Educational Administration: Theory and Practice
2024,30(4), 6658- 6667
ISSN:2148-2403
https://kuey.net/

Research Article

Using Maps As A Factor To Increase The Accuracy Of Collaborative Filtering In Providing Recommendations Regarding Cluster-Based Diseases Covid-19, Varicella And Dengue.

Husni Iskandar Pohan[1*] , Sutoto[2] , Yaya Heryadi[3] , Harjanto Prabowo[4]

[1*]Computer Science Department, BINUS Graduate Program-Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia, 11480, husni.pohan@binus.ac.id
[2]Indonesian Hospital Accreditation Commission, Jakarta, Indonesia Jakarta, Indonesia 12960, sutoto@kars.or.id
[3]Computer Science Department, BINUS Graduate Program-Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia, 11480, husni.pohan@binus.ac.id, yayaheryadi@binus.edu
[4] Management Department, BINUS Business School Doctor of Research in Management Jakarta, Indonesia 11480, harprobowo@binus.edu

Citation: Husni Iskandar Pohan et.al(2024), Using Maps As A Factor To Increase The Accuracy Of Collaborative Filtering In Providing Recommendations Regarding Cluster-Based Diseases Covid-19, Varicella And Dengue.*Educational Administration: Theory And Practice*, *30*(4), 6658-6667 Doi: 10.53555/kuey.v30i4.2460

ARTICLE INFO **ABSTRACT**

In previous research, we classified disease types from a dataset of 66 thousand patient visits from 1/1/2019 to 12/31/2021 at Nadhifa Al Ghiffari, a health service institution in West Java, Indonesia. Using this data, we obtained data for diseases whose transmission is based on geographic clusters, namely Varicella (VAR), COVID-19 (COV) and Dengue Fever (DHF).
We tried to carry out several experiments using machine learning to classify types of disease and classification of referral/non refferal types. In addition, we also conduct synthetic data experiments to increase the population of health data samples which are limited in number due to regulations related to medical confidentiality,
However, to strengthen accuracy, we also tried to process visualization of map data formed by the patient's address coordinates mapped with polygons in that area, compared with the location of old patients in the active period of transmission. This research shows what can be created with the data coordinates of patients from the three diseases above.
All of this process is to provide early warning to health workers about the type of patient's disease, and whether to be referred or not, while from the government side it is necessary to observe the spread of the disease and the areas affected, in case special measures are needed such as isolating certain areas.

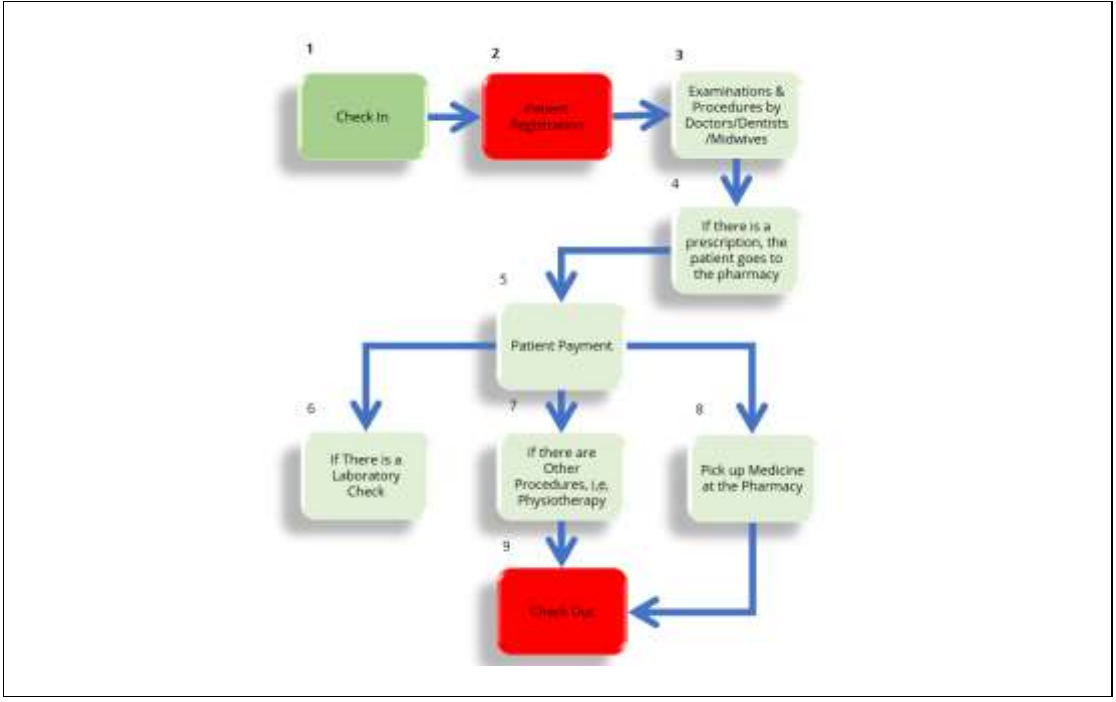
Keyword: Collaborative Filtering, Covid-19, Varicella, and Dengue.

I.INTRODUCTION

In this research, we will use the terms COV for Covid-19, VAR for Varicella and DHF for Dengue Fever. In recent years, the pandemic, notably COV, has spurred the development of diverse solutions across various sectors. For instance, in healthcare, solutions encompass vaccines and test kits (Antigen and Polymerase Chain Reaction), while the pharmaceutical field has seen the emergence of antiviral medications. Additionally, the manufacturing industry has contributed healthcare equipment like odor detectors and ventilators. Diseases such VAR and DHF often exhibit geographical proximity in terms of transmission impact.
Predicting contagious diseases based on geographic clusters necessitates caution, particularly in densely populated areas. Within these clusters, illnesses like DHFs, spread by mosquitoes [1], COV, transmitted through respiratory droplets and surface contact [2], and VAR, transferred via direct touch [3], can rapidly propagate. Early identification of these diseases by medical services can mitigate transmission rates. For instance, patients can be directed to facilities with ample resources, and immediate geographic isolation measures can be implemented as needed.

The process flow below, describes the general pattern of services at first level health facilities. The second stage is the registration process, is a critical point in selecting patients to be treated directly or referred to health services that have facilities for dangerously infectious patients.

Figure 1 standard pattern of process flow, from patient admission to discharge

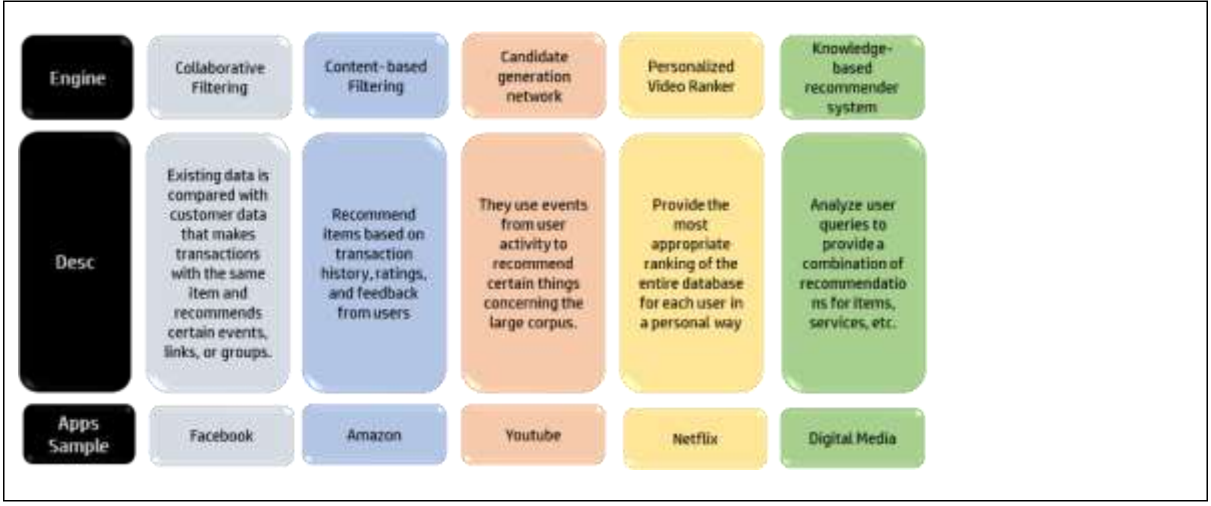

1
Check In
2
Patient Registration
3
Examinations & Procedures by Doctors/Dentists /Midwives
4
If there is a prescription, the patient goes to the pharmacy
5
Patient Payment
6
If There is a Laboratory Check
7
If there are Other Procedures, i.e. Physiotherapy
8
Pick up Medicine at the Pharmacy
9
Check Out

Despite the crucial role of early prediction, this research tries to utilize a collection of previous patient addresses (especially during periods of active transmission) to provide a map visualization that can help health workers determine what actions should be taken for new patients [4].

II.RELATED WORKS

In general, what we are trying to research is how recommendation systems, especially collaborative filtering, can be an early detection solution [5]. Incidentally, in recent years Collaborative Filtering has often used geographic solutions, for example, providing tourist recommendations for tourists who come from the same country, or providing recommendations for tourists staying in the same hotel, regarding destinations around the hotel.

Figure 2. Six types of recommender system engines



Engine	Collaborative Filtering	Content-based Filtering	Candidate generation network	Personalized Video Ranker	Knowledge-based recommender system
Desc	Existing data is compared with customer data that makes transactions with the same item and recommends certain events, links, or groups.	Recommend items based on transaction history, ratings, and feedback from users	They use events from user activity to recommend certain things concerning the large corpus.	Provide the most appropriate ranking of the entire database for each user in a personal way	Analyze user queries to provide a combination of recommendations for items, services, etc.
Apps Sample	Facebook	Amazon	Youtube	Netflix	Digital Media

In the study, this location was also used, but to help provide an idea of whether around the location of the new patient, there were old patients who also suffered from the same disease and were in a period of active transmission. Thus, overall prediction of the type of disease, prediction of whether it needs to be referred or not, as well as an overview of the geographic distribution of the disease can help health workers control the outbreak [6].

Collaborative filtering divides into two main strategies: the model-based method and the memory-based method. The model-based approach employs data mining techniques, whereas the memory-based approach relies on predictions derived from both item and user data. Both approaches prioritize item and user data equally to produce the best recommendations [7].

An issue frequently faced when applying collaborative filtering is the cold-start problem, occurring when recommendations can't be given because there are no ratings for a new item or user, or insufficient information is available. To tackle this obstacle, a popular solution involves employing a deep learning approach to enhance Quality of Service (QoS) prediction. One method is to integrate matrix factorization models with geographic data. In such instances, when confronted with a cold-start scenario, where ratings are absent, the most pertinent geographic features are utilized [8].

To tackle this obstacle, the suggestion is to employ the Spatiotemporal Dilated Convolutional Generative Network (ST-DCGN). This network processes check-in data by considering two types of recurring time intervals: hourly within a day (reflecting short-distance check-ins) and daily within a week (reflecting long-distance check-ins). The rationale behind utilizing check-in data stems from the tendency of individuals to opt for nearby venues like malls or fitness centers due to time efficiency compared to distant locations. This observation aligns with Tobler's First Law of Geography, asserting that "Everything is related to everything else, but near things are more related than distant things." Nevertheless, besides check-ins, other pertinent features such as user activities, text comments, and image data can also be leveraged [9].

Based on a number of studies above, by maximizing the existing dataset, the addresses of old patients are tried to be linked to the addresses of new patients during the active transmission period, so that visualizations can be created that help increase the accuracy of disease type predictions. However, this method is only used for three diseases whose transmission is based on geographic clusters.

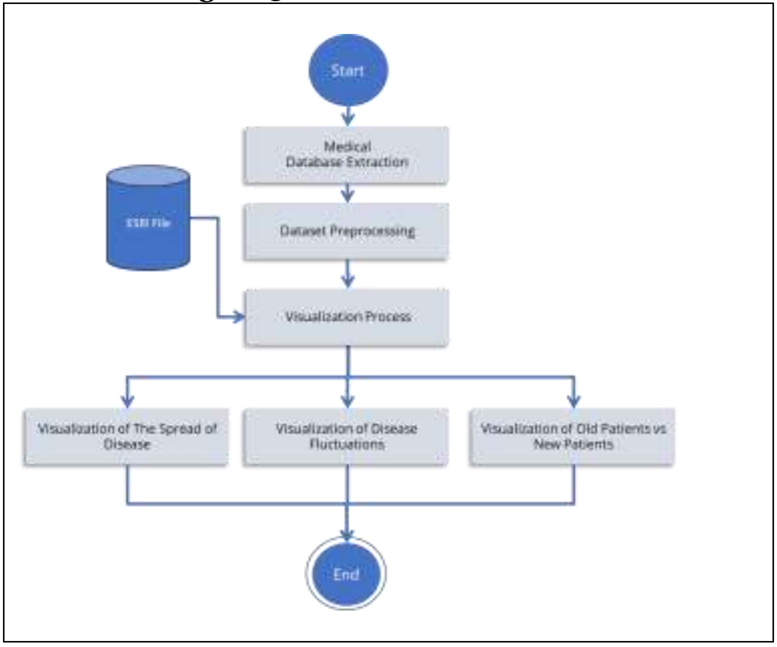
III.RESEARCH METHODS.

III.1.Research Frameworks

The process carried out is as follows

- Ensure that the patient data used has successfully identified whether the patient's disease consists of DHF, VAR or COV
- Preprocessing patient address data so that it matches the spatial function parameters used
- Ensure that spatial data is in accordance with ESRI standards and in accordance with patient distribution coordinate data at the location where health services operate.
- Conduct visualization experiments on the distribution of patients in district and sub-districts, fluctuations in disease within a certain period, and ensure the number of patient with the same disease in a certain radius.

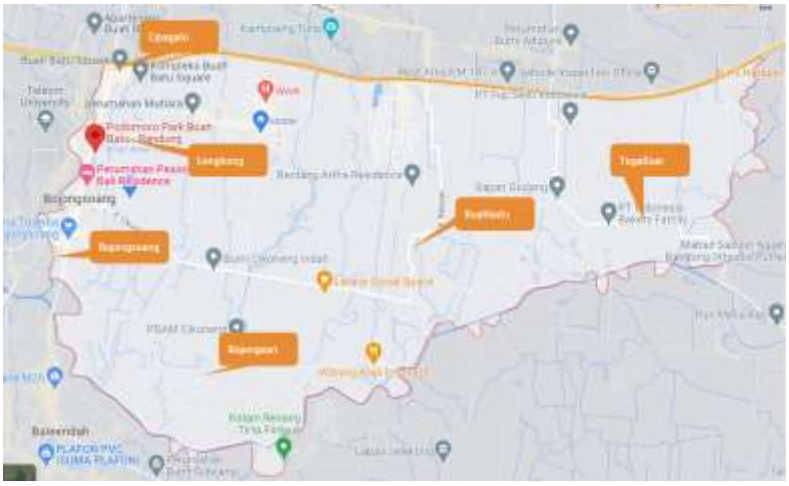
Figure 3 Research Frameworks


Start
Medical Database Extraction
ESRI File
Dataset Preprocessing
Visualization Process
Visualization of The Spread of Disease
Visualization of Disease Fluctuations
Visualization of Old Patients vs New Patients
End

III.2.Bojongsoang District

Bojongsoang District, which is also the location of Nadhifa Al Ghiffari health services which is the object of research, is located in Bandung Regency, West Java, Indonesia. It has an area of 2,622,192 hectares, with a population of 126,045 people based on the 2017 census. This sub-district has six sub-district, namely Bojongsari, Bojongsoang, Buahbatu, Cipagalo, Lengkong and Tegalluar.

Figure 4 Bojongsoang District



III.3.Map Data

The map data required is in the form of an ESRI file, which has a set of points that form a district polygon and 6 sub-district, and consists of

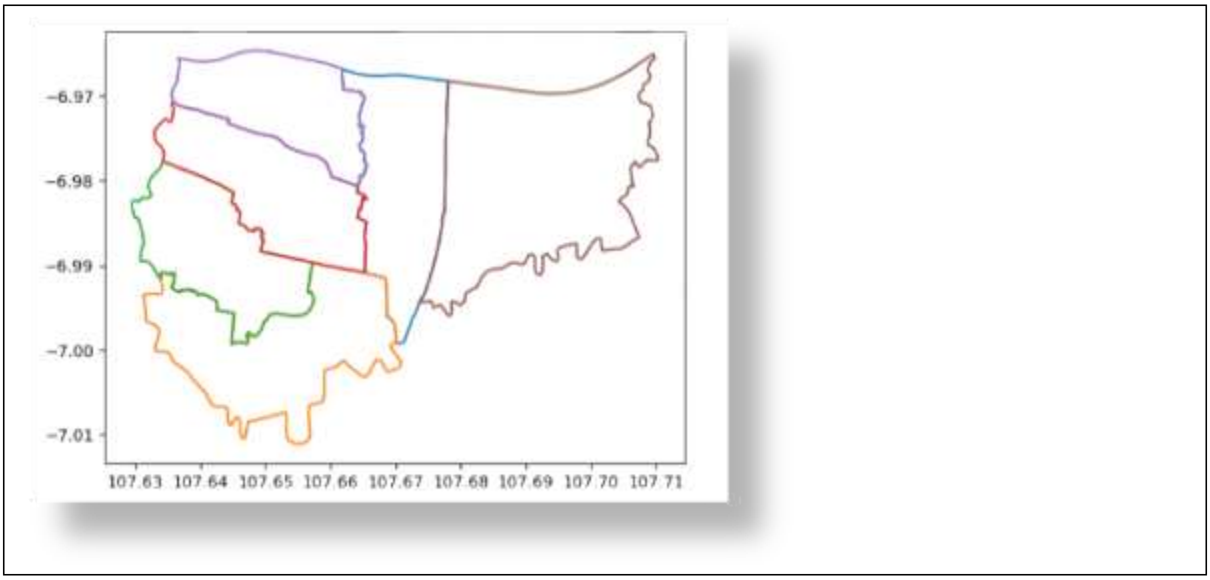
- Bojongsoang.shp
- Bojongsoang.dbf
- Bojongsoang.prj
- Bojongsoang.shx

By using the following algorithm

```
sf = shapefile.Reader("bojongsoang")
shapes = sf.shapes()
#menampilkan polygon semua kelurahan
plt.figure()
for shape in sf.shapeRecords():
   for i in range(len(shape.shape.parts)):
      i_start = shape.shape.parts[i]
      if i==len(shape.shape.parts)-1:
         i_end = len(shape.shape.points)
      else:
         i_end = shape.shape.parts[i+1]
      x = [i[0] for i in shape.shape.points[i_start:i_end]]
      y = [i[1] for i in shape.shape.points[i_start:i_end]]
      plt.plot(x,y)
```

With this ESRI map data, we can display the polygons of all district and sub-district polygons as follows.

Figure 5 Bojongsoang Polygon



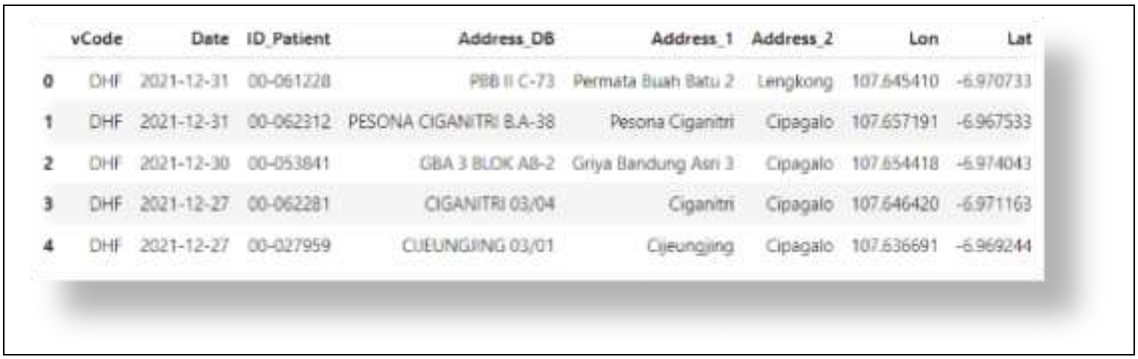
III.4.Search for Longitude and Latitude

The following is an example of sample data that has been processed with Nominatim, a tool that allows us to get the longitude and latitude coordinates of an address. Apart from Nominatim, there are several other tools, namely Google Maps Platform, Mapbox API, Bing API, and Yandex Map API. This research uses Nominatim, because the service is currently free, but if you want to increase accuracy you can use Google Maps Platform, but of course additional costs are required.

The following is the algorithm for getting the coordinates from the address parameters

```
from geopy.geocoders import Nominatim
geolocator = Nominatim(user_agent="example app")
location = geolocator.geocode("address#1, address#2")
lon_loc = location.longitude
lat_loc = location.latitude
print('Longitude : ',lon_loc,' and Latitude :  ',lat_loc)
date_start  = datetime.datetime(2019,1,1) # start
date_end = datetime.datetime(2021,12,31) # end
```

with the following results



	vCode	Date	ID_Patient	Address_DB	Address_1	Address_2	Lon	Lat
0	DHF	2021-12-31	00-061228	PBB II C-73	Permata Buah Batu 2	Lengkong	107.645410	-6.970733
1	DHF	2021-12-31	00-062312	PESONA CIGANITRI B.A-38	Pesona Ciganitri	Cipagalo	107.657191	-6.967533
2	DHF	2021-12-30	00-053841	GBA 3 BLOK AB-2	Griya Bandung Asri 3	Cipagalo	107.654418	-6.974043
3	DHF	2021-12-27	00-062281	CIGANITRI 03/04	Ciganitri	Cipagalo	107.646420	-6.971163
4	DHF	2021-12-27	00-027959	CIJEUNGJING 03/01	Cijeungjing	Cipagalo	107.636691	-6.969244

Table 1 Longitude and Latitude

IV.EXPERIMENT RESULTS AND DISCUSSION

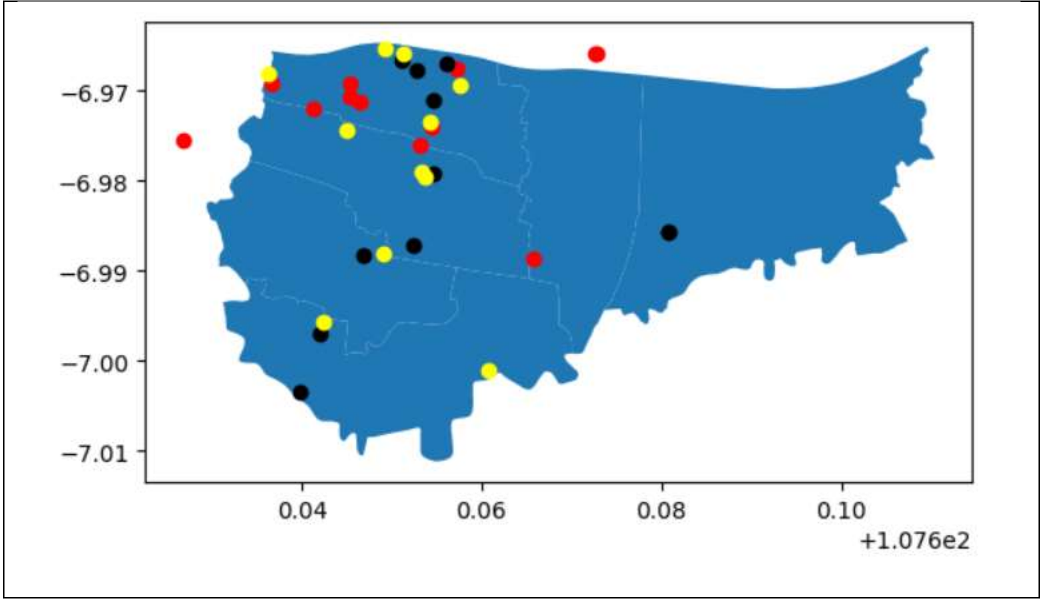
IV.1.Patient Distribution

With this data, the cluster-based distribution of disease in certain geographic areas can be displayed as shown in the image below. Where yellow is VAR, red is DHF and black is COV. This process is generated from the following algorithm

```
BBT = BSR = BJS = LNG = CPG = TGL = OTH = 0
for i in range(0, len(df)):
   z = (df.loc[i, 'Date'])
```

```
    date_df = datetime.datetime(int(z[:4]),int(z[5:-3]),int(z[-2:]))
    _Lon = df.loc[i, 'Lon']
    _Lat = df.loc[i, 'Lat']
    if (date_df >= date_start) and (date_df <= date_end):
        if PolyBBT.contains(Point(_Lon,_Lat)): BBT += 1
        elif PolyBSR.contains(Point(_Lon,_Lat)): BSR += 1
        elif PolyBJS.contains(Point(_Lon,_Lat)): BJS += 1
        elif PolyLNG.contains(Point(_Lon,_Lat)): LNG += 1
        elif PolyCPG.contains(Point(_Lon,_Lat)): CPG += 1
        elif PolyTGL.contains(Point(_Lon,_Lat)): TGL += 1
        else: OTH +=1
```

Figure 6 Disease Distribution Sample

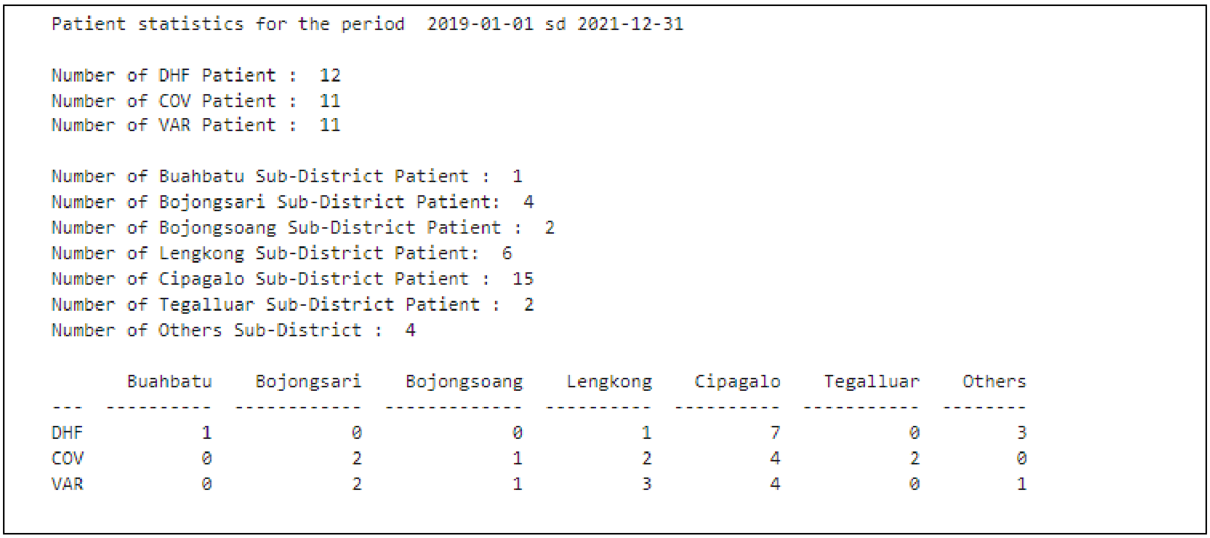


IV.2.Calculation of the Number of Patients

If data is needed to monitor the spread of disease per sub-district, where sub-district data is represented by BBT (Buahbatu), BSR (Bojongsari), LNG (Lengkong), BJS (Bojongsoang), CPG (Cipagalo), and TGL (Tegalluar), then The following algorithm can be used

```
DHF = COV = VAR = 0
BBT = BSR = BJS = LNG = CPG = TGL = OTH = 0
baris = kolom = 0
tabsum = [[‘DHF',0,0,0,0,0,0,0,0],[‘COV',0,0,0,0,0,0,0,0],['VAR',0,0,0,0,0,0,0,0]]
for i in range(0, len(df)):
    z = (df.loc[i, 'Date'])
    date_df = datetime.datetime(int(z[:4]),int(z[5:-3]),int(z[-2:]))
    if (date_df >= date_start) and (date_df <= date_end):
        _Lon = df.loc[i, 'Lon']
        _Lat = df.loc[i, 'Lat']
        xxCode = df.loc[i, 'xCode']
        if PolyBBT.contains(Point(_Lon,_Lat)):
            BBT += 1
            kolom = 1
        elif PolyBSR.contains(Point(_Lon,_Lat)):
            BSR += 1
            kolom = 2
elif PolyBJS.contains(Point(_Lon,_Lat)):
            BJS += 1
            kolom = 3
elif PolyLNG.contains(Point(_Lon,_Lat)):
            LNG += 1
            kolom = 4
        elif PolyCPG.contains(Point(_Lon,_Lat)):
            CPG += 1
            kolom = 5
        elif PolyTGL.contains(Point(_Lon,_Lat)):
            TGL += 1
            kolom = 6
        else:
            OTH +=1
            kolom = 7
```

Locations outside the polygon are grouped into the OTH (Others) sub-district. With the existing sample data, we can produce the following tabulation



```
Patient statistics for the period  2019-01-01 sd 2021-12-31

Number of DHF Patient :  12
Number of COV Patient :  11
Number of VAR Patient :  11

Number of Buahbatu Sub-District Patient :  1
Number of Bojongsari Sub-District Patient:  4
Number of Bojongsoang Sub-District Patient :  2
Number of Lengkong Sub-District Patient:  6
Number of Cipagalo Sub-District Patient :  15
Number of Tegalluar Sub-District Patient :  2
Number of Others Sub-District :  4

       Buahbatu   Bojongsari   Bojongsoang   Lengkong   Cipagalo   Tegalluar   Others
---  ----------  ------------  -------------  ----------  ----------  -----------  --------
DHF           1             0              0           1           7            0         3
COV           0             2              1           2           4            2         0
VAR           0             2              1           3           4            0         1
```

Table 2 Calculation of The Number of The Patient

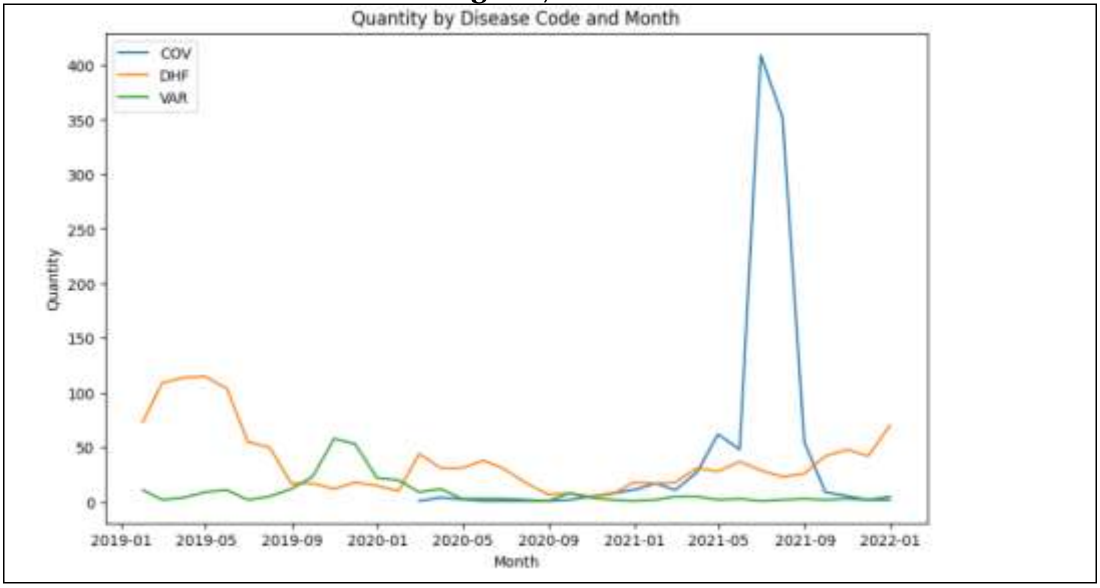
IV.3. Disease Fluctuations

Disease fluctuations within a certain period can be generated using the seaborn library and matplotlib.pyplot library as in the algorithm below

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Read the Excel file
data = pd.read_excel("TestSimpo2.xlsx","TestOri")

# Convert the 'date' column to datetime format
data['Date'] = pd.to_datetime(data['Date'])

# Group the filtered data by location and month, and calculate the sum of quantity
grouped_data = data.groupby(['vCode', pd.Grouper(key='Date', freq='M')])['Qty'].sum().reset_index()

# Plot the graph
plt.figure(figsize=(10, 6))
sns.lineplot(data=grouped_data, x='Date', y='Qty', hue='vCode')

# Customize the graph
plt.xlabel('Month')
plt.ylabel('Quantity')
plt.title('Quantity by Disease Code and Month')
plt.legend(loc='upper left')

# Show the graph
plt.show()
```

In the following image we can see the results, which show a spike in mid-2019 for DHF, then a spike in mid-2021 for COV.

Figure 7 Disease Fluctuations

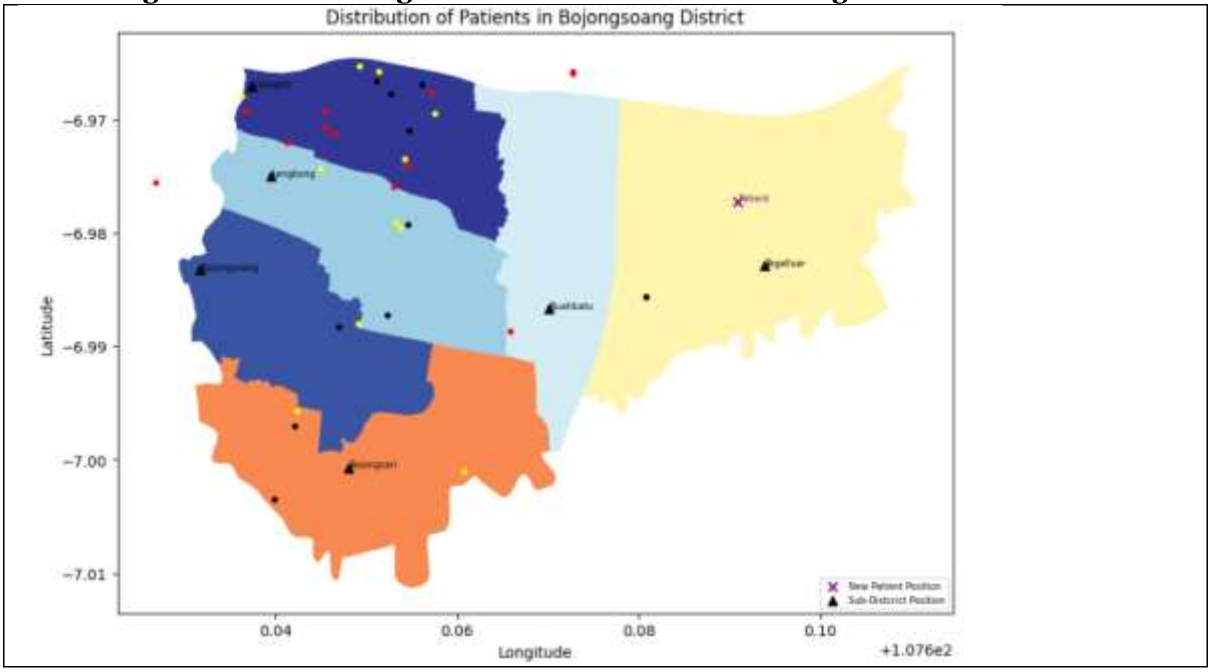


IV.4.Determining the Position of New Patients Against Old Patients

The following is the determination of the position of the new patient, given the symbol X, towards the old patient with the following visualization. This process requires

- New patient address to search for coordinates
- Types of disease
- Old patient database to search cluster-based coordinates and disease types
- District and Sub-District Polygon
- Sub-District Point
- Early and late periods of transmission

Figure 8 Determining the Position of New Patients Against Old Patients



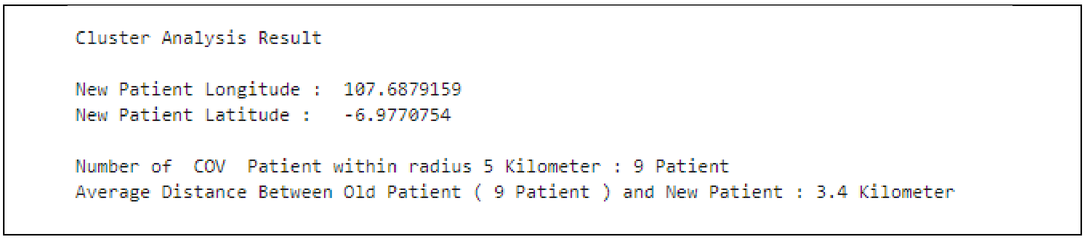
IV.5.Calculation of the Number of Patients in a Certain Radius

To calculate the number of patients with the same disease, within a certain radius and calculate the average distance to new patients, use the following algorithm.

```
_address = address in string
_xCode = disease code
_radius = radius (in kilometer)
geolocator = Nominatim(user_agent="example app")
location = geolocator.geocode(_address)
```

```
lon_loc = location.longitude
lat_loc = location.latitude
date_start  = start period
date_end = end period
from geopy.distance import geodesic
patient_qty = 0
total_distance = 0
for i in range(0, len(df)):
   z = (df.loc[i, 'Date'])
   date_df = datetime.datetime(int(z[:4]),int(z[5:-3]),int(z[-2:]))
   _Lon = df.loc[i, 'Lon']
   _Lat = df.loc[i, 'Lat']
   xxCode = df.loc[i, 'vCode']
   if (date_df >= date_start) and (date_df <= date_end) and (xxCode == _xCode):
      distance = geodesic((lat_loc, lon_loc), (_Lat, _Lon)).kilometers
      if distance <= _radius:
         patient_qty += 1
         total_distance += distance
```

with the following results



```
Cluster Analysis Result

New Patient Longitude :  107.6879159
New Patient Latitude :   -6.9770754

Number of  COV  Patient within radius 5 Kilometer : 9 Patient
Average Distance Between Old Patient ( 9 Patient ) and New Patient : 3.4 Kilometer
```

Table 3 Calculation of the Number of Patients in a Certain Radius

IV.6.Discussion

After carrying out a series of experiments above, we believe that patient address data map can be used for

- Monitor fluctuations in the number of patients within a certain period and a specific geographic area
- Monitor the distribution of patients within a specific geographic area
- Provides calculation results for the number of patients with the same disease in a certain radius and provides the average distance between new patients and old patients

However, as we know, even though in the experiment above for districts and sub-districts using polygons, the calculation of the number of sufferers in a certain area still uses a radius, which in this case ignores geographical separation factors such as rivers, roads and complex fences. So in the future it is possible to use polygons that have taken into account the geographical separation factor.

Another important thing is that the address data used may not necessarily be usable, because during registration the registration officer sometimes uses abbreviations that are familiar to that location but are not recognized by the coordinate search application. As feedback, for this visualization to be successful, the recording of address data must be done as well as possible.

REFERENCES

[1] H. I. Pohan, W. Suparta, Y. Heryadi, A. Wibowo, and Lukas, “Prediction of DHF (Dengue Hemorrhagic Fever) Severity Using Random Forest, KNN, Decision Tree and Naïve Bayes,” *Proc. 2022 IEEE 7th Int. Conf. Inf. Technol. Digit. Appl. ICITDA 2022*, 2022, doi: 10.1109/ICITDA55840.2022.9971377.

[2] L. M. Andersen, S. R. Harden, M. M. Sugg, J. D. Runkle, and T. E. Lundquist, “Analyzing the spatial determinants of local Covid-19 transmission in the United States,” *Sci. Total Environ.*, vol. 754, Feb. 2021, doi: 10.1016/j.scitotenv.2020.142396.

[3] A. A. Gershon *et al.*, “Varicella zoster virus infection,” *Nat. Rev. Dis. Prim.*, vol. 1, no. July, pp. 1–19, 2015, doi: 10.1038/nrdp.2015.16.

[4] T. Zhong, S. Zhang, F. Zhou, K. Zhang, G. Trajcevski, and J. Wu, “Hybrid graph convolutional networks with multi-head attention for location recommendation,” *World Wide Web*, vol. 23, no. 6, pp. 3125–3151, 2020, doi: 10.1007/s11280-020-00824-9.

[5] Sanam Malhotra, “Recommendation Systems With Machine Learning,” 2020. https://artificialintelligence.oodles.io/blogs/recommendation-systems-with-machine-learning/

[6] Y. Ma and M. Gan, “Exploring multiple spatio-temporal information for point-of-interest recommendation,” *Soft Comput.*, vol. 24, no. 24, pp. 18733–18747, 2020, doi: 10.1007/s00500-020-05107-z.

[7] A. Anjali, J. K. Sandhu, and D. Goyal, “User profiling in travel recommender system using hybridization and collaborative method,” *Proc. - IEEE 2021 Int. Conf. Comput. Commun. Intell. Syst. ICCCIS 2021*, pp. 143–148, 2021, doi: 10.1109/ICCCIS51004.2021.9397099.

[8] M. I. Smahi, F. Hadjila, C. Tibermacine, and A. Benamar, “A deep learning approach for collaborative prediction of Web service QoS,” *Serv. Oriented Comput. Appl.*, vol. 15, no. 1, pp. 5–20, 2021, doi: 10.1007/s11761-020-00304-y.

[9] C. Liu *et al.*, “A spatiotemporal dilated convolutional generative network for point-of-interest recommendation,” *ISPRS Int. J. Geo-Information*, vol. 9, no. 2, 2020, doi: 10.3390/ijgi9020113.