatile magician; and R, the alchemist of numbers. In its ritual, Jupyter Notebook combines the powers of these three languages into a mesmerizing interactive spectacle. Like a sorcerer turning objects into gold, this web application transforms thoughts into beautiful computational documents. Undisturbed, uncomplicated, solely focused on the magic of the document itself [15].

Python, the magical language that traverses various platforms like astral beings exploring the universe. Interactive like conversing with genies, it swiftly and gracefully responds to every command call. Its magical prowess is undeniable, comprehending human language with elegance and charm. The enchanting codes in this language will be transformed into secret codes known as byte code before the execution spell is cast. Like embarking on a journey to learn the art of magic, understanding classification and Python is the first step in mastering this mysterious world. Throughout the adventure, you will comprehend how to piece together clues from the past to unlock the gates of the future [16].

3. RESEARCH METHODOLOGY

3.1. Stages of Research

This research journey begins by crafting challenging questions and delving into the realm of hidden literature. Like a detective gathering clues from various angles, we gather data through observation and documentation methods, laying the foundation to unravel the existing mysteries [17].

Having completed the initial phase, we step into the next chapter, gathering the "trails" of data from the required students. Like assembling puzzle pieces, 395 sets of data from the 2018 batch of students who have completed their study journey are collected across 16 attributes. The subsequent action involves crafting and cleansing this data, akin to arranging bricks before constructing a house. Out of the 395 data points, 302 of them, complete with 14 relevant attributes, will serve as the main ingredients when the naive Bayes algorithm comes into play.

This is the spotlight moment, where the naive Bayes algorithm takes center stage as the hero of this narrative. Utilizing the magic of the Python programming language, this algorithm is activated to meticulously unravel the data and provide potential hidden answers. Not dissimilar to the power of a wizard concocting magical potions [18].

As the experiments unfold, the constructed model is tested as if facing a magical trial. And behind the curtain's veil, evaluation and validation take on the leading roles. Like a sorcerer assessing the success of an incantation, we carefully evaluate the testing results and ensure their authenticity.

Thus, from scientific steps to magical performances, this research is a journey to unveil mysteries, gather truths, and carve new pathways in the realm of knowledge.

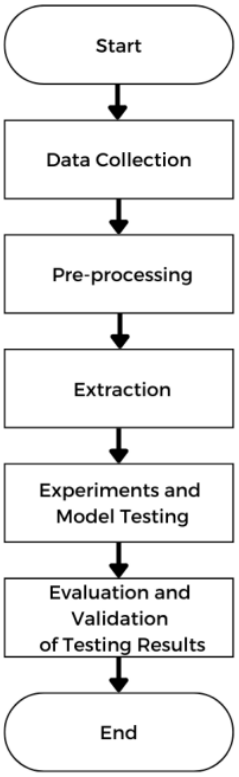

Figure 1. Research Flowchart

3.2. Data Collection

As a strategy to gather valuable information resources, the researcher opted for the following steps:

1. Observation Method
   Embark on a journey of direct observation, immersing yourself in every hidden detail of the field. The insights gained from this observation are not merely visual snapshots, but rather the core components that will paint the path forward for a dynamic system under the company's spirited umbrella [19]. The author also assumes the role of an observer within the observed university campus.
2. Interview Method
   Engaging in direct interviews through a carefully crafted set of questions, we venture into the realm of inquiry, guided towards the Bureau of Academic Administration and Student Affairs. Like an explorer unearthing treasures from conversations, we discern the pieces of information needed.
3. Library Method
   Delve into the realm of knowledge by excavating intellectual treasures through the pages of books, scholarly journals, and the digital footprints in the vast sea of the internet. Like a mind archaeologist, we unearth valuable artifacts that fortify the foundations and outcomes of this research.

3.3. Data Analysis

The student landscape that emerges in this research paints a portrait of the students from the 2018 cohort at Universitas. They are captured through intriguing variables such as gender traits, student statuses, marital journeys, age, records of Semester Grade Point Averages (IPS) spanning from the first semester to the eighth, and also Cumulative Grade Point Averages (IPK) that provide a glimpse into the extended journey of academic achievements. Amidst this center stage of attention, the target class dances with joy, depicting the graduation destination that awaits at the end of the journey, whether it lands accurately or might encounter a minor time obstacle [20].

4. RESULTS AND DISCUSSION

4.1. Data Collection

Like tracing the footsteps of an adventure, the process of data collection brings its narrative to life in the field and from the expanse of the observed university's website. Much like painting piece by piece of a puzzle, data is meticulously gathered and poured into Microsoft Excel files in the xlsx format. As if weaving the tales of the 2018 cohort that has traversed their academic journey, this data arrives in a total of 395 entities woven with 16 unique attributes. As if presenting a beautiful painting, examples of the showcased data can be found within mysterious images under the sunlight[21].

SIN	Name	Gender	Student Status	Age	Marital Status
18314252601	ADAM SUJATMIKO	Male	Student	22	Not yet
18314252602	ADITYA RAMADHAN	Male	Student	22	Not yet
18314252603	AGAM	Male	Student	22	Not yet
18314252604	ANGGI SEPTIANI	Female	Student	22	Not yet
18314252605	AYU PRIHARTINI	Female	Student	21	Not yet
18314252606	BUDIMAN FAJAR PEBRIANTO	Male	Student	21	Not yet
18314252607	DIAN FAHMI SEPTIAWAN	Female	Student	22	Not yet
18314252608	ELEN ANGELICA	Female	Student	22	Not yet
18314252609	FARHAN ALI RAMADHAN	Male	Student	21	Not yet
18314252610	INDRA SAPUTRA RAMADHAN	Male	Student	22	Not yet
18314252611	MIKEL YUKOZU	Male	Student	22	Not yet
18314252612	MOEHAMAD RIZKY FADILAH	Male	Student	21	Not yet
18314252613	MUHAMAD IKHSAN PRATAMA	Male	Student	22	Not yet
18314252614	MUHAMAD RAFLI	Male	Student	21	Not yet
18314252615	MUHAMMAD RIFKI	Male	Student	21	Not yet
18314252616	NATASYA NUR SAFITRI	Female	Student	22	Not yet
18314252617	RIYAN ADI PRASETYA	Male	Student	22	Not yet
18314252618	ROBIATUL HADAWIYAH	Male	Student	22	Not yet
18314252619	ROY ADI LESMANA	Male	Student	21	Not yet
18314252620	WAHYU	Male	Student	21	Not yet

Figure 2. Sample Data

IPS1	IPS2	IPS3	IPS4	IPS5	IPS6	IPS7	IPS8	IPK	Graduated
3,45	3,50	3,65	3,63	3,63	3,81	3,65	3,32	3,58	Late
3,73	3,50	2,96	3,63	3,84	3,76	3,82	3,24	3,56	Late
3,14	3,48	3,65	3,75	3,53	3,86	3,82	3,41	3,58	On time
2,86	3,48	2,91	3,38	3,58	3,90	3,82	2,79	3,34	Late
3,00	3,38	3,61	3,38	3,68	4,00	3,82	2,57	3,43	On time
3,36	3,76	3,87	3,63	4,00	3,86	4,00	3,00	3,78	On time
3,45	3,33	3,39	3,5	3,42	3,43	3,82	3,18	3,44	On time
3,23	3,68	3,65	4,00	3,84	4,00	3,82	4,00	3,73	On time
3,14	3,41	3,65	3,88	4,00	3,89	3,65	3,34	3,62	On time
3,55	4,00	3,87	4,00	3,84	3,81	3,65	4,00	3,84	On time
3,50	4,00	3,87	4,00	3,68	4,00	3,65	4,00	3,84	On time
3,36	3,67	3,65	3,75	4,00	3,90	4,00	4,00	3,73	On time
3,64	3,86	3,87	4,00	3,84	4,00	3,65	3,00	3,86	On time
3,73	3,71	3,91	4,00	4,00	4,00	3,69	4,00	3,88	On time
3,36	3,62	4,00	3,88	3,74	4,00	3,65	3,00	3,77	On time
3,27	3,48	3,87	3,75	3,58	3,76	3,65	3,76	3,64	Late
3,59	3,76	3,91	3,63	3,84	4,00	4,00	4,00	3,80	On time
2,91	3,48	3,17	3,63	3,26	3,57	3,82	2,56	3,30	Late
3,50	3,48	3,57	3,75	3,58	3,62	3,00	3,74	3,53	On time
3,50	3,71	3,57	3,63	3,84	4,00	4,00	4,00	3,69	On time

Figure 3. Additional Data Samples

4.2. Pre-processing

Based on the collected research findings, some intriguing revelations come to light. A total of 395 datasets of students who have completed their academic journey are recorded. When categorized by study duration, three main groups emerge. Firstly, there are those who graduated on time within 7 semesters (3.5 years) or 8 semesters (4 years). Then, there is a group of students who surpassed these timeframes, concluding their studies in more than 8 semesters[22].

However, just as sifting for gems in the sand, not all data and information gathered can be readily utilized. An initial, creative process is necessary to refine this data, akin to tending a garden to yield better results. And, do not miss it – under the spotlight's beam, there are 16 data attributes that have not undergone this process. Behold, the list of

treasures to be further unearthed[23].

Table 1. Data Before Pre Processing

No	Date Name	Data Type
1	SIN	Character
2	Name	Character
3	Gander	Categorical
4	Student Status	Categorical
5	Age	Numerical
6	Marital Status	Categorical
7	IPS 1	Numerical
8	IPS 2	Numerical
9	IPS 3	Numerical
10	IPS 4	Numerical
11	IPS 5	Numerical
12	IPS 6	Numerical
13	IPS 7	Numerical
14	IPS 8	Numerical
15	IPK	Numerical
16	Graduated	Categorical

As for the preprocessing techniques employed by the author, they encompass:

1. Polishing the Data Brilliance, by sweeping away all vacant and incomplete data. For instance, the records of inactive or departed students, their data erased due to the incomplete payload of course grades. Consequently, a mere 302 usable data remain from the initial 395, implying a data cleansing of 23.54%. This process acts as a cleansing beam in the data preprocessing phase, ensuring no gaps are left behind.
2. Squeezing the Data Spectrum, aims to grasp the relevant traces within records along with the suitable count of attributes engaged in the mining process. This resembles not inviting attributes like SIN and name to the gathering, deemed somewhat unrelated or less impactful. This signifies that in the swift mining dance, only a handful of attributes are embraced – such as gender, age, student status, marital status, Semester Grade Index (IPS) from semester 1 to 8, as well as Cumulative Grade Index (CGPA), and Graduation status.
3. Interweaving Data: Shifting from Categorical Flair to Numeric Charm. The "gender" attribute, once entwined with the words "male" and "female," is now transformed into 0 for males and 1 for females. The "student status" attribute, previously branching into "working" and "student," is now swapped to 0 for those working and 1 for students. Next, the "marital status" attribute, previously linked to "married" and "unmarried," is ignited to 0 for those already married and 1 for those still waiting. Lastly, the "graduation" attribute, formerly narrating "on time" and "late," is

reshaped to 0 for the late ones and 1 for the punctual achievers.

After the preprocessing waltz reaches its final bow, the next act unfolds – a dance into the mining process upon the cluster of 302 student data. They all play their parts across 14 attributes that have passed through the scale harmonization and evaded the potential of missing values. Here are the intricate movements engraved on the following page:

Table 2. Data After Preprocessing

No	Date Name	Data Type
1	Gander	Numerical
2	Student Status	Numerical
3	Age	Numerical
4	Marital Status	Numerical
5	IPS 1	Numerical
6	IPS 2	Numerical
7	IPS 3	Numerical
8	IPS 4	Numerical
9	IPS 5	Numerical
10	IPS 6	Numerical
11	IPS 7	Numerical
12	IPS 8	Numerical
13	IPK	Numerical
14	Graduated	Numerical

❒ P-ISSN: 2963-6086, E-ISSN: 2963-1939

4.3. Mining

Undergoing a numerical metamorphosis, categorical data undergoes a transformation, as illustrated in the informational canvas below:

Gender	Student Status	Age	Marital Status
0	1	22	1
0	1	22	1
0	1	22	1
1	1	22	1
1	1	21	1
0	1	21	1
1	1	22	1
1	1	22	1
0	1	21	1
0	1	22	1
0	1	22	1
0	1	21	1
0	1	22	1
0	1	21	1
0	1	21	1
1	1	22	1
0	1	22	1
0	1	22	1
0	1	21	1
0	1	21	1

Figure 4. Data After Transformation

IPS1	IPS2	IPS3	IPS4	IPS5	IPS6	IPS7	IPS8	IPK	Graduated
3,45	3,50	3,65	3,63	3,63	3,81	3,65	3,32	3,58	0
3,73	3,50	2,96	3,63	3,84	3,76	3,82	3,24	3,56	0
3,14	3,48	3,65	3,75	3,53	3,86	3,82	3,41	3,58	1
2,86	3,48	2,91	3,38	3,58	3,90	3,82	2,79	3,34	0
3,00	3,38	3,61	3,38	3,68	4,00	3,82	2,57	3,43	1
3,36	3,76	3,87	3,63	4,00	3,86	4,00	3,00	3,78	1
3,45	3,33	3,39	3,5	3,42	3,43	3,82	3,18	3,44	1
3,23	3,68	3,65	4,00	3,84	4,00	3,82	4,00	3,73	1
3,14	3,41	3,65	3,88	4,00	3,89	3,65	3,34	3,62	1
3,55	4,00	3,87	4,00	3,84	3,81	3,65	4,00	3,84	1
3,50	4,00	3,87	4,00	3,68	4,00	3,65	4,00	3,84	1
3,36	3,67	3,65	3,75	4,00	3,90	4,00	4,00	3,73	1
3,64	3,86	3,87	4,00	3,84	4,00	3,65	3,00	3,86	1
3,73	3,71	3,91	4,00	4,00	4,00	3,69	4,00	3,88	1
3,36	3,62	4,00	3,88	3,74	4,00	3,65	3,00	3,77	1
3,27	3,48	3,87	3,75	3,58	3,76	3,65	3,76	3,64	0
3,59	3,76	3,91	3,63	3,84	4,00	4,00	4,00	3,80	1
2,91	3,48	3,17	3,63	3,26	3,57	3,82	2,56	3,30	0
3,50	3,48	3,57	3,75	3,58	3,62	3,00	3,74	3,53	1
3,50	3,71	3,57	3,63	3,84	4,00	4,00	4,00	3,69	1

Figure 5. Data After Transformation (Continuation)

4.4. Experimentation and Model Testing

The author harnesses the power of Jupyter software to embark on a journey of data experimentation concerning student graduation, employing the Naive Bayes methodology. Much like a digital alchemy expert, they skillfully blend information in pursuit of shimmering discoveries[24], [25].

1. Summoning the Required Library

```
import pandas as pan
import numpy as npy
```

Figure 6. Command to Invoke a Library

2. Reading student graduation data from an Excel file

```
mahasiswa = pan.read_excel("Student Graduation Data.xlsx")
mahasiswa.head(302)
```

Figure 7. Command to Read Excel Data

3. The displayed results of the data are as follows:

Out [2] :

	Gender	StudentStatus	Age	MaritalStatus
0	0	1	22	1
1	0	1	22	1
2	0	1	22	1
3	1	1	22	1
4	1	1	21	1
...	...	...	...	...
297	1	1	22	1
298	0	1	22	1
299	0	1	22	1
300	0	1	21	1
301	0	1	21	1

302 rows x 14 columns

Figure 8. Python Data Output

IPS1	IPS2	IPS3	IPS4	IPS5	IPS6	IPS7	IPS8	IPK	Graduated
3.45	3.50	3.65	3.63	3.63	3.81	3.65	3.32	3.58	0
3.73	3.50	2.96	3.63	3.84	3.76	3.82	3.24	3.56	0
3.14	3.48	3.65	3.75	3.53	3.86	3.82	3.41	3.58	1
2.86	3.48	2.91	3.38	3.58	3.90	3.82	3.79	3.34	0
3.00	3.38	3.61	3.38	3.68	4.00	3.82	2.57	3.43	1
...	...	...	...	...	...	...	...	...	...
3.78	3.78	3.40	3.70	3.73	3.76	3.71	4.00	3.69	0
3.78	3.91	3.65	3.70	3.86	3.79	3.71	4.00	3.79	0
3.11	3.05	2.92	3.09	3.63	3.38	2.89	3.00	3.14	0
3.58	3.80	2.83	3.09	4.00	3.50	3.06	3.00	3.20	0
3.68	3.70	3.83	3.82	3.94	4.00	3.44	3.00	3.75	0

Figure 9. Extended Python Data Display

4. As if orchestrating a dance of thoughts, these attributes are embraced within two shimmering circles: X and Y. Variable X cradles in its arms the attributes of gender, student status, age range, marital status, along with notes traversing from semester one to eight, adorned with the jewel of GPA. Meanwhile, within the embrace of Variable Y, only one splendid story resides: the story of graduation.

```
x = mahasiswa.drop(["Lulus"], axis=1)
x.head(302)
```

Figure 10. Command to Discard Passed Attribute

Out [3] :

	Gender	StudentStatus	Age	MaritalStatus
0	0	1	22	1
1	0	1	22	1
2	0	1	22	1
3	1	1	22	1
4	1	1	21	1
...	...	...	...	...
297	1	1	22	1
298	0	1	22	1
299	0	1	22	1
300	0	1	21	1
301	0	1	21	1

302 rows x 14 columns

Figure 11. Variable Y Data

IPS1	IPS2	IPS3	IPS4	IPS5	IPS6	IPS7	IPS8	IPK
3.45	3.50	3.65	3.63	3.63	3.81	3.65	3.32	3.58
3.73	3.50	2.96	3.63	3.84	3.76	3.82	3.24	3.56
3.14	3.48	3.65	3.75	3.53	3.86	3.82	3.41	3.58
2.86	3.48	2.91	3.38	3.58	3.90	3.82	3.79	3.34
3.00	3.38	3.61	3.38	3.68	4.00	3.82	2.57	3.43
...	...	...	...	...	...	...	...	...
3.78	3.78	3.40	3.70	3.73	3.76	3.71	4.00	3.69
3.78	3.91	3.65	3.70	3.86	3.79	3.71	4.00	3.79
3.11	3.05	2.92	3.09	3.63	3.38	2.89	3.00	3.14
3.58	3.80	2.83	3.09	4.00	3.50	3.06	3.00	3.20
3.68	3.70	3.83	3.82	3.94	4.00	3.44	3.00	3.75

Figure 12. Continued X Variable Data

```
y = mahasiswa["Lulus"]
y.head(302)
```

Figure 13. Command for Grouping Passed Attributes into Variable Y

```
Out [4] :  0      0
           1      0
           2      0
           3      1
           4      1
                  ..
           297    1
           298    0
           299    0
           300    0
           301    0
           Name :  Lulus,  Length:  302,  dtype:  int64
```

Figure 14. Variable Y Data

1. Importing Gaussian Naive Bayes Model

```
from sklearn.naive_bayes import GaussianNB
```

Figure 15. Command to Import GaussianNB Model

2. Calling the GaussianNB Function

```
nbc = GaussianNB()
```

Figure 16. Command to Invoke GaussianNB

3. Generating Training Data

```
data_training = nbc.fit(x,y)
```

Figure 17. Command to Create Training Data

4. Performing Data Prediction on Training Data

```
y_predict = data_training.predict(x)
print(y_predict)
```

Figure 18. Command for Predicting Training Data

5. Prediction Results of Training Data

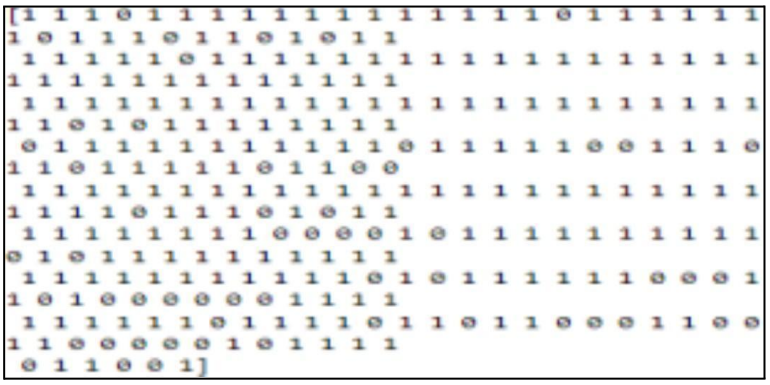

Figure 19. Training Data Prediction Results

6. Envisioning Input Data Prediction, by peering into the future with the addition of the following data:
   Gender = 1
   StudentStatus = 0
   Age = 23
   MaritalStatus = 1
   IPS1 = 3.6
   IPS2 = 3.5
   IPS3 = 3
   IPS4 = 3.8
   IPS5 = 3.4
   IPS6 = 3.21
   IPS7 = 3.5
   IPS8 = 4
   FinalGPA = npy.average([GPA1, GPA2, GPA3, GPA4, GPA5, GPA6, GPA7, GPA8])
   TestingData = [[Gender, StudentStatus, Age, MaritalStatus, GPA1, GPA2, GPA3, GPA4, GPA5, GPA6, GPA7, GPA8, FinalGPA]]

7. Crafting Predictions Based on Input Data: Deciphering Signs from Numbers

```
y_pred = data_training.predict(Data_Testing)
```

Figure 20. Input Data Prediction Command

8. Printing the Prediction Results

```
if y_pred == 0 :
  hasil = "Terlambat"
elif y_pred == 1 :
  hasil = "Tepat"
else :
  hasil = "Error"

print("Hasil Prediksi Kelulusan Mahasiswa = ", hasil)
```

```
Hasil Prediksi Kelulusan Mahasiswa = Tepat
```

Figure 21. Printing the Prediction Results

9. Based on the infused data, the prediction results of students graduating on time are unveiled through the enchanting strokes of data.

4.5. Evaluation and Validation of Test Results

From the outcome of predicting training data against the core dataset, emerges a gleaming accuracy score of 0.85 or 85%.

```
from sklearn.metrics import accuracy_score
print("Nilai Akurasi = %0.2f % accuracy_score (y, y_predict)")
```

```
Nilai Akurasi = 0.85
```

Figure 22. Printing Accuracy Score

With the staged report of classification results as follows, unveiling the grand spectacle where 88% of students successfully complete their studies on time:

```
from sklearn.metrics import accuracy_score
print(Classification_report(y, y_predict))
```

	precision	recall	fi-score	support
0	0.42	0.65	0.51	37
1	0.95	0.88	0.91	205
accuracy			0.85	302
macro avg	0.68	0.76	0.71	302
weighted avg	0.88	0.85	0.86	302

Figure 23. Prediction Result

5. CONCLUSION

In the course of this research involving the utilization of the Naive Bayes algorithm, various aspects have been examined and analyzed comprehensively. The evaluation and validation of these findings indicate that the Naive Bayes algorithm exhibits an impressive accuracy rate, reach ing 85% out of a total of 302 student data records. Specifically, the late submission precision value reaches 0.42, while the on-time precision stands at 0.95. Similarly, the late submission recall rate is 0.65 and the on-time recall rate is 0.88. The late submission F1-score achieves 0.51, and the on-time F1-score is 0.91.

In its application, the Naive Bayes algorithm has demonstrated the ability to predict student graduation statuses accurately, whether they are on time or delayed. This

undoubtedly holds significant benefits for the university environment, providing valuable additional insights for decision-making processes.

As a guidance for future endeavors, the author provides valuable recommendations. Researchers exploring similar topics are advised to delve into alternative algorithms alongside Naive Bayes, with the aim of achieving superior predictive outcomes. Furthermore, increasing the volume of data within the testing dataset will enhance the overall quality and accuracy of predictive results. By doing so, these steps will undoubtedly make positive contributions to the advancement of research and applications in the future.

REFERENCES

[1] Kamagi, D.H. and Hansun, S. (2014) 'Implementasi Data mining Dengan Algoritma C4.5 Untuk Memprediksi Tingkat Kelulusan mahasiswa', Jurnal ULTIMATICS, 6(1), pp. 15–20. doi:10.31937/ti.v6i1.327.

[2] Prawiyogi, A.G. and Widayanti, R., 2023. Exploratory Activities in Educational Games using Fuzzy Logic. International Transactions on Artificial Intelligence, 1(2), pp.188-194.

[3] Gaol, L.Y.L., Safii, M. and Suhendro, D., 2021. Prediksi Kelulusan Mahasiswa Stikom Tunas Bangsa Prodi Sistem Informasi Dengan Menggunakan Algoritma C4. 5. Brahmana: Jurnal Penerapan Kecerdasan Buatan, 2(2), pp.97-106.

[4] S. Royan, A. Yulian, and Syaechurodji, "Implementasi Data Mining Menggunakan Metode Naive Bayes Dengan Feature Selection Untuk Prediksi Kelulusan Mahasiswa Tepat Waktu," J. Ilm. Sains dan Teknol., vol 6, no. 1, pp. 50–61, 2022, doi:10.47080/saintek.v6i1.1467.

[5] L. Y. L. M. S. D. S. Gaol, "Prediksi Kelulusan Mahasiswa Stikom Tunas Bangsa Prodi Sistem Informasi Dengan Menggunakan Algoritma C4.5," Brahmana J. Penerapan Kecerdasan Buatan, vol. 2, no. 2, pp. 97–106, 2021, doi:10.30645/brahmana.v2i2.71.

[6] R. M. Sagala, "Prediksi Kelulusan Mahasiswa Menggunakan Data mining Algoritma K-means," J. TeIKa, vol. 11, no.2, pp. 131–142, 2021.

[7] N. Wati, "PREDIKSI KELULUSAN MAHASISWA MENGGUNAKAN K NEAREST NEIGHBOR BERBASIS PARTICLE SWARM OPTIMIZATION Nursetia Wati," Jtii, vol. 6, no. 2, pp. 118–127, 2021.

[8] D. P. Utomo and M. Mesran, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Dataset Penyakit Jantung," J. Media Inform. Budidarma, vol. 4, no. 2, p. 437, 2020, doi:10.30865/mib.v4i2.2080.

[9] N. A. Sudibyo, Ardymulya Iswardani, Kartika Sari, and Siti Suprihatiningsih, "Penerapan Data Mining Pada Jumlah Penduduk Miskin Di Indonesia," J.Lebesgue J. Ilm. Pendidik. Mat. Mat. dan Stat., vol. 1, no. 3, pp. 199–207, 2020, doi:10.46306/lb.v1i3.42.

[10] I. P. Sari, A. Syahputra, N. Zaky, R. U.Sibuea, and Z. Zakhir, "Perancangan Sistem Aplikasi Penjualan dan Layanan Jasa Laundry Sepatu Berbasis Website," Blend Sains J. Tek., vol. 1, no. 1, pp. 31–37,2022, doi: 10.56211/blendsains.v1i1.67.

[11] N. Khesya, "Mengenal Flowchart dan Pseudocode Dalam Algoritma dan Pemrograman," Preprints, vol. 1, pp. 1–15, 2021, [Online]. Available:https://osf.io/dq45ef.

[12] T. S. Alasi, A. T. Al, and A. Siahaan, "Algoritma Vigenere Cipher Untuk Penyandian Record Informasi Pada Database," J. Inf. Komput. Log., vol. 1, no. 4, 2020, [Online].

Available:http://ojs.logika.ac.id/index.php/jikl.

[13] R. Harun, K. C. Pelangi, and Y. Lasena, “Penerapan Data Mining Untuk Menentukan Potensi Hujan Harian Dengan Menggunakan Algoritma Naive Bayes,” J. Manaj. Inform. dan Sist. Inf., vol. 3, no. 1, pp. 8–15, 2020, [Online]. Available: http://mahasiswa.dinus.ac.id/docs/skripsi/jurnal/19417.

[14] R. Thaniket, Kusrini, and E. T. Luthf, “Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Algoritma Support Vector Machine,” J. FATEKSA J. Teknol. dan rekayasa, vol. 13, no. 2, pp. 69–83, 2019.

[15] D. Ghassa, Aji; Wahyudi, Adi; Tampubolon, Silvia Ovella, Putri, Nurul Afrilia; Rasywir, Errisya; Kisbianty, “Penerapan Data Mining Algoritma Naive Bayes Clasifier Untuk Mengetahui Minat Beli Pelanggan Terhadap INDIHOME,” J. Inform. …, vol. 2, no. 2, pp. 240–247, 2022, [Online]. Available: https://ejournal.unama.ac.id/index.php/jakakom/article/view/33%0Ahttps://ejournal.unama.ac.id/index. php/jakakom/article/download/33/56.

[16] M. Idris, “Implementasi Data Mining Dengan Algoritma Naïve Bayes Untuk Memprediksi Angka Kelahiran,” J. Pelita Inform., vol. 7, no. 3, pp. 1–33, 2019.

[17] Anwar, M.R., 2023. Analysis of Expert System Implementation in Computer Damage Diagnosis with Forward Chaining Method. International Transactions on Artificial Intelligence, 1(2), pp.139-155.

[18] Hudiono, R.K. and Watini, S., 2023. Remote Medical Applications of Artificial Intelligence. International Transactions on Artificial Intelligence, 1(2), pp.182-187.

[19] Jayanagara, O. and Wuisan, D.S.S., 2023. An Overview of Concepts, Applications, Difficulties, Unresolved Issues in Fog Computing and Machine Learning. International Transactions on Artificial Intelligence, 1(2), pp.213-229.

[20] Meria, L., 2023. Development of Automatic Industrial Waste Detection System for Leather Products using Artificial Intelligence. International Transactions on Artificial Intelligence, 1(2), pp.195-204.

[21] Yang, L., & Liu, J. (2022). The Influence of Cultural Communication on the Psychological Health of University Students in the Environment of Big Data. Journal of Environmental and Public Health, 2022.

[22] Sha, W., Guo, Y., Yuan, Q., Tang, S., Zhang, X., Lu, S., ... & Cheng, S. (2020). Artificial intelligence to power the future of materials science and engineering. Advanced Intelligent Systems, 2(4), 1900143.

[23] Arokiaraj, P., Sandeep, D. K., Vishnu, J., & Muthurasu, N. (2023). Movie Recommendation System Using Machine Learning. Advances in Science and Technology, 124, 398-406.

[24] Haffner, M., Hagge, P., Brown, C., Heyrman, R., & Perkins, C. (2022). Fusing machine learning with place-based survey methods: revisiting questions surrounding perceptual regions. International Journal of Geographical Information Science, 36(11), 2226-2247.

[25] Meenaz, A. (2022). Predicting the Price of Cryptocurrency Using Machine Learning Algorithm.