JURIKOM (Jurnal Riset Komputer), Vol. 9 No. 3, Juni 2022
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v9i3.4215
Hal 564−572
http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Analysis of Expertise Group Using The Fuzzy K-NN Classification Algorithm (Case Study: School of Computing Telkom University)

Jodi Kusuma[1,*], Angelina Prima Kurniati[2], Ichwanul Muslim Karo Karo[3]

[1,2]Informatika, Universitas Telkom, Bandung, Indonesia
[3]Ilmu Komputer, Universitas Negeri Medan, Medan, Indonesia
Email: [1,*]jodikusuma@student.telkomuniversity.ac.id, [2]angelina@telkomuniversity.ac.id, [3]imkarokaro@gmail.com
Email Penulis Korespondensi: jodikusuma@student.telkomuniversity.ac.id

Submitted **08-06-2022**; Accepted **24-06-2022**; Published **30-06-2022**

Abstract

The School of Computing at Telkom University has four Expertise Groups that defines the lectures taken by students. Deciding the Expertise Group, will be influential in deciding elective courses and raising the topic of the Final Project. There are many students who are still having difficulty in deciding the Expertise Group and finally only decide based on the most popular Expertise Group without seeing their potential and abilities. The impact of wrong decision of the Expertise Group are delays in graduation time. It will then affect accreditation of study program and university rank, especially in the timely graduation indicator. Therefore, it is necessary to have a system that can predict the decision of the Expertise Group for the School of Computing students based on their academic scores. In this study, prediction using the Fuzzy K-Nearest Neighbor classification algorithm was chosen because it can determine the class based on the nearest neighbor and consider ambiguous data because of the weighting value in each class. There are five tests carried out to get the best model, namely (1) examine the best split training and validation data, (2) examine the best K value, (3) compare Fuzzy K-Nearest Neighbor with Naïve Bayes and Decision Tree (C4.5) which is a commonly used classification algorithm, (4) examine the values of accuracy, precision, recall, f1-score, and (5) examine the values of accuracy using Cross-Validation method. The result is that the model made using Fuzzy K-Nearest Neighbor has an accuracy value of 72% in the case of imbalance data, 62% in the case of applying the undersampling technique, and 56% in the case of applying oversampling. Based on experiments with the other two algorithms, it was found that compared to the other two algorithms, the Fuzzy K-Nearest Neighbor has a higher accuracy value in the case of imbalance data and the case of applying to undersampling, but it has a lower accuracy in the case of applying oversampling, due to the lack of Fuzzy K-Nearest Neighbor in handling small minority data variations.

Keywords: School of Computing; Expertise Group; Classification; Fuzzy K-Nearest Neighbor; Undersampling; Oversampling

1. INTRODUCTION

Every student at the School of Computing will be required to take or decide elective courses that have been provided by the study program for both regular and fast-track students. There are 19 elective courses that are grouped based on Expertise Groups, namely Intelligent Systems, Cyber Physical Systems, Software Engineering, and Data Science [1]. The Expertise Group is useful for grouping and directing students to have special skills according to their chosen field of Informatics, as well as supervising on the Final Project later.

The main problem is whether the student chosen a Expertise Group based on their academic abilities. This allegation is also reinforced by the results of a survey conducted on Telkom University students to prove that many students are still having difficulties in determining the field of specialization or Expertise Group in the study [2]. In addition, if this continues, it can have an impact on students and experience delays in graduating on time. The Final Assignment socialization program held by the Bachelor of Informatics study program explains that students who have taken Final Project courses and have not graduated in the Odd semester 2020/2021 are 46.03%, in the Even semester 2020/2021 are 43.73%, and in the last semester, namely the Odd semester 2021/2022 are 86.17%. If this continues to occur and increases, it can affect the accreditation assessment and achievement in the graduation time of the students from the School of Computing at Telkom University.

One way to overcome or assist students in deciding an Expertise Group according to their abilities is requiring a system that can predict the decision of Expertise Groups based on students score of the compulsory courses in Informatics that has been by every student from semester 1 to semester 5. In the 2020 curriculum in semester 6, students are required to decide elective courses that will assist them in doing their Final Project. Therefore, by assisting students in decide Expertise Groups according to their academic scores, the system can assist students in deciding elective courses and Final Project topics.

In the field of data mining, there are various methods, one of which is classification. Classification is an algorithm that can classify a new object based on existing characteristics or variables, to predict an object whose class or category is still unknown [3]. Classification algorithm is considered capable of considering good and appropriate specializations or Expertise Groups [3]. There are various classification algorithms, one of which is the Fuzzy K-Nearest Neighbor [4]. The algorithm is a variant of the K-Nearest Neighbor algorithm with the Fuzzy Technique.

Several previous studies also discussed the decision of study specialization or Expertise Groups, such as in research [3] that used the K-Nearest Neighbor classification algorithm to determine the specialization of study in STMIK Amik Riau students with an accuracy of 98%. Research [5] and [6] also used a classification algorithm to determine the area of interest for Telkom University students in the Information Systems study program with an accuracy of 94.81% and 90.17%. Research [7] made predictions on determining the interest of students at Amikom University Yogyakarta using the Support Vector Machine (SVM) which is also one of the classification algorithms in data mining with an accuracy of

66%. From several studies that have been mentioned, it can be concluded that the case raised in this study, namely the decision of Expertise Groups for the Bachelor of Informatics study program students, is very likely to be solved using a classification algorithm. Beside that, making predictions using a classification algorithm does not always produce a high accuracy value, because it could be that the algorithm chosen is not in accordance with the case raised [7]. The reason for choosing the Fuzzy K-Nearest Neighbor classification algorithm is because research [4] implemented the Fuzzy K-Nearest Neighbor algorithm to determine whether students graduate on time or not with an accuracy value of 98%. And in research [8] a comparison of the Fuzzy K-Nearest Neighbor classification algorithm with the Decision Tree (C4.5) and Naïve Bayes classification algorithms has been carried out. It is found that the highest accuracy of Fuzzy K-Nearest Neighbor is 98% and the average accuracy is 96%, in Decision Tree (C4.5) the highest accuracy is 86% and the average accuracy is 79.5%, and in Naïve Bayes the highest accuracy is 90% and the average the accuracy is 87.5%. Therefore, the Fuzzy K-Nearest Neighbor classification algorithm in this study is suitable to be applied to the case raised. Previous study [8] also found that the accuracy obtained by Fuzzy K-Nearest Neighbor is higher than Decision Tree (C4.5) and Naive Bayes. In addition, the Fuzzy K-Nearest Neighbor has stable accuracy and a higher accuracy value when compared to K-Nearest Neighbor. This due to the fuzzy initialization value that plays a role in obtaining membership values for determining the output class [4].

This research aims to create a system that can predict the decision of Expertise Groups (Intelligent System, Cyber Physical System, Software Engineering, and Data Science) for students at the School of Computing at Telkom University based on their academic scores using the Fuzzy K- Nearest Neighbor classification algorithm. This system is useful for students to get an idea of deciding elective courses and raising the topic of the Final Project according to their Expertise Group. In addition, it also indirectly helps the Bachelor of Informatics study program in reducing the occurrence of delays in graduating students on time due to their wrong choice of specialization.

2. RESEARCH METHODOLOGY

In this study, we will create a system that can predict the decision of Expertise Groups based on students academic score from semesters 1 to 5 using the Fuzzy K-Nearest Neighbor. There are several stages carried out in building the system to be made, which are preprocessing, splitting data, building a classification model, conducting validation tests on the model, and making predictions on data whose class or category is not known. An overview of the system, in general, can be seen in Figure 1.

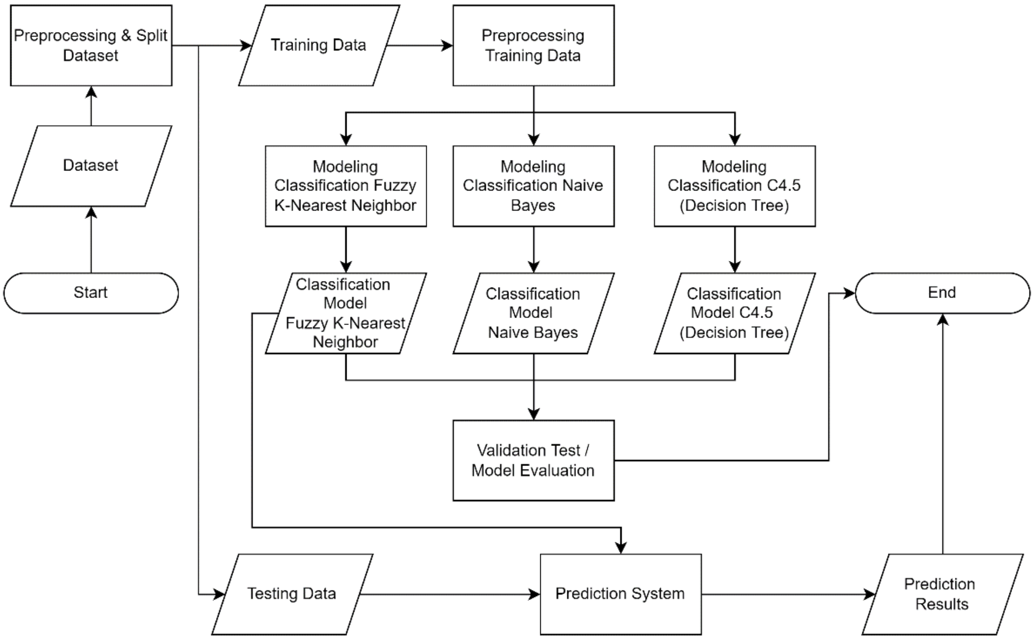
Start
Dataset
Preprocessing & Split Dataset
Training Data
Preprocessing Training Data
Modeling Classification Fuzzy K-Nearest Neighbor
Modeling Classification Naive Bayes
Modeling Classification C4.5 (Decision Tree)
Classification Model Fuzzy K-Nearest Neighbor
Classification Model Naive Bayes
Classification Model C4.5 (Decision Tree)
Validation Test / Model Evaluation
End
Testing Data
Prediction System
Prediction Results

Figure 1. Classification System Flow Schematic

2.1 Data Set

In this study, we used data on the scores of bachelor of Informatics students at Telkom University batch 2016 and 2017 as Training and 2018 as Testing. The reason for choosing students in this batch for testing data is because in the batch 2018 and above there are still many students who have not graduated in the odd semester of 2021/2022 so there are many students who have not taken or retaken a course. Students of the 2018 class are expected to graduate on time in year 2022, because the normal study period for bachelor is 4 years. Besides that, the batch meets the criteria for students who have reached or passed semester 5, so the score from semester 1 to 5 that are still empty will decrease, when compared to using data from the batch 2019 or later. The system predicts which Expertise Group a student is more suitable for based on his academic score from semesters 1 to 5.

The student score data obtained from the School of Computing study program until the most recent semester data, namely the odd semester 2021/2022. The dataset consists of a total of 364365 data in .csv format. However, the dataset is still unstructured so it is necessary to do the next step, namely data preprocessing.

2.2 Preprocessing Dataset

Data preprocessing is the initial stage to make sure that the dataset can be used or processed at a later stage. Data preprocessing changes the dataset that is unstructured and does not match the needs of this study.

Table 1. Sample Data Before Preprocessing

ID	BATCH	SUBJECT_CODE	SUBJECT_NAME	SEMESTER	…	INDEXPOINT
sfff91ba5fc6	1617	CSH2B3	Language Theory & Automata	1718-1	…	A
sfff91ba5fc6	1617	CCH1D4	Data Structure	1819-1	…	AB
sfff91ba5fc6	1617	CCH4D4	Final Project	1819-1	…	A
…	…	…	…	…	…	…
s0024bd259	1819	CII3E3	Cyber Security	2021-2	…	C

The dataset obtained from the School of Computing study program can be seen in Table 1. Each row of data contains student data with one course and its index. The goal of this preprocessing stage is to make the data more efficient for this study, by combining all course scores taken by a student into one data row (courses are used as columns). The preprocessing stages carried out in this study are as follows.

a. Dropping columns that are not used, namely SEMESTERS, PRESENCE, COMPONENT NAME, POINT, TOTAL_POINT, PLO_NAME, PLO_NUMBER, and COURSE_ID.
b. Checking Missing Values (empty/null data) and Duplicate Data on the dataset, then drop the data.
c. Splitting data based on their batch (1617, 1718, and 1819).
d. Changed course names that have been changed in the 2016 curriculum to the 2020 curriculum in each batch.
e. Changing the Index Value from an alphabetic to numeric (encoding process) with the provisions of the value A = 4, AB = 3.5, B = 3, BC = 2.5, C = 2, D = 1, E and T = 0 in each batch.
f. Create and enter data in the previous dataset into a new dataset that only contains the student ID and the names of the courses as columns.

Changing the data structure can reduce the number of rows of data in the dataset from 364365 to 2213 rows, where 807 rows in the 2016 batch, 774 in the 2017 batch, and 632 in the 2018 batch. The results of the preprocessing in one of batch can be seen in Table 2.

Table 2. Sample Data After Preprocessing

ID	OPERATING SYSTEM	STATISTICS	DATA STRUCTURE	INTERACTION DESIGN	BASIC PHYSICS	…	CALCULUS
sff7860618	4.0	3.5	3.5	3.0	4.0	…	3.5
sfeef8b458	0.0	3.0	3.0	3.5	2.5	…	4.0
sfe3a916bf	1.0	2.0	2.0	3.0	3.0	…	2.0
…	…	…	…	…	…	…	…
s01d2ea2f6	4.0	4.0	3.5	2.5	3.0	…	2.5

2.3 Preprocessing Training Data

This preprocessing data focuses on the 2016 and 2017 batch as training data that has been processed in the previous stage. This stage produced output in the form of a dataset that already has a class label on each data , that has been set based on the Expertise Group of the completed Final Project. The steps involved in this process are as follows.

a. Combining the 2016 and 2017 batch into one dataset with the condition only takes data of students who have graduated based on score in the Final Project course. So the total data in training data is 981 rows.
b. Dropping courses that are not included in semesters 1 to 5 of the 2020 curriculum.
c. Because there is no prior information on which a student belongs to the Expertise Group, the authors do a manual class determination based on their academic scores from semesters 1 to 5 using equation (1). The authors find out which subject belongs to an Expertise Group by using the plotting data of the courses in each Expertise Group obtained from one of the Head of Expertise Group.

$$\frac{\left((the\ score\ of\ each\ Course\ *\ Credits)\ +\ Other\ Subjects\ in\ the\ same\ Expertise\ Group\right)}{total\ Credits} \tag{1}$$

d. After obtaining the total scores for each Expertise Group, the highest score is taken as the class in the data, and if there are 2 or more of the same highest scores will be dropped because it is ambiguous. Then combine the dataset with the class (Intelligent System = IS, Cyber Physical System = CPS, Software Engineering = SE, and Data Science = DS).

e. Check and drop the data if one of the course score is 0 (null), because the dataset only contains data on the score of students who have graduated, while the courses in the dataset are mandatory courses that will be taken by every student in Bachelor of Informatics study program.

After passing the stages above, the result is that a student is included in an Expertise Group which is based on his scores. To create a classification model, training data is needed that has a class label on each data, then to find out how accurate the predictions is, validation data is needed which is taken from training data. Therefore, this dataset will be further divided into Training and Validation data to train the model that has been made, based on the best split training and validation data. The results of the preprocessing training data stages can be seen in Table 3.

Table 3. Sample of Training and Validation Data

ID	ADVANCED CALCULUS	COMPUTER NETWORK	DATA STRUCTURE	MATH LOGIC	DIGITAL SYSTEM	…	EXPERTISE GROUP
sfb76e562d	2.0	4.0	2.0	3.5	4.0	…	IS
sf94d59ce7	4.0	4.0	3.5	2.0	4.0	…	SE
sfc1b668f5	4.0	3.5	2.5	2.0	3.5	…	DS
…	…	…	…	…	…	…	…
sf447e5828	2.0	2.0	4.0	2.5	2.0	…	CPS

2.4 Testing Scenarios

Table 4 shows that there is an imbalance of data in each class, that it can affect the results obtained [9]. Therefore, the author also adds several testing scenarios in the application of the Fuzzy K-Nearest Neighbor, namely in the presentation of training data that will be used in the model to be created. Oversampling and Undersampling are two of many ways to overcome data imbalance as was done in research [10]. Scenario 1 will use the existing dataset without applying undersampling or oversampling techniques, Scenario 2 will use a dataset with undersampling techniques, and Scenario 3 will use a dataset with oversampling techniques.

Tabel 4. Students In Each Expertise Group on Training and Validation Data

Expertise Group	Number of Students
Intelligent System	138
Cyber Physical System	459
Software Engineering	125
Data Science	259

2.5 Testing Data

Testing data uses the 2018 batch of 632 student score, which serves to prove the classification model that has been created using the Fuzzy K-Nearest Neighbor algorithm can predict class based on training data, to see the difference in the prediction results obtained in the model that has been made in each test scenario. In testing data, preprocessing was done by dropping courses that are not included in semesters 1 to 5 of the 2020 curriculum only. So form testing data is the same as in Table 2, but the column score of the courses in semesters 1 to 5 only.

2.6 Classification Fuzzy K-Nearest Neighbor

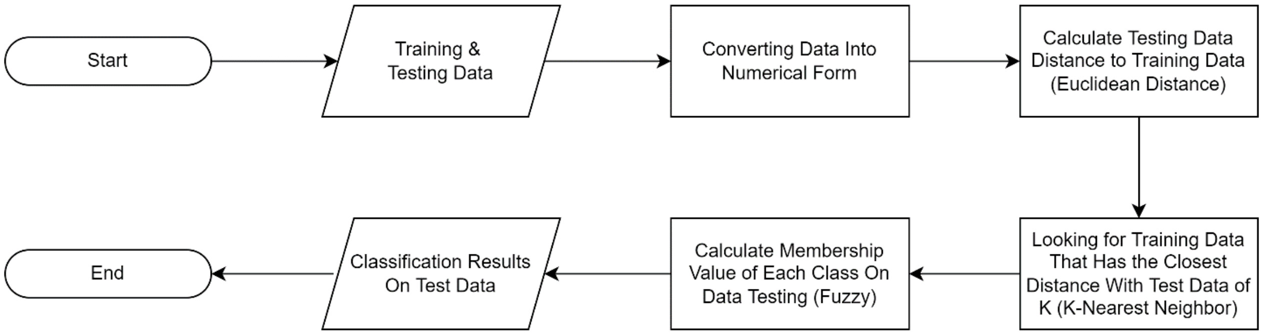
Start
Training & Testing Data
Converting Data Into Numerical Form
Calculate Testing Data Distance to Training Data (Euclidean Distance)
Looking for Training Data That Has the Closest Distance With Test Data of K (K-Nearest Neighbor)
Calculate Membership Value of Each Class On Data Testing (Fuzzy)
Classification Results On Test Data
End

Figure 2. Overview of the Fuzzy K-Nearest Neighbor Classification Process

Fuzzy K-Nearest Neighbor is one of the classification algorithms that combines the Fuzzy technique with the K-Nearest Neighbor Classifier [8]. Fuzzy K-Nearest Neighbor has two advantages compared to the K-Nearest Neighbor algorithm, which are: (1) it can consider if there are ambiguous properties of neighbors and (2) each object will have a degree of membership in each class so that it will give more strength or confidence to an object in a class [8]. The Fuzzy K-Nearest Neighbor algorithm is a classification algorithm used to predict the class label of the testing data using the nearest neighbor and the membership value of the testing data in each class [4].

Fuzzy K-Nearest Neighbor classification algorithm was chosen in this study because this algorithm can classify data based on the closest distance, can consider if there are ambiguous objects, and has a high level of accuracy based on research [8] when compared with the Decision Tree (C4.5) and Naïve Bayes classification. In addition, the Fuzzy K-Nearest Neighbor assigns each object with membership degree value in each class which can increase object confidence in a class that is not done in K-NN algoritm, so if there are two or more neighboring distances that are the same, it will be ambiguous. In general, the classification uses Fuzzy K-Nearest Neighbor by following the steps taken based on the reference [4] :

a. Performing Fuzzy initialization calculations, as in equation (2) :

$$u_{ij} = \begin{cases} 0{,}51 + \left(\frac{n_j}{k}\right) * 0{,}49 & \text{, if } j = i \\ \left(\frac{n_j}{k}\right) * 0{,}49 & \text{, if } j \neq i \end{cases} \tag{2}$$

Where :
u_{ij} = class membership value i on vector j
n_j = number of class members j on a dataset K
k = number of nearest neighbors
j = target class

b. Calculate the Euclidean distance of the testing data to the training data.
c. Sort by smallest euclidean value.
d. Determine the closest k records.
e. Calculate the degree of membership of the new data for each class using the equation (3).

$$u_i(x) = \frac{\sum_{j=1}^{k} u_{ij}\left(1/\|x-x_j\|^{\frac{2}{m-1}}\right)}{\sum_{j=1}^{k}\left(1/\|x-x_j\|^{\frac{2}{m-1}}\right)} \tag{3}$$

Where :
$u_i(x)$ = membership value data x to class i
k = number of nearest neighbors
$x - x_j$ = data distance difference x to data x_j in K nearest neighbor
m = weight exponent whose magnitude is m > 1

f. Select the class that has the greatest membership value as a result.

2.7 Validation Test / Model Evaluation

To measure the performance of classification model that has been made, several validation tests or model evaluations will be carried out by calculating the values of accuracy, precision, recall, and f1 score, using a confusion matrix (can be seen in Table 5).

Tabel 5. Confusion Matrix

	Prediction Class	
True Class	True Positive (TP)	False Negative (FN)
	False Positive (FP)	True Negative (TN)

Where True Positive (TP) is a correctly predicted result, True Negative (TN) is incorrect result, False Positive (FP) is an unexpected result, and False Negative (FN) is a missed result [11] [12].

a. Accuracy measures how accurate a model is in carrying out the classification process and determining which new data belongs to which class correctly [13], using the equation (4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

b. Precision measures the level of accuracy between the information requested by the user and the answer given by the system [14], using the equation (5).

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

c. Recall measures the success rate of the system in retrieving information [14], using the equation (6).

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

d. F1 Score is useful for describing the comparison of the average results of weighted precision and recall [13], using the equation (7).

$$F1\ Score = \frac{2*TP}{2*TP+FP+FN} = 2*\frac{precision*recall}{precision\ +\ recall} \qquad (7)$$

3. RESULTS AND DISCUSSION

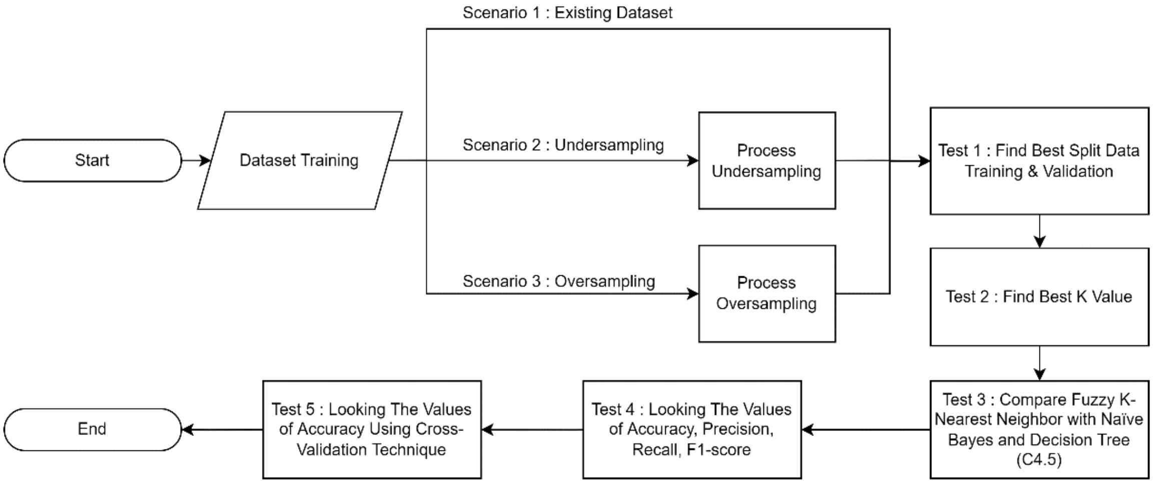
Scenario 1 : Existing Dataset
Start
Dataset Training
Scenario 2 : Undersampling
Process Undersampling
Test 1 : Find Best Split Data Training & Validation
Scenario 3 : Oversampling
Process Oversampling
Test 2 : Find Best K Value
End
Test 5 : Looking The Values of Accuracy Using Cross-Validation Technique
Test 4 : Looking The Values of Accuracy, Precision, Recall, F1-score
Test 3 : Compare Fuzzy K-Nearest Neighbor with Naïve Bayes and Decision Tree (C4.5)

Figure 3. The Flow of Tests Carried Out

3.1 Testing

In classification model that has been made using Fuzzy K-Nearest Neighbor, several tests were carried out to get the best model performance results and solve the problems that have been mentioned. The tests are: (1) testing the best split data in the training and validation data, (2) testing the best K value (number of nearest neighbors), (3) testing comparisons Naïve Bayes and C4.5 based on the highest accuracy value obtained in each test, (4) testing at the values of accuracy, precision, recall, and f1-score base on best split data and best K value, and (5) testing the values of accuracy using cross-validation, in three test scenarios.

In scenario 1, the model is tested using the existing dataset without applying undersampling or oversampling techniques or the dataset are still imbalance because data distribution in each class are 1:5:1:3 for DS:SE:IS:CPS respectively.

Scenario 2 is testing the model using a training and validation data with undersampling technique, by generalizing the amount of data in each class based on the data in the least class [10]. Then based on Table 4 the data on student at least in the Software Engineering class with 125 data. By applying the undersampling technique, it produced a total of 500 data with 125 data in each class.

Scenario 3 is testing the model using a training and validation data with oversampling technique, by generalizing the amount of data in each class based on the data in the most populated class . The assumption is that rather than discarding important data, it is better to duplicate the data in the least class to balance the amount of data in the most class [15]. So based on Table 4, the most populated class is the Cyber Physical System class with 459 rows, by applying the oversampling technique it produced a total of 1836 rows with 459 rows in each class.

The first test is to find the best split data based on the accuracy value that will be obtained, using K = 5 in each scenario. Determining the best split data will affect the results of the classification model that will be made [16]. Based on Figure 4, the results of the best split data on scenarios 1 and 2 is 0.8 (80%) training data and 0.2 (20%) validation data with an accuracy of 72% and 62%. And scenario 3 shows that the best split data is 0.2 (20%) training data and 0.8 (80%) validation data with an accuracy of 56%.

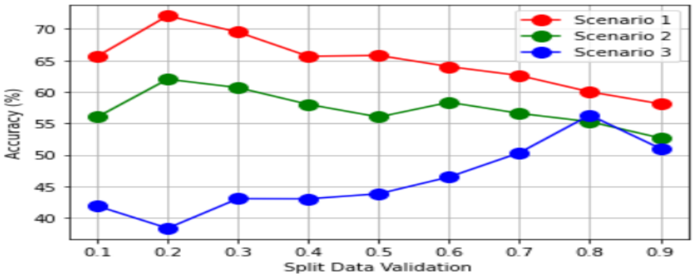

Figure 4. Best Split Training and Validation Data in Each Scenario

The second test is to find the best K value or the number of the closest neighbors in each scenario and find out the effect of the K value on the Fuzzy K-Nearest Neighbor. Find the best K value from 3 to 39, we take only odd values, to avoid ambiguity on the class voting stage [17]. Scenarios 1, 2, and 3 has been done using the best splits data that have been searched for previously in each scenario. In Figure 5 it can be seen that in for scenarios 1, 2, and 3 the best K value is 5, although the accuracy obtained in each scenario varies. It can be concluded that the higher iteration value causes a decrease in accuracy, this is related to a high K. Because K is high, it causes the value of the validity of the training data to be low. Beside that, the comparison of data from the results of weighted voting also increases, then causes system errors in classifying [18].

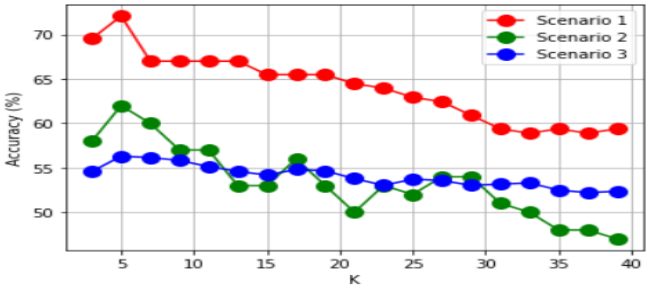

Figure 5. Finding the Best K Value in Each Scenario

The third test is doing a comparison of Fuzzy K-Nearest Neighbor with Naïve Bayes and Decision Tree (C4.5) at the same cases. These two algorithms were chosen as the two most commonly used classification algorithms [8]. The results of the accuracy comparison in each scenario can be seen in Figure 6, where in scenario 1 and 2 the highest accuracy is Fuzzy K-Nearest Neighbor (FKNN), but in scenario 3 the highest accuracy is Decision Tree (C4.5).

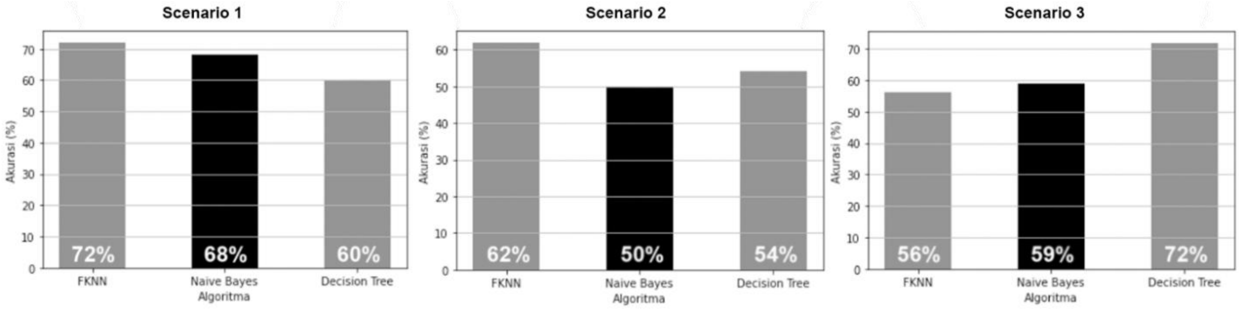

Figure 6. Comparison the Accuracy of Classification Algorithm in Each Scenario

The fourth test is looking the values of accuracy, precision, recall, and f1-score in each scenario. It can be seen in Table 6 that the highest accuracy is found scenario 1. Scenario 2 has a more stable recall value, because the data in each class has the same amount and the data is more varied (there is no duplicate data), and if averaged for each class, it produced a higher recall value than scenarios 1 and 3. In scenario 3, the low precision, recall, and f1-score values affects the accuracy value. This is due to the oversampling technique in scenario 3 which duplicates the fewest classes, causing a lack of data variation in the oversampling selected class. However, the accuracy value obtained in each scenario only taken based on the best split data, so it is necessary to check the accuracy value with the cross-validation method.

Tabel 6. Accuracy, Precision, Recall, And F1-Score In Each Scenario

Scenario	Class	Precision	Recall	F1-Score	Accuracy
1	Intelligent System	0.68	0.44	0.54	72%
	Cyber Physical System	0.75	0.88	0.81	
	Software Engineering	1.00	0.31	0.47	
	Data Science	0.64	0.85	0.73	
2	Intelligent System	0.73	0.67	0.70	62%
	Cyber Physical System	0.59	0.55	0.57	
	Software Engineering	0.77	0.57	0.65	
	Data Science	0.44	0.80	0.57	
3	Intelligent System	0.44	0.85	0.58	56%
	Cyber Physical System	0.71	0.61	0.66	
	Software Engineering	0.62	0.28	0.39	
	Data Science	0.69	0.51	0.59	

The fifth test is looking the values of accuracy using Cross-Validation in each scenario and classification algorithm. Cross-Validation is a method used to get a more valid accuracy value when compared using a confusion matrix. The way cross-validation works is by dividing the dataset into training and testing based on the number of folds [19], then take a few percent of the data in the training data as validation data. By using the cross-validation, all the data in the dataset act as a training and testing data so that it produced a valid accuracy value based on the average accuracy of the fold. Using cross-validation in this study uses fold = 5, so the dataset will be divided into training and testing data five times. The accuracy values using the cross-validation can be seen in Table 7, where is the same as the results in the third test. The highest accuracy value in scenarios 1 and 2 is Fuzzy K-Nearest Neighbor, and in scenario 3 is Decision Tree (C4.5). But the difference is that the accuracy value obtained is different from the third test, namely the Decision Tree (C4.5) scenario 3 algorithm has a higher accuracy value than the others.

Table 7. Accuracy Results for Each Classification Model Using Cross-Validation

Scenario	Algorithm	Average Accuracy
1	FKNN	**62%**
	Naïve Bayes	61%
	Decision Tree (C4.5)	59%
2	FKNN	**55%**
	Naïve Bayes	53%
	Decision Tree (C4.5)	40%
3	FKNN	42%
	Naïve Bayes	66%
	Decision Tree (C4.5)	**85%**

3.2 Prediction Results on Testing Data

After making a classification model using Fuzzy K-Nearest Neighbor in each scenario, predictions on new data whose class labels are not yet known by using testing data, using a classification model that has been trained and carried out various tests using training and validation data to get the best model in each scenario. The prediction results obtained from each scenario in testing data can be seen in Table 8, where scenario 2 and 3 have more balance proportion for each Expertise Groups.

Table 8. Plotting Prediction Results on the Testing Data for Each Scenario

Scenario	Class	Number of Students (Persons)	Total Students (%)
1	Intelligent System	32	5%
	Cyber Physical System	398	63%
	Software Engineering	21	3%
	Data Science	181	29%
2	Intelligent System	80	13%
	Cyber Physical System	139	22%
	Software Engineering	197	31%
	Data Science	216	34%
3	Intelligent System	92	15%
	Cyber Physical System	211	32%
	Software Engineering	174	28%
	Data Science	155	25%

3.3 Analysis of Test Results

Based on the tests that have been carried out in each scenario, scenarios 1, 2, and 3 have the same best K value that is K = 5, so the number of close neighbors to an object that is used as a comparison from testing data to training data is 5. The best split data from training data, where training data will be used to train the classification model and validation data will be used to find out how accurate the predictions is. In scenarios 1 and 2 have the same value that is 80% as training data and 20% as validation and at scenario 3 is 20% as training data and 80% as validation data, this happens because the difference in the amount of data in the data is more in scenario 3. Scenario 1 has a higher accuracy value when compared to scenarios 2 and 3, but scenario 2 has a more stable recall value in each class. Recall is the ratio of true positive predictions compared to the overall data that is true positive. It means that the higher the recall value, more data items are predicted correctly in each class. Then, using the undersampling and oversampling techniques to overcome the imbalanced training data, the prediction results obtained in the testing data.

4. CONCLUSION

Fuzzy K-Nearest Neighbor algorithm the highest accuracy is 72% in scenario 1, and the accuracy obtained by applying cross-validation method is 62%. Compared to the Naïve Bayes and Decision Tree (C4.5) algorithms for scenarios 1 and 2, Fuzzy K-Nearest Neighbor has the highest accuracy value. But, in scenario 3 the highest accuracy is the Decision Tree (C4.5) algorithm and the lowest is Fuzzy K-Nearest Neighbor. This is due to the limitation of Fuzzy K-Nearest Neighbor in handling small minority data variations. So it can be concluded Fuzzy K-Nearest Neighbor can be used in this study, and the accuracy obtained is better in the case of imbalance data (scenario 1) and cases of undersampling application (scenario 2) when compared to the other two classification algorithms. But if looking for the best classification models in this study, using Decision Tree (C4.5) with oversampling techniques produced a higher accuracy value on cross-validation method when compared to other scenarios and classification algorithms, namely 85%.

The results of the study show that the highest accuracy obtained by applying Fuzzy K-Nearest Neighbor in cross-validation method is not as high as those of the Decision Tree (C4.5) algorithm. This could be the Fuzzy K-Nearest Neighbor is not suitable for the case in this study, based on the finding that the accuracy value is lower in the use of cross-validation method. Another possible cause is that the amount of data in the training data is still lacking, because the total number of student data for the 2016 and 2017 batch is 1581 rows, after only taking data from students who have graduated in that batch, it becomes 981 rows. Therefore further studies can add the number of data scores of students who have passed, use or add other variables and make comparisons on the prediction results obtained with the actual data, compare the proportion of student distribution in each Expertise Group from the prediction results with the proportion of lecturers, or can try to focus on doing further research using the Decision Tree (C4.5) in the same case, as the algorithm with the highest accuracy value in the cross-validation method in this study.

REFERENCES

[1] “Kelompok Keahlian - Telkom University,” Telkom University, 2020. [Online]. Available: https://telkomuniversity.ac.id/kelompok-keahlian/. [Accessed 29 Apr 2022].

[2] A. C. Febryanti, I. Darmawan and R. Andreswari, “Modelling Of Decision Support System For Fields Of Interest Selection With Simple Additive Weighting Method Case Study : Bachelor Program Of Information System Telkom University,” e-Proceeding of Engineering, vol. 4, no. 2, pp. 3114-3121, 2017.

[3] N. Lizarti and A. N. Ulfah, “Penerapan Algoritma K-Nearest Neighbor Untuk Penentuan Peminatan Studi STMIK Amik Riau,” Fountain of Informatics Journal, vol. 4, no. 1, pp. 1 - 7, 2019.

[4] A. S. P. Anugerah, Indriati and C. Dewi, “Implementasi Algoritme Fuzzy K-Nearest Neighbor untuk Penentuan Lulus Tepat Waktu (Studi Kasus : Fakultas Ilmu Komputer Universitas Brawijaya),” Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 2, no. 4, pp. 1726-1732, 2018.

[5] S. N. Latifah, R. Andreswari and M. A. Hasibuan, “Prediction Analysis of Student Specialization Suitability using Artificial Neural Network Algorithm,” dalam International Conference on Sustainable Engineering and Creative Computing (ICSECC), 2019.

[6] A. R. Manurung, R. Andreswari and M. A. Hasibuan, “Analisis Prediksi Pemilihan Bidang Peminatan Berdasarkan Rekam Data Akademik Menggunakan Algoritme C4.5 (Studi Kasus : Mahasiswa Sistem Informasi Universitas Telkom),” dalam Conference on Information Technology and Electrical Engineering (CITEE 2019), Yogyakarta, 2019.

[7] F. K. Wattimury and E. Seniwati, “PENENTUAN PEMINATAN MAHASISWA PRODI INFORMATIKA DI UNIVERSITAS AMIKOM YOGYAKARTA MENGGUNAKAN SVM,” INTECHNO Journal - Information Technology Journal, vol. 1, no. 4, pp. 15-18, 2019.

[8] P. E. Mas`udia, R. Rismanto and A. Mas`ud, “Analysis of Comparison of Fuzzy Knn, C4.5 Algorithm, and Naïve Bayes Classification Method for Diabetes Mellitus Diagnosis,” International Journal of Computer Applications Technology and Research, vol. 7, no. 8, pp. 363-369, 2018.

[9] R. D. Fitriani, H. Yasin and Tarno, “PENANGANAN KLASIFIKASI KELAS DATA TIDAK SEIMBANG DENGAN RANDOM OVERSAMPLING PADA NAIVE BAYES (Studi Kasus: Status Peserta KB IUD di Kabupaten Kendal),” JURNAL GAUSSIAN, vol. 10, no. 1, pp. 11-20, 2021.

[10] Y. B. Wah, H. A. A. Rahman, H. He and A. Bulgiba, “Handling Imbalanced Dataset Using SVM and k-NN Approach,” dalam AIP Conference Proceedings, 2016.

[11] I. M. K. Karo, A. Khosuri and R. Setiawan, “Effects of Distance Measurement Methods in K-Nearest Neighbor Algorithm to Select Indonesia Smart Card Recipient,” dalam International Conference on Data Science and Its Applications (ICoDSA), 2021.

[12] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” BMC Genomics, vol. 21, no. 6, pp. 1-13, 2022.

[13] I. M. K. Karo, A. T. R. Dzaky and M. A. Saputra, “Comparative Analysis of K-Nearest Neighbor and Modified K-Nearest Neighbor Algorithm for Financial Well-Being Data Classification,” Indonesia Journal of Computing, vol. 6, no. 3, pp. 25-34, 2021.

[14] I. M. K. Karo, A. Khosuri, J. S. I. Septory and D. P. Supandi, “Pengaruh Metode Pengukuran Jarak pada Algoritma k-NN untuk Klasifikasi Kebakaran Hutan dan Lahan,” JURNAL MEDIA INFORMATIKA BUDIDARMA, vol. 6, no. 2, pp. 1174-1182, 2022.

[15] W. Ustyannie and Suprapto, “Oversampling Method To Handling Imbalanced Datasets Problem In Binary Logistic Regression Algorithm,” Windyaning Ustyannie, vol. 14, no. 1, pp. 1-10, 2020.

[16] A. Rácz, D. Bajusz and K. Héberger, “Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification,” Multidisciplinary Digital Publishing Institute, vol. 26, no. 4, p. 1111, 2021.

[17] A. B. Hassanat, M. A. Abbadi and G. A. Altarawneh, “Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach,” International Journal of Computer Science and Information Security (IJCSIS), vol. 12, no. 8, pp. 33-39, 2014.

[18] F. Wafiyah, N. Hidayat and R. S. Perdana, “Implementasi Algoritma Modified K-Nearest Neighbor (MKNN) untuk Klasifikasi Penyakit Demam,” Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 1, no. 10, pp. 1210-1219, 2017.

[19] F. Tempola, M. Muhammad and A. Khairan, “PERBANDINGAN KLASIFIKASI ANTARA KNN DAN NAIVE BAYES PADA PENENTUAN STATUS GUNUNG BERAPI DENGAN K-FOLD CROSS VALIDATION,” Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), vol. 5, no. 5, pp. 577-584, 2018.