Credit scoring analysis using pseudo nearest neighbor

H Pratiwi¹, M A Mukid¹, A Hoyyi¹ and T Widiharih¹
¹Department of Statistics, Diponegoro University
Jl. Prof. Soedharto, SH, Tembalang, Semarang 50275, Indonesia
E-mail: mamukid@live.undip.ac.id

Abstract. Credit scoring is one of the crucial task and a core responsibility for financial institutions in their risk management. This study aims to apply the pseudo nearest neighbour (PNN) method as a tool to identify which prospective borrowers are eligible for their loan proposals. If a new borrower has characteristics closer to a good historical borrower then the loan proposal is worthy to approval. But if not, the proposed loan will be refused. The historical data in this paper are credit data from a national bank in Indonesia. The characteristics of historical debtors consist of age, amount of a child, length time of business, income, loans amount, and the period of credit. The best classification of k-NN is using k =1, because it makes the smallest error 1.89%. While the best classification of PNN is using k = 13 with the smallest error 20.75%. Based on total accuracy of classification shows that the credit classification of debtors using k-NN is more appropriate than PNN.

1. Introduction
Credit is an important catalyst for economic growth. However, credit cannot be done carelessly. The desire of financial institutions to obtain substantial benefits from granting credit will increase the risk of default. Therefore, credit risk must be managed properly so that the potential for default loans does not damage the liquidity of the financial institution. The provision of credit in the form of business loans and other loans to prospective borrowers is carried out by passing the credit application process and the analysis of the credit disbursement submitted. The analysis used by the banking sector is 5C’s namely Character, Capacity, Capital, Collateral, and Conditional of Economics [1]. By using 5C’s analysis, the ability of prospective borrowers will be known in paying off their credit. Although it has gone through the process of analyzing the provision of credit to credit, it is often found to have problem loans or often called Non-Performing Loans (NPLs) which are substandard, doubtful and loss. To be able to find out the ability of debtors to repay credit, it is necessary to examine the factors that distinguish the ability of debtors. So that these variables can be used to classify prospective debtors. A good classification method is a method that produces small errors [2]. Currently, there are many statistical methods used to classify, and two of the existing classification methods are the k-Nearest Neighbor (k-NN) and Pseudo Nearest Neighbor (PNN) methods.

According to [3], the k-Nearest Neighbor (k-NN) is a method that classifies based on the proximity of a data with other data. The k-NN method is a fairly simple method but has a high degree of accuracy, but one of the problems faced by K-NN is the selection of the right k value. The way of voting for the majority of k-neighbors for large k values can result in large data distortions because each k-neighbor has the same weight for the test data, while k which is too small can cause the algorithm to be too sensitive to noise.
According to [4], the Pseudo Nearest Neighbor (PNN) method is an extension of the k-NN method which is enhanced in the accuracy of classification in a small data. The PNN method basically uses a combination of the closest weighting k-NN proposed and the closest local average method ([5], [6]).

Research relates to the comparison of the k-Nearest Neighbor (k-NN) method with other classification methods that have been done before. The study raised the problem of Comparative Classification Analysis Using the k-Nearest Neighbor (k-NN) Method and Multivariate Adaptive Regression Spline (MARS) on Public Primary School Accreditation Data in Semarang City [7]. The study classified public elementary school data based on school accreditation, namely A and B accreditation. Based on the results and analysis of the study, it was concluded that the MARS method was better than K-NN in classifying public elementary schools in Semarang based on their accreditation. It prompted the writer to compare the k-NN method with the other classification method, PNN. This research was conducted to find out which method is more appropriate in classifying debtor credit status which consists of credit status of good, special attention, and bad.

2. Literature review

2.1. Micro credit

Based on Law Number 7 of 1992 concerning banking, credit is the provision of money or equivalent claims, based on a loan agreement between banks and other parties requiring the borrower to repay the debt after a certain period of time with interest. According to [8], credit is the ability to carry out a purchase or make a purchase or make a loan with the promise of payment to be made or deferred on an agreed period.

According to the Coordinating Ministry of Economic Affairs of the Republic of Indonesia [9], people's business loans are working capital financing or investment loans to UMKMK in productive and feasible business fields but not yet having additional collateral or insufficient collateral. The purpose of the KUR program is to accelerate the development of access to finance to improve the empowerment of micro, small and medium enterprises with the business sector financed by the sectors of agriculture, fisheries, processing, trade and services [10].

2.2. K-Nearest Neighbor (k-NN)

According to [3], k-Nearest Neighbor (k-NN) is a method that classifies based on the proximity of a data with other data. The value of k on k-NN means the closest data from the training data. If k = 1, the class from one training data as the first closest neighbor will be given as the class label for the test data which is class 1. If k = 2, then 2 closest neighbors are taken from the training data, as well as k = 3, 5, 7 and so on. If there are two different classes in the neighboring class, a class with the majority rule will be taken. The choice of k value is determined by the researcher. Determination of this k value can affect the accuracy of the prediction done. Adding the value of k means that the distance of each test data to the nearest neighbor is increasing.

According to [11] and [12], the k-NN algorithm was formulated as follows:

1. Calculate the distance from each test data against all training data.
2. Sort the distance values from the smallest value to the largest value.
3. Determine the value of k.
4. Assign the most classes to appear from the nearest neighbor as the test data class

2.3. Pseudo Nearest Neighbor (PNN)

Based on the theory of the local mean learning method of PNN, let $y_j$ be the weighted sum of the distance of k-nearest neighbor of data $x$ in the $j$-th class, then $y_j$ is defined as follows:

$$y_j = w_1 \cdot d_{j}^{(1)} + w_2 \cdot d_{j}^{(2)} + \cdots + w_k \cdot d_{j}^{(k)}$$

where $d_{j}^{(i)}$ is an Euclidean distance of i-th nearest neighbor in the $j$-th class and $w_i = 1/i$ is the weight of i-th nearest neighbor. According to [4], the pseudo nearest neighbor algorithm with the parameter $k$, as follows:
1. calculate the distance between the \( k \)-nearest neighbors of training data \( d_i \) in the \( j \)-th class and the test data \( d_j \), where \( d_j = \| x_j - x \|_2 \).
2. get the arranged distances in increasing order, \( d_j^{(1)}, \ldots, d_j^{(k)} \).
3. calculate \( y_j = \sum_{i=1}^{k} w_i \cdot d_j^{(i)} \).
4. get the pseudo nearest neighbor \( y_{PNN} = \arg\min_j \{ y_j \} \), for \( j = 1, 2, \ldots, M \).

2.4. Evaluation criteria
According to [2], a good classification method will produce a slight misclassification. In this study to test the accuracy of classification, a confusion matrix is used. Confusion matrix is a classification table of work results. The following is a confusion matrix that classifies the three classes in table 1.

| \( f_{uv} \) | Class = 1 | Class = 2 | Class = 3 |
|-------------|-----------|-----------|-----------|
| Actual (u)  |
| Class = 1   |
| \( f_{11} \) | \( f_{12} \) | \( f_{13} \) |
| Class = 2   |
| \( f_{21} \) | \( f_{22} \) | \( f_{23} \) |
| Class = 3   |
| \( f_{31} \) | \( f_{32} \) | \( f_{33} \) |

Each \( f_{uv} \) cell in a matrix states the amount of data from class \( u \) whose prediction results into class \( v \). \( f_{11} \) is the amount of data in class 1 that is correctly classified in class 1, \( f_{12} \) is the amount of data in class 1 which is incorrectly classified into class 2, \( f_{13} \) is the amount of data in class 1 that is incorrectly classified in class 3, \( f_{21} \) is the amount of data in class 2 which is incorrectly classified into class 1, \( f_{22} \) is the amount of data in class 2 that is correctly classified into class 2, \( f_{23} \) is the amount of data in class 2 which is incorrectly classified into class 3, \( f_{31} \) is the amount of data in class 3 that is incorrectly classified into class 1, \( f_{32} \) is the amount of data in class 3 that is incorrectly classified into class 2, \( f_{33} \) is the amount of data in class 3 that is correctly classified to class 3. So based on table 1 above, the amount of data predicted correctly is \( f_{11} + f_{22} + f_{33} \) and the data is classified incorrectly, \( f_{12} + f_{13} + f_{21} + f_{23} + f_{31} + f_{32} \). The quantity of the confusion matrix is divided into two, namely accuracy and prediction error or Apparent Error Rate (APER) [3].

\[
\text{Accuracy} = \frac{(f_{11} + f_{22} + f_{33})}{(f_{11} + f_{12} + f_{13} + f_{21} + f_{22} + f_{23} + f_{31} + f_{32} + f_{33})}
\]

\[
\text{APER} = \frac{(f_{12} + f_{13} + f_{21} + f_{23} + f_{31} + f_{32})}{(f_{11} + f_{12} + f_{13} + f_{21} + f_{22} + f_{23} + f_{31} + f_{32} + f_{33})}
\]

3. Data and methods
The type of data used in this study is secondary data. The data is a loan for People’s Business Credit in a national bank from Wonogiri Regency, Central Java in 2016-2017. The data used in this study consist of 265 debtors. In this study, the data were divided into two parts, namely training data (training) and test data (testing) with a ratio of 80% and 20%. The distribution of training data (training) in this study is 80% as much as 212 data and test data 20% as much as 53 data. The variables used in this study consist six variables including age, number of children, long in business, income, installment, and credit period.

Before using KNN and PNN methods, the first step that must be done is standardizing data for each variable. Data standardization is done so that the difference in the scale of data on each variable whose size is different from each other, has the same weight in calculating the Euclid distance. Standardization is done using the Z score which subtracts the value of the data by the overall average and then divides by the standard deviation of each variable.

4. Result and Discussion
This study examines the application of the pseudo nearest neighbor method for credit scoring analysis with a case study of micro credit in Indonesia where the data come from a national bank in Wonogiri district, Central Java. We use 6 variables which are the characteristics of borrowers including age, number of children, length of business, income, installment amount and length of credit. Table 3 is a statistical summary of the variables used for credit scoring analysis. Based on table 2, it is known that age, length of business, income and installment amount of good category borrowers tend to be higher than other credit statuses. However, the number of children and credit duration is lower than that of borrowers with credit status in special attention, although it is still higher than the bad category borrowers.

**Table 2. Statistical summary of variables used for credit scoring**

| Variable            | Credit Status | Mean    | Std. Deviation | Median | Modus | Max | Min |
|---------------------|---------------|---------|----------------|--------|-------|-----|-----|
| age (year)          | Good          | 39,943  | 10,928         | 39     | 35.00 | 68.00 | 21  |
|                     | Special Attention | 38,680 | 11,960 | 35     | 34    | 64  | 19  |
|                     | Bad           | 35,103  | 9,420          | 34     | 25    | 54  | 21  |
| Number of Children  | Good          | 1,270   | 0.95           | 1      | 1     | 4   | 0   |
|                     | Special Attention | 1,640  | 1.846          | 1      | 1     | 9   | 0   |
|                     | Bad           | 0.828   | 0.759          | 1      | 1     | 2   | 0   |
| Long in Business    | Good          | 6,436   | 5,407          | 5      | 5     | 40  | 1   |
|                     | Special Attention | 5,960  | 7,613          | 4      | 3     | 40  | 1   |
|                     | Bad           | 3,724   | 1,771          | 4      | 5     | 8   | 1   |
| Income (Million Rupiah) | Good         | 1,413   | 8,964          | 1.2    | 1.2   | 8.25 | 0.1 |
|                     | Special Attention | 1,042  | 0.377          | 1.25   | 0.75  | 1.825 | 0.3 |
|                     | Bad           | 1,357   | 0.715          | 1.20   | 1.20  | 4.10 | 0.35|
| Installment Amount  | Good          | 18,687  | 7,077          | 20,000 | 25,000 | 25,000 | 3,000 |
| (Million Rupiah)    | Special Attention | 16,120  | 7,271          | 15,000 | 10,000 | 25,000 | 4,000 |
|                     | Bad           | 14,552  | 5,944          | 15,000 | 20,000 | 20,000 | 3,000 |
| Credit Period       | Good          | 27,953  | 8,171          | 24     | 36    | 36  | 12  |
| (Month)             | Special Attention | 30,240  | 6,119          | 36     | 36    | 36  | 24  |
|                     | Bad           | 25,655  | 8,913          | 24     | 24    | 36  | 12  |

In the K-NN method, the majority rule is used in decision making. Determination of the value of k in this study assisted with the Matlab program. By using training data, the KNN method error rate is calculated based on different k values where odd values are chosen. The error results for the experiment k value in k-NN are presented in table 3.

Table 3 is the classification error rate based on KNN using different k. It can be seen that at k = 1 produces the lowest classification error rate and the use of k more than 1 will tend to increase the level of misclassification. Using k = 1 the classification results of each test data can be seen in Table 4. There were 42 observations derived from good credit status which were predicted to be properly included in credit status, and there were no observations of credit status either classified as poor credit status or in special attention. Meanwhile, there were 4 observations of credit status in special attention which are predicted to be properly included in credit status in special attention. But there is an observation of credit status in the special attention predicted to enter good credit status and there is no
observation of credit status in the special attention predicted to enter bad credit status. Furthermore, all observations from bad credit status are predicted to be correctly included in the bad credit status.

**Table 3.** Error rate of KNN in some values of k

| k   | Error rate |
|-----|------------|
| 1   | 0.0189     |
| 3   | 0.075      |
| 5   | 0.0566     |
| 7   | 0.0755     |
| 9   | 0.0755     |
| 11  | 0.0943     |
| 13  | 0.0943     |
| 15  | 0.0943     |
| 17  | 0.0943     |
| 19  | 0.0943     |

**Table 4.** Classification result Using KNN with k = 1

| Actual class | Predicted class | Good | Special attention | Bad |
|--------------|----------------|------|-------------------|-----|
| Good         | Good           | 42   | 0                 | 0   |
| Special Attention | 1       | 4    | 0                 | 0   |
| Bad          | 0              | 0    | 6                 |     |

Based on Table 4, the APER value and accuracy of the KNN method can be calculated as follows:

\[
\text{APER} = \frac{(f_{12}+f_{13}+f_{21}+f_{23}+f_{31}+f_{32})}{(f_{31}+f_{32}+f_{33}+f_{21}+f_{22}+f_{23}+f_{13}+f_{12}+f_{33})} = \frac{1}{53} = 0.0189 = 1.89\% \\
\text{Accuracy} = 1 - \text{APER} = 1 - 0.0189 = 0.9811 = 98.11\% 
\]

By using the KNN method, overall 98.11% of micro credit can be classified properly.

Furthermore, the classification accuracy value of the PNN method will be calculated on different k values. Because PNN uses a local average, for some k values, in certain credit classes accuracy cannot be calculated because the observations are not enough. In this case we can only calculate the classification accuracy up to k = 21.

Table 5 is the classification error rate based on PNN using different k values. It can be seen that at k = 13 yields the lowest classification error rate and the use of k less than 13 will tend to increase the level of misclassification. Using k = 13 the classification results of each test data can be seen in Table 6. There were 42 observations of good credit which were predicted to be properly, and there were not observations of credit status either classified as poor credit status or in special attention. Meanwhile all debtors with special or bad status by the PNN method are classified as good debtors.

**Table 5.** Error rate of PNN in some values of k

| Value of k | Error rate |
|------------|------------|
| 1          | 0.3396     |
| 3          | 0.2642     |
Table 6. Classification Result Using PNN with $k = 13$

| Actual class  | Predicted class |         |
|---------------|-----------------|---------|
|               | Good            | Bad     |
| Good          | 42              | 0       |
| Special attention | 5          | 0       |
| Bad           | 6               | 0       |

Based on table 6, the APER value and accuracy of the PNN method can be calculated as follows:

$$\text{APER} = \frac{\sum_{i=1}^{11} \left( f_{i2} + f_{i3} + f_{21} + f_{22} + f_{31} + f_{32} \right)}{\sum_{i=1}^{15} \left( f_{i1} + f_{i2} + f_{i3} + f_{21} + f_{22} + f_{31} + f_{32} + f_{33} \right)}$$

$$= \frac{20.75}{325}$$

$$= 0.2075$$

$$= 20.75\%$$

$$\text{Accuracy} = 1 - \text{APER}$$

$$= 1 - 0.2075$$

$$= 0.7925$$

$$= 79.25\%$$

From the evaluation of the accuracy of classification for each method that uses APER calculations, it can be seen that the accuracy for the $k$-NN method is 98.11\% with $k = 1$ while for the PNN method is 79.25\% with $k = 13$. With the APER value, the data classification of the People's Business Credit in Bank X of Wonogiri Regency, Central Java should be done using the K-NN method rather than PNN.

5. Conclusion
Inevitably, the use of credit scoring models in managing credit risk is a must for banks and financial institutions. It is done solely so that the institution can survive in an increasingly competitive business world. This study examines the application of the pseudo nearest neighbor method as an alternative method in credit scoring modeling. Unfortunately in our case study, the PNN method has not shown better performance than its predecessor, KNN.

Acknowledgment
This research was supported by the research and technology and higher education ministry of Indonesia based on contract number 007/SP2H/LT/DRPM/IV/2017.

References
[1] Kasmir 2010 Dasar-Dasar Perbankan (Jakarta : PT. Raja Grafindo Persada).
[2] Johnson R A and Wichern D W 2007 Applied Multivariate Statistical Analysis. Sixth Edition (New Jersey: Prentice Hall International. Inc.).
[3] Prasetyo E 2012 Data Mining Konsep dan Aplikasi Menggunakan MATLAB (Yogyakarta: Penerbit ANDI Yogyakarta).
[4] Zeng Y, Yang Y and Zhao L 2008 Pseudo Nearest Neighbor Rule For Pattern Classification. Department of Automation (Shanghai Jiao Tong University: Elsevier Ltd.).
[5] Dudani S A 1976 IEEE Transaction on System, Man and Cybernetics 6 325.
[6] Mitani Y and Hamamoto Y 2006 Pattern Recognition Letters 27 1151.
[7] Merluarini B 2014 Perbandingan Analisis Klasifikasi Menggunakan Metode k-Nearest Neighbor (k-NN) dan Multivariate Adaptive Regression Spline (MARS) Pada data Akreditasi Sekolah Dasar Negeri Di Kota Semarang. Semarang.
[8] Mulyono T P 2007 Manajemen Perkreditan Bagi Perbankan Komersil Yogyakarta.
[9] [Kemenko] Kementrian Koordinator Bidang Perekonomian Republik Indonesia Kredit Usaha Rakyat. http://kur.ekon.go.id/ (download 25 November 2017).
[10] Askrindo 2016 Penjaminan KUR .https://askrindo.co.id/new/id/penjaminankur (download 30 November 2017)
[11] Wahbeh A H, Al-Radaideh Q A, Al-Kabiand M N and Al-Shawakfa E M 2011 International Journal of Advanced Computer Science and Application Special Issue on Artificial Intelligence (IJACSA) 35 18.
[12] Fix E and Hodges J L 1951 Discriminatory Analysis Nonparametric Discrimination. Consistency Properties Technical Report TR4 USAF School of Aviation Medicine, Randolph Field TX.