Implementation of Data Mining for Prediction of Student Waiting Time for Students Industrial Engineering Undergraduate Study Program at Telkom University Using Naive Bayes

1st Ishfahan Dzilalin Nuha
School of Industrial Engineering
Telkom University
Bandung, Indonesia
ishfahannuha@student.telkomuniversity.ac.id

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

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

Abstrak—**Telkom University, a leading institution in Indonesia, aims to produce competitive graduates. The Industrial Engineering program at the university is focused on maintaining a low average waiting time for graduates to preserve accreditation. In 2023, the Faculty of Industrial Engineering (FRI) saw an increase in the average waiting time to 3.89 months, compared to 3.72 months in 2022. For the Industrial Engineering program, the waiting time rose to 4.16 months. Factors contributing to this increase include poor English proficiency, extended study duration, and involvement in non-academic activities. To address this issue, the final project aims to develop a prediction model using the Naïve Bayes Algorithm to forecast student waiting times. The model employs data mining techniques, utilizing attributes such as gender, study duration, GPA, English Proficiency (EPrT) Score, and Student Activity Transcript (TAK) points. Data from the tracer study of 2016-2018 alumni were used, split into 80% for training and 20% for testing. The model achieved an accuracy of 65.95%, with precision and recall rates of 65% and 97%, respectively. A predictive dashboard was developed, allowing manual input and Excel data uploads. This tool helps the Head of the Industrial Engineering Program monitor and predict waiting times, aiding in decision-making and strategy development for improved academic management.**

Keyword— **Dashboard, Data Mining, Naïve Bayes Classifier, Prediction, Waiting Time**

I. INTRODUCTION

Telkom University is one of the higher education institutions in Indonesia that provides quality educational services to its students to produce competent, integrity-driven graduates who are competitive at both national and international levels. One of the offered study programs is the bachelor's program in Industrial Engineering. The Bachelor's (S1) Industrial Engineering Study Program at Telkom University specializes in strengthening Industrial Engineering knowledge in the field of Information and Communication Technology (ICT).

The Industrial Engineering study program aims to maintain its students and alumni to have first job waiting time no more than 6 months. This is the case because in the 2023 tracer study book, the industrial engineering study program experienced a longer average waiting time of 4.16 months [1]. Even though the number is below 6 months, The industrial Engineering is the study program with the longest waiting time among the other two study programs under the School of Industrial Engineering (FRI). [1]

To tackle this issue, the Head of the Industrial Engineering Study Program plans has a desire to create a dashboard using the Naïve Bayes method to predict student waiting time outcomes. This initiative aims to enhance the early identification of at-risk students, thereby improving waiting time of Students' rates and overall academic management.

II. LITERATURE REVIEW

A. Data Mining

Data mining is the process of discovering interesting patterns, models, and knowledge from large amounts of data [2]. It is a discipline with the primary goal of discovering, unearthing, or mining knowledge from the data or information we possess. Often referred to as Knowledge Discovery in Databases (KDD), data mining is the process of collecting, utilizing, and analyzing historical data to uncover relationships, patterns, or regularities in very large datasets. [3]

B. Naïve Bayes

Naive Naive Bayes Classifier is a simple probabilistic classification technique that calculates various probabilities based on the frequency and combinations of values in the dataset [4]. Naive Bayesian classifiers enable the modeling of relationships among specific attribute groups, providing a detailed representation of their interdependencies. [2] This algorithm is capable of efficiently processing data in various formats. It works by calculating the probability of each new data point for every existing class [5]. The Naive Bayes theorem can be represented by the following equation*:*

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

C. Decision Tree

A decision tree is a predictive model that employs a tree-like or hierarchical structure, which helps in simplifying complex decision-making processes [6]. Additionally, decision trees classify data samples with unknown classes into one of the existing classes using various algorithms designed for classification [7]. Consequently, decision trees provide an efficient and precise solution for data analysis and management

D. Support Vector Machine

SVM (Support Vector Machines) is an effective method for categorizing research studies, making searches more efficient for users and achieving high accuracy in research evaluations [9]. Support Vector Machine (SVM) is a learning system that uses a hypothesis space of linear functions in a high-dimensional feature space, trained with learning algorithms based on optimization theory [8].

E. Google Colab

Google Colab is a Python programming application available online, which has access to a Graphics Processing Unit (GPU) for use in research [9]. Google Colab has provided most of the libraries required for various purposes. With a wide selection of libraries available, users can easily access and utilize a variety of tools and resources to expand analysis and development capabilities. 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. [10]

F. Dashboard

The term dashboard refers to a system designed to present data in a visual format that aids decision-making. Dashboards aim to provide informative insights without diverting users from their primary tasks. They summarize data using charts, tables, gauges, and other visual elements. To ensure users interpret dashboard elements accurately, these systems usually provide access to the underlying raw data that was used for the summarization [13]

G. Entity Relationship Diagram (ERD)

ERD is one of the primary diagrams representing a conceptual data model that reflects user data requirements in a database system. ERD is the initial step in database design. There are several important aspects to consider when creating an ERD for database design. Every database must have entities that are related to each other, and each entity must have attributes that include a primary key as an entity characteristic and descriptive attributes. [11]

H. Unified Modelling Language

Unified Modeling Language or UML is a versatile visual tool widely used in software engineering. It provides a standard method for depicting, specifying, building, and documenting components of software systems.. Its diverse functions support system comprehension, design exploration, configuration control, maintenance, and information management. This makes UML invaluable for handling and examining complex software projects.[12] Types of UMLs are sequence diagram, activity diagram, and usse case diagram.

I. User Interface

The user interface is a computer-facilitated method that enables communication between individuals or between a person and a device. This includes interactive activities and the integration of physical objects with computer systems, encompassing hardware and software components such as applications, operating systems, and networks. [13]

III. METODE

The initial phase in designing a student graduation prediction system involves several key processes, including identifying the problem, setting objectives and boundaries, and conducting preliminary studies such as literature reviews and field research. Below is a description of each step in this preliminary stage of the final project:

A. Preprocessing Step

1. Identify Problem Formulation

In this final project, two sources of data are utilized. The first is primary data, collected through interviews with stakeholders, particularly the head of the study program. The second source is secondary data, obtained from Telkom University's academic services and relevant literature.

2. Objectives

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

3. Boundaries

The data used for training the prediction models comprises information on industrial engineering students from Telkom University, spanning the years 2016 to 2018.

4. Literature Studies

Preliminary studies for this final project were conducted in two ways: by performing literature reviews through journal research and by collecting data via interviews with the Head of the Study Program and from Telkom University's academic services, Directorate CAE, FRI and LaC.

B. Data Collection Step

The data collection stage for the final project starts by identifying data requirements and determining collection strategies based on data types. Primary data is obtained through interviews with stakeholders, particularly the Head of the Bachelor of Industrial Engineering program. Secondary data is sourced from journals, books, past research, and student data from Telkom University's Industrial Engineering program (2016-2018) acquired through the university's academic services, Directorate CAE, FRI and LaC. The collected data encompasses six variables:

1. Gender
2. Study Duration
3. GPA
4. Score EprT
5. TAK

IV. RESULT AND DISCUSSION

At this stage, data and information are gathered to be utilized in the data processing phase for developing a student graduation prediction plan. This research incorporates various data sources, including:

A. Data Collection

Gambar dinomori secara berurutan. Letak penulisannya di bawah gambar yang dijelaskan. Contoh: Gambar 1(A)

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 emerged when the head recognized the necessity of an efficient and accurate platform for predicting student graduation, rather than relying on previous manual methods or simply reviewing existing data. This platform will enable the Head of Industrial Engineering to monitor students more effectively by utilizing the generated prediction results.

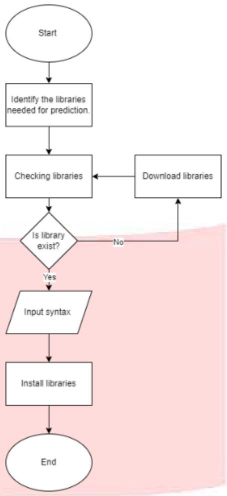
2. Secondary Data

Data was obtained through Telkom University's academic services, LaC, CAE, and FRI and literature studies. The secondary data includes gender, study duration, GPA, Score EPrT, TAK and Waiting Time labels indicating whether students from the 2016 to 2018 classes having waiting time less than 6 months or not.

Gender	Study Duration	GPA	EPrT Score	TAK	Waiting Time (Label)
PRIA	TTW	3,38	490	96	> 6 Bulan
PRIA	TW	3,24	503	91	> 6 Bulan
PRIA	TW	3,23	453	85	<= 6 Bulan
WANITA	TTW	3,8	623	92	> 6 Bulan
PRIA	TW	3,65	450	88	<= 6 Bulan
WANITA	TW	3,81	490	67	<= 6 Bulan
WANITA	TW	3,93	510	81	<= 6 Bulan

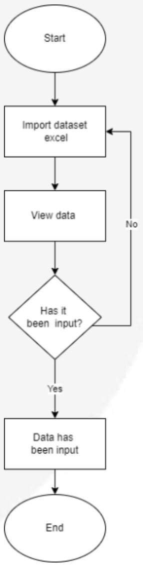
B. Design Process

1. Import Libraries



When using libraries, start by importing the 'files' module from the 'google.colab' library using the syntax ‘from google.colab import files’. This module allows you to interact with the file system in a Google Colab session, such as uploading and downloading files. Then, with the syntax ‘upload = files.upload()’, call the ‘upload()’ function from the ‘files’ module to upload files from your local computer to the Google Colab session. When the ‘upload()’ function is executed, a prompt will appear asking you to select files from your local system. After selecting the files, they will be uploaded to the Google Colab session and stored in the variable ‘uploaded’

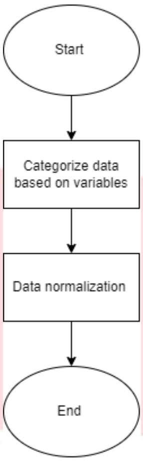
2. Import Data



First, the syntax import pandas as pd is used to import the pandas library and assign it the alias pd. This alias allows us to use functions from the pandas library more concisely Next, the syntax df = pd.read_excel('Dataset_WaitingTime.xlsx') is used to read an Excel file named Dataset_WaitingTime.xlsx and store its contents in a DataFrame called df. The pd.read_excel() function, part of the pandas library, reads Excel files. The file name 'Dataset_WaitingTime.xlsx' must be correct and located in the current working directory, or you must provide the full path to the file. The variable df will hold the DataFrame generated from reading the Excel file. A

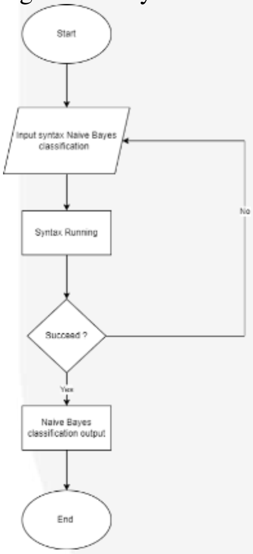
DataFrame is a two-dimensional data structure in pandas used to store data in a table format, similar to a spreadsheet.

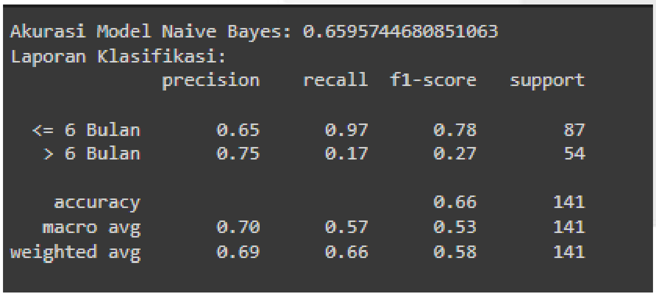
3. Data Categorization



In data categorization, there are two categories of data: predictor attributes and label/class attributes. The predictor attributes consist of X = df[['Gender', 'Lama Masa Studi', 'IPK', 'Score EPRT', 'TAK']], which will be used as the prediction variables. The label/class attribute is y = df['Masa Tunggu'].

4. Determining Naïve Bayes Classification Results


Start
Input syntax Naive Bayes classification
Syntax Running
Succeed ?
No
Yes
Naive Bayes classification output
End



```
Akurasi Model Naive Bayes: 0.6595744680851063
Laporan Klasifikasi:
              precision    recall  f1-score   support

  <= 6 Bulan       0.65      0.97      0.78        87
   > 6 Bulan       0.75      0.17      0.27        54

    accuracy                           0.66       141
   macro avg       0.70      0.57      0.53       141
weighted avg       0.69      0.66      0.58       141
```

FIGURE II.1
Accuracy

This Naive Bayes model achieved an accuracy of 65.95%, indicating that most of its predictions were correct. For the “CEPAT” class, the model had a precision of 65% which means the model predicts a student as ‘CEPAT’ is correct 65% of the time, a recall of 97% means the model correctly identifies 97% of all actual ‘CEPAT’ students, with an F1-score of 78% indicating the model’s balanced performance in predicting ‘CEPAT’ class. For the “LAMBAT” class, it had a precision of 75% meaning the model predicts a student as ‘LAMBAT’ is correct 75% of the time, a recall of 17% indicating the model correctly identifies only 17% of all actual ‘LAMBAT’ students and with an F1-score of 27% indicating poor performance in predicting this class. In terms of effectiveness, the model shows decent performance in predicting the ‘CEPAT’ class with high recall and moderate precision. However, the performance for the ‘LAMBAT’ class is poor, with low recall and a significantly lower F1-score.

C. User Interface

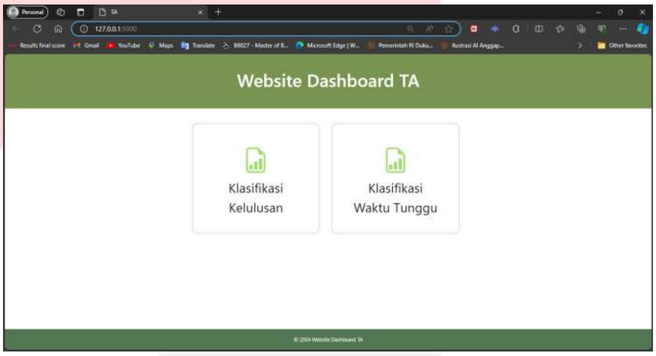


FIGURE II.2
User Interface 1

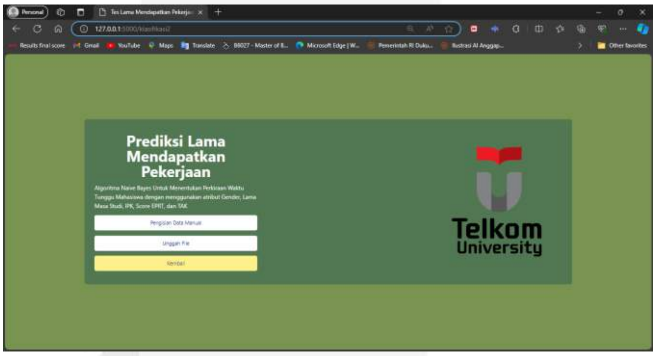

Prediksi Lama Mendapatkan Pekerjaan
Algoritma Naive Bayes Untuk Menentukan Perkiraan Waktu Tunggu Mahasiswa dengan menggunakan atribut Gender, Lama Masa Studi, IPK, Score EPRT, dan TAK
Pengisian Data Manual
Unggah File
Kembali
Telkom University

FIGURE II.3
User Interface 2

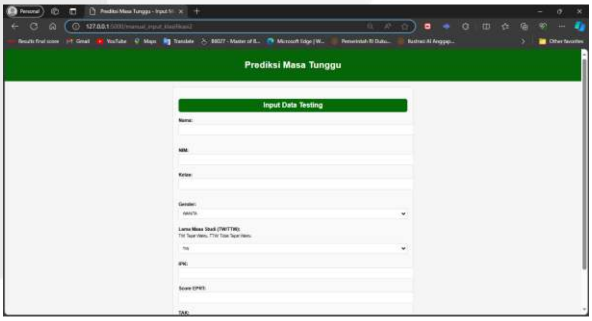
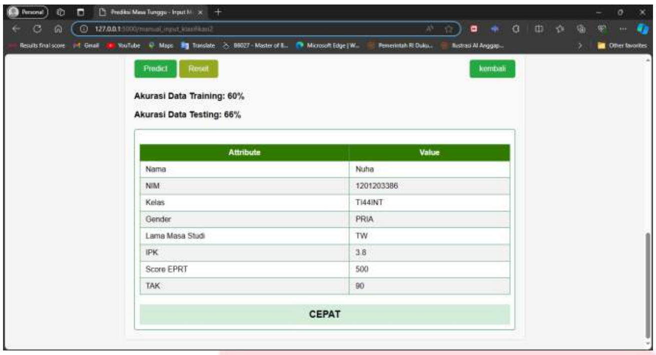


FIGURE II.4
User Interfcae 3



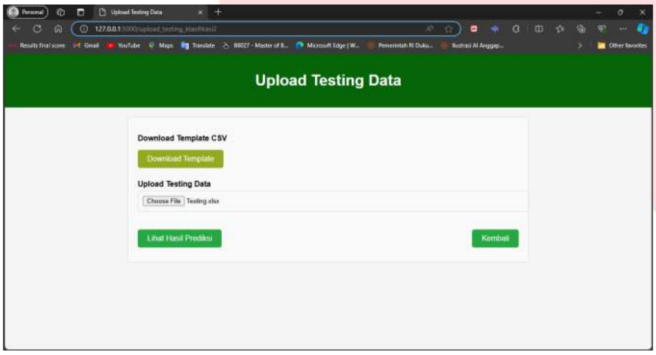
Akurasi Data Training: 60%

Akurasi Data Testing: 66%

Attribute	Value
Nama	Nuha
NIM	1201203386
Kelas	TI44INT
Gender	PRIA
Lama Masa Studi	TW
IPK	3.8
Score EPRT	500
TAK	90

CEPAT

FIGURE II.5
User Interface 4



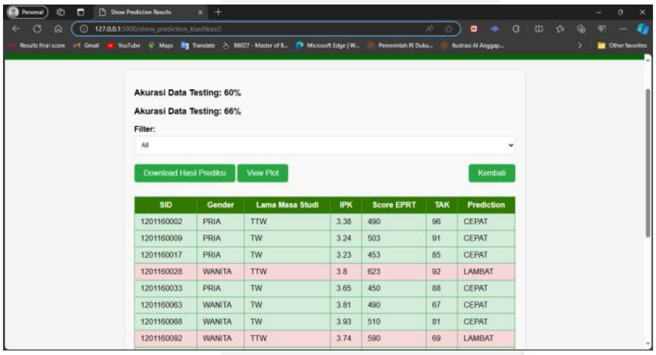
Upload Testing Data

Download Template CSV

Upload Testing Data

Testing.xlsx

FIGURE II.6
User Interface 5



Akurasi Data Testing: 60%

Akurasi Data Testing: 66%

Filter: All

SID	Gender	Lama Masa Studi	IPK	Score EPRT	TAK	Prediction
1201160002	PRIA	TTW	3.38	490	96	CEPAT
1201160009	PRIA	TW	3.24	503	91	CEPAT
1201160017	PRIA	TW	3.23	453	85	CEPAT
1201160028	WANITA	TTW	3.8	623	92	LAMBAT
1201160033	PRIA	TW	3.65	450	88	CEPAT
1201160063	WANITA	TW	3.81	490	67	CEPAT
1201160068	WANITA	TW	3.93	510	81	CEPAT
1201160092	WANITA	TTW	3.74	590	69	LAMBAT

FIGURE II.7
User interface 6

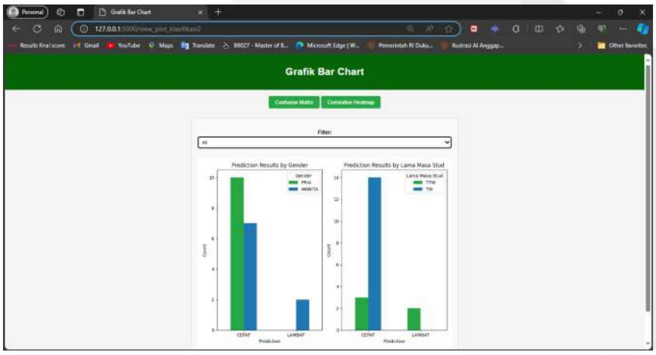


FIGURE II.8
User Interface 7

V. CONCLUSION

From the design results of the information system outcomes that predict student graduation at Telkom University Industrial Engineering program, the following conclusions can be drawn:

1. The designed dashboard is developed to predict student waiting time and enhance the monitoring and evaluation process for the Head of the Industrial Engineering Study Program, addressing specific needs. This simplification boosts efficiency within the Study Program of Industrial Engineering and fosters improved learning outcomes in the future.
2. To predict the waiting time of students in Telkom University's Industrial Engineering Study Program, the dashboard employs a data mining methodology that takes into account factors such as gender, study durartion, GPA, EPrT Score, TAK and waiting time class. By applying predictive modeling and algorithmic analysis, it provides more accurate projections of waiting time and helps identify potential areas for improvement.

REFERENCE

[1] A. dan E. Direktorat Pengembangan Karir, *Buku Tracer Study Direktorat Pengembangan Karir, Alumni, dan Endowment 2023*. 2023.

[2] J. Han, J. Pei, and H. Tong, “Data Mining: Concepts and Techniques,” 2023.

[3] M. Ridwan, H. Suyono, and M. Sarosa, “Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier,” *jurnal EECCIS*, vol. 7, no. 1, pp. 59–64, 2013.

[4] F. Dwiramadhan, M. I. Wahyuddin, and D. Hidayatullah, “Sistem Pakar Diagnosa Penyakit Kulit Kucing Menggunakan Metode Naive Bayes Berbasis Web,” *Jurnal Teknologi Informasi dan Komunikasi)*, vol. 6, no. 3, p. 2022, 2022, doi: 10.35870/jti.

[5] J. Sulaksono and Darsono, “SISTEM PAKAR PENENTUAN PENYAKIT GAGAL JANTUNG MENGGUNAKAN METODE NAIVE BAYES CLASSIFIER,” pp. 6–8, 2015.

[6] A. H. Nasrullah, “Implementasi algoritma Decision Tree untuk klasifikasi produk laris,” *Jurnal Ilmiah Ilmu Komputer Fakultas Ilmu Komputer Universitas Al Asyariah Mandar*, vol. 7, no. 2, pp. 45–51, 2021, [Online]. Available: http://ejournal.fikom-unasman.ac.id

[7] L. Qadrini, A. Seppewali, and A. Aina, “Decision Tree dan Adaboost pada Klasifikasi Penerima Program Bantuan Sosial,” *Jurnal inovasi penelitian*, vol. 2, no. 7, pp. 1959–1966, 2021, doi: 10.47492/jip.v2i7.1046.

[8] D. Irawan, E. B. Perkasa, Y. Yurindra, D. Wahyuningsih, and E. Helmud, “Perbandingan Klassifikasi SMS Berbasis Support Vector Machine, Naive Bayes Classifier, Random Forest dan Bagging Classifier,” *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 10, no. 3, pp. 432–437, Dec. 2021, doi: 10.32736/sisfokom.v10i3.1302.

[9] R. T. Handayanto and Herlawati, “Prediksi Kelas Jamak dengan Deep Learning Berbasis Graphics Processing Units,” *Jurnal Kajian Ilmiah*, vol. 20, no. 1, pp. 67–76, 2020, doi: 10.31599/jki.v20i1.71.

[10] R. G. Guntara, “Visualisasi Data Laporan Penjualan Toko Online Melalui Pendekatan Data Science Menggunakan Google Colab,” *ULIL ALBAB: Jurnal Ilmiah Multidisiplin*, vol. 2, no. 6, pp. 2091–2100, 2023, doi: 10.56799/jim.v2i6.1578.

[11] S. M. Pulungan, R. Febrianti, T. Lestari, N. Gurning, and N. Fitriana, “Analisis Teknik Entity-Relationship Diagram Dalam Perancangan Database,” *Jurnal Ekonomi Manajemen dan Bisnis (JEMB)*, vol. 1, no. 2, pp. 143–147, 2023, doi: 10.47233/jemb.v2i1.533.

[12] James. Rumbaugh, Ivar. Jacobson, and Grady. Booch, *The unified modeling language reference manual*. Addison-Wesley, 1999.

[13] A. Marcus, “Dare We Define User-Interface Design?,” *Interactions*, vol. 9, no. 5, pp. 19–24, 2002. doi: 10.1145/566981.566992.

.