Analysis of classification algorithm in pension types

A S Fitrani¹, *¹, I R I Astutik² and M A Rosid³
Universitas Muhammadiyah Sidoarjo, Sidoarjo, Indonesia

*asfjim@umsida.ac.id

Abstract. The concept of data mining in its implementation is used in handling and analysing data in large capacity. Steps to extract information and to gain knowledge in supporting decision making. Data mining, in its implementation can be used in various sectors of data analysis, namely about the condition of the type of pension. The purpose of this study is to provide information, where a civil servant within the retirement limit is taken normally or earlier. The technique of using the classification method, with the decision tree C4.5 (J48), Naïve Bayes and k Nearest Neighbour algorithm. Presentation of assets of 1,316 pension data in 2012 and 2013 from Bukopin Bank customers, which are divided into training data by 65% and data testing by 35%. From the testing of three algorithms, the highest classification was produced, namely Naïve Bayes with 91%. The results of the three classifications indicate the quality of attribute determination can affect the results. And the predictive level of the three algorithms shows different results. For classification techniques with better results, improvements are needed to determine attributes and develop existing datasets.

1. Introduction
Retirement is a condition of someone who has stopped working because his term is finished. Retirement is caused by several factors ranging from age and early retirement. A retired person receives monthly allowances in accordance with the provisions of the company or agency. As for pension funds to manage and run the programs set by the pension fund regulations. The retirement age for civil servants is divided into three: 58 years old for administrative officials, 60 years old for high ranking officials, and 65 years for functional officials [1].

The types of pensions held for civil servants are normal, accelerated, postponed and disabled pensions. For a normal pension, the tenure is over 30 years, while the pension is accelerated under 30 years. At retirement postponement of work by reaching normal retirement age and disability pension is a termination of work that is unable to continue work. The concept of this study uses a type of normal and accelerated pension. Because the retirement data at the Bank Bukopin Surabaya Branch is a type of pension in the normal and accelerated category.

Pension is a condition of someone who is not allowed to work in a company or agency because the term of the assignment is complete [2]. Retirement is caused by several factors ranging from age and early retirement. Someone who has retired will get a monthly allowance in accordance with the provisions of the company or agency. As for pension funds to manage and run programs stipulated by pension fund regulations. Pensioners can receive a monthly pension by continuing to become productive workers through pension credit [3]. Thus, pension credit is used to finance the needs of the applicant regularly every month. The important role of this credit card is to make transactions easier for the company and a big concern in the company [3].
In the form of data mining in the collection and storage of data taken from the past based on the data warehouse. Where, the data can be used to find information on the method to be selected. The method learns to extract knowledge or find patterns from large data. Data mining will be processed through several stages of data, namely selection, pre-processing, transformation, data mining, and evaluation. Technical pre-processing in preparing data in the form of removing duplicate data, removing data inconsistencies, and correcting data errors. In the set of data can be divided into attributes, types of datasets, and public datasets. Thus, the use of data mining techniques in existing datasets will be calculated according to the algorithm chosen to achieve accurate levels in a maximum percentage value. Data mining techniques can be applied in various fields that are used to support a wide range of applications [4].

The dataset and several attributes found in the Surabaya branch of Bank Bukopin customers, which then the dataset is processed by a classification method to estimate the class of objects whose labels are unknown. This effort was made so that the Bank Bukopin Surabaya Branch could predict the decisions related to the type of pension in the future. Then can estimate prospective customers who are projected to be customers of Bank Bukopin Surabaya Branch.

2. Classification
Classification method that is studying a set of data that can be produced by a new classification of data. The classification process in data mining techniques is a set of data that can produce a classification model (target function). So, it requires a dataset on the set for the classification process. The dataset used is attributes and features using training data and testing data. Classification techniques can be grouped into two categories, namely the classification technique globally calculates all training data and classification locally taking into account some training data [5].

2.1. C4.5 algorithm
At the learning stage of the data, the C4.5 algorithm constructs decision trees from training data, in the form of cases or records (tuples) in the database. The three working principles of the C4.5 algorithm at the learning stage of the data are:

- Making a decision tree.
- The decision tree pruning and evaluation (optional).
- Making rules from decision trees (optional).

The formula for the C4.5 algorithm is:

\[ Gain(S, A) = Entropy(S) \sum_{i=1}^{n} \frac{S_i}{S} \times Entropy(S_i) \]  

Where:
S: set of cases
A: attribute
n: number of partition attributes A
Si: number of cases on the i-partition
S: number of cases in S

\[ Entropy(S) \sum_{i=1}^{n} \frac{S_i}{S} \times p_i \times log_2 p_i \]  

Where :
S: set of cases
2.2. Naïve bayes algorithm

Naïve Bayes algorithm is a classification method by calculating probabilities in determining the number of classes and values of a dataset. The advantage of using the Naïve Bayes algorithm is that the small amount of training data can determine the required parameter estimates. Assume a simplification of the Naïve Bayes algorithm with attribute values that are mutually independent if given their output values [6]. Naïve Bayes has a very strong level of accuracy and speed when applied to a database with large data. Based on the Bayes theorem which has the ability to classify the same method as the Decision Tree algorithm and Neural Network [7]. In the Bayes theorem equation, the conditional probability is expressed as:

$$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)}$$  \hspace{1cm} (3)

Where,
- $X$: Data with unknown classes.
- $H$: Data hypothesis with specific classes.
- $P(H)$: Probability of $H$. hypothesis
- $P(X)$: Probability $X$.
- $P(H|X)$: Posterior probability $H$ with condition $X$.
- $P(X|H)$: Posterior probability $X$ with the condition $H$.

2.3. K nearest neighbour classification

The K-Nearest Neighbour (NN) is the simplest method of machine learning. It is a type of instance base learning in which object is classified based on the closest training example in the feature space. It implicitly computes the decision boundary however it is also possible to compute the decision explicitly. So the computational complexity of K NN is the function of the boundary complexity [8].The k-NN algorithm is sensitive to the local structure of the data set. The special case when $k = 1$ is called the nearest neighbour algorithm. The best choice of $k$ depends upon the data set; larger values of $k$ reduce the effect of noise on the classification [9] but make boundaries between classes less distinct. The various heuristic techniques are used to select the optimal value of $K$. KNN has some strong consistent results. As the infinity approaches to data, the algorithm is guaranteed to yield an error rate less than the Bayes error rate [9].

3. Methods

The condition of the problem is based on data mining techniques to predict the type of customer retirement. Designing a classification diagram using diagrams:
3.1. Data collection
The dataset is used to predict the types of customer pensions from 2012 and 2013. With the dataset, it can be processed through the attributes of the pension type data based on the work period of the civil servants. So, it is processed through taking training data and testing data. Retrieval of retirement data attribute data used. Variable is a collection of variables consisting of entity parts. The variables used for retirement data are 11 attributes. Whereas, features are the contents of variables. Data that has been obtained from Bank Bukopin Surabaya Branch can be known variables and features as follows:

| No | Variable | Future |
|----|----------|--------|
| 1  | JK       | \{L,P\} |
| 2  | Umur     | Numeric |
| 3  | Agama    | \{Islam,Protestan,Katholik,Hindu\} |
| 4  | Kab      | \{Surabaya,Sidoarjo,Gresik,Bangkalan,Lamongan,Jombang,Mojokerto\} |
| 5  | Jabatan  | \{Guru,Pengawas,Kasi,Staf,'Staf',KaKel,KaUPTD,PengUPTD,Sekretaris,'As Ap,PenjUPTD,Dokter,Penilik,Kasubag,'Sekretaris ' ,Bidan,Perawat,Sanitarian\} |
| 6  | Instansi | \{DP,KEC,Dinkominfo,DKP,DPUKMP,DinKes,DK,DPPK,DinSos,DPhub,RUSD,DCKTR,DTK,SKPULPP,DKUMKM,DKebPar,DPO,DP,DPP,DKS, DinPer\} |
| 7  | Pekerjaan | \{Guru,'Non Guru'\} |
| 8  | UnitKerja | \{SDN,DPend,Seksi,SMAN,SuBag,UPTD,'SMPN ' ,Sekretaris,Kelurahan,Puskesmas,'SMKN ',SMPN,'UPTD ',SMKN,'Seksi ',Dinaker,RUSD,'Puskesmas ' ,SMAN ', 'SMPN ' , 'SMPN ' \} |
| 9  | TMTPengab | Numeric |
| 10 | TMTPens  | Numeric |
| 11 | Jenis Pensiun | \{Normal,Dipercepat\} |

3.2. Data processing
Data obtained from PT. Bank Bukopin Surabaya Branch with 1,299 retirement data. 845 training data and 454 testing data. Variables used for the calculation process are gender, age, name of religion, district, position, name of institution, work status, work unit name, service level, pension plan, and type of pension.
Table 2. Data training.

| No | Data pension |
|----|--------------|
| 1 | L,26,Islam,Surabaya,Guru,DP,Guru,SDN,1978,2012,Normal |
| 3 | L,26,Islam,Surabaya,Pengawas,DP,Guru,DPend,1979,2013,Normal |
| 4 | P,24,Islam,Surabaya,Guru,DP,Guru,SDN,1977,2013,Normal |
| ... | ... |
| ... | L,36,Islam,Surabaya,Staf,Dinkominfo,'Non Guru',Seksi,1993,2013,Dipercepat |

3.3. Set algorithm model
After the Processing stage, then for each classification algorithm (C4.5, NaiveBayes and k-NN) the model is determined at 845 training data. Where table 3 describes the time needed for these 2 stages (Build and Test Model).

Table 3. Build and test model on training data.

| Classification | build model | test model |
|----------------|-------------|------------|
| C4.5           | 0.05 seconds| 0.27 seconds|
| NaiveBayes     | 0 seconds   | 0.36 seconds|
| k-NN           | 0 seconds   | 0.31 seconds|

3.4. Testing data results
The results of the classification algorithm (C4.5, Naive Bayes and kNN) for the "Jenis Pensiun" class. Table 4 explains the comparison of class prediction.

Table 4. Result testing data.

| ...   | Jenis Pensiun | C4.5 | NaiveBayes | k-NN |
|-------|---------------|------|------------|------|
| ...   | Normal        | Dipercepat | Normal    | Normal |
| ...   | Dipercepat    | Dipercepat | Dipercepat | Dipercepat |
| ...   | Normal        | Normal   | Normal    | Normal |
| ...   | Normal        | Normal   | Normal    | Normal |
| ...   | Dipercepat    | Dipercepat | Dipercepat | Dipercepat |
| ...   | Dipercepat    | Dipercepat | Dipercepat | Normal |

And the total score for the results of 454 testing data in table 5, the results of the class prediction are obtained where the best is 97.14%, then 92.51% for the NaiveBayes algorithm and for the C4.5 algorithm and the lowest is 86.12% for the k-NN algorithm.

Table 5. Class prediction result

| Status | Normal | Accelerated | Total | Percentage |
|--------|--------|-------------|-------|------------|
|        |        |             |       |            |
| NaiveBayes |
| Benar   | 261    | 159         | 420   | 92.51%     |
| Salah   | 28     | 6           | 34    | 7.49%      |
| C45 |
| Benar   | 286    | 155         | 441   | 97.14%     |
| Salah   | 3      | 10          | 13    | 2.86%      |
| k-NN |
| Benar   | 264    | 127         | 391   | 86.12%     |
| Salah   | 24     | 28          | 52    | 11.45%     |
4. Conclusion

Analysis of research from three algorithms (C4.5, Naive Bayes and k-NN), which uses a dataset of 1,299 from PT. Bank Bukopin Surabaya Branch which was divided into 845 training data and 454 testing data. From the comparison of three algorithms, the best results are obtained for C4.5 Algorithm, namely with the correct prediction class value of 97.14% for C4.5 Algorithm, then the Naive Bayes algorithm is 92.51% and k-NN is 86.12%. So that the application is possible related to the analysis of the type of pension for customers of PT. Bank Bukopin Surabaya Branch can be relevantly implemented.

Acknowledgements

We hereby thank you to Universitas Muhammadiyah Sidoarjo for supporting the publication of this research.

References

[1] Argo P J K 2014 Analisis Peraturan Batas Usia Pensiun Pns Dalam Uu No. 5 Tahun 2014 Tentang Aparatur Sipil Negara
[2] Bagaskara R G and Lestianawan H Implementasi Algoritma Fuzzy C-Means Untuk Menemukan Pengelompokan Data Pensiun Di Badan Kepegawaian Daerah Kota Semarang 1–18
[3] Riza M, Kudang B M and Agus A 2018 Pembentukan Target Pasar Berdasarkan Data Stream Transaksi Kartu Kredit (Clustering Dan Association Rule) Pada Pt Bank Bukopin 4 1
[4] Widayu H, Nasution S D, Silalahi N and Mesran 2017 Data Mining UntukMemprediksi Jenis Transaksi Nasabah Pada Koperasi Simpan Pinjam Dengan Algoritma C4.5 Media Informatika Budidarma 1 2 pp 32–37
[5] Rosid M A, Rachmadany A, Multazam M T, Nandiyanto A B D, Abdullah A G and Widiaty I 2018 Integration Telegram Bot on E-Complaint Applications in College IOP Conf. Ser. Mater. Sci. Eng. 288 1
[6] Arif S F Prediksi Kelulusan Mahasiswa (Strata 1) Universitas Muhammadiyah Sidoarjo Menggunakan Algoritma C4.5
[7] Jananto A 2013 Algoritma Naive Bayes Untuk Mencari Perkiraan Waktu Studi Mahasiswa 18 1 9–16
[8] Everitt B S, Landau S, Leese M and Stahl D Miscellaneous
[9] Cover T and Hart P 1967 Nearest neighbor pattern classification Information Theory, IEEE Transactions on 13 1 21-27