Using KNN Algorithms for Determining the Recipient of Smart Indonesia Scholarship Program

1st Purwanto
Magister Teknik Komputer
Politeknik Caltex Riau
Pekanbaru, Indonesia
purwanto20s2tk@mahasiswa.pcr.ac.id

2nd Dadang Syarif Sihabudin Sahid
Magister Teknik Komputer
Politeknik Caltex Riau
Pekanbaru, Indonesia
dadang@pcr.ac.id

***Abstract*—The Smart Indonesia Card (KIP) scholarship program is a government scholarship program through the Ministry of Religion of the Republic of Indonesia which is given to students who have a good academic level but have a weak economic level. Sultan Syarif Kasim State Islamic University, Riau accepts new students every year, but the quota for the KIP scholarship program is limited. With the limited quota for the KIP program, a system is needed that is able to classify submission data from students who register for the KIP program, so that the selection process can be carried out, quickly, precisely, and in accordance with the required quota. In this study, the K-Modes and K-Nearest Neighbor (KNN) Algorithms were used by using the achievement variables, report cards, and national exam scores when high school, father's income, parental status, and homeownership status. Reprocessing is carried out before the testing stage, testing is carried out by performing the initial stages, namely clustering using the K-Modes algorithm, then validating or testing data by applying the Grid Search Cross-Validation (GSCV) method, and finally predicting using the KNN algorithm. The test resulted in a performance value of 66.79%.**

Keywords—KIP scholarship, Classification, Clustering, Validation, Prediction.

I. INTRODUCTION

There is no term "poor children are prohibited from going to school or college" in this country. Those who are less able and have achievements must continue to study up to the level of higher education through the Smart Indonesia Program (PIP). PIP is assistance in the form of cash, expansion of access, and learning opportunities from the government given to students and students who come from families unable to pay for education. KIP Lectures are proof of the state's presence to help its citizens obtain the right to higher education. The nation's children at college-age do not lose hope to sit as low and stand as high. KIP College will ensure the continuity of student studies and it is hoped that it will break the chain of poverty with the emergence of a profile of the nation's children who are characterized, intelligent, and prosperous [1].

Sultan Syarif Kasim State Islamic University, Riau the registrants for the smart Indonesia card scholarship program has increased from year to year, while the quota for the KIP scholarship program is limited by the Ministry of Religion. The selection process for the KIP scholarship program has so far been carried out in a semi-online and manual way, where students register for the KIP registration application through the iRaise(academic information system) portal, then after registration is closed, the KIP scholarship program manager will conduct an interview process to validate the data. which has been filled in by the student who has registered, after the verification process, then a manual selection will then be carried out to determine whether this student is eligible or not to receive the KIP scholarship program.

With the manual selection system, the selection process becomes difficult and takes a long time, with the manual system, the right system is needed to carry out the KIP scholarship program selection process so that it can produce a fast and targeted selection process according to the government's program in giving equal rights in terms of education.

The framework in this research is to combine the K-Modes clustering method to assign classes based on cluster results. After the class is set, then validate or test the data by applying Grid Search Cross-Validation. Finally, predictions will be made using the KNN algorithm. KNN is the best prediction algorithm with a high level of efficiency when compared to other methods such as Naive Bayes [2].

II. RELATED WORKS

Govindarajan, [3]. With high data variance, KNN is highly recommended. KNN is an algorithm that is very efficient in using resources but still produces predictions with high accuracy and precision.

Tun, [4]. Similar to the research conducted by Govindarajan, [3], This study also applies KNN to select prospective scholarship recipients. The advantages of KNN with low computation so that it can be easily integrated with other applications. In this study, KNN was run using an application built with the programming language C#NET.

Surarso, [5]. This research was carried out for the classification of the study program's work budget. KNN is

applied with validation with the K-Fold Cross Validation method so that it gets a value for the performance of 77.96%.

Kurniadi, [6]. This work has the same purpose as the research to be carried out, namely predictions for scholarship recipients at universities. The variables used in this work include; student name, semester, GPA, Parent Income, and Dependents. Predictions by applying the KNN algorithm get a performance value of 95.83%.

Zhou, [7]. This research was conducted to analyze the advantages and disadvantages of the k-mode algorithm on categorical data. By using the complex K-Modes Based on Global-Relationship Dissimilarity (KMBGRD) dataset that has been analyzed previously, K-Modes works very effectively and stably.

Chaturvedi, [8]. This study compares several cluster algorithms, including K-Means, K-Medians, K-Modes. The performance of K-Means and K-Medians is limited in contrast to K-Modes which can group more than 10,000 respondents and 100 more variables including 50 categorical data.

Inspired by some of the studies above, with our case, it is more appropriate if we apply the clustering process using the K-Modes algorithm and the classification stages by applying the KNN algorithm.

III. THEORETICAL BASIC

A. Program scholarship Kartu Indonesia Pintar (KIP)

The Indonesian Smart College Card, hereinafter referred to as KIP Lecture, is social assistance in the form of tuition fees provided by the government to students who are economically disadvantaged and have the good academic potential to continue their studies at the diploma (D3) and undergraduate (S1) programs.

The requirements given by the Ministry of Religion to be able to take part in this KIP scholarship program are as follows:

- New students graduated from MA/MAK/SMTK/SMAK/SMA/equivalent in the current year and a maximum of 2 years before;
- Students who are currently studying in the previous year's batch;
- Has economic limitations but has good academic potential supported by valid documentary evidence; and
- Not involved and/or indicated to participate in activities/organizations that are contrary to Pancasila and the Unitary State of the Republic of Indonesia as evidenced by the signing of the integrity pact.

Proof of fulfillment of requirements:

- Economic limitations are evidenced by the ownership of a national assistance program in the form of the Smart Indonesia Card (KIP) or Prosperous Family Card (KKS), or Jakarta Smart Card (KJP).
- If a student does not have a KIP or a parent/guardian does not yet have a KKS, then they can still register to get a KIP Tuition provided that they meet the requirements of being economically incapable by the provisions, as evidenced by the combined gross income of the parents/guardians of IDR 4,000,000.00 (four million rupiahs) or the combined gross income of parents/guardians divided by the maximum number of family members of Rp. 750,000.00 (seven hundred and fifty thousand rupiah).
- The final decision of the recipient will be taken by the respective PTK.

B. Data Normalization

Normalization of data is needed in the preprocessing stage with the use of narrowing the range of training data and distributing data evenly. In normalization, there are several techniques, including min-max normalization, z-score normalization, decimal scaling and sigmoidal normalization. In this study, the focus is on min-max normalization. The general equation for min-max normalization is as follows [9]:

$$V' = \frac{v-A}{A-A} . (new_A - new_A) + new_A \quad (1)$$

V' = The value of the new data is the result of normalization
v = Value of data before normalization
new_{maxA} = Latest maximum value limit
new_{minA} = Latest minimum score
Max A = Maximum value in column
Min A = Minimum value in column

C. Algoritma K-Nearest Neighbor

KNN is one of the algorithms with the simplest and most intuitive technique in the field of statistical discrimination. The process of calculating the Euclidean distance in this algorithm is first to create a dataset from training data whose class has been known in advance, the next step is that the dataset from the test data will be classified based on the closest distance from each training data and depending on the value of k used. Euclidean equation to calculate the distance between test data points and training data points where i is a representation of attribute values and n attribute dimensions are shown in equation (2) [10].

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (2)$$

d(x,y) = Euclidean distance between training data points x and test data
x_i = Sample training data
y_i = Testing Data
n = Dimension attribute

D. K-Modes Clustering

K-modes were first introduced by Huang as a clustering method which was developed from the k-means method. Therefore k-modes are efficient like k-means but are used on categorical data.

The following are the steps for k-modes clustering [11] :

1. Select the starting mode some k
2. Allocate data objects to the nearest cluster based on a simple dissimilarity measure.

3. After all data objects have been allocated to a cluster, recheck the value of an object against the mode. If a data object turns out to be the closest modelto be in another cluster, move the object to the appropriate cluster and update the second mode of the cluster.
4. Repeat Step 3 until no data objects change cluster.

IV. RESEARCH METHODOLOGY

A. Flowchart

The following is a flow chart of the research methodology carried out. This study aims to provide an assessment of predictions in awarding KIP scholarships to new students.

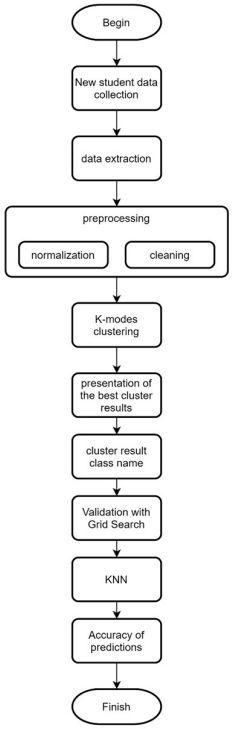

Fig. 1. Flowchart research methodology

Figure 1 shows some of the algorithms used in doing this research. The goal with a combination concept like this is to increase the value of objectivity in the system for the selection of KIP scholarship recipients.

B. Sources of data and research variables

The sample data in this study came from the Student Affairs section of the Sultan Syarif Kasim State Islamic University, Riau in 2019/2020.

Variable data in this KIP dataset include ;

```
Index(['nim', 'rata2_un', 'rata2_raport', 'status_ayah', 'status_ibu',
       'penghasilan_ayah', 'kepemilikan_rumah'],
      dtype='object')
```

Fig. 2. List of KIP dataset variables or columns

The description of the dataset is as follows:

	nim	rata2_un	rata2_raport	status_ayah	status_ibu	penghasilan_ayah	kepemilikan_rumah
count	3.258000e+03	3258.000000	3258.000000	3258.000000	3258.000000	3.258000e+03	3258.000000
mean	1.145212e+10	117.138046	146.446470	1.217004	1.045427	1.268812e+06	1.702271
std	1.671295e+09	1017.726912	509.253807	0.690649	0.408415	2.102468e+07	1.328602
min	1.526201e+09	0.000000	0.000000	0.000000	0.000000	-1.000000e+06	0.000000
25%	1.161613e+10	49.175000	84.890000	1.000000	1.000000	1.000000e+00	1.000000
50%	1.172120e+10	58.660000	87.570000	1.000000	1.000000	1.000000e+06	1.000000
75%	1.185042e+10	70.682500	90.000000	1.000000	1.000000	1.500000e+06	1.000000
max	1.198032e+10	39590.000000	9035.330000	4.000000	3.000000	1.200000e+09	5.000000

Fig. 3. Description of the KIP dataset

Based on Figures 2 and 3, it can be seen a list of variables will be used in the clustering and classification process. And it can be seen also the spread of data for each variable.

V. RESULTS AND DISCUSSION

In this section, the stages of the method used will be explained to obtain an accuracy value from the prediction of the selection of KIP scholarship recipients.

A. Preprocessing

This stage is an important stage, where the status of the readiness of the dataset depends on this stage. First, in this study, the data cleaning process will be carried out and then data preparation and produce a table as shown in the following figure :

	nim	rata2_un	rata2_raport	status_ayah	status_ibu	penghasilan_ayah	kepemilikan_rumah
0	0	1070	1053	1	1	76	1
1	1	1096	1129	1	1	81	1
2	2	865	1363	1	1	56	1
3	3	1162	1048	3	3	2	1
4	4	931	1095	1	2	72	1

Fig. 4. Results from the preprocessing stage

B. K-Modes Initialization

After preprocessing the data, the next step is to determine the initialization of k-modes that will be used for clustering. In this study, the initializations compared include Huang and Cao initialization. Based on the lowest cost, then Cao initialization will be used.

Using K-Mode with "Cao" initialization

```
[ ] km_cao = KModes(n_clusters=2, init = "Cao", n_init = 1, verbose=1)
    fitClusters_cao = km_cao.fit_predict(kip)

    Init: initializing centroids
    Init: initializing clusters
    Starting iterations...
    Run 1, iteration: 1/100, moves: 6, cost: 13317.0
```

Fig. 5. Cao Initialization

Using K-Mode with "Huang" initialization

```
[ ] km_huang = KModes(n_clusters=2, init = "Huang", n_init = 1, verbose=1)
    fitClusters_huang = km_huang.fit_predict(kip)

    Init: initializing centroids
    Init: initializing clusters
    Starting iterations...
    Run 1, iteration: 1/100, moves: 31, cost: 13807.0
```

Fig. 6. Huang Initialization

C. Determination of the number of clusters

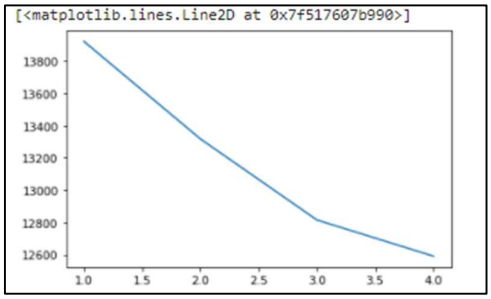

Fig. 7. Setting the number of clusters

Figure 7 is a visualization of the elbow method commonly used to determine the number of clusters in the clustering process. According to the visualization, the best cluster is 3.

D. K-Modes Clustering

	nim	rata2_un	rata2_raport	status_ayah	status_ibu	penghasilan_ayah	kepemilikan_rumah	cluster_predicted
0	1526201056	74.53	91.43	1	1	1600000	1	0
1	1526201408	75.89	92.42	1	1	1800000	1	0
2	1526204357	67.00	3116.33	1	1	1000000	1	1
3	1572101873	79.78	91.37	3	3	0	1	0
4	1572104865	69.05	91.95	1	2	1500000	1	2

Fig. 8. Results of the clustering process by K-Modes

Based on the results of the clustering above, it can be seen that the cluster_predicted column has given a label according to each cluster.

E. Cluster identification

The following process is a process that must be carried out so that the results of the clustering can be used in the classification process.

Before giving identity, it is necessary to analyze using visualization as reference material for giving identity.

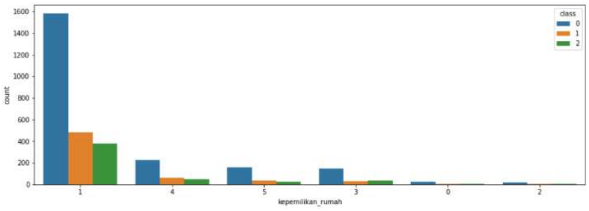

Fig. 9. Visualization of a cluster deployment on the status variable of kepemilikan_rumah column.

Figure 9 describes the distribution of data in each specific cluster for the kepemilikan_rumah column status variable. It can be seen clearly that the majority of cluster 0 is in part 1, namely the status of homeownership is private property. And it can be assumed that cluster 0 is a student whose economic status is middle to upper, therefore cluster 0 is a group of students who are not entitled to receive KIP scholarships.

```
[ ] combinedDf.loc[combinedDf['class'] == 0, 'class'] = 'tidak diterima'

[ ] combinedDf.loc[combinedDf['class'] == 1, 'class'] = 'butuh pertimbangan'

[ ] combinedDf.loc[combinedDf['class'] == 2, 'class'] = 'diterima'

[ ] combinedDf.head()
```

	nim	rata2_un	rata2_raport	status_ayah	status_ibu	penghasilan_ayah	kepemilikan_rumah	class
0	1526201056	74.53	91.43	1	1	1600000	1	tidak diterima
1	1526201408	75.89	92.42	1	1	1800000	1	tidak diterima
2	1526204357	67.00	3116.33	1	1	1000000	1	butuh pertimbangan
3	1572101873	79.78	91.37	3	3	0	1	tidak diterima
4	1572104865	69.05	91.95	1	2	1500000	1	diterima

Fig. 10. Giving status to the class variable.

F. Min-Max Normalization

```
x=df_to_class.iloc[:,1:9].values
y=df_to_class['class']
# normalization
x=(x-np.min(x))/(np.max(x)-np.min(x))

#train test split
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test=train_test_split(x,y,test_size=0.3)
```

Fig. 11. Min-Max Normalization

This stage is carried out to normalize variable data so that the data range is narrower and easier to analyze.

G. Validation with Grid Search Cross-Validation

In the classification process, the validation process is a step that must be done. Validation serves to maximize data sharing for further training and testing.

```
# Grid search cross validation
from sklearn.model_selection import GridSearchCV

grid ={"n_neighbors":np.arange(1,50)}
knn= KNeighborsClassifier()
knn_cv=GridSearchCV(knn,grid,cv=10) #GridSearchCV
knn_cv.fit(x_train,y_train)
```

Fig. 12. Grid Search Cross-Validation

H. Accuracy value

```
# test your model
knn.fit(x_train,y_train)
print("test accuracy :",knn.score(x_test,y_test))

test accuracy : 0.5838445807770961
```

Fig. 13. The level of accuracy of the method if not using validation.

```
# print hyperparameter KNN
print("hyperparameter K:",knn_cv.best_params_)
print("Accuracy (best score):",knn_cv.best_score_)

hyperparameter K: {'n_neighbors': 33}
Accuracy (best score): 0.6679824561403509
```

Fig. 14. The level of accuracy of the method if using validation Grid Search Cross-Validation.

VI. CONCLUSION

From the results of the research above, it can be concluded that testing the predictions of students who are entitled to receive KIP scholarships using a combination method starting from clustering, classification, and validation has an accuracy rate of 66.79%. The effect of validation on the KNN classification method is very significant, the level of accuracy without the validation method is 58.38%. The level of accuracy mentioned above is quite good for datasets with high variance.

REFERENCES

[1] K. A. R. I. Direktorat Jenderal Pendidikan Islam, Keputusan Menteri Agama Republik Indonesia Nomor 361 Tahun 2020 tentang Pedoman Program Kartu Indonesia Pintar Kuliah pada Perguruan Tinggi Keagamaan. Indonesia, [Online]. Available: https://kemenag.go.id/archive/keputusan-menteri-agama-nomor-361-tahun-2020-tentang-pedoman-program-kartu-indonesia-pintar-kuliah-pada-perguruan-tinggi-keagamaan, 2020.

[2] H. Parvin, H. Alizadeh, and B. Minaei-bidgoli, "MKNN : Modified K-Nearest Neighbor," *Proc. World Congr. Eng. Comput. Sci. WCECS*,

pp. 22–25, 2008.

[3] M. Govindarajan and R. Chandrasekaran, “Evaluation of k-Nearest Neighbor classifier performance for direct marketing,” *Expert Syst. Appl.*, vol. 37, no. 1, pp. 253–258, 2010, doi: 10.1016/j.eswa.2009.04.055.

[4] K. T. Tun and A. M. Aye, “Selection of Appropriate Candidates for Scholarship Application Form using KNN Algorithm,” *Int. J. Sci. Eng. Technol. Res.*, vol. 03, no. 06, pp. 1019–1026, 2014.

[5] B. Surarso and R. Gernowo, “Implementation of the K-Nearest Neighbor Method to determine the Classification of the Study Program Operational Budget in Higher Education,” *Proceeding of ICOHETECH*, pp. 201–204, 2019, [Online]. Available: http://ojs.udb.ac.id/index.php/icohetech/article/view/803.

[6] D. Kurniadi, E. Abdurachman, H. L. H. S. Warnars, and W. Suparta, “The prediction of scholarship recipients in higher education using k-Nearest neighbor algorithm,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 434, no. 1, 2018, doi: 10.1088/1757-899X/434/1/012039.

[7] H. Zhou, Y. Zhang, and Y. Liu, “A Global-Relationship Dissimilarity Measure for the k -Modes Clustering Algorithm,” *Comput. Intell. Neurosci.*, vol. 2017, 2017, doi: 10.1155/2017/3691316.

[8] A. Chaturvedi, K. Foods, P. E. Green, and J. D. Carroll, “K-modes clustering,” *Journal of Classification*, vol. 18, no. 1. pp. 35–55, 2001, doi: 10.1007/s00357-001-0004-3.

[9] Y. K. JAIN and S. K. BHANDARE, “Min Max Normalization Based Data Perturbation Method for Privacy Protection,” *Int. J. Comput. Commun. Technol.*, vol. 4, no. 4, pp. 233–238, 2013, doi: 10.47893/ijcct.2013.1201.

[10] K. Schliep, K. Hechenbichler, and A. Lizee, “Weighted k-Nearest Neighbors,” *2016*, vol. 399, p. 15, 2016, [Online]. Available: https://cran.r-project.org/web/packages/kknn/kknn.pdf.

[11] Z. Huang, “Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values,” *Data Min. Knowl. Discov.*, vol. 2, pp. 283–304, 1998, doi: 10.1023/A:1009769707641.