Implementation of Naive Bayes for Classification and Potentially MSMEs Analysis

Meta Amalya Dewi[1,1,,] *Tri Wahyu Widyaningsih*[2]

[1]Tanri Abeng University, Information System Department, Jakarta Selatan, Indonesia
[2]Tanri Abeng University, Informatic Engineering Department, Jakarta Selatan, Indonesia

Abstract. Micro Small and Medium Enterprises (MSMEs) have an important role for a country in significantly increasing its economic growth, its high absorptive capacity to labor can reduce unemployment, and has become the biggest contributor of gross domestic product value. Therefore, the government should give more attention to MSMEs. However, the government does not have any information on the results of clustering analysis and prediction of potential business types from existing MSMEs data. This study aims to assist the government by presenting the results of potential MSMEs processing analysis in Tangerang region based on business characteristics in each region, using Naive Bayes. From the data of the number of MSMEs in Tangerang region, it has been successfully classified and the result of its analysis has become recommendation for the government in establishing the grow up as well as the provision of business assistance for the potential business field.

1 Introduction

MSMEs (Micro, Small and Medium Enterprises) is an easy-to-use term for business segmentation and other organizations that are between the sizes of "small office-home office" (SOHO) and larger companies [1]. In its development perspective, MSMEs are classified into 4 groups, namely [2]:

1. *Livelihood Activities*, is the MSMEs used to make a living, which is more popular as an informal sector. Examples are street vendors.
2. *Micro Enterprise*, is the MSMEs that has the nature of craftsmen but does not have the entrepreneurial nature
3. *Small Dynamic Enterprise*, is the MSMEs that has an entrepreneurial spirit and is able to accept subcontracting job and export
4. *Fast Moving Enterprise*, is the MSMEs that has an entrepreneurial spirit and will transform into a Big Business

MSMEs are going to be one of the fields that contribute significantly to spurring economic growth in Indonesia. MSMEs present have a direct impact on the economy, reduce unemployment and become the largest contributor to the value of gross domestic product [3]. In Europe, MSMEs are an important part of the economy [4] which is the

[1] Corresponding author: meta.amalya@tau.ac.id

driving force for industrial growth and economic development. The growth of MSMEs in Indonesia is increasing rapidly from year to year, data from the Ministry of Cooperatives and MSMEs in 2013 were 57,900,787 MSMEs increasing by 1,361,227 from 2012 [5].

Tangerang has 1850 officially registered MSMEs spread across 29 region, 28 sub-districts and 246 villages [6] and will continue to grow like the vision of the Department of Cooperatives and MSMEs to move the populist economy in Tangerang, but there has been no mapping of potential MSME analysis and business results. which can help provide information and knowledge to the government regarding potential types of businesses in certain areas to be upgraded.

Data mapping can be done by various methods, one of which is by using data mining for the process of extracting information from a very large collection of data through the use of algorithms and withdrawal techniques in the statistics, machine learning and database management systems [7].

Naive Bayes is used as probability and statistical method to classify data and predict future opportunities based on previous experience [8], easily implemented and proven to provide good predictions [9]. The advantage of using Naive Bayes is that it only requires small training data to determine the estimated parameters needed in the classification process. Naive Bayes often works much better in many complex real-world situations than expected [10]. Extracting information on a data that has a large number of records and fields cannot be done easily. Data mining techniques are one of the tools for extracting data in large databases and with complexity level specifications that are widely used in many application domains such as banking and telecommunications [11].

Data mining tools can help hidden knowledge in large amounts of data [12]. Data mining is a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract useful information and then get knowledge from large databases [13]

2 Research Methodology

This research was carried out in the following steps:

1. Designed Measurement Research
   a. Determined Research Variables
   b. Determined Performance Parameters
2. Data Retrieval
   This is the stage of gathering the required data. Data obtained from secondary sources, through the official website of the cooperative service and the Tangerang government MSME.
3. Data Mining Mapping
   As a series of processes, data mining has several stages that are interactive, where users can interact directly or through an intermediary knowledge base. Among the stages of the process are [14]:
   a. Data cleaning and data integration
   b. Data selection and data transformation
   c. Mining process
   d. Pattern evaluation
   e. Knowledge presentation
4. Analysis
5. Documentation

The steps of this research are illustrated through the diagram in Figure 1 below :

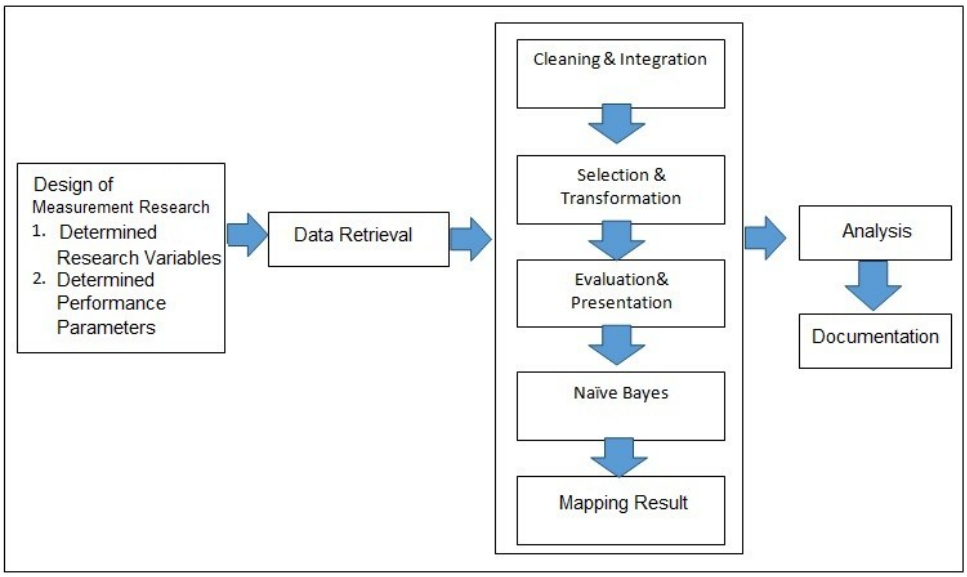

Fig. 1. Research Methods

3 Results And Discussion

The MSMEs data in Tangerang has 7 fields including id, business name, business name, business address, telephone number, business form, and business fields as in table 1 below.

Table 1. Tangerang MSMEs Field Data

No	Field Name	Information
1	Id	Serial number
2	Name of Entrepreneur	Name of registered business owner
3	Business Name	Business name registered
4	Business Address	Complete address of the place of business including the name of the road, RT / RW, region, sub-distric
5	Phone Number	Phone number or cell number (HP)
6	Form of Business	Individual, cooperative, or CV / FA
7	Business fields	Types of businesses that are run

3.1 Cleaning and Integration

This activity is a process of cleaning up data from imperfect entries, inconsistent writing, invalid data, typographical errors, and removing items with empty values. Data adjustments are made to ensure that all data has a complete address to facilitate the mapping process based on sub-district and business fields. Correction techniques can be seen in table 2 below:

Table 2. Inconsistent Correcting Technique Data For Each Table Attribute

Atributte	Information
Address	Eliminate data records with incomplete addresses and addresses with sub-districts outside Tangerang

Business fields	Eliminate data records that are business fields with incorrect data, for example: selling, trading, traveling, and so on.

From this cleaning process, all data records are generated with complete contents by adding sub-district attributes which are the results of the fraction of the address attribute, and adding business classification attributes which are the result of business data grouping. All attributes can be seen in table 3 below

Table 3. Data Integration

No	Field Nama	Information
1	Id	Serial number
2	Name of Entrepreneur	Name of registered business owner
3	Business Name	Business name registered
4	Business Address	Complete address of the business place
5	Sub- districts	The area where the business is run
6	Phone Number	Phone number or cell number (HP)
7	Form of Business	Individual, cooperative, or CV / FA
8	Business fields	Types of businesses that are run
9	Business Classification	Specifications of business groupings

3.2 Evaluation and Presentation

After the data integration process, information is obtained on the percentage of business fields per sub-district as shown in Figure 2 below :

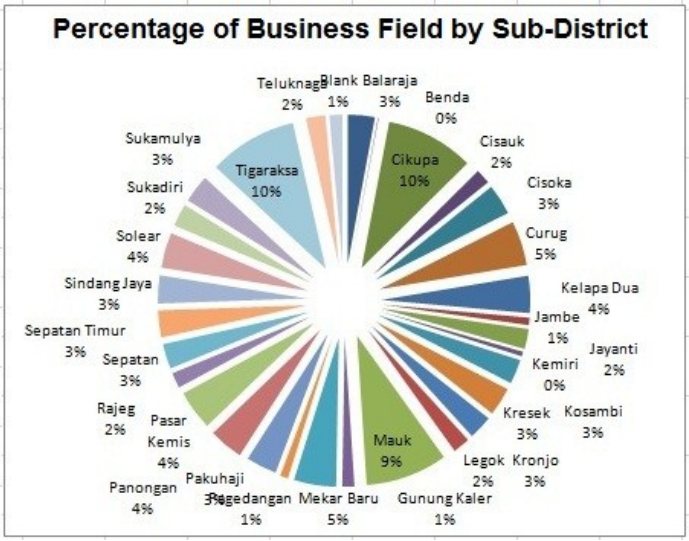

Fig. 2. Percentage of Business Field by Sub-District

And the percentage of business information based on business classification can be seen in Figure 3 below :

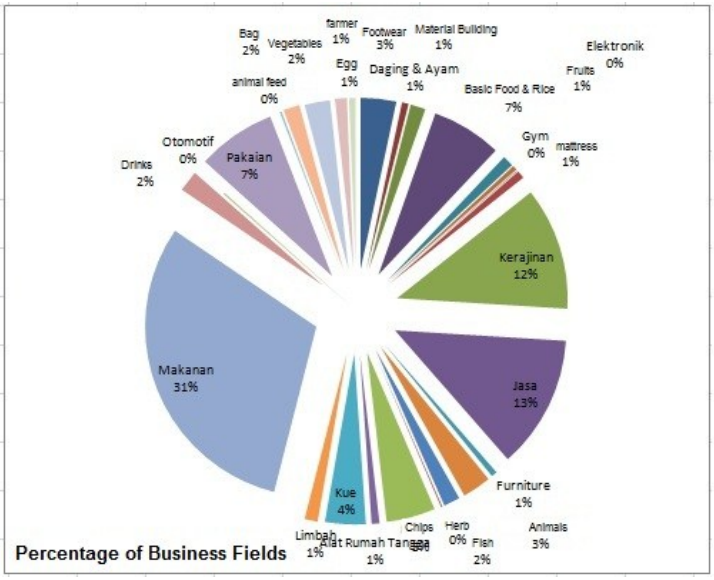

Fig. 3. Percentage of Business Fields

3.3 Naive Bayes

In this process the file and select column are in the data menu to select the file to be analyzed, select column functions to select the variable target. Naive bayes is on the Model menu as a statistical classification model, test and score is in the Evaluate menu to see the results of the analysis with data input and output evaluation result. The process can be seen in Figure 4 below :

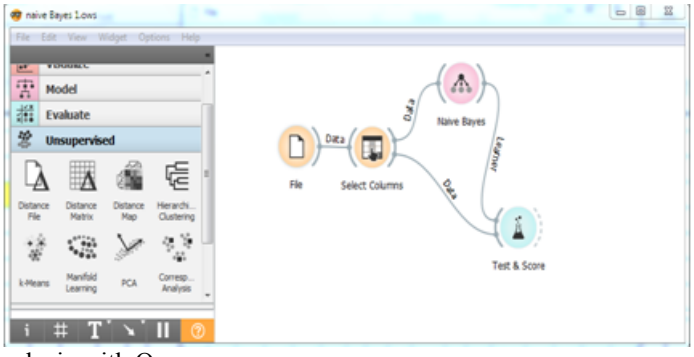

Fig.4. Data Analysis with Orange

The next step is select column as a variable target. In select column there are features, target variables and meta attribus. The target variable must be determined first to determine the results of a better analysis. In this process, a business form field is selected as shown in Figure 5 below :

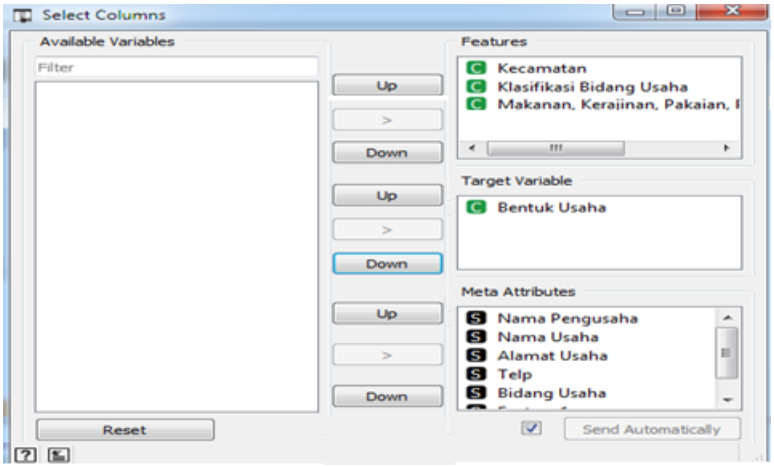

Fig.5. Select Column as Target Variabel

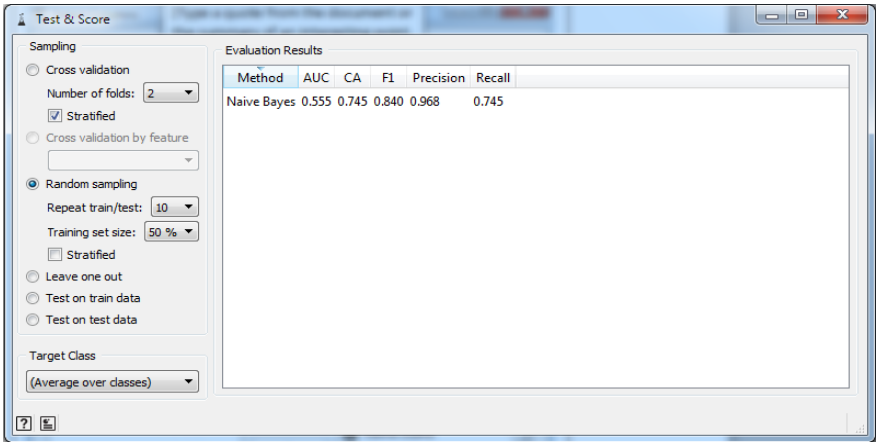

Fig.6. Result of Evaluation Test and Score

Based on Figure 6 above, the level of data accuracy is measured from 5 criteria [15] namely AUC, CA, F1, Precision, and Recall. F1 score results are valid based on the following formula [16] :

$$2\,x\,\frac{precision * recall}{precision + recall}$$

AUC (Area Under Curve) : 0.55 shows the accuracy of the diagnostic test in general (getting closer to 1 is getting better). CA (Capability Analytics) : 0.74, F1 (matrix that combines recall and precision using harmonic mean) : 0.84, precision (the value of a classification model to return only the relevant instances) : 0.968 and recall (the value of a classification model to identify all relevant instances) : 745

3.4 Classification Analysis Results

Based on all the processes that have been completed, the results of ranking 9th District in Tangerang which are considered the most productive are:

1. Tigaraksa with a productivity level of 10%
2. Cikupa with a productivity level of 10%
3. Mauk with a productivity level of 9%
4. Curug with a productivity level of 5%
5. Mekar Baru with a productivity level of 5%
6. Solear with a productivity level of 4%
7. Pasar with a productivity level of 4%
8. Panongan with a productivity level of 4%
9. Kelapa with a productivity level of 4%

Based on the results of the classification analysis of business fields that are in demand by the public, shows the ratings:

1. Food is 31%, with the most productive area is Cikupa
2. Services is 13%, with the most productive area Mekar Baru
3. Crafts is 12%, with the most productive area Tigaraksa
4. Basic food and rice are 7%, with the most productive area Tigaraksa and Cikupa
5. Clothing is 7%, with the most productive area Curug

4 Conclusion

After completing research on potential MSMEs classification and analysis, some conclusions can be drawn as follows :

a. The Naive Bayes method has been successfully carried out in generating the most productive MSMEs classification and analysis based on sub-districts and regions based on business specifications from 1850 Tangerang MSMEs data containing 7 attributes (id, name of enterpreneur, business name, business address, telephone number, business form, and business fields) which results in 9 data integrity (id, name of enterpreneur, business name, business address, telephone number, sub-seconds, business forms, business fields, and business classifications)
b. From the analysis results can be used as recommendations in establishing a grow up policy and business assistance based on potential business fields in areas that have high business productivity.

References

1. M. Rouse (2011) small to medium enterprise (SME), Business intelligence - business analytics retrieved from http://whatis.techtarget.com/definition/small-to-medium-enterprise-SME Accessed Maret, 17th 2018
2. Sudaryanto, Regimun, and R.R. Wijayanti,"Srategi Pemberdayaan UMKM Menghadapi Pasar Bebas ASEAN", 2014.www.kemenkeu.go.id/sites/default/files/strategi%20pemberdayaan%20umkm.pdf
3. D. Sena, M. Ozturkb, and O. Vayvay, "An Overview of Big Data for Growth in SMEs", Elsavier 12th International Strategic Management Conference, pp. 159-167 (2016)
4. B.O. Ogbuokiri, C.N. Udanor, and M.N. Agu, "Implementing bigdata analytics for small and medium enterprise (SME) regional growth", IOSR Journal of Computer Engineering, Vol. 17 (6), pp.35-43 (2015)
5. Departemen Koperasi, http://www.depkop.go.id/pdf-viewer/?p=uploads/tx_rtgfiles/sandingan_data_umkm_2012-2013.pdf diakses tanggal 18 Maret 2018

6. website kab tangerang https://tangerangkab.go.id/
7. Taruna R., S., Hiranwal, S., "Enhanced Naive Bayes Algorithm for Intrusion Detection in Data Mining", *International Journal of Computer Science and Information Technologies*, Vol.6 (4), pp. 960-962 (2013)
8. Bustami., "Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi", Jurnal Penelitian Teknik Informatika, Vol. 3 (2), pp. 127-146 (2013)
9. S. Sarkar, R.S. Sriram, "Bayesian models for early warning of bank failures. Management Science", Management Science,Vol. 47 (11), pp. 1457–1475 (2001)
10. S. A. Pattekari, A. Parveen, Prediction System for Heart Disease Using Naive Bayes, *International Journal of Advanced Computer and Mathematical Sciences,* ISSN 2230-9624, Vol. 3 (3), pp. 290-294. (2012)
11. Jananto, Arief. *Memprediksi Kinerja Mahasiswa Menggunakan Teknik Data Mining (Studi kasus data akademik mahasiswa UNISBANK*. Tesis Tidak Terpublikasi. Yogyakarta: Universitas Gajah Mada (2010)
12. E.W.T. Ngai a,*, Li Xiu b , D.C.K. Chau a, Application of data mining techniques in customer relationship management: A literature review and classification, Expert Systems with Applications 36. 2592–2602 (2009)
13. Turban, E., Aronson, J. E., Liang, T. P., & Sharda, R. Decision support and business intelligence systems (Eighth ed.). Pearson Education (2007)
14. J. Han and M. Kamber, Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc, San Francisco (2006)
15. https://www.scribd.com/doc/15123416/Kurva-Receiver-Operating-Character diakses Agustus 2018
16. https://mragungsetiaji.github.io "Machine learning: matrics recall and precision" diakses Agustus 2018