Knowledge Engineering and Data Science (KEDS)
Vol 6, No 2, October 2023, pp. 114–128

pISSN 2597-4602
eISSN 2597-4637

Multivariate Analysis Approach to Factor-Affected Tuberculosis Disease

Zuli Agustina Gultom [1], Farid Akbar Siregar [2], Mahardika Abdi Prawira Tanjung [3*], Al-Hamidy Hazidar [4]

Universitas Muhammdiyah Sumatera Utara, Jalan Kapten Muchtar Basri, Medan 20238, Indonesia
[1] zuliagustina@umsu.ac.id; [2] faridakbar@umsu.ac.id; [3] dika.abdi@gmail.com; [4] alhamidy@umsu.ac.id*
** corresponding author*

ARTICLE INFO

Article history:
Received 14 June 2023
Revised 14 September 2023
Accepted 29 September 2023
Published online 19 October 2023

Keywords:
Tuberculosis
Multivariate
Analysis
Surabaya

ABSTRACT

Tuberculosis is a disease caused by infection with the mycobacterium tuberculosis complex. Tuberculosis attack organ besides the lung, such as the pleura, lining of the brain, lining of the heart, lymph gland, bones, joint, skin, intestines, kidney, urinary tract, and genital. This disease is found in densely populated settlements with poor sanitation, lack of ventilation and sunlight and lack of rest. Moreover, the factors that will be analyzed in this research are Population Density (X1), Number of HIV/AIDS (X2), number of toddlers who experience nutrition (X3), Number of toddlers who experience BCG immunization (X4), number of toddlers who get exclusive breastfeeding (X5), Total families with PHBS (X6), number of residents with healthy homes (X7), number of families with clean water facilities (X8), number of families with ownership of latrine sanitation (X9), number of families with have landfills (X10), number of families have management waste place (X11), number of elementary education facilities (X12), Number of junior school education facilities (X13), Number of senior school education facilities (X14), Number of institutions fostered by neighborhood health (X15), Number of Posyandu (X16), Number Life Expectancy (X17), Literacy Rate (X18), Human Development Index (X19), Number of Tuberculosis sufferers (X20). This research aims to analyze what variables influence each other on the prevalence rate of tuberculosis in the city of Surabaya. The method used in this research is a multivariate analysis using factor analysis, cluster analysis, biplot analysis and discriminant analysis. This discriminant analysis determines accuracy by calculating the value (1-APER). The resulting research the Number of HIV/AIDS, number of residents with healthy homes, and Number of families with ownership of Sanitation (latrine, landfills, waste management) have a high correlation with the spread of tuberculosis in Surabaya. Meanwhile, areas with a high rate of tuberculosis are Tambaksari, Wonokromo, Sawahan, and Semampir. The classification analysis accuracy level was 90.32% and the accuracy of the resulting model or discriminant function was very high. So that discriminant analysis can be used for predicting the accuracy of tuberculosis prevalence rates.

I. Introduction

Tuberculosis is an infectious disease caused by Mycobacterium tuberculosis, which attacks organs other than the lungs [1]. This disease is a problem for developing countries with declining socio-economic conditions [2]. The prevalence rate of cases of pulmonary tuberculosis in Indonesia is 130/100,000 [3]. Every year, there are 539,000 new cases, and the number of deaths is around 101,000 people per year [4]. AFB (Acid Fast Bacilli) pulmonary tuberculosis (+) incidence is around 110/per 100,000 population [5]. TBC (Tuberculosis) is the third leading cause of death, after heart disease and respiratory disease [6]. According to [5], Indonesia is fifth after India, China, South Africa, and Nigeria.

The leading causes of increased tuberculosis problems are declining socio-economic conditions in people in developing countries [7], environmental conditions inside and outside the home that are very supportive for the occurrence of TB (Tuberculosis) disease [8], demographic changes due to the increasing world population and changes in the age structure of the population [9], the impact of the

https://doi.org/10.17977/um018v6i22023p114-128

HIV/AIDS pandemic [1]. The tuberculosis program has not been optimally implemented, which includes poor health infrastructure in countries that experienced an economic crisis, lack of implementation of tuberculosis services (less accessible to the public, not guaranteed provision of OAT, and non-standard monitoring, recording, and reporting [1].

The prevalence rate of tuberculosis is not only a medical problem; socio-economic conditions and environmental factors also have an influence [11]. For example, those with a low socioeconomic status will have a house in a slum area, an unhealthy house with a lack of air circulation, no sanitation, poor nutritional conditions, and a lack of clean water in their environment. According to research conducted by Sejati and Sofiana [12], people with family incomes below the minimum wage have a 1.123 times higher risk of being infected with TB(Tuberculosis) than those above the minimum wage. Factors that influence the prevalence rate of tuberculosis are Population Density, Number of HIV/AIDS, number of toddlers who experience nutrition, Number of toddlers who experience BCG (Bacillus Calmette Guerin) immunization, number of toddlers who get exclusive breastfeeding, Total families with PHBS (Clean and Healthy Living Behavior), number of residents with healthy homes, number of families with clean water facilities, number of families that have latrine sanitation, Number of families that have landfills, number of families that have waste management sites, number of Basic Education Facilities, Number of Middle School Education Facilities, Number of High School Education Facilities, Number of Institutions fostered by Environmental Health, Number of *Posyandu* (Integrated Healthcare Center), Expectation Rate Life, Literacy Rate, Human Development Index, Number of TB Sufferers.

Education level is one of the factors that influences the incidence of tuberculosis [13]. The higher a person's education level, the lower the incidence of tuberculosis [14], this happens because someone who has a good education will get more information and be able to absorb information about tuberculosis well and be able to treat it well. Apart from education, lighting or sunlight entering the house and the ventilation conditions of the house are also factors that influence the incidence of pulmonary tuberculosis [15].

Surabaya is the second largest city in Indonesia, with an area of approximately 326.37 km2; administratively, it is divided into 31 districts and 163 sub-districts with a population of approximately 2,912,197 people [16]. Based on [17], the highest tuberculosis in East Java is in Surabaya. At least 4,493 residents living in Surabaya have tuberculosis. This disease is found in densely populated settlements with poor sanitation, lack of ventilation and sunlight, and lack of rest [18]. TB cases in Surabaya are pretty significant compared to other cities [19], so there is a need for research or theoretical studies on the factors influencing the tuberculosis prevalence rate. Different characteristics, such as economic conditions and sociocultural factors in each region in Surabaya, will cause different health quality [20] so it is necessary to group areas with tuberculosis incidence characteristics. The goal of this research is to find out what factors influence the prevalence rate of tuberculosis in the city of Surabaya and to group regions based on the characteristics of the incidence of tuberculosis, with the hope that this research can help the Surabaya city government in handling tuberculosis prevalence rates quickly and accurately

The analysis technique used is multivariate analysis. This analysis is used to test more than two variables simultaneously. The multivariate analysis approach is divided into two main methods, namely dependency and interdependence [21]. This research was carried out using an interdependence approach. The types of multivariate analysis methods used are factor analysis, cluster analysis, biplot analysis and discriminant analysis. Factor analysis is used to reduce variables into new variables with fewer numbers. Cluster analysis groups observe areas based on the variable number of tuberculosis cases and the factors influencing them. Biplot analysis shows the closeness between objects, characteristics, or variables that characterize each object and the relationship between variables. Discriminant analysis was conducted to determine the differentiating variable and classification accuracy of the groupings obtained. All factors that will be examined are independent variables. The variables will be grouped into new variables, grouping sub-districts based on characteristics, knowing the mapping of the sub-district area and the accuracy of the classification of each factor used. The appropriate analysis in this research is multivariate analysis, by knowing what factors influence the prevalence rate of tuberculosis and knowing which areas have the number of tuberculosis sufferers, it is hoped that the government will be more responsive and quick in its handling of tuberculosis sufferers.

II. Methods

The analysis step for the method in this research present in Figure 1. The data used in this research is secondary data from the health services, Badan Pusat Statistika (BPS) and Badan Perencanan Pembanguna Kota Surabaya (BAPPEKO) [22]. The data taken is data related to the prevalence rate of tuberculosis in the city of Surabaya. The observation units studied were 31 sub-districts in the city of Surabaya. namely Krembengan, Gubeng, Tegalsari, Bubutan, Simokerto, Kenjeran, Tandes, Rungkut, Sukolilo, Mulyorejo, Sukomanunggal, Lakasantri, Gayungan, Genteng, Tenggilis, Karang Pilang, Wonocolo, Gunang Anyar, Dukuh Pakis, Jambangan, Bulak, Wiyung, Asemrowo, Benowo, Pakal, Sambikerep, Pabean Cantikan, Tambaksari, Wonokromo, Sawahan, Semampir.

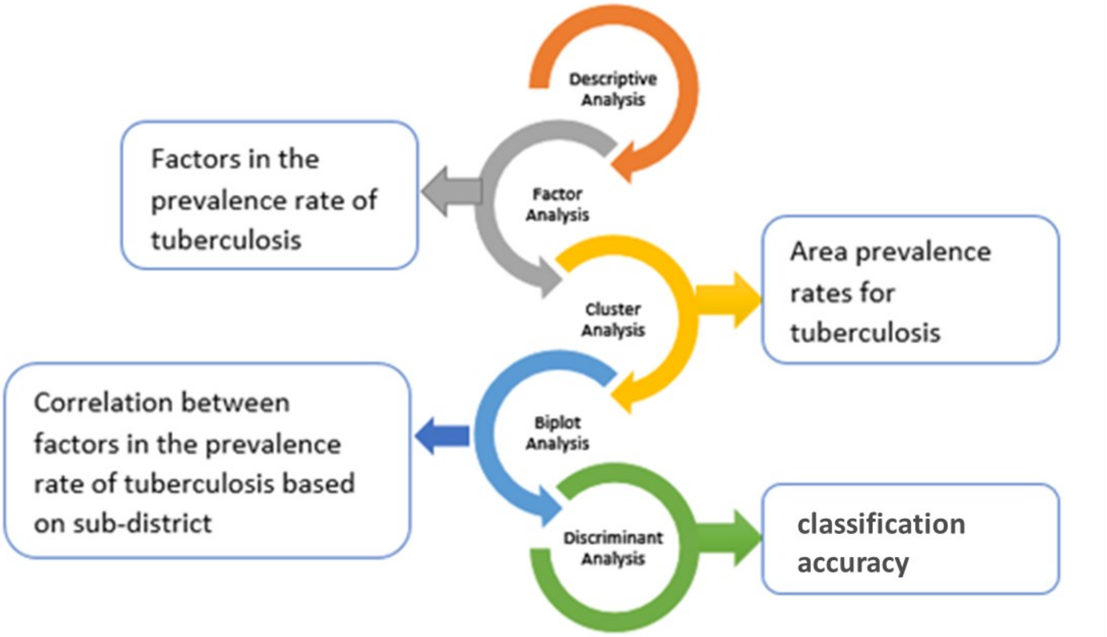

Descriptive Analysis
Factors in the prevalence rate of tuberculosis
Factor Analysis
Area prevalence rates for tuberculosis
Cluster Analysis
Correlation between factors in the prevalence rate of tuberculosis based on sub-district
Biplot Analysis
Discriminant Analysis
classification accuracy

Fig. 1. Analysis steps

The epidemiological factors for tuberculosis are BCG vaccination, inaccurate diagnosis, inadequate treatment, and control programs not implemented. Appropriately, endemic HIV infection, migration residents, self-medicate (self-treatment), increasing poverty, and services inadequate health [23]. A factor that is no less important in TB epidemiology is socioeconomic status, low income, low income, overcrowded housing, unemployment, and low education [24]. So the variables that will be examined in this research are Population Density (X1), Number of HIV/AIDS (X2), number of toddlers who experience nutrition (X3), Number of toddlers who experience BCG immunization (X4), number of toddlers who get exclusive breastfeeding (X5), Total families with PHBS (Clean and Healthy Living Behavior) (X6), number of residents with healthy homes (X7), number of families with clean water facilities (X8), Number of families that have latrine sanitation (X9), Number of families that have landfills (X10), Number of families that have waste management sites (X11), number of elementary education facilities (X12), Number of Middle school education facilities (X13), Number of High school education facilities (X14), Number of institutions fostered by neighborhood health (X15), Number of Posyandu (X16), Number Life Expectancy (X17), Literacy Rate (X18), Human Development Index (X19), Number of Tuberculosis sufferers (X20).

The clustering method used is Single Linkage, Complete Linkage, Average Linkage and Ward's Method. The Single Linkage method determines the distance between clusters by knowing the distance between two existing clusters and then choosing the closest distance or close neighbor rule [25]. The Complete linkage method (farthest-neighbor method) is used for the furthest inter-cluster distance (farthest-neighbor) between two objects in different clusters [26]. Ward's method aims to obtain clusters with the smallest possible cluster internal variance [27]. This method is very commonly used in determining clusters. This method is obtained by calculating the average value of each cluster and then calculating the Euclidean distance between each object.

III. Result and Discussion

Figure 2 shows a map of the number of tuberculosis patients in Surabaya. Marked in purple is the sub-district group that has the lowest number of tuberculosis, namely ranging from which ranges from 61 to 114, with the sub-districts of Tandes, Sukomanunggal, Customs Cantikan, Bubutan, Simokerto, Genteng, Tegalsari, Gubeng, Wonokromo, Wonokolo, Rungkut, Sukolilo. The color brown indicates the classification of the sub-district with the highest number of tuberculosis, ranging from 114 to 201, with the sub-district of Sawahan, Krembengan, Semampir, Kenjeran, and 16 to 61, with the Districts of Pakal, below, Asemrowo, Sambikerep, Lakasntri, Dukuh Pakis, Wiyung, Karang Pilang, pots, Gayungan, Gunung Anyar, Mulyorejo, Bulak, Tenggilis Mejoyo. White color is the classification of sub-districts with tuberculosis in the moderate category, Tambaksari.

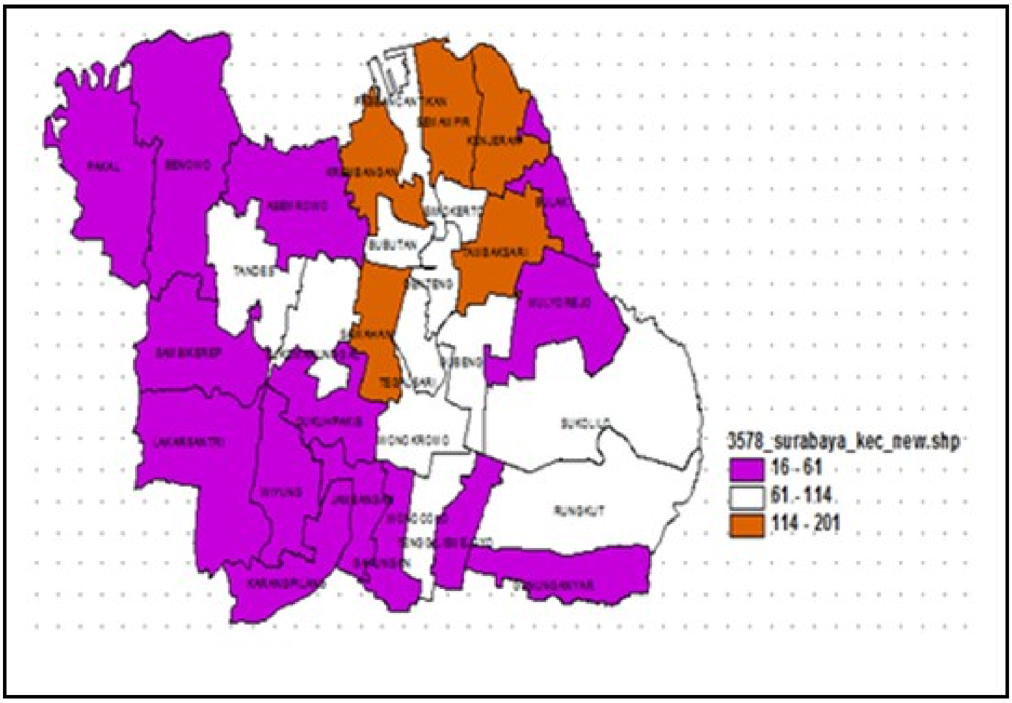


Fig. 2. A map of the number of tuberculosis patients in Surabaya

Reduce data dimensions that can explain as much as possible the diversity of data with several sets of variables that are fewer than the initial variable without losing the important information contained therein.

The inter-correlation test uses the Barlett test and data adequacy with KMO. The Kaiser–Meyer–Olkin (KMO) test is a statistical measure to determine how suited data is for factor analysis [28]. The test measures sampling adequacy for each variable in the model and the complete model. The statistic measures the proportion of variance among variables that might be common variance. The higher the proportion, the higher the KMO value, and the more suited the data is to factor analysis [29]. The following is the correlation testing hypothesis.

H0 : $\rho = I$ (between variables from the data of the factors that influence tuberculosis disease are not correlated)

H1 : $\rho \neq I$ (between variables from the data of the factors that influence tuberculosis disease are correlated)

Table 1. Correlation test and data adequacy

Method	Correlation test and data adequacy	
	Test Statistic	**Value**
KMO (Kaiser Mayor Olkin)		0.771
Bartlett's Test	Approx. Chi-Square	649.145
	Df	190
	Sig.	0.000

Table 1 shows that the Chi-Square value of the factors that influence tuberculosis is 649.145 with a P_value of 0.000. It was decided that P_value rejects H0, because the value of P value (0.000) < α (0.05). So it can be concluded that there is a correlation between the data variable that affect tuberculosis. The KMO value of the data is 0.771. From this value, it can be decided that it failed to reject H0, because the value of KMO (0.771) > 0.5, which means that the data on the factors that influence tuberculosis have accepted the data adequacy test to be analyzed further.

Table 2. Initial eigenvalues

Component	Initial Eigenvalues		
	Total	**Varian (%)**	**Cumulative variance**
1	10,856	54,281	54,281
2	2,473	12,363	66,644
3	1,310	6,548	73,452
4	1,033	5,165	78,357

From Table 2, it is known that there are four mutually independent factors, with a cumulative variance of 78.357%. The variable is divided into certain factor groups by selecting the most considerable loading factor value between loadings 1, 2, 3 and 4. The loading factor used is the loading factor, which is rotated varimax.

Table 3. Loading factor

Variable	Factor 1	Factor 2	Factor 3	Factor 4
X2	**0,829**	0,117	0,061	-0,021
X4	**0,742**	0,445	0,191	0,116
X7	**0,813**	0,291	0,018	0,279
X8	**0,601**	0,208	0,144	0,586
X9	**0,875**	0,364	0,001	0,249
X10	**0,801**	0,364	0,036	0,366
X11	**0,839**	0,365	0,068	0,300
X16	**0,779**	0,459	0,165	0,088
X20	**0,622**	0,505	-0,115	0,455
X1	0,126	**0,663**	0,070	0,396
X5	0,444	**0,583**	-0,173	-0,191
X6	0,307	**0,813**	-0,046	0,064
X12	0,497	**0,687**	0,108	0,417
X13	0,458	**0,668**	0,266	0,245
X14	0,398	**0,620**	0,387	-0,085
X15	0,561	**0,582**	0,128	-0,062
X17	-0,082	-0,064	**0,846**	-0,088
X18	0,054	0,060	**0,823**	-0,002
X19	0,220	0,164	**0,820**	-0,005
X3	0,189	0,027	-0,189	**0,879**

Table 3 shows the variable grouped in factor 1 have HIV/AIDS, the number of children under five who received BCG immunization, the number of residents who have healthy homes, families who have clean water facilities, the number of families with ownership of sanitation (latrines, landfills, Waste Management Sites), Number of Posyandu, Number of TB Patients. Factor 1 reviews the quality of a person's health. Factor 1 is very prominent in the development of the spread of tuberculosis in Surabaya. Factor 2 includes population density, exclusive breastfeeding, clean and healthy living behavior (PHBS), and educational facilities (elementary school, junior high school, senior high school). Factor 2 reviews demography and education. Factor 3 includes Life Expectancy, Literacy Rate, and Human Development Index. Factor 3 reviews the Human Development Index. Factor 4 includes the number of toddlers who experience nutrition.

The Cluster Analysis that will be explicitly used in this study is Ward's Linkage method with Square Euclidian Distance. In Figure 3, the dendrogram is cut into four groups, and the 31 subdistricts in Surabaya are grouped as in Table 4.

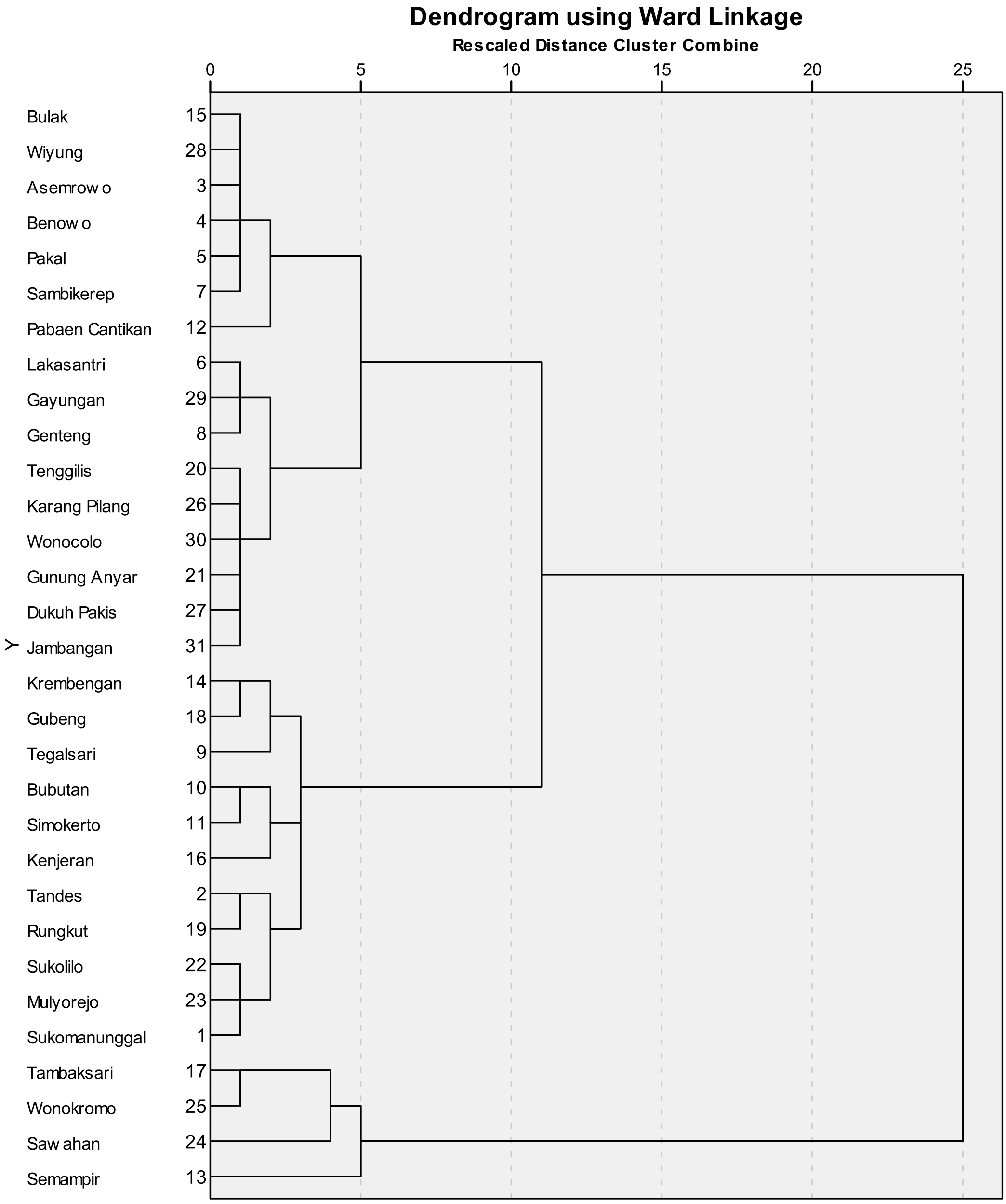


Fig. 3. A map of the number of tuberculosis patients in Surabaya

Table 4 shows that Group 1 consists of 11 sub-districts, group 2 consists of 9 sub-districts, group 3 consists of 7 sub-districts, and Group 4 consists of 4 sub-districts, namely Tambaksari, Wonokromo, Sawahan, Semampir sub-districts.

Table 4. Results of subdistrict grouping in Surabaya city

Group	District
1	Krembengan, Gubeng, Tegalsari, Bubutan, Simokerto, Kenjeran, Tandes, Rungkut, Sukolilo, Mulyorejo, Sukomanunggal
2	Lakasantri, Gayungan, Genteng, Tenggilis, Karang Pilang, Wonocolo, Gunang Anyar, Dukuh Pakis, Jambangan
3	Bulak, Wiyung, Asemrowo, Benowo, Pakal, Sambikerep, Pabean Cantikan
4	Tambaksari, Wonokromo, Sawahan, Semampir.

The Figure 4 is a picture of health, demographics and education, HDI and nutrition. Figure 4 shows that the Krembengan, Semampir, Sukolilo, Kenjeran, Wonokromo, and Tambaksari districts have high-quality health, high demography and high education. Tegalsari, Mulyorejo, Pabean Cantikan, Genteng, Simokerto, Gubeng, Wonocolo, and Sukomanunggal sub-district have high demographics and education, while the quality of health is low. Tenggilis, Sambikerep, Dukuh Pakis, Benowo, Pakal, Jambangan, Gayungan, Lakasantri, Bulak, and Asemrowo sub-district have the characteristic of low health quality, low demography and low education. Rungkut, Tandes, Gunung Anyar, Karang Pilang, and Sawahan sub-districts have high-quality health, high demography and low education.

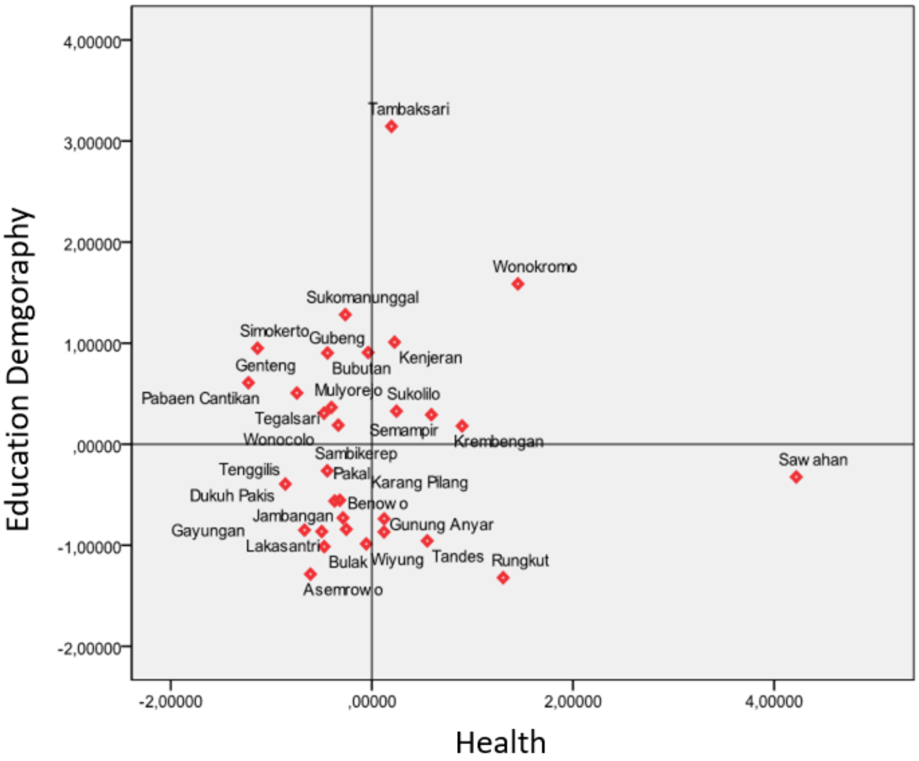


Fig. 4. Relationship between Health and Education Demographic

Figure 5 have Tegalsari, Tendes, Tenggilis, Krembengan, Dukuh Pakis, and Gubeng sub-district have high HDI characteristics and nutritional deficiencies. Semampir, Kenjeran, Simokerto, Bubutan, and Asemrowo sub-districts have high HDI (Human Development Index) characteristics and low malnutrition. Pabean Cantikan, Wiyung, Bulak, Karang Pilang, Mulyorejo, Jambangan, Gununganyar, Benowo, Pakal, Sukomanunggal, and Sambikerep districts have regional characteristics of HDI and low malnutrition. Genteng, Gayungan, Lakasantri, Wonokromo, Tambaksari, and Sawahan sub-districts have high HDI characteristics and low malnutrition.

Biplot analysis of that area has been formed to find out the sub-district mapping seen from the tendency of the variable that influences tuberculosis.

Figure 6 shows that the variable waste management sites (X11) have the most incredible diversity because the vector length is the longest among the others. At the same time, the nutrition variable (X3) has a minor diversity or tends to be homogeneous because the vector length is the shortest. Variables that have a positive correlation are the number of toddlers who experience nutrition (X3), literacy rate (X18), clean water facilities (X8), HIV/AIDS (X2), healthy homes (X7), latrine sanitation (X9), landfills (X10), waste management sites (X11), BCG immunization (X4), number of posyandu (X16), environmental health development institution (X15), number of TB sufferers (X20), exclusive breastfeeding (X5), elementary education facilities (X12), junior high school education facilities (X13), senior high school education facilities (X14), population density (X1), clean and healthy living behavior (X6), HDI (X19). At the same time, the variable that has a negative correlation is life expectancy (X17).

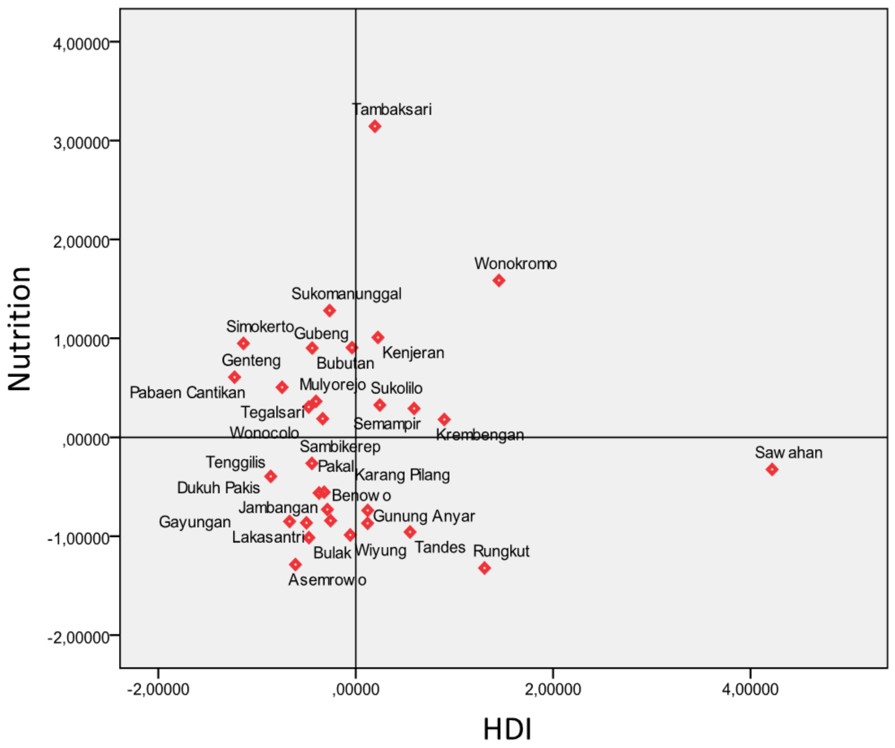


Fig. 5. Relationship between Human Development Index (HDI) and Nutrition

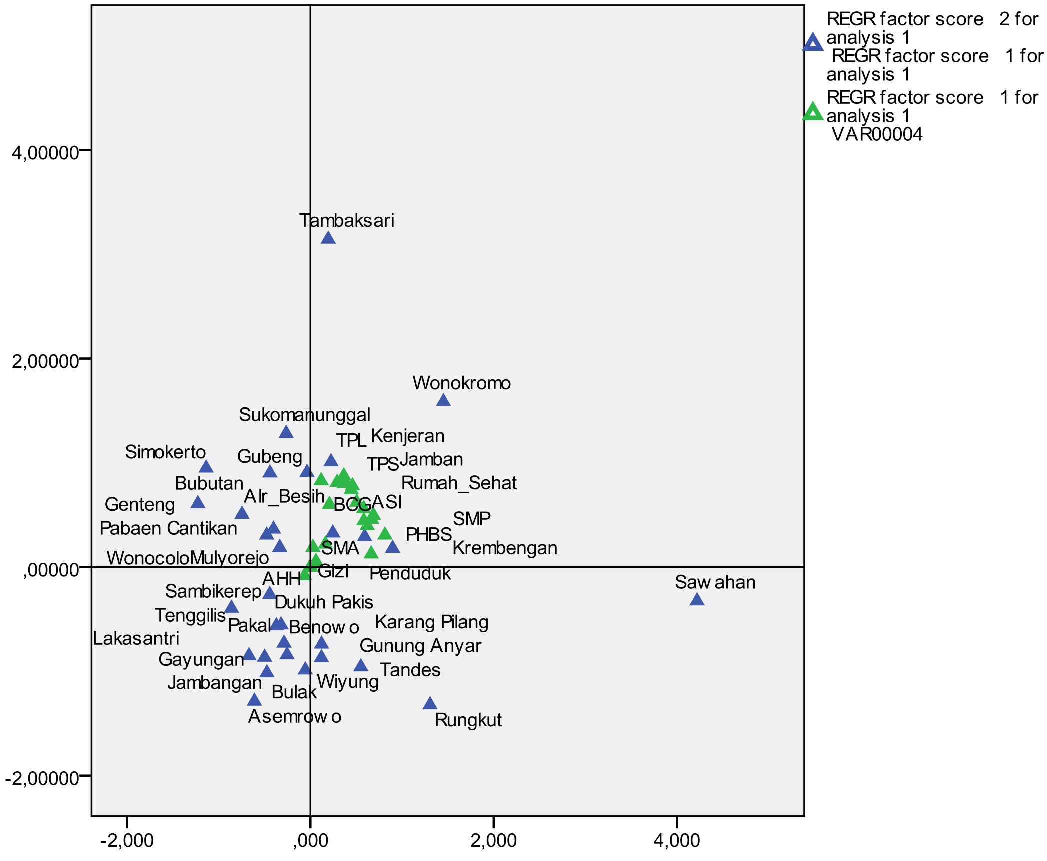


Fig. 6. Biplot between variable 1 factor

Variable waste management sites, landfills, latrine sanitation , healthy homes, BCG immunization, exclusive breastfeeding, clean and healthy living behavior, elementary, junior high school education facilities, population, senior high school education facilities, clean water facilities, nutrition, literacy rate, HDI, HIV'AIDS, number of posyandu, health development institutions, the number of TB contributes a lot to the sub-district of Tambaksari, Wonokromo, Kenjeran, Semampir, Krembengan. Life expectancy variable contributes to the sub-district of Asemrowo, Bulak, Jambang, Gayungan, Lakasantri, Pakal, Tenggilis, Benowo, Dukuh Pakis, and Sambikerep. Meanwhile, the sub-district of Sawahan, Rungkut, Tandes, Gunung Anyar, Karang Pilang, Benowo, Tenggilis Mejoyo, Pabaen Cantikan, Genteng, Bubutan, Tandes, Tegalsari, Simokerto, Gubeng, Sukomanunggal do not dominate the variables that affect tuberculosis.

Figure 7 shows that life expectancy (X17) has the most incredible diversity because the length of the vector is the longest among the others. The population variable (X1) has a minor vector diversity or tends to be homogeneous because the vector length is the shortest. Variables that have a positive correlation are the human development index (X19), junior high school education facilities (X13), senior high school (X14), population density (X1), health development institutions (X15), number of *posyandu*-Integrated Healthcare Center (X16), HIV/AIDS (X2), BCG immunization (X4), waste management site (X11), latrine sanitation (X9), healthy homes (X7), landfills (X10), clean water facilities (X8), exclusive breastfeeding (X5), elementary education facilities (X12), number of toddlers who experience nutrition (X3), literacy rate (X18), number of TB sufferers (X20), clean and healthy living behavior (X6). The variable that has a negative correlation is life expectancy (X17).

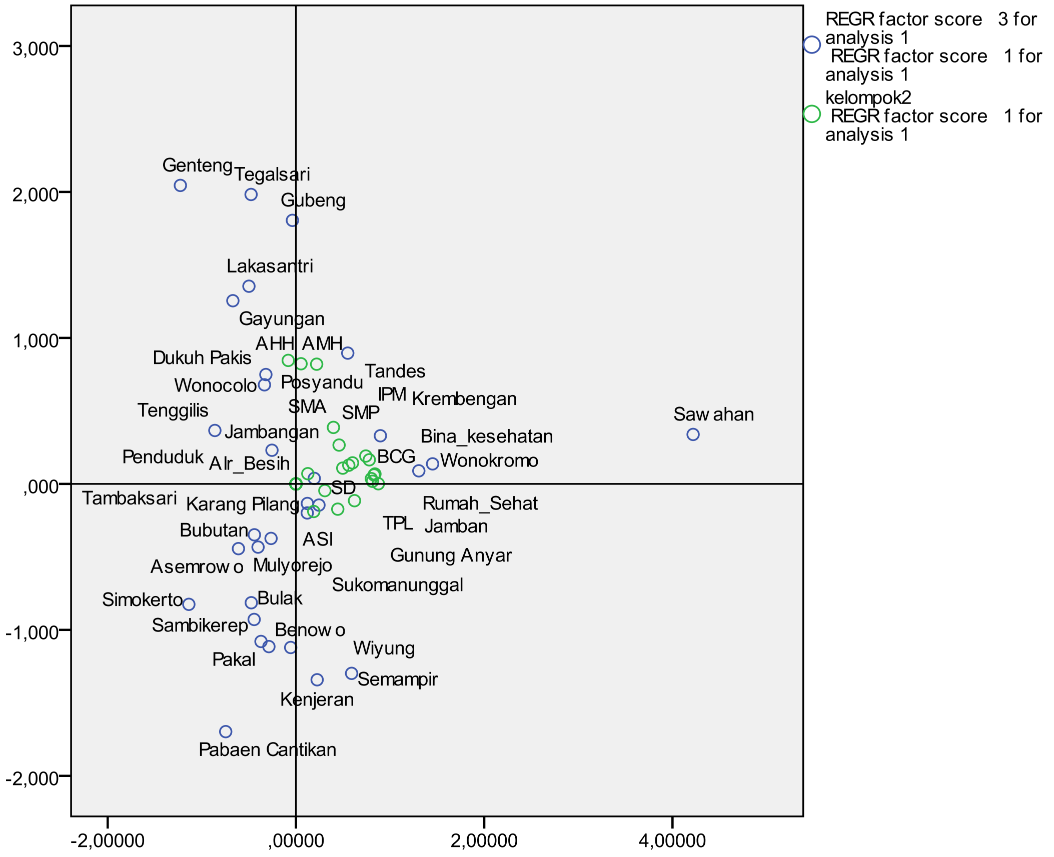


Fig. 7. Biplot between variable 2 factor

Variables Human development index, educational facilities for junior and senior high schools, population density, a health development institution, number of *posyandu*, HIV/AIDS, number of babies immunized with BCG, waste management sites, latrine sanitation, healthy homes, landfills, number of families with facilities clean water, the number of babies who are exclusive breastfeeding, elementary school education facilities, number of toddlers who experience nutrition, the literacy rate, the number of TB sufferers, the number of clean and healthy living behavior have contributed a lot to the sub-district of Sawahan, Tandes, Krembengan, Wonokromo, Rungkut, Sukolilo, Tambaksari, Gunung Anyar, Kenjeran, Semampir, Sukomanunggal. The life expectancy variable significantly contributes to the sub-district of Gubeng, Tegalsari, Genteng, Lakasantri, Gayungan, Wonocolo, Dukuh Pakis, Jambangan, and Tenggilis Mejoyo. Whereas for the district of Bubutan, Asemrowo, Karang pilang, Mulyorejo, Bulak, Simokerto, Pakal, Benowo, Pabean Cantikan, Wiyung, Sambikerep did not dominate the variables that affect tuberculosis.

Figure 8 shows that the nutritional variable (X3) has the greatest diversity because the vector length is the longest among the others. The development index has the smallest diversity because it has the shortest vector. positive correlation with number of toddlers who experience nutrition (X3), clean water facilities (X8), healthy homes (X7), elementary education facilities (X12), junior high schools (X13), senior high school (X14), latrine sanitation (X9), landfills (X10), waste management sites (X11), HIV/AIDS(X2), health development institution (X15), exclusive breastfeeding (X5), number of TB (X20), population density (X1), BCG immunization (X4), clean and healthy living behavior (X6), HDI (X19), literacy rate (X18), number of posyandu (X16). The life expectancy variable (X17) negatively correlates with the factors influencing tuberculosis.

Subdistricts of Tandes, Semampir, Kenjeran, Rungkut, Krembengan, Sukolilo, Gunung Anyar, Sawahan, Tambaksari, and Wonokromo have contributed to the variables of nutrition, clean water facilities, healthy homes, educational facilities for elementary school, junior high school, senior high school, latrine sanitation, landfills, waste management, number of people living with HIV/AIDS, health development institutions, exclusive breastfeeding, number of TB, population density, BCG immunization, clean and healthy living behavior, HDI, literacy rate, number of *posyandu*. Karang Pilang, Jambangan, Sukomanunggal, Wiyung, Benowo, Sambikerep, Lakasantri, Pakal, Bulak, Pabean Cantikan, and Genteng contribute significantly to the life expectancy variable. The districts of Gubeng, Tegalsari, Asemrowo, Bubutan, Wonocolo, Mulyorejo, Gayungan, Simokerto, and Tenggilis Mejoyo have no contribution to the factors that influence tuberculosis.

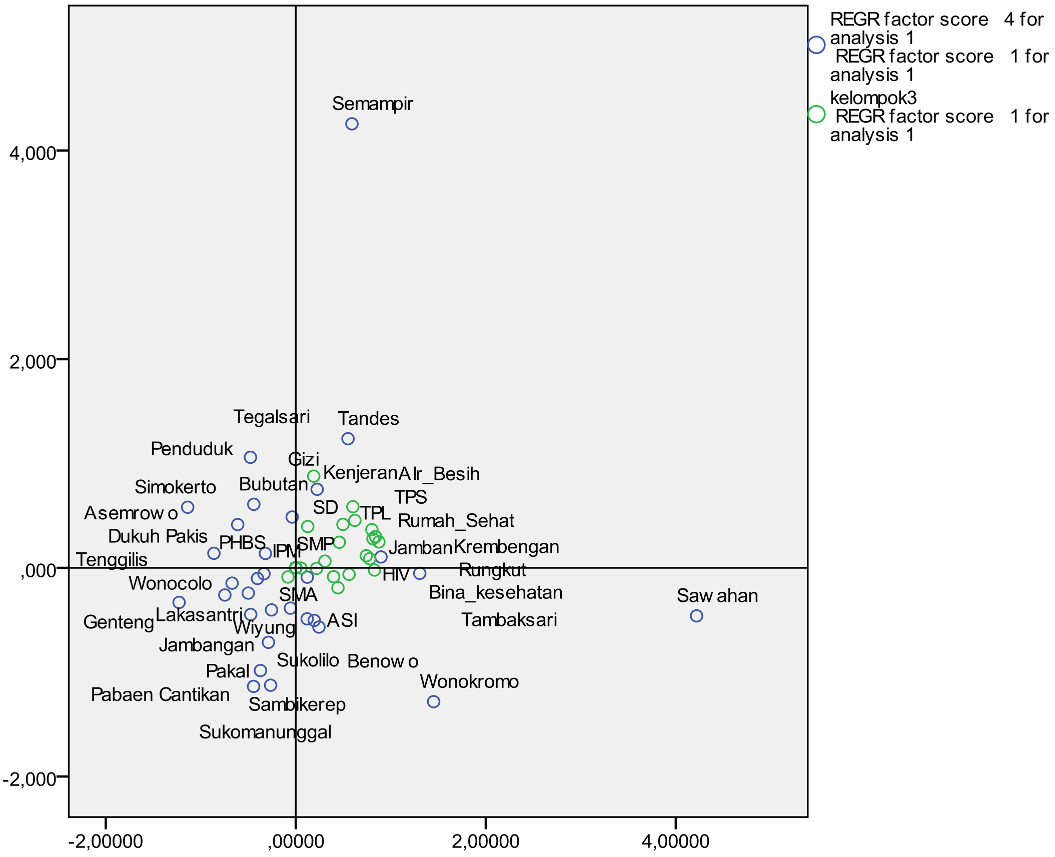
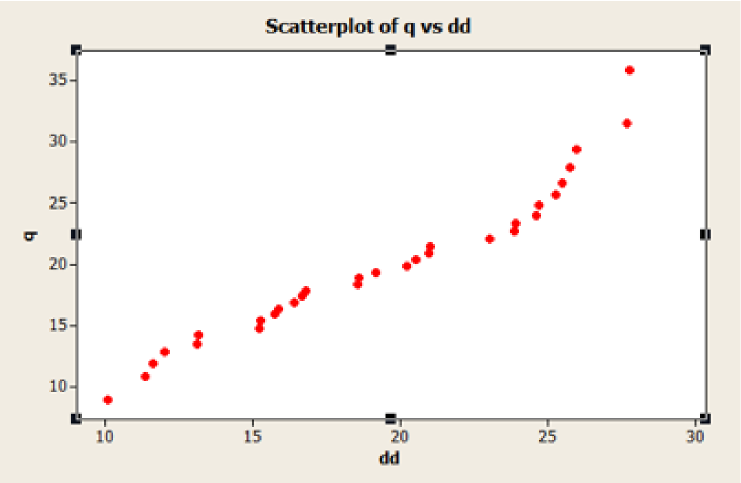


Fig. 8. Biplot between variable 3 factor

Before proceeding to discriminant analysis, the multivariate normal assumptions and assumption of homogeneity of covariance variant matrix. The average multivariate assumptions are tested to determine whether the data used is usually distributed [30]. The primary requirement in conducting multivariate analysis is that data is multi-normally distributed.

From Figure 9, it can be concluded that the data is usually distributed. Visually, the QQ plot tends to form a straight line so that it can be concluded that the data assumptions follow a multivariate normal distribution and have been accepted. The results of the covariance variant matrix can be seen in Table 5.



```
t     0,516129
Data Multinormal Distribution
```

Fig. 9. Multinormally distributed

Test the homogeneity of the covariance variant matrix using the Box'M test statistic with the hypothesis:

$$H_0 : \sum_1 = \sum_2 = \sum_3 = \sum_4$$

$$H_0 : \text{at least one is different } \sum_j$$

Table 5. The results of the covariance variant matrix

Method	Covariance Variant Matrix	
	Test Statistic	Value
Box'M		60.093
F	Approx. Chi-Square	2.211
	df1	190
	df2	1508.643
	Sig.	0.002

Reject H_0 if the P_value is less than 0.05 (this study uses a 95% confidence level). From the test results, it can be concluded that the data analyzed have the same covariance matrix. Then, the discriminant analysis can be continued.

From the discriminant analysis using the stepwise method, the following Table 6 shows the obtained results. Table 6 shows that 19 variables are confirmed in the grouping, and only four variables meet the criteria as differentiators. These variables are elementary education, landfills, exclusive breastfeeding, and literacy rate (AMH). So, it can be concluded that the groups distinguishing tuberculosis in Surabaya are education, exclusive breastfeeding, literacy and sanitation.

Table 6. The results of the covariance variant matrix

Variable	Wilks'Lambda
Elementary education	0,179
Landfills	0,224
Exclusive breastfeeding	0,511
Literacy Rate	0,696

Table 7. The function of the discriminant equation

Variable	Function 1	Function 2	Function 3
Exclusive breastfeeding	0,667	-0,021	0,827
Landfills	0,714	0,376	-0,311
Elementary Education	0,867	-0,337	-0,105
Literacy rate	-0,224	1.019	0.265

After obtaining the discriminant equation, it is obtained in the discriminant equation function, as shown in Table 7. Based on Table 7, the function of the discriminant equation can be described as follow.

Function 1 = 0.667 Exclusive breastfeeding + 0.714 landfills + 0.867 elementary education – 0.224 literacy rate

Function 2 = -0.021 Exclusive breastfeeding + 0.376 landfills – 0.337 elementary education + 1.019 literacy rate

Function 3 = 0.827 Exclusive breastfeeding – 0.311 landfills – 0.105 elementary education + 0.265 literacy rate

The variable with the most significant coefficient contributes to differentiating groups. Based on the above, function one shows that elementary school education is a factor that plays a role in distinguishing the first and second groups. The second function shows that the literacy rate variable significantly differentiates the second and third groups. The third function shows that the exclusive breastfeeding variable has a role in differentiating the third and fourth factors.

Figure 10 shows that the grouping based on the discriminant function is correct because not all group members are spread around the centroid point of the group. In Group 1, there are members of Group three who enter Group One. In Group 2, two members enter Group 3. Groups 3 and 4 are around the group centroid point. In determining the results of discriminant analysis, the results of the total accuracy value (1-APER) are needed which are based on the classification table (Hosmer & Lemeshow, 2000).

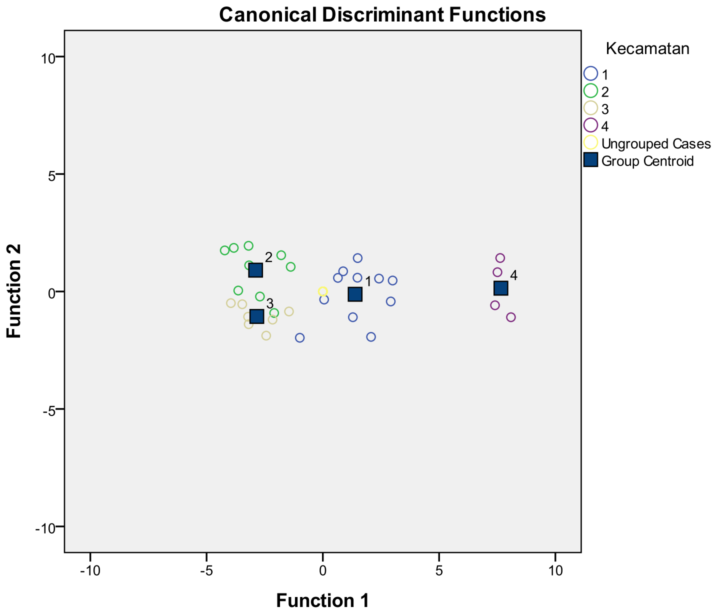


Fig. 10. Plot of the discriminant function

Table 8 shows that the accuracy of the classification result for the four groups that have been formed is 0.9032 or 90.32%. There was a classification error in the grouping of variables that affected tuberculosis in Surabaya at 0.9032. so the APER value is known to be 0.0968, which means the error level in the data using discriminant analysis is 0.0968. There is an incorrect unit of observation (district) in the grouping. There is one observation in group 3 (Simokerto) that must be included in the first row of actual group 1. That is, if we look at the cluster analysis results, Simokerto is in Group 1 (by grouping TB patient numbers by region). Based on cluster analysis on actual data, the grouping of the Gunung Anyar area is in Group 2. In the observations, there is Group Three (Gunung Anyar). Wiyung, based on actual data in cluster analysis, they are in Group 3, while the observation group is in row two.

Table 8. Accuracy of classification

The Real Group	Predicted group			
	1	2	3	4
1	10	0	1	0
2	0	7	2	0
3	0	0	7	0
4	0	0	0	4
Accuracy	0.9032			

IV. Conclusion

The study's extensive analysis has yielded significant findings regarding the tuberculosis situation in Surabaya. It is worth mentioning that the sub-districts exhibiting the most significant tuberculosis burden have been identified as Sawahan, Krembengan, Semampir, Kenjeran, and Tambaksari. The categorization as mentioned above, plays a crucial role as an essential initial step in developing focused intervention tactics within these domains.

Significantly, the study has shed light on the complex relationship between an individual's health status and the spread of tuberculosis prevalence in Surabaya—the discovery, as mentioned earlier results from a rigorous factor analysis, which considered multiple variables. The factors examined in this study included the population of individuals living with HIV/AIDS, the rate of immunization coverage among toddlers for BCG, the prevalence of households with adequate living conditions, access to clean water facilities, availability of sanitation facilities such as latrines and waste disposal sites (TPS), the provision of posyandu services, and the incidence of tuberculosis cases. The convergence of these factors has shown the complex network of elements that contribute to the tuberculosis situation in the city.

Furthermore, the study has identified regions exhibiting a pronounced susceptibility to the exacerbation of tuberculosis prevalence. Tambaksari, Wonokromo, Sawahan, and Semampir Districts have been identified as areas of significant concern. This conceptualization of vulnerability enables the implementation of proactive actions in various domains, which may encompass the intensification of healthcare provision, the dissemination of public health awareness campaigns, and the improvement of healthcare facility accessibility.

The study's implementation of discriminant analysis has produced a noteworthy degree of precision, surpassing the criterion of 0.5. The discriminant technique has demonstrated a noteworthy accuracy of 90.32% in predicting tuberculosis prevalence data. The translation results in a meager error rate of only 0.0968, which highlights the strong performance of the model utilized in this study.

However, although this study represents a substantial advancement in our comprehension of the prevalence of tuberculosis in Surabaya, it also emphasizes the necessity for more investigation. In order to enhance our understanding and improve the effectiveness of intervention approaches, it is recommended that future analyses consider including supplementary health variables that were not within the purview of this study. In addition, it is essential to conduct comparisons with various analytical methodologies in order to determine their effectiveness and accuracy, ensuring the utilization of the most efficient strategies to address the issue of tuberculosis in Surabaya. This study establishes the groundwork for a more comprehensive and efficient approach to addressing the issue of tuberculosis prevalence in urban areas

Declarations

Author contribution

All authors contributed equally as the main contributor of this paper. All authors read and approved the final paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare no known conflict of financial interest or personal relationships that could have appeared to influence the work reported in this paper.

Additional information

Reprints and permission information are available at http://journal2.um.ac.id/index.php/keds.

Publisher's Note: Department of Electrical Engineering and Informatics - Universitas Negeri Malang remains neutral with regard to jurisdictional claims and institutional affiliations.

References

[1] A. A. Chackerian, J. M. Alt, T. V Perera, C. C. Dascher, and S. M. Behar, "Dissemination of Mycobacterium tuberculosis is influenced by host factors and precedes the initiation of T-cell immunity," Infect. Immun., vol. 70, no. 8, pp. 4501–4509, 2002.

[2] M. Buheji et al., "The extent of covid-19 pandemic socio-economic impact on global poverty. a global integrative multidisciplinary review," Am. J. Econ., vol. 10, no. 4, pp. 213–224, 2020.

[3] E. A. Wikurendra, N. Herdiani, Y. G. Tarigan, and A. A. Kurnianto, "Risk factors of pulmonary tuberculosis and countermeasures: A literature review," Open Access Maced. J. Med. Sci., vol. 9, no. F, pp. 549–555, 2021.

[4] S. Andarmoyo, "The Effect Of Home Ventilation On The Incidence Of Lung TuberculosisIn Ponorogo Regency," Dev. Nurs. Curric. to Improv. Qual. Nurs. Educ. Islam. values Int. Perspect., pp. 85–91, 2015.

[5] W. H. Organization, "Tuberculosis surveillance and monitoring in Europe 2021: 2019 data," 2021.

[6] A. L. Byrne, B. J. Marais, C. D. Mitnick, L. Lecca, and G. B. Marks, "Tuberculosis and chronic respiratory disease: a systematic review," Int. J. Infect. Dis., vol. 32, pp. 138–146, 2015.

[7] L. C. Rodrigues and P. G. Smith, "Tuberculosis in developing countries and methods for its control," Trans. R. Soc. Trop. Med. Hyg., vol. 84, no. 5, pp. 739–744, 1990.

[8] W. H. Organization, "Intersectoral collaboration to end HIV, tuberculosis and viral hepatitis in Europe and central Asia: a framework for action to implement the United Nations Common Position," 2020.

[9] J. B. Dowd et al., "Demographic science aids in understanding the spread and fatality rates of COVID-19," Proc. Natl. Acad. Sci., vol. 117, no. 18, pp. 9696–9698, 2020.

[10] T. Togun, B. Kampmann, N. G. Stoker, and M. Lipman, "Anticipating the impact of the COVID-19 pandemic on TB patients and TB control programmes," Ann. Clin. Microbiol. Antimicrob., vol. 19, no. 1, pp. 1–6, 2020.

[11] A. K. Kashyap and J. C. Stein, "Monetary Policy When the Central Bank Shapes Financial-Market Sentiment," J. Econ. Perspect., vol. 37, no. 1, pp. 53–75, 2023.

[12] A. Sejati and L. Sofiana, "Faktor-faktor terjadinya tuberkulosis," KEMAS J. Kesehat. Masy., vol. 10, no. 2, pp. 122–128, 2015.

[13] A. Mollalo, L. Mao, P. Rashidi, and G. E. Glass, "A GIS-based artificial neural network model for spatial distribution of tuberculosis across the continental United States," Int. J. Environ. Res. Public Health, vol. 16, no. 1, p. 157, 2019.

[14] S. K. Singh, G. C. Kashyap, and P. Puri, "Potential effect of household environment on prevalence of tuberculosis in India: evidence from the recent round of a cross-sectional survey," BMC Pulm. Med., vol. 18, no. 1, p. 66, 2018.

[15] I. Yuniar, S. Rusmindarti, and S. Sarwono, "The Level of Lighting and Ventilation on the Incidence Rate of Pulmonary TB," 2021.

[16] S. C. N. Tang, M. Rusli, and P. Lestari, "Climate Variability and Dengue Hemorrhagic Fever in Surabaya, East Java, Indonesia," 2019.

[17] N. N. Juliasih, N. M. Mertaniasih, C. Hadi, Soedarsono, R. M. Sari, and I. N. Alfian, "Factors affecting tuberculosis patients' quality of life in Surabaya, Indonesia," J. Multidiscip. Healthc., pp. 1475–1480, 2020.

[18] P. Pardeshi et al., "Association between architectural parameters and burden of tuberculosis in three resettlement colonies of M-East Ward, Mumbai, India," Cities Heal., vol. 4, no. 3, pp. 303–320, 2020.

[19] D. S. Rachmawati, N. Nursalam, M. Amin, and R. Hargono, "Developing Family Resilience Models: Indicators and Dimensions in the Families of Pulmonary TB Patients in Surabaya," 2019.

[20] S. Hawken and R. Y. Sunindijo, "City of Kampung: risk and resilience in the urban communities of Surabaya, Indonesia," Int. J. Build. Pathol. Adapt., vol. 36, no. 5, pp. 543–568, 2018.

[21] Y. Yulianto, N. Robihaningrum, and B. D. Elinda, "Management Multivariate Analysis Methods for Variables Measurement in Scientific Papers," Aptisi Trans. Manag., vol. 3, no. 1, pp. 65–72, 2019.

[22] B. P. Statistik, "Statistik lingkungan hidup indonesia," Jakarta. BPS Indones., 2018.

[23] D. Sharma, J. Sharma, N. Deo, and D. Bisht, "Prevalence and risk factors of tuberculosis in developing countries through health care workers," Microb. Pathog., vol. 124, pp. 279–283, 2018.

[24] R. Duarte et al., "Tuberculosis, social determinants and co-morbidities (including HIV)," Pulmonology, vol. 24, no. 2, pp. 115–119, 2018.

[25] Ö. Akay and G. Yüksel, "Clustering the mixed panel dataset using Gower's distance and k-prototypes algorithms," Commun. Stat. Comput., vol. 47, no. 10, pp. 3031–3041, 2018.

[26] A. E. Ezugwu et al., "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," Eng. Appl. Artif. Intell., vol. 110, p. 104743, 2022.

[27] S. Sharma and N. Batra, “Comparative study of single linkage, complete linkage, and ward method of agglomerative clustering,” in 2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon), 2019, pp. 568–573.

[28] N. Shrestha, “Factor analysis as a tool for survey analysis,” Am. J. Appl. Math. Stat., vol. 9, no. 1, pp. 4–11, 2021.

[29] V. Victor, J. Joy Thoppan, R. Jeyakumar Nathan, and F. Farkas Maria, “Factors influencing consumer behavior and prospective purchase decisions in a dynamic pricing environment—an exploratory factor analysis approach,” Soc. Sci., vol. 7, no. 9, p. 153, 2018.

[30] J. W. Osborne and E. Waters, “Four assumptions of multiple regression that researchers should always test,” Pract. assessment, Res. Eval., vol. 8, no. 1, p. 2, 2019.