**International Journal of Social Learning**
August 2024, Vol. 4 (3), 291-310
e-ISSN 2774-4426 and p-ISSN 2774-8359
DOI: https://doi.org/10.47134/ijsl.v4i3.277

# K-Nearest Neighbors (K-NN) Algorithm Model in Predicting the Graduation Rate of Teacher Professional Education Students in Indonesia

Musthofa[1*], Dwi Yunita Sari[2], Nasikhin[3], Juanduo Wang[4]

[1,2,3]*Universitas Islam Negeri Walisongo Semarang, Indonesia*
[4]*University of Science and Technology of China, China*

*e-mail: thofa@walisongo.ac.id

## ABSTRACT

*Predicting the graduation rate of the PPG program has an important significance in analyzing the factors that affect students' success in completing the PPG program. This study uses the K-Nearest Neighbor model in online learning to predict the pass rate of students in the Teacher Professional Education Program (PPG) at UIN Walisongo Semarang. The study analyzed data from 423 students, focusing on input quality variables, such as pedagogical competence and teaching innovation. Results showed the Wave 1 pass rate in 2023 was 86.7%, with 13.3% failure, a 1.7% decrease from Wave 3 in 2022. The confusion matrix showed significant improvement in True Positives (TP) and True Negatives (TN), with an accuracy of 0.916, precision of 0.3, and recall of 0.9725 students' academic achievement.*

***Keywords*:**

Input Quality; K-Nearest Neighbor; Teacher Professional Education; Online Learning.

## ABSTRAK

*Memprediksi tingkat kelulusan program PPG memliki signifikansi penting untuk menganalisa faktor-faktor yang mempengaruhi keberhasilan mahasiswa dalam menyelesaikan program PPG. Studi ini bertujuan untuk memprediksi tingkat kelulusan mahasiswa Program Pendidikan Profesi Guru (PPG) di UIN Walisongo Semarang menggunakan model K-Nearest Neighbor dalam pembelajaran daring. Studi menganalisis data 423 mahasiswa dengan fokus pada variabel kualitas input, seperti kompetensi pedagogik dan inovasi mengajar. Hasil menunjukkan tingkat kelulusan Gelombang 1 tahun 2023 sebesar 86,7%, dengan kegagalan 13,3%, turun 1,7% dari Gelombang 3 tahun 2022. Confusion matrix menunjukkan peningkatan*



**Submitted: 2023-09-04; Accepted: 2024-08-10; Published: 2024-08-31**
***Corresponding author:** thofa@walisongo.ac.id

*signifikan pada True Positives (TP) dan True Negatives (TN), dengan akurasi 0,916, presisi 0,3, dan recall 0,9725.asilan akademik mahasiswa*.

***Kata kunci*:**

Kualitas Input; K-Nearest Neighbor; Pendidikan Profesi Guru; Pembelajaran Online.

## 1. Introduction

The quality of student input in implementing the Teacher Professional Education program is a key factor in determining graduation. However, some challenges must be overcome in this regard. First, students have various educational backgrounds, levels of technological ability, and motivations for online learning (Pangestika & Alfarisa, 2015). This can affect how they learn and whether they graduate. Second, assessing the quality of student input in online learning is often difficult due to the lack of an objective assessment method (Zulfitri et al., 2019). This can lead to unfairness in assessments and graduation decisions because the host university only receives data from the government. Third, the rapid development of technology and online learning methods can be problematic because students and lecturers may have difficulty adapting (Arifa & Prayitno, 2019). This can affect the quality of input and learning outcomes (Zulfitri et al., 2019). In addition, online accessibility and infrastructure can also be obstacles, especially in underdeveloped areas. This is important because it can make it difficult for students to access online learning properly (Muslim, 2010). Therefore, this study must explore ways to overcome these problems to develop a fair and accurate way to assess whether students pass based on the quality of their input in online learning (Ratnasari, 2019).

So far, the study of online learning in professional teacher education programs has discussed three issues. First, the research focuses mainly on the effectiveness of online teaching methods in developing the pedagogical skills of prospective teachers as efficiently as possible (Sukardi et al., 2019). Second, there has been more in-depth research on the challenges and obstacles faced by students in participating in online professional teacher education programs, such as the lack of physical interaction with instructors and fellow students (Palloff & Pratt, 2007). Third, the study also includes comparing the learning outcomes of students who took professional teacher education programs online with those who took traditional programs on campus (Nguyen, 2015). This study has provided many benefits in understanding the dynamics of online learning in professional teacher education programs. However, it should be noted that specific research using K-Nearest Neighbor is still rare. Identifying factors that can affect student success is important (Adnan & Anwar, 2020). With a deeper understanding of these factors, we can develop more effective strategies and approaches to supporting students' success in facing the challenges of online learning (Abel, 2005).

This study aims to predict the graduation rate of professional teacher education program students in online learning based on a K-Nearest Neighbor analysis. This is important as it provides solutions to a series of problems within professional teacher education programs based on online

learning. This model can help measure the quality of student input more objectively. Using relevant data, such as educational background, technological ability, and motivation for online learning, the K-Nearest Neighbor analysis can predict students' chances of graduating (Keller et al., 1985). This will help universities identify students who need additional support in the online learning process (Jiang et al., 2007). In addition, by monitoring the development of technology and online learning methods, this model can provide recommendations for improvement based on the needs of students and lecturers. Finally, by considering online accessibility and infrastructure, the K-Nearest Neighbor analysis can help universities design more effective solutions to overcome these barriers so that students can better access online learning (Jiang et al., 2007). Thus, the K-Nearest Neighbor analysis model can be a valuable tool in improving the quality of student input and learning outcomes in the Teacher Professional Education program.

## 2. Methods

### *2.1. Research Design*

This quantitative study aims to predict the graduation rate of students from the Teacher Professional Education Program (PPG) at UIN Walisongo using the K-Nearest Neighbor (K-NN) model. The K-NN method was chosen because it effectively handles distance-based classification problems, which are relevant for predicting graduation outcomes based on student characteristics (Kang, 2021). K-NN also provides accurate predictions when applied to datasets with clear labels, such as the Teacher Professional Education Student Competency Test (UKMPPG) graduation data. The relevance of K-NN in this context is due to its non-parametric nature and its flexibility in handling various types of data without requiring specific distribution assumptions (Dann et al., 2022). Student graduation data will be analyzed using R software, which supports comprehensive modeling and visualization of analysis results, thereby providing deeper insight into the factors that influence graduation.

### *2.2 Research Procedures*

This research consists of data management, exploration, and analysis. The first is data management. These steps are done through transformation, selecting variables, improving the data structure, and cleaning unnecessary data using Microsoft Excel software. The second stage of this research is data exploration, which aims to obtain as much information as possible from the data and is useful for determining the right strategy for processing the data (Faisal & Nugrahadi, 2017). The third step is data analysis, which consists of descriptive statistical analysis and K-NN classification. Descriptive statistical analysis is used to evaluate the UKMPPG graduation rate based on factors such as education level, employment status, province, wave, and district where PPG students work. The data types used in this analysis are categorical and numerical (Lu, 2021). Categorical data includes education level, full name, employee status, province of origin, district of origin, and graduation status, while numerical data includes various variables such as age, GPA, and competency scores (Boateng et al., 2020). The K-NN classification procedure consists of preparing classification data,

sharing classification data, and implementing the K-NN classification algorithm. Data sharing is carried out using the k-fold cross-validation method. The data-sharing results will be saved into four files for the training process and 4 for the testing process.

### *2.3 Population*

This research data was collected through documentation techniques using secondary data on the official website https://ppg.siagapendis.com/. The data collected includes important information such as PPG students' RPL (Recognition of Past Learning) scores, PPG students' data, and UKMPPG (Teacher Professional Education Student Competency Test) graduation data. With its large dimensions and complex structure, this data provides a comprehensive picture of the profile of PPG students and their performance during the program. This approach allows researchers to obtain rich and varied information (Cunningham & Delany, 2021). This is then processed for further analysis, especially in predicting student graduation rates in professional teacher education programs at Islamic universities.

### *2.4 Data Collection*

This research data was collected through documentation techniques using secondary data on the official website https://ppg.siagapendis.com/. The data collected includes important information such as PPG students' RPL (Recognition of Past Learning) scores, PPG students' data, and UKMPPG (Teacher Professional Education Student Competency Test) graduation data. With its large dimensions and complex structure, this data provides a comprehensive picture of the profile of PPG students and their performance during the program. This approach allows researchers to obtain rich and varied information (Cunningham & Delany, 2021). This is then processed for further analysis, especially in predicting student graduation rates in professional teacher education programs at Islamic universities.

### *2.5 Data Analysis*

This research analyzed data by applying the K-Nearest Neighbor (K-NN) algorithm to a dataset collected over two academic years. The K-NN algorithm was chosen because of its ability to classify data based on similarities with other data whose results are already known. Using sophisticated statistical software allows for more in-depth and accurate analysis, thereby identifying important patterns in the data (Isnai et al., 2021). Through this analysis, it is hoped that the research can provide valuable insight into the main factors that influence the graduation rate of students from the professional teacher education (PPG) program at UIN Walisongo. These insights will greatly benefit institutions in designing and implementing more effective educational strategies, thereby increasing overall graduation rates and ensuring that students receive the support they need to succeed (Bakia et al., 2012).

## 3. Results and Discussion

### *3.1 Data collection*

The first stage of this research is data collection. Data collection is carried out by downloading data from the website: https://ppg.siagapendis.com/. The data collected consisted of PPG student RPL score data in Figure 1 (a), PPG Student Personal Data in Figure 1 (b), and UKMPPG Graduation Data in Figure 1 (c). The data collected is secondary data. This data has large dimensions and has a complex structure.

**Table 1.** (a). RPL Assessment Data, (b) Student Personal Data, and (c). Student Graduation Data

| No | Respondent Code | A | B | C | D | And | F | G | H | I | Information |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | R-1 | | | | | | | | | | A pedagogical competency development pedagogic B preparation of learning tools C development of professional competencies D= management of learning administration E= Learning innovation F= deepening of pedagogical materials G= deepening of professional materials Development of learning tools I=Field Experience Practice |
| **2** | R-2 | 75 | 90 | 79 | 90 | 90 | 86.130 | 82.930 | 82.86 | 86.88 | |
| **3** | R-3 | 90 | 90 | 90 | 90 | 90 | 78.180 | 83.180 | 78.29 | 86.67 | |
| **4** | R-4 | | | | | | | | | | |
| **5** | R-5 | | | | | | | | | | |
| **6** | R-6 | | | | | | | | | | |
| **7** | R-7 | | | | | | | | | | |
| **….** | …. | …. | …. | …. | …. | …. | …. | …. | …. | …. | |
| **911** | R-911 | 90 | 79 | 67 | 90 | 68 | 87.640 | 83.880 | 79.97 | 84.84 | |

(a). Past Experience Recognition Assessment Data (RPL)

| No | Ladder | Respondent Code | Status Pegawai | Province | Regency |
|---|---|---|---|---|---|
| **1** | SD | R-1 | Non-PNS | Central Java | Temanggung |
| **2** | SMK | R-2 | Non-PNS | Central Java | Holy |
| **3** | SD | R-3 | Non-PNS | Central Java | Temanggung |
| **4** | SD | R-4 | PNS | Central Java | Demak |
| **5** | SD | R-5 | PNS | Central Java | Pati |
| **6** | SD | R-6 | PNS | Central Java | Semarang City |
| **7** | SD | R-7 | Non-PNS | Central Java | Pati |
| **….** | …. | …. | …. | …. | …. |
| **911** | SD | R-911 | 79 | 67 | 90 |

(b). Student Personal Data

| No | Respondent Code | Graduation Status |
|---|---|---|
| **1** | R-1 | Pass |
| **2** | R-2 | Pass |
| **3** | R-3 | Pass |
| **4** | R-4 | Pass |
| **5** | R-5 | Pass |
| **6** | R-6 | Pass |
| **7** | R-7 | Not Passed |
| **….** | …. | …. |
| **911** | R-911 | Not Passed |

(c) Student Graduation Data

*3.2 Data Management*

Using Microsoft Excel software, data management is done through transformation, variable selection, data structure improvement, and unnecessary data cleaning. RPL score data includes RPL scores for four waves of PPG of the Islamic Religious Education (PAI) study program, which consists of 911 student score data from the aspects of assessing pedagogical competency development, preparation of learning tools, development of professional competencies, management of learning administration, learning innovation, deepening of pedagogical materials, deepening of professional materials, development of learning tools. Student identity data is downloaded separately according to the following table descriptions:

**Table 2.** Student identity

| No. | Batch | Number of Identities | Detected Graduation |
|---|---|---|---|
| 1 | I | 199 | 197 |

| | | | |
|---|---|---|---|
| 2 | II | 110 | 105 |
| 3 | III | 265 | 260 |

Data management in Excel involves using functions such as getting and modifying data to combine, change, and organize variables so that the data is more concise and complete. It simplifies research by allowing the selection of the necessary variables and data. Changing numeric variables can improve the data structure, such as changing the value from tens to thousands or changing the data type from date to numerical, making the analysis easier. In addition, data management is also useful for eliminating duplicate data on students' data. Well-managed data will facilitate further research and analysis, as seen in Table 3:

**Table 3.** Data Management Results

| No | Ladder | Status | Regency | Phase | IPK3 | A | B | C | D | E | F | G | H | Graduation |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | SD | Not PNS | Temanggung Regency | 1 | 3.4 | 90 | 80 | 65 | 90 | 65 | 79.5 | 81.04 | 80.3 | Pass |
| **2** | SMK | Not PNS | Sragen Regency | 2 | 3.3 | 90 | 85 | 90 | 85 | 78 | 83 | 82.75 | 51.6 | Pass |
| **3** | SD | Not PNS | Salatiga City | 3 | | | | | | | | | | Pass |
| **4** | SD | Not PNS | Kudus Regency | 1 | 3.9 | 90 | 90 | 80 | 90 | 85 | 85.7 | 85.86 | 79.5 | Pass |
| **5** | SD | PNS | Jepara Regency | 2 | 3.2 | 90 | 90 | 90 | 90 | 90 | 84.7 | 84.71 | 77.4 | Pass |
| **6** | SD | Not PNS | Pati Regency | 3 | 3.9 | 90 | 90 | 80 | 90 | 80 | 86 | 87.21 | 52.3 | Pass |
| **7** | SD | Not PNS | Temanggung Regency | 1 | 3.8 | 90 | 90 | 80 | 90 | 60 | 84.7 | 80.67 | 83 | Pass |
| **8** | …. | Not PNS | Demak Regency | 3 | 3.8 | 90 | 92 | 90 | 80 | 90 | 86.3 | 85.56 | 79 | Pass |
| **9** | SD | PNS | Jepara Regency | 2 | 3.3 | 90 | 90 | 80 | 90 | 60 | 85.6 | 83.79 | 53 | Pass |

### *3.3 Data Exploration*

The third stage of this research is data exploration. The exploration stage aims to obtain as much information as possible from the data and is useful for determining the right strategy to process the data. (Faisal & Nugrahadi, 2017). The results of data exploration obtained from the data management stage show that student personal data, RPL score data, and student graduation data are data sets with dimensions of 573 x 19 (attachments), meaning that the data consists of 19 input variables (Figure 3) and the number of students who registered for PPG in 2022 batches 1, 2 and 3 is 573 students. Both student personal data, RPL score data, and graduation data are structurally incomplete.

```
> names(data)
 [1] "Nomor.Akun"
 [2] "Jenjang"
 [3] "Nama.Lengkap"
 [4] "Status.Pegawai"
 [5] "Umur"
 [6] "Provinsi"
 [7] "Kabupaten"
 [8] "Gelombang"
 [9] "IPK3"
[10] "Pengembangan.Kompetensi.Pedagogik"
[11] "Penyusunan.Perangkat.Pembelajaran"
[12] "Pengembangan.Kompetensi.Profesional"
[13] "Pengelolaan.Administrasi.Pembelajaran"
[14] "Inovasi.Pembelajaran"
[15] "Pendalaman.Materi.Pedagogik"
[16] "Pendalaman.Materi.Profesional"
[17] "Pengembangan.Perangkat.Pembelajaran"
[18] "Praktek.Pengalaman.lapangan"
[19] "Status.Kelulusan"
```

**Figure 1.** Research variables

Student personal data is web-based data entered by students, most likely not entered incompletely in the input process. Likewise, with RPL data, some input it completely, some input it incompletely, and some do not enter data. Meanwhile, some students do not take part in the exam stages for student graduation status data, so the graduation status data is also incomplete. Figure 1 lists the completeness status of research data, partially and combined.

**Table 4**. Completeness of PPG Student Data on Personal Data, RPL Assessment and UKMPPG Exam

| Activities | Complete | Incomplete | Entire |
|---|---|---|---|
| **Personal data** | 567 | 6 | 573 |
| **RPL Assessment** | 480 | 70 | 573 |
| **UKMPPG Exam** | 562 | 11 | 573 |
| **Combined** | 479 | 94 | 573 |

Based on Table 4, 567 students (99% of the total 573) filled in their data completely, 480 students (84% of 573) filled out the RPL completely, and 562 students (98%) took the UKMPPG exam stage. However, only 479 students (84%) completed the data perfectly. Incomplete data will be considered a missing value, marked with NA. In addition, an outlier can be seen in the boxplot Figure 2. Descriptive statistical analysis does not require handling NA and outliers, but classification analysis requires clean data from both. Therefore, the data used for classification analysis must be complete and free from outliers.

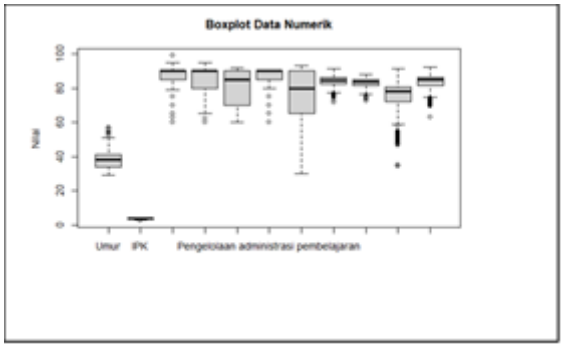
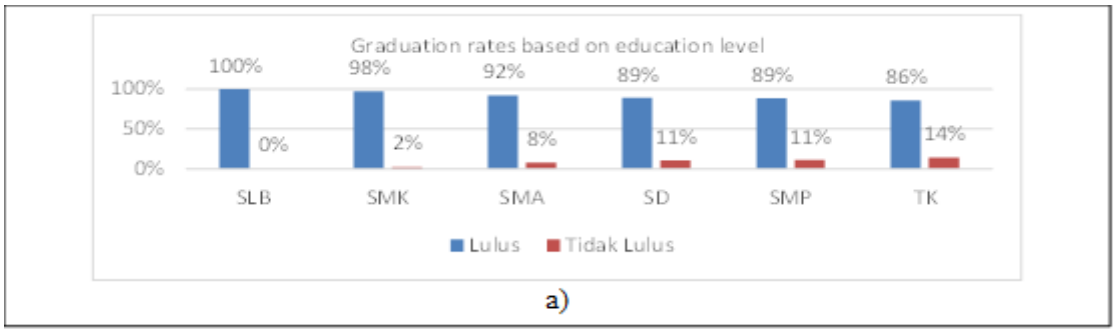


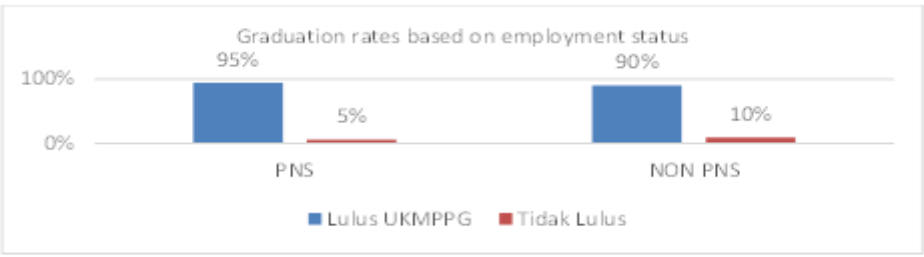

**Figure 2.** Boxplot to Detect The Presence of Outliers in The Data

### *3.4 Descriptive Statistical Analysis*

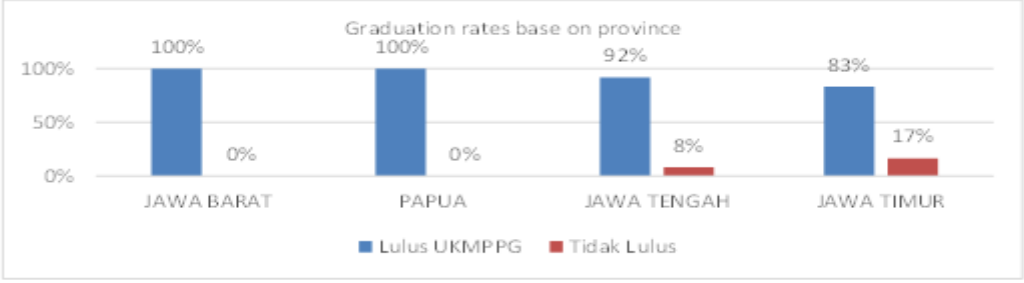
The fourth stage of this study is to conduct descriptive statistical analysis using several methods such as minimum, maximum, average, mode, median, frequency, and bar charts. This analysis is adjusted to the data type, both numerical and categorical. The research data structure consists of 13 numerical data, including account number, age (29-59), UKMPPG implementation wave (1-2), GPA (2-4), pedagogical competency development value (60-99), learning tool preparation value (60-95), professional competency development value (60-92), learning administration management value (60-90), learning innovation value (30-93), pedagogical material deepening value (71-91), professional material deepening value (72.30-88.14), learning tool development value (34.08-91.35), and practical field experience value (62.94-92.53). There are also six categorical data, such as the level of education taught, full name, employee status, province of origin, district of origin, and graduation status. The results of the descriptive statistical analysis can be found in the appendix.

The results of the descriptive statistical analysis in this study show that the difference in graduation rates based on education level, employment status, province, wave, and district where PPG students work is found at the SLB level, civil servant status, West Java and Papua provinces, wave 2, and certain districts. This information can be useful in planning, implementing, and evaluating UKMPPG to improve the overall graduation rate Figure 3.

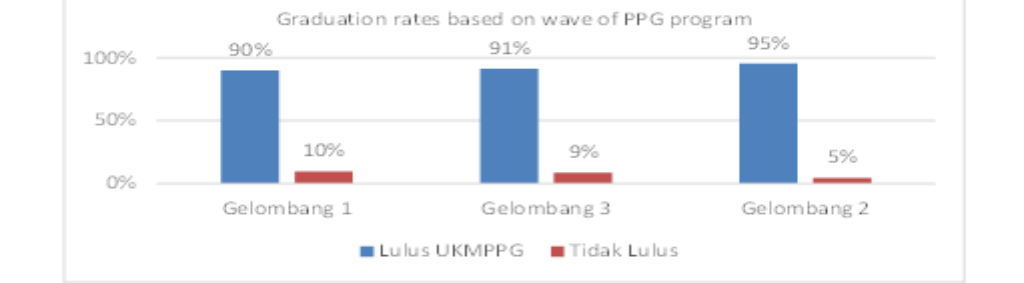



Graduation rates based on employment status
95%
90%
100%
0%
5%
10%
PNS
NON PNS
Lulus UKMPPG
Tidak Lulus

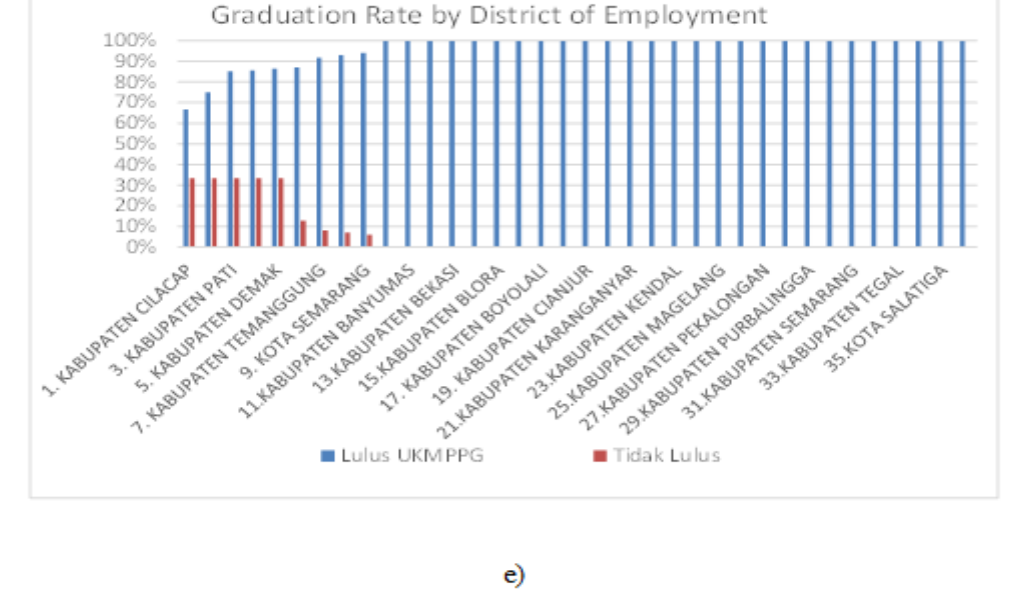
b)


Graduation rates base on province
100%
100%
92%
83%
100%
50%
0%
0%
0%
8%
17%
JAWA BARAT
PAPUA
JAWA TENGAH
JAWA TIMUR
Lulus UKMPPG
Tidak Lulus

c)


Graduation rates based on wave of PPG program
90%
91%
95%
100%
50%
0%
10%
9%
5%
Gelombang 1
Gelombang 3
Gelombang 2
Lulus UKMPPG
Tidak Lulus

d)


Graduation Rate by District of Employment
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1. KABUPATEN CILACAP
3. KABUPATEN PATI
5. KABUPATEN DEMAK
7. KABUPATEN TEMANGGUNG
9. KOTA SEMARANG
11.KABUPATEN BANYUMAS
13.KABUPATEN BEKASI
15.KABUPATEN BLORA
17. KABUPATEN BOYOLALI
19. KABUPATEN CIANJUR
21.KABUPATEN KARANGANYAR
23.KABUPATEN KENDAL
25.KABUPATEN MAGELANG
27.KABUPATEN PEKALONGAN
29.KABUPATEN PURBALINGGA
31.KABUPATEN SEMARANG
33.KABUPATEN TEGAL
35.KOTA SALATIGA
Lulus UKMPPG
Tidak Lulus

e)

**Figure 3.** UKMPPG graduation rate is based on (a) education level in the workplace, (b) Employment status, (c) Province, (d) Waves, and (e) Districts


**Figure 4.** Distribution of the number of students participating in PPG based on (a) education level in the workplace, (b) Employee status, (c) Province); (d). UKMPPG wave; and (e) Districts

The distribution of PPG program participant data, which in this case consists of data on education level, employment status, province, wave of UKMPPG, and districts, needs to be seen to inform stakeholders about the characteristics of PPG participants. This data is presented in the bar chart Figure 4.

Based on Figure 4(a), the majority of PPG students work at the elementary school level (278 students), while the smallest number is at the SLB level (2 students). There are also a large number of PPG students working in vocational schools (104 students) and a significant number working in high schools (43 students) and junior high schools (45 students). Meanwhile, only a few work at the kindergarten level, namely seven students. Figure 4(b) shows that of the total PPG students, 330 are civil servants, while 149 are non-civil servants. This indicates that the majority of PPG students work as civil servants. Figure 4(c) depicts the work sites of PPG students, with the majority (463 out of 479) working in Central Java Province. A small number work outside this province, such as in West Java (8 students), East Java (6 students), and Papua (2 students).

Meanwhile, Figure 4(d) shows the distribution of PPG students based on the wave of PPG implementation. Wave 1 and Wave 3 had more participants than wave 2, with 183 students in wave 1, 209 students in wave 3, and 87 students in wave 2, respectively. Figure 4(e) shows where PPG students work by district. Pati Regency (81 students), Demak Regency (67 students), Semarang City (51 students), and Kudus Regency (43 students) are some of the locations with the highest number of PPG students. These characteristics are important as a reference for related parties to consider the infrastructure needs needed during the PPG program to provide maximum benefits for all parties involved.

### *3.5 Preparation of Classification Data*

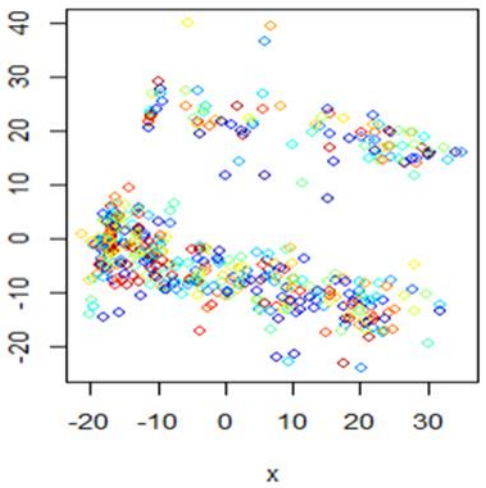
The data used in the classification analysis must be good quality, i.e. free from missing values caused by incomplete data input processes. Therefore, in this classification analysis, only complete data is used, namely data from 479 students. Classification analysis of student graduation data at UKMPPG using R software. Figure 5 is the variable data type used in the UKMPPG graduation classification process. From Figure 5, the variables used consist of 8 predictor variables (feature) and one response variable (target). Variables 1 to 8 are variables whose data type is numerical, and variable 9 is a variable whose data type is categorical, consisting of pass and fail categories. Variables 1 to 8 are called features, and variable 9 is the target variable. Table 25 shows the number of PPG students who passed and did not pass UKMPPG. 439 students passed, and 40 students did not pass UKMPPG.

```
'data.frame':   479 obs. of  12 variables:
 $ data.Umur                                 : num  37 45 33 30 40 33 39 40 35 42 ...
 $ data.IPK3                                 : num  3.44 3.33 3.92 3.17 3.89 3.83 3.81 3.25 3.83
2.92 ...
 $ data.Pengembangan.Kompetensi.Pedagogik    : num  90 90 90 90 90 90 90 90 85 90 ...
 $ data.Penyusunan.Perangkat.Pembelajaran    : num  80 85 90 90 90 90 92 90 85 90 ...
 $ data.Pengembangan.Kompetensi.Profesional  : num  65 90 80 90 80 80 90 80 80 65 ...
 $ data.Pengelolaan.Administrasi.Pembelajaran: num  90 85 90 90 90 90 80 90 90 90 ...
 $ data.Inovasi.Pembelajaran                 : num  65 78 85 90 80 60 90 60 60 65 ...
 $ data.Pendalaman.Materi.Pedagogik          : num  79.5 83 85.7 84.7 86 ...
 $ data.Pendalaman.Materi.Profesional        : num  81 82.8 85.9 84.7 87.2 ...
 $ data.Pengembangan.Perangkat.Pembelajaran  : num  80.3 51.6 79.5 77.4 52.3 ...
 $ data.Praktek.Pengalaman.lapangan          : num  85.5 81.4 87 76.2 81.7 ...
 $ data.Status.Kelulusan                     : chr  "Lulus UKMPPG" "Lulus UKMPPG" "Lulus UKMPPG"
"Lulus UKMPPG" ...
  'data.frame':   479 obs. of  13 variables:
   $ Jenjang                                : Factor w/ 6 levels "SD","SLB","SMA",..: 1 1 4 1 1 1 4 1
  1 1 ...
   $ Status.Pegawai                         : Factor w/ 2 levels "NON PNS","PNS": 1 1 1 2 1 1 1 2 2 1
  ...
   $ Umur                                   : num  37 45 33 30 40 33 39 40 35 42 ...
   $ Provinsi                               : Factor w/ 4 levels "JAWA BARAT","JAWA TENGAH",..: 2 2 2
  2 2 2 2 2 2 ...
   $ Kabupaten                              : Factor w/ 36 levels "KABUPATEN BANJARNEGARA",..: 31 29
  19 14 22 31 12 14 12 19 ...
   $ Gelombang                              : Factor w/ 3 levels "1","2","3": 1 2 1 2 3 1 3 2 1 2 ...
   $ IPK3                                   : num  3.44 3.33 3.92 3.17 3.89 3.83 3.81 3.25 3.83 2.92
  ...
   $ Pengembangan.Kompetensi.Pedagogik      : num  90 90 90 90 90 90 90 90 85 90 ...
   $ Penyusunan.Perangkat.Pembelajaran      : num  80 85 90 90 90 90 92 90 85 90 ...
   $ Pengembangan.Kompetensi.Profesional    : num  65 90 80 90 80 80 90 80 80 65 ...
   $ Pengelolaan.Administrasi.Pembelajaran: num  90 85 90 90 90 90 80 90 90 90 ...
   $ Inovasi.Pembelajaran                   : num  65 78 85 90 80 60 90 60 60 65 ...
   $ Status.Kelulusan                       : Factor w/ 2 levels "Lulus UKMPPG",..: 1 1 1 1 1 1 1 1 1
  1 ...
```

**Figure 5.** Graduation Classification Variable Data Types

**Table 5.** Student Graduation Status Data

| Colum | Pass | Not Passed |
|---|---|---|
| **Number of Students** | 439 | 40 |



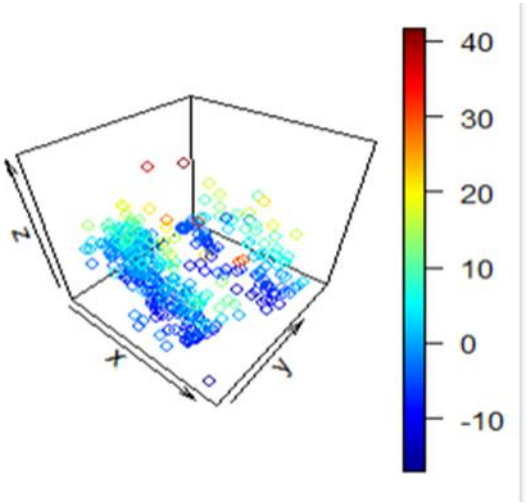



**Figure 6**. 2-Dimensional And 3-Dimensional Graphs of PCA Classification Data

Figure 6 is a 2-dimensional and 3-dimensional graph of the major component analysis (PCA). The PCA process aims to summarize the variable dimensions of a large classification into smaller dimensions, in this case**,** two and three dimensions. In addition, the PCA process also aims to see if there is any overlap between classes. From Figure 6, it can be seen that there is no overlap between the variable classes of the PCA classification, so the classification analysis process can be continued

### *3.6 Classification Data Sharing*

The distribution of the number of training data and test data is presented in Table 6. The training and testing data distribution uses the k-fold cross-validation method with four folds. The number of folds of 4 ensures that the computation runs effectively and efficiently and has stable modeling results. If the number of instances is limited, then selecting too many folds can reduce the compute performance because it requires larger compute resources, in addition to causing the number of samples in the folds to be smaller, which has an impact on decreasing the stability of the classification model estimation performance, but this will not be a problem if the computing resources are of high quality and the sample size is very large.

An illustration of the classification data division is as follows: if fold 1 becomes test data, then the rest will be training data. If fold 2 becomes test data, the rest will be training data; if fold 3 becomes test data, then the rest will be training data, and if fold 4 becomes test data, the rest will be training data. Table 6 shows that Fold 1 has 109 trajectory data and ten pass data, ten pass data fold 2 has 110 trajectory data and ten pass data, and Fold 3 has 110 trajectory data and ten trajectory data. Fold 4 has 110 trajectory data and ten trajectory data. If fold 1 is the test data, then the test data is 119, and the remaining 360 is the training data. If the fold of 2 is the test data, then the test data is 120, the training data is 350, and so on.

**Table 6.** Distribution of UKMPPG graduation classification data

| | ***Fold* 1** | ***Fold* 2** | ***Fold* 3** | ***Fold* 4** |
|---|---|---|---|---|
| | **109 Passed** | 110 Passed | 110 Passed | 110 Passed |

| | | | | |
|---|---|---|---|---|
| **10 Not Passed** | 10 Not Passed | 10 Not Passed | 10 Not Passed | |

### *3.7 Implementation of the K-Nearest Neighbors (K-NN) Classification Algorithm*

Table 7 is the student data used to implement the K-NN classification algorithm consisting of data from students who passed and did not pass in the training data group and test data. The K-NN algorithm creates a classification model using training data and tests the classification performance using test data. In fold 1, the training data consisted of 330 students who passed and 30 students who did not pass, while the testing data consisted of 109 data who passed and ten who did not pass. In fold 2, the training data consisted of 329 data from students who passed and 30 students who did not pass, while the testing data consisted of 110 data from students who passed and 10 data from students who did not pass. In fold 3, the training data consisted of 329 data from students who passed and 30 students who did not pass, while the testing data consisted of 110 data from students who passed and 10 data from students who did not pass. In fold 4, the training data consisted of 329 data of students who passed and 30 students who did not pass, while the testing data consisted of 110 data of students who passed and 10 data of students who did not pass.

**Table 7.** Amount of Training Data and Test Data on Fold 1 to Fold 4

| | | Fold 1 | Fold 2 | Fold 3 | Fold 4 |
|---|---|---|---|---|---|
| **Data Training** | Pass | 330 | 329 | 329 | 329 |
| | Not Passed | 30 | 30 | 30 | 30 |
| **Data Testing** | Pass | 109 | 110 | 110 | 110 |
| | Not Passed | 10 | 10 | 10 | 10 |

The PPG Student Graduation Classification Method is a binary class classification method because the response variable consists of two classes: passing and not passing UKMPPG. The features used for classification are age, GPA, Pedagogical Competency Development score, Learning Tool preparation value, Professional Competency Development value, Learning Administration Management value, Learning Innovation value, Pedagogical Material Deepening value, Professional Material Deepening value, and Learning Media Value Development value. The K-Nearest Neighbors algorithm can only be used if the data has features that have numerical values. This is important for identifying response data and features that are both numerical. The K-Nearest Neighbors algorithm predicts new instance categories based on information from nearby instances or closest neighbors. The distance metric commonly used in the K-NN algorithm is Euclidean Distance; therefore, the K-NN analysis in this study will use the Euclidean Distance metric.

Based on the comparison of the performance of the classification algorithm on the training data and the test data (Table 7), the classification results show that there is overfitting when applying the K-NN classification algorithm; this is shown in the performance difference between the prediction of the classification of the training data and the test data which is quite large; namely, the performance of the classification model is very good when using training data, but not so good when using test data. The values of TP, TN, FP, FN, accuracy, precision, and drawdown indicate the performance of

the K-NN classification algorithm. The higher the TP, TN, accuracy, precision, and recall, the better the algorithm, and vice versa. The lower the FP and FN, the better, and the higher the FP and FN, the worse the performance of the K-NN classification algorithm will be. For this reason, the number of closest neighbors selected in this study is small, namely 1, because increasing the number of closest neighbors will increase the potential for overfitting, and choosing a small number of closest neighbors will reduce the potential.

**Table 8.** Performance of K-NN algorithm on training data and test data

| **Data Testing** | | | | | |
|---|---|---|---|---|---|
| **Classification Performance** | Fold | | | | Average |
| | 1 | 2 | 3 | 4 | |
| **HCMC** | 106 | 100 | 97 | 98 | - |
| **TN** | 3 | 2 | 2 | 2 | - |
| **FP** | 7 | 8 | 8 | 8 | - |
| **FN** | 3 | 10 | 13 | 12 | - |
| **Accuracy** | 0.916 | 0.85 | 0.825 | 0.833 | 0,856 |
| **Precision** | 0.3 | 0.2 | 0.2 | 0.2 | 0,225 |
| **Remember** | 0.9725 | 0.9091 | 0.8818 | 0.8909 | 0,9136 |
| **Data Training** | | | | | |
| **HCMC** | 330 | 329 | 329 | 329 | - |
| **TN** | 30 | 30 | 30 | 30 | - |
| **FP** | 0 | 0 | 0 | 0 | - |
| **FN** | 0 | 0 | 0 | 0 | - |
| **Accuracy** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| **Precision** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| **Remember** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

Table 8 displays the results of the classification analysis conducted using the K-NN algorithm. These results include a confusion matrix, which is a set of values such as TP (True Positives), TN (True Negatives), FP (False Positives), FN (False Negatives), accuracy, precision, and recall. From the data in the confusion matrix in Table 8, it can be seen that in fold one there are the highest TP and TN values compared to other folds, namely TP 106 and TN 3. In addition, fold one indicates the accuracy, precision, and recall level. Highest, with an accuracy value of 0.916, precision of 0.3, and drawdown of 0.9725. Because fold one shows the best performance, the K-NN classification algorithm using data from fold one will be used to predict the graduation classification of PPG Wave 1 students in 2023.

**Table 9.** Completeness of PPG Student Data Batch 1 of 2023

| **Activity** | **Complete** | **Incomplete** | **Entire** |
|---|---|---|---|
| **RPL Assessment** | 314 | 33 | 347 |

Table 9 describes the completeness of PPG student data batch 1 in 2023. The data consists of 314 complete data and 33 incomplete data. Incomplete data means that the student has complete identity data on the PPG ww.siagapendis.com website, but RPL data for that student is unavailable. Predicting the graduation of PPG students will only be carried out on students with a complete RPL score, which is 314 students.

**Table 10.** Prediction of the Proportion of PPG Student Graduation Batch 1 in 2023

| **Graduation Status** | **Entire** | **Percentage** |
|---|---|---|
| Pass | 277 | 86,7 % |

| | | |
|---|---|---|
| Not Passed | 37 | 13,3 % |

Table 10 predicts the proportion of PPG students who graduated and failed in Wave 1 of 2023. The proportion of PPG students who passed Wave 1 in 2023 is 86.7%, and the proportion of PPG students who did not pass Wave 1 in 2023 is 13.3%. Based on the academic report on the inauguration of PPG Wave 3 in 2022, the proportion of PPG failures is 15%. A comparison of the proportion of failures between wave 3 in 2022 and wave 1 in 2023 shows that there is predicted to be a decrease in the proportion of failures in UKMPPG wave 1 in 2023, which is 1.7%.

This study revealed that based on several assessment indicators such as pedagogical competence, learning tool development, professional competence, learning administration management, learning innovation, pedagogical material understanding, professional material understanding, and learning tool development, it is estimated that 86.7% of PPG students will graduate in Wave 1 in 2023, while around 13.3% will not graduate. This is good news because compared to Wave 3 in 2022, there was a 1.7% decrease in the proportion of failures. This result reflects the improvement in PPG student performance in terms of grades such as TP, TN, FP, FN, accuracy, precision, and recall, with fold 1 showing a significant improvement, including accuracy of 0.916, precision of 0.3, and recall of 0.9725, as well as TP number 106 and TN number 3.

This study illustrates similarities with the results of the research of Monireh et al., which shows that the quality of student input significantly impacts the graduation rate of students in the Teacher Education Program (Toosi et al., 2019). Several important indicators, such as pedagogical competence, learning tool development, professional competence, learning administration management, and learning innovation, play an important role in determining students' ability to complete the final project (Ningrum, 2016). This is in line with the findings of Avalos Beatrice, which confirms that input quality has strong implications for academic outcomes (Avalos, 2011). However, this study also adds a different dimension to Irwan's analysis, arguing that the most important factor in student graduation is the management of the organizing institution, not only the quality of input (Irwan, 2016). In this context, it should be noted that each college can have students with different levels of quality (Sunaryo et al., 2020). Therefore, rankings that only consider the graduation percentage can be considered unfair because they do not consider the difference in the quality of students each university accepts.

This happens because several key factors may affect students' ability to solve the technical problems they encounter (Beauchamp & Thomas, 2009). First, there is a noticeable improvement in developing pedagogical competence, professionalism, and learning administration management skills (Stürmer et al., 2015). This interpretation means that PPG students of the 2023 phase 1 batch are more pedagogically and professionally prepared to face learning challenges (Nasikhin, Shodiq, et al., 2022). In addition, the increase in learning innovation and the development of learning tools shows that the teaching methods are more effective and attractive to students. In addition, improving the understanding of pedagogical and professional material provides a strong foundation for more effective learning (Nasikhin et al., 2021). These results also reflect the efforts made to improve the PPG program's teaching and evaluation quality. Thus, the decline in non-graduation by 1.7% in one year is evidence of a significant increase in the PPG program.

Based on the significant strength of the PPG student graduation rate in Wave 1 of 2023, several

steps may need to be taken so that the 1.7% of students who do not graduate can be pursued to achieve maximum results. First, it is important to continue to encourage the development of pedagogical competence, professionalism, and learning administration management skills for PPG students (Mushtaha et al., 2022). This can be done by increasing training and support (Simamora, 2020). Furthermore, it is necessary to continue encouraging learning innovation and developing learning tools to motivate and engage students (Kearns, 2012). Improving the understanding of pedagogical and professional material must also be considered through a more in-depth teaching approach and regular evaluation (Junaedi et al., 2022). In addition, PPG programs must continue to be monitored and improved to improve the quality of teaching and evaluation (Nasikhin, Ikhrom et al., 2022). In this way, the non-graduation reductions that have been achieved can be maintained and even improved, creating a stronger foundation for a better future in teacher education.

## 4. Conclusion

This study proves that the quality of PPG student input with indicators of pedagogical competence, professional competency development, learning management, innovation, and material understanding can be used as a prediction model to predict the future PPG implementation graduation rate. The study results show that in Batch 1 of 2023, it is estimated that 86.7% of PPG students graduated, an increase from 85% in Wave 3 in 2022. This signifies an improvement in student performance, indicated by a confusion matrix with higher TP and TN scores and significant accuracy, precision, and recall at fold 1 (accuracy 0.916, precision 0.3, recall 0.9725). However, this study is weak in generalizing the results because it was conducted only at UIN Walisongo Semarang, so the results may not be widely applicable to other institutions with different characteristics. The variability factor of student input and curriculum differences can affect the research results, and data quality is also important. To overcome this limitation, larger studies with more diverse samples from various educational institutions must apply the results more widely.

## 5. References

Abel, R. (2005). Implementing Best Practices in Online Learning. *Educause Quarterly, 28(3),* 75–77. https://www.learntechlib.org/p/103719/.

Adnan, M., & Anwar, K. (2020). Online Learning amid the COVID-19 Pandemic: Students' Perspectives. *Online Submission, 2(1),* 45–51. https://doi.org/10.33902/JPSP.2020261309.

Arifa, F. N., & Prayitno, U. S. (2019). Peningkatan Kualitas Pendidikan: Program Pendidikan Profesi Guru Prajabatan dalam Pemenuhan Kebutuhan Guru Profesional di Indonesia. *Aspirasi: Jurnal Masalah-Masalah Sosial, 10(1),* 1–17. https://doi.org/10.22212/aspirasi.v7i1.1084.

Avalos, B. (2011). Teacher professional development in Teaching and Teacher Education over ten years. *Teaching and Teacher Education, 27(1),* 10–20. https://doi.org/10.1016/j.tate.2010.08.007.

Bakia, M., Shear, L., Toyama, Y., & Lasseter, A. (2012). Understanding the Implications of Online Learning for Educational Productivity. Office of Educational Technology, US Department of Education. http://www2.ed.gov/about/offices/list/os/technology/index.html.

Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. *Decision Analytics Journal, 3,* 100071. https://doi.org/10.1016/j.dajour.2022.100071.

Beauchamp, C., & Thomas, L. (2009). Understanding teacher identity: An overview of issues in the literature and implications for teacher education. *Cambridge Journal of Education, 39(2),* 175–189. https://doi.org/10.1080/03057640902902252.

Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: A review. *Journal of Data Analysis and Information Processing, 8(4),* 341-357. DOI: 10.4236/jdaip.2020.84020.

Cunningham, P., & Delany, S. J. (2021). K-nearest neighbour classifiers-a tutorial. *ACM computing surveys (CSUR), 54(6),* 1-25. https://doi.org/10.1145/3459665.

Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D., & Marioni, J. C. (2022). Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nature Biotechnology, 40(2),* 245-253. DOI: 10.1038/s41587-021-01033-z.

Faisal, M. R., & Nugrahadi, D. T. (2017). Belajar Data Science: Klasifikasi dengan Bahasa Pemrograman R (Vol. 1). M Reza Faisal. From https://www.researchgate.net/publication/312160783_Belajar_Data_Science_Klasifikasi_dengan_Bahasa_Pemrograman_R.

Irwan, I. (2016). Kualitas Input Mahasiswa Baru UIN Alauddin Makassar Tahun 2014. *Teknosains: Media Informasi Sains dan Teknologi, 10(1),* Article 1. https://doi.org/10.24252/teknosains.v10i1.1876.

Isnain, A. R., Supriyanto, J., & Kharisma, M. P. (2021). Implementation of K-Nearest Neighbor (K-NN) algorithm for public sentiment analysis of online learning. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 15(2),* 121-130. https://doi.org/10.22146/ijccs.65176.

Jiang, L., Cai, Z., Wang, D., & Jiang, S. (2007). Survey of Improving K-Nearest-Neighbor for Classification. *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007), 1,* 679–683. https://doi.org/10.1109/FSKD.2007.552.

Junaedi, M., Nasikhin, N., & Hasanah, S. (2022). Issues in the Implementing of Online Learning in Islamic Higher Education During the Covid-19 Pandemic. *Ta'dib, 25(1),* 33–46. https://doi.org/10.31958/jt.v25i1.5365..

Kang, S. (2021). K-nearest neighbor learning with graph neural networks. *Mathematics, 9(8),* 830. https://doi.org/10.3390/math9080830.

Kearns, L. R. (2012). Student Assessment in Online Learning: Challenges and Effective Practices. 8(3), 198. https://jolt.merlot.org/vol8no3/kearns_0912.pdf.

Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy K-nearest neighbor algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-15(4),* 580–585. IEEE Transactions on Systems, Man, and Cybernetics. https://doi.org/10.1109/TSMC.1985.6313426.

Larose, D. T., & Larose, C. D. (2014). k-Nearest Neighbor Algorithm. In Discovering Knowledge in Data: An Introduction to Data Mining (pp. 149–164). Discovering Knowledge in Data: An Introduction to Data Mining. Wiley. https://doi.org/10.1002/9781118874059.ch7.

Lu, J., Qian, W., Li, S., & Cui, R. (2021). Enhanced K-nearest neighbor for intelligent fault diagnosis of rotating machinery. *Applied Sciences, 11(3),* 919.

https://www.researchgate.net/publication/348636817_Enhanced_K-Nearest_Neighbor_for_Intelligent_Fault_Diagnosis_of_Rotating_Machinery.

Mushtaha, E., Abu Dabous, S., Alsyouf, I., Ahmed, A., & Raafat Abdraboh, N. (2022). The challenges and opportunities of online learning and teaching at engineering and theoretical colleges during the pandemic. *Ain Shams Engineering Journal, 13(6),* 101770. https://doi.org/10.1016/j.asej.2022.101770.

Muslim, S. B. (2010). Supervisi Pendidikan Meningkatkan Kualitas Profesionalisme Guru. Alfabeta.

Nasikhin, Nasikhin, & Shodiq. (2021). Different Perspective of Religious Education in Islamic Theology and West Theology. *Al-Fatih: Jurnal Pendidikan Dan Keislaman, 4(2),* 328–342. https://jurnal.stit-al-ittihadiyahlabura.ac.id/index.php/alfatih/article/view/157/132.

Nguyen, T. (2015). The Effectiveness of Online Learning: Beyond No Significant Difference and Future Horizons. *MERLOT The Journal of Online Teaching and Learning, 11(2),* 309–319. http://jolt.merlot.org/Vol11no2/Nguyen_0615.pdf.

Ningrum, E. (2016). Membangun Sinergi Pendidikan Akademik (S1) Dan Pendidikan Profesi Guru (PPG). *Jurnal Geografi Gea, 12(2).* https://doi.org/10.17509/gea.v12i2.1783.

Palloff, R. M., & Pratt, K. (2007). Building Online Learning Communities: Effective Strategies for the Virtual Classroom. John Wiley & Sons. San Francisco, CA: Jossey-Bass.

Pangestika, R. R., & Alfarisa, F. (2015). Pendidikan profesi guru (PPG): Strategi pengembangan profesionalitas guru dan peningkatan mutu pendidikan Indonesia. *Prosiding Seminar Nasional, 19(1),* 671–683. https://core.ac.uk/download/pdf/33518888.pdf.

Ratnasari, Y. T. (2019). Profesionalisme Guru Dalam Peningkatan Mutu Pendidikan. Revitalisasi Manajemen Pendidikan Anak Usia Dini (PAUD) Di Era Revolusi Industri 4.0, 0, Article 0. http://conference.um.ac.id/index.php/apfip2/article/view/404.

Simamora, R. M. (2020). The Challenges of Online Learning during the COVID-19 Pandemic: An Essay Analysis of Performing Arts Education Students. *Studies in Learning and Teaching, 1(2),* 86–103. https://doi.org/10.46627/silet.v1i2.38.

Stürmer, K., Könings, K. D., & Seidel, T. (2015). Factors Within University-Based Teacher Education Relating to Preservice Teachers' Professional Vision. *Vocations and Learning, 8(1),* 35–54. https://doi.org/10.1007/s12186-014-9122-z.

Sukardi, Giatman, M., Haq, S., Sarwandi, & Pratama, Y. F. (2019). Effectivity of Online Learning Teaching Materials Model on Innovation Course of Vocational and Technology Education. *Journal of Physics: Conference Series, 1387(1),* 012131. https://doi.org/10.1088/1742-6596/1387/1/012131.

Sunaryo, H., Zuriah, N., & Handayani, T. (2020). Kesiapan Mahasiswa Pendidikan Profesi Guru (PPG) Dalam-Jabatan untuk Menempuh Program Praktik Pengalaman Lapangan. *Jurnal Pendidikan Profesi Guru, 1(1),* 29–38. https://doi.org/10.22219/jppg.v1i1.12430.

Toosi, J. F., Jamil, A. I., & Zulkifli, M. Y. (2019). Moral Autonomy and Habituation Method: A Study Based on Islamic Teachings. *Kemanusiaan the Asian Journal of Humanities, 26(1),* 47–61. https://doi.org/10.21315/kajh2019.26.s1.3.

Zulfitri, H., Setiawati, N. P., & Ismaini, I. (2019). Pendidikan Profesi Guru (PPG) sebagai Upaya Meningkatkan Profesionalisme Guru. *LINGUA: Jurnal Bahasa Dan Sastra, 19(2),* 1–130.