Implementation of Data Mining for Predicting Graduation of Industrial Engineering Students at Telkom University Using Naïve Bayes

1st Alvita Juliana Ambarwati
Faculty of Industrial Engineering
Telkom University
Bandung, Indonesia
alvitajuliana@student.telkomuniversity.ac.id

2nd Afrin Fauzya Rizana
Faculty of Industrial Engineering
Telkom University
Bandung, Indonesia
afrinfauzya@telkomuniversity.ac.id

3rd Rayinda Pramuditya Soesanto
Faculty of Industrial Engineering
Telkom University
Bandung, Indonesia
raysoesanto@telkomuniversity.ac.id

***Abstract*— In higher education, students play a central role, and their academic progress is essential for Telkom University's evaluation. To improve graduation prediction accuracy, the Head of the Study Program is implementing the Naïve Bayes method through a dashboard. This method considers attributes like gender, semester grades (*IPS*1 to *IPS*6), and graduation status, achieving an 83.11% accuracy rate. The designed dashboard not only streamlines monitoring but also allows for intervention and improvement strategies, leading to enhanced learning outcomes and better overall academic management effectiveness.**

***Keyword*— Data Mining, Naïve Bayes Classifier, Student Graduation**

I. INTRODUCTION

Telkom University, located in Bandung, West Java, is a leading private university in Indonesia offering academic excellence through seven faculties and 50 study programs, including the Industrial Engineering Study Program. Established on September 28, 1990, this program has achieved Superior accreditation by the National Accreditation Board for Higher Education of Indonesia No. 2555/SK/BAN-PT/AK-ISK/S/IV/2022.

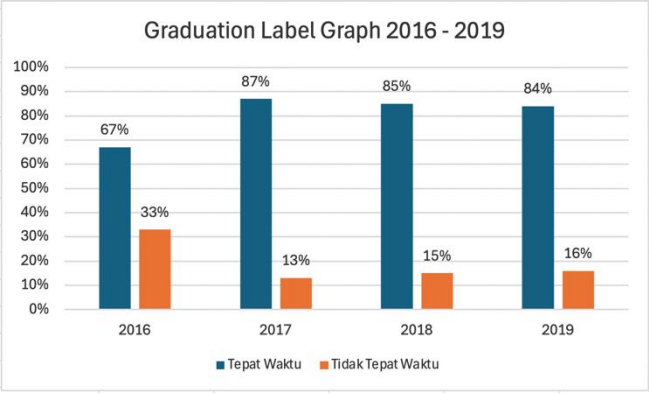

Figure 1.1
Graduation Label Graph

Based on the data that obtained from academic services Telkom University in Figure I.1 show that the graph illustrates a downward trend in on time graduation (TW) rates from 87% in 2017 to 84% in 2019, while late graduation (TTW) rates increased from 13% to 16% during the same period. This indicates that there are still many students who need support to complete their studies on time. The gradual increase in TTW from 2017 to 2019 is a cause for concern because it can obstruct the achievement of on time graduation and have a negative impact on students. Therefore, further research is needed to identify the factors that caused and to develop effective solutions to help students complete their studies on time.

The Industrial Engineering program aims to become a world-class study program actively contributing to the development of information technology-based industrial engineering. Despite its rigorous curriculum and supportive environment, predicting student graduation on time remains challenging due to various influencing factors, such as delays in starting final projects and semester grades.

To address this issue, the Head of the Study Program seeks to develop a dashboard using the Naïve Bayes method to predict student graduation outcomes. This initiative aims to enhance the ability to identify at-risk students early, thereby improving on-time graduation rates and overall academic management.

II. LITERATURE RERVIEW

A. Data Mining

Data mining is the process of automatically discovering valuable information in large data stores. Data mining techniques are used to search large databases with the aim of finding new and valuable patterns that may have previously been unknown. This technique also provides the ability to predict the results of future observations. [1]

B. Naïve Bayes

The Naive Bayes Classifier is a straightforward probabilistic classification method that can calculate various probabilities by considering the frequency and combinations of values from the data in the dataset. [2]

C. Decision Tree

A decision tree is a predictive model that utilizes a tree-like or hierarchical structure to make decisions or predictions based on input data[3]. This structure consists of nodes,

branches, and leaves. Each internal node represents a decision point or test on an attribute, each branch represents the outcome of the test.

D. Support Vector System (SVM)

SVM (Support Vector Machines) serves as an effective method for categorizing research studies, streamlining searches for those in need and achieving a high in research evaluations[4].

E. Neural Network

Neural networks are advanced statistical tools used for non-linear data modeling and pattern recognition, assisting data warehousing firms in extracting valuable insights through data mining processes to facilitate informed decision making [5].

F. Google Colab

Google Colab comes equipped with various Python libraries such as Pandas, Matplotlib, and Plotly that can be used for data processing and creating data visualizations. The advantage of Colab is that users don't need to install software on their local computer to generate the required data visualizations [6].

G. Python

Python is a multipurpose interpreted programming language known for its focus on code readability in its design philosophy and it's praised for blending capabilities and skills, offering a clear syntax and an extensive, comprehensive set of standard library functionalities [7].

H. Information System

An information system refers to an organized and interacting collection of elements, components, or variables, emphasizing the importance of paying attention to each aspect that makes up a system to increase managerial effectiveness[8].

I. Dashboard

Dashboards act as a crucial tool in navigating the overwhelming amount of data and information, offering clear visibility for effective decision-making and management [9].

III. METHOD

The preliminary stage in designing a student graduation prediction system is carried out by carrying out several processes, including identifying the formulation of the problem, creating objectives and boundaries of the problem, then carrying out preliminary studies such as literature studies and also field studies. The following is a description of each step in the preliminary stage of this final project...

A. Preprocessing Step

1. Identify Problem Formulation

In the final project, there are two sources of data. The first source is primary data, gathered through interviews with stakeholders, specifically the head of the study program. The second source is secondary data, acquired from Telkom University's academic services and relevant literature.

2. Objectives

The project aims to design a dashboard for predicting the graduation of students in the Industrial Engineering Study Program at Telkom University and to develop prediction models using data mining techniques, specifically employing the Naïve Bayes method.

3. Boundaries

The data used for training the prediction models consists of industrial engineering student data from Telkom University, covering the years 2016 to 2019. Predictions tend to achieve high accuracy when the educational curriculum remains consistent across the dataset.

4. Literature Studies

Preliminary studies in this final assignment research were carried out in 2 ways, including conducting literature studies by searching for data through journals, interviews with the Head of Program Study and collecting data from Telkom University academic services.

B. Data Collecting Step

C. The data collection stage for the final project begins with identifying data needs and determining data collection strategies based on the data type. Primary data is gathered through interviews with stakeholders, specifically the Head of the Bachelor of Industrial Engineering program. Secondary data is collected from journals, books, past research, and student data from Telkom University's Industrial Engineering program (2016-2019) obtained through the university's academic services. The collected data includes three variables:

1. Gender
2. Semester achievement index 1 to 6
3. Graduation status (on time or not)

After collecting the data, it undergoes cleaning to remove unwanted or empty values, resulting in 1,508 datasets. Identifying user needs and stakeholder requirements helps in developing the predictive model using the Naive Bayes method, addressing challenges, and creating effective solutions.

D. Integrated System Design Step

This stage is carried out to carry out data processing so that we can design a prediction system for student graduation with good accuracy using google colab.

1. Call Packages
2. Import Data
3. Categorize data
4. Determining Naïve Bayes Classification Results
5. Distribution of Training and Testing Data
6. Encoding and Normalization

E. Validation Step of Designing Result

The verification stage is a stage to ensure whether the design results that have been designed are appropriate or not.

F. Closing Step

Conclusions and suggestions are the final stage in this final assignment which contains a design for a student graduation prediction system

IV. RESULT AND DSICUSSION

At this stage, data and information are collected to be used in the data processing phase for creating a student graduation prediction plan. This research utilizes various data sources, including:

A. Data Collection

1. Primary Data

Primary data was collected through direct interviews with the Head of the Industrial Engineering Study Program, who explained the current situation and the initial approach to making predictions. The need for this research arose when the head recognized the necessity of a platform to predict student graduation efficiently and accurately, as opposed to the previous manual methods or simply reviewing existing data. With this platform, the Head of Industrial Engineering can monitor students more effectively by utilizing the generated prediction results.

2. Secondary Data

Data obtained through Telkom University academic services and study literature. For the secondary data it contains gender, IPS 1 to 6 and information label graduation time it's on time or not for students from the 2016 to 2019 class.

TABLE III. 1
Example Student Data

Gender	IPS 1	IPS 2	IPS 3	IPS 4	IPS 5	IPS 6	Label
Male	2.68	2.47	2.66	2.18	3.29	3.35	Late
Female	3.58	3.69	3.53	3.58	3.55	3.53	On Time
Female	3.82	3.78	3.05	3.87	3.65	3.76	On Time
Male	1.87	2.53	1.8	2.58	2.55	2.18	Late
Female	3.39	3.75	3.45	3.5	3.62	2.95	On Time
Male	3.74	3.72	3.74	3.61	3.39	3.3	On Time
Male	3.62	2.61	2.82	2.55	3.29	3.23	Late

B. Design Process

1. Call Packages

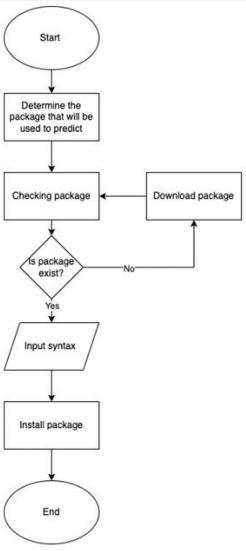

FIGURE III. 1
Call Packages

To use packages in Google Colab, first import the files module with `from google.colab import files`. This module allows interaction with the file system in a Colab session. Next, use `uploaded = files.upload()` to call the `upload()` function, which prompts the user to select a file from their local computer. The selected file is then uploaded to the Colab session and stored in the `uploaded` variable.

2. Import Data

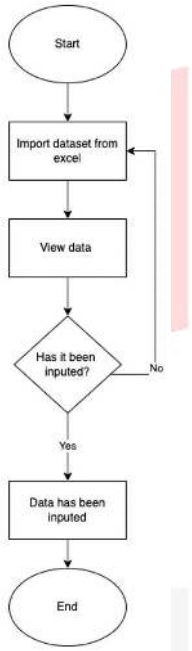

FIGURE III. 2
Import Data

First, the syntax `import pandas as pd` is used to import the pandas library with the alias `pd`, allowing for concise usage of its functions. Pandas is widely used in Python for data analysis and manipulation. Next, the syntax `df = pd.read_excel('DATA FIX.xlsx')` reads an Excel file named `DATA FIX.xlsx` into a DataFrame called `df`. The `pd.read_excel()` function reads the file, which must be in the current working directory or have its full path provided. The resulting DataFrame, `df`, stores the data in a table format similar to a spreadsheet.

3. Categorize data

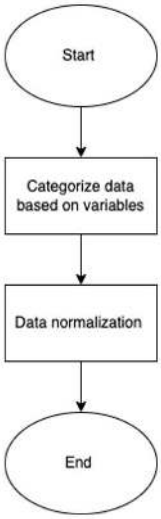

FIGURE III. 3
Categorize Data Flow

In categorizing data, there are 2 types, namely predictor attributes and outcome attributes. The predictor attribute consists of X = df[['Gender', 'IPS 1', 'IPS 2', 'IPS 3', 'IPS 4', 'IPS 5', 'IPS 6']] which is designated as a variable for prediction.

4. Determining Naïve Bayes Classification Results

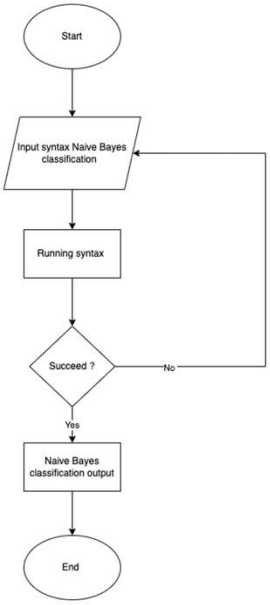

FIGURE III. 4
Determining Naïve Bayes Classification Flow

This code splits the dataset into two parts: training data (80%) and testing data (20%) using the `train_test_split` function from the `scikit-learn` library. The first line imports the necessary function, while the second line performs the dataset split for `X` (features) and `y` (labels). The parameter `test_size=0.2` specifies that 20% of the data will be used for testing, and `random_state=42` ensures that this split is consistent every time the code is run, providing the same results each time.

```
Akurasi Naive Bayes: 0.8311258278145696
              precision    recall  f1-score   support

         TTW       0.59      0.60      0.59        62
          TW       0.90      0.89      0.89       240

    accuracy                           0.83       302
   macro avg       0.74      0.74      0.74       302
weighted avg       0.83      0.83      0.83       302
```

FIGURE III. 5
Accuration Result

The Naive Bayes model achieved an 83.11% accuracy, showing mostly correct predictions. Specifically, for the "On Time" class (TW), the model demonstrated a 90% precision and 89% accuracy with an F1-score of 89%, indicating its strong ability to identify timely graduates. However, its performance was lower for the "Not on Time" class (TTW) with 59% accuracy and 60% precision, resulting in an F1-score of 59%. Overall, the model showed an average precision, accuracy, and F1-score of 74% (macro avg) and 83% (weighted avg), indicating good overall performance but with room for improvement in predicting the TTW class.

C. User Interface

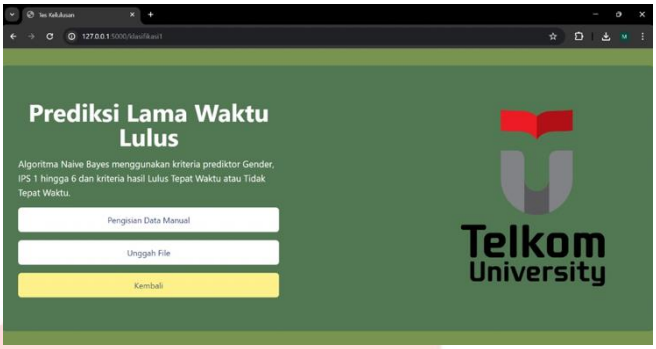

FIGURE III. 6
User Interface 1

shows menu of this predicting. There are *Pengisian Data Manual* that in this menu will be predict as one personal. The other one is *Unggah* File menu that user can predict student as many data that needs.

FIGURE III. 7
User Interface 2

Shows testing data once user choose for *Pengisian Data Manual menu.* In here user needs to fill student data such as nama, wali dosen, gender, kelas, IPS1, IPS2, IPS3, IPS4, IPS5, IPS6 and after fill student data user can choose to predict menu.

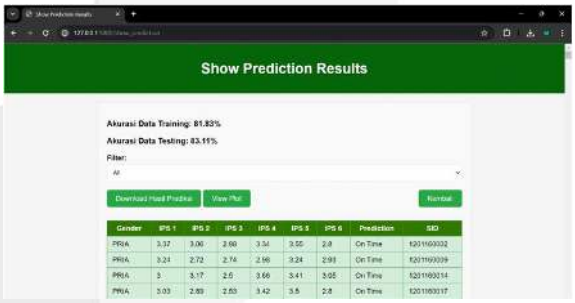

FIGURE III. 8
User Interface 3

Shows *Lihat Hasil Prediksi* interface that user can see which student that on time marked in green colour but for late student marked in red colour. If user has done to predict user can be downloaded the prediction and choose *Kembali* button to finish the prediction

V. CONCLUSION

The information system designed to predict student graduation at Telkom University's Industrial Engineering program yielded the following conclusions:

1. To predict the graduation of Telkom University Industrial Engineering Study Program students on time using the factors obtained, the dashboard uses a Naïve Bayes methodology that considers factors such as gender, IPS 1 to 6 and graduation label. Using Naïve Bayes method this model gains an accuracy of 83%. Applying predictive modeling helps identify potential areas for intervention or improvement.
2. The designed dashboard is created to predict student graduation and improve the monitoring and evaluation process for the Head of the Study Program addressing specific needs. The dashboard designed can predict student graduation as individually and also large data. This simplification enhances efficiency within the School of Industrial Engineering and promotes better learning outcomes in the future.

Overall, the dashboard's design using the Naive Bayes method significantly aids in predicting student graduation and refining monitoring processes. Leveraging data mining techniques and specific factors improves efficiency and identifies areas for enhancement, contributing to better learning outcomes and academic management.

REFERENCE

1 Dwi Cahya, P., & Durbin Hutagalung, D. (2023). Penerapan Data Mining Menggunakan Algoritma Apriori Pada Penjualan Produk Sembako Berbasis Web (Studi Kasus: Warung Abah Murdika). *Jurnal Ilmu Komputer Dan Pendidikan*, *1*(6), 1465–1469.

2 Dwiramadhan, F., Wahyuddin, M. I., & Hidayatullah, D. (2022). Sistem Pakar Diagnosa Penyakit Kulit Kucing Menggunakan Metode Naive Bayes Berbasis Web. *Jurnal JTIK (Jurnal Teknologi Informasi Dan Komunikasi)*, *6*(3), 429–437.

3 Nasrullah, A. H. (2021). Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik. Jurnal Pilar Nusa Mandiri, 7(2), 217.

4 Jumeilah, F. S. (2017). Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *1*(1), 19–25.

5 Singh, Y., & Chauhan, A. S. (2009). Neural Networks in Data Mining. *Journal of Theoretical and Applied Information Technology*, *5*(1), 37–42.

6 Guntara, R. G. (2023). Visualisasi Data Laporan Penjualan Toko Online Melalui Pendekatan Data Science Menggunakan Google Colab. *Jurnal Ilmiah Multidisiplin*, *2*(6), 2091–2100.

7 Syahrudin, A. N., & Kurniawan, T. (2018). Input dan Output pada Bahasa Pemrograman Python. *Jurnal Dasar Pemrograman Python STMIK*, *June 2018*, 1–7.

8 Sutabri, T. (2012). *Analisis Sistem Informasi*.

9 Malik, S. (2005). *Enterprise Dashboard*.