IJEMD

IJEMD-CSAI, 1 (2) (2022), 1 – 12 https://doi.org/10.54938/ijemdcsai.2022.01.2.100

International Journal of Emerging Multidisciplinaries:
Computer Science & Artificial Intelligence

Research Paper
Journal Homepage: www.ojs.ijemd.com
ISSN (print): 2791- 0164 ISSN (online): 2957-5036

IJEMD

COMPUTER SCIENCE & ARTIFICIAL INTELLIGENCE

Verification of Covid-19 Social Assistance Recipients using Naïve Bayes Classifier

Ramzi Kamali[1], Yunita S. Sari[2], Ismat Aldmour[3] and Rahmat Budiarto[1,*]

[1]Department of Informatics, Faculty of Computer Science, Universitas Mercu Buana, Jakarta, Indonesia
[2]Department of Information System, Faculty of Computer Science, Universitas Mercu Buana, Jakarta, Indonesia
[3] Department of Computer Science, Faculty of Computer Science and IT, Albaha University, Albaha, Saudi Arabia
**Corresponding author*

Abstract

The Indonesian government launches the Covid-19 social assistance program to reduce the impacts of the economic downturn during the pandemic. The recipients of social assistance in Sukabumi Selatan District of Jakarta Province is collected form Neighborhood Association (RT/RW). However, this mechanism has limitations in terms of feasibility assessment through direct verification which is not optimal due to social restriction activities. At the same time, data is also collected through the regular recipients of social aid program, so there is a data discrepancy that causing a bias in determining the recipients' feasibility. Therefore, a mechanism is required to assess the eligibility of the recipients. This study aims to assist Social Service Agency of Sukabumi Selatan district, in assessing the eligibility of the recipients using Naïve Bayes classifier and K-Nearest Neighbors (K-NN) classifier as comparison. Experiments using Cross-Industry Standard Process for Data Mining (CRISP-DM) model were carried out on a collected dataset, and the results show that Naïve Bayes classifier shows the best result with 93% accuracy, 86% precision and 100% recall, while K-NN has 90% accuracy, 82% precision and 98% recall. This research may assist the Social Service Agency of the district to determining more accurately the target recipients.

Keywords: Covid-19 Pandemic; Social assistance; Data Classification; Naïve Bayes; K-NN.

1. Introduction

A new type of Corona virus that was first discovered in the city of Wuhan, China at the end of December 2019 known as Corona Virus Disease 2019 (COVID-19). This virus is not only spreading in China, even massively in number of countries in the world including Indonesia. WHO then declared the Covid-19 as a pandemic on March 11, 2020. The outbreak of Covid-19 around the world has created a global crisis. The Indonesian government itself took steps by implementing Community Activity Restrictions (PPKM) to

Email addresses: 41518010153@student.mercubuana.ac.id (R. Kamali), yunita.sartika@mercubuana.ac.id (Y. S. Sari), mutdm@yahoo.com (I. Aldmour), rahmat.budiarto@mercubuana.ac.id (R. Budiarto)

reduce the spread of Covid-19 through the instruction of the Minister of Home Affairs Number 15 of 2021 concerning Restrictions on Emergency Community Activities for Corona Virus Disease 2019 in Java and Bali regions. The restriction affects economic aspects of the people. In response to this situation, the Government immediately implemented social policies to overcome the impacts that emerged such as loss of people's income, increasing poverty rates and long-term economic crisis.

The impact of the pandemic has been tremendous for household incomes. Statistical Central Bureau (BPS) recorded an increase in the number of unemployed people from previously 6.93 million people in February 2020 to 8.75 million people in February 2021. That is why the Jakarta Provincial Government prepare assistance scheme to minimize the impacts of economic decline by issuing a Covid-19 social assistance program for communities affected by the pandemic. The program is handled directly by the Social Service Agency of Jakarta Province. The distribution of Covid-19 social aid programs administered under the policy of the Governor with regards to the implementation of PPKM activities on the distribution of Covid-19 social aid programs.

The recipients data of Covid-19 soscial assistance in Sukabumi Selatan District of Jakarta Province is based on data collected form Neighborhood Association (RT/RW) and handed over to the local government. However, this mechanism has limitations in terms of feasibility assessment through direct verification which is not optimal due to social restriction activities. At the same time, data is also collected through the regular recipients of social aid program, so there is a discrepancy since this data is relatively old. The pandemic has also caused people who previously had good finances or were categorized as unfit to receive social assistance to experience a decrease in finances, making them included in the category of being eligible to receive social assistance. Thus, there is a bias in determining the feasibility of the people as recipients of Covid-19 social assistance.

To overcome this problem, a study was conducted to classify accurately recipients of covid-19 social aid program in Sukabumi Selatan district. This study adopts Naïve Bayes method as classifier and considers K-NN as a comparison. Some research works have been carried out using Naïve Bayes method. The research work by Naeni, et al [1] discussed the prediction of poor student beneficiaries (BSM) in MAN 2 high school, Lampung. The researchers use three methods, i.e.: Naïve Bayes, Decision Tree and K-NN. The student’s attributes: siblings, parental work, parental income, KIP recipients, family status are considered. The number of sampling was 393 students. The results of the study on the precision value show that Naïve Bayes method outperformed the other two methods while result on accuracy and recall show that Decision Tree is the best.

Another study conducted by Firasati, et al [2] discussed the classification of poor people receiving social assistance in Somokerto village, Central Java Province, by comparing two algorithms, namely, Naïve Bayes and K-NN methods. The experimental results showed that Naïve Bayes produced a higher accuracy than K-NN (89.04% compare to 87.67%).

Research work by Safri, et al [3] discussed the feasibility of the Healthy Indonesia Card (KIS). The researchers worked on total of 200 records with 15 determinants of feasibility in 2017 taken at the Pekalongan Regency Social Service, Central Java Province using K-NN method and a combination of K-

NN and Naïve Bayes method. The results showed the accuracy of the feasibility using K-NN method was 64%, while the combination of K-NN and Naïve Bayes was 96%. Thus, the combination of K-NN and Naïve Bayes methods performs very well in determining the eligibility of healthy Indonesia card recipients (KIS).

Research work conducted by Tempola, et al [4] discussed the classification of smart Indonesian card recipients (KIP) using also Naïve Bayes and K-NN on 150 datasets. The experimental results showed that the classification system without Naïve Bayes validation had better accuracy, where an average accuracy of 85.66% for Naïve Bayes and an average accuracy of K-NN was 84.89%. However, when validation is applied, the accuracy of K-NN (88.7%) is better compared to Naïve Bayes which was only 81.3%.

This study aims to assist Social Service Agency, especially in Sukabumi Selatan district, in determining the eligibility of Covid-19 social aid recipients and increasing the accuracy of Covid-19 social aid distribution. With this study, it is hoped that the implementation of the Naïve Bayes method in determining covid-19 social aid recipients can be used as a reference for Social Service Agency in warning Covid-19 social aid recipients to be right on target to residents affected by the pandemic.

2. Research Method

This research work utilizes the Cross-Industry Standard Process for Data Mining (CRISP-DM) model. CRISP-DM is a method that provides a standard for data mining and can be applied to a common problem-solving strategy [5]. The CRISP-DM model consists of six phases, namely business understanding, data understanding, data preparation, modeling, evaluation and deployment [6]. The CRISP-DM stages are presented in Figure 1. In this study, it was only carried out until the evaluation phase.

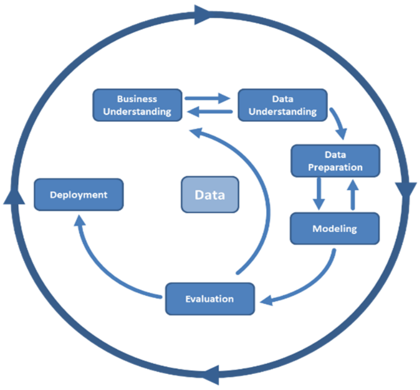


Figure 1. CRISP-DM Model

The CRISP-DM model was applied in this study, so that the research stage was adjusted as shown in Figure 2. The research stage consists of: (i) collecting data on Covid-19 social aid recipients, where the dataset comes from Social Service Agency, Sukabumi Selatan district; (ii) conducting data preprocessing; (iii) splitting data; (iv) implementing Naïve Bayes classifier method then comparing to the K-NN classifier method; and finally the evaluation and validation stages of the model.

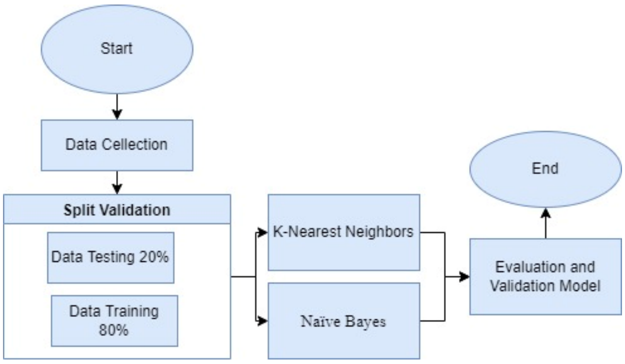


Figure 2. Research Flow

Business Understanding

This stage is a stage of understanding the object of the study [7]. The study was conducted on data of recipients of *Covid-19* social assistance which indicated that there was a lack of target for recipients of the aid program. In this study, a comparison of the classification algorithms of K-NN and *Naïve Bayes* was carried out to help Social Service Agency determines the eligibility of *Covid-19* social aid recipients. At this stage, an understanding is also made to find the best classification method so that it can help during the data processing process.

Data Understanding

At this stage, data collection is carried out, starting form understanding the data, analyzing the data and the parameters to be used [8]. This study used dataset of recipients of the Covid-19 social aid program in Sukabumi Selatan district as of April 2021. The dataset consists of 6471 records with 12 initial attributes shown in Table 1.

Table 1. Dataset’s attributes

NO	Dataset	
	Attribute Name	**Description**
1	Idpusdatin	Recipient ID
2	Full Name	Recipient's address
3	Village Name	Distribution point
4	RW	Recipient's RW address
5	RT	Recipient's RT address
6	Siak_No_KK	Registered KK number
7	Siak_NIK	Registered NIK number
8	Nama_LNGKP	Recipient's name
9	Gender	Gender of the recipient
10	Place of Birth	Place of birth of the recipient
11	Date of Birth	Date of birth of the recipient
12	KET	The work of the recipient

Data Preparation

This stage is a process for data preparation to be carried out by preprocessing data through re-examining the correlation of each attribute so that it becomes quality data and is ready to be modeled [9]. The data preparation stage can also be referred to as the preprocessing stage. The stages for data preparation include: (i) Data cleaning that aims to check and clean unneeded blank values; (ii) Data reduction that adjusts the number of attributes used, because not all attributes will be a condition for the determining attribute; (iii) Data transformation is process by which data changes with a certain format are carried out. This process aims to make the data more suitable for the classification stage.

Modeling

At this stage, the results of the data preparation process will be proceeded with the modeling that has been proposed for classification using Naïve Bayes classifier and K-Nearest Neighbors classifier as a comparison. The modeling process will test the two models with the aim to obtaining the most accurate model, so that the level of comparison of its accuracy can be seen immediately.

Evaluation

The evaluation stage aims to determine the usefulness of the model that has been successfully created in the previous modeling step [10]. This study uses the evaluation stage with five-fold cross validation tests, where the values for *k* are: 5, 10, 15, 20, and 25.

The validation process consists of two sub-processes, namely, training data (training set) and testing data (testing set). The training sub process is used to train a predefined classifier model obtained during the modeling stage. After the classifier model is trained at the subprocess training stage, the model will then be tested in the testing subprocess.

This evaluation process also looks at the accuracy results on Naïve Bayes and K-NN classifier models based on their confusion matrix, to determine whether the models have produced an appropriate assessment. There are number of metrics used to evaluate or assess the classification models, including accuracy (1) as the degree of proximity between the predicted value and the actual value, precision (2) as the level of accuracy between the information requested by the user and the answer given by the system and recall (3) as the success rate of the system in rediscovering information [11].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

Thus, the evaluation metrics for the classification model there are: TP (True Positive), i.e.: the result of the correct classification, TN (True Negative), i.e.: the result of improper classification, FP (False Positive), i.e.: the result of the right classification but the fact is not correct, and FN (False Negative), i.e.: the result of improper classification but the fact is correct [12].

3. Results And Discussion

This section describes the experimental results on the performance of each classifier. The experiments are carried out on a computer with the following specifications. Processor 110 Intel Celeron N3060, 4 GB RAM, running Windows 10 operating system and Python language programming on Google Colab platform. The dataset is split into composition of 80% for training data and 20% for testing data. Prior to performing experiments on classification, the following data preparation steps are conducted.

Preprocessing

Preprocessing is process that serves to prepare raw data into data that can be used in building computer learning models [13]. This study used the following stages of preprocessing:

- Data Cleaning

  At this stage it is intended to clean the data from noise such as data duplication, missing value, outlier [14]. There are several outliers in the data of covid-19 social aid recipients and error in the recipient's name, therefore deletion is carried out on the outlier.

- Data Reduction

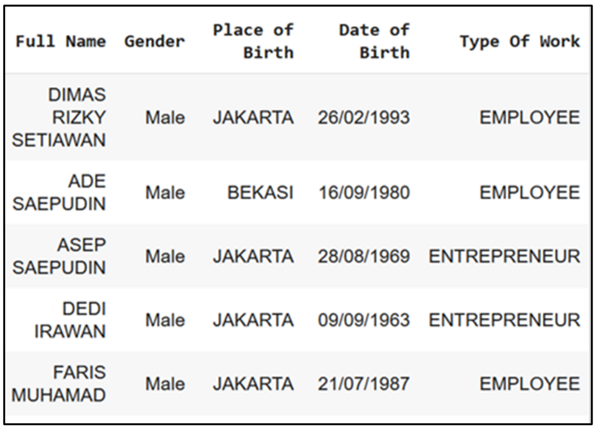
  At this stage, researchers reduced the number of data of the dataset by eliminating the attributes of village names and places of birth because they are not determining factor of whether they are feasible or unfit in receiving Covid-19 social assistance. The results of the data reduction process are presented in Table 2.

Table 2. Data Attributes reduction result

No	Attribute Name
1	Idpusdatin
2	Full Address
3	RW
4	RT
5	Siak_NO KK
6	Siak_NIK
7	Nama_LNGKP
8	Gender
9	Date of Birth
10	KET

- Data Transformation

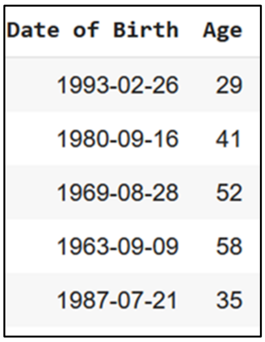
  Data transformation is a process by which changes to data of a certain format are made [15]. This process aims to make the data more suitable for the classification stage. The transformation data is carried out to convert the *KET* and *Nama_LNGKP* into the type of work and the *Full Name* is presented in Figure 3.



Full Name	Gender	Place of Birth	Date of Birth	Type Of Work
DIMAS RIZKY SETIAWAN	Male	JAKARTA	26/02/1993	EMPLOYEE
ADE SAEPUDIN	Male	BEKASI	16/09/1980	EMPLOYEE
ASEP SAEPUDIN	Male	JAKARTA	28/08/1969	ENTREPRENEUR
DEDI IRAWAN	Male	JAKARTA	09/09/1963	ENTREPRENEUR
FARIS MUHAMAD	Male	JAKARTA	21/07/1987	EMPLOYEE

Figure 3. Screenschot of transformation of *KET* and *Nama_LNGKP* attributes.

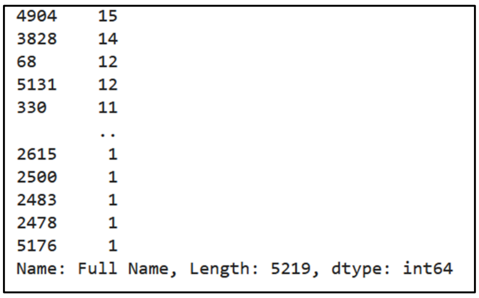
Then, transform the *Date of Birth* in the dataset into the *Age.* The results of the process are presented in Figure 4.



Date of Birth	Age
1993-02-26	29
1980-09-16	41
1969-08-28	52
1963-09-09	58
1987-07-21	35

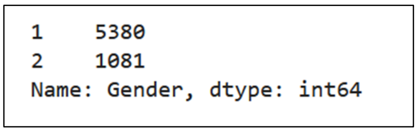
Figure 4. Screenshot of *Age* attribute

After that, changing the data type that originated the object to integer, the process of changing the *Full Name* is presented in Figure 5, the change in *Gender* is presented in Figure 6, and the change in the *Type of Work* is presented in Figure 7.



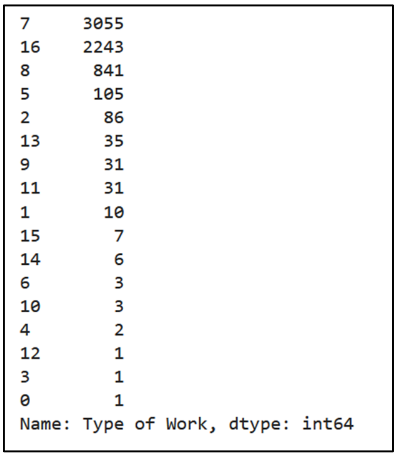
```
4904    15
3828    14
68      12
5131    12
330     11
        ..
2615     1
2500     1
2483     1
2478     1
5176     1
Name: Full Name, Length: 5219, dtype: int64
```

Figure 5. Full Name Transformation



```
1    5380
2    1081
Name: Gender, dtype: int64
```

Figure 6. Gender Transformation



```
7     3055
16    2243
8      841
5      105
2       86
13      35
9       31
11      31
1       10
15       7
14       6
6        3
10       3
4        2
12       1
3        1
0        1
Name: Type of Work, dtype: int64
```

Figure 7. Job Type Transformation

Upon completion of data preparation steps, dataset with the appropriate format is obtained for the classification experimentation purpose. The new attributes with clean data are presented in Table 3.

Table 3. Attributes of the clean dataset

NO	Dataset	
	Attribute Name	*Data Type*
1	Idpusdatin	Object
2	Full Name	Int
3	Gender	Int
4	Full Address	Object
5	Date of Birth	Datetime
6	Age	Int
7	Siak_No_KK	Int
8	Siak_NIK	Int
9	Type of Work	Int
10	RT	Int
11	RW	Int
12	Information	Int

3.1. Results for Naïve Bayes Classifier

The process carried out on the Naïve Bayes classifier is that the entered data will be assigned a label or class. Then from the class will be calculated the probability of each class. Then, the class results will be compared, if the probability value of each class on the label 0 (feasible) is greater than 1 (not feasible) then the data is classified as worthy of receiving Covid-19 social assistance and if vice versa, then the data is classified as unfit to receive Covid-19 social assistance. The experimental results for Naïve Bayes classifier with a composition of 80% training data and 20% testing data are presented in Table 4.

Table 4. Naïve Bayes Classification Result

Label	Metric		
	Accuracy	**Precision**	**Recall**
0	93%	86%	100%
1	93%	100%	87%

Figure 8 shows the confusion matrix of the Naïve Bayes classifier.

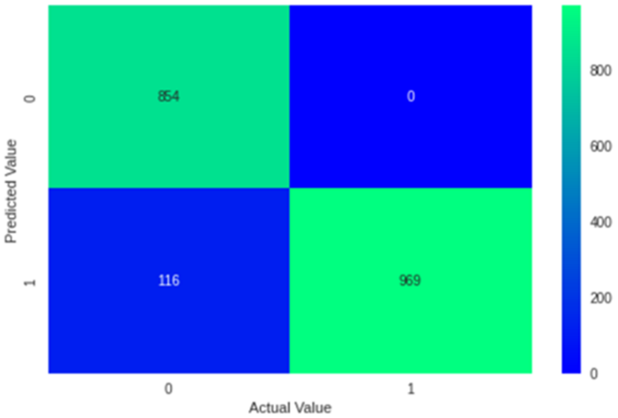


Figure 8. Confusion Matrix for Naïve Bayes classifier

From Figure 8, we obtain the value of TP is 854, TN is 969, FP is 0 and FN is 116. Five experiments on 5-fold cross validation are also carried out and the results are shown in Table 5.

Table 5. k-Fold cross validation result for Naïve Bayes classifier

Experiment	K-Fold Naïve Bayes	
	k-Value	Accuracy
1	5	94%
2	10	94%
3	15	94%
4	20	90%
5	25	94%

The results showed that for k values = 5, 10, 15, and 25, k-fold validation got the same accuracy result of 94% while for k value = 20 got an accuracy of 90%.

3.2. Results for KNN Classifier

The research conducted on the K-NN classifier is that the dataset used will be used as learning data where there are labels 0 (feasible) and 1 (not feasible). Then determine the nearest neighbor. Having done the calculation of K value, then the distance of each item in the training data will be calculated. Next, the closest distance to the specified K is seen, and then group the testing data based on the majority label on K. Results from experiment for the K-NN classifier on the 80% to 20% composition of training data and testing data are presented in Table 6.

Table 6. K-NN Classification Result

Label	Metric			
	K	Accuracy	Precision	Recall
0	10	90%	82%	98%
1	10	90%	98%	83%

Figure 9 shows the confusion matrix of the K-NN classifier.

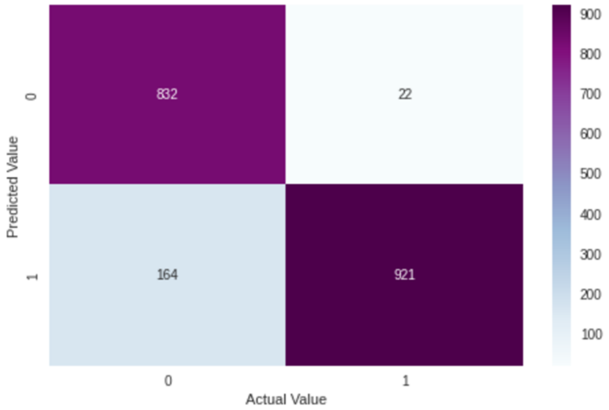


Figure 9. Confusion Matrix for K-NN classifier

From Figure 8, we obtain the value of TP is 832, TN is 921, FP is 22 and FN is 164. Five experiments on 5-fold cross validation are also carried out and the results are shown in Table 6.

Table 7. k-Fold cross validation result for K-NN classifier

Experiment	K-Fold K-NN	
	K-Value	Accuracy
1	5	89%
2	10	90%
3	15	88%
4	20	89%
5	25	89%

The results showed that k-fold validation with a value of k = 10 was the best with an accuracy of 90%, the value of k = 5, 20 and 25 achieved an accuracy of 89% while the value of k = 15 gave an accuracy of 88%.

3.3. Discussion

The experimental results of the Naïve Bayes classifier and K-NN classifier on the accuracy of classification using the confusion matrix and cross validation showed that the Naïve Bayes classifier has a 3% higher accuracy value than K-NN classifier. The Naïve Bayes classifier produces an accuracy of 93% while the accuracy value for K-NN classifier is 90%. Overall comparison of the performance results of Naïve Bayes and K-NN classifiers can be seen in Table 8.

Table 8. Performance comparison between Naïve Bayes and K-NN classifiers

Metric	Classifier	
	K-NN	Naïve Bayes
Accuracy	90%	93%
Precision	98%	100%
Recall	82%	85%

The Naïve Bayes algorithm is superior to the K-NN classifier because the use of the conditional probability of input variables owned by Naïve Bayes makes it performs better than K-NN. The independent data also made Naïve Bayes run well in this testing phase.

4. Conclusion

This study conducted model testing by comparing two methods, namely, the K-Nearest Neighbors algorithm and the Naïve Bayes algorithm for data on Covid-19 social aid recipients in Sukabumi Selatan Village. Then the classification results are compared to find out which algorithm is the best in determining the recipients of the Covid-19 social aid program. To measure the performance of the two algorithms, cross validation and confusion matrix testing methods are used. It was found that the Naïve Bayes algorithm had better results than the K-NN algorithm with an accuracy value of 93%, 100% precision and 85% recall. The test used the best cross validation with values k=5,10,15, and 25 resulted in the same accuracy value of 94%, while the K-NN algorithm produced an accuracy value of 90%, precision of 98% and recall of 82%. Testing using cross validation resulted in the best accuracy value of 10 k-fold of 90%. Thus, the two algorithms are equally good for use in classifying Covid-19 social aid recipients judging from the results of the tests carried out because the difference in values from their accuracy is not too significant.

Acknowledgement

Authors thanks to Research & Technology Directorate, Universitas Mercu Buana for partially support on this research, Grant no: 02-5/1337/B-SPK/V/2022.

References

[1] O. Naeni and R. A. Sari. *Comparison Of Data Mining Methods For Recipient Prediction Poor Student Assistance (BSM) In MAN 2 North Lampung*, 3rd Int. Conf. Inf. Technol. Bus., no. 7th Dec 2017, 207–213 (2017).

[2] E. Firasari, N. Khasanah, U. Khultsum, D. N. Kholifah, R. Komarudin, and W. Widyastuty. *Comparation of K-Nearest Neighboor (K-NN) and Naive Bayes Algorithm for the Classification of the Poor in Recipients of Social Assistance*, J. Phys. Conf. Ser. **1641**(1) (2020). doi: 10.1088/1742-6596/1641/1/012077.

[3] Y. F. Safri, R. Arifudin, and M. A. Muslim. *K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor,* Sci. J. Informatics. **5**(1), page 18 (2018). doi: 10.15294/sji.v5i1.12057.

[4] F. Tempola, R. Rosihan, and R. Adawiyah. *Holdout Validation for Comparison Classfication Naïve Bayes and KNN of Recipient Kartu Indonesia Pintar,* IOP Conf. Ser. Mater. Sci. Eng. **1125**(1), page 012041 (2021). doi: 10.1088/1757-899x/1125/1/012041.

[5] C. Schröer, F. Kruse, and J. M. Gómez. *A systematic literature review on applying CRISP-DM process model,* Procedia Comput. Sci. **181**(2019), 526–534 (2021). doi: 10.1016/j.procs.2021.01.199.

[6] T. Darmawan. *Credit Classification Using CRISP-DM Method On Bank ABC Customers,* Int. J. Emerg. Trends Eng. Res. **8**(6), 2375–2380 (2020). doi: 10.30534/ijeter/2020/28862020.

[7] H. Mousa and A. Maghari. *School Students' Performance Predication Using Data Mining Classification*, Int. J. Adv. Res. Comput. Commun. Eng. **6**(8), 136–141(2017). doi: 10.17148/IJARCCE.2017.6824.

[8] R. Irawan and A. E. Setiawan. *Comparison of Data Mining Classification Methods for Predicting Credit Appropriation through Naïve Bayes and Decision Tree Methods.* 10, 294–301 (2020).

[9] A. I. Sasmito and Y. Ruldeviyani. *Comparison of the Classification Data Mining Methods to Identify Civil Servants in Indonesian Social Insurance Company*, 3rd Int. Semin. Res. Inf. Technol. Intell. Syst. ISRITI. **2020**, 111–116 (2020). doi: 10.1109/ISRITI51436.2020.9315444.

[10] N. Dengen. *Comparison Performance of C4.5, Naïve Bayes and K-Nearest Neighbor,* Int. Conf. Sci. Inf. Technol.112–117 (2019).

[11] M. H. Effendy, D. Anggraeni, Y. S. Dewi, and A. F. Hadi. *Classification of Bank Deposit Using Naïve Bayes Classifier (NBC) and K –Nearest Neighbor (K -NN),* Proc. Int. Conf. Math. Geom. Stat. Comput. (IC-MaGeStiC 2021). **96**, 163–166 (2022). doi: 10.2991/acsr.k.220202.031.

[12] K. Kusrini, E. T. Luthfi, M. Muqorobin, and R. W. Abdullah. *Comparison of naive bayes and K-NN method on tuition fee payment overdue prediction,* 4th Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2019. **6**, 125–130 (2019). doi: 10.1109/ICITISEE48480.2019.9003782.

[13] M. Sadikin and F. Alfiandi. *Comparative Study of Classification Method on Customer Candidate Data to Predict its Potential Risk*, Int. J. Electr. Comput. Eng. **8**(6), page 4763 (2018). doi: 10.11591/ijece.v8i6.pp4763-4771.

[14] H. Elmunsyah, R. Mu'awanah, T. Widiyaningtyas, I. A. E. Zaeni, and F. A. Dwiyanto. *Classification of Employee Mental Health Disorder Treatment with K-Nearest Neighbor*

Algorithm, ICEEIE 2019 - Int. Conf. Electr. Electron. Inf. Eng. Emerg. Innov. Technol. Sustain. Futur. 211–215 (2019). doi: 10.1109/ICEEIE47180.2019.8981418.

[15] M. Imron and S. A. Kusumah. *Application of Data Mining Classification Method for Student Graduation Prediction Using K-Nearest Neighbor (K-NN) Algorithm*, IJIIS Int. J. Informatics Inf. Syst. **1**(1), 1–8 (2018). doi: 10.47738/ijiis.v1i1.17.