IJAIT (International Journal
of Applied Information Technology)

http://journals.telkomuniversity.ac.id/ijait/

Poverty Level Prediction Based on E-Commerce Data Using Naïve Bayes Algorithm and Similarity-Based Feature Selection

Pramuko Aji [a, *], Dedy Rahman Wijaya [a], Elis Hernawati [a], Sherla Yualinda [a], Sherli Yualinda [a], Muhammad Akbar Haikal Frasanta [a], Rathimala Kannan [b]

[a] *School of Applied Science, Telkom University, Bandung, Indonesia*

[b] *PSG Institute of Management, PSG College of Technology, Coimbatore, India*

pramuko@telkomuniversity.ac.id, dedyrw@telkomuniversity.ac.id, elishernawati@telkomuniversity.ac.id, sherla@student.telkomuniversity.ac.id, sherli@student.telkomuniversity.ac.id , jerichoholic535@gmail.com, rathimala@psgim.ac.in

ARTICLE INFO

Received August 24th, 2022
Revised June 16th, 2023
Accepted July 18th, 2023
Available online October 20th, 2023

Keywords
Poverty, BPS, naive bayes, feature selection, e-commerce data

ABSTRACT

The poverty rate is an important measure of any country because it indicates how well the economy develops and how well the economic prosperity distributes among citizens. The Central Statistics Agency, or BPS, measures the poverty rates in Indonesia using the concept of the ability to meet demands (basic needs approach). Using this approach, spending becomes a measure of poverty, defined as an economic incapacity to satisfy food and non-food requirements. Thus, the poor are individuals whose monthly per capita spending is less than the poverty threshold. In this study, the machine learning method using Naive Bayes with similarity-based feature selection and e-commerce data has been proposed to predict the poverty level in Indonesia. We proposed the method to be used as a complement to the results of the costly surveys and censuses conducted by BPS. Our experiments show that the classifier shows little relevance between the predicted and the original values or actual poverty prediction based on BPS data. A limited number of features does not necessarily result in poor accuracy, however great accuracy is not always achieved if a lot of features are being used.

* Corresponding author at:
School of Applied Science, Telkom University,
Jl. Telekomunikasi No. 1, Terusan Buah Batu, Bandung, 40257,
Indonesia.
E-mail address: pramuko@telkomuniversity.ac.id

ORCID ID:
- First Author: 0000-0003-4356-7157
- Second Author: 0000-0003-0351-7331
- Third Author: 0000-0002-5855-3615

https://doi.org/10.25124/ijait.v7i02.5374
Paper_reg_number IJAIT000070204 2023 © The Authors. Published by School of Applied Science, Telkom University.

1. Introduction

The poverty rate is an important measure of any country because it indicates how well the economy develops and how well the economic prosperity distributes among citizens. Poverty has been said to be linked with a lot of social problems, including crimes and health. Therefore, the goal of any economic growth is to reduce unemployment and poverty rates [1].

According to the Central Statistics Agency (BPS), poverty rate in Indonesia reached 27.54 million people or 10.3 percent of the total population of Indonesia [2], which reached 270 million lives in 2020 [3]. The rates have been measured by the Central Statistics Agency (BPS) using the idea of the ability to fulfill demands (basic needs approach). Using this concept, poverty is defined as an economic incapability to meet the needs of food and non-food, which are measured in terms of expenditure, so that the low-income persons are residents with per capita spending is less than the poverty line. The poverty line concept contains these two aspects, namely the Food Poverty Line (FPL) and Non-Food Poverty Line (NFPL) [4].

The FPL is the value or outcome of spending on minimum food necessities, which includes 2100 kilocalories per capita per day. Fifty-two commodity types represent commodity packages of basic food needs (tubers, grains, meat, eggs, milk, vegetables, nuts, fruits, oils, etc.). The NFPL is a bare minimum for clothing, housing, health, and education. In rural areas, 47 species and 51 kinds of non-food basic needs are represented. The Poverty Line (PL) is the result of adding the FPL and the NFPL. The category of poor population is the population whose average monthly expenditure is below the Poverty Line. Rice and cigarettes are among the biggest contributors to the total Poverty Line (PL) in urban and rural areas with the amount spent reaching 73.48%.

The Central Statistics Agency (BPS), especially BPS Bandung, conducts a National Socio-Economic Survey (SUSENAS) which aims to obtain data as a picture of the socioeconomic conditions. Starting in 2015, data collection on the National Socio-Economic Survey was carried out in March, and in 2018 SUSENAS was carried out in all provinces in Indonesia (34 Provinces) with a sample size of up to 300,000 households spread across 514 districts/cities in Indonesia. The sample does not include households that live in special households or special blocks such as dormitories and prisons [5].

Data collection from selected households will be conducted through a direct interview between the enumerator and the respondent (head of household, husband/wife, and other family members) selected. The processing process carried out by BPS staff to obtain data begins with the stage of storing data through data recording, and then checking the suitability of the contents of the data with the results of the questionnaire to the tabulation stage with the help of a computer. Before performing the steps above, the officer will first check the completeness of the list of questions, editing the data contents deemed unnatural, including the relationship or consistency between the answers of one answer to another.

Based on the statements above, it can be concluded that the problem in determining the poverty level in a region comprises of many aspects. Therefore, another perspective is needed so that poverty can be seen more deeply and accurately. Data collections by BPS are lengthy and the steps involved are quite complicated. In addition, the interview process conducted with the head of the household is generally difficult because some heads of the household were reluctant to meet up or avoidant for fear of fraud.

Thus, another method proposed by researchers to augment survey results in predicting poverty in an area is to use Naive Bayes with Similarity-Based e-commerce methods. E-commerce data is used because, among Southeast Asian countries, Indonesia has the largest e-commerce market, contributing about 50% of all running transactions. This contribution will continue to increase because Indonesian residents often use the internet for daily activities.

E-Commerce is a flow of goods buying and selling online. According to McKinsey, e-commerce falls into two categories. The first category is E-tailing, that is formal buy and sell through online platforms built specifically to facilitate transactions. The second category is Social Commerce that utilizes social media such as Facebook or Instagram as a medium for trading goods with payment and shipping via other platforms [6]. In 2018, research results disclosed by management consulting firm McKinsey and Company include growth in the value of the Indonesian e-commerce market until 2022 and the potential impact of such growth on the Indonesian economy and society. McKinsey stated that the e-commerce market is predicted to increase eightfold by 2022 [6]. The following is an illustration of the predicted increase in e-commerce data in Indonesia. Indonesia's e-commerce market value growth in 2017 in E-Tailing amounts to around IDR 70 trillion and Social Commerce around amounting to more than IDR 42 trillion while in 2022, E-Tailing reaches IDR 563 trillion and Social Commerce around IDR 211 to 351 trillion. This value is converted from USD where USD 1 = IDR 14.075. E-commerce market growth in Indonesia is predicted to generate around USD 65 billion equivalent with IDR 910 trillion.

The purpose of this study is to investigate the accuracy of machine learning, particularly the Naïve Bayes algorithm and similarity-based feature selection, in predicting the property level.

This paper is structured as follows: Section 2 explains some theories related to this discussion. Section 3 illustrates the basic theories and methods of implementation. In Section 4, we explain the results and discussion. Finally, in Section 5, we explain the conclusions of all the results of the implementation produced.

2. Related Works

In [7], the authors predicted the poverty level in the African continent from 2000 to 2010 from Landsat 7 satellites (night lights) using the CNN method. The study concluded that determining poverty predictions utilizing this satellite data is suitable for predicting wealth in a country. However, it can also extrapolate national boundaries [8].

Another study in [7] explained that cell phone use can be used to find out which names are relatively rich and which ones are relatively poor in Rwanda. This study has a three-step method. The first step is comparing the overall demographic composition from the location of Rwanda to measure differences between those who have cell phones and those who do not have cell phones. The second step is checking survey data from the newly used cell phone. The last step is analyzing the respondent's calls. The results produced from this study show that in Rwanda, the people are far from uniform, there are significant and systematic differences, and more calls are made by the one who are educated and come from wealthy families [7].

Poverty prediction using communication networks is explained in [9]. The results can benefit mobile users because in this study the user's private data is protected [9].

Meanwhile, [10] incorporate the classification process in the dimensionality curse using the Equality Measure, entropy size, feature selection, and Fuzzy k-NN Classification. From the experimental results, an average accuracy of 80.5% is obtained [10].

Another approach in the newly developed system for multimodal biometric authentication using the ReliefF algorithm is used by selecting optimal features and then classifying using the supervised classifier MSVM producing accuracy reaching 97.09% [11] Furthermore, there is a hybrid SOMI method combined with GANB using the NB (Naïve Bayes classification) to achieve high accuracy in several heterogeneous datasets [12]. The Naïve Bayes approach is also utilized in the spam detection system on e-mails which produces an accuracy of around 83.5% by using the hybrid bagged approach for its application [13].

A paper entitled "Electronic Nose for Classifying Beef and Pork using Naïve Bayes" obtained 75% accuracy based on k-fold cross-validation to distinguish beef from pork [14].

Another paper entitled "Sensor Array Optimization for Mobile Electronic Nose: Wavelet Transform and Filter Based Feature Selection Approach" contributes to reducing noise from gas sensors. The results of this experiment indicate that this method can reduce noise by 14.41% and produce the best combination [15]. The FSA or single feature selection algorithm described in a paper indicates that a single FSA cannot guarantee stable sensor recommendations in the optimization of sensor arrays [16]. E-commerce data can be used as a proxy for calculating poverty levels in cities and can monitor poverty-level development indicators because e-commerce data has the potential to represent real household expenditure described in a paper entitled "Estimating city-level poverty rate based on e-commerce data with machine learning" [17].

3. Materials and Methods

3.1. Dataset

A dataset from an Indonesian e-commerce company has been used in this study. The dataset consists of eight items, that is; (1) houses for sale, (2) apartments for sale, (3) apartments for rent, (4) houses for sale, (5) cars, (6) motorcycles, (7) lands for sale, and (8) land purchases. Items number 1, 2, 3, 4, 7, and 8 were advertisements related to property, which is considered a good representation of the housing category in SUSENAS. While items number 5 and 6 represent the household consumption module in SUSENAS since they imply routine gasoline and maintenance expenses.

The data has been limited to advertisements originating from Java Island since the island was regarded as the greatest contributor to e-commerce, with approximately 18,881,913 advertisements coming from 118 cities/regencies within. Table 1 shows the number of advertisements for each item.

Table 1 Details of Dataset

Items	Number of Advertisements
Houses for sale	3,594,545
Houses for rent	336,758
Lands for sale	1,179,972
Lands for rent	13,916
Motorbikes	6,313,016
Cars	6,933,513

Items	Number of Advertisements
Apartments for sale	250,504
Apartments for rent	259,689

3.2. Proposed Method

Figure 1 depicts the methodology used in this study.

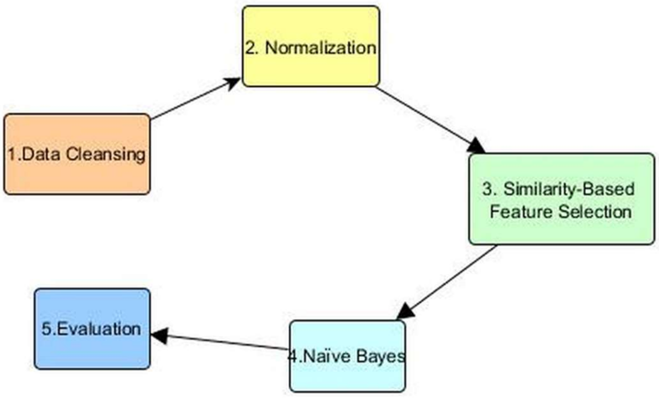


Figure 1 Proposed Method

Figure 1 The methodology consists of five stages. The first two stages would serve as data preparation. The third stage is the process of selecting the most suitable features. The fourth stage is the where the poverty level predictions will be calculated. Finally, the final stage is the evaluation of the outcomes of the previous stages.

3.2.1. Data Cleansing

First, the data went through the preparation or cleaning process. This stage is essential because it aims to clean up incomplete data or missing values which may interfere with the process in its entirety, so that the overall data cannot be processed into the next stage because they were considered unready. In the dataset used, missing or null values were still found. These missing or null values were then replaced with zero, and further, each of them had to be classified as a feature or a label.

3.2.2. Normalization

After going through the first step, which is data cleansing, then the dataset proceeded to the normalization process, where the data that has been determined is scaled from 0-10. In the process of normalization, the authors use a method of readjustment called the min-max normalization. A min-max method is carried out to transform linear to original data. Equation 1 is the formula for Min-Max normalization [17].

$$\text{MinMax} = \frac{x - min\ (x)}{max(x) - min\ (x)} \times 10 \qquad \text{Equation 1}$$

Where x represents the value of each characteristic, min (x) represents its lowest value, and max (x) represents its highest value.

3.2.3. Similarity-Based Feature Selection

The following stage, known as the Similarity-Based Feature Selection, will be initiated when the data has been normalized. The Similarity-Based Feature Selection algorithm exploits various types of features to determine which features are suitable. Similarity-based data may be produced from label information in supervised feature selection. However, most unsupervised features employ various metric stages to collect data. In Similarity-Based Feature Selection, there are several feature selection methods, namely fisher score, lap score, relief, SPEC, and trace ratio. Below are some formulas from the similarity-based feature selection algorithm. Fisher Score is a supervised feature selection algorithm in which the value of features in the same class is the same and the value of features in different classes is not the same. Equation 2 depicts the formula [18].

$$\text{Fisher\_score } (f_i) = \frac{\sum_{j=1}^{c} n_j(\mu_{ij}-\mu_i)^2}{\sum_{j=1}^{c} n_j \sigma_{ij}^2} \qquad \text{Equation 2}$$

It is generally established that n_j is the number of samples available in class j, μ_i is the average value in feature f_i, μ_{ij} is the value of feature f_i for sample in class j and finally, σ_{ij}^2 is variance of feature fi for samples in class j. While RelieffF chooses the features that are used to be separated. Equation 3 is the formula from ReliefF [18].

$$\text{ReliefF\_score } (f_i) = \frac{1}{c}\sum_{j=1}^{l}(-\frac{1}{m_j}\sum_{x,\epsilon NH(j)} d\big(X(j,i) - X(r,i)\big) + \sum_{y \neq y_j} \frac{1}{h_{jy}} \frac{p(y)}{1-p(y)} \sum_{x,\epsilon NM(j,y)} d\big(X(j,i) - X(r,i)\big)) \qquad \text{Equation 3}$$

$NH(j)$ and $NM(j,y)$ are the closest examples to x_j of the same class and in class y, the sizes of each are m_j and h_{jy}. $p(y)$ is the ratio of instances in class y. And the next is the trace ratio which directly selects globally the subnet feature based on the appropriate score. Equation 4 is the formula for trace ratio [18].

$$\text{Trace\_ratio } (\mathcal{S}) = \frac{tr(W'X'L_bXW)}{tr(W'X'L_wXW)} \qquad \text{Equation 4}$$

where L_b and L_w are the Laplacian matrices of S_a and S_b.

3.2.4. Naïve Bayes

After the data proceeds into the Similarity-Based Feature Selection, the data will be processed into machine learning Naïve Bayes where Naïve Bayes is a classification method that uses probability and statistical methods. The Naïve Bayes algorithm can predict future opportunities from a previous experience known as the Bayes Theorem [19]. Equation 5 is Bayes' Theorem [20].

$$P(H \mid X) = \frac{P(X \mid H).P(H)}{P(X)} \qquad \text{Equation 5}$$

Where P (H | X) is the posterior probability of the H hypothesis based on condition X and P (H) is the hypothesis probability, and X is data with an undefined class. P (X | H) is the possibility of hypothesis X given the conditions of hypothesis H, whereas P (X) is the possibility of hypothesis X in separately.

3.2.5. Evaluation

The last step involves two metrics for evaluating the goodness of fit of the regression. These are coefficient of determination (R^2) and root mean squared error (RMSE). RMSE is used to measure differences and errors between actual and predictive values. Low RMSE indicates little discrepancies between the actual and predicted values. The formula of RMSE is described in Equation 6 [21].

$$RMSE(y,\hat{y}) = \sqrt{\frac{\sum_{i=1}^{L}(y_i - \hat{y}_i)^2}{L}} \qquad \text{Equation 6}$$

The Coefficient of determination (R^2) is used measure the strength of the relationship between the actual value (poverty prediction value from BPS) and the predicted results using this method. It ranges from 0 to 1. R^2 value of 1 indicates that the predicted results fully explain the variance of the observed values. Thus, regression is well suited for predicting values. If the value of R^2 close to zero, it means that almost no variance in the data can be explained by the regression model. Negative R^2 values indicates that the model does not correctly predict the values. The formula of R^2 is described in Equation 7 [21].

$$R^2\,(y,\hat{y}) = 1\text{-}\frac{\sum_{i=1}^{L}(y_i-\hat{y_i})^2}{\sum_{i=1}^{L}(y_i-\bar{y})^2} \qquad \text{Equation 7}$$

3.3. Software Used

In this study we have used several software. Our tools are mainly based on the Python platform. Main libraries for machine learning were scikit-learn and scikit-feature. Scikit-learn has been used to calculate the Naïve Bayes classification while scikit-feature has been used to perform similarity-based feature selection. Scikit-learn and scikit-feature are popular machine learning libraries which depend on NumPy and SciPy scientific computing library. Experimental codes were written using Jupyter Notebook.

4. Result and Discussion

We collected 96 features after data cleansing and normalization process, and then the feature selection process was performed. Experimentation was conducted by varying the number of selected features, started with 10, 20, 30, 40, 50, 60, 70, 80, 90, up to 96 features. The goodness of fit of these different sets of features were calculated, and results are described below.

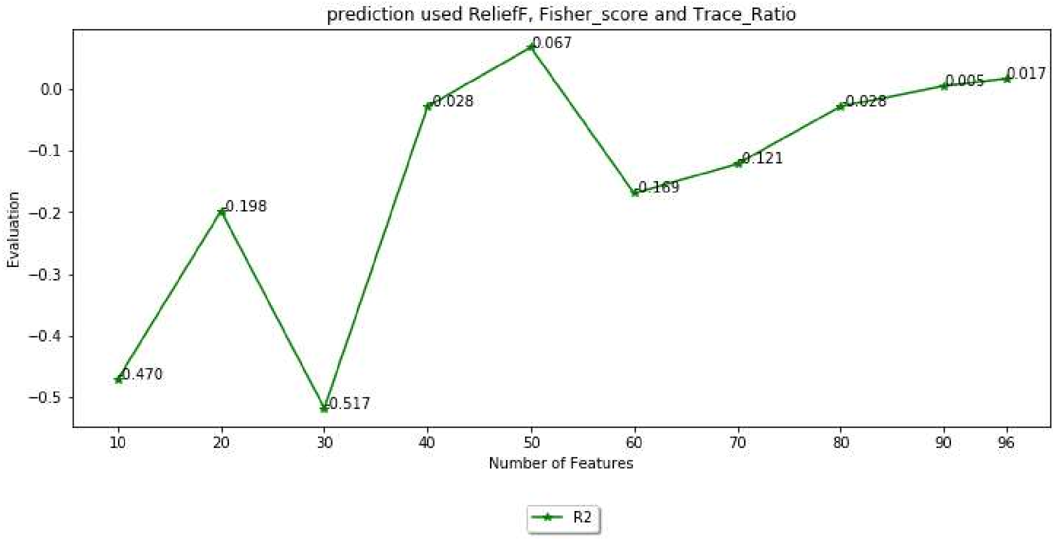


Figure 2 Coefficient of Determination Values of Different Number of Features

Figure 2 is an evaluation using R^2 and in Figure 3 is an evaluation using RMSE. In Figure 2, it appears that the highest R^2 falls on the use of 50 features with a result of 0.067. These results indicate that using too few features will reduce performance, but it does not guarantee that using many features will increase the value of R^2.

In Figure 3, it appears that the highest RMSE falls on the use of 30 features with a result of 6.046. Therefore, it can be concluded that the highest accuracy of E-commerce data predictions in this study is the use of 50 features.

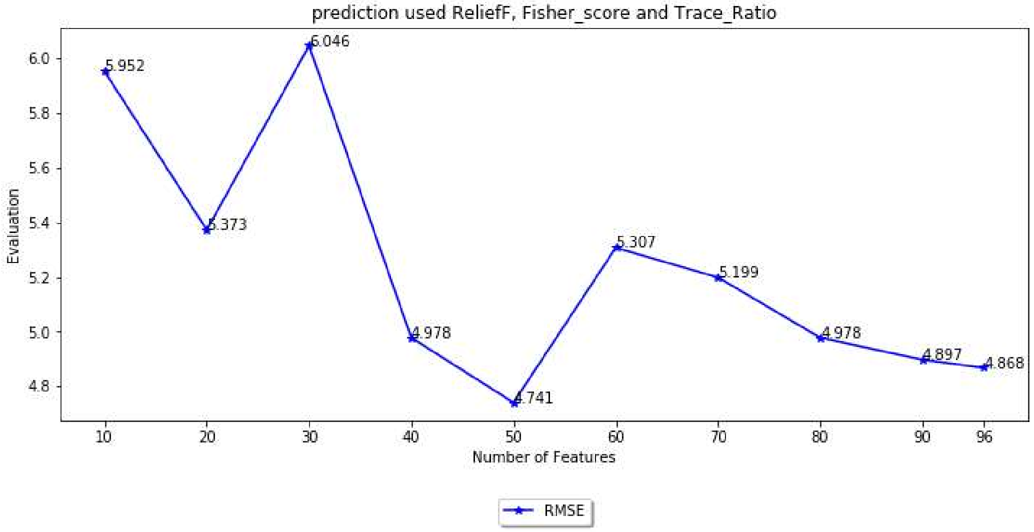
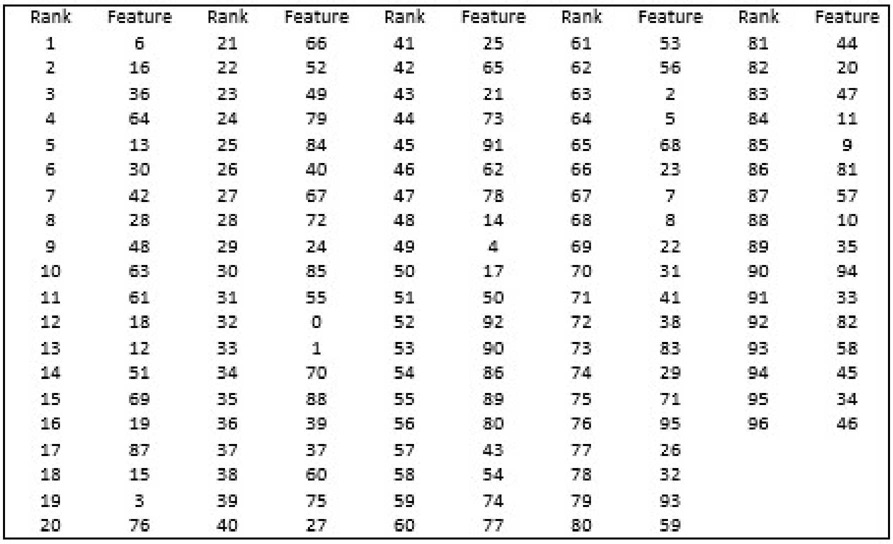


Figure 3 Evaluation With RMSE

Evaluations shown in Figure 2 and Figure 3 are the result of the ReliefF, Fisher Score, and Trace Ratio algorithms. These three algorithms produce identical R^2 and RMSE scores because in the Naive Bayes (GaussianNB) machine learning, Leave-One-Out process cannot accept float data types but can accept integer data types. However, these identical results were still coming from different feature rankings. The following are the results and discussion of each algorithm.

4.1. ReliefF Feature Selection

Figure 4 is the result of using the ReliefF algorithm, which can be explained as follows.



Rank	Feature	Rank	Feature	Rank	Feature	Rank	Feature	Rank	Feature
1	6	21	66	41	25	61	53	81	44
2	16	22	52	42	65	62	56	82	20
3	36	23	49	43	21	63	2	83	47
4	64	24	79	44	73	64	5	84	11
5	13	25	84	45	91	65	68	85	9
6	30	26	40	46	62	66	23	86	81
7	42	27	67	47	78	67	7	87	57
8	28	28	72	48	14	68	8	88	10
9	48	29	24	49	4	69	22	89	35
10	63	30	85	50	17	70	31	90	94
11	61	31	55	51	50	71	41	91	33
12	18	32	0	52	92	72	38	92	82
13	12	33	1	53	90	73	83	93	58
14	51	34	70	54	86	74	29	94	45
15	69	35	88	55	89	75	71	95	34
16	19	36	39	56	80	76	95	96	46
17	87	37	37	57	43	77	26		
18	15	38	60	58	54	78	32		
19	3	39	75	59	74	79	93		
20	76	40	27	60	77	80	59		

Figure 4 Feature Ranking using ReliefF Feature Selection

Figure 4 is a ranking sequence of e-commerce data features using the ReliefF algorithm. Figure 4 shows that feature 6 is ranked first, feature 16 is ranked second, and so on.

Figure 5 illustrates features among 10, 20, 30, 40, 50, 60, 70, 80, 90, and 96. Each number of features represents a different value generated from the ReliefF algorithm. The graph with 50 features produces the highest R^2 value. The highest value is valued with many points approaching straight lines or original data.

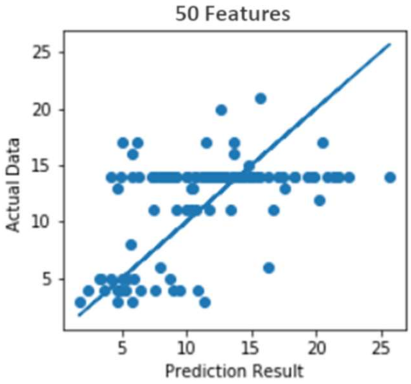
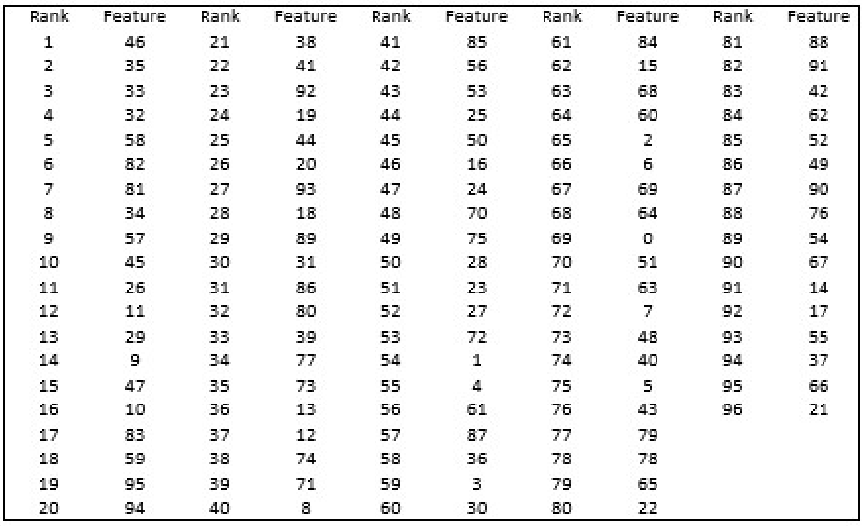


Figure 5 The Best predictive picture from Relief F Feature Selection

4.2. Fisher_Score Feature Selection

Figure 6 shows the result of using the ReliefF algorithm, which can be explained as follows.



Rank	Feature	Rank	Feature	Rank	Feature	Rank	Feature	Rank	Feature
1	46	21	38	41	85	61	84	81	88
2	35	22	41	42	56	62	15	82	91
3	33	23	92	43	53	63	68	83	42
4	32	24	19	44	25	64	60	84	62
5	58	25	44	45	50	65	2	85	52
6	82	26	20	46	16	66	6	86	49
7	81	27	93	47	24	67	69	87	90
8	34	28	18	48	70	68	64	88	76
9	57	29	89	49	75	69	0	89	54
10	45	30	31	50	28	70	51	90	67
11	26	31	86	51	23	71	63	91	14
12	11	32	80	52	27	72	7	92	17
13	29	33	39	53	72	73	48	93	55
14	9	34	77	54	1	74	40	94	37
15	47	35	73	55	4	75	5	95	66
16	10	36	13	56	61	76	43	96	21
17	83	37	12	57	87	77	79		
18	59	38	74	58	36	78	78		
19	95	39	71	59	3	79	65		
20	94	40	8	60	30	80	22		

Figure 6 Feature Ranking using Fisher Score Feature Selection

Figure 6 is a ranking sequence of e-commerce data features using the ReliefF algorithm. In Figure 6, it is displayed that feature 46 is ranked first, feature 35 is ranked second, and so on.

Figure 7 below illustrates the results of the Fisher Score algorithm with several different features. The highest value of R^2 falls on the use of 50 features, same result as ReliefF algorithm.

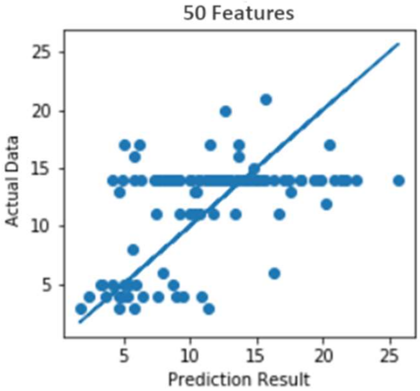
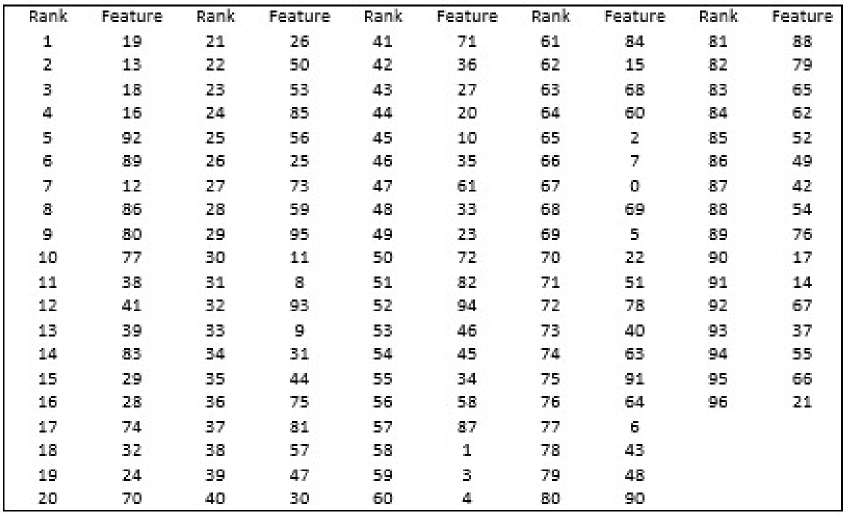


Figure 7 The Best Prediction result using Fisher_Score Feature Selection

4.3. Trace Ratio Feature Selection

Figure 8 shows the result of using the Trace Ratio algorithm which can be explained as follows. Figure 8 is a ranking sequence of e-commerce data features using the ReliefF algorithm. In Figure 8, it is shown that feature 19 is ranked first, feature 13 is ranked second, and so on. Figure 9 shows the results of the Trace Ratio algorithm with several different features. The Trace Ratio algorithm also yields the same result with the former two algorithm, with 50 features produce the highest R^2.



Rank	Feature	Rank	Feature	Rank	Feature	Rank	Feature	Rank	Feature
1	19	21	26	41	71	61	84	81	88
2	13	22	50	42	36	62	15	82	79
3	18	23	53	43	27	63	68	83	65
4	16	24	85	44	20	64	60	84	62
5	92	25	56	45	10	65	2	85	52
6	89	26	25	46	35	66	7	86	49
7	12	27	73	47	61	67	0	87	42
8	86	28	59	48	33	68	69	88	54
9	80	29	95	49	23	69	5	89	76
10	77	30	11	50	72	70	22	90	17
11	38	31	8	51	82	71	51	91	14
12	41	32	93	52	94	72	78	92	67
13	39	33	9	53	46	73	40	93	37
14	83	34	31	54	45	74	63	94	55
15	29	35	44	55	34	75	91	95	66
16	28	36	75	56	58	76	64	96	21
17	74	37	81	57	87	77	6		
18	32	38	57	58	1	78	43		
19	24	39	47	59	3	79	48		
20	70	40	30	60	4	80	90		

Figure 8 Feature Ranking using Trace Ratio Feature Selection

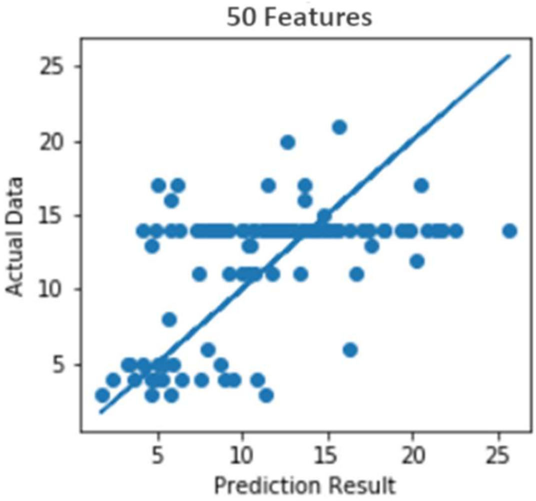


Figure 9 The Best Prediction result using Trace Ratio Feature Selection

5. Conclusions

In this study, we use an e-commerce data set, Naive Bayes, and several similarity-based feature selection algorithms to identify the order of most important features. From the evaluation it can be concluded that the similarity-based feature selection algorithm shows excellent performance on handling problems in the non-supervised field. Calculations that focus on building affinity matrices make this method easy and simple, and after calculating matrices, this method can get a score from each feature. Besides, this method does not depend on the learning algorithm.

The weakness of this algorithm is that it usually cannot handle the problem of feature redundancy, or it often finds highly correlated features during the selection phase. In the predicted results of each feature selection algorithm, R2 with the same RMSE is obtained because Naïve Bayes machine learning (Gaussian NB) would accept integer numbers and would not accept floating point numbers when processing Leave-One-Out. Therefore, data containing floating point numbers were automatically converted into integers. However, the feature ratings for each algorithm feature differ according to the algorithm. In this case, we obtain unsatisfactory results, thus another algorithm to ensure better results will be used in the next plan.

Bibliography

[1] N. Zuhdiyaty and D. Kaluge, “Analisis Faktor-faktor Yang Mempengaruhi Kemiskinan di Indonesia Selama Lima Tahun Terakhir,” *Jurnal Ilmiah Bisnis dan Ekonomi Asia*, vol. 11, no. 2, pp. 27–31, Sep. 2018, doi: 10.32812/jibeka.v11i2.42.

[2] Badan Pusat Statistik, “Jumlah Penduduk Miskin Menurut Wilayah (Juta Jiwa), 2021-2022.” https://bps.go.id/indicator/23/183/1/jumlah-penduduk-miskin-menurut-wilayah.html (accessed Aug. 24, 2022).

[3] Badan Pusat Statistik, “Sensus Penduduk 2020.” https://sensus.bps.go.id/main/index/sp2020 (accessed Aug. 24, 2022).

[4] B. P. Statistik, “Kemiskinan dan Ketimpangan,” *Badan Pusat Statistik*, 2019.

[5] Badan Pusat Statistik, *Statistik Kesejahteraan Rakyat 2021*. 2021. Accessed: Aug. 24, 2022. [Online]. Available: https://bps.go.id/publication/2021/11/19/36c2f9b45f70890edb18943d/statistik-kesejahteraan-rakyat-2021.html

[6] Diaz Praditya, “Prediksi Perkembangan Industri E-commerce Indonesia pada Tahun 2022,” *techinasia*, 2019.

[7] J. Blumenstock and N. Eagle, “Mobile divides: Gender, socioeconomic status, and mobile phone use in rwanda,” *ACM International Conference Proceeding Series*, 2010, doi: 10.1145/2369220.2369225.

[8] A. Perez, C. Yeh, G. Azzari, M. Burke, D. Lobell, and S. Ermon, “Poverty Prediction with Public Landsat 7 Satellite Imagery and Machine Learning,” no. Nips, 2017.

[9] C. Smith-Clarke, A. Mashhadi, and L. Capra, “Poverty on the cheap,” pp. 511–520, 2014, doi: 10.1145/2556288.2557358.

[10] J. Y. Sari, M. Muchtar, M. Zarkasi, and A. Z. Arifin, “Similarity Based Entropy on Feature Selection for High Dimensional Data Classification,” *Jurnal Ilmu Komputer dan Informasi*, vol. 7, no. 2, p. 101, 2014, doi: 10.21609/jiki.v7i2.263.

[11] G. Gavisiddappa, S. Mahadevappa, and C. Patil, “Multimodal Biometric Authentication System Using Modified ReliefF Feature Selection and Multi Support Vector Machine,” *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 1, pp. 1–12, 2020, doi: 10.22266/ijies2020.0229.01.

[12] B. Khotimah, M. Miswanto, and H. Suprajitno, “Optimization of Feature Selection Using Genetic Algorithm in Naïve Bayes Classification for Incomplete Data,” *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 1, pp. 334–343, 2020, doi: 10.22266/ijies2020.0229.31.

[13] P. Sharma and U. Bhardwaj, “Machine learning based spam E-mail detection,” *International Journal of Intelligent Engineering and Systems*, vol. 11, no. 3, pp. 1–10, 2018, doi: 10.22266/IJIES2018.0630.01.

[14] D. R. Wijaya, R. Sarno, and A. F. Daiva, “Electronic nose for classifying beef and pork using Naïve Bayes,” *Proceedings - 2017 International Seminar on Sensor, Instrumentation, Measurement and Metrology: Innovation for the Advancement and Competitiveness of the Nation, ISSIMM 2017*, vol. 2017-Janua, pp. 104–108, 2017, doi: 10.1109/ISSIMM.2017.8124272.

[15] D. R. Wijaya, R. Sarno, and E. Zulaika, “Sensor array optimization for mobile electronic nose: Wavelet transform and filter based feature selection approach,” *International Review on Computers and Software*, vol. 11, no. 8, pp. 659–671, 2016, doi: 10.15866/irecos.v11i8.9425.

[16] D. R. Wijaya and F. Afianti, “Stability Assessment of Feature Selection Algorithms on Homogeneous Datasets: A Study for Sensor Array Optimization Problem,” *IEEE Access*, vol. 8, pp. 33944–33953, 2020, doi: 10.1109/ACCESS.2020.2974982.

[17] D. R. Wijaya, N. L. P. S. P. Paramita, A. Uluwiyah, M. Rheza, A. Zahara, and D. R. Puspita, “Estimating city-level poverty rate based on e-commerce data with machine learning,” *Electronic Commerce Research*, 2020, doi: 10.1007/s10660-020-09424-1.

[18] C. E. Queiros and E. S. Gelsema, “on Feature Selection.,” *Proceedings - International Conference on Pattern Recognition*, vol. 50, no. 6, pp. 128–130, 1984, doi: 10.1145/3136625.

[19] INFORMATIKALOGI, “Algoritma Naive Bayes,” *informatikalogi.com*, 2017. https://informatikalogi.com/algoritma-naive-bayes/ (accessed Jan. 04, 2020).

[20] V. Ratnasari, “Pengoptimalan Naïve Bayes Dan Regresi Logistik Menggunakan Algoritma Genetika Untuk Data Klasifikasi,” p. 86, 2017.

[21] D. R. Wijaya, R. Sarno, and E. Zulaika, “Noise filtering framework for electronic nose signals: An application for beef quality monitoring,” *Comput Electron Agric*, vol. 157, no. December 2018, pp. 305–321, 2019, doi: 10.1016/j.compag.2019.01.001.