OPEN ACCESS
Indo-JC
Indonesia Journal of Computing
ISSN 2460-9056
socj.telkomuniversity.ac.id/indojc

Ind. Journal on Computing
Vol. 6, Issue. 3, December 2021. pp. 25-34
doi:10.34818/indojc.2021.6.3.593

Comparative Analysis of K-Nearest Neighbor and Modified K-Nearest Neighbor Algorithm for Financial Well-Being Data Classification

Ichwanul Muslim Karo Karo [#1], Athallah Tsany Rakha Dzaky [*2], Muhammad Alfhi Saputra [#3]

School of Computing, Telkom University
Bandung, Indonesia

[1] ichwanulkarokaro@telkomuniversity.ac.id
[3] rakhadzaky9@gmail.com
alfhi24@gmail.com

Abstract

Financial Well-Being is the condition that a person has been able to meet current and future financial obligations. There are many parameters in determining people who have obtained financial well-being. Classification is a data mining function that can identifies someone into financial well-being class. One of the most popular classification algorithms is K Nearest Neighbor (KNN). However, there is also a Modified K Nearest Neighbor (MKNN) classification algorithm which is a extend KNN. In this paper we will analyze comparison of KNN and MKNN algorithms to classify financial well-being datasets. Comparative analysis is based on accuracy and running time of both algorithms. Prior to the classification process, k-Fold Cross Validation was performed to find the optimal data modeling. The comparative analysis is based on the accuracy, precision, recall and execution time of the two algorithms. To obtain optimal classification results, the KNN and MKNN algorithms are run with K-Fold cross validation. Based on the test results, the performance of the KNN algorithm is higher than that of the MKNN in all test parameters, with an average gap of 25 percent. In addition, it was also found that the execution time of the KNN algorithm was faster than that of the MKNN.

Keywords: K Nearest Neighbor, Modified K Nearest Neighbor, Financial Well-Being

Abstrak

Financial Well-Being adalah kondisi seseorang memperoleh kemapanan keuangan kini dan diwaktu yang akan datang. Terdapat banyak parameter dalam menentukan orang yang sudah memperoleh financial well-being. Klasifikasi adalah satu task data Data mining yang dapat digunakan untuk mengidentifikasi sesorang dengan ketercapaian *financial well-being*. Salah satu algoritma klasifikasi yang paling populer adalah *K Nearest Neighbor* (KNN). Namun ada juga algoritma klasifikasi *Modified K Nearest Neighbor* (MKNN) yang merupakan modifikasi KNN. Dalam makalah ini kami akan menganalisis perbandingan algoritma KNN dan MKNN untuk mengklasifikasikan financial well-being dataset. Analisis komparatif didasarkan pada keakuratan, tingkat presisi, *recall, F-1 score* dan waktu eksekusi. kedua algoritma. Untuk memperoleh hasil klasifikasi yang optimal, algoritma KNN dan MKNN di lakukan dengan *k-Fold cross validation*. Berdasarkan hasil tes tersebut, performansi algoritma KNN lebih tinggi dibandingkan dengan MKNN di semua parameter pengujian, dengan rata rata gap 25 persen. Selain itu didapati juga waktu eksekusi algoritma KNN lebih cepat dibandingkan dengan MKNN.

Kata Kunci: K Nearest Neighbor, Modified K Nearest Neighbor, Financial Well-Being

I. Introduction

Received on October, 2021. Accepted on December, 2021

FINANCIAL Well-Being is the condition that a person has been able to meet current and future financial obligations [1]. Financial Well-Being is an abstract concept used to describe the financial situation of an individual or family. This means that the parameters used in defining financial well-being are not very clear. The Consumer Financial Protection Bureau (CFPB) is an organization that defines financial well-being with several parameters [2]. There are at least 215 parameters used to determine one's financial well-being whether it has been established or not.

One of the data mining tasks that identify objects is classification. Classification is the process of identifying objects by predicting labels/classes based on attributes or features without previously unknown label [3]. Task classification is very useful for categorizing objects, in order to obtain a model that is able to distinguish objects specifically. Classification algorithms have been implemented in the financial sector, such as identifying potential consumers [4], fraud detection on credit cards [5], anomaly detection on withdrawals [6], et cetera.

One of popular classification methods is KNN from Linear regression, Logistic regression, Decision tree, SVM algorithm, Naive Bayes algorithm, Random Forest algorithm, neural network, Gradient boosting algorithm and AdaBoosting algorithm [7]. KNN is known to be very simple and easy to implement. The idea of the KNN algorithm is to label new objects based on their similarity to surrounding objects. The method commonly used to measure the similarity between objects in KNN is Euclidean distance. Many classification cases used KNN algorithm because it is good at handling noise, simple, uncomplicated, easy and cheap computerization [7,8].

Parvin, et.al proposed Modified K-Nearest Neighbor (MKNN) to increase the accuracy of KNN [7]. The idea is to calculate the multiplication of the validity of the data by the weights used. Validity is used to test the validity of the train data, and the test is based on the number of neighbors on all samples of the train data. Voting weight is used to determine the highest weight of the multiplication calculation between validity and new sample data (testing data) to determine the final prediction class. Previously, KNN only calculated the neighbor distance between training data and test data, and determined the environment based on the number of K. However, MKNN will carry out a validity process in the training data before calculating the training data with test data, then the highest weight. voting from the nearest neighbor based on the number of K or neighbors will be calculated from the results of the multiplication of training data and test data. They tested the MKNN algorithm on 9 datasets by using K =3,5 and 7. Although the MKNN algorithm is claimed to be better than KNN, in fact we found from the paper that the MKNN algorithm has lower accuracy than KNN for the Bupa dataset at some amount of K. In addition, this study does not discuss the execution time aspect of the algorithm.

Based on the condition that not all datasets and the number of K on MKNN produce better performance than KNN, this study tries to re-analyze the comparison of KNN and MKNN by using a different dataset than before. Both algorithms run to identify a person's financial well-being. The dataset is obtained from the CFPB website [2], through the normalization process, the classification process and the last is the evaluation process. The results of this study are structured as follows: section II we describe related work by comparison of both algorithms and the concept of financial well-being. Section III describes the research methodology, such as the data pre-processing and classification stages. Section IV then presents the results obtained. Finally, conclusions and further work are detailed in section V.

II. LITERATURE REVIEW

A. Related Work

The initial study comparing the MKNN and KNN algorithms was proposed by [8]. They proposed the MKNN algorithm to improve the shortcomings of the KNN algorithm. The fundamental idea is to classify test samples depending on their neighbors. This technique is a form of weighting KNN using a certain procedure. The procedure is to calculate the fraction of the same false classification results to the total of various neighbors.

They used composition of the training and test data used is 90: 10 on five dataset; Isodata, Wine, Bupa and two variant monk datasets to run MKNN algorithm. Generally, the MKNN algorithm can increase 1-6 percent the accuracy of the model. However, we found that there is a dataset of five datasets where the classification results using MKNN are not better than KNN. MKNN's performance is not better in classifying Bupa datasets. Besides, in this study, there is no comprehensive analysis of the aspects of running time or computerization of the MKNN algorithm. So, it is very possible that the accuracy of the MKNN algorithm is better, but the execution time takes longer.

Another analysis comparing the KNN and MKNN algorithms was run to classify the data of Conditional Cash Transfer Implementation Unit (Unit Pelaksana Program Keluarga Harapan) [9]. The dataset consists of 7395 records. Comparative analysis is based on the accuracy of both algorithms. Before classification, k-Fold Cross Validation was done to search for the optimal data modeling resulted in data modeling on cross 2 with accuracy of 93.945 percent. The results of K-Fold Cross Validation modeling will be the model for training data samples and testing data to test KNN and MKNN for classification. Classification result produced accuracy based on the rules of confusion matrix. The test resulted in the highest accuracy of KKN by 94.95 percent with average accuracy during the test was 93.94 percent and the highest accuracy of MKNN was 99.5percent with the average accuracy during the test was 99.2 percent, almost all testing from the first test up to the tenth, MKNN algorithm is superior and has better accuracy value than KNN so it can be analyzed that the ability of MKNN algorithm in accuracy is better than KNN. Although the accuracy of MKNN is better than KNN algorithm, this research does not have a comprehensive analysis from the aspect of computerization and execution time of the MKNN algorithm.

With another dataset, a comparative analyzes the performance of the KNN and MKNN algorithms in classifying the UKT (tuition fee) of a new student university [10]. In their research, they used the MKNN method and supported by the k-Fold Cross Validation method. The results define that the best composition of k-Fold cross validation on the research is 80: 20. On the paper, the accuracy of MKNN algorithm is just 1 percent better than the KNN algorithm. However, they did not analyze the time consumption of MKNN in predicting UKT for new student universities. If the time consumption of MKNN is much greater than KNN, while the performance only increases by 1 percent than KNN. We think that the MKNN algorithm is not the right choice to identify the UKT (tuition fee) of a new student university. In other words, it is very possible that the MKNN algorithm is not over outcome KNN when viewed from a more comprehensive process aspect.

B. Financial Well-Being

Financial Well-Being is a condition where a person has been able to fulfill current and future financial obligations, is prepared to meet financial needs in the future, and is able to make choices that can be enjoyed in their life [2]. The financial well-being community is a society that can manage financially well, could develop assets well and is able to maintain financial resilience and is satisfied with the financial or financial condition that is owned. Financial Well-Being is an abstract concept used to explain the financial situation of an individual or family. So, there are no clear parameters in defining financial well-being.

The Consumer Financial Protection Bureau (CFPB) is a 21st century agency that helps consumer finance markets work by making rules more effective, by consistently and fairly enforcing those rules, and by empowering consumers to take more control over their economic lives. They define a financial well-being society with several parameters [2]. There are at least 215 parameters used to determine one's financial well-being whether it has been established or not.

III. Research Method

In this section, we will explain the research framework. The process is shown in Fig. 1. The research begins with collecting Financial Well-Being datasets from [2]. After obtaining the dataset, the next process is to normalize the data. The dataset normalization process is carried out because there are differences in the scalability value of each attribute of the Financial Well-Being dataset. Next stage is cross validation, in parallel,

the next process classifies the dataset using MKNN and KNN. Last of the process is evaluation and compared the performance of both algorithms.

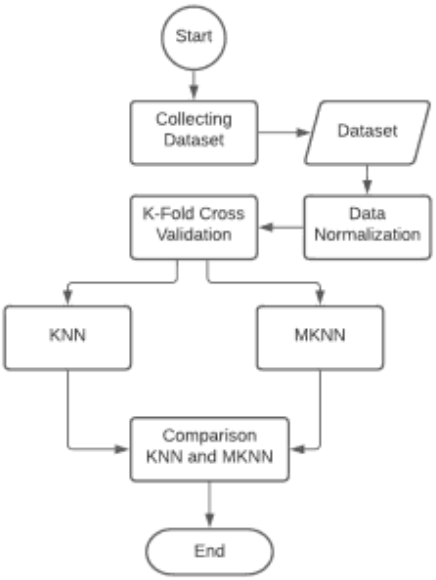

Fig. 1 Framework Research

A. Dataset

The dataset used in this study is the 2017 financial well-being dataset, which is obtained from [2]. The dataset consists of 6,394 records and 215 attributes to measure one's financial well-being. The explanation of financial well-being parameters can be simplified into 4 categories (Table 1). The next phase is that the data are normalized using the Min-Max scaler method by following the guidelines by [11].

TABLE I
DESCRIPTION OF VARIABLE

Categories	Description
Objective (qualitative)	A person's economic status such as income, debt, net worth, and household wealth.
Subjective (quantitative)	Driving force for financial behavior
Financial satisfaction	Could measure through several items such as satisfaction with income, spend of income for vacation, savings amount, emergency fund amount
Financial behavior	Cash, credit, and debt management behavior, planning various life activities

B. K-Fold Cross Validation

k-Fold Cross validation is used to find the best parameters from a model. In cross validation, the dataset is divided into k samples of the same size. Number of sample $k - 1$ will be used as data training and rest as data testing. An illustration of k-Fold cross validation shown on Fig. 2. However, there is no exact composition to determine how many k-folds are needed, perhaps 33%, 50%, and 60% of the dataset [8][12]. Therefore, in this research, the financial well-being dataset will be tested in 3 samples. Two samples as training data, and a sample as testing data. Cross validation run 10 times test to determine the value of K parameter and a good model to use in the algorithm test. Next, the training and testing process will calculate the average error (error mean).

Each run will find errors for testing the data, the model that gives the smallest average error is chosen to be the best method. In cross-validation, it has uncertainty in determining the cross-test [12][13].

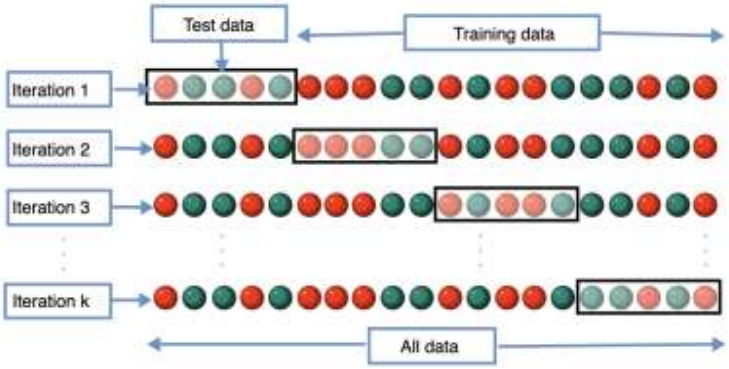

Fig. 2 K-Fold Cross Validation Illustration [12]

C. KNN Algorithm

K-Nearest Neighbor (KNN) algorithm is a supervised algorithm that uses geometer distance to classify objects [14], the algorithm shown on Fig. 3. The idea of algorithm is calculated similarity between objects and grouped into highest similarity. The final state of K-NN is by finding the k group of objects.

Algorithm 1: Basic KNN algorithm

Input: Training samples *D*, test sample *d*, *K*
Output: Class label of test sample
1: Compute the distance between *d* and every sample in *D*
2: Choose the *K* samples in *D* that are nearest to d; denote the set by *P* ($\in D$)
3: Assign *d* the class that is the most frequent class (or the majority class)

Fig. 3 KNN Algorithm

D. MKNN Algorithm

The MKNN algorithm was proposed by [8]. MKNN is a refinement of KNN for optimizing classification results. The fundamental difference between the MKNN and KNN algorithms is that the MKNN algorithm uses data training to classify the model and uses test data to test the model [8]. The MKNN algorithm is modified from the KNN algorithm by adding several steps. Here are the stages of the process.

1) Validity
The KNN algorithm does not go through the training data validation process, while the MKNN algorithm goes through that process. Validation of each object is measured based on the nearest k objects. The validity of the training data serves to determine the number of objects with the same table for all data in the training data. In this process all function values in *S* depends on the nearest neighbor being equal or not on the training data. The data validity equation is shown in equation (1)[10]. *K* is number of closest objects, *label(x)* is class label *x* and *Ni(x)* = nearest point class label *x*, meanwhile *S* = 1 when class is the same or worth 0 when the class is not the same.

$$Validity\ (x)\ = \frac{1}{K}\sum_{i=1}^{K} S(label(x))(N_i(x)) \quad (1)$$

2) Weight Voting
Determining weight voting is by using the validity of each data in the training data multiplied by the weight based on Euclidean distance. Weight voting is useful on training data with high validity and the closest distance to the test data. In MKNN method, weight vote calculation of each neighbor is in Equation (2) [10]. $W\ (x)$ is weighting of object *x*, d_e represent distance training data and test data, a is alpha value.

$$W\ (x)\ = validity(x)\frac{1}{d_e + a} \tag{2}$$

E. Evaluation

The last stage is to evaluate the classification results using the KNN and MKNN algorithms. The evaluation tool used is the Confusion matrix (Table II). The tool evaluates the classification model by measuring the precision, recall and F-1 measure of the true or false object. In addition, the performance of the two algorithms will be compared, which includes precision, recall, F-1 measure and running time. In this study, we followed the confusion matrix guide that has been used by [9][14].

TABLE II
CONFUSION MATRIX

		Predict	
		Positive	**Negative**
Actual	**Positive**	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True Negative (TN)

Precision (P) is the ratio of accurately predicted positive observations to total predicted positive observations which are shown in the Equation 3.

$$P = \frac{TP}{TP + FP}\ .100\% \tag{3}$$

Recall is the ratio of accurately predicted positive observations to the number of all relevant samples. which is shown in the Equation 4.

$$Recall(R) = \frac{TP}{TP + FN}\ .100\% \tag{4}$$

F1 measure, which is shown in Equation 5, is a method used to measure the performance of the model by combine precision and recall.

$$F1\ score = \frac{2PR}{P + R}\ .100\% \tag{5}$$

The percentage of correctly classified instances is called accuracy, which is shown in Equation 6.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}\ .100\% \tag{6}$$

IV. RESULTS AND DISCUSSION

A. *Preprocessing Data and K-Fold Cross Validation*

In the preprocessing stage, the normalization process uses a Min-Max scaler with the aim of standardizing the data. Next is the analysis cross validation process. In cross validation, the financial well-being dataset was tested in 3 samples. Two samples as training data, and 1 sample as testing data. Depend on several research, the best cross validation was conducted 10 times [12,13] to determine the value of the *k* parameter and a good model to be used in the algorithm test. The results of the validation test can be seen in Fig. 4. Based on the figure, the best *k* is 6 and 7 with an average accuracy of 42 percent. In addition, a good Cross is Cross1 with an average accuracy of 43 percent.

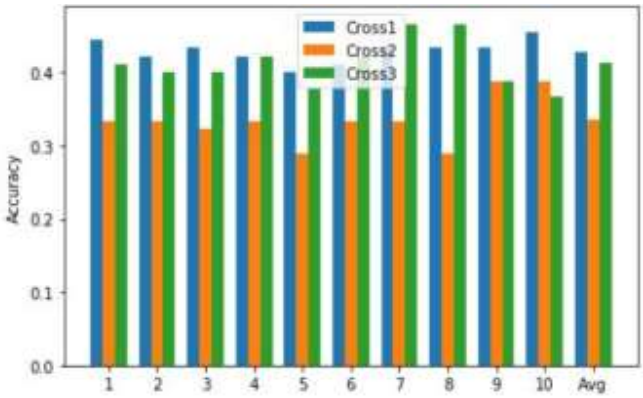

Fig. 4 Accuracy Measurement Results

B. *Result KNN Algorithm*

The implementation of the KNN algorithm is run with the same dataset, with *K* from 1 to 10. The results of the KNN algorithm can be seen in Figure 5. The best *K* is 5 with performance precision, recall and F-1 measure respectively 25, 42 and 32 percent. While the average performance of the KNN algorithm is 23.44, 42.2 and 30.1 percent for precision, recall and F-1 measure.

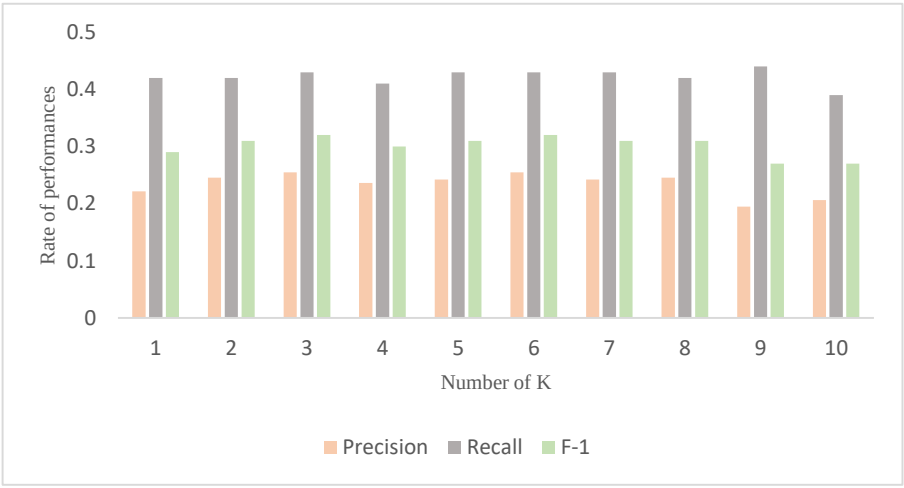

Fig. 5 Performance of KNN Algorithm

C. *Result MKNN Algorithm*

By using the same scenario as the previous experiment, the implementation of the MKNN algorithm is run with the same dataset, with *K* from 1 to 10. The results of the MKNN algorithm can be seen in Figure 6. The best *K* is 1 with performance precision, recall and F-1 measure respectively 10.56, 33 and 16 percent. While the average performance of the KNN algorithm is 7.45, 17.2 and 10.2 percent for precision, recall and F-1 measure.

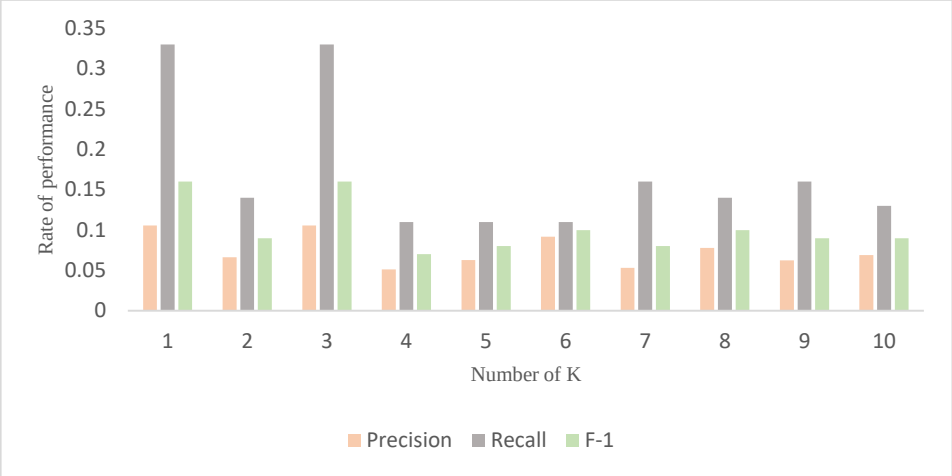

Fig. 6 Performance of MKNN Algorithm

D. *Comparative Analysis of the MKNN and KNN Algorithms*

The comparative analysis of the KNN and MKNN algorithms is reviewed depend on two aspects, that is the performance of the classification model and the running time. Performance of algorithm is seen based on precision, recall and F-1. Table III presents a detailed comparison of the performance of the KNN and MKNN algorithms based on the values of precision, recall and F-1. In this experiment, we found a significant performance gap between the KNN and MKNN algorithms (Fig. 7). Gap precision, recall and F-1 were obtained respectively 15.99, 25 and 19.9 percent. In other words, in this study the KNN algorithm is better than the MKNN algorithm. These results may not be in accordance with previous studies [7-10]. However, we see that the MKNN algorithm has over-outcome to identify the dataset which is the type of dataset that is just categorical or numerical [7-10]. In other words, MKNN is better at identifying single data types. We examine the performance of MKNN to classify Bupa dataset [8]. The Bupa KNN dataset consists of numerical, categorical, and Boolean data types with high scalability. Unfortunately, MKNN performance is lower than KNN in any number of K.

TABLE III
PERFORMANCE OF BOTH ALGORITHMS

K	KNN			MKNN		
	Precision	**Recall**	**F-1**	**Precision**	**Recall**	**F-1**
1	0.22	0.42	0.29	0.11	0.33	0.16
2	0.25	0.42	0.31	0.07	0.14	0.09
3	0.25	0.43	0.32	0.11	0.33	0.16
4	0.24	0.41	0.3	0.05	0.11	0.07
5	0.24	0.43	0.31	0.06	0.11	0.08
6	0.25	0.43	0.32	0.092	0.11	0.1
7	0.24	0.43	0.31	0.05	0.16	0.08
8	0.24	0.42	0.31	0.078	0.14	0.1
9	0.19	0.44	0.27	0.063	0.16	0.09
10	0.21	0.39	0.27	0.069	0.13	0.09

We see that there are similarities in characteristics between the financial well-being dataset and the Bupa dataset, where each dataset variable consists of more than one data type. In addition, between one variable and another, the scalability differs greatly.

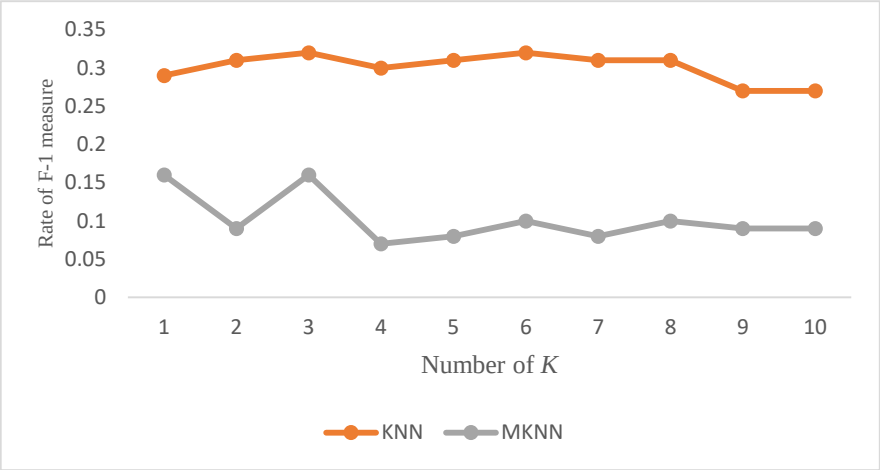

Fig. 7 Comparison of Performance MKNN and KNN Algorithm

In the second aspect, we analyzed the running times of both algorithms in each *K* (Shown in Fig. 8). Based on the figure, the running time of both algorithms is linear with respect to the number of *K*. In other words, if the number of neighbors being tested increases, the execution time of both algorithms will be longer, and vice versa. Running time analysis did not conducted in previous studies [7-10] did not do this. So, this analysis becomes important to know the computerization of the algorithm.

Comparison of the running time of the two algorithms is presented in Figure 8. The execution time of the MKNN algorithm for classifying financial well-being datasets is in the range of 20-25 minutes, while the execution time of the KNN algorithm for classifying financial well-being datasets is in the range of 5-10 minutes. Thus, the cost required by MKNN is doubled compared to KNN. In other words, the MKNN algorithm has a much larger time consumption than KNN. We found that the main cause of the swelling time consumption on MKNN was in two stages (validation and weight voting), where each data would be validated and given a weight. Imagine if there are 100 data and each stage takes 1 second, then it takes 200 seconds to complete the two phases. Thus it will be taken into consideration when using MKNN to classify some data with cheap cost.

Fig. 8 Comparison of Running Time Algorithm (Minute)

V. Conclusion

Based on the experiments, the best k-Fold cross validation from the financial well-being dataset resulted in 43 percent performance. The results of the classification of the financial well-being dataset using the two algorithms found results that were not good enough. The performance of both algorithms is below 40 percent. However, the average gap between the performance of the KNN algorithm and the MKNN algorithm is 25 percent adrift. In other words, the KNN algorithm is less than MKNN, because in this case dataset consist of several type of data large scalability between variables. In addition, the execution time of the KNN algorithm is also faster than the MKNN algorithm, because there are additional two stages (validation and weight voting). Lastly, we see that there are considerable research opportunities to obtain better performance.

References

[1] Brüggen, E. C., Hogreve, J., Holmlund, M., Kabadayi, S., & Löfgren, M. (2017). Financial well-being: A conceptualization and research agenda. Journal of Business Research, 79, 228-237.
[2] CFPB. (2015). Financial well-being; The goal of Financial Education. US: CFPB
[3] Alasadi, S. A., & Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining. Journal of Engineering and Applied Sciences, 12(16), 4102-4107.
[4] Elsner, P. (2018). Classification of consumer products under the EU CLP Regulation: what to consider when caring for contact dermatitis patients. Contact dermatitis, 78(1), 1-6.
[5] Itoo, F., & Singh, S. (2021). Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. International Journal of Information Technology, 13(4), 1503-1511.
[6] Mahdi, W. S., & Maolood, A. T. (2020). Banking Intrusion Detection Systems based on customers behavior using Machine Learning algorithms: Comprehensive study. Journal of Al-Qadisiyah for computer science and mathematics, 12(4), Page-1.
[7] Parvin, Hamid, Alizadeth, Hoseinali, and M. Behrouz. (2010). "A Modification on K-Nearest Neighbor Classifier". Global Journal of Computer Science and Technolgy. Vol. 10 No. 14, pp.37-41.
[8] Parvin, Hamid, Alizadeth, Hoseinali, Minati, M. Behrouz, and Bidgoli.(2008). "MKNN: Modified K-Nearest Neighbor". Proceedings of the World Congress on Engineering and Computer Science WCECS. ISBN: 978-988-98671-0-2 pp. 1-4.
[9] Gazalba, I., & Reza, N. G. I. (2017, November). Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification. In 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE) (pp. 294-298). IEEE.
[10] Adli, D., & Sahid, D. S. S. (2021). UKT (Single Tuition) Classification Prediction uses MKNN (K-Nearest Neighbor Modification) algorithm. International ABEC, 81-84.
[11] Pandey, A., & Jain, A. (2017). Comparative analysis of KNN algorithm using various normalization techniques. International Journal of Computer Network and Information Security, 11(11), 36.
[12] Wani, F. J., Rizvi, S. E. H., Sharma, M. K., & Bhat, M. I. J. (2018). A study on cross validation for model selection and estimation. International Journal of Agricultural Sciences, 14(1), 165-172.
[13] Wong, T. T., & Yeh, P. Y. (2019). Reliable accuracy estimates from k-fold cross validation. IEEE Transactions on Knowledge and Data Engineering, 32(8), 1586-1594..
[14] Karo, I. M. K., Ramdhani, R., Ramadhelza, A. W., & Aufa, B. Z. (2020, October). A Hybrid Classification Based on Machine Learning Classifiers to Predict Smart Indonesia Program. In 2020 Third International Conference on Vocational Education and Electrical Engineering (ICVEE) (pp. 1-5). IEEE.