Identifying Demographic Factors Attributed to the Infection Rate of Covid-19 in Malaysia

Jun-Ting Chan[1], Keng-Hoong Ng[1](✉), Gee-Kok Tong[1], Choo-Yee Ting[1], and Kok-Chin Khor[2]

[1] Faculty of Computing and Informatics, Multimedia University Cyberjaya, 63100 Selangor, Malaysia
khng@mmu.edu.my

[2] Lee Kong Chian Faculty of Engineering and Science, UTAR Sungai Long, 43000 Selangor, Malaysia

Abstract. Since 2020, the Covid-19 pandemic has spread like wildfire across many countries, including Malaysia. The disease has caused disastrous impacts on the country's economy, public health system, and the livelihoods of its citizens. Hence, there is an urgent need to investigate and determine the underlying factors attributed to the high infection rate of Covid-19. This research aims to study and identify demographic factors attributed to the high infection rate of Covid-19 in Malaysia using regression models. The preliminary results show that the *labour force participation rate*, *unemployment rate*, and *average household income* contribute to Malaysia's high COVID-19 infection rates.

Keywords: Demographic factors · COVID-19 · Boruta attribute selection · Regression models

1 Introduction

Coronavirus disease (Covid-19) is an infectious disease caused by the SARS-CoV-2 virus [1]. The disease's symptoms are fever, headache, fatigue, cough, difficulty breathing, and loss of smell. The virus becomes more infectious when it evolves into variants, such as alpha, beta, gamma, delta, and omicron strains. According to [2], three foreign tourists were quarantined at the Sungai Buloh hospital in January 2020 after testing positive for COVID-19. It started the pandemic in Malaysia, and our country had to impose lockdowns to prohibit all citizens and foreign tourists from going out by staying at home. Many countries have to enforce the same measure to quarantine their people after the uncontrolled spread to the entire world.

This disease has taken many lives, and most of us cannot go out to work but work from home. Aside from that, business owners suffered a significant loss and were forced to close their businesses. Doctors and nurses are working days and nights to save the infected patients. From the scientific perspective, many factors could contribute to the high infection rate of Covid-19, such as population, social contact, socioeconomic, demographic factors, etc. Malaysia hit the new high of cases per day, i.e., 11,079, on July 13,

S.-C. Haw and K. Sonai Muthu (Eds.): CITIC 2022, 10, pp. 92–103, 2022.
https://doi.org/10.2991/978-94-6463-094-7_8

2021, and it was followed by hitting the 20,000 mark on August 5, 2021, even though the country had begun the mass vaccination program for its citizens [3]. This fact has driven the concern of many Malaysians that the pandemic may not end soon. As such, many mitigation measures are still needed to tackle the pandemic. The primary objective of this study is to apply data mining techniques to investigate and identify demographic factors that are responsible for the high infection rate of Covid-19 in Malaysia.

2 Related Work

According to [4], the authors applied statistical analysis using a heat map to display the total COVID-19 incidences and death rates in the United States and place them over the Social Vulnerability Index (SVI) for each county. Besides, mixed-effects negative binomial regression was used to model the total number of COVID-19 cases and changes in weekly COVID-19 cases. For mixed-effects models, zero-inflated negative binomial regression was used to model mortality counts because 29% of the counties have zero death cases. The results showed that the increase in SVI is associated with the hike in the Covid-19 mortality rate.

Research conducted by [5] evaluated the population density factor towards the spread of the disease in Algeria. The researchers used hierarchical clustering and regression analysis to investigate the problem. They clustered the cities based on the COVID-19 infection rate and then compared the disease's spread with the contamination rates of Algeria's various cities. They also created city groups based on the COVID-19 infection rates and population density to allow comparisons across areas with the greatest and lowest rates. The latter was intended to find the correlation between population density and the number of infected cases. A linear regression model was used to illustrate the impact of population density on the number of infected cases. The outcome of the research suggested that the population density is positively correlated with the spread of the Covid-19 cases.

Another research [6] conducted a thorough study to investigate the connection of COVID-19 infection with factors, i.e., mortality rate with race, health, and economic inequality in the United States. They adopted statistical analysis, i.e., bivariate, partial correlation, and regression, to find associations between factors. The study revealed that counties with a high population were prone to higher infection risk. However, a high mortality rate correlates to counties with high poverty and disability rates. Research [7] applied a regression technique to find the associations between the Covid-19 fatality rate with demographic and financial factors. The research outcome indicated that the household debt is a significant factor attributed to the disease's fatality rate.

3 Research Methodology

Figure 1 shows the research overview. There are seven stages in the research methodology, i.e., data collection, data cleaning, data normalisation, attribute selection, regression analysis, and evaluation.

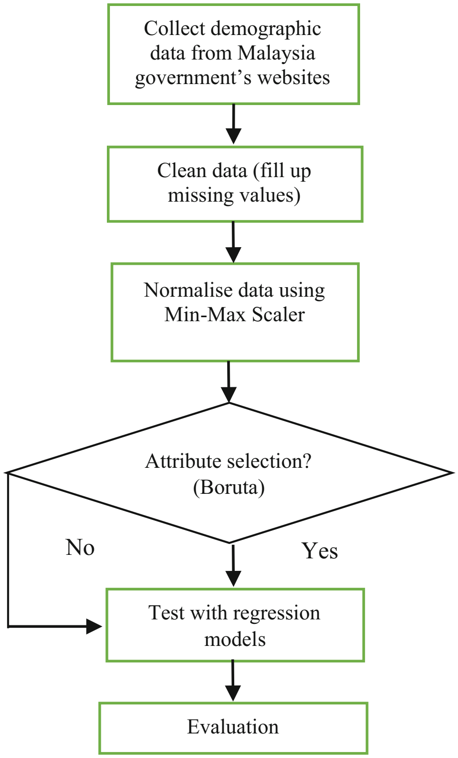

Fig. 1. The overall methodological flow of this research

3.1 Data Collection

This study collected various demographic factors (attributes) relevant to the high infection rate of Covid-19 in Malaysia. The demographic attributes used in this study were identified from [8–12].

We identified 14 demographic groups in the Malaysian context. The collected demographic attributes were then categorised into 14 demographic groups (refer to Table 1).

Besides the demographic variables listed in Table 1, the data set also included the Covid-19 Infection Rate (%) from January 25, 2020, until August 10, 2021. The Covid-19 Infection Rate (%) was calculated based on the Covid-19 cases divided by the population and multiplied by 100. This infection rate could measure the cumulative number of persons whom Covid-19 infected in a district of Malaysia.

The dependence ratio is an age-population ratio of people not in the working population (aged 0–14 or over 65) to those typically in the working population (aged 15–65). It is presented as a percentage and calculates the pressure level on the productive population [13]. In this project, a district-level demographic data set has been collected, and it is denoted by District_C19. The data were collected from the Department of Statistics Malaysia (https://www.dosm.gov.my) and the Malaysia Open Data Portal (https://www.data.gov.my).

Table 1. 14 Demographic groups and their corresponding attributes with descriptions

Demographic	Attributes	Meaning
1. Population	Population	Number of population for each district in Malaysia
2. Age group	Age group (0–14 years)	Number of people age groups from age 0 to age 14
	Age group (15– 64 years)	Number of people age groups from age 15 to age 64
	Age group (65 years and older)	Number of people age groups from age 65 and above
3. Gender	Sex (male)	Number of people for males in each district
	Sex (female)	Number of people for females in each district
4. Dependency ratio	Dependency ratio	The age-to-population ratio of people who are not normally in the working population
5. Ethnic group	Ethnic group (native)	Number of people based on the native ethnic group in each district
	Ethnic group (Chinese)	Number of people based on the Chinese ethnic group in each district
	Ethnic group (Indian)	Number of people based on the Indian ethnic group in each district
	Ethnic group (others)	Number of people based on the others ethnic group in each district
	Ethnic group (non citizen)	Number of people based on the Non citizen ethnic group in each district
6. Education level	Education__Primary and Secondary_School	Number of primary and secondary schools based on each district
	Education_Primary and Secondary_Teacher	Number of primary and secondary teachers based on each district

(*continued*)

Table 1. (*continued*)

Demographic	Attributes	Meaning
	Education_Primary and Secondary_Pupil	Number of the primary and secondary students based on each district
7. Working status	Working status (labour force)	Number of the labour force which means is either employed or unemployed based on each district
	Working status (employed person)	Number of the Employed person based on each district
	Working status (unemployed)	Number of Unemployed people based on each district
	Working status (outside labour force)	Number of not classified as an employed or unemployed person based on each district
	Working status (labour force participation rate (%))	Number rate of labour force divided by the total working-age population based on each district
	Working status (unemployment rate (%))	Number rate of Unemployment rate population based on each district
8. Marital status	Marital status (widowed)	Number of people who lose one's spouse through death based on each state
	Marital status (never married)	Number of people who never married which they are a couple or single based on each state
	Marital status (married)	Number of people who married based on each state
	Marital status (divorced)	Number of people who divorced based on each state
9. Number of household	Number of household	The total number of households with residential units as their permanent residence in each district
10. Average household Size	Average household size	Average per person who lift in the house based on each district
11. Average household income	Average household income	Average per household income based on each district

(*continued*)

Table 1. (*continued*)

Demographic	Attributes	Meaning
12. Profession	Profession_Manager_Male	Number of the profession of male managers
	Profession_Manager_Female	Number of professions of female managers
	Profession_Professional_Male	Number of professions of male professionals (exp: lawyer)
	Profession_Professional_Female	Number of professions of female professionals (exp: lawyer)
	Profession_Technicians and associate professionals_Male	Number of the profession of male technicians and associates (exp: accounting, engineer, technician)
	Profession_Technicians and associate professionals_Female	Number of professions of female Technicians and associates (exp: accounting, engineer, technician)
	Profession_Clerical support workers_Male	Number of professions of male clerical support workers (exp: secretaries)
	Profession_Clerical support workers_Female	Number of professions of female clerical support workers (exp: secretaries)
	Profession_Service workers and sales_Male	Number of profession of male service workers and sales (exp: housekeeping, salesman)
	Profession_Service workers and sales_Female	Number of professions of females sales (exp: housekeeping, saleswomen)
	Profession_Skilled agricultural, forestry and fishery workers_Male	Number of professions of males who does the planting, managing and harvesting crops; raising livestock; managing forests and catching fish
	Profession_Skilled agricultural, forestry and fishery workers_Female	Number of the profession of females who does the planting, managing and harvesting crops; raising livestock; managing forests and catching fish

(*continued*)

Table 1. (*continued*)

Demographic	Attributes	Meaning
	Profession_Craft and related trades workers_Male	Number of professions of males who does mining and construction
	Profession_Craft and related trades workers_Female	Number of professions of females who does mining and construction
	Profession_Plant and machine operators and assemblers_Male	Number of professions of males who do operate and supervise industrial and agricultural gear and equipment
	Profession_Plant and machine operators and assemblers_Female	Number of the profession of females who do operate and supervise industrial and agricultural gear and equipment
	Profession_Elementary occupation_Male	Number of professions of males who do perform simple and routine tasks (exp: cleaner)
	Profession_Elementary occupation_Female	Number of profession of male who does perform simple and routine tasks (exp: cleaner)
13. Birth rate	Birth rate	Number of birth rate based on each district
14. Average life span	Average life span (male)	Number of the average age of death in a population for males based on each state
	Average life span (female)	Number of the average age of death in a population for females based on each state

3.2 Data Preprocessing

3.2.1 Data Cleaning

The missing values in the data sets were handled by (1) searching and finding the values on other government's official websites or (2) deleting the missing values of attributes if option (1) failed. After data cleaning, the total number of records in the data set is 76.

3.2.2 Normalisation

The collected data are numerical types. All data were normalised using Min-Max Scaler, except for the COVID-19 infection rate. Hence, the attributes' values in the dataset were transformed into the range of [0,1]. This process ensures that each attribute contributes equally to model fitting.

Table 2. The result of different regression models using the District_C19 data set

LR R^2 score	DT R^2 score	RF R^2 score	KNN R^2 score	SVM R^2 score	MLP R^2 score
−32.572	0.009	**0.646**	0.395	0.358	0.391

3.2.3 Attribute Selection

The last step for data pre-processing is the attribute selection process. According to [14], using the Boruta attribute selection method could improve the classification accuracy of their experimental data sets. The training process becomes more time-consuming if the number of attributes is large. We applied Boruta, which works as a wrapper method with Random Forest because it is accurate and robust. The Boruta method shows scores by ranking each attribute from 0 to 1 with the dependent variable. The higher the number, the more important the attribute is in predicting the class label and vice-versa.

3.3 Regression Model

This research applied six commonly used regression models to determine which model delivers the best outcome. Regression was chosen to be applied in this research due to its wide usage in disease infection studies [15–17]. The regression models include Linear Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP).

3.4 Data Evaluation

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} \tag{1}$$

Equation 1 shows the evaluation matrix used in this research, the Coefficient of determination (R^2) [18]. $\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$ is the sum of squares error where $\widehat{y_i}$ denotes the value of the objective variable (y) predicted by regression for the ith data point. $\sum_{i=1}^{n}(y_i - \overline{y_i})^2$, which is the sum of squares total where $\overline{y}$ is mean of y. The R^2 score contains the best value, +1, and its value can be zero or negative. For example, if the R^2 score appears near one or one in a regression model, it means a perfect linear connection between the independent and dependent variables. On the contrary, if the R^2 score appears near zero or negative, there is no relationship between the independent and dependent variables.

4 Results and Discussion

4.1 Regression Tests on the Data Set

Regression tests were performed on the District_19 data set. Table 2 shows that the highest R^2 score for District_C19 is only 0.646 by Random Forest (RF).

The results indicated that not all demographic attributes in the data sets (particularly in the District_C19 without categorical attributes) contribute to Malaysia's high COVID-19 infection rate. Hence, attribute selection was performed to identify the important attributes.

Table 3. Top 10 attribute scores of the full District_C19 data set. * are the attributes marked by MZS as the confirmed types

Attributes	Score
Working status (labour force participation rate (%))*	1.0
Average household income*	1.0
Working status (unemployment rate (%))*	1.0
Birth rate	0.99
Working status (unemployed)	0.97
Ethnic group (Indian)	0.96
Average household size	0.96
Ethnic group (native)	0.94
Ethnic group (non citizen)	0.93
Ethnic group (others)	0.92

4.2 Regression Tests on Reduced Data Set Yielded by Attribute Selection

The attribute selection method was applied to increase the opportunity to find and identify demographic attributes that are significant to Malaysia's high COVID-19 infection rate. The selected attribute selection method in this study is the Boruta algorithm. Only the top 10 attribute scores are shown in Table 3. The ranking score order starts from 1 (best attribute) to 0 (worst attribute). To keep the table in the appropriate size, We do not include the scores for other less significant attributes.

To identify the attributes that can produce excellent results on regression models, we used the Maximum Z Score (MZS) [19]. In the MZS method, shadow attributes were added to the dataset, and they were randomly permutated. Then, the MZS among shadow attributes was calculated and identified. The MZS is the threshold that represents the highest attribute importance recorded among the shadow attributes. The attributes that scored higher than the MZS (threshold) are treated as confirmed types and have a link with the target attribute. Attributes with scores much lower than the MZS (threshold) are treated as rejected types. The remaining attributes are treated as tentative types. Hence, the Boruta algorithm will classify the attributes into three types, i.e., confirmed, tentative, and rejected. Those attributes not present in both sections (confirmed and tentative) are discarded completely. Attributes in the confirmed type were used to form the attribute subsets (attributes marked by * in Table 3).

4.2.1 Regression Results Using the Identified Attribute Subsets

Only the identified attribute subset was run with the target attribute, the COVID-19 infection rate, on six regression models to determine which models yielded good regression results. The training and testing size was split into 2–1 because the data set size is not large. Table 4 shows the regression results for the selected attribute subset.

Table 4 Comparing the regression results utilising the full data set and the reduced data sets containing only the identified attributes

	LR R^2 score	DT R^2 score	RF R^2 score	KNN R^2 score	SVM R^2 score	MLP R^2 score
Full data set	−32.572	0.009	0.646	0.395	0.358	**0.391**
Reduced data set	**0.391**	**0.281**	**0.766**	**0.731**	**0.743**	0.377

Table 4 indicated that the attribute selection process had successfully identified the attributes that significantly impacted the regression outcome: the infection rate. Overall, the R^2 scores were improved for all the models except MLP. The scores were obtained using the attribute subset consisting of three attributes, i.e., *Labour force participation, Unemployment rate,* and *average household income*. The highest R^2 score, 0.766, was generated using Random Forest (RF). The result demonstrated that the three demographic attributes play an important role in the infection rate.

5 Conclusion and Future Work

This study focused on identifying demographic factors that are attributed to the Covid-19 infection rate in Malaysia. A district-level data set was collected and compiled from the official websites of departments or agencies in the Malaysian government. The attribute selection and regression results indicated that attributes like *Labour force participation, Unemployment rate,* and *average household income* in the District_19 data set are the significant attributes that influence the Covid-19 infection rate in Malaysia.

We may consider including more demographic attributes or adding environmental factors into the study for future work.

References

1. World Health Organization. (2020, January 10). *Coronavirus*. Retrieved April 3, 2022, from https://www.who.int/health-topics/coronavirus.
2. Ting, R. (2020, January 25). First coronavirus cases in Malaysia: 3 Chinese nationals confirmed infected, quarantined in Sungai Buloh Hospital. Retrieved January 10, 2022, from https://www.theborneopost.com/2020/01/25/first-coronavirus-cases-in-malaysia-3-chinese-nationals-confirmed-infected-quarantined-in-sungai-buloh-hospital/.
3. Ministry of Health Malaysia. (2021, September 11). Current Situation of COVID-19 in Malaysia. COVID-19 MALAYSIA. Retrieved January 10, 2022, from https://covid-19.moh.gov.my/terkini.
4. Karmakar, M., Lantz, P. M., & Tipirneni, R. (2021). Association of social and demographic factors with COVID-19 incidence and death rates in the US. JAMA network open, 4(1), e2036462–e2036462.
5. Kadi, N., & Khelfaoui, M. (2020). Population density, a factor in the spread of COVID-19 in Algeria: statistic study. Bulletin of the National Research Centre, 44(1), 1–7.

6. Abedi, V., Olulana, O., Avula, V., Chaudhary, D., Khan, A., Shahjouei, S., ... & Zand, R. (2021). Racial, economic, and health inequality and COVID-19 infection in the United States. Journal of racial and ethnic health disparities, 8(3), 732–742.
7. Zainudin, N. S., Ng, K. H., & Khor, K. C. (2021). Identifying the Important Demographic and Financial Factors Related to the Mortality Rate of COVID-19 with Data Mining Techniques. In International Conference on Soft Computing in Data Science (pp. 241–253). Springer, Singapore.
8. Cole, M. A., & Neumayer, E. (2004). Examining the impact of demographic factors on air pollution. Population and Environment, 26(1), 5–21.
9. Han, Y., Jin, R., Wood, H., & Yang, T. (2019). Investigation of demographic factors in construction employees' safety perceptions. KSCE journal of civil engineering, 23(7), 2815–2828.
10. Munir, K., & Shahid, F. S. U. (2020). Role of demographic factors in economic growth of South Asian countries. Journal of Economic Studies.
11. Amarasena, T. S. M., Ajward, A. R., & Haque, A. K. M. (2015). The effects of demographic factors on job satisfaction of university faculty members in Sri Lanka. International Journal of Academic Research and Reflection, 3(4), 89–106.
12. Bujang, A. A., Zarin, H. A., & Jumadi, N. (2010). The relationship between demographic factors and housing affordability. Malaysian Journal of Real Estate, 5(1), 49–58.
13. Wan-Ibrahim, W. A., & Zainab, I. (2014). Some demographic aspects of population aging in Malaysia. World Applied Sciences Journal, 30(7), 891–894.
14. Tang, R., & Zhang, X. (2020, May). CART Decision Tree Combined with Boruta Attribute Selection for Medical Data Classification. In 2020 5th IEEE International Conference on Big Data Analytics (ICBDA) (pp. 80–84). IEEE.
15. Mohammadi, F., et al. (2021). Artificial neural network and logistic regression modelling to characterise COVID-19 infected patients in local areas of Iran. Biomedical journal, 44(3), 304–316.
16. Car, Z., Baressi, Š. S., Anđelić, N., Lorencin, I., & Mrzljak, V. (2020). Modeling the spread of COVID-19 infection using a multilayer perceptron. Computational and mathematical methods in medicine, 2020.
17. Sujath, R. A. A., Chatterjee, J. M., & Hassanien, A. E. (2020). A machine learning forecasting model for COVID-19 pandemic in India. Stochastic Environmental Research and Risk Assessment, 34(7), 959–972.
18. Kasuya, E. (2019). On the use of r and r squared in correlation and regression 34(1), 235–236. Hoboken, USA: John Wiley & Sons, Inc.
19. Szul, T., Tabor, S., & Pancerz, K. (2021). Application of the BORUTA Algorithm to Input Data Selection for a Model Based on Rough Set Theory (RST) to Prediction Energy Consumption for Building Heating. Energies, 14(10), 2779.