Health forecasting and Medical Inventory Optimisation

Using Plaksha's Health Care Centre Data

PROBLEM STATEMENT

01 Predicting prevalent symptoms and health problems on campus

- Analyse Plaksha's Healthcare data to predict the prevalence of symptoms and health problems at various times of the year by utilising machine learning techniques

02 Developing a Medical Inventory Optimisation System

- Leverage the predictions of symptoms to optimise the management of healthcare resources such as medicines and medical supplies

Our goal is to enhance the campus healthcare system's ability to prepare for any significant health challenges, inventory adjustments, and ensure the timely availability of necessary medical resources.

APPLICATIONS AND IMPACT

INVENTORY OPTIMIZATION

Efficient procurement and management of medical supplies and resources to minimise waste and ensure on-time availability

DATA-DRIVEN CAMPUS HEALTH POLICIES

Evidence-based healthcare policies to ensure the health and safety of the campus community and be more aware and take preventative Healthcare measures based on future predictions

LITERATURE REVIEW

01 META-ANALYSIS AND SCOPING REVIEWS

Historical use of machine learning techniques being used in population healthcare forecasting

02 REPORTS AND STUDIES

Use of time series analysis as well as logistic regression; COVID-19 forecasting models

Models being used: LSTM, Regression models, ARIMA and SEIR horizons. Yet, these differences are important since forecasting future events is based on a strong assumption that the current drivers or predictors will also follow the trend over the future horizon. Hence, long-range forecasting

methods. For instance, the Box–Jenkins ARIMA model, is commonly used in fitting forecasting models when dealing with a non-stationary time series, and this model has been used extensively in health forecasting [27, 33, 52–55]. *Stationarity* is a feature of trend in a time series, and refers to the level of variation in the statistical properties (such as the mean, variance, auto-correlation, etc.) over time. Smoothing models have also been used in health

Lag refers to the lapse of time before an effect is manifested. Lags have proven useful in forecasting events globally, and are a feature of time series data that is widely exploited in many forecasting techniques, e.g. in auto regressive integrated moving averages (ARIMA) [27]. In

components in the data. For instance, in health data, an overall record of a progressively increasing incidence over a specified period would show an increasing trend, irrespective of any random or systematic fluctuations.

Soyiri, I.N., Reidpath, D.D. An overview of health forecasting. Environ Health Prev Med 18, 1–9 (2013). https://doi.org/10.1007/s12199-012-0294-6

Forecasting a health condition or situation for a population aggregate of a particular problem, or for groups of the same *family*, presents a lesser challenge than doing so for an individual case. This is because by pooling the variances of the population-related factors (which are usually broad and well known), the behavior of the aggregated data can have very stable characteristics, even when the individuals within exhibit high degrees of randomness [45]. It is therefore easier to obtain a higher degree of accuracy in forecasting specific health events when using pooled population data versus data for specific individuals.

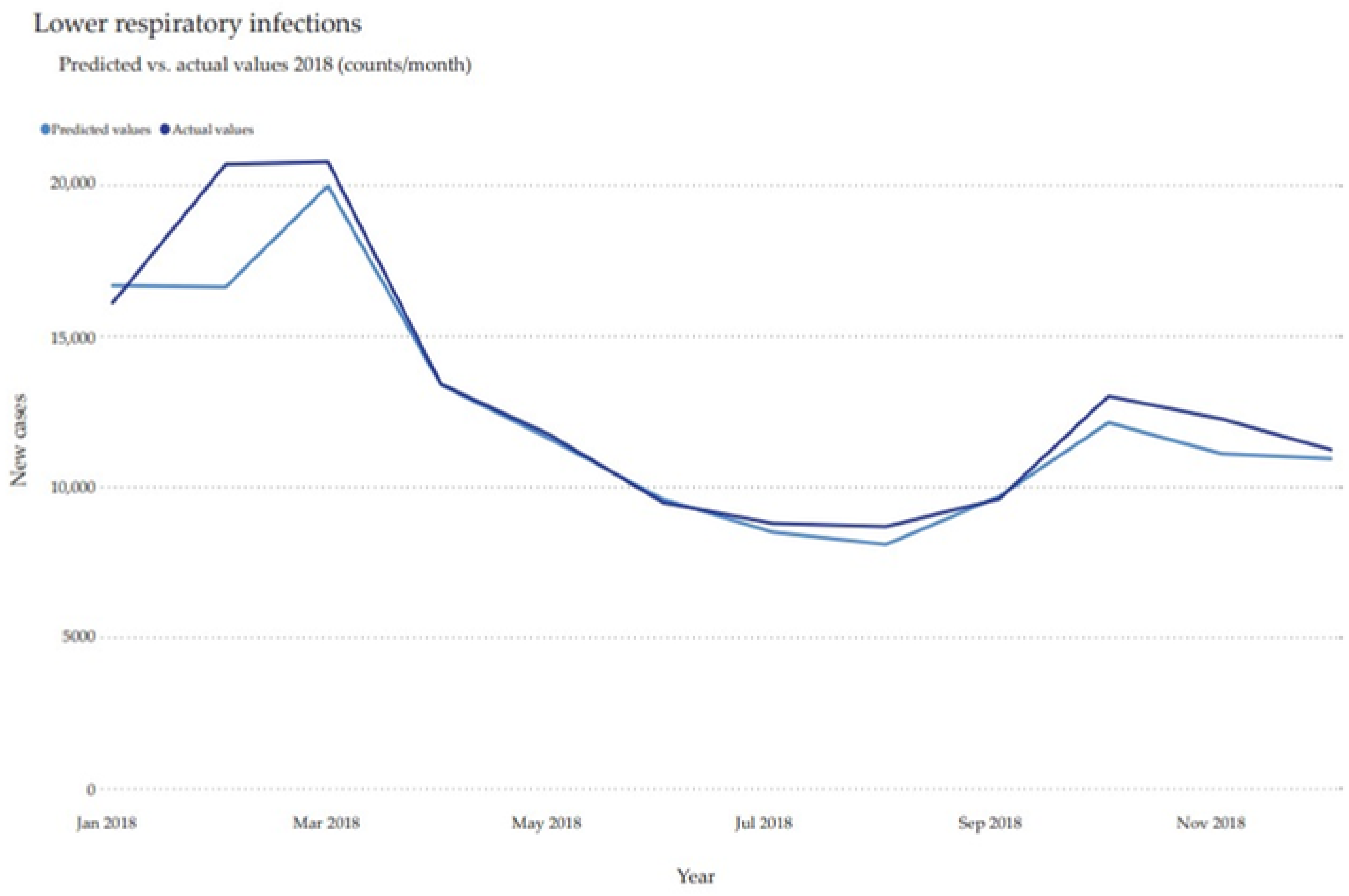
as emergency department visits [26, 27]. These individual studies adapted environmental, climatic and other factors as predictors in forecasting health. They are very specific

Morgenstern JD, Buajitti E, O'Neill M, et alPredicting population health with machine learning: a scoping reviewBMJ Open 2020;10:e037860. doi: 10.1136/bmjopen-2020-037860 model forms of regression analysis, such as Poisson regression models [34], the SARIMA (seasonal autoregressive integrated moving average) model developed from the autoregressive integrated moving average (ARIMA) model became one of the most used forecasting models [26,27]. Beyond

Data preparation of this sort allows for modern ML techniques such as XGBoost (eXtreme Gradient Boosting) and LightGBM (Light Gradient Boosting Machine), to be used, which again presents another challenge and layer of decisions to be made for researchers wishing to use these techniques, as the models' hyperparameters must be optimized, the model code implemented, and the models trained and validated appropriately on historical data. Automated time series machine learning (AutoTS) addresses

Olsavszky V, Dosius M, Vladescu C, Benecke J. Time Series Analysis and Forecasting with Automated Machine Learning on a National ICD-10 Database. International Journal of Environmental Research and Public Health. 2020; 17(14):4979. https://doi.org/10.3390/ijerph17144979

LITERATURE REVIEW



Olsavszky V, Dosius M, Vladescu C, Benecke J. Time Series Analysis and Forecasting with Automated Machine Learning on a National ICD-10 Database. International Journal of Environmental Research and Public Health. 2020; 17(14):4979. https://doi.org/10.3390/ijerph17144979

HOW IS OUR MODEL DIFFERENT?

We are employing regression on time series data to forecast the number of new cases within a specific symptom group over a three-month period (from September to November 2023). What sets our approach apart is the application of forecasting to the broader context of population health. We conducted time series forecasting for multiple symptom groups at the university level, incorporating techniques such as data aggregation and leveraging diverse ML models. Furthermore, our aim is to extend this application to optimize medicine inventory management, contributing to more efficient healthcare resource allocation.

DATASET AND FEATURES

Data collection, pre-processing

DATA COLLECTION

Plaksha's healthcare centre's data

We had to take permission from the Assistant Manager of Administration at Plaksha, Mr. Ankur Wadhwa, to access the HCC data.
We asked for anonymised data so as to maintain privacy of individuals.

NATURE OF DATA

Format	Num. of Features	Datapoints
Excel	**8**	**5431**

DATA CLEANING

Original Data

Sr.no	Date Of Visit	student/staff	department	age	gender	diagnosis	Treatment	vitals
1	01/08/22	student	TLP	25	F	Indigestion	Tab. Reciper*1	
2	01/08/22	student	TLP	25	F	Sore throat, Cold	Tab. Sinarest*4, Tab. Montair LC*3	
3	01/08/22	student	TLP	26	F	Sore throat, Cold, Cough	Tab. Sinarest*4, Tab. Montair LC*3,Rexcof DX*1	
4	02/08/22	student	TLP	35	M	Nausea,Vomiting	Tab.Emeset *1 ,ORS*1	
5	02/08/22	student	TLP	31	F	Sore Throat	Tab.Montair Lc*1	
6	02/08/22	staff	TA	21	M	Sprain on both legs and hand	Volini Spray applied, Tab. Dolomed MR*1	
7	02/08/22	staff	labour	35	M	Gastric problrm, acidity	Eno*1, inj. Pantop*1	
8	02/08/22	student	UG	26	F	Loose Motion	Tab. Sporolac DS*10, Tab. Drotin*1	
9	02/08/22	student	TLP	28	M	Sore Throat, Cough	Tab. Azee*3, Tab. Montair LC*1, Syp. Rexcof DX	
10	02/08/22	staff	TLP	31	M	Lt. foot wound	Dressing done	
11	03/08/22	student	TLP	36	F	Fever ,Spre Throat ,Mild Cuogh	Tab.Montair Lc*4,Tab.Dolo*1,Tab.Reciper*6 ,Tab.Anxit*1	
12	03/08/22	staff	TLP	25	M	Lt. foot wound	follow up done	
13	03/08/22	staff	TLP	35	M	ILI syptoms	Tab.Azee*3,Tab.Montair Lc*3,Tab.Dolo*6	
14	03/08/22	student	labour	21	F	Loose Motion	Tab. Sporolac DS*1, Tab. Reciper*1, Tab. Dolo*1	
15	03/08/22	staff	TA	22	F	bodyache	Tab.Dolo*1	
16	03/08/22	staff	HK staff	24	M	bodyache	Tab.Dolo*1	
17	03/08/22	staff	HK staff	26	M	Toothache	Tab.Brufen*1	
18	03/08/22	student	TLP	19	F	Runny Nose	Tab.Sinarest*7,Tab.Montair Lc*3,Limcee*6	
19	03/08/22	student	Maintanance	20	F	Headache, Bodyache	Tab. Dolo*1	
20	03/08/22	student	TLP	20	F	bodyache	Tab. Dolo*1	
21	04/08/22	student	TLP	24	M	Fever ,Spre Throat ,Mild Cuogh	Tab.Combiflam*2,Tab.Azee*3,Tab.Montair LC*2,Tab.Limcee*6,	Temp-100F
22	04/08/22	student	TLP	26	M	Sore Throat, Cough	Betadine Gangle*1	
23	04/08/22	sdudend	TLP	28	F	Cut Wound Lt.Foot	Dressing done,soft Swab*1,Bandage4inch*1	
24	04/08/22	staff	Security Staff	31	M	Bodyache	Tab. Dolo*2,	
25	04/08/22	staff	Maintanance	22	M	Headache	Tab. Crocin*1	
26	04/08/22	staff	HK staff	32	F	Anxiety	Tab. Dolo*1, Tab . Reciper*1	
27	05/08/22	staff	BCMS	23	F	Cold	Tab. Sinarest*1	
28	05/08/22	staff	Faculty staff	23	M	pain in abdomen	Tab. Metrogyl*9, Tab. Dolo*9, Tab. Panticid*9	
29	05/08/22	staff	HK staff	26	M	Headache	Tab.Dolo*1, Tab. Emset*1	
30	05/08/22	staff	BCMS	24	F	Sore throat,Fever	Tab.Azee*3,Tab.Montair Lc*3,Tab.combiflam*6,tab.pan DSR*3	
31	05/08/22	staff	BCMS	28	F	COVID	tab.pan DSR*1,tab.Montair lc*3	
32	05/08/22	student	TLP	29	M	Acidity	Tab.Pan DSR*2	
33	06/08/22	staff	BCMS	21	M	vomiting	Tab.Emset*3	
34	06/08/22	staff	TLP	19	F	ILI syptoms	TAB Azee*2,tab.montair*4,tab.Pan DSR*2	
35	06/08/22	staff	HK staff	19	F	bodyache, acidity	Tab.Dolo*1,tab.Pan DSR*1	
36	06/08/22	staff	HK staff	20	M	Dysmenorhea	tab.maftal Spas*1	
37	06/08/22	student	Tlp	20	M	Cut Wound Lt.Foot	Dressing Done Soft Swab *1	
38	06/08/22	student	UG	18	F	Dysmenorhea	Tab.Meftal spas*1	
39	06/08/22	student	TLP	35	M	Acidity	tab.pan DSR*1	
40	06/08/22	staff	PRF	21	M	ILI syptoms	Tab.combiflam1,Montair Lc*1	
41	06/08/22	student	TLP	22	M	Cut Wound Lt.Foot	soft Swab*1,Bandage4inch*1,Dressing done	
42	07/08/22	staff	PRF	20	M	ILI syptoms	tab.Azee*3,tab.Dolo*7,tab.pan dsr*3	

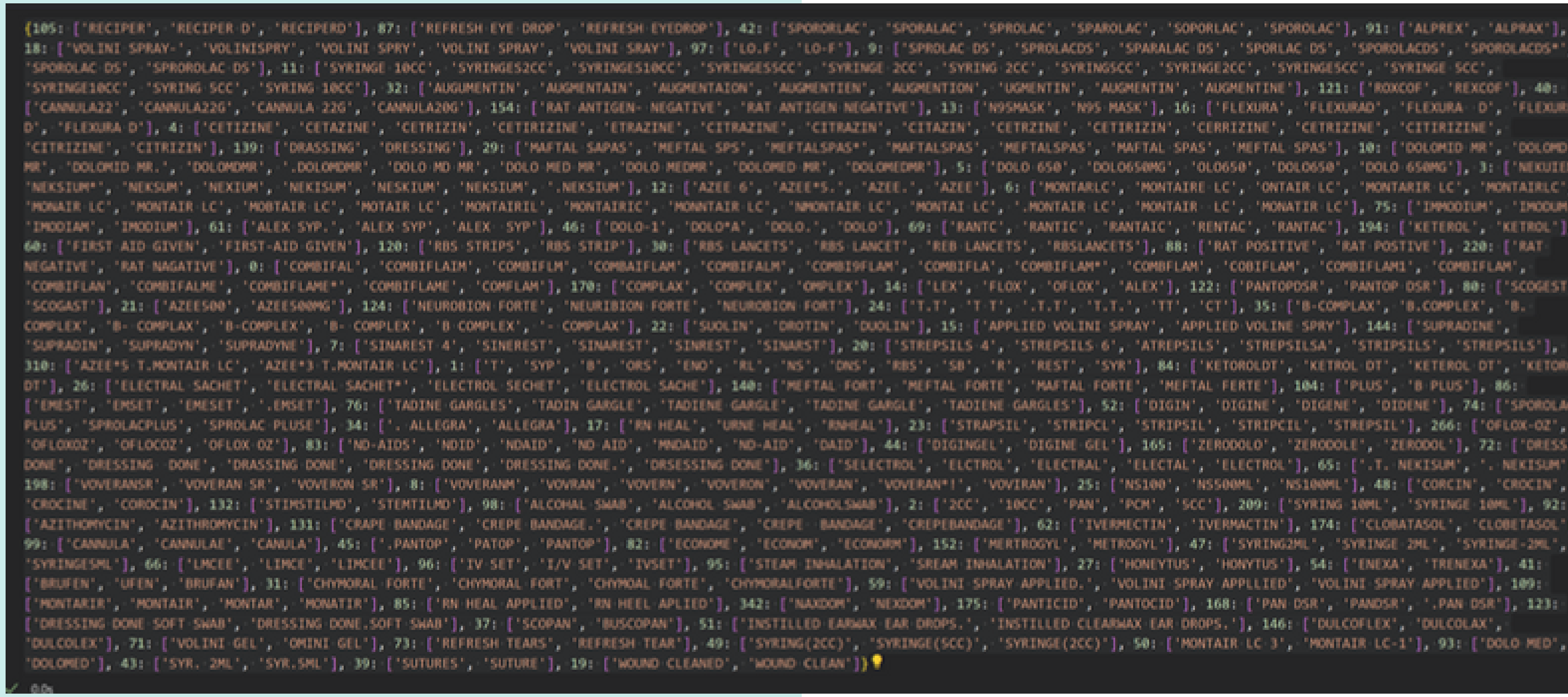
FEATURE - SYMPTOM

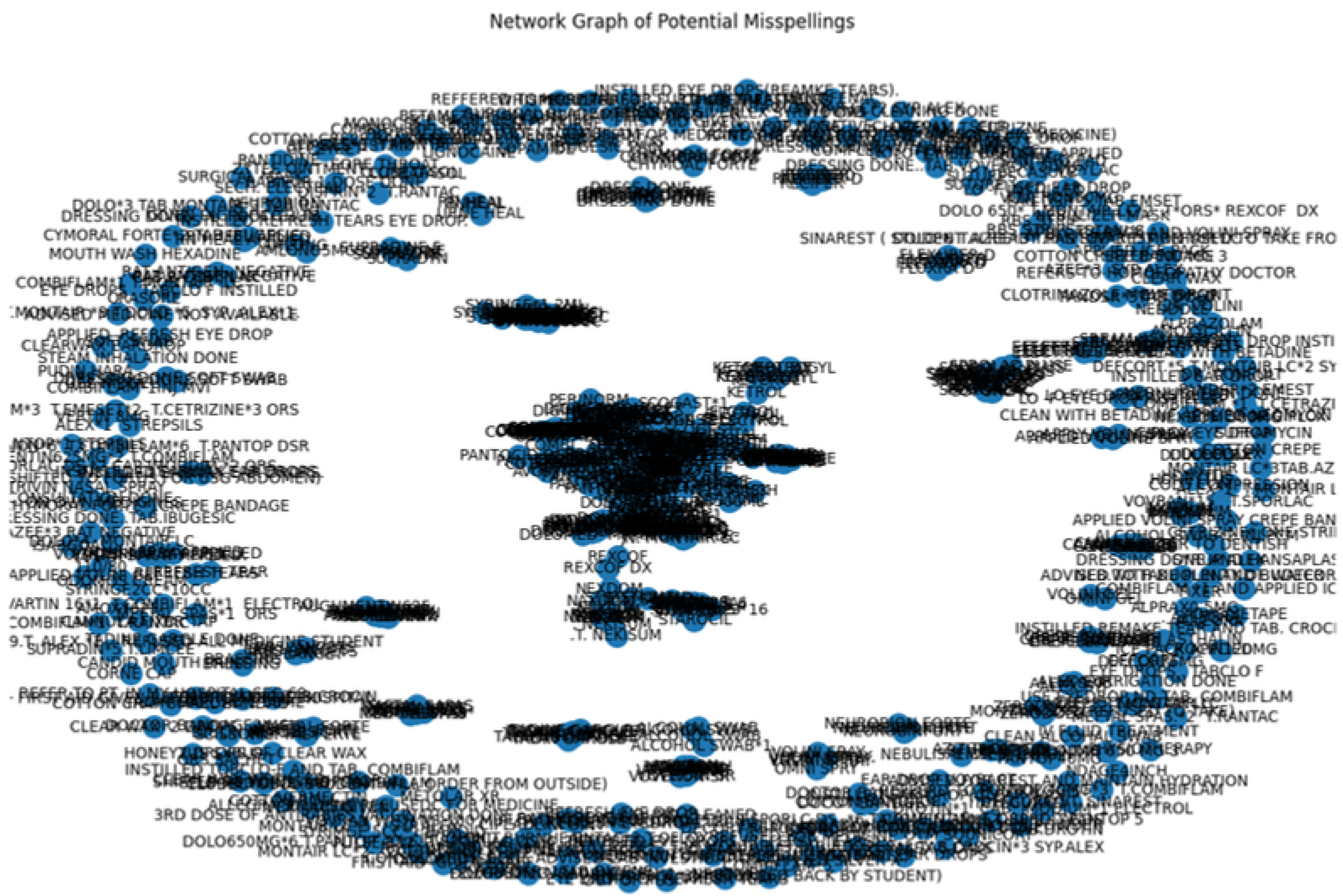
- Separated symptoms into six aggregate symptom groups based on literature survey - simplifying assumptions made due to limited data availability.
- For example, URI (Upper respiratory infection) includes cold, cough, etc.
- Performed feature encoding
- For each date, calculated frequency of occurence of each aggregate group

index	URI	HEADACHE	GASTROINTESTINAL DISORDER	FEVER	AQI	Max Temp	Min Temp	FATIGUE_C	OTHERS
02-11-2021	0	1	0	0	100	30	17	0	0
03-11-2021	0	0	0	0	96	31	16	0	0
04-11-2021	0	0	0	0	84	31	16	0	0
05-11-2021	0	0	0	0	118	30	17	0	0
06-11-2021	0	0	0	0	79	30	15	0	0
07-11-2021	0	1	0	0	97	30	16	0	0
08-11-2021	0	0	0	0	93	28	15	1	0
09-11-2021	0	0	0	0	87	27	16	0	0
10-11-2021	1	0	0	1	91	25	16	1	0
11-11-2021	1	0	0	1	93	29	13	0	0
12-11-2021	0	0	0	0	94	28	13	0	0
13-11-2021	2	2	2	0	88	29	11	2	0

FEATURE - TREATMENT

- Split treatment into different columns of medicines
- Create a list of medicine names that appear only once - potential misspellings
- Calculate Levenshtein distance between all pairs of medicine names
- Organise the potential misspellings into clusters based on the clustering labels obtained
- Map the labelled clusters manually to correct spelling names
- Update the data frame with corrected medicine names




Network Graph of Potential Misspellings
INSTILLED EYE DROP(REAMKE TEARS).
MOUTH WASH HEXADINE
APPLIED REFRESH EYE DROP
STEAM INHALATION DONE
EYE DROPS CARCLO F INSTILLED
SINAREST (
DOLO 650*
CLOTRIMAZOLE
REFERS TO ALLOPATHY DOCTOR
DEFCORT *5 T.MONTAIR LC*2 SY
CLEAN WITH BETADINE
APPLIED VOLINI SPRAY CREPE BAN
DRESSING DONE
REXCOF
REXCOF DX
NEXDOM
.T. NEXISUM
KETROL
PERINORM
CYMORAL FORTE
CORNE CAP
HONEY
CLEAR WAX
INSTILLED
METOLAR XR
3RD DOSE OF ANTI
DOLO650MG*6 T.
MONTAIR LC
FIRST AID
NEUROBION FORTE
ALCOHOL SWAB
ALCOHOL SWAB*1
VOLINI SPRAY
NEBULISE
IV FLUID TREATMENT
CLEAN ALCOHOL SWAB
COMBIFLAM
ALPRAZOLAM
NEBULISATION
COTTON CREPE
IRRIGATION DONE
CETRIZINE
ALEX
T.RANTAC

FEATURE ENGINEERING

Air Quality Index

Data Source:
Central Pollution Control Board, Ministry of Environment, Forests and Climate Change

Major Pollutant

Data Source:
Central Pollution Control Board, Ministry of Environment, Forests and Climate Change

Min. Temperature

Data Source:
AccuWeather - 3rd party weather forecast provider. Gets data from Google Cloud and other sources

Max. Temperature

Data Source:
AccuWeather - 3rd party weather forecast provider. Gets data from Google Cloud and other sources

MISSING VALUES

Air Quality Index

Applied regression to fill in missing values

Min/Max Temperature

Interpolation: averaging previous and next value to get an estimate of the value

Symptoms

Mapped treatments to most probable symptom values

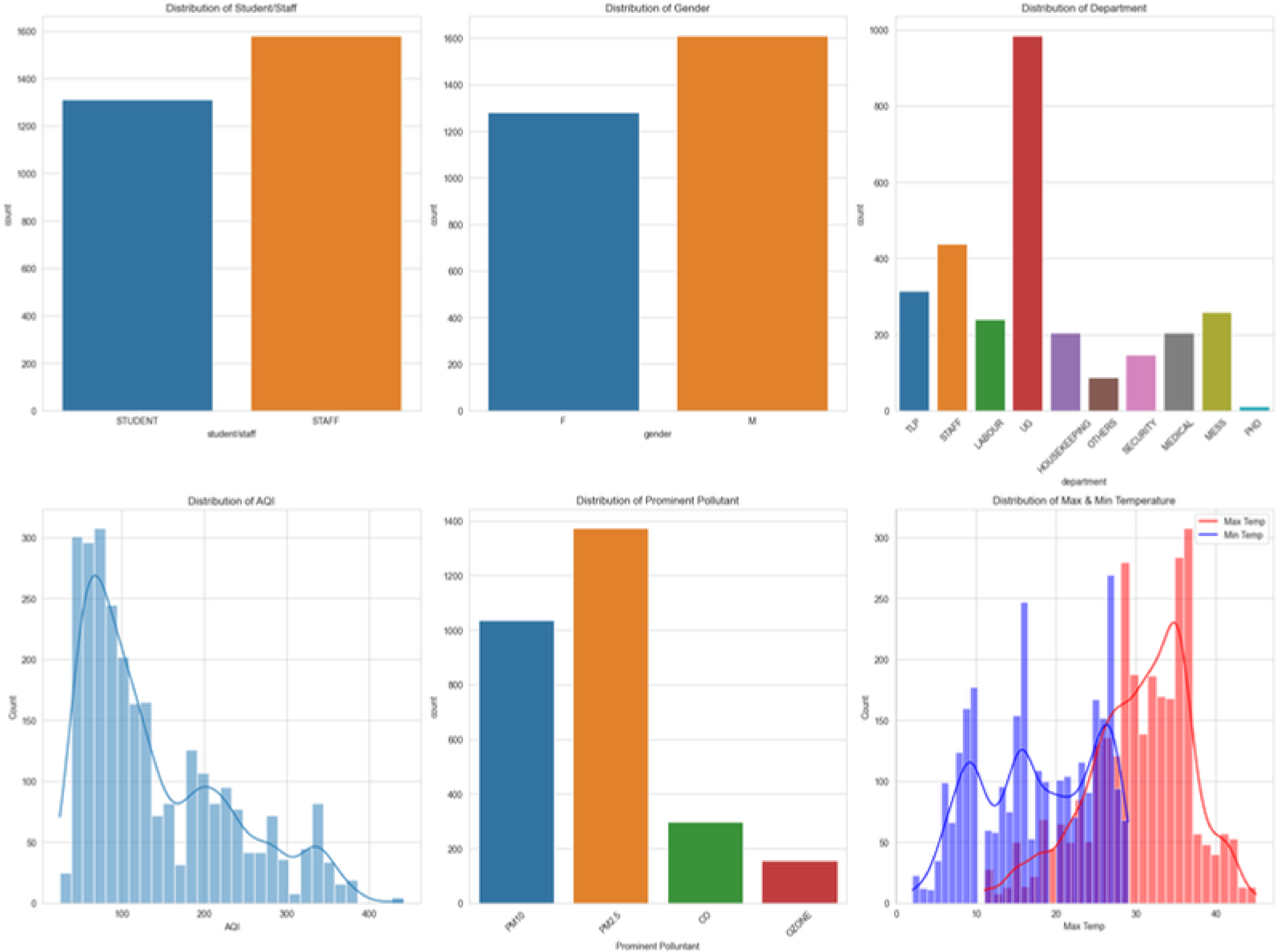
CLEANED DATA

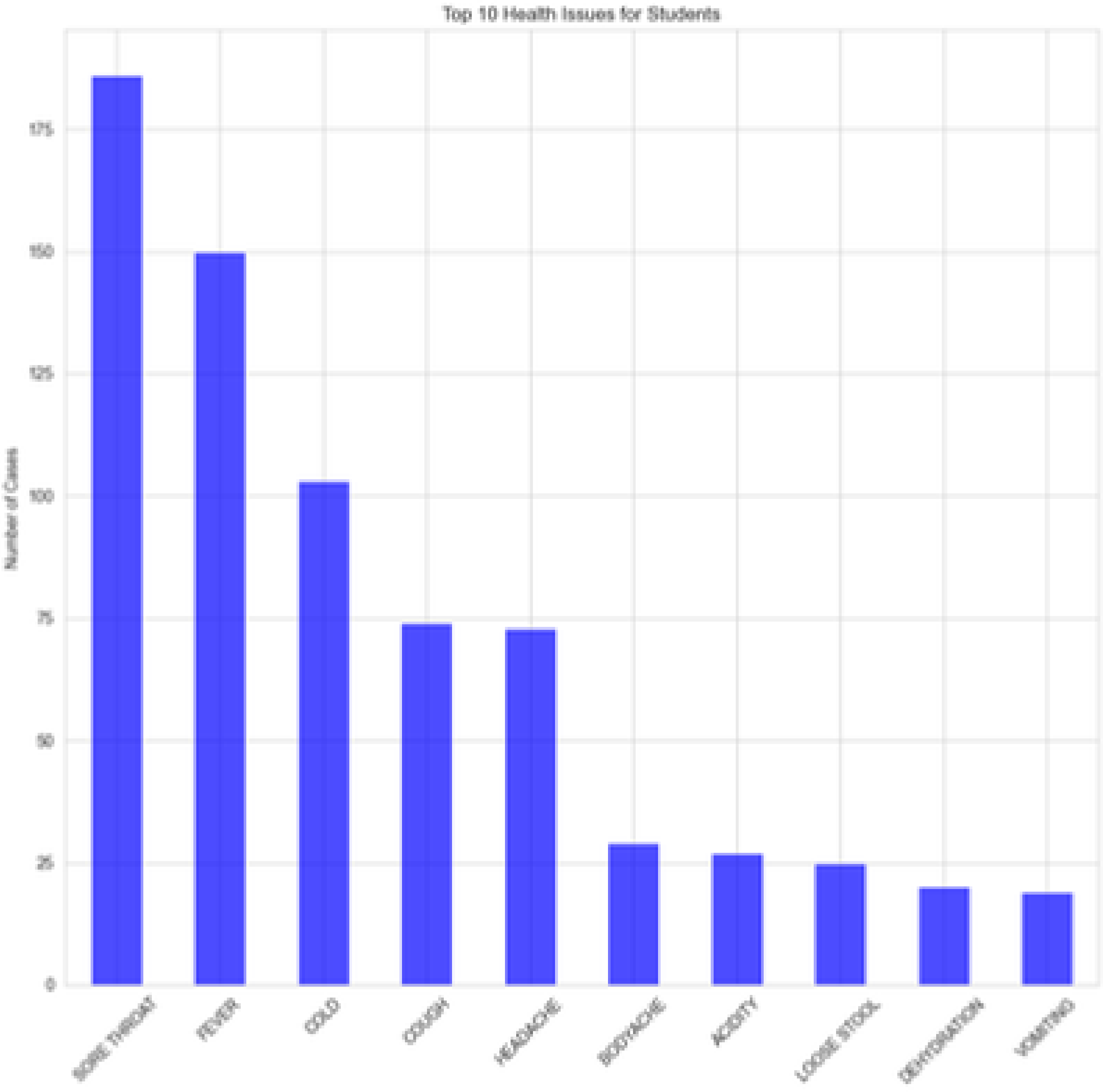
Date Of Visit	student/staff	department	gender	AQI	Prominent Pollutant	Max Temp	Min Temp	Diagnosis_1	iagnosis	iagnosis	iagnosis	Treatment_1
2021-11-02 00:00:00	8.00AM	TLP	F	100	PM10	30	17	HEADACHE				DOLO
2021-11-07 00:00:00	7.00AM	TLP	M	97	PM10	30	16	HEADACHE				SARIDON
2021-11-08 00:00:00	2PM	TA	M	93	PM10	28	15		NAUSEA			CONSULTATION
2021-11-10 00:00:00	7.00AM	MAINTANANCE	M	91	PM10	25	16	FEVER				DOLO650
2021-11-10 00:00:00	12.00PM	TLP	F	91	PM10	25	16	MYALGIA				COMBIFLAM
2021-11-10 00:00:00	9.30AM	HK STAFF	M	91	PM10	25	16	URI				
2021-11-11 00:00:00	10PM	TLP	M	93	PM10	29	13	FEVER				OVSERVATION DONE FOR 1/2 HRS
2021-11-11 00:00:00	10PM	SECURITY STAFF	M	93	PM10	29	13	URI				MONTAIRLC
2021-11-13 00:00:00	3:52PM	HK STAFF	F	88	PM10	29	11	GASTROINTESTINAL DISORDER				DOLO /650MG
2021-11-13 00:00:00	5,30PM	BCMS	F	88	PM10	29	11	GASTROINTESTINAL DISORDER				
2021-11-13 00:00:00	2:55PM	BCMS	M	88	PM10	29	11	HEADACHE				COMBIFLAM
2021-11-13 00:00:00	4:26PM	PRF	M	88	PM10	29	11	HEADACHE				COMBIFLAM
2021-11-13 00:00:00	8PM	UG	M	88	PM10	29	11	MYALGIA				COMBIFLAM
2021-11-13 00:00:00	2PM	BCMS	F	88	PM10	29	11	URI	MYALGIA			STEAM INHALATION
2021-11-13 00:00:00	2,50PM	TLP	M	88	PM10	29	11	URI				MONTAIRLC
2021-11-14 00:00:00	11:55AM	ACADEMICS	M	90	PM10	26	13	ULCER				CAP
2021-11-14 00:00:00	2:50PM	BCMS	M	90	PM10	26	13	URI				MONTAIRLC
2021-11-15 00:00:00	3:20PM	VC SON	M	89	PM10	29	13	FATIGUE				ELECTRAL GIVEN
2021-11-15 00:00:00	4:12PM	UG	F	89	PM10	29	13	HEADACHE				COMBIFLAM
2021-11-15 00:00:00	8:55PM	DINNING	M	89	PM10	29	13	URI				MONTAIRLC
2021-11-15 00:00:00	9:30PM	BCMS	M	89	PM10	29	13	URI				MONTAIRLC
2021-11-16 00:00:00	1:16PM	HK STAFF	F	109	PM10	28	13	URI				MONTAIRLC
2021-11-16 00:00:00	7:15PM	TLP	M	109	PM10	28	13	URI				SYRUP HONEYTUS
2021-11-17 00:00:00	3:15PM	TLP	M	108	PM10	26	13	GASTROINTESTINAL DISORDER				DOLO / 650MG
2021-11-17 00:00:00	10AM	TLP	M	108	PM10	26	13	URI				MONTAIRLC
2021-11-18 00:00:00	5:30PM	UG	F	114	PM10	29	11	MYALGIA				COMBIFLAM
2021-11-18 00:00:00	8.00AM	BCMS	M	114	PM10	29	11	URI				STEAM
2021-11-18 00:00:00	10:10AM	BCMS	M	114	PM10	29	11	URI				SVICKS
2021-11-19 00:00:00	9:37PM	SECURITY STAFF	F	112	PM10	28	11	MYALGIA	URI			COMBIFLAM
2021-11-19 00:00:00	11AM	TLP	F	112	PM10	28	11	URI				CITRAZIN
2021-11-21 00:00:00	15:40:00	TA	F	103	PM10	28	11	NAUSEA	GASTRI	HEADACHE		COMBIFLAM
2021-11-21 00:00:00	9.30AM	HORTICULTURE	M	103	PM10	28	11	URI				MONTAIRLC
2021-11-21 00:00:00	11.00AM	HK STAFF	F	103	PM10	28	11	URI				MONTAIRLC
2021-11-21 00:00:00	11:30AM	LABOUR	M	103	PM10	28	11	URI				MONTAIRLC
2021-11-21 00:00:00	3:30PM	LABOUR	F	103	PM10	28	11	URI				MONTAIRLC
2021-11-22 00:00:00	10:30PM	TLP	M	94	PM10	28	11	GASTROINTESTINAL DISORDER				COMBIFLAM
2021-11-22 00:00:00	17:54:00	AUDIT VISITOR	F	94	PM10	28	11	URI				CITRAZIN
2021-11-22 00:00:00	18:30:00	TA	M	94	PM10	28	11	URI				SYRUP HONEYTUS
2021-11-23 00:00:00	1:21PM	PHD	M	102	PM10	27	8	FEVER	HEADA	NAUSEA		COMBIFLAM
2021-11-23 00:00:00	17:00:00	TUCK SHOP	M	102	PM10	27	8	NAUSEA	FEVER			DOLO650
2021-11-23 00:00:00	11AM	TLP	M	102	PM10	27	8	URI				MONTAIRLC
2021-11-24 00:00:00	8PM	TLP	M	111	PM10	28	9	NAUSEA	NAUSEA			EMEST
2021-11-24 00:00:00	5:15PM	TLP	M	111	PM10	28	9	URI				MONTAIRLC
2021-11-24 00:00:00	6PM	TLP	M	111	PM10	28	9	URI				CITRAZIN
2021-11-25 00:00:00	11:05:00	TLP	M	92	PM10	27	9	FEVER	URI			COMBIFLA
2021-11-25 00:00:00	3:50PM	TLP	M	92	PM10	27	9	GASTROINTESTINAL DISORDER				DIGIN
2021-11-25 00:00:00	7PM	MAINTANANCE	M	92	PM10	27	9	GASTROINTESTINAL DISORDER				SPROLACDS
2021-11-25 00:00:00	12:30:00	TLP	M	92	PM10	27	9	HEADACHE				COMBIFLAM
2021-11-26 00:00:00	1:10PM	TLP	F	92	PM10	26	10	GASTROINTESTINAL DISORDER				SPROLACDS
2021-11-26 00:00:00	10:30:00	TLP	F	92	PM10	26	10	URI				AZEE
2021-11-26 00:00:00	4:55PM	TLP	F	92	PM10	26	10	URI				MONTAIR LC
2021-11-27 00:00:00	5:55PM	LABOUR	M	112	PM10	23	13	ALLERGY				ALLEGRA
2021-11-29 00:00:00	2AM	TA	F	107	PM10	27	12	NAUSEA	URI			CITRAZIN
2021-11-30 00:00:00	2:30PM	UG	F	103	PM10	25	11	GASTROINTESTINAL DISORDER				DIGIN

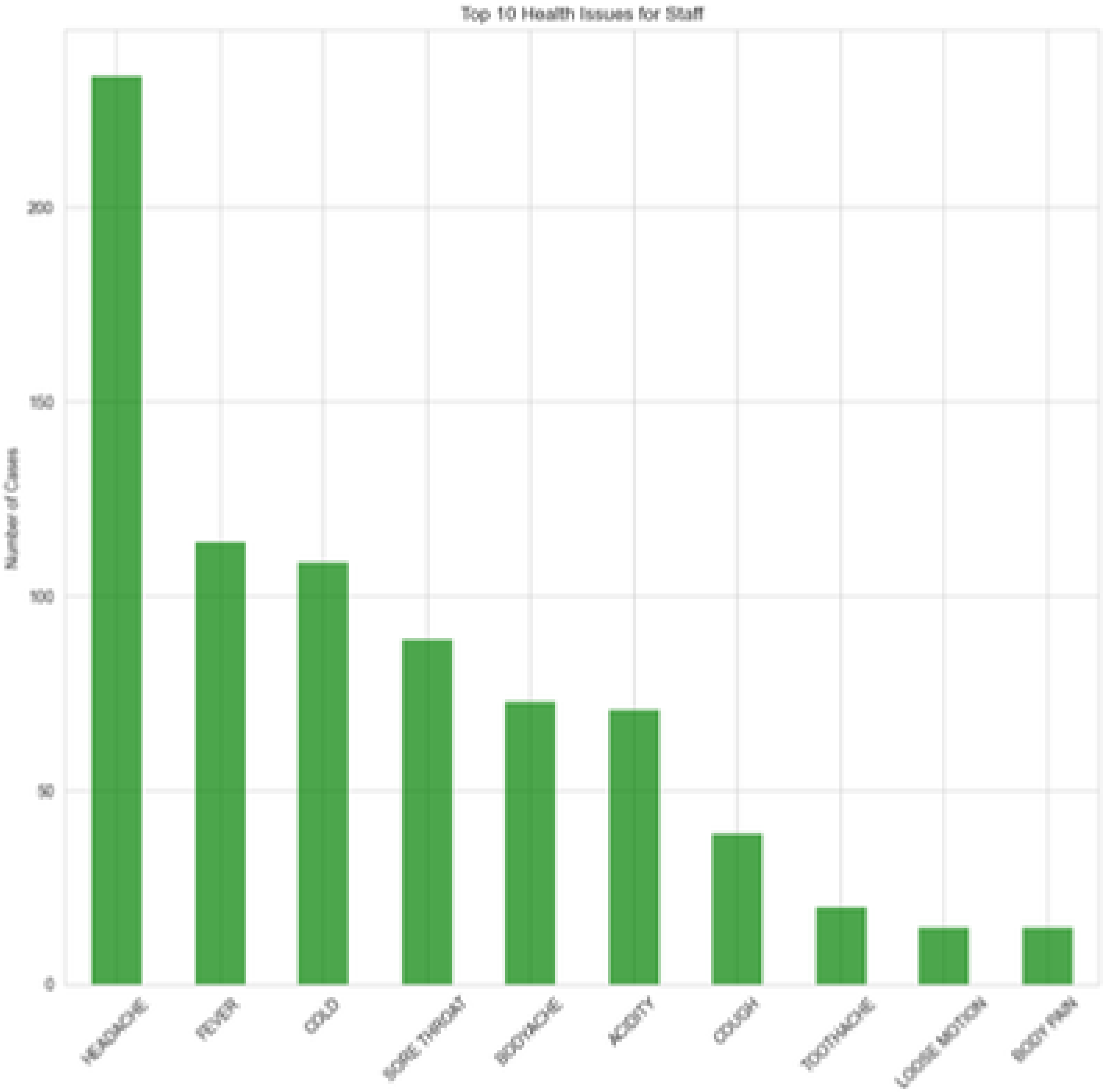
TRANSFORMED DATA

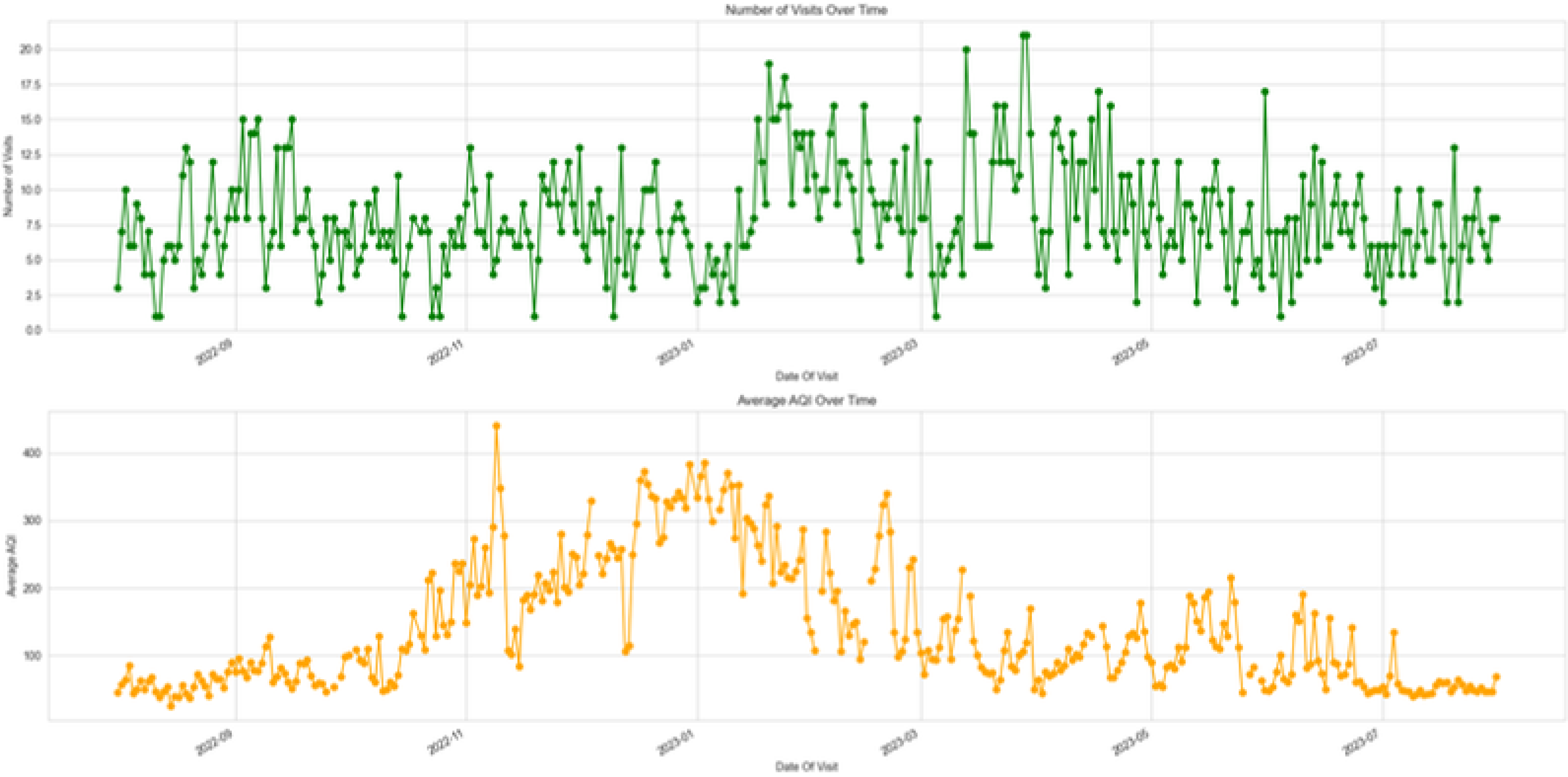
index	URI	HEADACHI	GASTROIN	FEVER	AQI	Max Temp	Min Temp	FATIGUE_(	OTHERS
02-11-2021	0	1	0	0	100	30	17	0	0
03-11-2021	0	0	0	0	96	31	16	0	0
04-11-2021	0	0	0	0	84	31	16	0	0
05-11-2021	0	0	0	0	118	30	17	0	0
06-11-2021	0	0	0	0	79	30	15	0	0
07-11-2021	0	1	0	0	97	30	16	0	0
08-11-2021	0	0	0	0	93	28	15	1	0
09-11-2021	0	0	0	0	87	27	16	0	0
10-11-2021	1	0	0	1	91	25	16	1	0
11-11-2021	1	0	0	1	93	29	13	0	0
12-11-2021	0	0	0	0	94	28	13	0	0
13-11-2021	2	2	2	0	88	29	11	2	0
14-11-2021	1	0	0	0	90	26	13	0	1
15-11-2021	2	1	0	0	89	29	13	1	0
16-11-2021	2	0	0	0	109	28	13	0	0
17-11-2021	1	0	1	0	108	26	13	0	0
18-11-2021	2	0	0	0	114	29	11	1	0
19-11-2021	2	0	0	0	112	28	11	1	0
20-11-2021	0	0	0	0	111	28	13	0	0
21-11-2021	4	1	1	0	103	28	11	1	0
22-11-2021	2	0	1	0	94	28	11	0	0
23-11-2021	1	1	0	2	102	27	8	2	0
24-11-2021	2	0	0	0	111	28	9	2	0
25-11-2021	1	1	2	1	92	27	9	0	0
26-11-2021	2	0	1	0	92	26	10	0	0

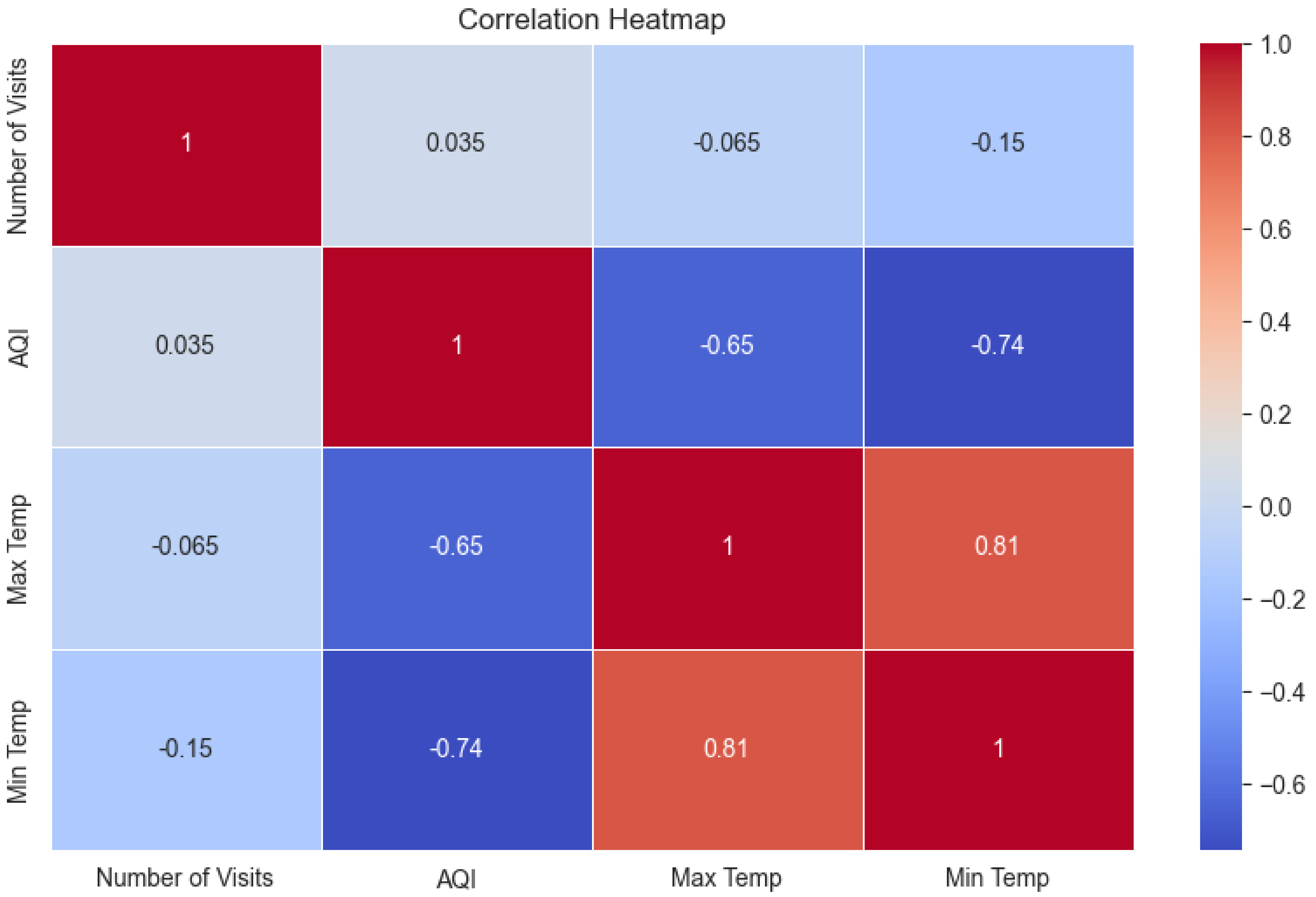
Exploratory Data Analysis


Distribution of Student/Staff
count
STUDENT
STAFF
student/staff
Distribution of Gender
count
F
M
gender
Distribution of Department
count
TLP
STAFF
LABOUR
UG
HOUSEKEEPING
OTHERS
SECURITY
MEDICAL
MESS
PHD
department
Distribution of AQI
Count
AQI
Distribution of Prominent Pollutant
count
PM10
PM2.5
CO
OZONE
Prominent Pollutant
Distribution of Max & Min Temperature
Max Temp
Min Temp
Count
Max Temp


Top 10 Health Issues for Students
Number of Cases
0
25
50
75
100
125
150
175
SORE THROAT
FEVER
COLD
COUGH
HEADACHE
BODYACHE
ACIDITY
LOOSE STOOL
DEHYDRATION
VOMITING


Top 10 Health Issues for Staff
Number of Cases
0
50
100
150
200
HEADACHE
FEVER
COLD
SORE THROAT
BODYACHE
ACIDITY
COUGH
TOOTHACHE
LOOSE MOTION
BODY PAIN


Number of Visits Over Time
Number of Visits
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
2022-09
2022-11
2023-01
2023-03
2023-05
2023-07
Date Of Visit
Average AQI Over Time
Average AQI
400
300
200
100
2022-09
2022-11
2023-01
2023-03
2023-05
2023-07
Date Of Visit



Correlation Heatmap
Number of Visits
AQI
Max Temp
Min Temp
1
0.035
-0.065
-0.15
0.035
1
-0.65
-0.74
-0.065
-0.65
1
0.81
-0.15
-0.74
0.81
1
Number of Visits
AQI
Max Temp
Min Temp
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

ML Methodology

model forms of regression analysis, such as Poisson regression models [34], the SARIMA (seasonal autoregressive integrated moving average) model developed from the autoregressive integrated moving average (ARIMA) model became one of the most used forecasting models [26,27]. Beyond

Data preparation of this sort allows for modern ML techniques such as XGBoost (eXtreme Gradient Boosting) and LightGBM (Light Gradient Boosting Machine), to be used, which again presents another challenge and layer of decisions to be made for researchers wishing to use these techniques, as the models' hyperparameters must be optimized, the model code implemented, and the models trained and validated appropriately on historical data. Automated time series machine learning (AutoTS) addresses

Olsavszky V, Dosius M, Vladescu C, Benecke J. Time Series Analysis and Forecasting with Automated Machine Learning on a National ICD-10 Database. International Journal of Environmental Research and Public Health. 2020; 17(14):4979. https://doi.org/10.3390/ijerph17144979

WHY ML AND NOT TIME SERIES?

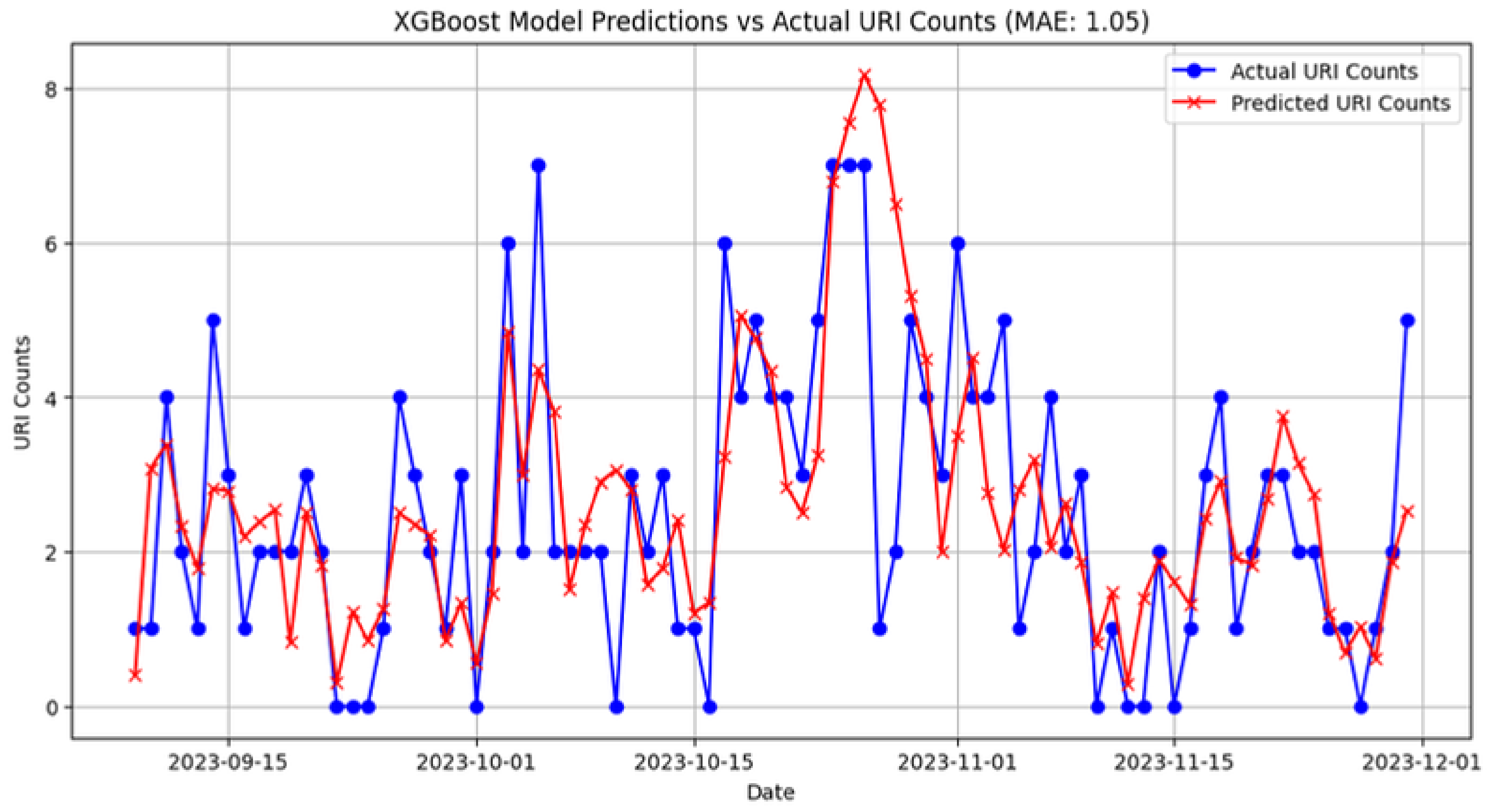
Zero-Inflation and Sparsity: The dataset has many zeros and low-frequency counts. Machine learning models can be equipped with techniques to handle zero-inflation and sparse data more effectively than SARIMA, which assumes a certain level of continuity in the data.

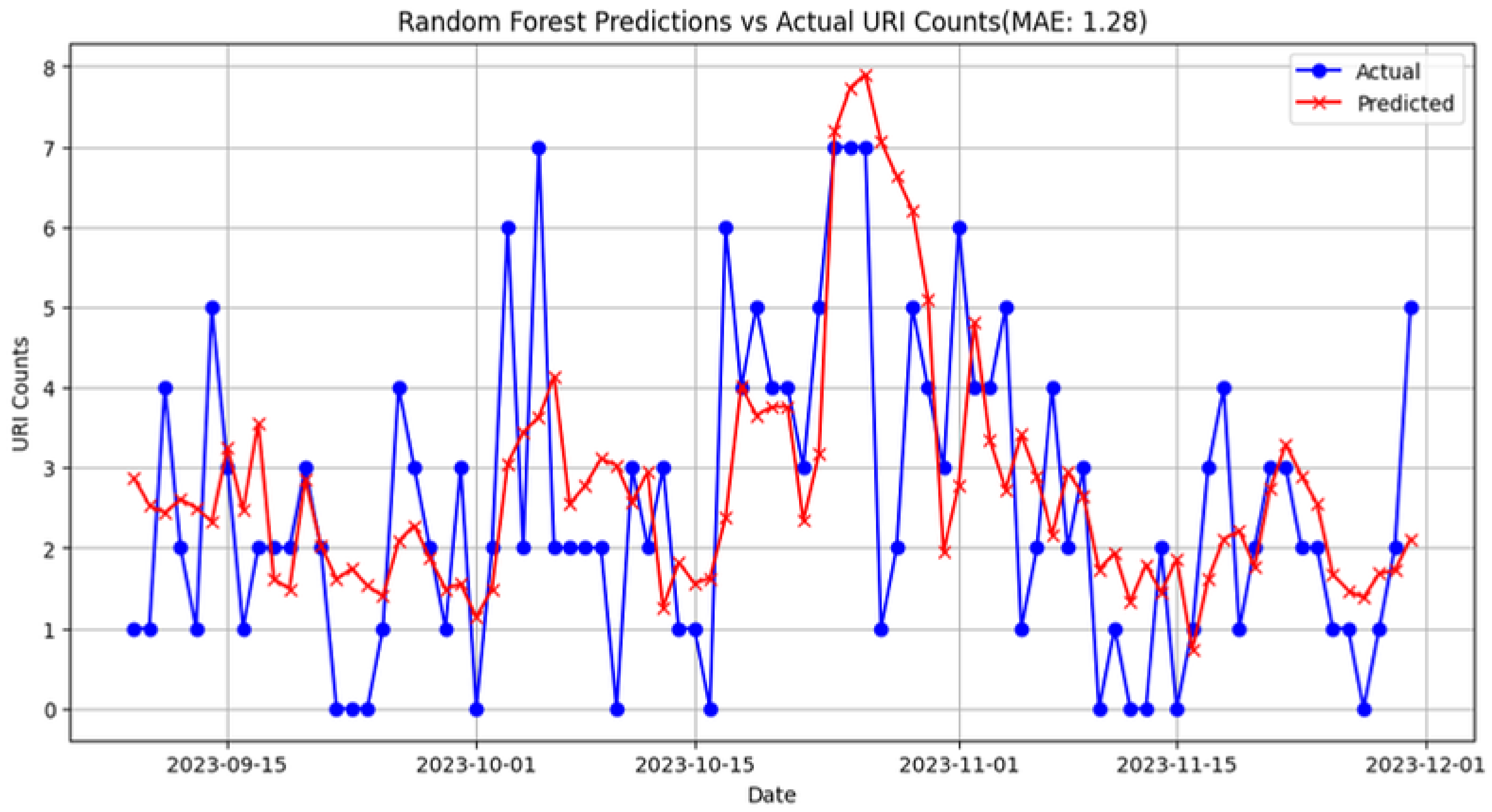
Multivariate Forecasting: Unlike SARIMA, which is univariate, machine learning models can naturally handle multiple input features simultaneously, allowing for multivariate forecasting. This means they can use all available data to predict the number of visits, rather than relying on the historical counts alone.

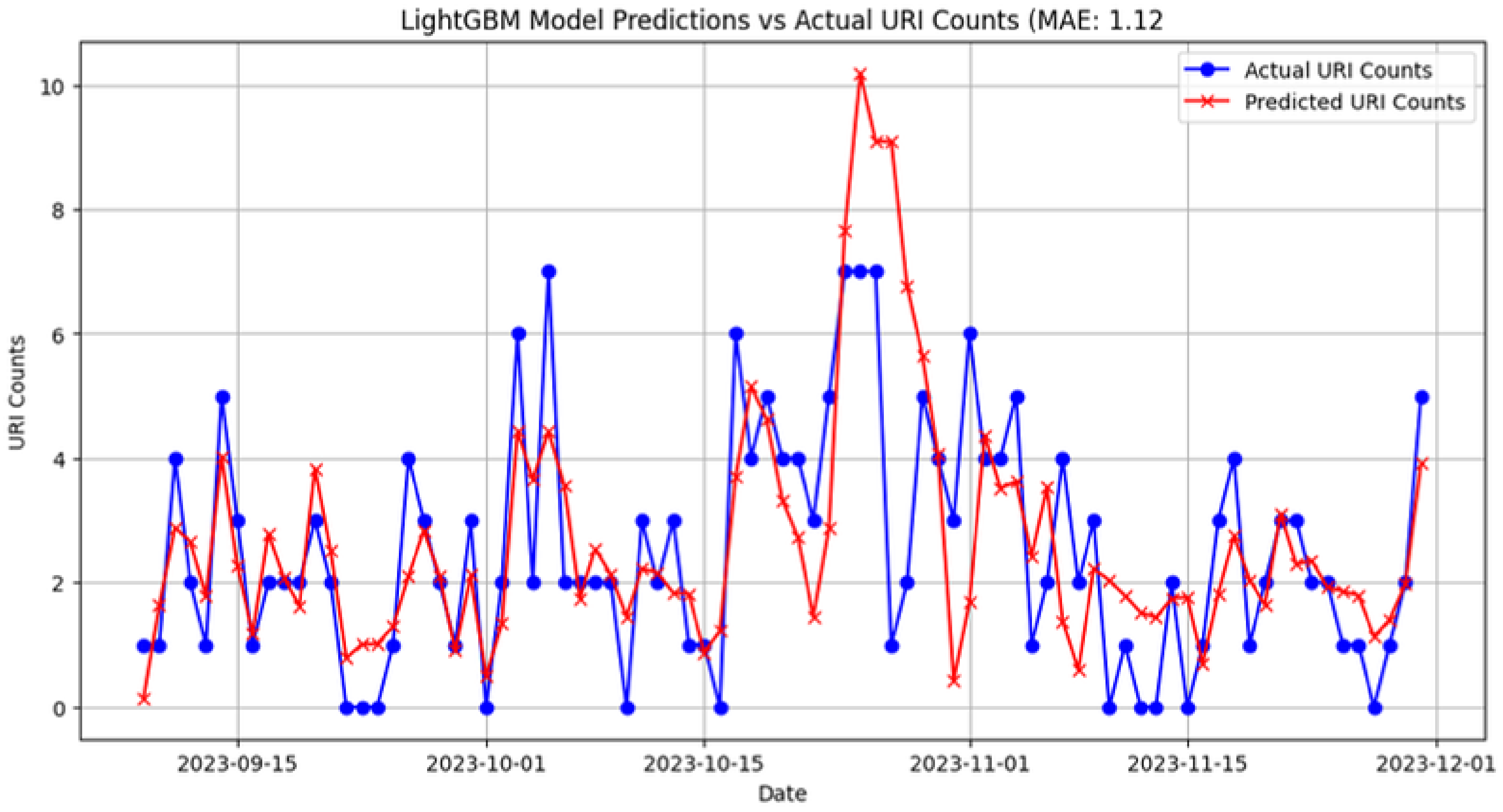
Scalability: These machine learning models are highly scalable and can handle larger datasets efficiently, which is beneficial if the amount of data grows over time.

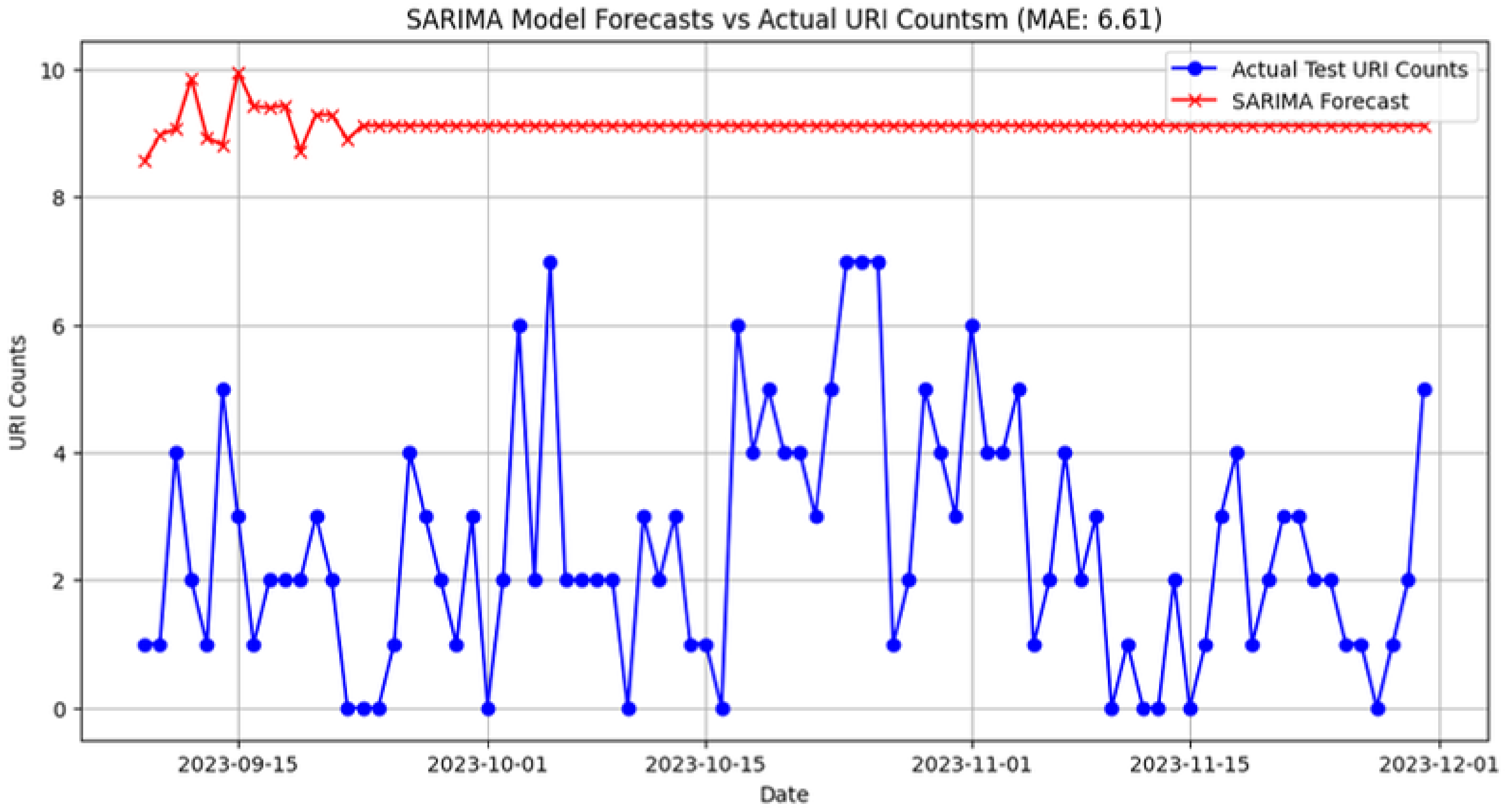
Feature Interactions and non-linear relationships: Factors such as AQI and temperature can interact in non-linear ways to affect health outcomes like 'FEVER' visits. ML models can automatically learn these, whereas SARIMA would require manual intervention to model such relationships, if it can model them at all.

Model performance for URI

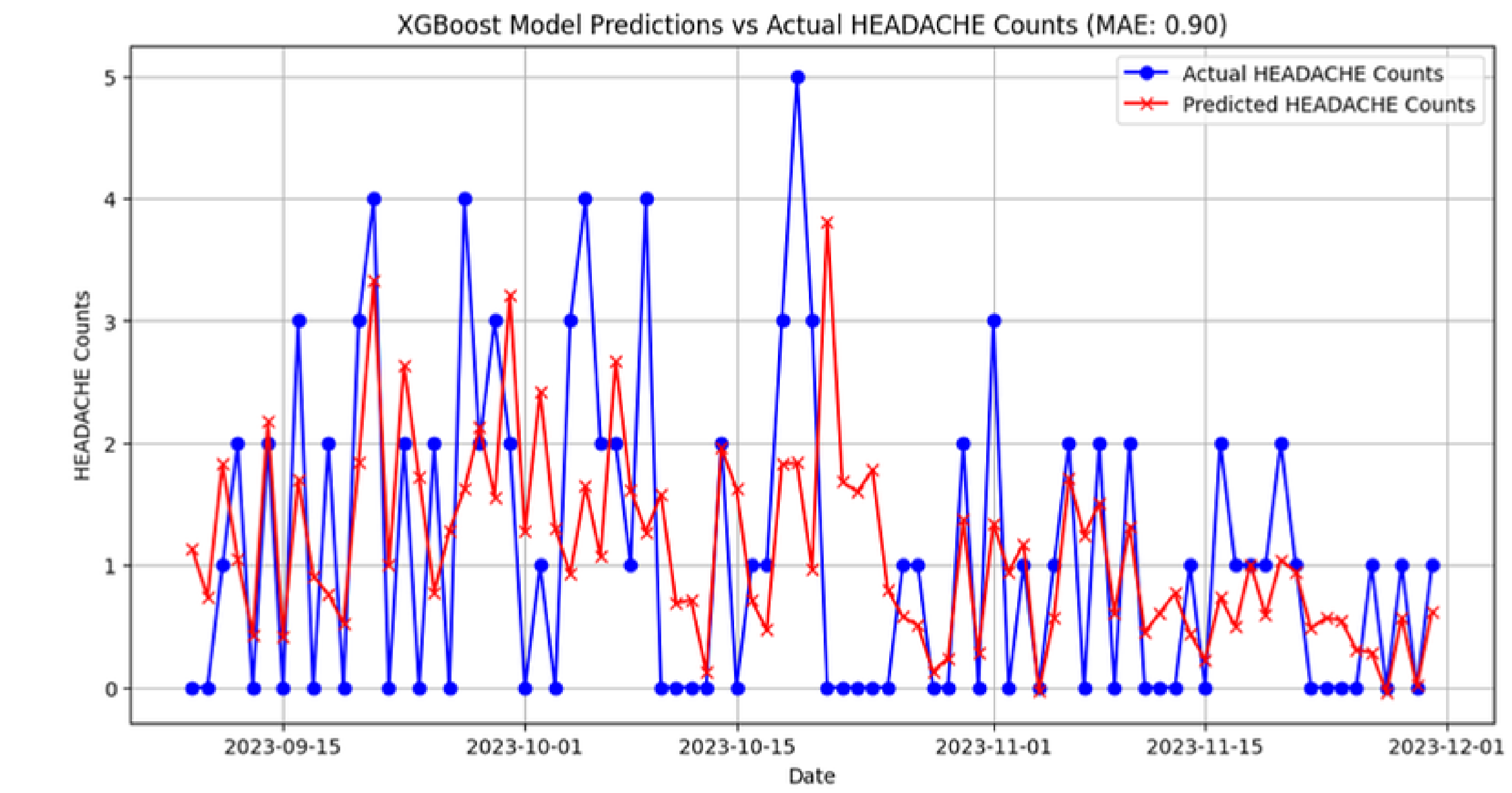


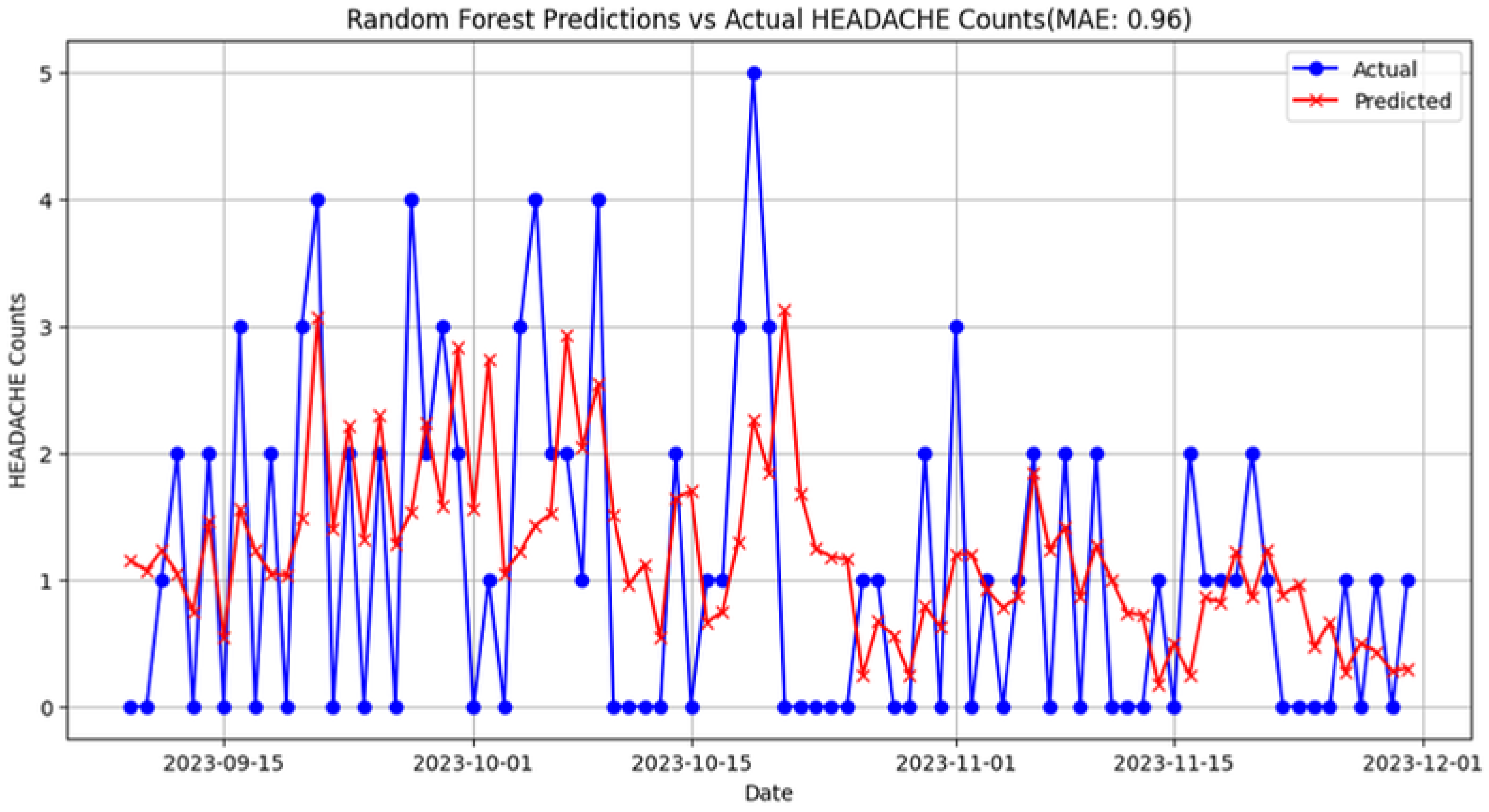


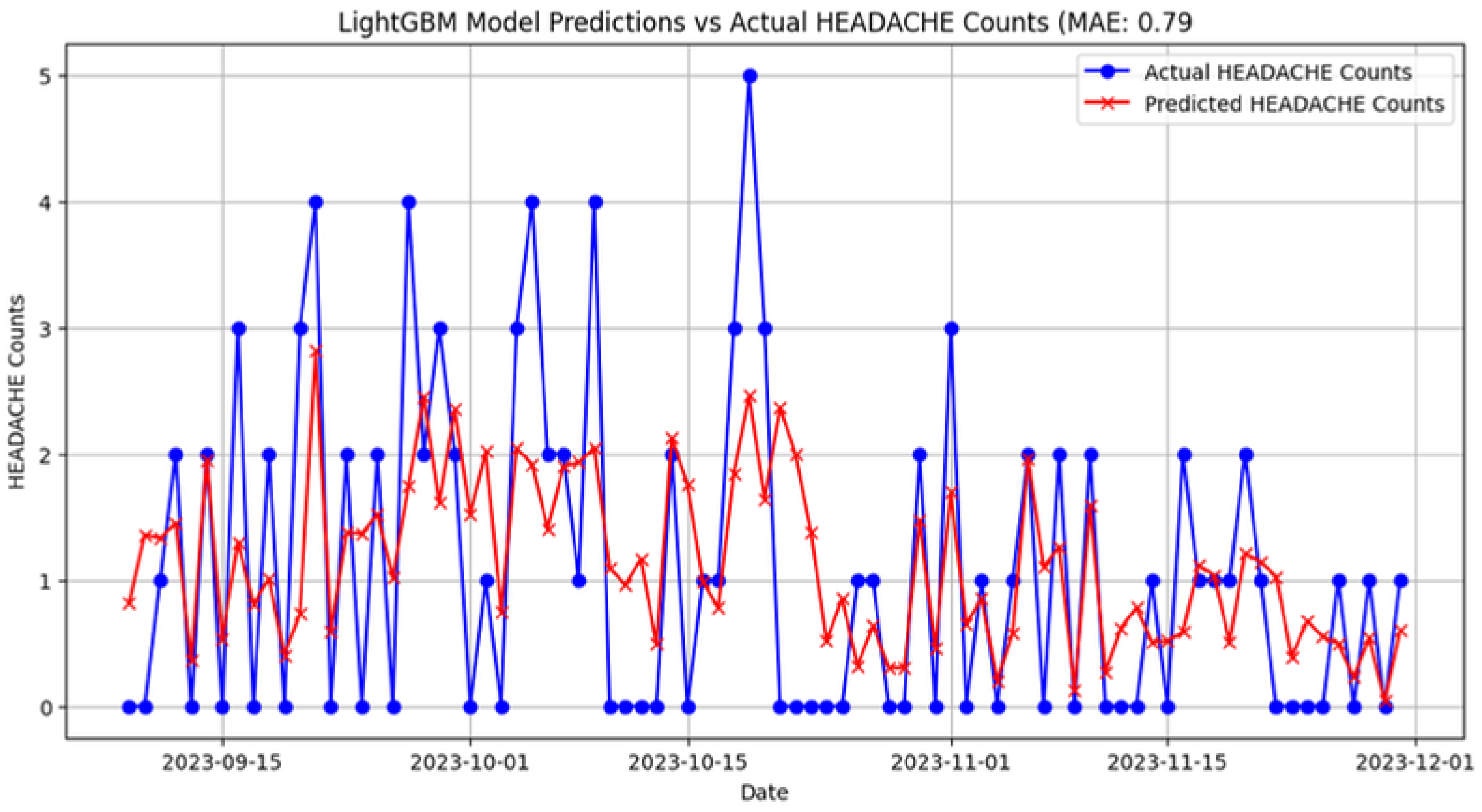


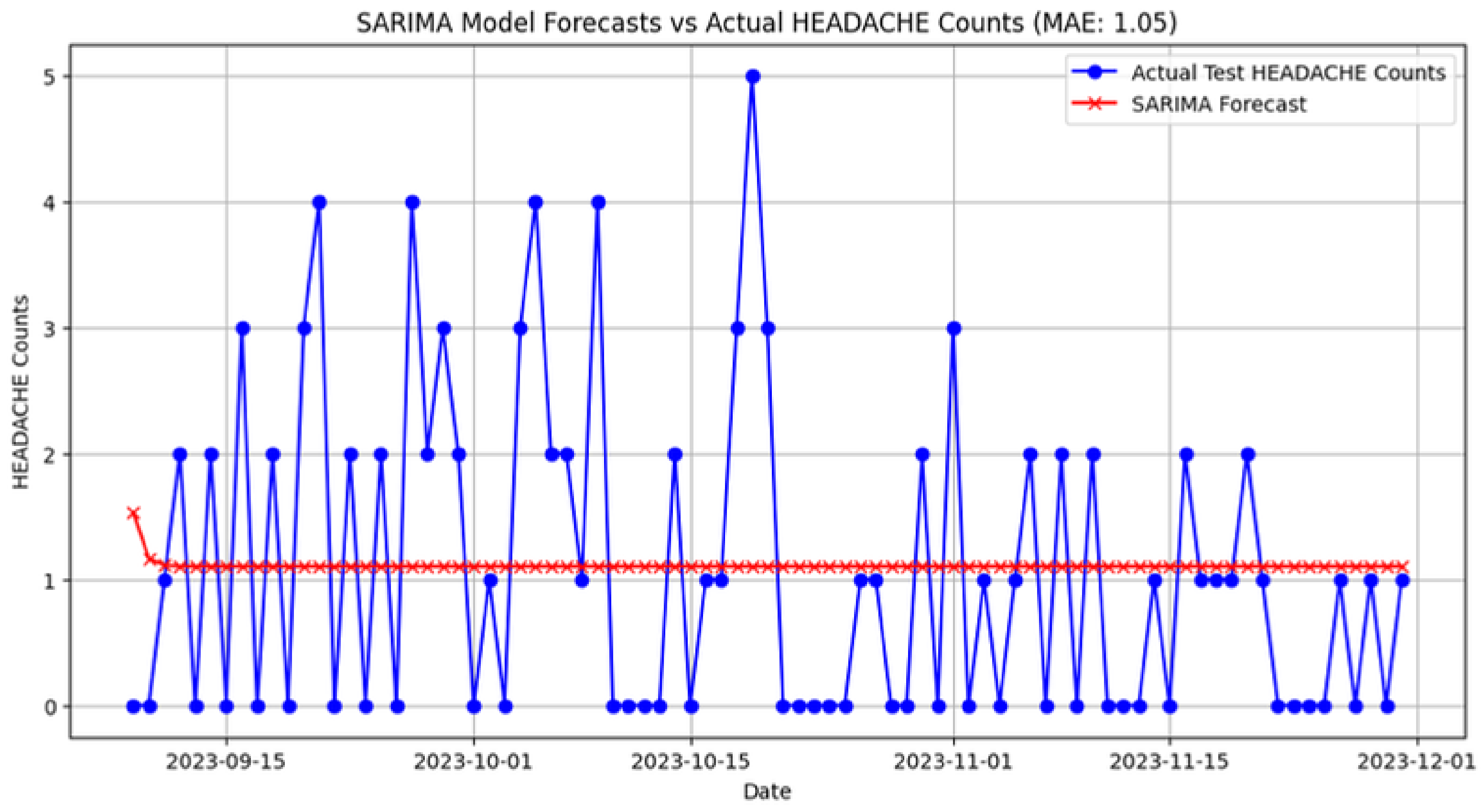


Model performance for Headache

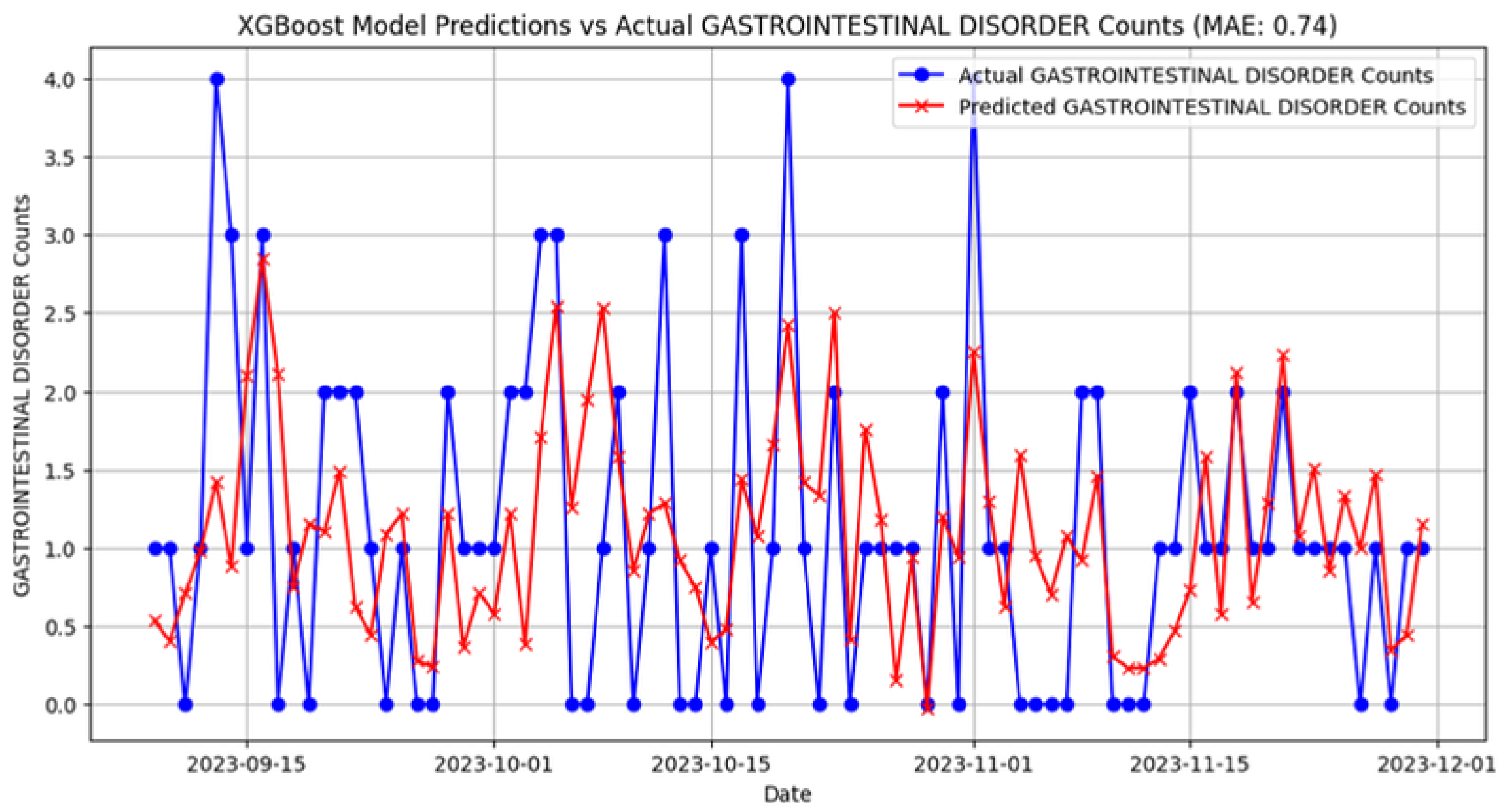


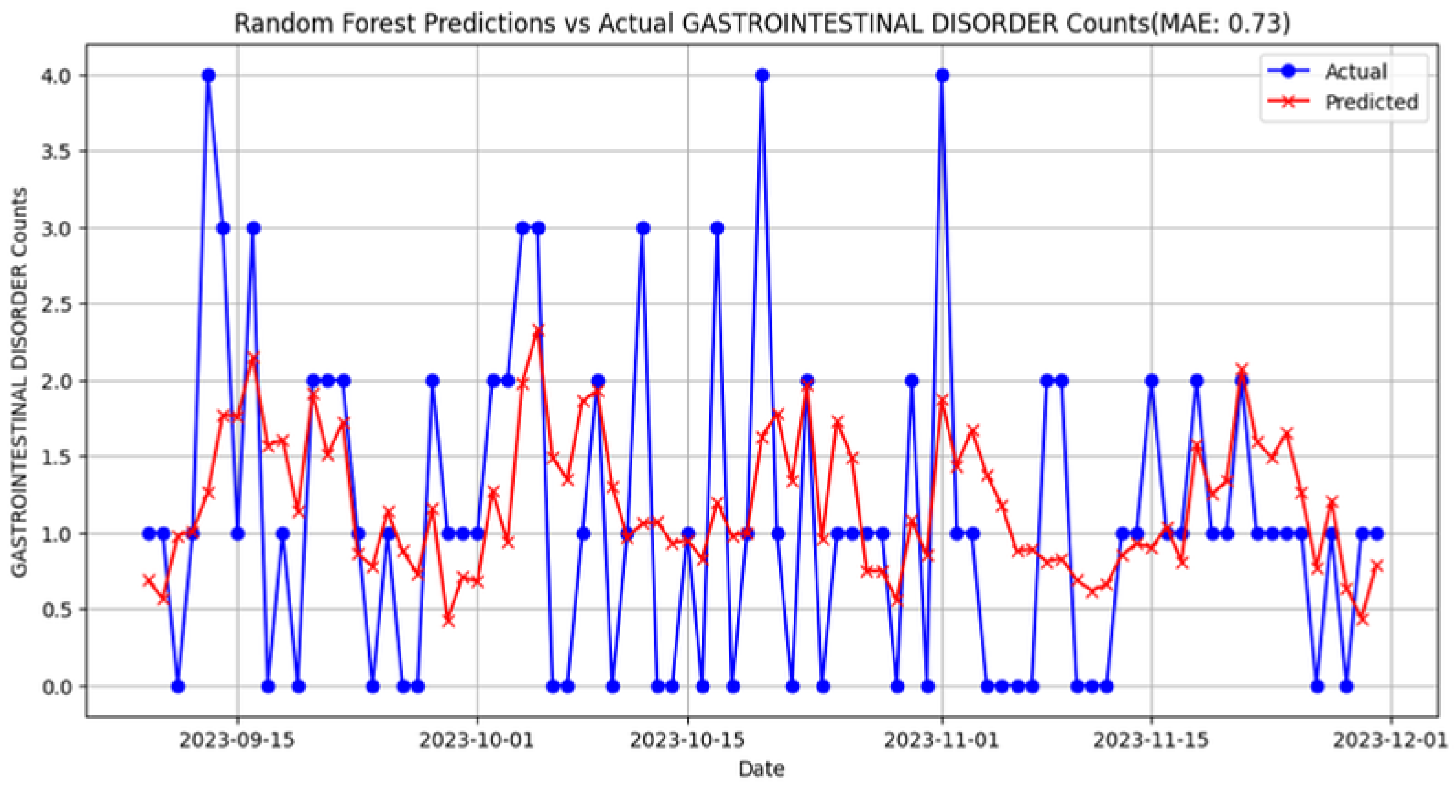


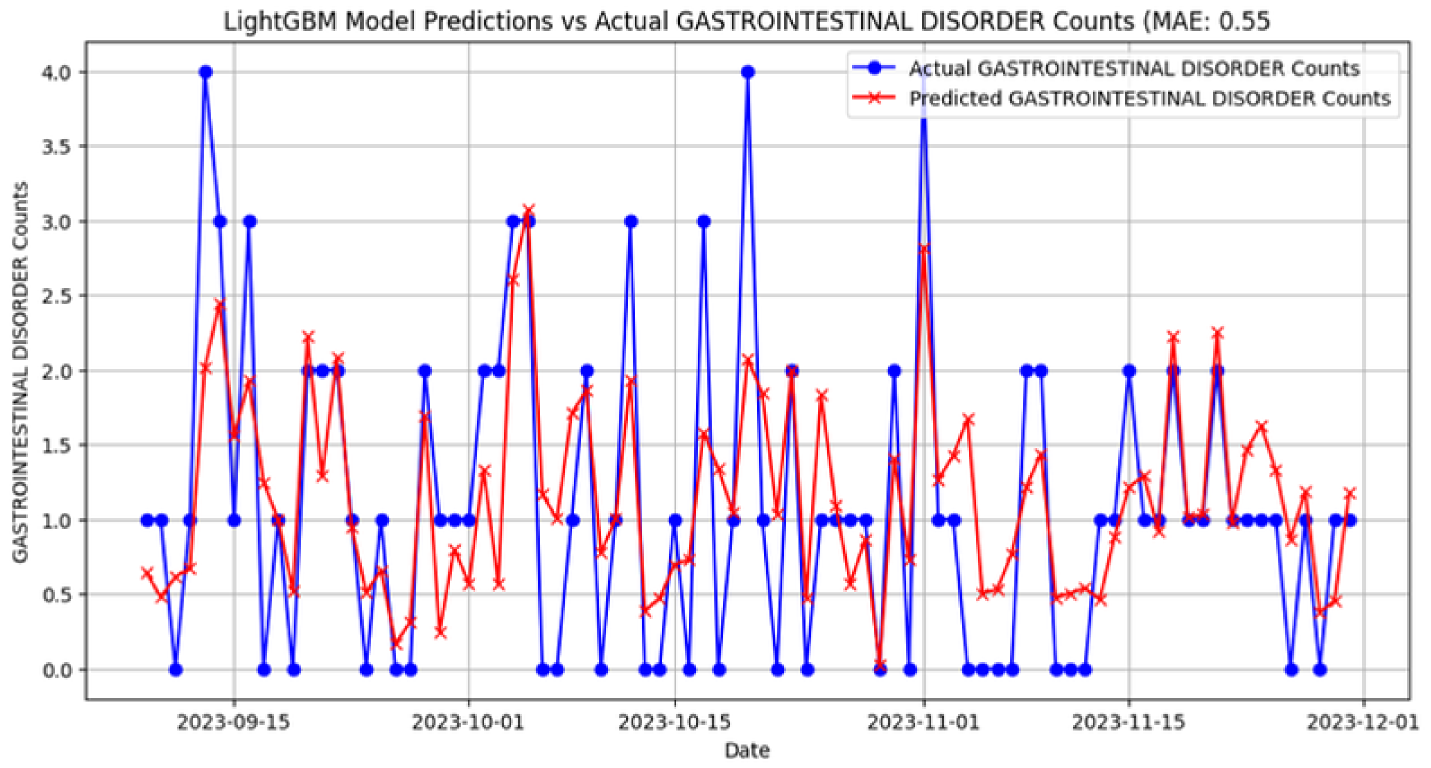


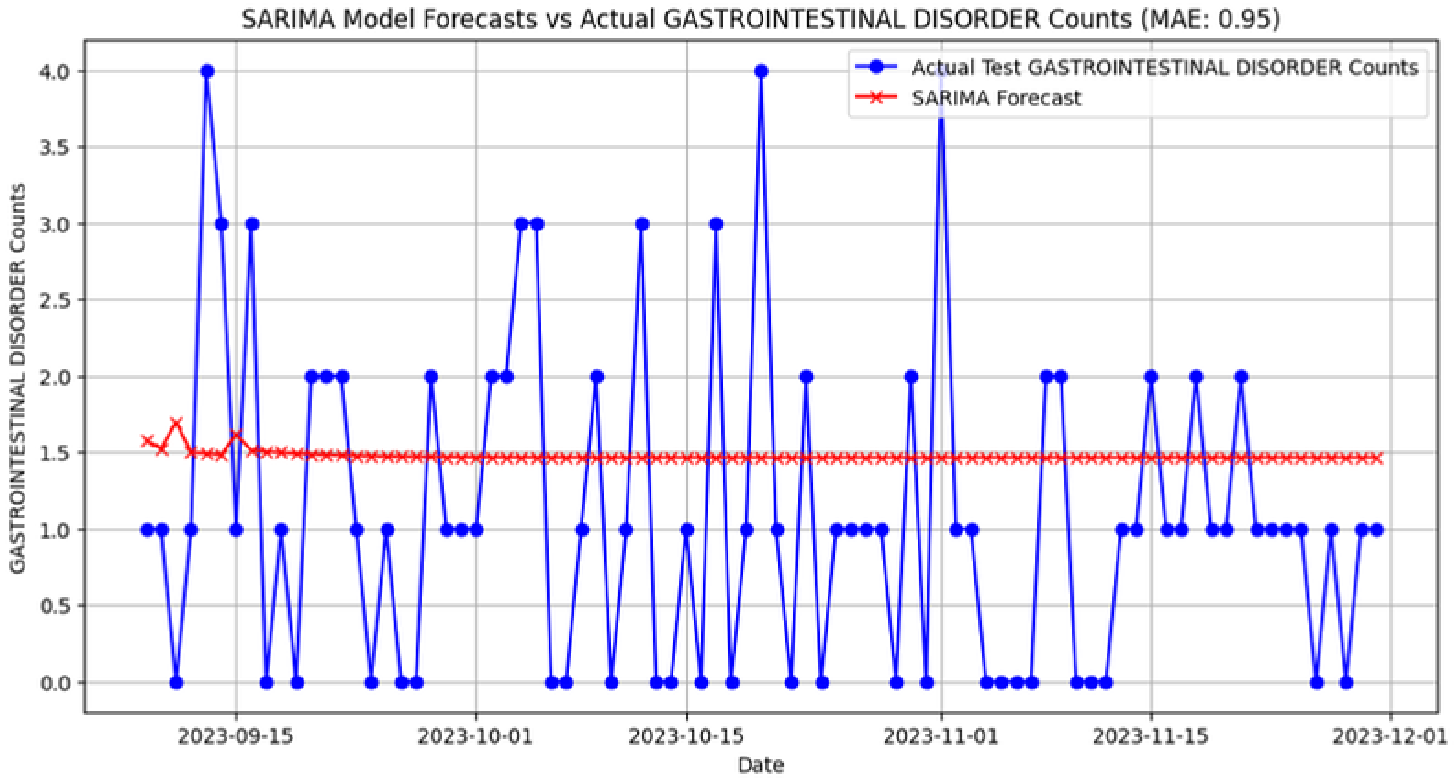


Model performance for Gastrointestinal Disorder

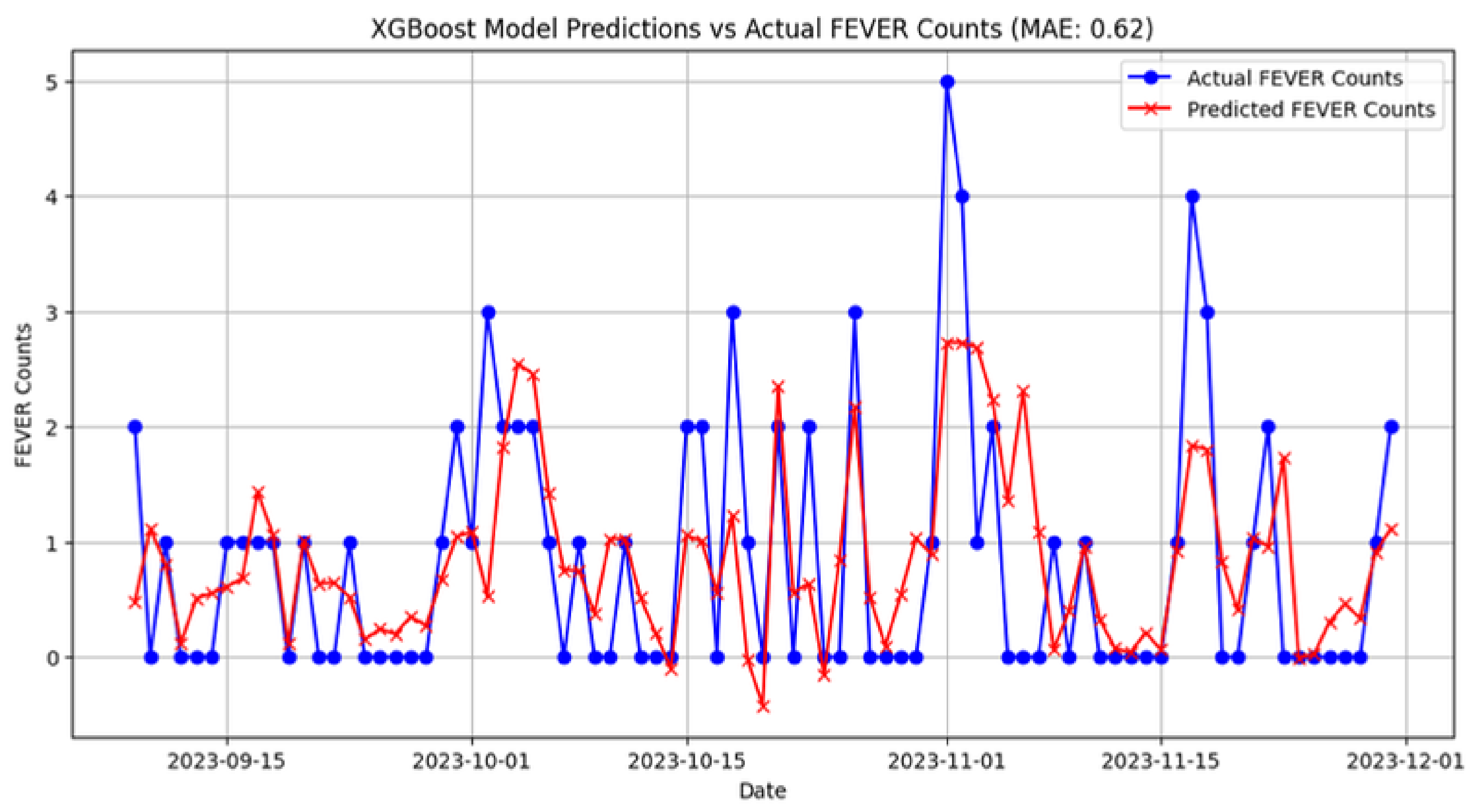


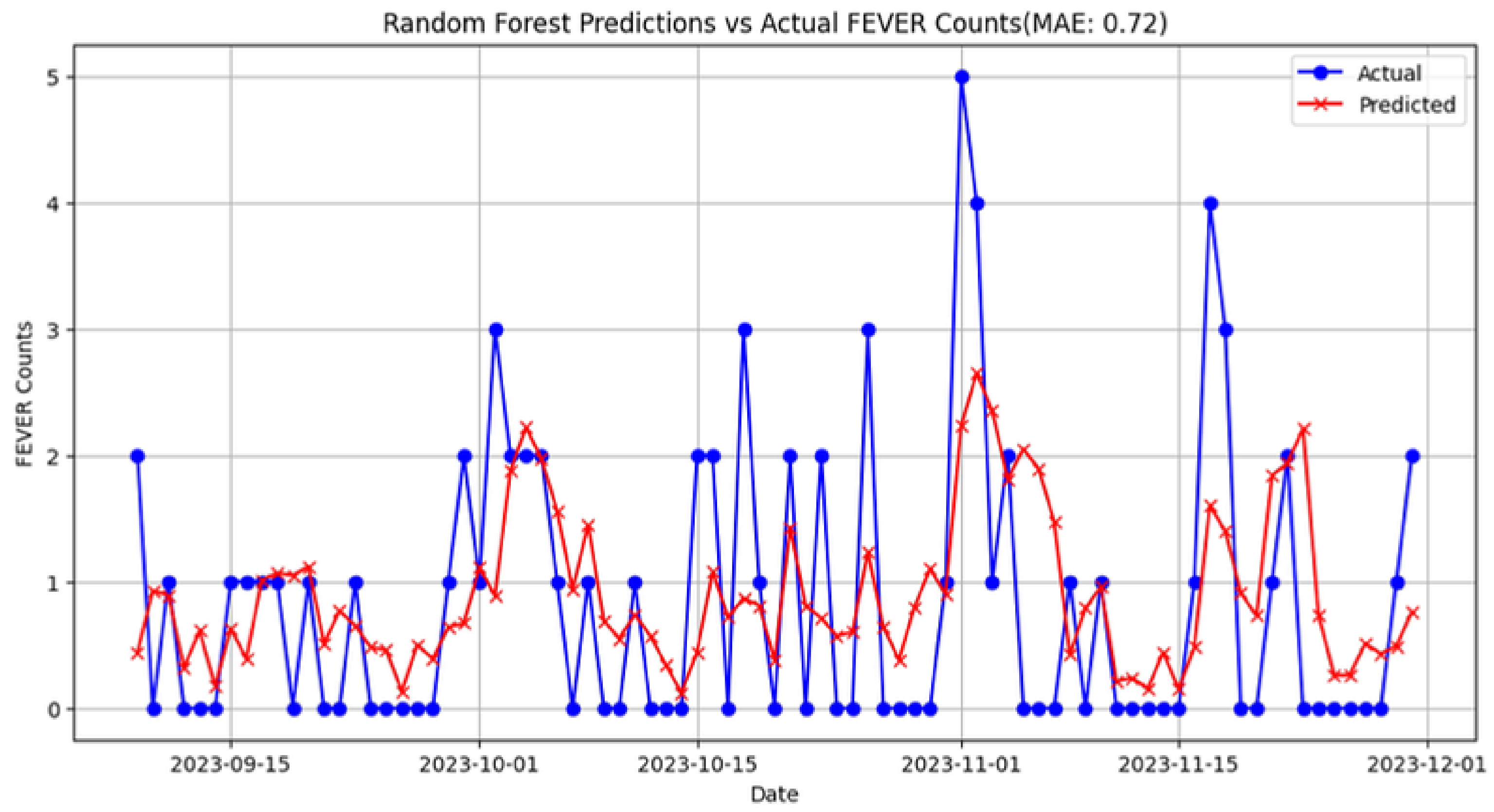


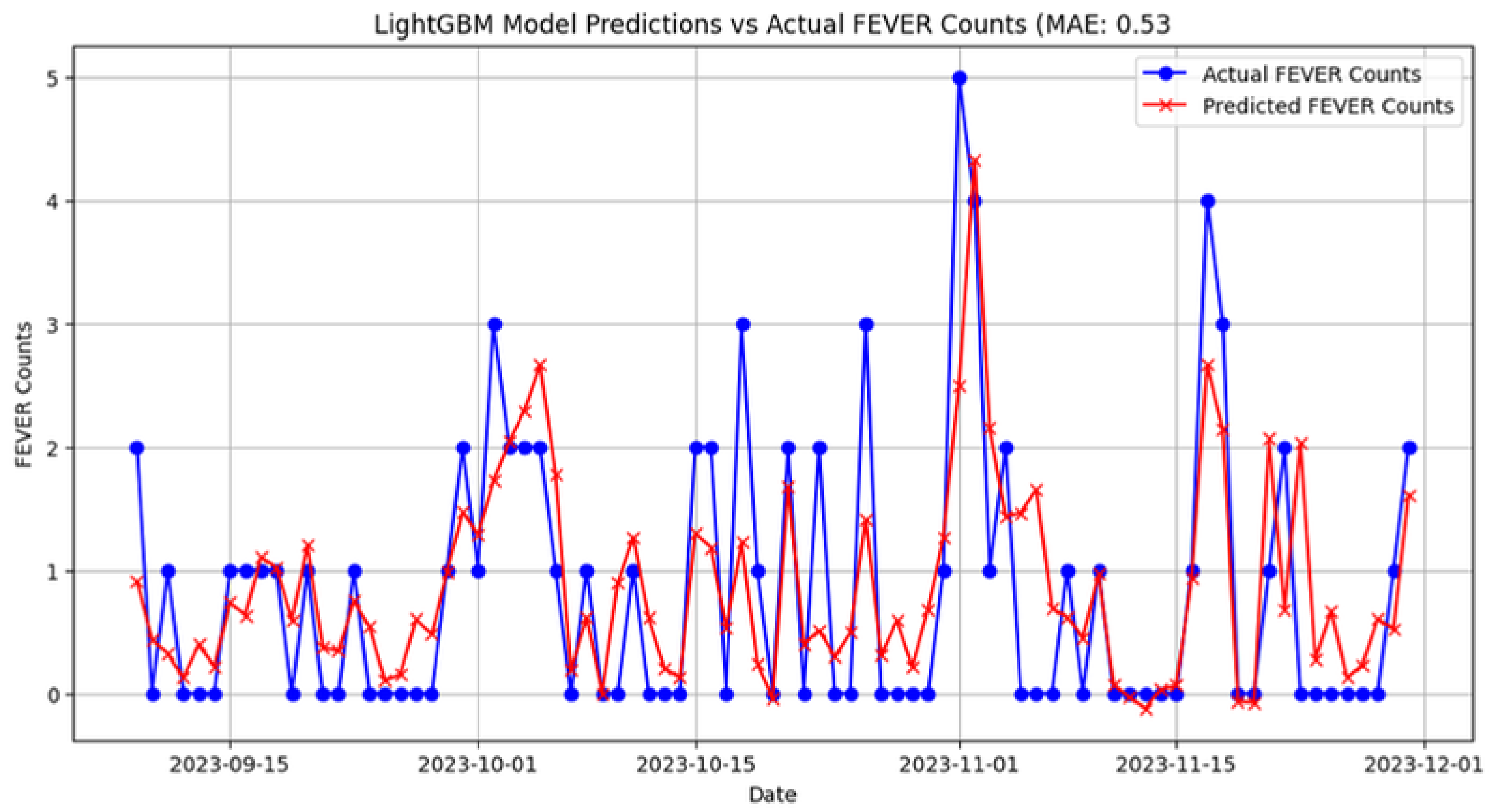


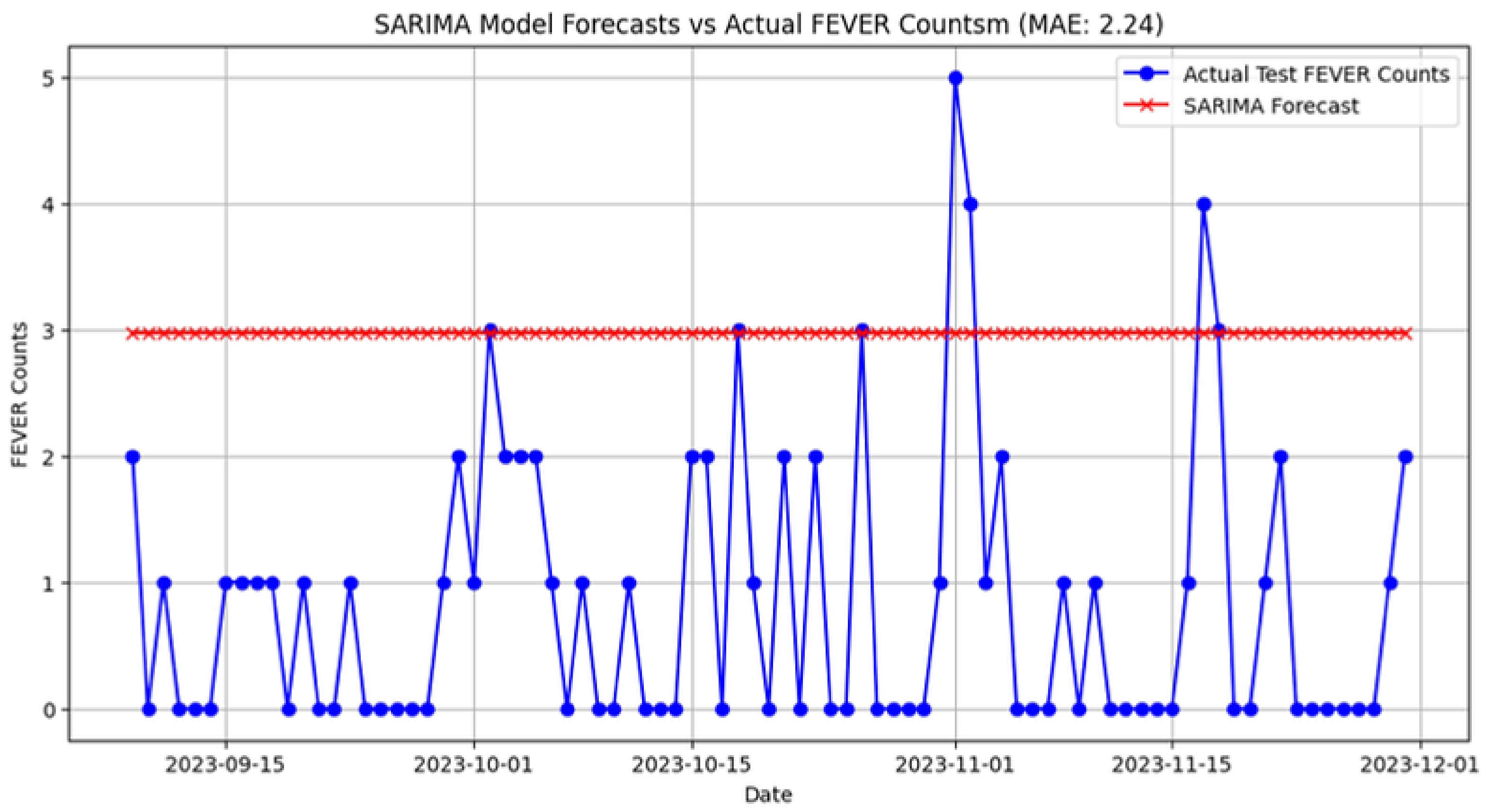


Model performance for Fever

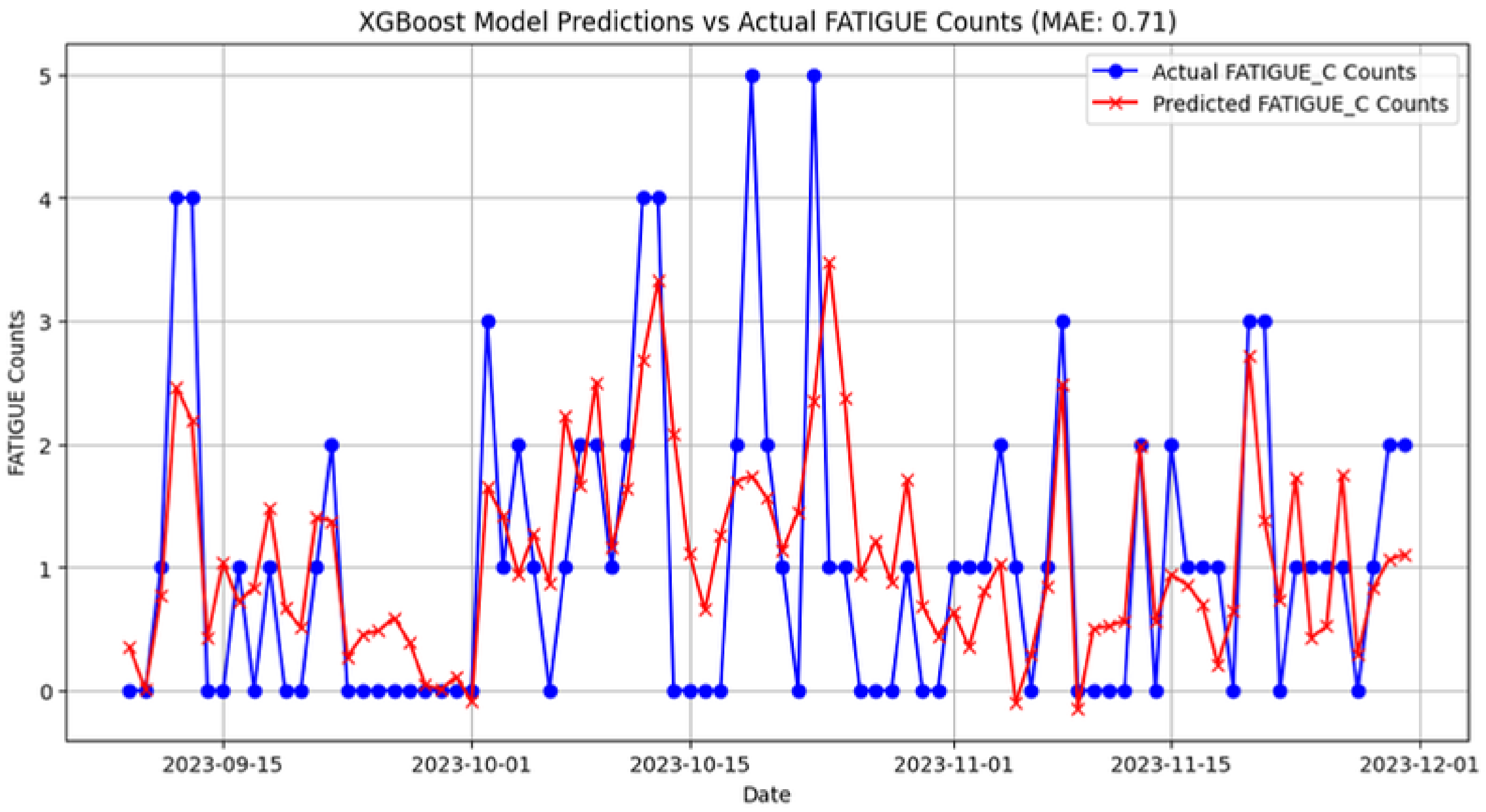


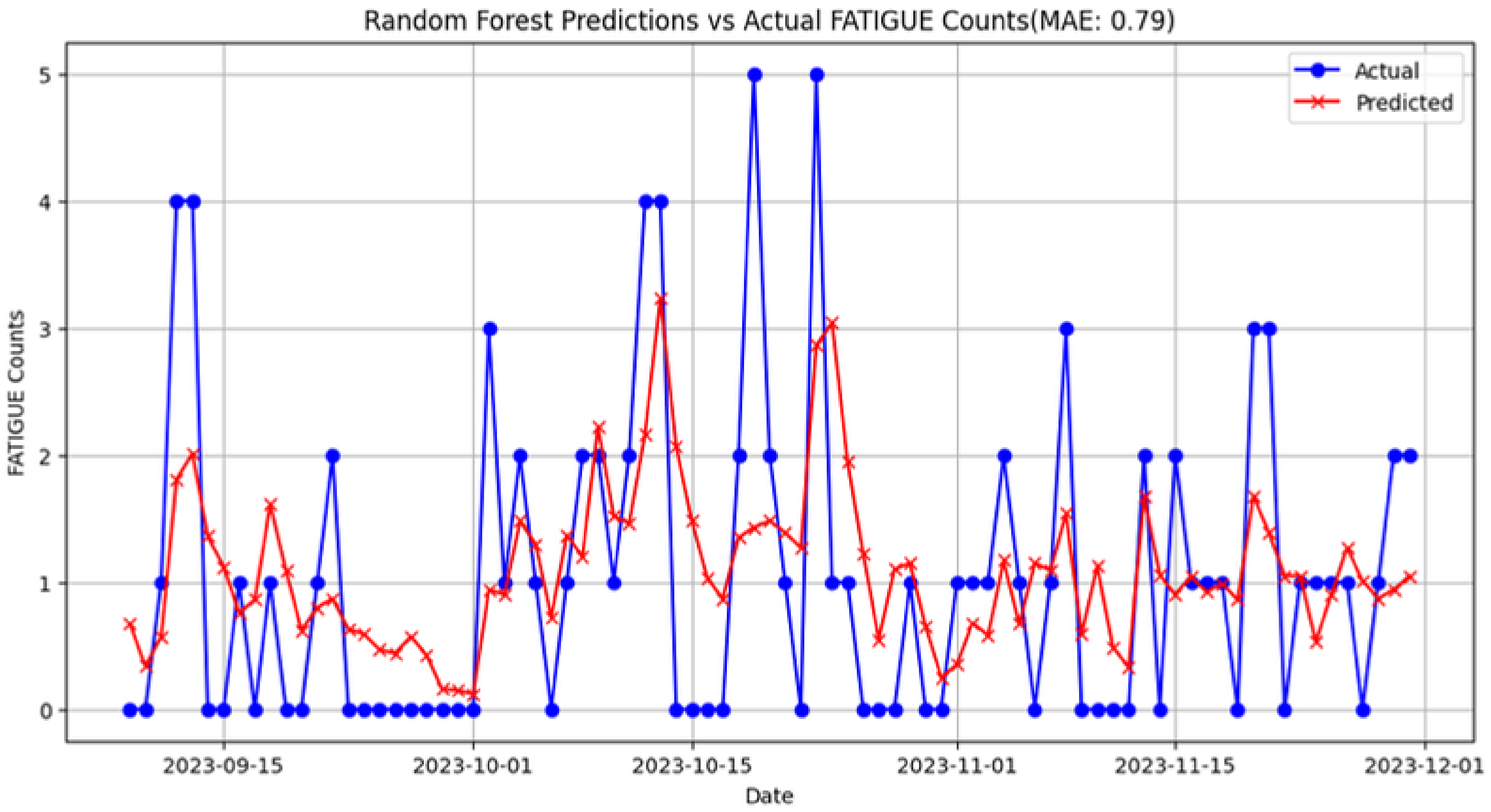


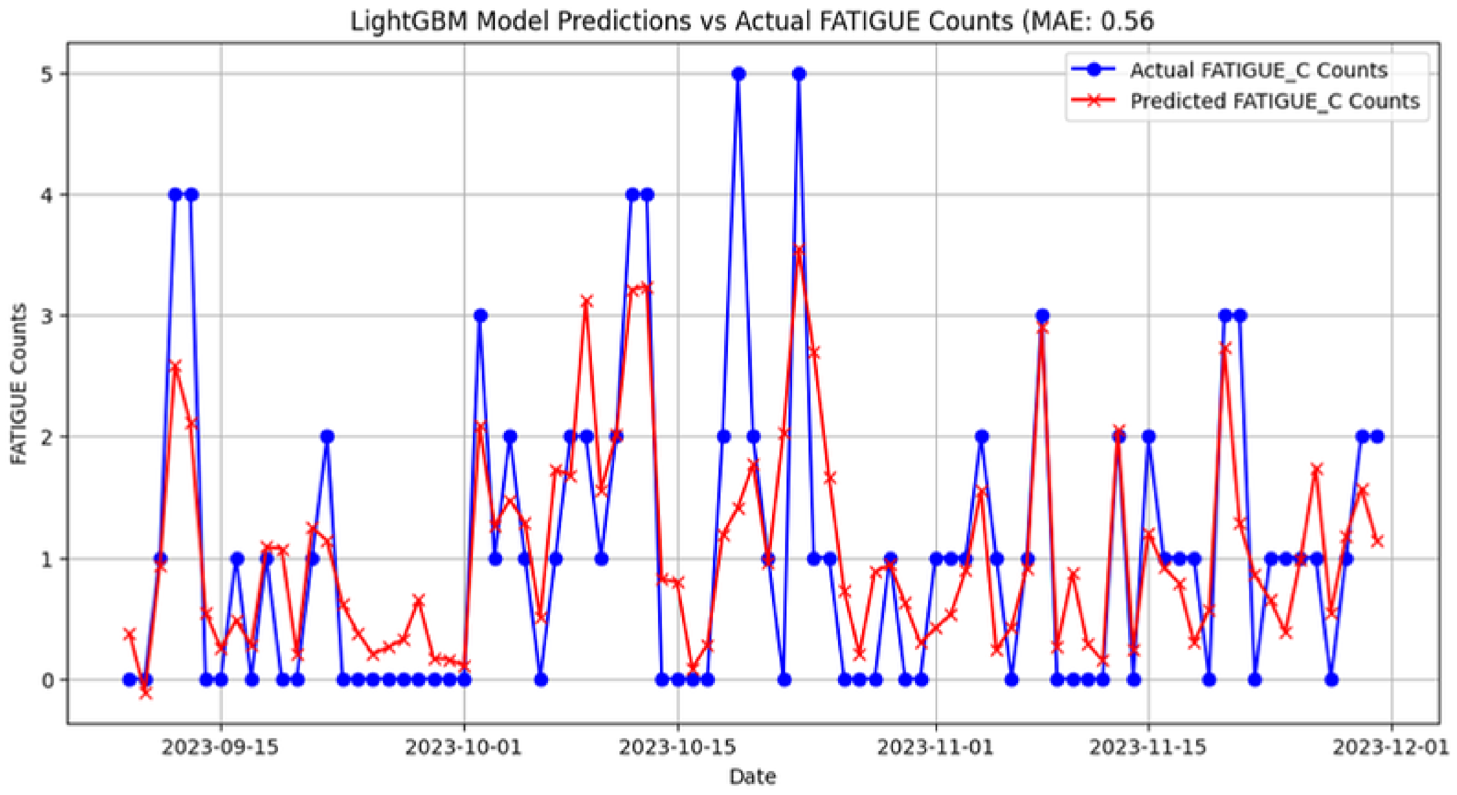


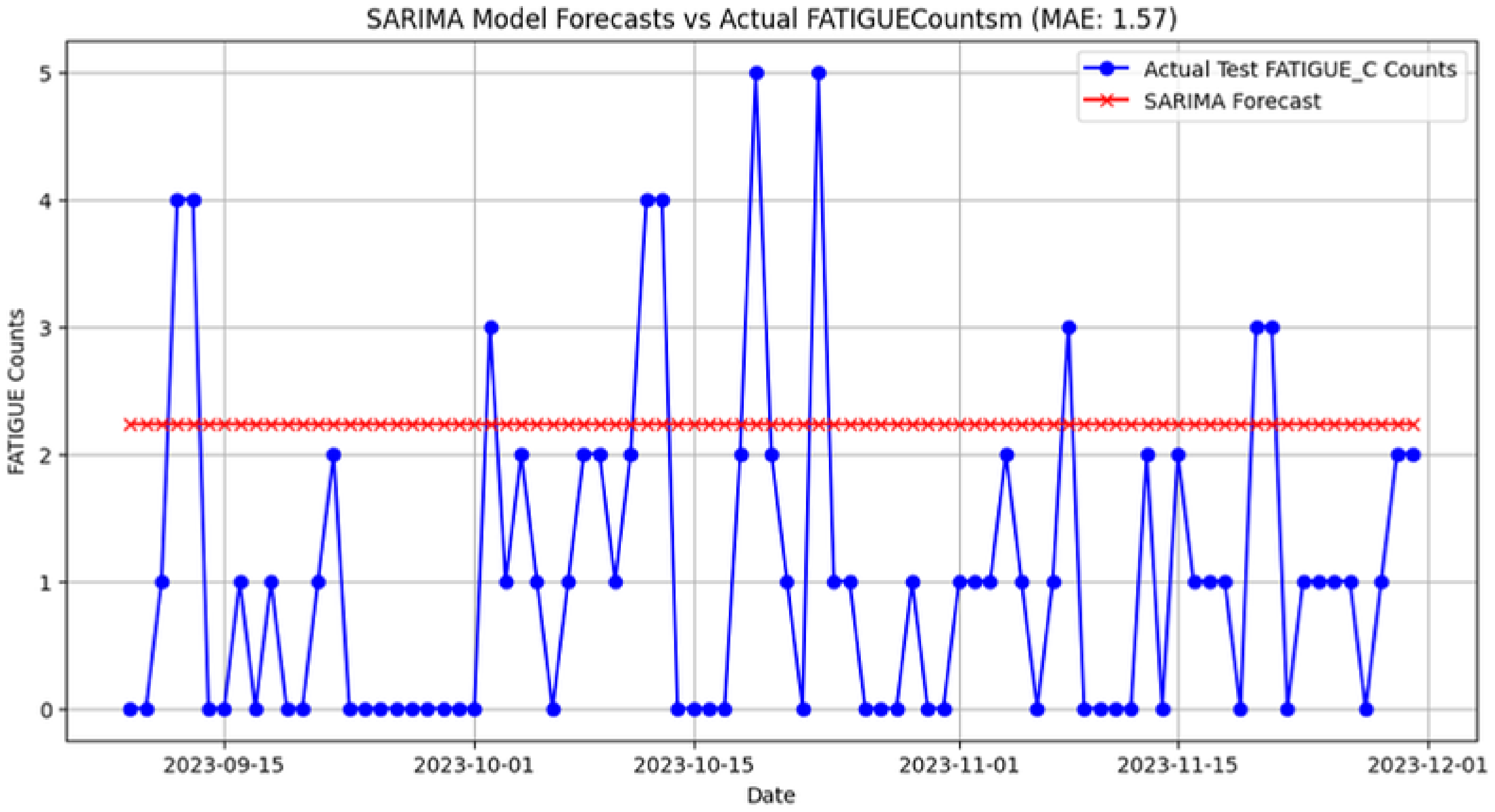


Model performance for Fatigue

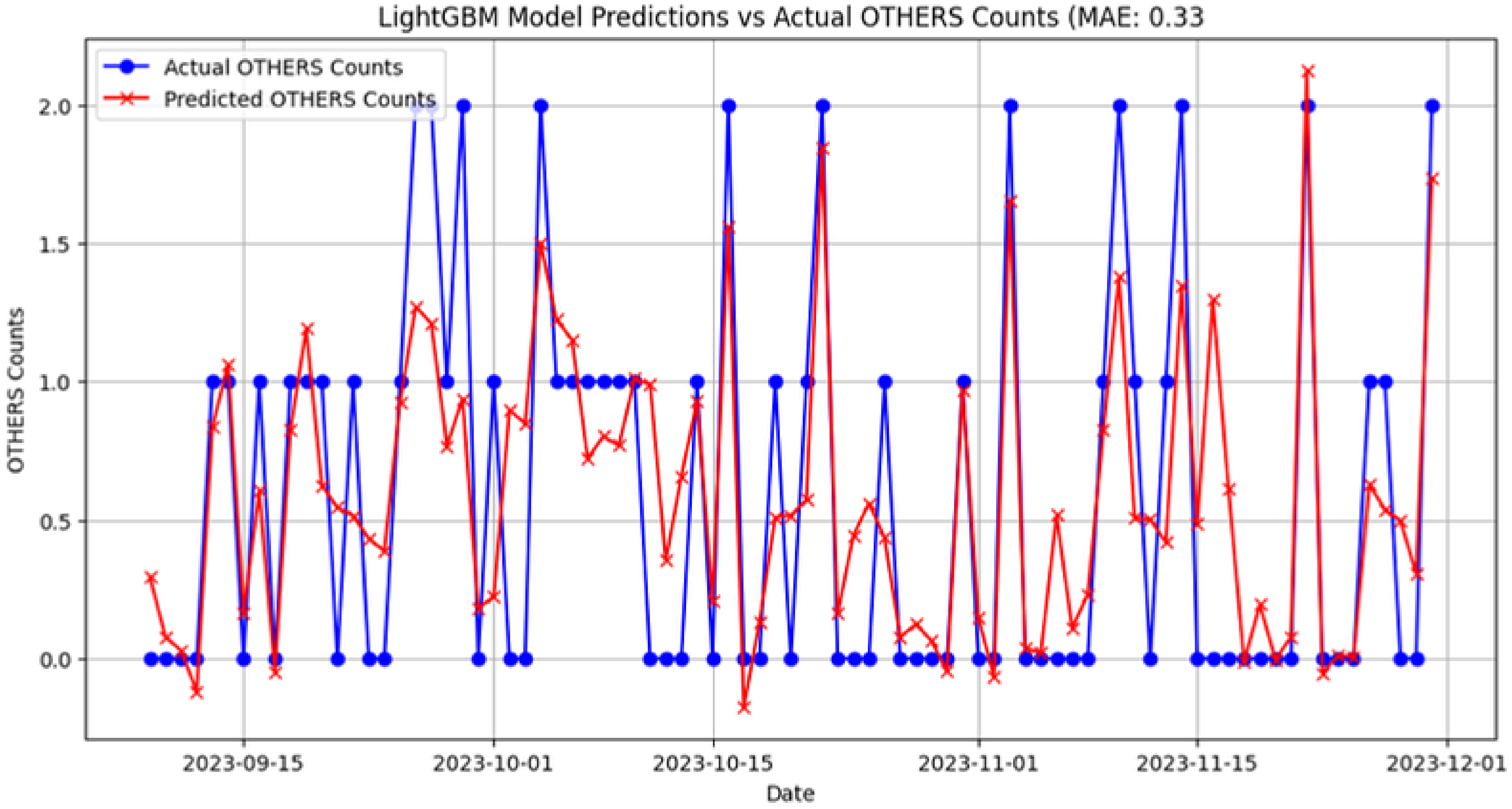




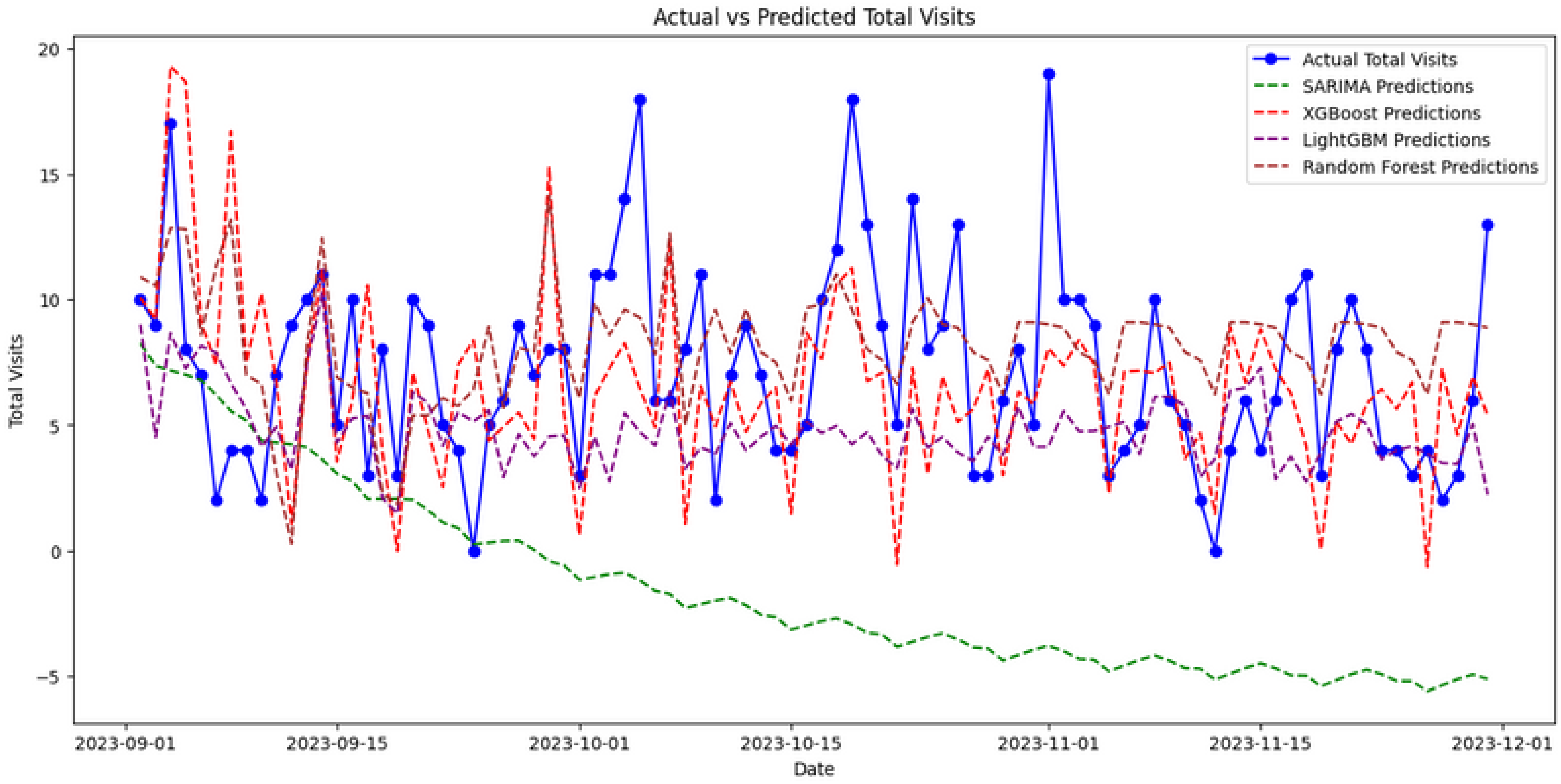




Model performance for all Categories of Symptoms



TOTAL NO. OF VISITS



COMPARISON OF MAE SCORES OF DIFFERENT MODELS

	XGBoost	LightGBM	SARIMA	Random forest
URI	1.05	1.12	6.61	1.28
HEADACHE	0.9	0.79	1.05	0.96
GASTROINTESTINAL DISORDER	0.74	0.55	0.95	0.73
FEVER	0.62	0.53	2.24	0.72
FATIGUE	0.71	0.56	1.57	0.79

LIGHTGBM VS XGBOOSTING

LightGBM

Pros:

- High efficiency and speed: It can handle large amounts of data more quickly than XGBoost, making it a good choice for our Project
- Requires less memory than XGBoost

Cons:

- With smaller datasets, LightGBM can overfit more easily compared to XGBoost
- Might not be as robust as XGBoost in certain scenarios

XGBoost

Pros:

- Model Robustness: More robust and can handle a variety of data types and distributions
- Fine-grained Tuning: More options for parameter tuning to improve the model's performance.

Cons:

- Can be computationally expensive, especially with large datasets
- Generally requires more memory than LightGBM

PERFORMANCE METRICS OF LIGHTGBM

Performance	URI	HEADACHE	GASTROINTESTINAL DISORDER	FEVER	FATIGUE
Mean Absolute Error	1.1228800666007932	0.7893278046742257	0.545100626332439	0.5342199749955514	0.5616901003042772
Root Mean Squared Error	2.8047928291929294	1.241371152458712	1.1599727512865052	0.7340474865021169	0.7838846136333598
Mean Absolute Percentage Error	1.0025902307030914	0.4642807504164691	0.4669016362682631	0.37631194616607133	0.3523159742841836
Mean Squared Error	7.866862814692077	1.5410023381566713	1.3455367837271843	0.5388257124400756	0.6144750874911218

PERFORMANCE METRICS

MEAN ABSOLUTE ERROR (MAE):

- provides an easily interpretable measure of forecasting accuracy by representing the average magnitude of errors in the same unit as the target variable
- less sensitive to outliers than RMSE
- easy to implement
- gives equal weight to both underestimation and overestimation, providing a balanced assessment of forecasting errors.
- scale independence

ROOT MEAN SQUARED ERROR (RMSE):

- sensitive to the magnitude of errors, emphasizing larger errors more than smaller ones.
- squaring the errors in RMSE penalizes larger deviations between predicted and actual values, making it particularly relevant in situations where large errors are of concern - such as healthcare
- allows for easy comparison between different models or different forecasting horizons
- scale independant

HYPERPARAMETER TUNING

XGBOOST:

- LEARNING RATE
- NUMBER OF TREES
- MAXIMUM DEPTH
- MINIMUM WEIGHT
- SUBSAMPLE
- COLSAMPLE BYTREE/COLSAMPLE BYLEVEL/COLSAMPLE BYNODE

LIGHTGBM:

- LEARNING RATE
- NUMBER OF TREES
- MAXIMUM DEPTH
- MINIMUM WEIGHT
- BAGGING FRACTION
- FEATURE FRACTION

RANDOM FOREST:

- NUMBER OF TREES
- MAXIMUM DEPTH
- MINIMUM SAMPLES SPLIT
- MINIMUM SAMPLES LEAF
- MAXIMUM FEATURES

VARIED WITH LITTLE TO NO CHANGE IN ACCURACY

CHALLENGES

Limited Data

Data Processing

Small Population on Campus

Predicted number of medicines for September 2023 - November 2023

Category	Month	Occurrences	Min Range (3x)	Max Range (5x)	Medicines
FEVER	September	24	72	120	COMBIFLAM, DOLO, MEFTAL FORTE, AZEE
HEADACHE	September	38	114	190	COMBIFLAM, NAXDOM, SARIDON, DISPRIN
GASTROINTESTINAL DISORDER	September	35	105	175	DIGENE, SPOROLAC DS, ORS, RACIPER, PANTOP, DROTIN, RANTAC
URI	September	62	186	310	MONTAIR LC, AZEE, ALLEGRA, CITRIZINE, DOLO, SINAREST
FATIGUE	September	19	57	95	COMBIFLAM, ELECTRAL, FLEXONMR, DOLO, DROTIN, VERTIN
FEVER	October	24	72	120	COMBIFLAM, DOLO, MEFTAL FORTE, AZEE
HEADACHE	October	34	102	170	COMBIFLAM, NAXDOM, SARIDON, DISPRIN
GASTROINTESTINAL DISORDER	October	34	102	170	DIGENE, SPOROLAC DS, ORS, RACIPER, PANTOP, DROTIN, RANTAC
URI	October	105	315	525	MONTAIR LC, AZEE, ALLEGRA, CITRIZINE, DOLO, SINAREST
FATIGUE	October	41	123	205	COMBIFLAM, ELECTRAL, FLEXONMR, DOLO, DROTIN, VERTIN
FEVER	November	23	69	115	COMBIFLAM, DOLO, MEFTAL FORTE, AZEE
HEADACHE	November	22	66	110	COMBIFLAM, NAXDOM, SARIDON, DISPRIN
GASTROINTESTINAL DISORDER	November	27	81	135	DIGENE, SPOROLAC DS, ORS, RACIPER, PANTOP, DROTIN, RANTAC
URI	November	56	168	280	MONTAIR LC, AZEE, ALLEGRA, CITRIZINE, DOLO, SINAREST
FATIGUE	November	27	81	135	COMBIFLAM, ELECTRAL, FLEXONMR, DOLO, DROTIN, VERTIN

Deployability

DEPLOYABILITY

01 Possible scenarios for deployment:

- Plaksha's Healthcare Centre can deploy the model, especially in the future when there are more people on campus. The model would have to be trained in intervals to improve accuracy and scale to a larger database
- Other Universities can also potentially use our mode to predict symptom prevalence and optimise their medical inventory

02 Tools for deployability:

- Use GitLab
- Plaksha's Sharepoint platform
- Creating a user-friendly web interface / EHRS system

POSSIBLE CHALLENGES

- Black swan events (i.e., epidemics such as COVID-19)
- Cannot predict injuries and therefore cannot help optimize for related medical stock
- More specificity required for aggregated symptoms
- Data does not include follow-up details to evaluate whether the treatment was successful

POLICY IMPLICATIONS FOR PLAKSHA

- Put efficient and strict guidelines in place for data-entering in the Healthcare Centre
- Further in the future, data-driven campus health policies based on trends in health
- Making Plaksha community more aware of health risks
- Extend Healthcare Centre’s ability to cater to more serious health conditions as well

THANK YOU