REVOLUTIONIZING BUS ROUTES: A DATA-DRIVEN APPROACH

SHEA YEE, KHOO

AUG 2024

AGENDA

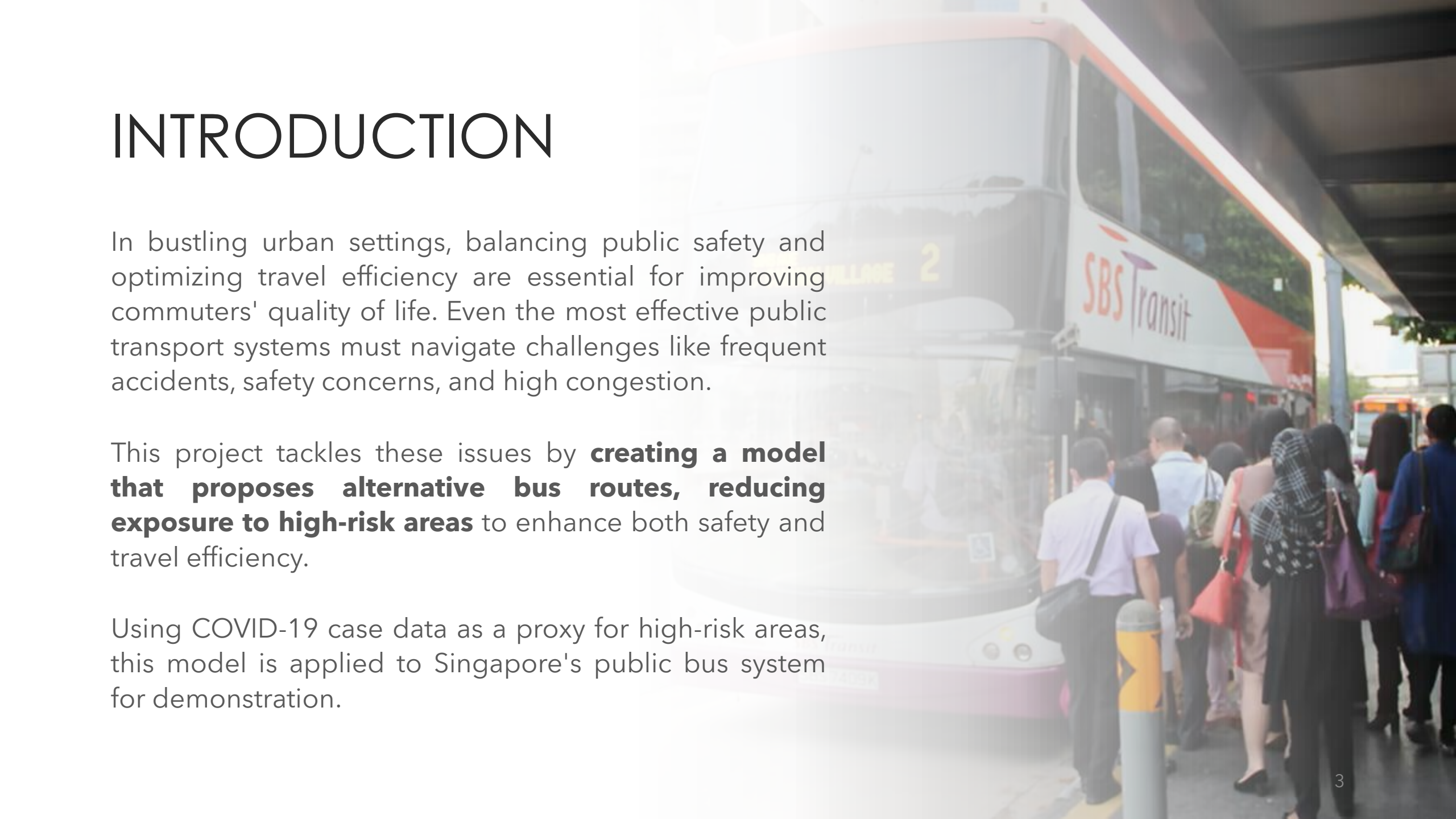
- Introduction
- Problem Statement
- Proposed Solution & Benefits
- Datasets Used
- Workflow Overview
- Model Development & Performance Assessment
- Summary & Next Steps
- Model Limitation & Suggestion

INTRODUCTION

In bustling urban settings, balancing public safety and optimizing travel efficiency are essential for improving commuters' quality of life. Even the most effective public transport systems must navigate challenges like frequent accidents, safety concerns, and high congestion.

This project tackles these issues by **creating a model that proposes alternative bus routes, reducing exposure to high-risk areas** to enhance both safety and travel efficiency.

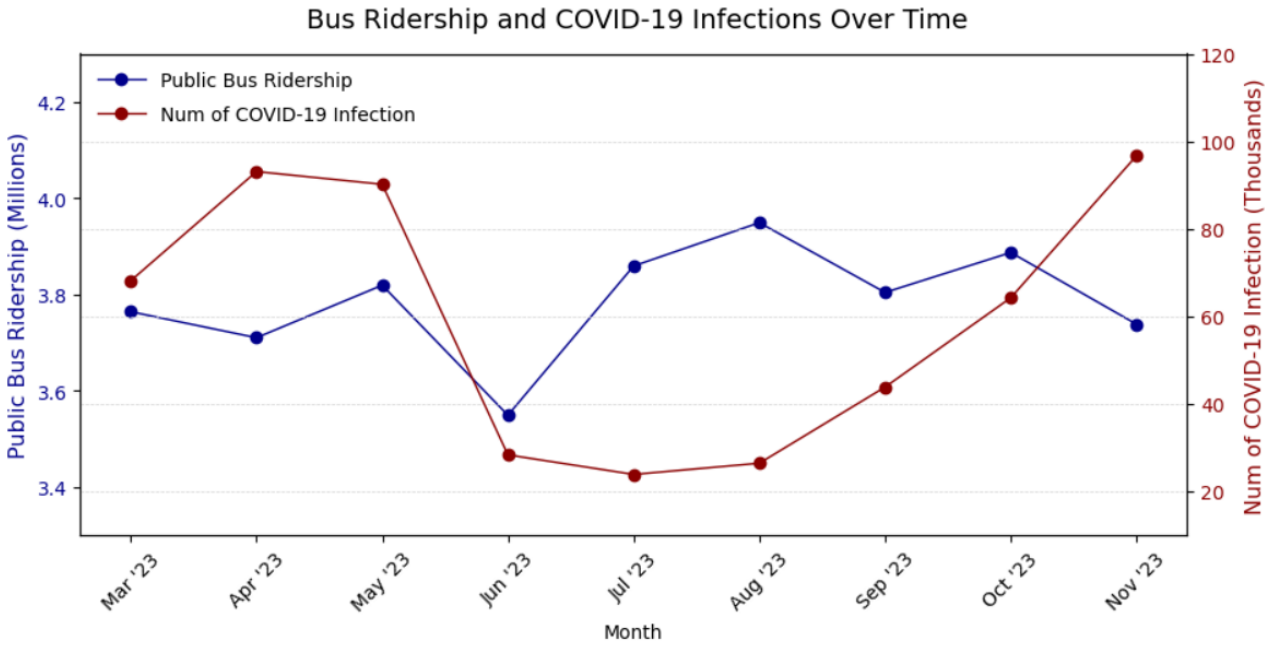
Using COVID-19 case data as a proxy for high-risk areas, this model is applied to Singapore's public bus system for demonstration.



PROBLEM STATEMENT

With the continued presence of COVID-19 cases and significant passenger volumes at certain locations and times in Singapore, there is a challenge in adjusting bus routes to minimize exposure to high-risk areas.

Effectively navigating around high congestion and frequent incidents is crucial for improving commuter safety and efficiency.



- **Singapore's public buses experience a substantial monthly ridership** of ave. 3.8 million, marking a 0.3 million increase compared to the same period last year.
- **COVID-19 cases remain significant**, with monthly ave. of 59k cases (Mar to Nov 2023).
- Data sources:
  - Land Transport DataMall (https://datamall.lta.gov.sg/content/datamall/)
  - Ministry of Health (https://beta.data.gov.sg/collections/522/view/)

PROPOSED SOLUTION & BENEFITS

The proposed solution employs a dual approach to enhance bus route optimization.

- First, **unsupervised machine learning techniques** are used to cluster Covid-19 high-risk areas.
- Second, a **supervised machine learning model** predicts passenger volume at specific locations and times.
- These models work together to suggest alternative bus routes that minimize both exposure to COVID-19 clusters and high passenger volume areas, while also minimizing travel distance.

This project offers the following benefits:

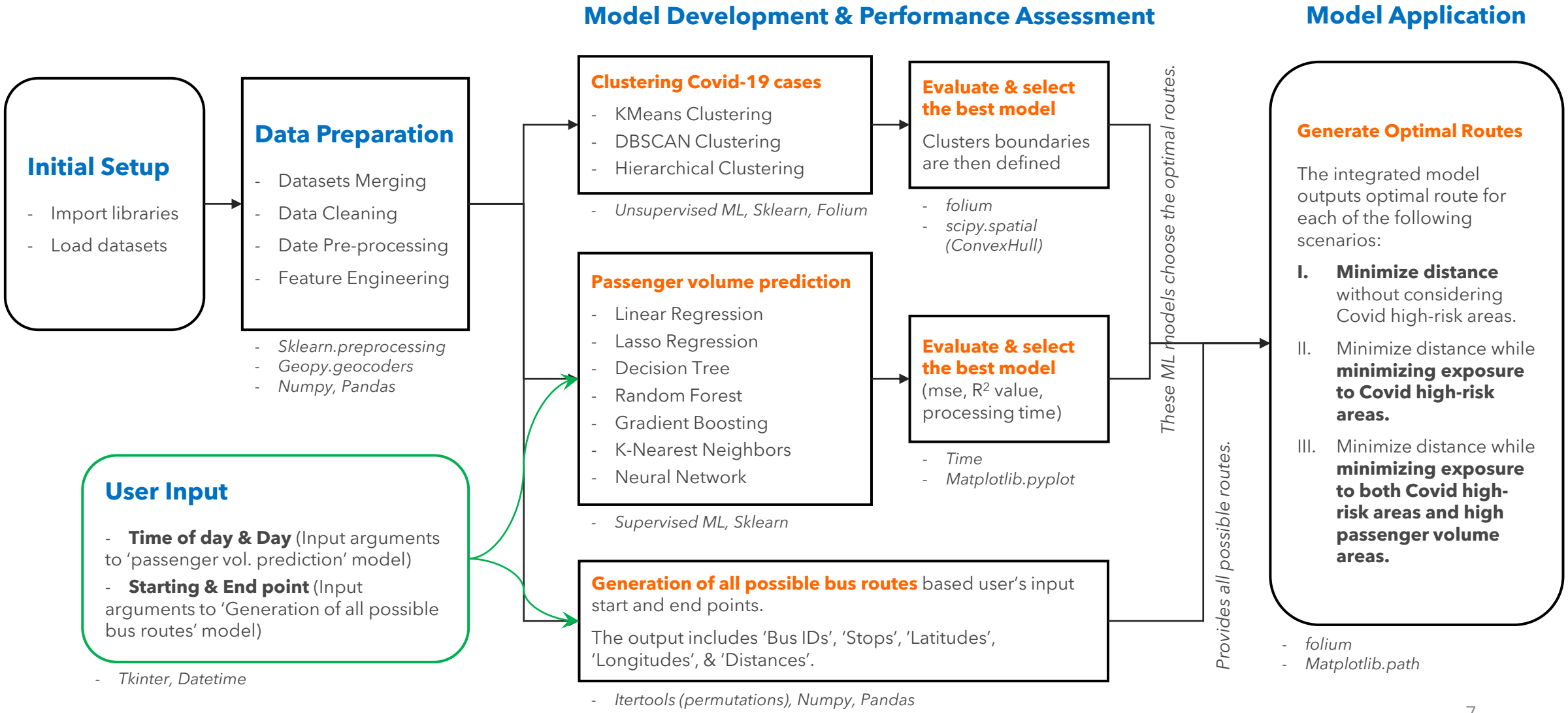
- **Public Bus Commuters**: The model assists in **selecting the optimal bus routes** based on different scenarios.
- **Government and Bus Operators:** Collaborate with both government agencies and bus operators to integrate the model into existing systems. This collaboration would **enhance scheduling and routing**, creating a more efficient and effective public transit system.

DATASETS USED

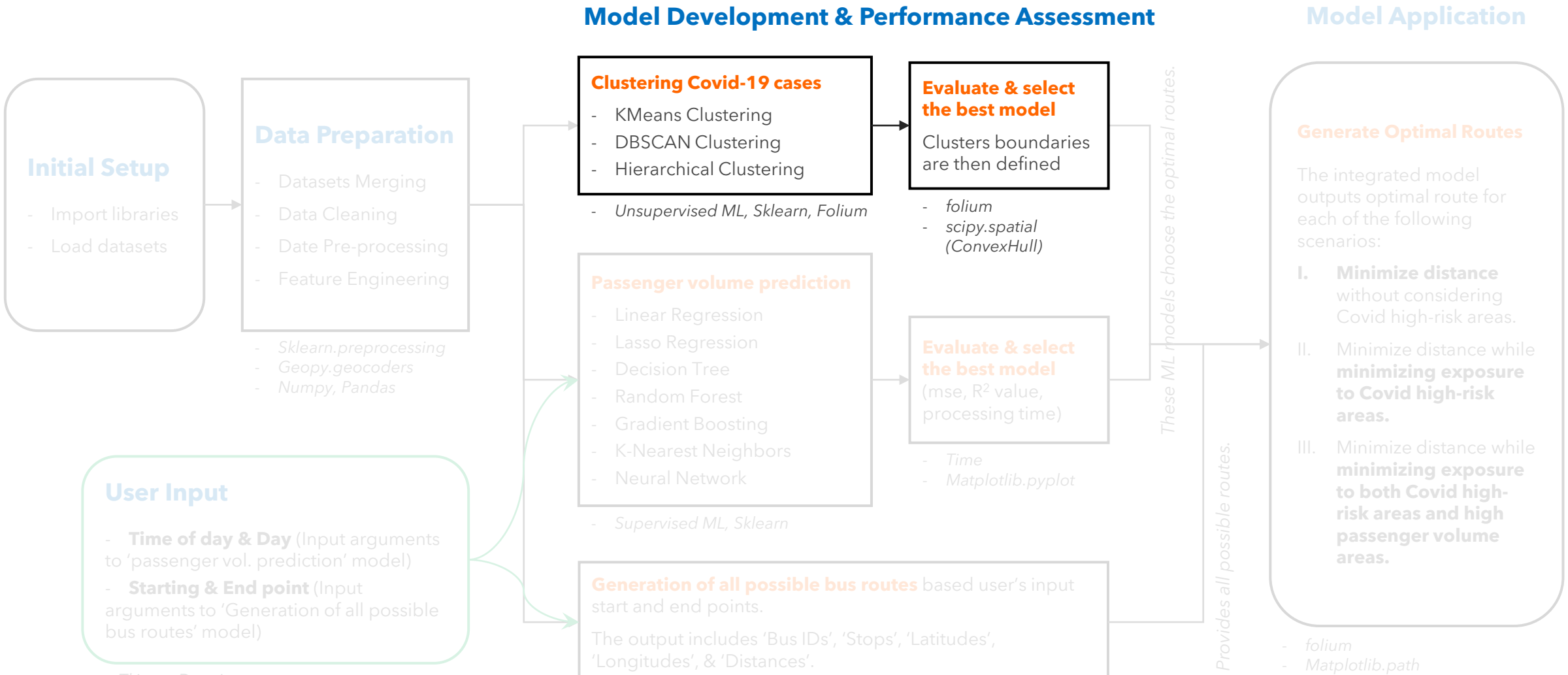
- This project utilizes a total of 7 datasets from 4 different sources. Details are summarized in table below.
- Datasets schema are included in the appendix slides.

#	Dataset Source	Dataset URL	Dataset Name	Dataset Description	Dataset Size	Columns Used	Dataset Purpose
1	Ministry of Health (MOH)	https://beta.data.gov.sg/datasets/d_554627df56037a1296507f35c374f79d/view/	*'Covid19CaseDetails.csv'*	Singapore covid-19 cases details.	(77 rows, 14 cols)	- Places visited by patients - Residing location of patients	To **cluster COVID-19 cases** into groups and identify the locations of these clusters.
2	Land Transport DataMall	https://www.kaggle.com/datasets/yorkyong/singapore-passenger-volume-by-train-stations	*'transport_node_train_202308.csv'* *'transport_node_train_202309.csv'* *'transport_node_train_202310.csv'*	Singapore passenger volume by train stations.	Each dataset consists of (6820 rows, 7 cols)	- Day - Time of day - Station Code - Total tap in - Total tap out	Datasets are merged on the 'Station Code' column, and then used to train a **machine learning model that predicts passenger volumes** at specific locations and times.
3	Land Transport DataMall	https://www.kaggle.com/datasets/shengjunlim/singapore-mrt-lrt-stations-with-coordinates	*'MRT Stations.csv'*	List of MRT& LRT stations in Singapore with geographic coordinates in decimal degrees.	(170 rows, 7 cols)	- Station Code - Station Latitude - Station Longitude	
4	Land Transport Authority (LTA)	https://www.kaggle.com/datasets/gowthamvarma/singapore-bus-data-land-transport-authority	*'bus_routes.csv'*	Singapore public bus routes details.	(26317 rows, 13 cols)	- Bus service No. - Bus direction - Bus Stop Code - Distance	Datasets are merged on the 'Bus Stop Code' column. A function is then created to **generate all possible bus route options given starting and ending points** as inputs.
			'bus-stops.csv'	List of bus stops in Singapore with geographic coordinates in decimal degrees.	(5021 rows, 6 cols)	- Bus Stop Code - Road Name - Stop Latitude - Stop Longitude	

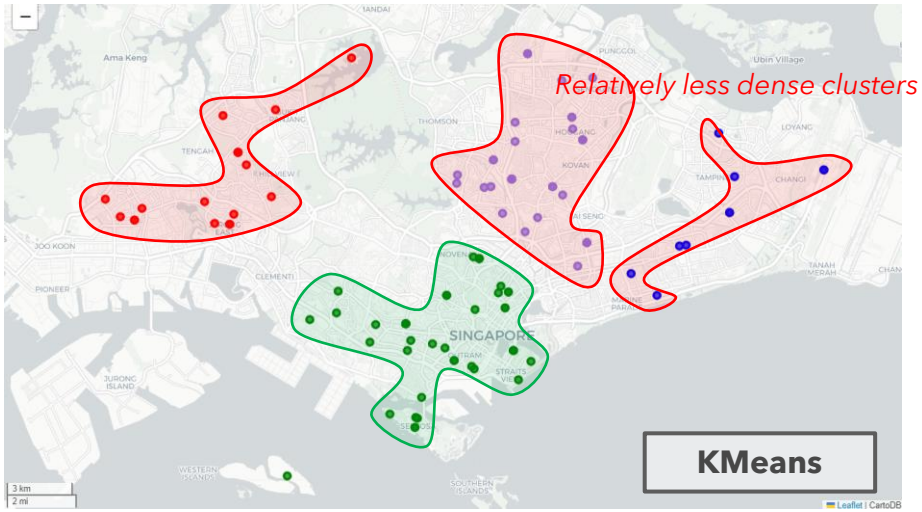
WORKFLOW OVERVIEW

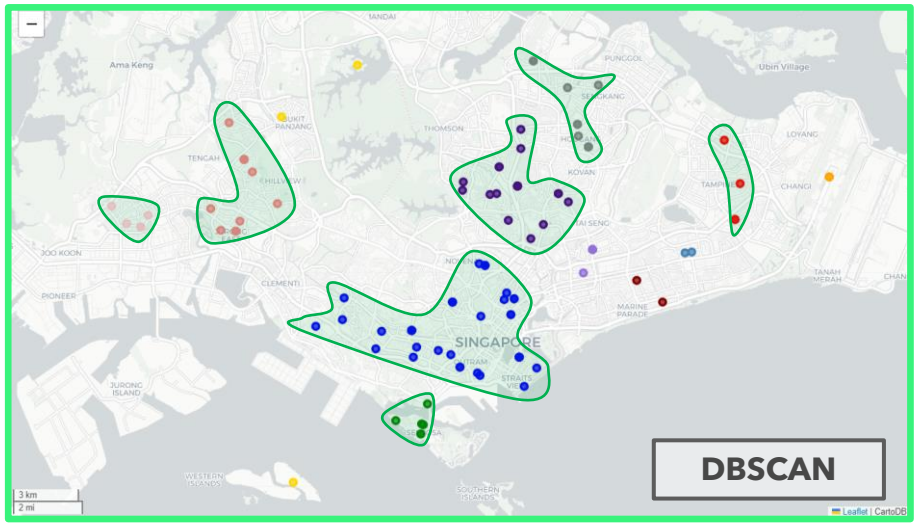


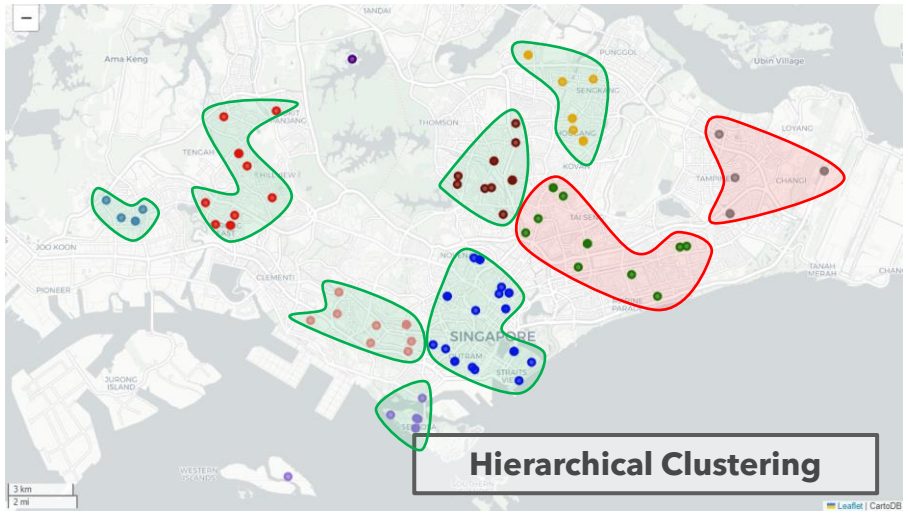
WORKFLOW OVERVIEW



CLUSTERING COVID-19 CASES





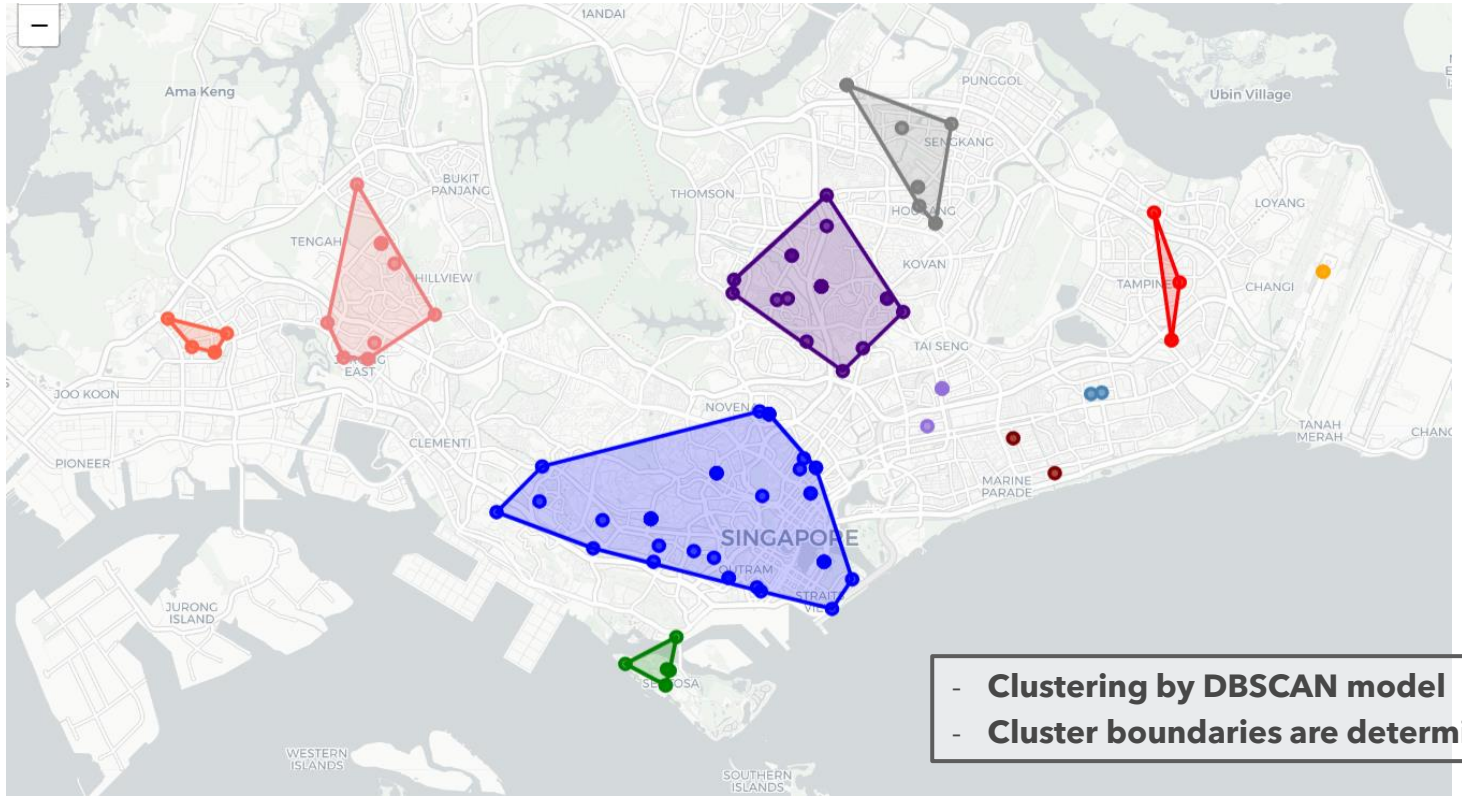


Three **unsupervised machine learning** models are employed to cluster COVID-19 cases by locations in Singapore.

DBSCAN is chosen as the best model due to its ability to create the most densely packed clusters. Dense clusters are crucial because overly large clusters can overly restrict route options, potentially leading to unnecessary detours when generating optimal routes.

In contrast, clusters defined by KMeans or Hierarchical clustering are less dense, making them less suitable for accurately defining high-risk areas for COVID-19.

DEFINING CLUSTERS BOUNDARIES

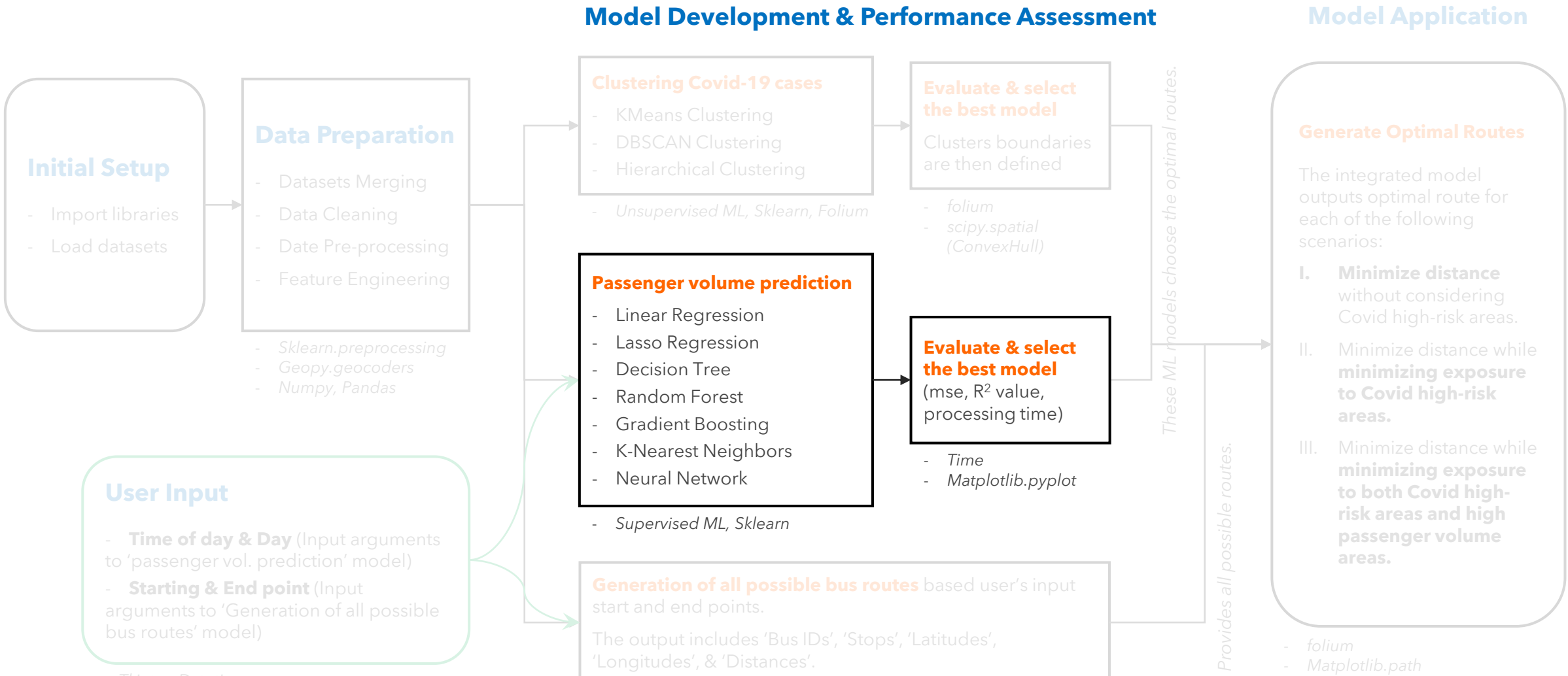


- **Clustering by DBSCAN model**
- **Cluster boundaries are determined using the 'ConvexHull' method**

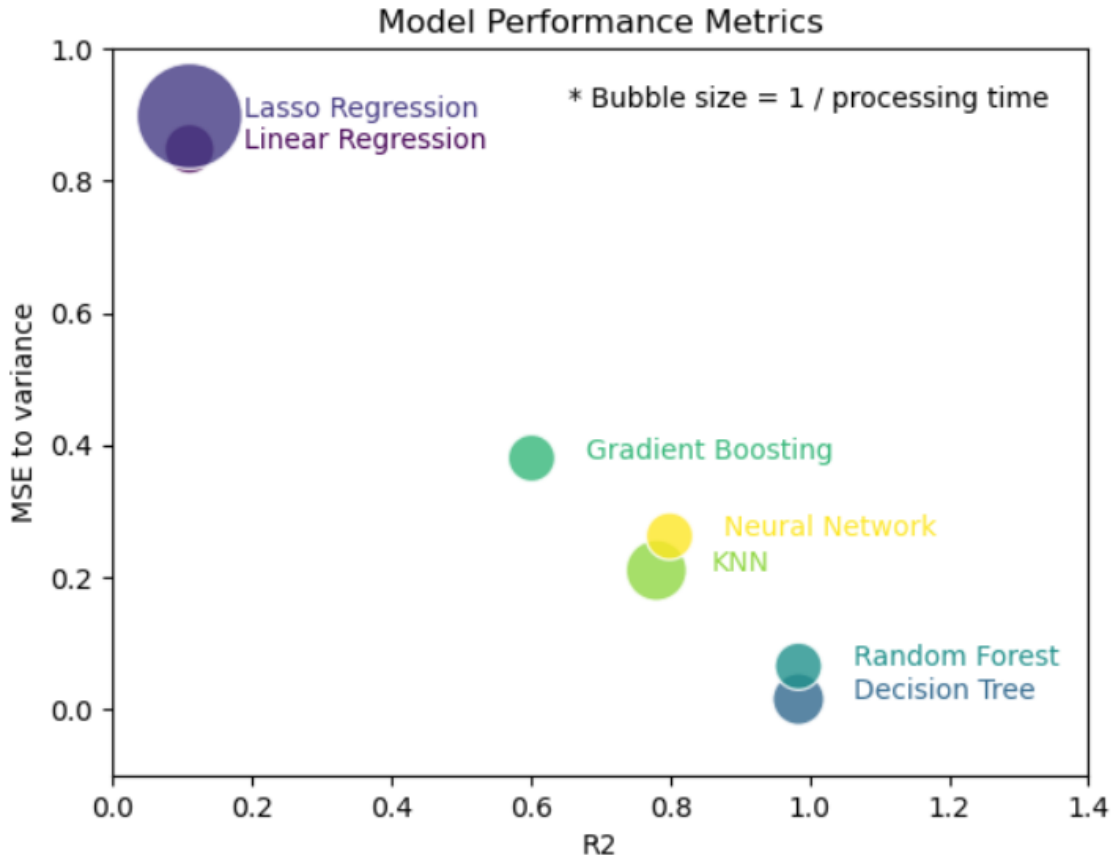
Cluster boundaries are then defined as polygons to enable the final model to assess whether specific bus stop coordinates fall within COVID-19 cluster regions.

The 'ConvexHull' algorithm is used to outline these boundaries, providing the smallest convex shape that encloses all the points within each cluster.

WORKFLOW OVERVIEW



PASSENGER VOLUME PREDICTION



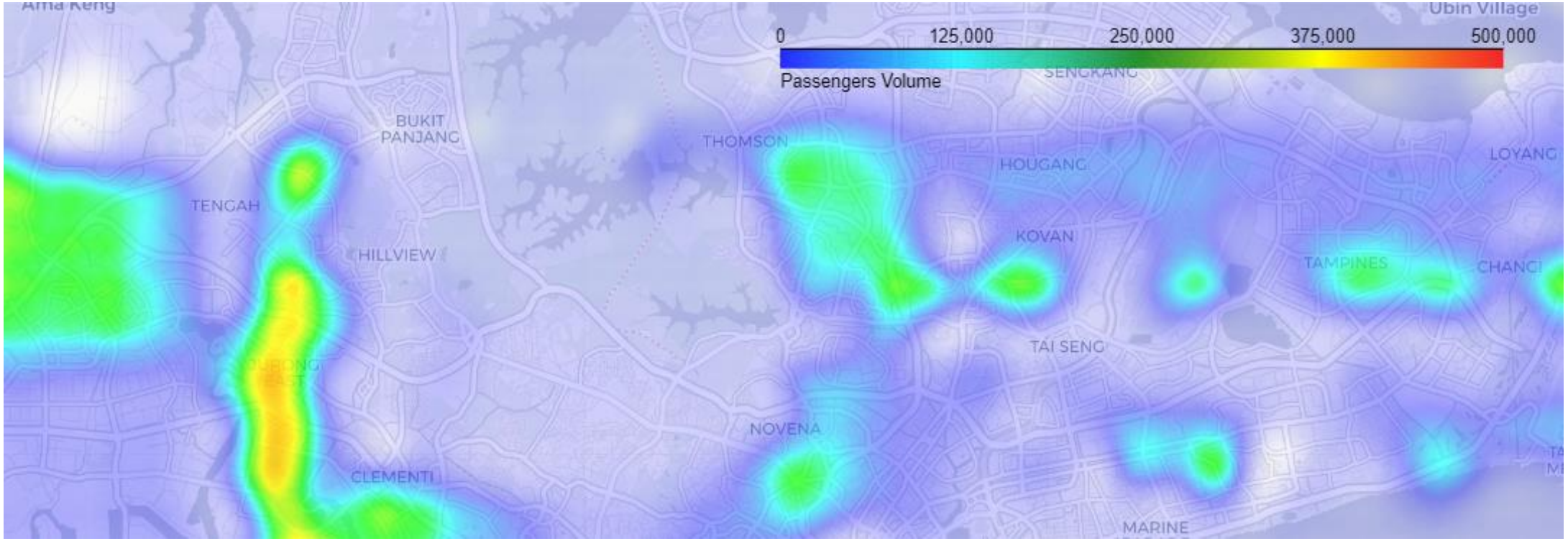
To predict passenger volume at specific locations and times, **supervised machine learning models** are trained using datasets from Land Transport DataMall.

Seven regression models are trained on four features: "Day", "Time of Day", "Latitude", and "Longitude".

The Decision Tree Regression model is selected as the best due to its:

- Lowest Mean Squared Error to variance ratio (MSE / variance = 0.015)
- Highest R-squared value ($R^2 = 0.984$)
- Efficient processing time (~0.081 seconds) compared to other models

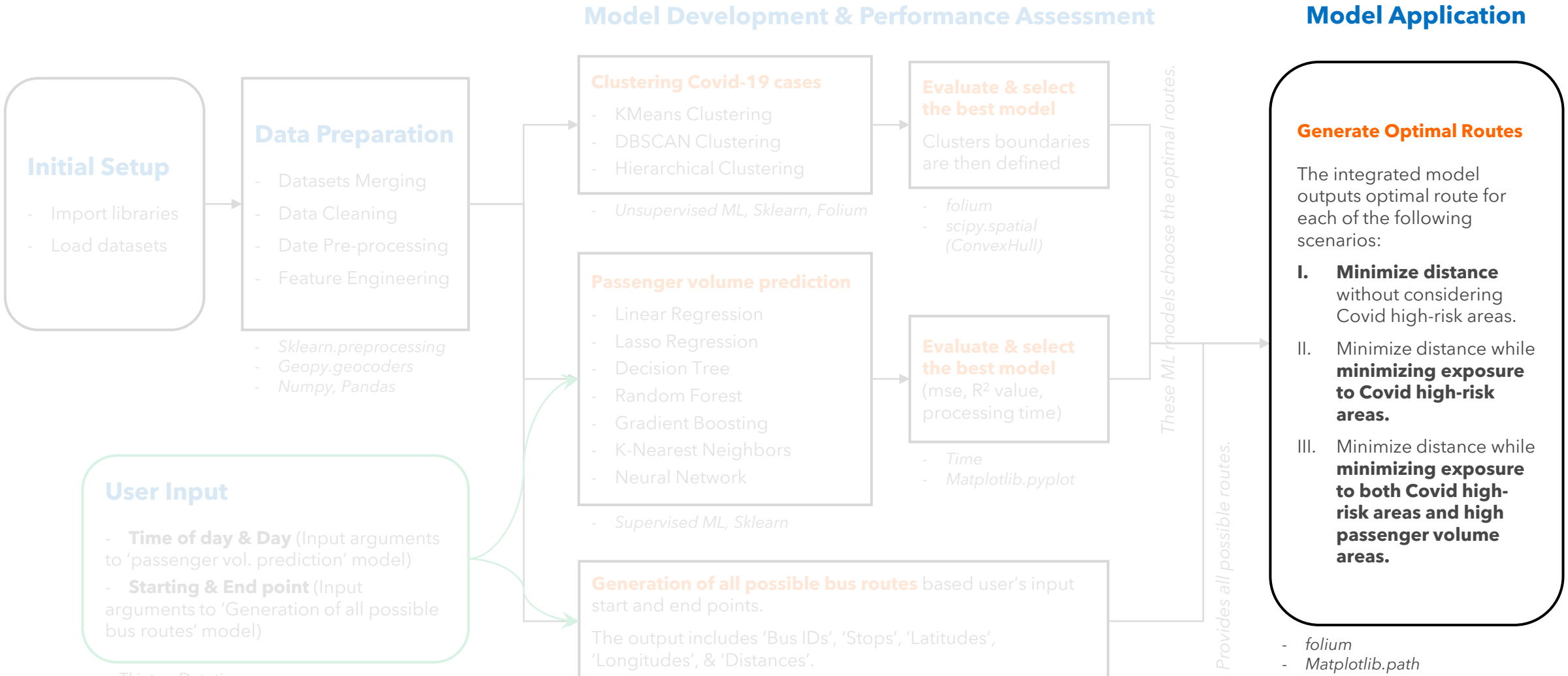
PASSENGER VOLUME PREDICTION



Passenger volume predictions from the Decision Tree Regression Model for specific areas in Singapore are shown for illustrative purposes. These predictions vary based on the day, time of day, and locations provided by the model user.

The example illustrates passenger volumes between 18:00 and 19:00 on a weekday, highlighting regions with high density (Jurong East, Clementi, Novena, AMK, Serangoon, Tai Seng, Tampines, etc.) due to the typical after-work rush when people are traveling from their workplaces.

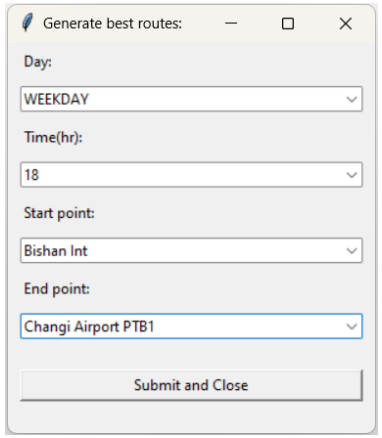
WORKFLOW OVERVIEW



GENERATE OPTIMAL ROUTES

Case example:

Input



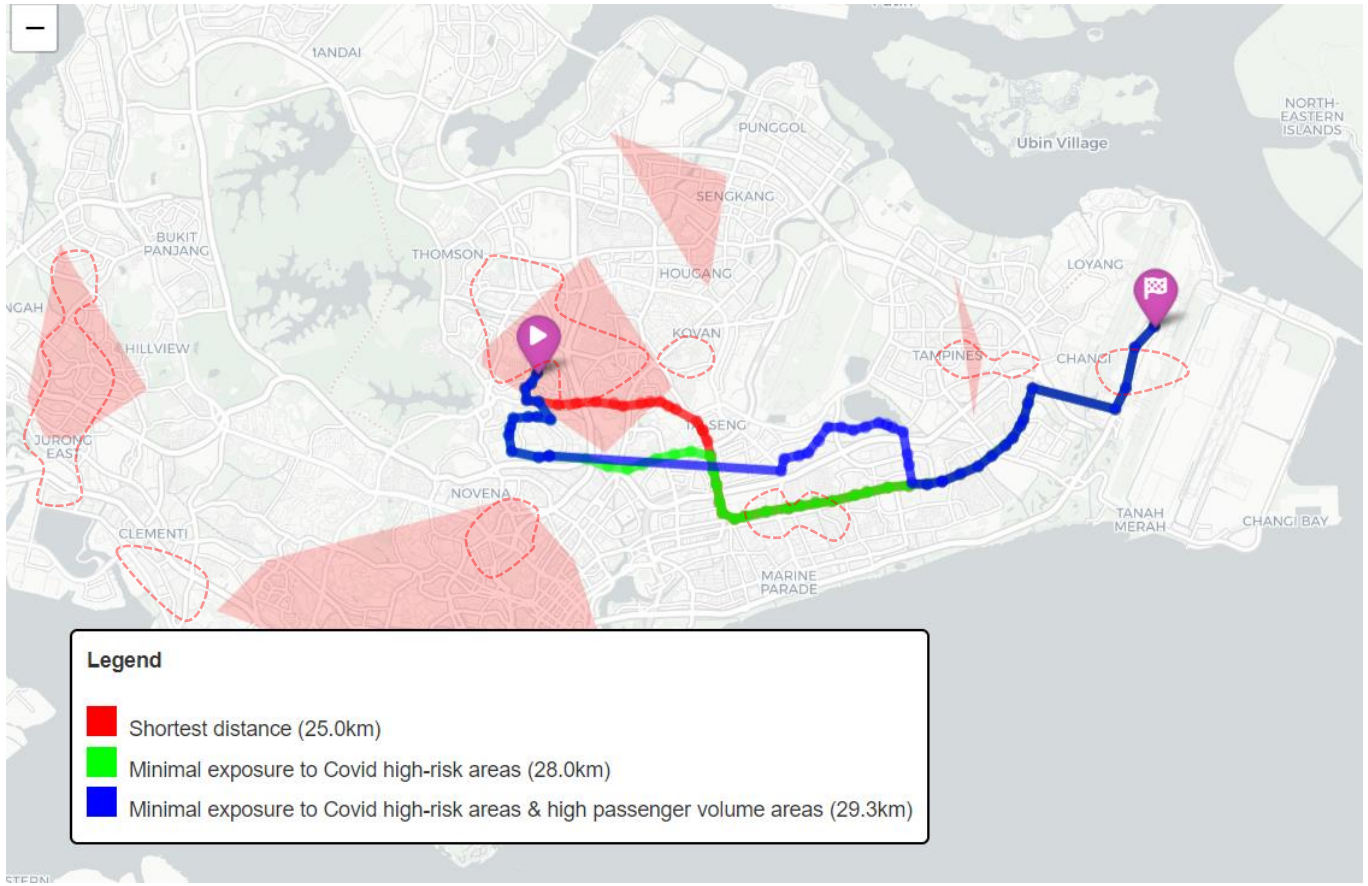
Output

```
Scenario #1. Optimal route with shortest distance.

Route: Bus59 (dir_1) Bishan Int --> Bus59 (dir_1) OPP BLK 998 --> Bus93 (dir_1) OPP BLK 998 --> Bus93 (dir_1) Opp Asiawide Ind Bldg --> Bus24 (dir_1) Opp
Asiawide Ind Bldg --> Bus24 (dir_1) Changi Airport PTB1

- Stops in cluster (count): 12
- Stops with high passenger vol. (count): 51
- Total distance (km): 25.0

================================================================================================================================================================

Scenario #2. Optimal route that minimizes exposure to Covid high-risk areas.

Route: Bus59 (dir_1) Bishan Int --> Bus59 (dir_1) Opp Toa Payoh Stadium --> Bus8 (dir_2) Opp Toa Payoh Stadium --> Bus8 (dir_2) Citipoint Ind Cplx --> Bu
s24 (dir_1) Citipoint Ind Cplx --> Bus24 (dir_1) Changi Airport PTB1

- Stops in cluster (count): 4
- Stops with high passenger vol. (count): 54
- Total distance (km): 28.0

================================================================================================================================================================

Scenario #3. Optimal route that minimizes exposure to both Covid high-risk areas and high passenger volume areas.

Route: Bus59 (dir_1) Bishan Int --> Bus59 (dir_1) Opp Toa Payoh Stadium --> Bus506 (dir_1) Opp Toa Payoh Stadium --> Bus506 (dir_1) The Tanamera Condo --
> Bus24 (dir_1) The Tanamera Condo --> Bus24 (dir_1) Changi Airport PTB1

- Stops in cluster (count): 4
- Stops with high passenger vol. (count): 47
- Total distance (km): 29.3
```

- The final model generates optimal route for each of the three scenarios.
- The output includes: (1) **route**, (2) **number of stops in clusters**, (3) **number of stops with high passenger volume**, and (4) **total distance in kilometers**.

GENERATE OPTIMAL ROUTES

Case example (continued):



Covid high-risk area

Areas of high passenger volume, where “high” is defined as a volume greater than the 75th percentile of all recorded station volumes in the dataset.

The optimal routes for each scenario are visualized on a Folium map.

- **Scenario #1** aims to achieve the shortest possible travel distance, hence it may include bus stops in Covid cluster regions.
- **Scenario #2** minimizes exposure to Covid cluster regions while minimizing travel distance.
- **Scenario #3** minimizes exposure to both Covid clusters and areas with high passenger volume (near Marina Parade) but results in a longer travel distance.

The model provides flexibility by offering different options for users to choose from based on their specific needs.

SUMMARY & NEXT STEPS

This project has developed a model that potentially enhances public bus system by suggesting safer routes and providing flexible options for commuters. The model uses unsupervised learning to identify hazardous clusters and supervised learning to predict passenger volumes, offering a comprehensive approach to route optimization.

It is recommended to integrate real-time data feeds into the system to allow for dynamic adjustments of bus routes. Incorporating real-time information would help address emerging safety issues and adapt to changes in passenger volume.

Further developments:

- **Deploy in a Public App**: Implement the model in a mobile app for real-time route optimization for commuters.
- **Collaborate with Bus Operators**: Work with bus operators to integrate the model, enhancing scheduling and routing efficiency for a more effective public transportation system.

MODEL LIMITATION & SUGGESTION

Limitation:

The supervised regression model used to predict passenger volume is trained on data from train stations. Consequently, the passenger volume at bus stops is an estimation derived from patterns observed at nearby train stations.

Suggestion:

To improve the accuracy of passenger volume predictions at bus stops, it is necessary to collect data specifically about bus stop passenger volumes to train the supervised regression model more effectively. The steps involved in training and using the model will remain unchanged.

Thank You

https://www.linkedin.com/in/shea-yee-khoo-0b14b930/

https://github.com/khoosheayee/data-science-projects

khoosheayee@live.com

Appendices

DATASETS SCHEMA

- **Dataset file name**: 'Covid19CaseDetails.csv'
- **Dataset URL:** https://beta.data.gov.sg/datasets/d_554627df56037a1296507f35c374f79d/view/
- **Dataset Description:** Singapore covid-19 cases details.
- **Dataset Size:** 77 rows, 14 columns
- **Purpose of using dataset:** To cluster COVID-19 cases into groups and identify the locations of these clusters.

case_id	age	gender	nationality	imported_local	place	public_healthcare_institution	status	date_of_confirmation	date_of_discharge	places_visited	residing_loca
Case 1	66	M	Chinese	Imported	Wuhan	Singapore General Hospital	Hospitalised	1/23/2020	-	N/A	Shangri-La Sentosa Res
Case 2	53	F	Chinese	Imported	Wuhan	National Centre for Infectious Disease	Discharged	1/24/2020	2/7/2020	Raffles Hospital, Tan Tock Seng Emergency Depa...	J8 ho Townshend
Case 3	37	M	Chinese	Imported	Wuhan	Singapore General Hospital	Hospitalised	1/24/2020	-	N/A	
Case 4	36	M	Chinese	Imported	Wuhan	Sengkang General Hospital	Discharged	1/25/2020	2/12/2020	USS, Vivocity	Village H Ser
Case 5	56	F	Chinese	Imported	Wuhan	National Centre for Infectious Disease	Hospitalised	1/27/2020	-	Tan Tock Seng Hospital	Ceylon

'places_visited' and 'residing_location' are used to extract their geographic coordinates, which will then serve as features for clustering models.

DATASETS SCHEMA

- **Dataset file name**: 'transport_node_train_202308.csv', 'transport_node_train_202309.csv', and 'transport_node_train_202310.csv'
- **Dataset URL:** https://www.kaggle.com/datasets/yorkyong/singapore-passenger-volume-by-train-stations
- **Dataset Description:** Singapore passenger volume by train stations.
- **Dataset Size:** Each dataset consists of (6820 rows, 7 cols)
- **Purpose of using dataset:** These three datasets are concatenated and then merged with 'MRT Stations.csv' (next slide) on the 'Station Code' column. The resulting dataset is used to train a machine learning model that predicts passenger volumes at specific locations and times.

	YEAR_MONTH	DAY_TYPE	TIME_PER_HOUR	PT_TYPE	PT_CODE	TOTAL_TAP_IN_VOLUME	TOTAL_TAP_OUT_VOLUME
0	2023-08	WEEKDAY	22	TRAIN	NS28	752	311
1	2023-08	WEEKENDS/HOLIDAY	22	TRAIN	NS28	612	223
2	2023-08	WEEKENDS/HOLIDAY	0	TRAIN	DT10/TE11	37	242
3	2023-08	WEEKDAY	0	TRAIN	DT10/TE11	86	445
4	2023-08	WEEKDAY	10	TRAIN	EW16/NE3/TE17	28179	39454

- Feature engineering includes converting 'PT_CODE' to geographic coordinates, and encoding 'DAY_TYPE' as integers.
- These features, along with other key columns, are used to train the supervised ML model that predicts passenger volume.

DATASETS SCHEMA

- **Dataset file name**: 'MRT Stations.csv'
- **Dataset URL:** https://www.kaggle.com/datasets/shengjunlim/singapore-mrt-lrt-stations-with-coordinates
- **Dataset Description:** List of MRT& LRT stations in Singapore with geographic coordinates in decimal degrees.
- **Dataset Size:** 170 rows, 7 columns
- **Purpose of using dataset:** This dataset provides station names and their geographic coordinates. It is merged with dataset of previous slide to train a model predicting passenger volumes at specific locations and times.

	Unnamed: 0	OBJECTID	STN_NAME	STN_NO	geometry	Latitude	Longitude
0	0	1	EUNOS MRT STATION	EW7	POINT (103.9032524667383 1.319778951553637)	1.319779	103.903252
1	1	2	CHINESE GARDEN MRT STATION	EW25	POINT (103.7325967380734 1.342352820874744)	1.342353	103.732597
2	2	3	KHATIB MRT STATION	NS14	POINT (103.8329799077383 1.417383370153547)	1.417383	103.832980
3	3	4	KRANJI MRT STATION	NS7	POINT (103.7621654109002 1.425177698770448)	1.425178	103.762165
4	4	5	REDHILL MRT STATION	EW18	POINT (103.816816670149 1.289562726402453)	1.289563	103.816817

Station names and corresponding geographic coordinates. Coordinates are used as a feature in training the ML model.

DATASETS SCHEMA

- **Dataset file name**: 'bus_routes.csv'
- **Dataset URL:** https://www.kaggle.com/datasets/gowthamvarma/singapore-bus-data-land-transport-authority/
- **Dataset Description:** Singapore public bus routes details.
- **Dataset Size:** 26317 rows, 13 columns
- **Purpose of using dataset:** Dataset is merged with dataset on the next slide, on the 'BusStopCode' column. A function is then created to generate all possible bus route options given starting and ending points as inputs.

	Unnamed: 0	ServiceNo	Operator	Direction	StopSequence	BusStopCode	Distance	WD_FirstBus	WD_LastBus	SAT_FirstBus	SAT_LastBus	SUN_FirstBus	SUN_LastBus
0	0	10	SBST	1	1	75009	0.0	0500	2300	0500	2300	0500	2300
1	1	10	SBST	1	2	76059	0.6	0502	2302	0502	2302	0502	2302
2	2	10	SBST	1	3	76069	1.1	0504	2304	0504	2304	0503	2304
3	3	10	SBST	1	4	96289	2.3	0508	2308	0508	2309	0507	2308
4	4	10	SBST	1	5	96109	2.7	0509	2310	0509	2311	0508	2309

Bus service number, direction, bus stop code, and distance are some of the features used to generate bus routes.

DATASETS SCHEMA

- **Dataset file name**: 'bus-stops.csv'
- **Dataset URL:** https://www.kaggle.com/datasets/gowthamvarma/singapore-bus-data-land-transport-authority/
- **Dataset Description:** List of bus stops in Singapore with geographic coordinates in decimal degrees.
- **Dataset Size:** 5021 rows, 6 columns
- **Purpose of using dataset:** This dataset is merged with dataset of previous slide, on the 'BusStopCode' column. A function is then created to generate all possible bus route options given starting and ending points as inputs.

	Unnamed: 0	BusStopCode	RoadName	Description	Latitude	Longitude
0	0	481	Woodlands Rd	BT PANJANG TEMP BUS PK	1.383764	103.758300
1	1	1012	Victoria St	Hotel Grand Pacific	1.296848	103.852536
2	2	1013	Victoria St	St. Joseph's Ch	1.297710	103.853225
3	3	1019	Victoria St	Bras Basah Cplx	1.296990	103.853022
4	4	1029	Nth Bridge Rd	Cosmic Insurance Bldg	1.296673	103.854414

Combined with the previous dataset, a new dataframe is created with the columns: ['ServiceNo', 'Direction', 'BusStopCode', 'Distance', 'RoadName', 'Latitude', 'Longitude']. These features are used to generate bus routes along with route information.