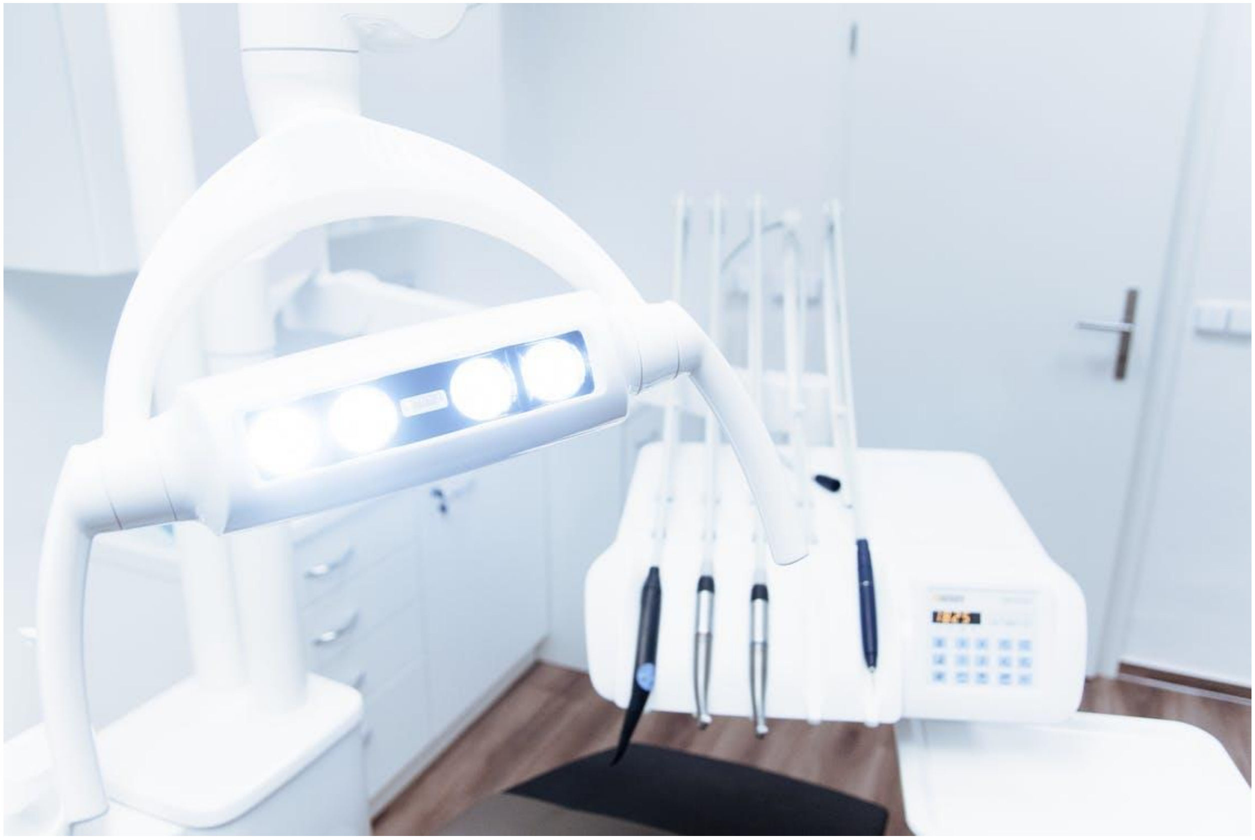


Clinic Go Where: KMeans-Based Clinic Recommendation System

Document: Project Report

Submitted By: Dong Meirong

Last Updated: 03-May-2020

Summary: One of the notable challenges in the primary care setting is the lack of informed choice. That is patients do not know which clinics to visit that can best serve their interest and general practitioners do not know where to set up clinics with their niche skills to help more patients. The project recognized the challenge and prototyped a KMeans-based clinics recommendation system to help both parties to make better decisions. The system can be expanded to make more precise recommendations if more information (e.g. population groups in an area, user feedback) are gathered.

Table of Contents

1.0 Introduction

1.1 Business Value

Healthcare is the focus of many governments around the world today because the world population is ageing at an unprecedented rate and this trend has led to higher demand on healthcare and aged care services and posed new challenges in the delivery of care.

According to the data from World Population Prospects: the 2019 Revision, in 2018, for the first time in history, people aged 65 or above outnumbered children under five years of age globally. And by 2050, one in six people in the world will be over age 65 (16%), up from one in 11 in 2019 (9%). The demographic shift has a direct effect on the growth of health spending. In the United States, the health share of Gross Domestic Product (GDP) is expected to rise from 17.9 percent in 2017 to 19.4 percent in 2027. In Singapore, the government allocated $10.7 billion to healthcare, the third largest spending by ministry, in the Financial Year 2017. And the amount is expected to “rise quite sharply” to about $14billion by Financial Year 2022 as the population ages.

However, an increase in healthcare spending does not always translate to better outcomes because the healthcare situation is complex, and the current support model is not very sustainable. Ideally, the revision of care models for the changing population structures should focus on reducing disease burden through early intervention and prevention via cost-effective healthcare integration, strengthening of primary healthcare, and collaboration for health in the community.

There are a number of challenges the industry must address to build such models. One of the notable challenges is the lack of choice. Each general practitioner has his/her niche specialties and a preferred way of working with patients and promoting their health, which is more compatible with some patients than others. For this reason, it is important that patients are given the information to select and find the general practitioner that best caters to their needs. But in some regions, there is a noticeable lack of choice and patients may not receive all the benefits of primary care as they do not have access to the general practitioner who best suited to their needs. So, there is a need of a platform to increase the information transparency to help both patients and medical practitioners in decision-making.

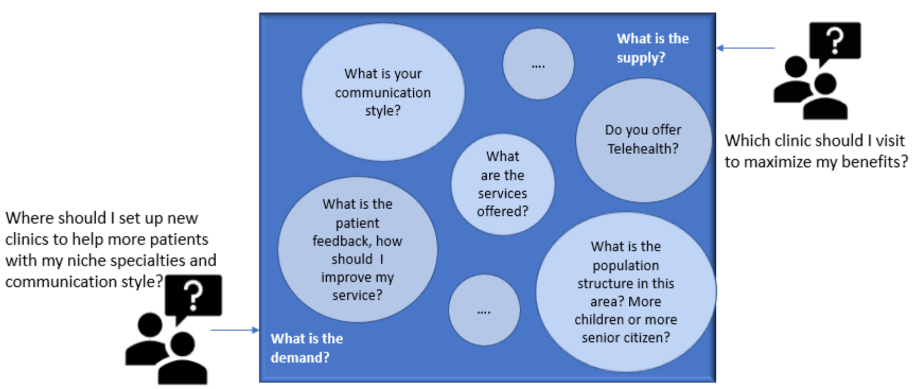


Figure 1: Care Integration Challenge

Through the platform, patients will gain a better knowledge of the clinics near them. And general practitioners can have a better idea of what are the niche specialties and communication style preferred in a community, to better serve patients.

The project recognized the need by prototyping a KMeans-based clinic recommendation system to address the concerns highlighted.

1.2 Project Scope

The product roadmap is shown below.

Figure 2: Product Roadmap

Due to a lack of time and resources and, the prototype is built in the context of Singapore healthcare system. Currently, the prototype only accepts user location as criteria for recommendation. In the future, the system can be enhanced to make more precise recommendations with more data such as population information (e.g. use the number of schools in an area to determine the population group and predict their healthcare needs) and user feedback.

2.0 System Design

The location-based clinics recommendation prototype can be broken down into 7 components. The different components are illustrated in the design below.

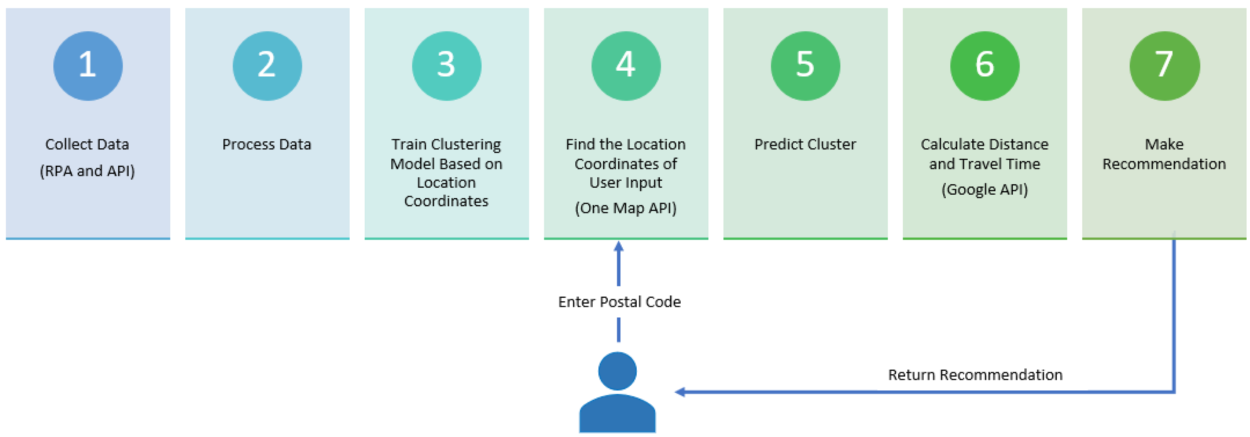


Figure 3: System Design

2.1 Data Collection

The clinic data used in this project are collected from the HCI Directory and Flu Go Where website via Robotic Process Automation (RPA) tools.

The HCI Directory is an initiative by the Ministry of Health (MOH) of Singapore to display all healthcare institutions (HCIs) licensed under the Private Hospitals and Medical Clinics (PHMC) Act to provide information about their services and operations to the public. Although this is an initiative by MOH, information featured is provided by the respective HCIs and there might be errors. The errors will be corrected via data processing in the next stage.

Flu Go Where website lists all the Public Health Preparedness Clinics (PHPCs), as well as polyclinics, which will provide special subsidies for Singapore Citizens and Permanent Residents diagnosed with respiratory illnesses (e.g. common cold). The information of PHPCs is not included in the HCI Directory because the PHPCs are primary care clinic set up in response to public health emergencies such as influenza pandemic, haze and COVID-19 (Coronavirus Disease 2019) pandemic only.

At the time the project was created, there was no Application Programming Interface (API) available for the provision of clinic details. Therefore, the project used Tagui, an RPA tool, to collect clinic information (such as names, address and phone number) from the HCI Directory and Flu Go Where Website.

RPA tools are mostly used to replace high-volume, repeatable tasks that previously required humans to perform such as copying data. The use of RPA as compared to manual work can improve the overall accuracy of data collection process, while reducing turnaround time and saving manhours for more meaningful tasks. The disadvantage is RPA scripts need consistent maintenance if the structures of the targeted websites change frequently.

There are about 4,000 clinics in the HCI Directory (4401 clinics found when the script "ExtractData _HCI.ipynb" was last run.) and 900+ clinics in Flu Go Where (924 general clinics and 20 polyclinics found when the script "ExtractData_PHPC.ipynb" was last run). The copying stage took about half a day because HCI Directory was slow to load.

The features collected from the HCI Directory are name, contact number, fax, address, opening hour, profile hyperlinks, offered services, fees and participated programs such as Chronic Disease Management Programme (CDMP), Community Health Assistance Scheme (CHAS) and Integrated Screening Programme (ISP).

Sno	Name	Contact Number	Fax	Address	Opening Hours	Link	Service	Programmes	Fees
1	DA CLINIC @ BISHAN	69541012.0	69541012	BLK 501 BISHAN STREET 11 ...	Eve of Chinese New Year (24/01/2020) : 09:00 a...	http://hcidirectory.sg/hcidirectory/clinic.do?...	General Medical	Community Health Assistance Scheme (CHAS)	NaN
2	PACIFIC ADVANCE RENAL CARE (CHOA CHU KANG)	68930617.0	69541012	BLK 488B CHOA CHU KANG AVENUE 5 ...	Please call the clinic for operating hours	http://hcidirectory.sg/hcidirectory/clinic.do?...	General Medical\n\t\t\t\t\t\t\t\t\n\t\t\t\t\...	NaN	NaN
3	SHANG DENTAL (KALLANG)	64433443.0	69541012	71 KALLANG BAHRU ...	Please call the clinic for operating hours	http://hcidirectory.sg/hcidirectory/clinic.do?...	General Dental	NaN	NaN
4	SINGHEALTH POLYCLINICS - TAMPINES	66436969.0	69541012	1 TAMPINES STREET 41 ...	Monday to Friday : 08:00 am to 01:00 pm, 02:00...	http://hcidirectory.sg/hcidirectory/clinic.do?...	General Dental\n\t\t\t\t\t\t\t\t\n\t\t\t\t\t...	NaN	NaN
5	1 MEDICAL TECK GHEE	62517030.0	69541012	BLK 410 ANG MO KIO AVENUE 10 ...	Eve of Chinese New Year (24/01/2020) : 07:00 a...	http://hcidirectory.sg/hcidirectory/clinic.do?...	General Medical	Community Health Assistance Scheme (CHAS)	NaN

Figure 4: HCI Data Features

The details of geographic coordinates were downloaded from OneMap Singapore, the authoritative national map of Singapore with the most comprehensive and timely updated information developed by the Singapore Land Authority.

The coordinates are downloaded via Application Programming Interface (API) therefore it is faster. The features collected are block number, road name, buildings, address, postal code, X coordinate, Y coordinate, latitude and longitude.


```
"SEARCHVAL":"INLAND REVENUE AUTHORITY OF SINGAPORE (IRAS)",
"BLK_NO":"55",
"ROAD_NAME":"NEWTON ROAD",
"BUILDING":"INLAND REVENUE AUTHORITY OF SINGAPORE (IRAS)",
"ADDRESS":"55 NEWTON ROAD, SINGAPORE 307987",
"POSTAL":"307987",
"X":"28983.7537272647",
"Y":"33554.4361084122",
"LATITUDE":"1.31972890510723",
"LONGITUDE":"103.842158118267",
"LONGTITUDE":"103.842158118267"
```


Figure 5: Screenshot from OneMap showing the details of coordinates

2.2 Data Processing

There are a few problems with the data gathered (e.g. wrong postal codes and empty fields). Data processing is implemented to correct the problmes and encode categorical features (e.g. services and programs) as a one-hot numeric array for easier interpretation.

Address	Opening Hours	Link	Service	24 Hour	Medical	Dental	CDMP	CHAS	ISP	PHPC
BLK 501 BISHAN STREET 11 #01-374 Singapore 570501	Eve of Chinese New Year (24/01/2020) : 09:00 a...	http://hcidirectory.sg/hcidirectory/clinic.do?...	General Medical	0	1	0	0	1	0	1
BLK 488B CHOA CHU KANG AVENUE 5 SUNSHINE GARDE...	Please call the clinic for operating hours	http://hcidirectory.sg/hcidirectory/clinic.do?...	General Medical,Specialist Medical	0	1	0	0	0	0	0
71 KALLANG BAHRU #01-531F Singapore 330071	Please call the clinic for operating hours	http://hcidirectory.sg/hcidirectory/clinic.do?...	General Dental	0	0	1	0	0	0	0
1 TAMPINES STREET 41 TAMPINES POLYCLINIC Singa...	Monday to Friday : 08:00 am to 01:00 pm, 02:00...	http://hcidirectory.sg/hcidirectory/clinic.do?...	General Dental,General Medical	0	1	1	0	0	0	0
BLK 410 ANG MO KIO AVENUE 10 TECK GHEE SQUARE	Eve of Chinese New Year (24/01/2020) : 07:00 a...	http://hcidirectory.sg/hcidirectory/clinic.do?...	General Medical	0	1	0	0	1	0	1

Figure 6: One Hot Encoding of Categorical Features

On top of that, the data collected from different sources (e.g. OneMap, Flu Go Where and the HCI Directory) are joined at this stage by postal code.

	Sno	Name	Contact Number	Fax	Address	Opening Hours	Link	Service	24 Hour	Medical	Dental	C	CDMP	CHAS	ISP	PHPC	Postal	Latitude	Longitude
0	1	DA CLINIC @ BISHAN	69541012.0	69541012	BLK 501 BISHAN STREET 11 #01-374 Singapore 570501	Eve of Chinese New Year (24/01/2020) : 09:00 a...	http://hcidirectory.sg/hcidirectory/clinic.do?...	General Medical	0	1	0		0	1	0	1	570501	1.348417	103.849077
1	2	PACIFIC ADVANCE RENAL CARE (CHOA CHU KANG)	68930617.0	69541012	BLK 488B CHOA CHU KANG AVENUE 5 SUNSHINE GARDE...	Please call the clinic for operating hours	http://hcidirectory.sg/hcidirectory/clinic.do?...	General Medical,Specialist Medical	0	1	0		0	0	0	0	682488	1.376974	103.739293
2	3	SHANG DENTAL (KALLANG)	64433443.0	69541012	71 KALLANG BAHRU #01-531F Singapore 330071	Please call the clinic for operating hours	http://hcidirectory.sg/hcidirectory/clinic.do?...	General Dental	0	0	1		0	0	0	0	330071	1.320993	103.870390
3	4	SINGHEALTH POLYCLINICS - TAMPINES	66436969.0	69541012	1 TAMPINES STREET 41 TAMPINES POLYCLINIC Singa...	Monday to Friday : 08:00 am to 01:00 pm, 02:00...	http://hcidirectory.sg/hcidirectory/clinic.do?...	General Dental,General Medical	0	1	1		0	0	0	0	529203	1.357301	103.945905
4	5	1 MEDICAL TECK GHEE	62517030.0	69541012	BLK 410 ANG MO KIO AVENUE 10 TECK GHEE SQUARE ...	Eve of Chinese New Year (24/01/2020) : 07:00 a...	http://hcidirectory.sg/hcidirectory/clinic.do?...	General Medical	0	1	0		0	1	0	1	560410	1.362715	103.854645

Figure 7: Processed Data

2.3 Model Training

The project uses unsupervised machine learning algorithms for modelling because unsupervised machine learning algorithms are capable of inferring patterns from a dataset without reference to known or labeled outcomes. There are 3 algorithms used in this project for study which are KMeans, Density-based spatial clustering of applications with noise (DBSCAN) and DBSCAN based hybrid system. The inputs used for modelling are geographic coordinates (latitude and longitude). The outputs are measure by Silhouette coefficients. The Silhouette coefficients ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

The geographic distribution of clinics in Singapore are illustrated below.

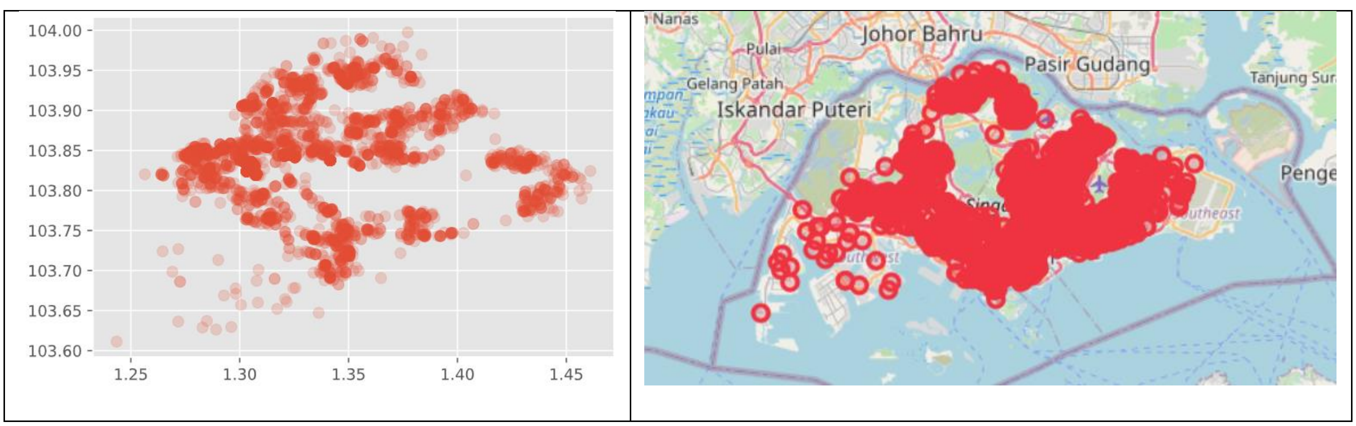


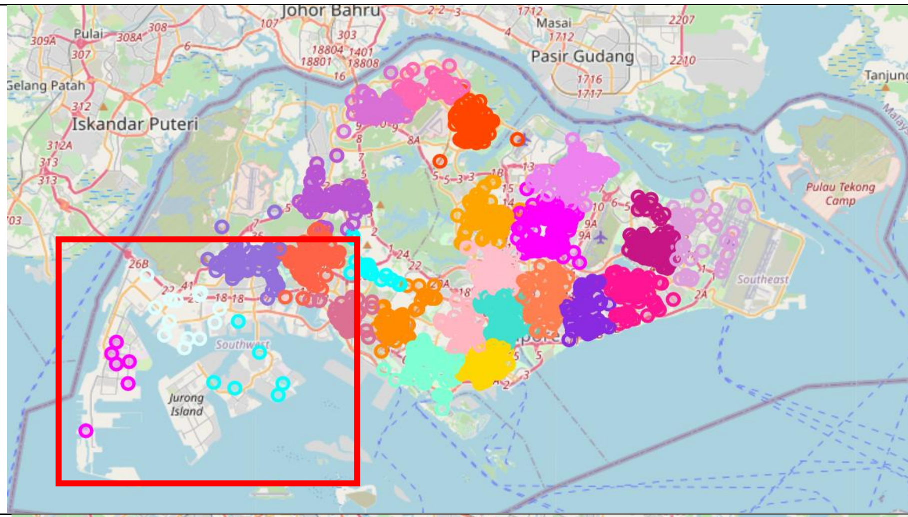
Figure 8: Singapore Clinics Distribution

KMeans

KMeans works by grouping similar data points together to discover underlying patterns. To use the algorithm, the number of centroids (mean value of each cluster) or the number of clusters, K, must be specified first. The algorithm then starts with K randomly selected centroids and allocates every data point to its nearest centroid. The algorithm will perform iterative calculations to optimize the positions of the centroids until the centroids have stabilized or the defined number of iterations has been achieved.

The project expects 20 to 50 clusters formed (the estimation is derived based on the number postal districts in Singapore). After running an exhaustive search over K, it was shown that the parameter K=25 would give the best Silhouette coefficient of 0.45. However, K=30 is used in this project because considering the actual geographic conditions such as the presence of water bodies, k=30 gives better clustering. The silhouette coefficient of K=30 is 0.445.

K = 25



K = 30

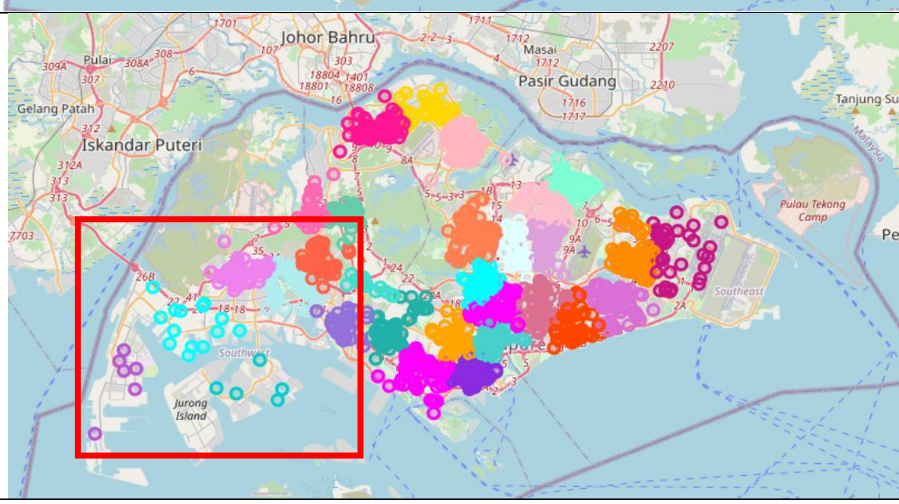


Figure 9: KMeans Clusters

DBSCAN

DBSCAN is used to differentiate datasets into high-density regions and low-density regions. Data points in low-density regions are considered outliers.

The algorithm has two mandatory parameters: eps and minPoints. Eps specifies how close points should be to each other to be considered a part of a cluster or number of points within a circle of Radius Eps (ϵ) from point P. MinPoints specifies the minimum number of points to form a dense region. In this project, minimally 10 clinics must be found in the cluster within a circle of 0.5km. So, the parameters are set as eps = 0.5 kilometers (which had been converted to radians for use by haversine) and the minPoints =10 clinics.

Haversine formula is adopted for distance measurement. The Haversine formula is usually used to calculate the shortest distance between two points on a sphere (e.g. Earth) using their latitudes and longitudes measured along the surface and is widely used in navigation.

Haversine formula:

$$a = \sin^2(\Delta\varphi/2) + \cos\varphi_1 \cdot \cos\varphi_2 \cdot \sin^2(\Delta\lambda/2)$$

$$c = 2 \cdot \mathrm{atan2}(\sqrt{a}, \sqrt{(1-a)})$$

$$d = R \cdot c$$

where φ is latitude, λ is longitude, R is earth's radius (mean radius = 6,371km); note that angles need to be in radians to pass to trig functions!

JavaScript:

```
const R = 6371e3; // metres
const φ1 = lat1 * Math.PI/180; // φ, λ in radians
const φ2 = lat2 * Math.PI/180;
const Δφ = (lat2-lat1) * Math.PI/180;
const Δλ = (lon2-lon1) * Math.PI/180;

const a = Math.sin(Δφ/2) * Math.sin(Δφ/2) +
          Math.cos(φ1) * Math.cos(φ2) *
          Math.sin(Δλ/2) * Math.sin(Δλ/2);
const c = 2 * Math.atan2(Math.sqrt(a), Math.sqrt(1-a));

const d = R * c; // in metres
```

Figure 10: Implementation Details from Movable Type Scripts

This setting categorizes data points into 32 clusters with Silhouette coefficient ignoring outliers = 0.465. However, there are also 612 points categorized as outliers (the red points). The Silhouette coefficient of outliers as singletons = -0.0754. The clusters produced are not very ideal.

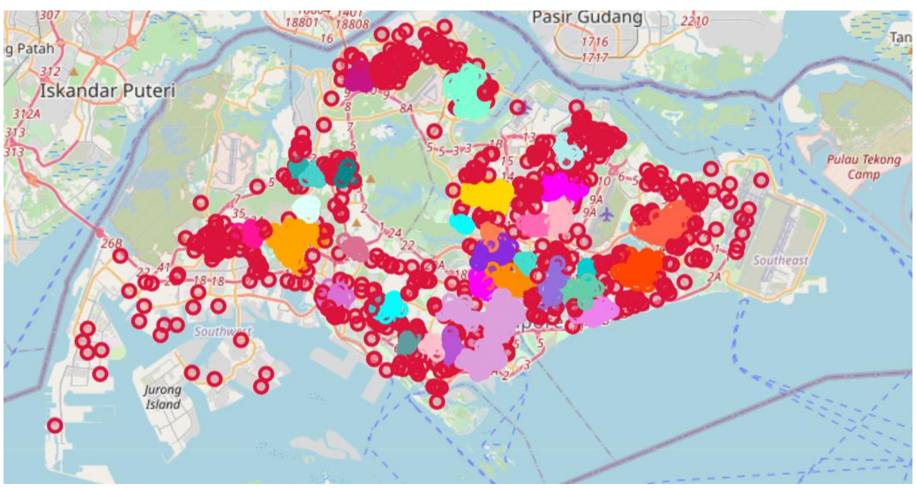


Figure 11: DBSCAN Clusters

Hybrid System: Hierarchical DBSCAN (HDSCAN) and K-Nearest Neighbor (KNN)

HDBSCAN is an algorithm extends DBSCAN by converting it into a hierarchical clustering algorithm, and then using a technique to extract a flat clustering based in the stability of clusters.

If the cluster_selection_epsilon is defined as = 0.5 kilometers (which will be converted to radians for use by haversine), min_samples=10 and distance measurement =Haversine formula, the model will form 38 clusters with Silhouette coefficient ignoring outliers =0.389 and 434 outliers with Silhouette coefficient of outliers as singletons =-0.126.

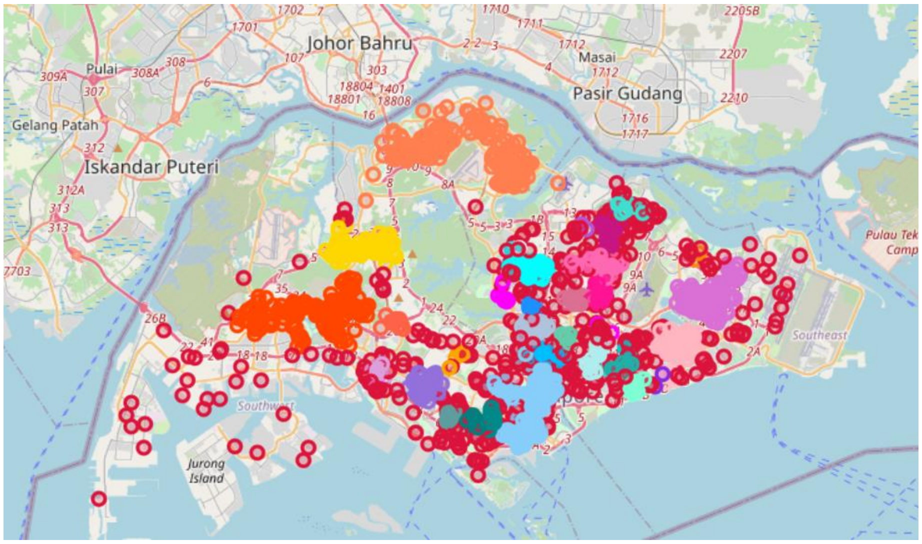


Figure 12: HDBSCAN Clusters

KNN is a supervised learning algorithm used to categorized outliers into clusters after HDBSCAN labelled the data points. The final Silhouette coefficient =0.186. The clusters produced are not very ideal also.

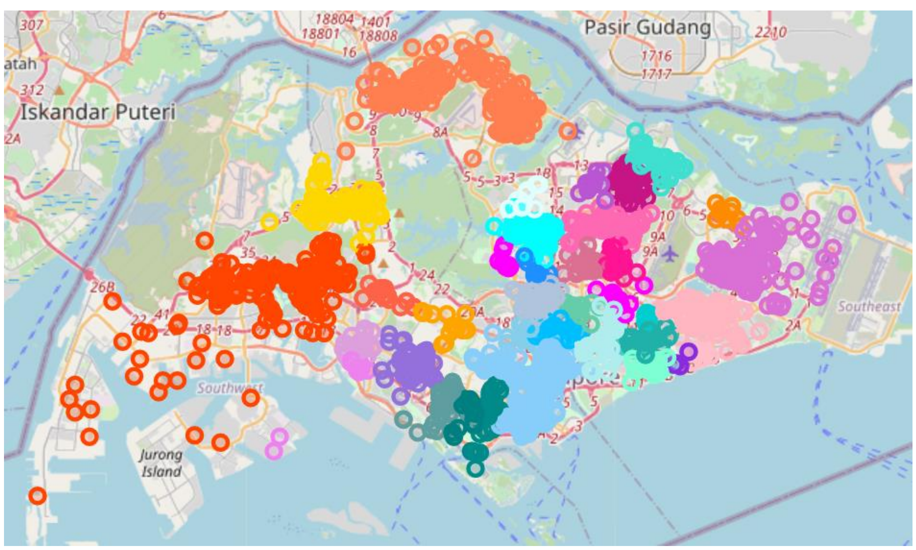


Figure 13: Hybrid System Clusters

In this case, DBSCAN and DBSCAN-based hybrid systems are more helpful in trend discovery than recommendation grouping. Therefore, KMeans is chosen as the model for location clustering.

2.4 Location Search

The input required in the search page is a postal code now. The feature can be expanded to accept address strings in the future.

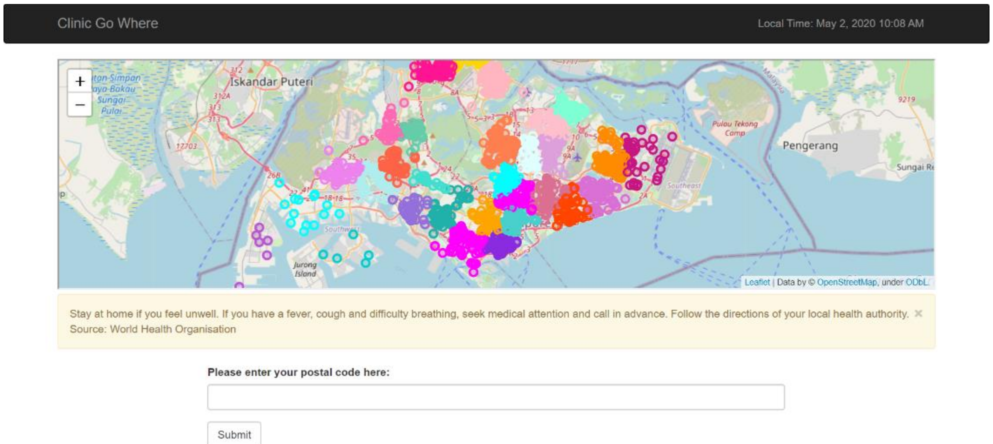


Figure 14: Search Page

The application will check the validity of an input. Users are allowed to enter 6 digits only.

Figure 15: Input Check

If the input entered is valid, OneMap API will be triggered. If the postal code entered is found in Singapore, API will return the coordinates. Else, no coordinates will be returned.

2.5 Cluster Prediction

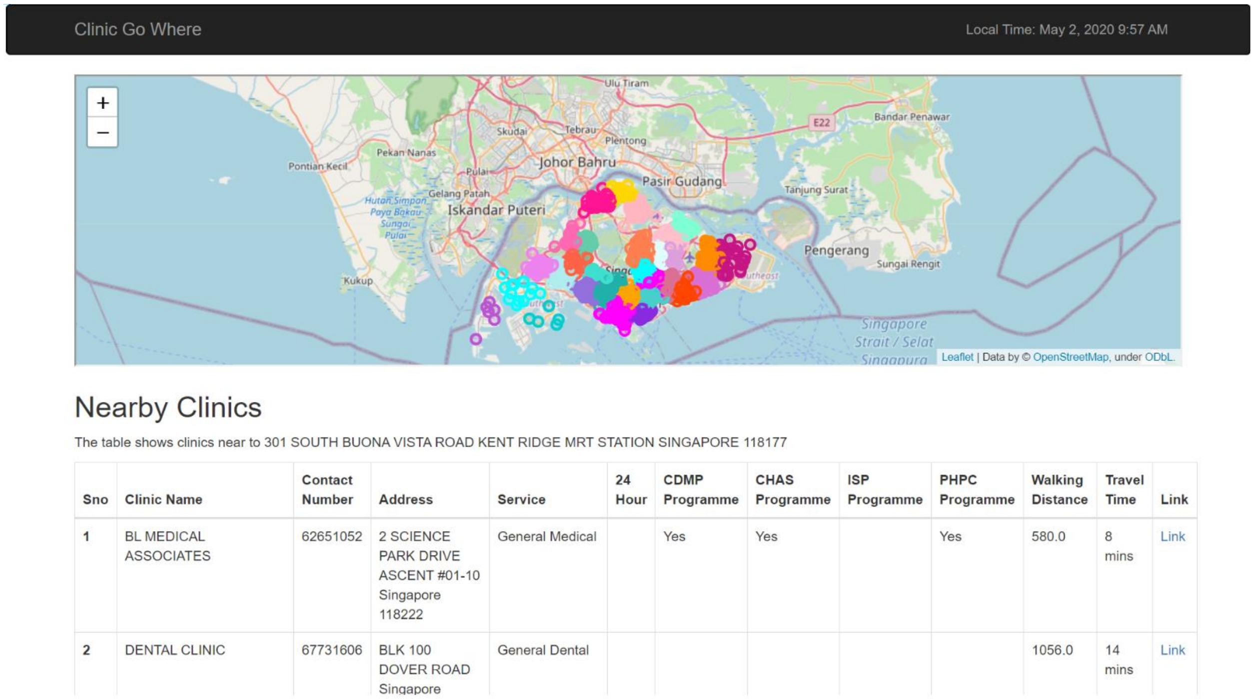
The coordinates returned by OneMap API will be feed into the pre-trained KMeans model to predict its cluster. Then all clinics under the cluster will be retrieved.

2.6 Distance and Travel Time Calculation

Google APIs are called to calculate the walking distance and travel time between a user defined location and the clinics retrieved. The calling of APIs to calculate the distance and travel time will take quite some time. Although the use of clusters has shortened the search time significantly, the component can be further improved by storing user queries into a database for faster retrieval in the future, instead of calling APIs each time.

2.7 Recommendation

The system will pick the top 10 results with shortest distance and returned to user.



Sno	Clinic Name	Contact Number	Address	Service	24 Hour	CDMP Programme	CHAS Programme	ISP Programme	PHPC Programme	Walking Distance	Travel Time	Link
1	BL MEDICAL ASSOCIATES	62651052	2 SCIENCE PARK DRIVE ASCENT #01-10 Singapore 118222	General Medical		Yes	Yes		Yes	580.0	8 mins	Link
2	DENTAL CLINIC	67731606	BLK 100 DOVER ROAD Singapore	General Dental						1056.0	14 mins	Link

Figure 16: Results Page

The system can be incorporated with chatbots and mobile messaging systems in the future to improve its usability.

3.0 Discussion and Future Enhancements

3.1 Benefits of Clustering-based Recommendation System

The main task of a recommendation system is to filter the data and recommend most relevant results to users. If the data volume is huge and is growing rapidly, the traditional techniques like filtering might suffer from problems like data sparsity and scalability. The design of clustering-based recommendation system can overcome such problems by dividing the data into smaller units based on similarity index for search and reduce the computational resources needed.

3.2 Use of Fuzzy Clustering to Enhance Recommendation

KMeans is not the most ideal model for data clustering because of its shortfall in the clustering of border points. For example, if I stay at block 484 in Jurongville, it is very likely that the model will categorise my location in orange cluster. And the system will only return the results in orange cluster to me and ignore the clinics which are in purple clusters but are also in close proximity to my house.

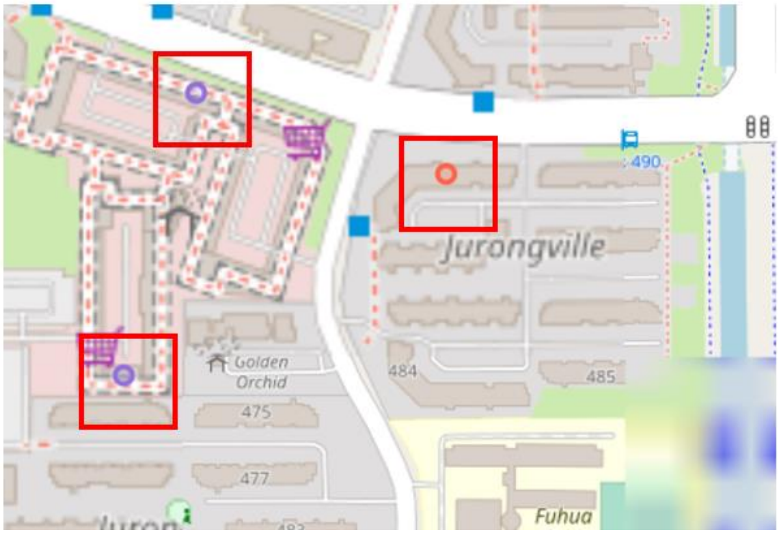


Figure 17: Discussion of Fuzzy Clustering

Fuzzy clustering is one of the possible solutions to address the limitation. Under fuzzy clustering, user inputs are assigned to each cluster with a degree of certainty. If my input is assigned with 100% certainty to a cluster, then only a cluster will be searched. If my input is assigned with a 30% certainty to purple cluster and 70% certainty to orange clusters, then both clusters will be searched. Search optimization algorithms should be implemented with the model to ensure fast speed.

3.3 Data Collection Analysis

In this project, API-based integrations are preferred for their stable performance (e.g. high response speed and consistent data format). However, when APIs are not available or available at an extremely high cost, the use of RPA tools can alleviate the pain of manual work. For example, the HCI Directory does not have an interface, it might take a person weeks to manually copy the data and the accuracy of data is not guaranteed. With RPA, the task took only took about half a day with warranted accuracy.

Additionally, if there is a need to consistently update the system data, RPA can lower the operational costs by eliminating the need to plan manpower for data maintenance.

3.4 Market Analysis

Currently, Google and various government websites are capable of recommending clinics based on user location. However, none of the services incorporate population demographic (e.g. the number of schools nearby to predict the population group) and user feedback for the general public to make well-informed choices. Therefore, there is an opportunity in the South East Asia market, as long as we can overcome the data collection concerns.

4.0 Conclusion

The final goal of this project is to design a comprehensive clinic recommendation system based on community user pattern, user feedback and geographic data. In this project, due to resources constraint, only geographic based recommendation is implemented. The aim to enable more efficient management of primary healthcare to address the problems modern-day healthcare settings.

5.0 Reference

1. Bhattacharyya, S. (2019, June 9). DBSCAN Algorithm: Complete Guide and Application with Python Scikit-Learn. Retrieved May 03, 2020, from https://towardsdatascience.com/dbscan-algorithm-complete-guide-and-application-with-python-scikit-learn-d690cbae4c5d

2. Bowman, D. (2020, January 09). Robotic Process Automation Vs API Integration. Retrieved May 02, 2020, from https://www.uipath.com/blog/rpa-vs-api-integration

3. Flu Go Where. (2020). Retrieved May 02, 2020, from https://www.flugowhere.gov.sg/

4. Choon, H. and Kwong. (2014) Healthcare in Singapore: the Present and Future , from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4293982/

5. K. (2020, April 28). TagUI. Retrieved May 02, 2020, from https://github.com/kelaberetiv/TagUI

6. Listing of Licensed Healthcare Institutions. (n.d.). Retrieved May 02, 2020, from http://hcidirectory.sg/hcidirectory/

7. Lutins, E. (2017, September 15). DBSCAN: What is it? When to Use it? How to use it. Retrieved May 03, 2020, from https://medium.com/@elutins/dbscan-what-is-it-when-to-use-it-how-to-use-it-8bd506293818

8. McInnes, L., Healy, J., & Astels, S. (n.d.). How HDBSCAN Works. Retrieved May 02, 2020, from https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

9. National Health Expenditure Projections 2018-2027. (n.d.). Retrieved May 03, 2020, from https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/ForecastSummary.pdf

10. Public Health Preparedness Clinic (PHPC). (2019). Retrieved May 03, 2020, from https://www.primarycarepages.sg/practice-management/moh-national-schemes/public-health-preparedness-clinic-(phpc)

11. Rana, A. (2019, April 18). Build Your Own Clustering Based Recommendation Engine in 15 minutes !! Retrieved May 02, 2020, from https://towardsdatascience.com/build-your-own-clustering-based-recommendation-engine-in-15-minutes-bdddd591d394

12. Robotic Process Automation and Application Programming Interface: A Match Made in Heaven or Not? (2019, January 07). Retrieved May 02, 2020, from

https://www.accelirate.com/robotic-process-automation-application-programming-interface-match-made-in-heaven-or-not/

13. Sharma.A, Estimating the Future Health and Aged Care Expenditure in Australia with Changes in Morbidity , https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0201697

14. Selecting the number of clusters with silhouette analysis on KMeans clustering. (n.d.). Retrieved May 03, 2020, from https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

15. Sharma, P. (2018, June 21). Comprehensive Guide to build Recommendation Engine from scratch. Retrieved May 02, 2020, from https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-recommendation-engine-python/?utm_source=blog

16. Smith, Y. (2019, February 27). Challenges in Primary Care. Retrieved May 02, 2020, from https://www.news-medical.net/health/Challenges-in-Primary-Care.aspx

17. T. (n.d.). Current health expenditure (% of GDP). Retrieved May 02, 2020, from https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS

18. United Nations, Ageing, https://www.un.org/en/sections/issues-depth/ageing/

19. Veness, C. (n.d.). Calculate distance, bearing and more between Latitude/Longitude points. Retrieved May 02, 2020, from https://www.movable-type.co.uk/scripts/latlong.html