MODELING AND PREDICTION OF INFECTIOUS DISEASE CASES IN ILOILO CITY USING MULTIPLE LINEAR REGRESSION

An Undergraduate Thesis
Presented to the Faculty of the
College of Information and Communications Technology
West Visayas State University
La Paz, Iloilo City

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Science in Computer Science

by
Joan Marie D. Arsenio
April Joy N. Gallano
Arle Kai Franco E. Gorriceta
Hiro G. Parcon
Regina Flor P. Tonogbanua

June 2023

Approval Sheet

MODELING AND PREDICTION OF INFECTIOUS DISEASE CASES IN ILOILO CITY USING MULTIPLE LINEAR REGRESSION

An Undergraduate Thesis for the Degree

Bachelor of Science in Computer Science

by

Joan Marie D. Arsenio
April Joy N. Gallano
Arle Kai Franco E. Gorriceta
Hiro G. Parcon
Regina Flor P. Tonogbanua

Approved:

DR. MA. LUCHE SABAYLE
Adviser

DR. MA LUCHE SABAYLE
Chair, CS Department

DR. MA. BETH S. CONCEPCION
Dean, CICT

June 2023

Acknowledgment

The researchers would like to express their outmost gratitude to the following:

Dr. Ma. Luche Sabayle, Research Adviser, for her selfless sharing of her time and knowledge, and for her constant encouragement that led the researcher to the successful completion of this academic endeavor;

Dr. Ma. Beth S. Concepcion, Dean of the College of Information and Communications Technology, for the approval of this thesis;

Dr. Regin Cabacas and Dr. Frank Elijorde, research teachers, for their expertise in research especially during the research proposal;

Their beloved teacher, for their guidance and lessons taught throughout the years;

Their parents, brothers, sisters, and relatives for their unconditional love and financial and moral support;

Their dearest friends and batchmates, for the great company, friendship and support that helped the researchers to finish this study, and

To the Lord God Almighty, for His divine guidance and wisdom that strengthened and moved the researchers to accomplish their research work.

Joan Marie D. Arsenio

April Joy N. Gallano

Arle Kai Franco E. Gorriceta

Hiro G. Parcon

Regina Flor P. Tonogbanua

June 2023

Arsenio, Joan Marie D.; Gallano April Joy N.; Gorriceta Arle Kai Franco E.; Parcon Hiro G.; Tonogbanua Regina Flor P.; "Modeling and Prediction of Infectious Disease Cases In Iloilo City using Multiple Linear Regression". Unpublished Undergraduate Thesis, BS Computer Science, West Visayas State University, Iloilo City, Philippines, June 2023.

Abstract

Local Government Units struggled with providing adequate equipment to health workers, facilities to handle those infected with the disease, as well as precautionary restrictions to halt the spread of the disease in certain areas. With this in mind, the researchers of this study focused on the utilization of tools and technologies to better understand and handle the situation of Covid-19 and infectious diseases in Iloilo. This study generally aimed to make use of newer algorithms such as PythonAF (Python Automatic Forecasting) to forecast future positivity rates, utilizing Multiple Linear Regression. It aimed to make use of newer algorithms such as PythonAF (Python Automatic Forecasting) to forecast future positivity rates, utilizing Multiple Linear Regression. This study used Steamlit for its visualization of the gathered data. It also provided information on how the data was processed. This study also used algorithms such

as PythonAF and Multiple Linear Regression. It was intended to understand the correlation of the variables and the impact positivity rates. As in the result this system model was effective and useful in predicting the positivity rates particularly here in Iloilo City. The proposed system was developed using data collection and manual coding. The system was submitted to the pertinent evaluators after it had been partially created so they could evaluate the most recent version of the created decision support system. The researchers therefore concluded that the proposed system had successfully accomplished the set of objectives that were specified on the first phase of the study.

Table of Contents

List of Tables

List of Figures

List of Appendices

CHAPTER 1 INTRODUCTION TO THE STUDY

Background of the Study and Theoretical Framework

A forecast to global disease outbreak happened over a large area, such as multiple countries or continents, and affected a sizable proportion of the global population. Coronavirus or Covid-19 was initially announced to the World Health Organization back in December 2019. It was not until March 2020 that the outbreak was declared a global pandemic (Coronavirus Disease (COVID-19) Pandemic, 2024). As countries scrambled to handle the outbreak, it became clear that although there were precautions taken - it made a devastating blow to the livelihoods of many of the citizens residing in countries that were less prepared for the outbreak. In the Philippines, different sectors suffered from the effects of the outbreak. Local Government Units struggled with providing adequate equipment to health workers, facilities to handle those infected with the disease, as well as precautionary restrictions to halt the spread of the disease in certain areas (Talabis et al., 2021b). With this in mind, the researchers of this study focused on the utilization of tools and technologies to better understand and handle the situation of Covid-19 and infectious diseases in Iloilo.

Theoretical Framework

This research is based on a study done by Alali (2022) which tackled an approach to forecasting COVID-19 spread via optimized dynamic machine learning models where he stated that "There is still more room for improvement, despite the satisfactory COVID-19 spread forecasting results using the dynamic machine learning models." Another study conducted by (Rath et al., 2020) tackles the use of the Multiple Linear Regression model in the Prediction of new active cases of coronavirus infectious disease COVID-19. Wherein they state "standard acquired uncommon accuracy in COVID-19 acknowledgment. A strong relationship between the factor determines the relationship among the dependent (active) with the independent variables (positive, deaths, recovered)."

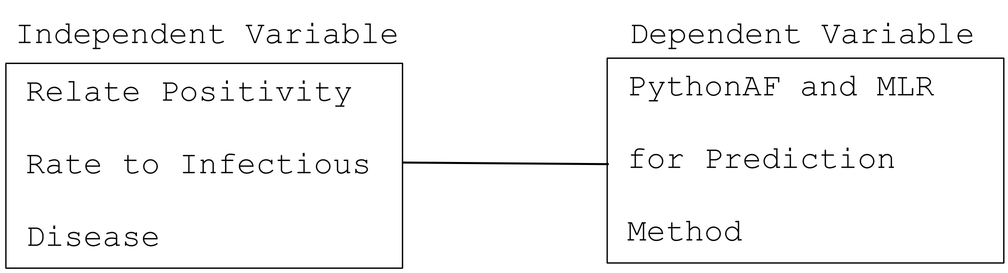


Figure 1. Conceptual Framework Diagram of the Study

As showed in Figure 1, researchers conducted the study to provide adequate equipment and facilities to LGU, the medical field, and health workers to utilize the tools and technologies to better understand and handle the situation of Covid-19 and infectious diseases in specific areas. The researchers developed a decision support system to predict the cases of infectious diseases in Iloilo City.

Data sets and attributes used are the following:

Total Cases. total recorded number of infectious disease cases of Covid-19.

Local Transmission. Recorded number of transmissions in public places.

ROFW. Number of returning OFW recorded in Iloilo City.

APOR. Authorize person to go out in residents such as frontliners.

New Cases. Recorded number of cases per day.

Death. Recorded number of deaths caused by Covid-19.

Lockdown Order. Restrictions given in an area for certain period of time.

Returning residents. Recorded number of persons returning from outside the province.

Positivity Rates. Data retrieved in Iloilo City Covid-19 Emergency Operations Center.

Increase in Cases. Compared data from the data from previous days.

Increase in positivity rates. Compared data of positivity rates from the data from previous days.

Data preprocessing

Extracting data provided by official DOH website and Iloilo City local government daily updates from Facebook page. Data was downloaded and time-frame was selected to fit with the study's focus. Data was not up to date due to some variables being omitted by data source.

Objectives of the Study

This study generally aimed to make use of newer algorithms such as PythonAF (Python Automatic Forecasting) to forecast future positivity rates, utilizing Multiple Linear Regression.

Specifically, it aimed to:

1. Determine whether or not the given variables are related to a disease's positivity rate.
2. Utilize Multiple Linear Regression to determine which factors have a significant relationship with positivity rate to infectious diseases.
3. Develop and Create a system that enables to predict cases of infectious disease.
4. Using Mean Absolute Percentage Error (MAPE), assess the accuracy of Multiple Linear Regression.

Significance of the Study

The study would be an essential contributions of medical field researchers and health workers, DOH, and LGU. With this, the forecast of the future positivity rate of the Covid-19 and may provide a meaningful discovery using the new algorithm used in this study. We may compare which algorithm has better optimal outcomes for forecasting future positivity rates. by means of the following:

Medical Field Researchers: subjects are easy to forecast the pandemic's rise and fall during activities or gatherings. It would be easy for the subjects to research and forecast the pandemic's positivity by using this kind of study. It will help a lot with their medical field research.

LGUs: This study on local government units (LGUs) can help spread early warnings to the people causing the outbreak of the pandemic. Through this research, LGU can forecast the future positivity rate of the pandemic early and can help warn national and local governments.

DOH: Since the purpose of this research is to forecast the future positivity rates of the pandemic, DOH can be able to forecast the future positivity rates of the pandemic by using this research. The DOH can utilize this research or implement a better algorithm generated from this manuscript in their own papers.

Future Researchers: This will serve as one of their bases in review of related studies. They can able to gain knowledge, ideas and outlooks to proceed with their chosen topic for research in the future.

Definition of Terms

For improvement of knowledge and understanding, the following terms were defined conceptually and operationally:

Infectious Diseases. Infectious diseases are conditions brought on by organisms like bacteria, viruses, fungi, or parasites. Numerous species live inside of our bodies. They are generally advantageous or even safe. But in specific circumstances, some bacteria have the capacity to cause disease. Some infectious diseases can transmit from one person to another. (*Infectious Diseases - Symptoms & Causes - Mayo Clinic*, 2022).

In the study, it will be used as the basis of how the researchers deal with and conduct the study's factors and variables contributing to the infectious diseases and will be identified based on the factors contributed.

Multiple linear regression. Utilizes Multiple Regression for prediction to generalize this understanding to a population, one is employing a sample to develop a regression equation that would optimally predict a given phenomenon within a particular population. The purpose here is to utilize the equation to anticipate the results for the sample that was used in the analysis. (Obsorne, J. W.,2019).

In this study, this is one of the algorithms that is used to predict the outcome of the positivity rates of the cases while taking into multiple variables with different values.

Positivity rates. Positive test results are a better predictor of disease transmission than confirmed cases. Using each confirmed case and positivity rate, it calculated the daily reproduction number for each of the eight nations studied. Each country's two sets of reproduction number values were statistically analyzed to see if they differed significantly. (Dallal et al.,2021).

In this study, this is the outcome of the diseases that are predicted and forecasted using the said algorithms. it includes the values that are collected in an area.

Prediction model. Is a mathematical process used to predict future events or outcomes by analyzing patterns in a given set of input data (Lawton G., et.al, 2022)

In this study, proposed model that the researchers used to forecast the future outcome of the positivity rates in a pandemic disease uses Python Automatic Forecasting with multiple regression algorithms.

Python Automatic Forecasting. This is an open-source Python software for automatically developing time-series forecasting models (either univariate or with exogenous data)(Antoinecarme, (n.d.)).

In this study, PyAF performs forecasting using a machine-learning approach. It begins by training a time series model on past values and then uses this model to generate future values (forecast).

Streamlit. A Python-based open source app framework is called Streamlit. It enables us to quickly develop web applications for data science and machine learning. Major Python libraries including scikit-learn, Keras, PyTorch, SymPy(latex), NumPy, pandas, and Matplotlib are compatible with it. (Dhir H.,2019)

In the study, it will be used as one of the languages to model and predict the infectious disease cases in Iloilo City using Multiple Linear Regression Algorithm model.

Delimitation of the Study

The main objective of the research is to create a prediction model algorithm based on positivity rates in the pandemic, employing Python AF with multiple regression algorithms as a tool in the campaign against COVID-19 and other infectious diseases in Iloilo City. The size of the target population was based on the gathered information from Ioilo Covid-19 Emergency Operations Center from December 2020 to May 2022. The study was conducted using google forms to gather the responses of medical personnel, students and IT experts. The study is only limited to the use of infectious diseases so the researchers ought to get the data available in the Emergency Operation Center only. Time constraint, resource and manpower are the reasons why the researchers did not indicate the legend and label of the predicted positivity rates in the system in order for the user to know how to use and understand the it.

CHAPTER 2 REVIEW OF RELATED STUDIES

Review of Existing and Related Studies

Healthcare systems, Infectious diseases, multiple linear Regression, pythonaf and prediction system

The timeliness of public health surveillance systems is an important performance indicator. The timeliness of data varies depending on the infectious disease, that intended to use of the data, and the level of the public health system. The timeliness of National Notifiable Disease Surveillance System (NNDSS) this system was to surveillance data was examined to determine the system that could support early notification and state the response to multistate outbreaks (Jajosky & Groseclose, 2004).

A large-scale infectious disease outbreak, the hospitals will play a key role on the society. The authors aim to understand the current status of the hospital preparedness for the infectious diseas outbreaks in country of Beijing and to provide the basic information's for the prevention and control of the infectious diseases in every hospitals (Zhang et al., 2007).

The COVID-19 pandemic, often known as SARS-CoV-2, According to the study it is considered one of the worst in recent history. The objective of this research study is to examined the multiple aspects of food access and also in health, as well as their spatial correlations and statistical associations with current disease outbreaks. Aside of the primary goal there is another goal that is to explore the regression models to analyze the spread of COVID-19 using these said variables (Almalki et al., 2022).

Malaria is a dangerous problem for the health of the population in Yunnan Province, it is located to the border areas of China. It is important to understand how to accurately measure the disease transmission. The objective of this research is to determine the role of slide positivity rate (SPR) in malaria transmission in Mengla County, Yunnan Province, in China. The Centers for Disease Control and Prevention (CDC) and Mengla Bureau of Statistics, China, it provides the data on annual malaria disease cases, SPR, and socioeconomic factors from year 1993 to 2008.These relationship between socio-ecological factors and malaria incidence was studied using multiple linear regression models (Bi et. al, 2012).

The Covid-19 virus outbreak was first to reported in late of December 2019, infects more than 7 million of people and killing more than 0.40 million of people worldwide. The first case it is discovered in late December year of 2019, infecting more than 7 million of people and killing more than 0.40 million of people worldwide. According to the authors The first case was discovered in January 30, 2020, in the country of India, and the number exceeded 0.24 million by June 6, 2020. This research paper it provides the a detailed examination of recently developed prognostic models and also to predicts the number of confirmed cases, recoveries, and deaths caused by COVID-19 in India. The Correlation coefficients and multiple linear regression was used to prediction, autocorrelation and autoregression was used to improve the accuracy. The predicted number of cases of said disease agreed well with the actual result values, with a 0.9992 R-squared score (Kumari et al., 2021).

The SARS-CoV-2 virus has a significant impact to the public health, social issues, and financial problems. Despite of the modern medical and technical technology were used, in predicting the spread of the infectious disease it has been extremely difficult. Healthcare systems, such as hospitals, are using predictive models to gain review into the impact of COVID-19 on outbreaks and possible sources. The Predictive research, it found that the LSTM-GRU stacked model forecast is more effective than existing models with better prediction results (Sah et al., 2022).

From the point of view of risk analysis, the detection of catastrophic events and the prediction of the final stage has become critical. When all data are provided, the statistical methods provide a accurate parameter estimates. However, if the raw data is incomplete, the accuracy of the estimates will suffer. As a result, the statistical methods used are ineffective at predicting future trends. Using these methods for predicting the spread of infectious diseases using the differential equations can sometimes provide the accurate estimates for the final stage. However, these methods required for some inspection time, which results in analysis delay of about a week when trying to predict future trends. The authors used the social networking systems to detect the disasters and predict the future trends much earlier. In this study, the researchers proposed a method to predict the future of flu trend using Twitter application. The researchers investigated the feasibility of constructing using regression model combining Twitter messages and CDC data on influenza-like illness (ILI) and found that a multiple linear regression model with regularized ridges outperforms a simple linear regression model and other unadjusted least squares methods. Multiple linear regression with a ridge model can significantly improve the prediction accuracy (Hirose and Wang,2012).

The total number of confirmed cases of COVID-19 since the outbreak of the Omicron variant in November 2021 continues to rise, in many countries, this infectious disease creates a significant challenge to prevention and control. Between November 1, 2021 and February 17, 2022, daily confirmed cases of COVID-19 worldwide were used as a modeling database, and ARIMA, MLR, and Prophet algorithm models were developed and compared. Mean absolute error (MAE), mean absolute percentage error (MAPE) and root mean squared error used to assess prediction performance (RMSE). In terms of fit, this study found that ARIMA (7, 1, 0) was the best model with lower MAE, MAPE, and RMSE values than MLR and Prophet (Zhao et al., 2022).

These studies aim to developed an assumption-free, the data-driven model to accurately predict the spread of COVID-19 disease. This first used Bayesian optimization to tune the Gaussian Process Regression (GPR) hyperparameters. The results highlighted the superior performance of dynamic GPR compared to other models. Finally, the researchers provided a confidence level for the predicted results and showed that the predictions are within a result of 95% confidence interval (Alali et al., 2022).

The Regression models used to predict the spread of the pandemic infectious disease, such as COVID-19. Multivariate linear regression (MvLR) it predicts mortality and the number of confirmed or active cases each day. The government and healthcare industry will benefit from the ability to stop the spread of the epidemic. Deep learning methods can be scaled up to expand the study and accurately predict its expansion (Suganya et al., 2020).

The Korea Centers for Disease Control (KCDC) a surveillance system is to reduced communicable diseases.

This study it predicts infectious diseases by optimizing the parameters of deep learning algorithms while taking into account big data, including social media data. The top-10 DNN and LSTM models improved their average performance in varicella prediction by 24% (Chae et al., 2018).

To investigate the relationship between effective distance and epidemic spread path and to improve the predictive ability of infectious disease spread, reporting date, and outbreak date of first diagnosed disease based on data from the Wuhan COVID-19 epidemic. Patients were taken as arrival times, respectively, to create a linear regression model of effective distance and arrival time. In different provinces and cities, the logarithm of the cumulative number of confirmed cases with a base of 5 was taken as the criterion for determining the level of cumulative confirmed cases. Based on this, a linear regression model of the effective distance and the level of cumulative confirmed cases in provinces and municipalities was created (Zhou, 2022).

Various governments have used community quarantine to combat the current COVID-19 pandemic; however, this is not the only non-therapeutic method to effectively control the spread of infection. We investigate the SEIR model with a non-linear incidence rate, introduce two parameters, and simulate the effect of quarantine (Q). the researchers compared this with the recently developed Q-SEIR model and show that COVID-19 can be controlled without strict community quarantine conditions. The researchers examined indices of sensitivity and elasticity of parameters about a reproductive number (Jm et al., 2020).

A non-linear model for the study of infectious diseases is discussed. To explain different levels of occurrence or changing pathogen transmission across regions, area-specific and sometimes spatially correlated random effects are introduced. This study, the model is applied to weekly influenza case counts in southern Germany and monthly meningococcal case counts in France (Paul & Held, 2011).

The government of Saudi Arabia has developed an expectation system model that help to recognize COVID-19 cases and normal cases from CT scans. This system model uses an artificial intelligence (AI) and also deep learning to work on the indicative interaction of chest CT scans. It would be contrasted with current models and considered more accurate (93% accurate) than current techniques (Mamoona & Alsayat, 2020).

This study, completed by the Corona Tracker System research team, aims to predict and also to forecast the cases, deaths, and recoveries from COVID-19 using predictive modeling. The goal is to predict and forecast COVID-19 cases, deaths, and recoveries using predictive modeling. The system helps in interpreting trends in public opinion regarding the dissemination of related health information. The real-time data query is complete and displayed on the website. The resulting data is then used for Susceptible-Exposed-Infectious-Recovered (SEIR) predictive modeling (Hamza et al., 2020)

The model tried to offer a solution by offering to preventing the spread of infectious diseases. Global data on COVID-19 is openly available for review through the prediction of COVID-19 spread patterns. Since most of the models now used in this field are black boxes, the glass box model can reveal the influence of the predictive functions on the prediction results. Compared to individual classifiers examined for the ensemble, STRIM is a robust, interpretable classifier model, providing 0.99 levels of prediction accuracy (Yadav et al., 2021).

The prevention measures taken by the local authorities to control the spread of the coronavirus disease (COVID-19) in Kuwait are being studied. Findings suggest that infection rates will rise exponentially if precautions are not strictly followed. This system is to build with Python language modules and runs on official Kuwaiti raw data from February to May 2020. For values in the range of 3 to 4, our results show that the SIR model is almost perfectly fitted to actual confirmed cases of infection and recovery (Alenezi et al.,2021).

This research study proposed a hybrid autoregressive integrated moving average (ARIMA) and Prophet model to predict the daily confirmed cases and cumulative confirmed cases. The auto- Arima built-in function was used for the first time to select the optimal values of the hyperparameters of the ARIMA system model. A modified ARIMA model was then used to find the best fit between the test and forecast data to find the best combination of model parameters. Articles, blog posts, and reports from virologists, scientists, and health experts related to the third wave of COVID-19 this collected using the Python web scraping package Beautiful Soup. The opinions (sentiments) about a potential third wave was analyzed using natural language processing (NLP) libraries (Mohan et al., 2022).

The coronavirus disease (COVID-19) pandemic began in December year of 2019 in the Wuhan province of mainland China and since of spread in worldwide. The researchers used several time-series models to predict the total number of cases and deaths per 30-day window. The autoregressive integrated moving average (ARIMA) and long short-term memory (LSTM) models and the susceptible-infected-recovered (SIR) model. It shows the forecast while locked. The total number of cases was increased, but in a controlled manner, with an accuracy result of 87%. After the relaxation of the blocking rules, forecasts indicated an obstreperous situation with 60% accommodation (Tulshyan et al., 2020).

The Future clinical research focus on the strategies to moderate the effects of elevated levels of these markers to improve treatment and reduced the costs. This study is retrospective in nature, with data collected from a single clinic. The researchers could not exclude the possibility that some treatments given before hospitalization affected the results of NLR, dNLR, MLR, PLR, SIRI, and SII. Both NLR, dNLR, and MLR algorithm was determined at hospital admission to have a high predictive value in patients with COVID-19 disease (Cîtu et al., 2022)

The crucial to assess the potential impact of COVID-19 in India and predict how it will behave in the coming days. In this study, genetic programming (GP)- based prediction models were developed for confirmed cases (CC) and death cases (DC) in the three most affected states, namely Maharashtra, Gujarat, and Delhi, as well as the country as a whole. This prediction models are presented in the form of explicit formulas, and the impotence of the prediction variables is investigated. These case, statistical parameters and metrics is used to evaluate and validate the developed models. According to the findings in study, the proposed GEP-based models use simple joint functions and are highly reliable for predicting the time series of COVID-19 cases in India (Salgotra et al., 2020).

The COVID-19 disease declared a pandemic by the World Health Organization in early month of March and now poses a serious threat to the human community in almost all countries. Everyone in related research community is using technology to figure out when it will stop and how to make the world healthy again. Multiple linear regression outperforms support vector regression in this comparison. As this result, the said algorithm could be used to predict outbreaks in real-world applications (Dash et al., 2022).

The only way to prevent the spread of COVID-19 disease is to anticipate outbreaks and take precautions. This study provides a machine learning technique based on the Adaptive Neuro-Fuzzy Inference System (ANFIS) that can predict possible outbreaks in India. The datasets obtained from there cloud computing, ANFIS-based prediction system tracks the spread of the infectious disease. Through cloud datasets, the ANFIS technique predicts the epidemic peak and the number of cases infected with COVID-19. ANFIS chosen for this study because it has both numerical and linguistic expertise, as well as the ability to classify data and identify patterns. The method only predicts outbreaks but also tracks the disease and suggests measurable policies to manage the COVID-19 epidemic. The prediction obtained shows that the proposed technique tracks the spread of the COVID-19 epidemic very effectively. The result shows that the growth of the infection rate decreases by the end of 2020 and that the peak of the epidemic is delayed by 40–60 days. The ANFIS technique provides a prediction result with a low mean square error (MSE) of 1.184 10-3 and an accuracy of 86%. The study provides a critical information to the public health providers and the government in their efforts to control the COVID-19 epidemic. (Kumar et al., 2021).

One of the factors contributing to the spread of the virus in Indonesia is the lack of public knowledge and awareness about the prevention and control of COVID-19. In addition, there are challenges in conducting surveillance, early detection, contact tracing, infection prevention or control, risk communication, and people's empowerment. This is due to the lack of public implementation and testing of artificial intelligence methods for diagnosing COVID-19. The goal of this study is to use the C4.5 algorithm to diagnose a surveillance classification that includes PDP, ODP, and OTG. The results showed that the diagnosis of the surveillance category COVID-19 disease is successfully modeled into a decision tree with PDP, ODP, and OTG classifications using the C4.5 algorithm. The three-class confusion matrix test procedure yields an accuracy rate of

92.86%, placing it in the category of excellent classification (Wiguna & Riana 2020).

PyAF is an open-source Python automatic prediction library built on top of popular Python data science modules such as NumPy, SciPy, Pandas, and SciKit-Learn. PyAF is an automated process that uses machine learning to predict future signal values. It offers features comparable to some popular commercial automatic forecasting products (Antoinecarme, n.d.)

This study explains that the one of the most important limitations of MAPE. To calculate this metric, it needs to divide the difference by the actual value. This means that if the actual values are close to or equal to 0, the MAPE score will either be divided by 0 or be extremely large. Consequently, it is not recommended to use MAPE when the true values are close to zero (Allwright, 2022)

A correlation is a relationship between two or more variables with values ranging from -1 to +1. It is usually calculated every month, with a minimum of one month. Correlation measures how quickly two stocks have historically moved relative to their average. It is tended to move in opposite directions and have a negative correlation are normally on opposite sides of the average. To have a little or no correlation there is no clear trend. Understanding correlation allows to diversify the portfolio by investing in uncorrelated underlying assets (Kiernan, n.d.).

The Data splitting a popular method for the model validation, in which it split a given data set into two disjoint sets: training and testing. Then statistical and machine learning models are fitted to the training set and validated using the test set. It can evaluate and compare the predictive performance of different models without worrying about possible overfitting of the training set by keeping the validation data set separate from the training. The most common approach to data partitioning is random subsampling, which involves random sampling without replacing some rows of the data set for testing and leaving the rest for training. The previous data splitting strategies can be executed once we indicate a part ratio. A common ratio 80:20, in which 80 percent of the data are used for training and 20 percent for testing. In practice, other ratios like 70:30, 60:40, and even 50:50 are also used. The optimal ratio for a given dataset does not appear to be clearly defined. The well-known Pareto rule serves as the basis for the 80:20 part, but experts still use it as a thumb rule. (Joseph, 2022)

Table 1: Related Systems that used different algorithm

Title	Description	Algorithm	Proposed System	Result
1. ANFIS for Prediction of epidemic peak and infected cases for Covid 19	Adaptive Neuro-fuzzy Inference System (ANFIS)-based machine learning technique to predict the possible outbreak in India.	ANFIS	forecast the future positivity and predict the cases of diseases	low Mean Square Error (MSE) of 1.184 × 10−3 with an accuracy of 86%.
2. Covid 19 Forecastin g using Multivaria te Linear Regression	propose a forecasting model using the COVID-19 available dataset from top affected regions across the world using machine learning algorithms.	MvLR	Prediction model to predict Infectious diseases in a region	Accuracy is 99.7%, Mean Absolute Error is 79818 and R2 score is 99.28%
3. Prediction Model for Coronaviru s Pandemic Using Deep Learning.	developed and provide a prediction model that would use Artificial Intelligence and Deep Learning to improve the diagnostic	Deep Learning	Determine the causes of positivity rate it uses Multi Linear Regression and PythonAf to visualized to result.	The accuracy, precision, and recall values were calculated for each part of the dataset,

	process by reducing unreliable diagnostic interpretation of chest CT scans and allowing clinicians to accurately discriminate between patients who are sick with COVID-19 or pneumonia, and also empowering health professionals to distinguish chest CT scans of healthy people			and results were compared with existing studies as well. The results show that the proposed model provides an efficient way of COVID-19 diagnostics from existing CT scans with better accuracy of 93%.
4. DIAGNOSIS OF CORONAVIRUS DISEASE 2019 (COVID-19) SURVEILLANCE USING C4.5 ALGORITHM	detection and response activities to identify conditions of PDP, ODP, OTG, or confirmed cases of COVID-19	C4.5 Algorithm	Decision Support System that predicts and forecast the positivity rates in a region	results in an accuracy of 0.9286 (92.86%).

The related system stated in Table 1 has the same objectives to control the spread of an infectious disease such as Covid 19 and uses other algorithms to predict the cases. These systems proposed are to measure, scan, tract, and predict the confirmed Covid 19 cases.

Table 2: Features of the Proposed System

Title	Description	Algorithm	Features	Comparisons Of Studies
Modeling and Prediction of Infectious Disease Cases in Iloilo City Using Multiple Linear Regression	Using of PythonAf a time series to forecast the future positivity rates with utilizing the Multiple linear regression to predict the	Multiple Linear Algorithm	Prediction Model Determining the accuracy of data Correlational Study Forecasting future rates	Predicts the possible outbreak and also tract the diseases in the study, it predicts the infectious diseases and forecast the future positivity rates. Predicts the mortality of number confirmed active cases a day. In our proposed study, our we predict the active positivity rate in certain time. Conducting a surveillance to

	cases of infectious diseases such as Covid 19			detect and testing artificial intellingence methods for diagnosing covid 19. In our study, prediction model the more data the more result to predict cases.

The Proposed system develop a prediction tool employed in linear regression to visualize the positivity rate in a region. Similar with the other related system mentioned in Table 1 but this proposed system used Two algorithms such as Multiple Linear Regression to predict the positivity rate of infectious disease like Covid 19.

CHAPTER 3 RESEARCH DESIGN AND METHODOLOGY

Description of the Proposed Study

PythonAF is an open-source Python application that uses time-series forecasting models automatically. To execute the forecasting task, a machine-learning technique is used. It takes place by training a time series model using previous data and then using this model to produce future values to forecast. Multiple linear regression (MLR) is a mathematical method for determining a mathematical relationship between various random variables. In other words, MLR examines how several independent variables are related to a single dependent variable. As a solution in the study, the PythonAF and MLR will assist us in predicting and analyzing data on future COVID-19 and other infectious disease positivity rates in Iloilo City. It helps establish and illustrate the relationship between the dependent variables. This prediction model shows the result of the positivity rate given data such as the number of individuals being tested, the number of tests performed each day, the location's status, the level of strictness of the regulations, the type of transmissions, and researchers would also like to consider the level of strictness of the regulations over a specific time period.

Assumptions and Preconditions

Multiple linear regression (MLR) is a statistical technique used to model the relationship between two MLR is a statistical technique for modeling the relationship between two or more independent variables and one dependent variable. PythonAF (Python Automation Framework) is a software development framework for task and process automation. These two concepts have distinct assumptions and prerequisites. The independent variable is the relationship of predicted positivity rates in the infectious diseases and the dependent variable is the Python Automatic Forecasting (PyAF) and Multiple Linear Regression (MLR) which in common have linear relationship with one another.

Multiple Linear Regression (MLR) assumptions include:

The independent variables and the dependent variable have a linear relationship. The model's errors are normally distributed and have a constant variance. There is no multicollinearity, which means the independent variables are not highly correlated. The residuals of the model are independent of one another.

Multiple Linear Regression (MLR) preconditions include:

The dependent variable should have a continuous value. The independent variables should be related to the dependent variable in a linear fashion. The sample size should be sufficient to yield meaningful results. Outliers and influential data points should be avoided in the data.

Assumption includes:

The automated system is a software system. The user is knowledgeable and experienced in programming. The user is well-versed in the Python programming language. The user is familiar with the automated system and its interfaces.

PythonAF prerequisites include:

The automated system should have appropriate software interfaces that can be accessed via Python. PythonAF requires the user to have access to the necessary software and hardware resources. The user should have a clear understanding of the automated tasks and processes. The user should be familiar with programming principles and practices.

Methods and Proposed Enhancements

The study proposed a predictive model to determine the outcome of the positive rates of the data collected at the COVID-19 Emergency Operations Center in Iloilo City. The researchers gathered the data, prepared it for processing, then used the Python AF and regression algorithm to forecast the outcome. Then, they presented the prediction model's results and visualization and made appropriate recommendations for future decisions based on factors contributing to the positivity rate. Using the information acquired at the Iloilo City COVID-19 Emergency Operations Center, the study developed a predictive model to determine the outcome of positive rates. Researchers gathered the information and then prepared to analyze the elements that influence the rate of positivity.

Components and Design

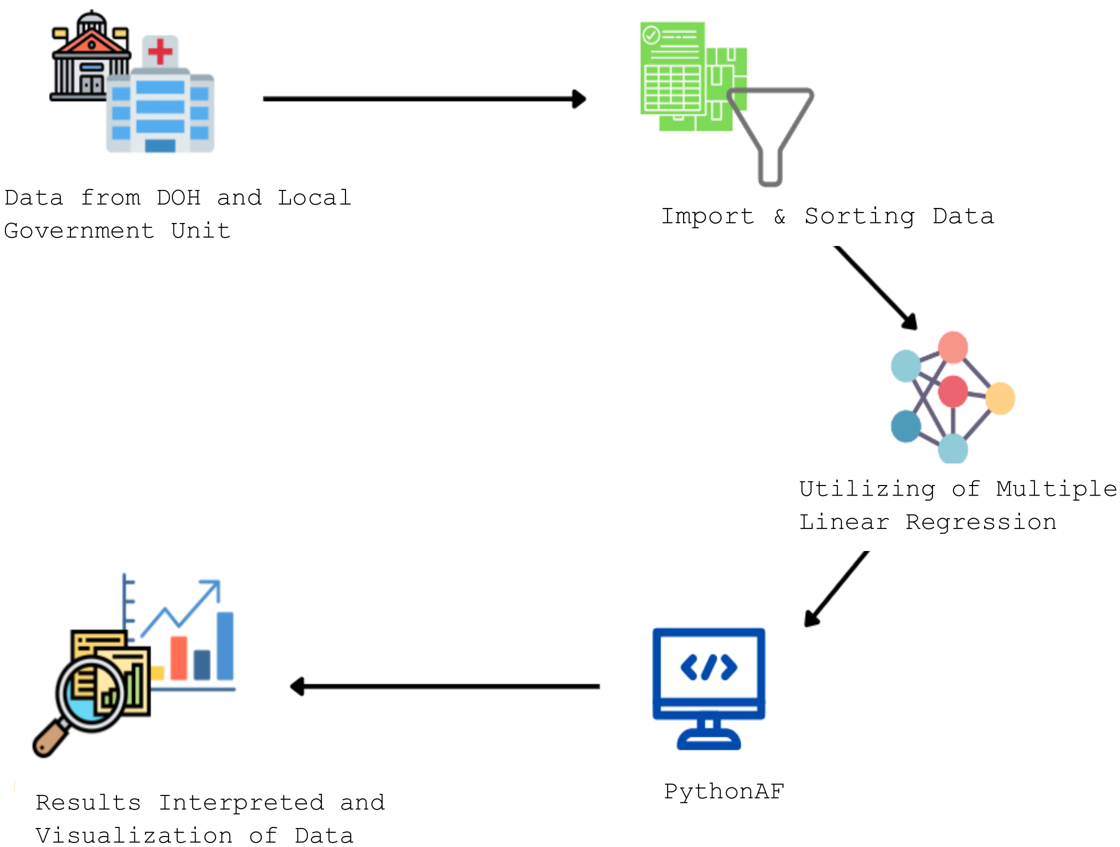


Figure 3.1 Software Architecture

The Figure 3.1 shows the software architecture of the proposed system that uses Pandas as a means for the visualization of data as well as the ease of providing information on how the data is being manipulated within the program. The researchers first use multiple linear regression to gather which factors or variables have a correlation to the dependent variable which is the positivity rate then use that data to provide further prediction with a newer algorithm which is Python Automatic Forecasting.

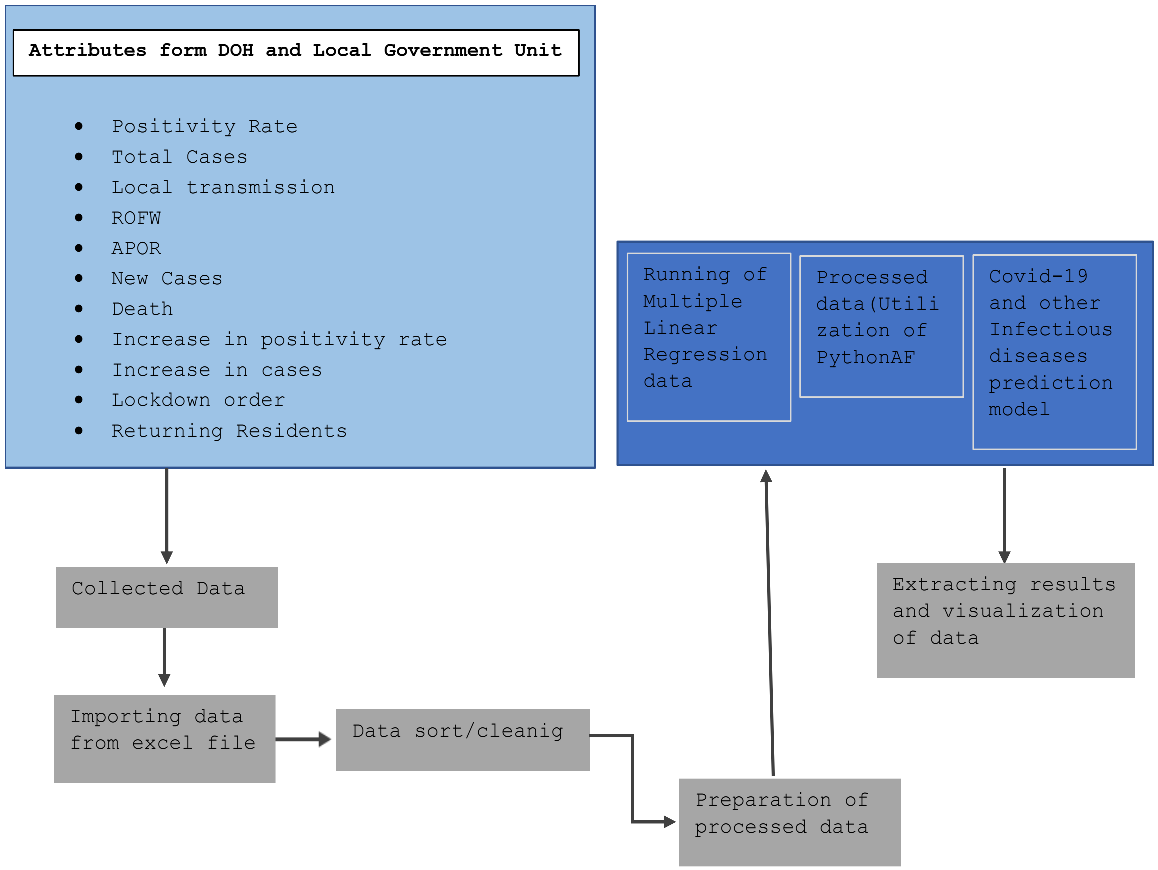


Figure 3.2 System Architecture

With the help of the data the researchers have gathered, this study suggests a prediction model for determining future positive rates. Figure 3.2 shows the process of the system architecture. After data has been collected and sorted, the users utilize the program to visualize any predictions and provide information on whether or not the used algorithm is accurate or has an acceptable margin of error or feasibility range in the study. The system was designed to only require an external excel file to be uploaded which can be then utilized by the program to provide visualizations.

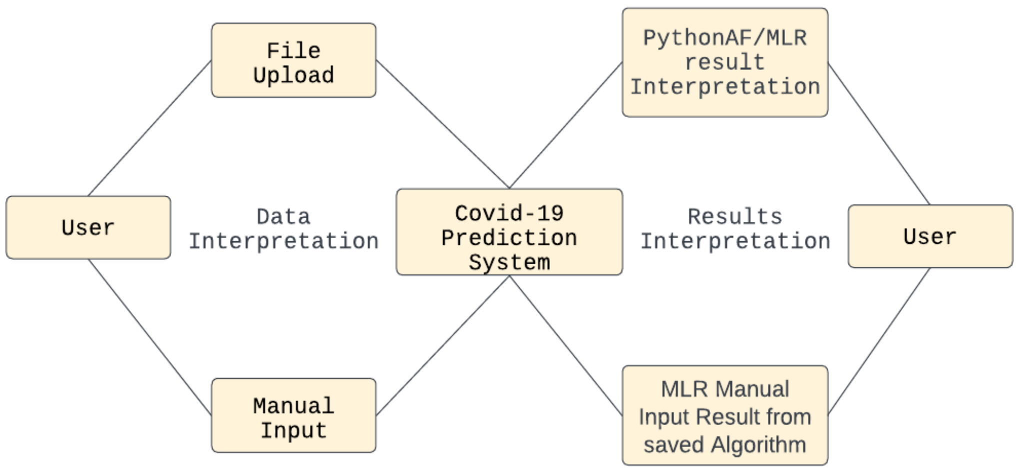


Figure 3.3 Database Design

Figure 3.3 shows the database consists of variables that are readily available to be extracted or provided by Iloilo City Covid-19 Emergency Operations Center. Over the duration of the study, researchers initially had factors such as age, sex, occupation (whether or not they worked in call centers), as well as returning OFWS and APORs. Eventually, said data was no longer being provided but the researchers still used this as a means to at least get a visualization on the earlier portions of the study. The database mostly consists of daily images that were provided by the government page that summarized weekly data.

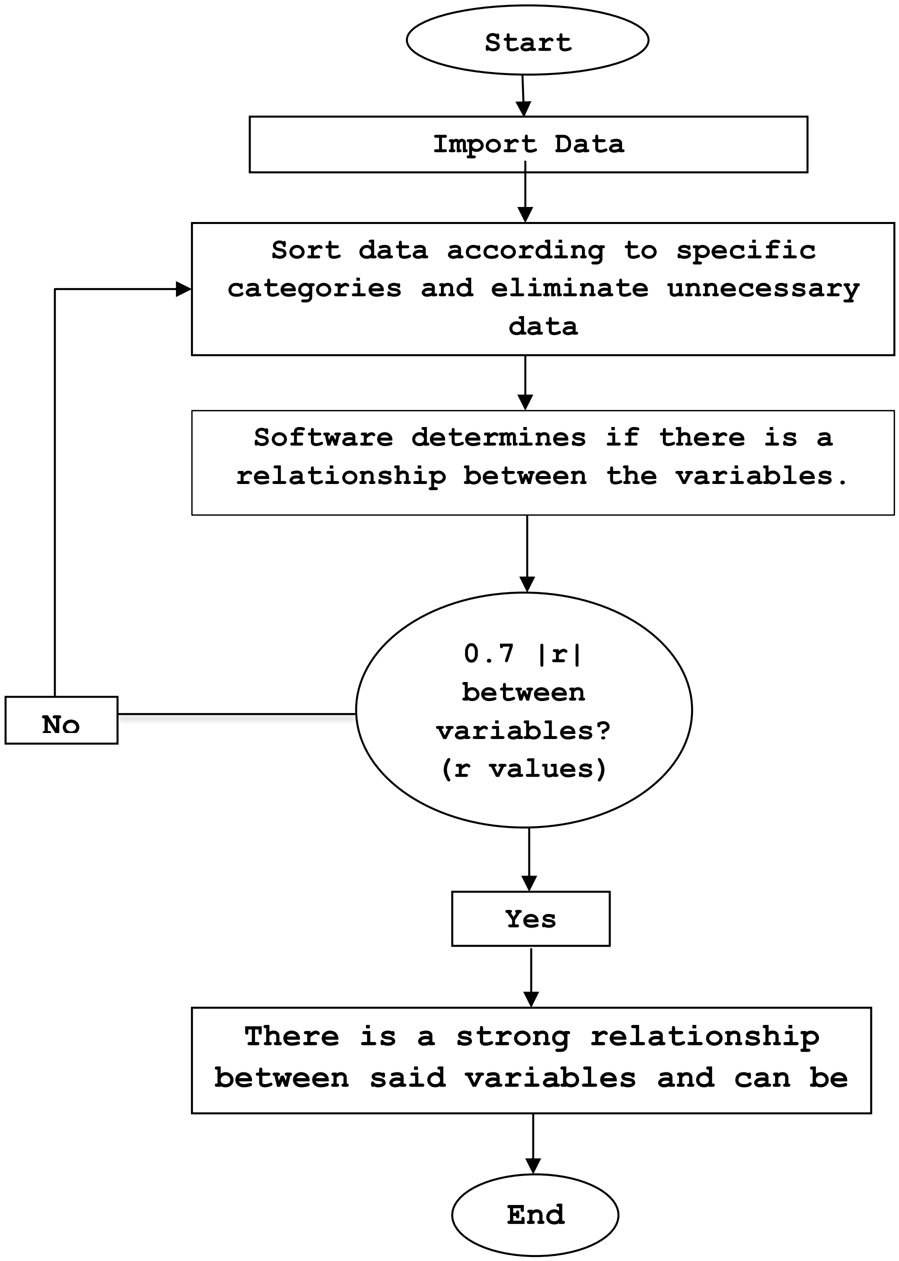


Figure. 3.4 Procedural object-oriented design

Figure 3.4 shows the model was created by combining intuition and judgment with a set of principles and algorithms, as well as an iterative process that results in final design specifications. A software system with insufficient design may fail to provide the intended service, necessitating time-consuming maintenance tasks. As a result, before beginning to implement the system, software developers must thoroughly complete the design phase.

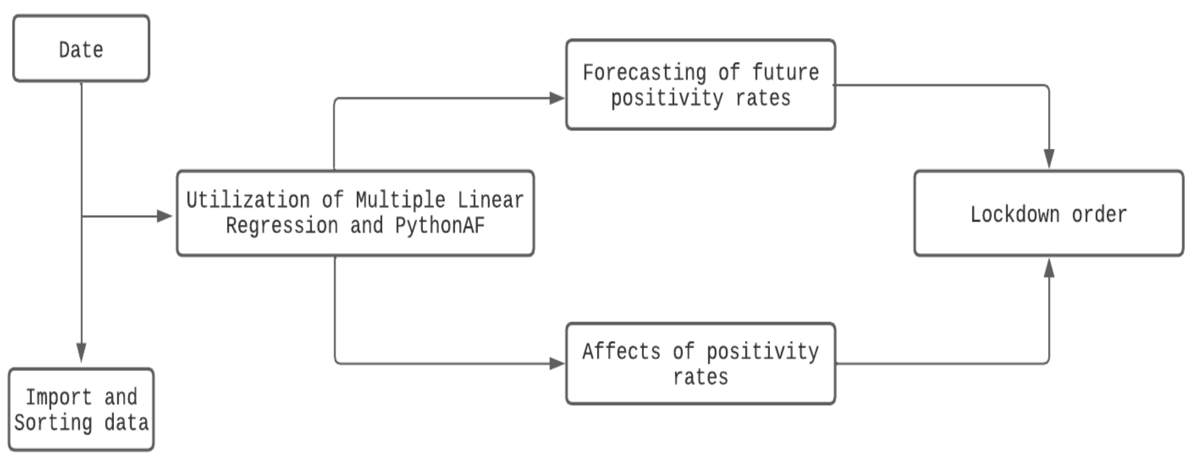


Figure 3.5 Process Design

Figure 3.5 shows the procedure describing the tasks that must be completed when using the system. These are readily available resources for both designers and users to use when performing supervisory tasks. This is a detail of all major system components' internal elements. Their characteristics, relationships, processes, and, in many cases, algorithms and data structures.

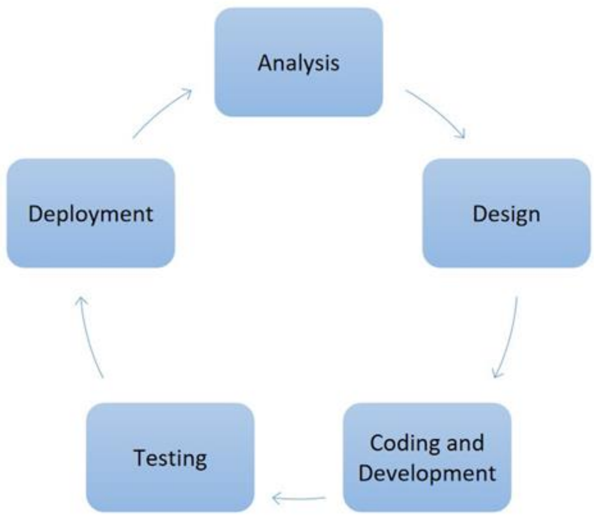


Figure 3.6 System Development Life Cycle

Figure 3.6 shows the SDLC procedure to produce a high-quality result, that describes the many phases of software development. It is a comprehensive strategy that outlines how to create, preserve, replace, alter, or improve a specific program.

Analysis. In this phase, all relevant data selected from the Iloilo City Covid-19 Emergency Operations Center to develop a system utilizing said algorithms. The following are data attributes.

Design. The data gathered which is compiled in a csv file is employed as a source of information for the software architecture that results in system development.

Coding and Development. The source code is created by translating the software design information system and it is used by python programming libraries that will be integrated into its environment and contrary installed. When this stage is done, the software is ready for testing and it provides visualization of data.

Testing. Upload the sample dataset to the model to be used for profiling and testing the data.

Deployment. The system evaluation using ISO standards.Depending on how the user interacts, it may be deployed in the production environment after testing or at the conclusion of acceptance testing.

CHAPTER 4 RESULTS AND DISCUSSION

Implementation

The study aimed to use newer algorithms such as PythonAF (Python Automatic Forecasting) to forecast future positivity rates, multiple linear regression to better understand which variables directly correlate with or actually impact positivity rates, and inferences based on the data collected and processed using the stated algorithms.The related system has the same goals of controlling the spread of an infectious disease like COVID-19 but uses different algorithms to predict cases. These proposed systems will evaluate and predict confirmed COVID-19 cases. They aim to use multiple linear regression to first address which aspects have a strong relationship with the independent variable (positivity rate); implement PythonAF, or Python automatic forecasting, to compare the prediction method used with multiple linear regression. Determine which points in time have a direct correlation with rising COVID-19 cases. Helps provide a basis from which the researchers can draw conclusions to help medical personnel or local governments make future decisions. This study developed a model for forecasting the outcome of positive interest rates. The researchers were to obtain data from the Department of Health or a local government unit, prepare it for processing, and then forecast the outcome using a regression technique. Based on the aspects that contribute to the positivity rate, the researchers then provide the prediction model's results and visualization, as well as make appropriate decisions for future decisions. This study developed a model for predicting the outcome of positive rates in DoH or local government data.It gives information for the prediction model to use in predicting positive rates in infectious diseases such as COVID-19 in Iloilo City. The data should be of excellent quality so that AI researchers can create data flows without having to spend time improving data quality. A system for managing data serves a similar purpose. This system model proved useful in forecasting future optimism rates, particularly within Iloilo City. It is accessible but requires an internet connection. It will be essential in the event of future infectious diseases. This algorithm can be used by researchers because it allows to test the accuracy and relationship of the variables and develop a prediction model that eventually only requires variables that should be plotted in the algorithm itself.This study intended to utilize algorithms with forecasting models such as Python Automatic Forecasting to produce results in line with future positivity rates as well as Multiple Linear Regression to better understand which variables directly correlated with one another or affected positivity rates, which would allow the researchers and users to make inferences based on the results given. The implementation of Streamlit as a means of our frontend was done due to the simplicity of implementing said algorithms towards a system that can be made available publicly, with the backend code being focused on the utilization of the algorithms. Both the frontend and backend of researchers system do not cache or hold any user data and mainly focus on the utility of either inputting values and utilizing the stored algorithm to make predictions or the ability of users to upload a file and allow the system to use the algorithms to make the prediction. With that in mind, the researchers did not see a need for a login system but welcomed the idea of a system where results are saved for easier viewing in the future, should the researchers choose to further improve the system.

System GUI

Figure 4.1 Welcome Window

Figure 4.1 shows the welcome screen, which contains a description of the system and the data used.

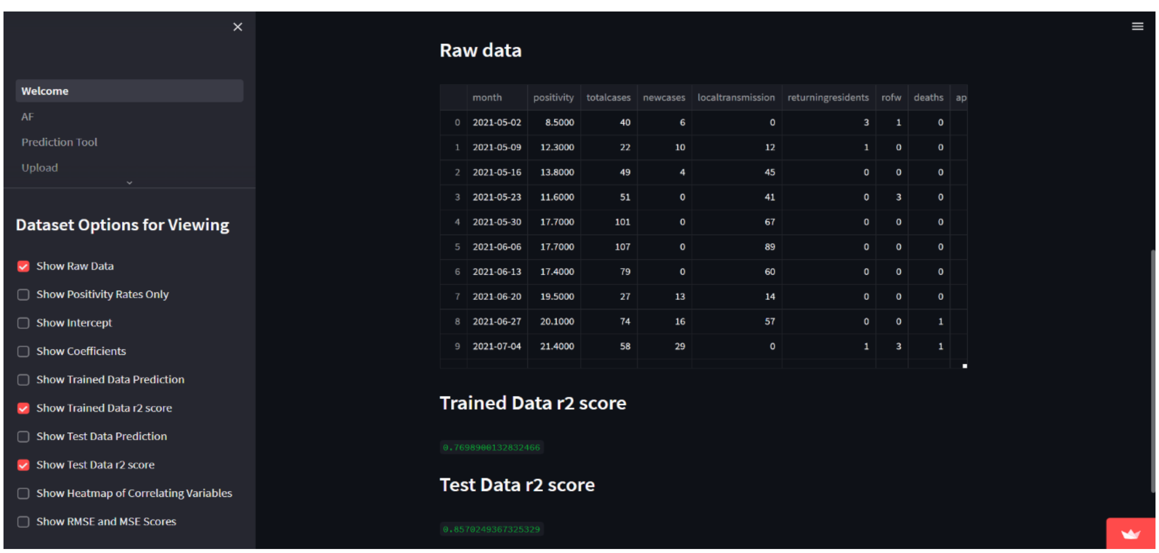


Figure 4.2 Dataset Options for Viewing

Figure 4.2 is intended for viewing data results, check the box of options in the left side of the window and results will show.

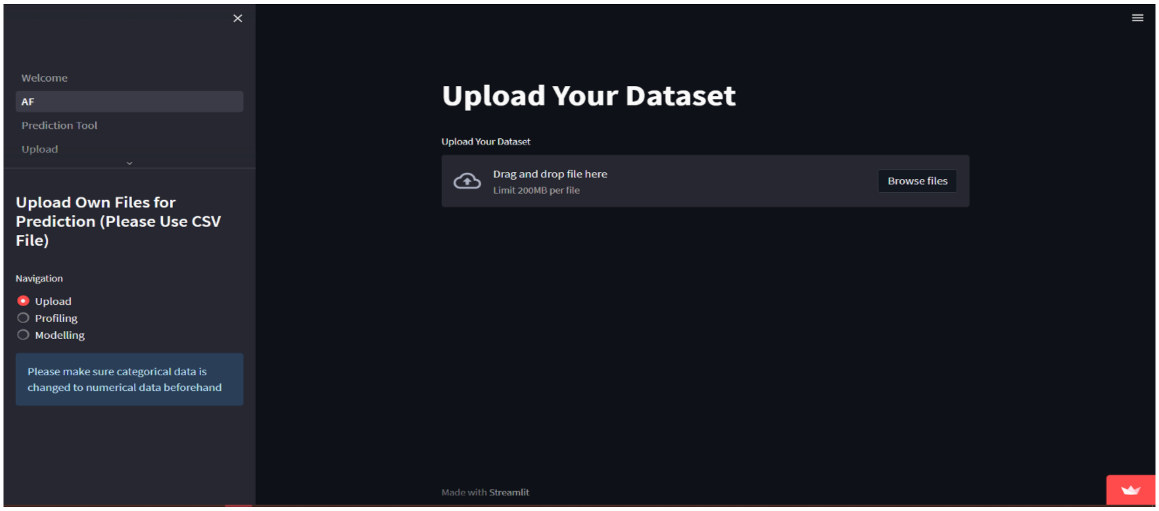


Figure 4.3 AF Upload Menu

Figure 4.3 shows the window where you can upload your own data. File to be used should be in CSV format.

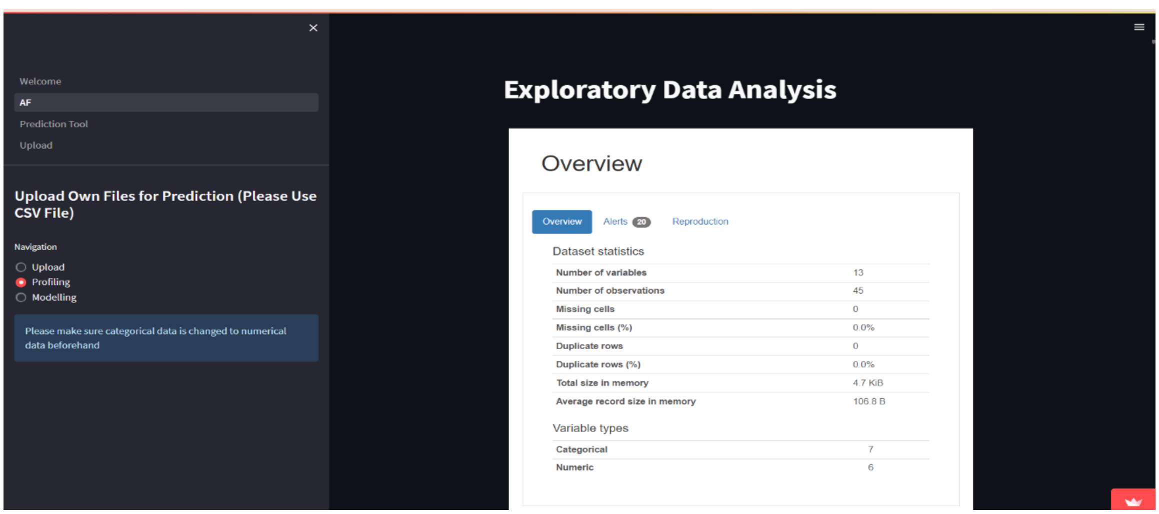


Figure 4.4 Profiling Menu

In profiling, the process of analyzing and summarizing a dataset in order to better understand its characteristics and relationships. It involves identifying patterns, trends, and relationships in the data and using visualizations, statistical tests, and other tools to help users better understand the data. It contains an overview of the data as what is shown on Figure 4.4.

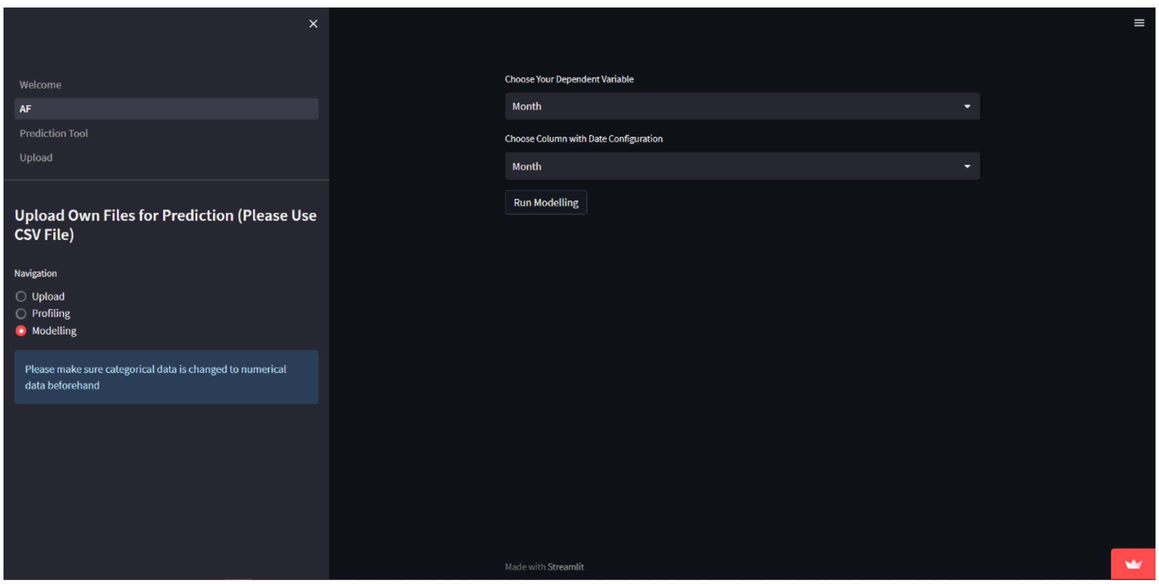


Figure 4.5 Modeling Menu

Figure 4.5 shows the modeling menu, wherein, once a data file is uploaded by the user, test the modeling by selecting dependent variable and date configuration and click the "Run Modeling" button.

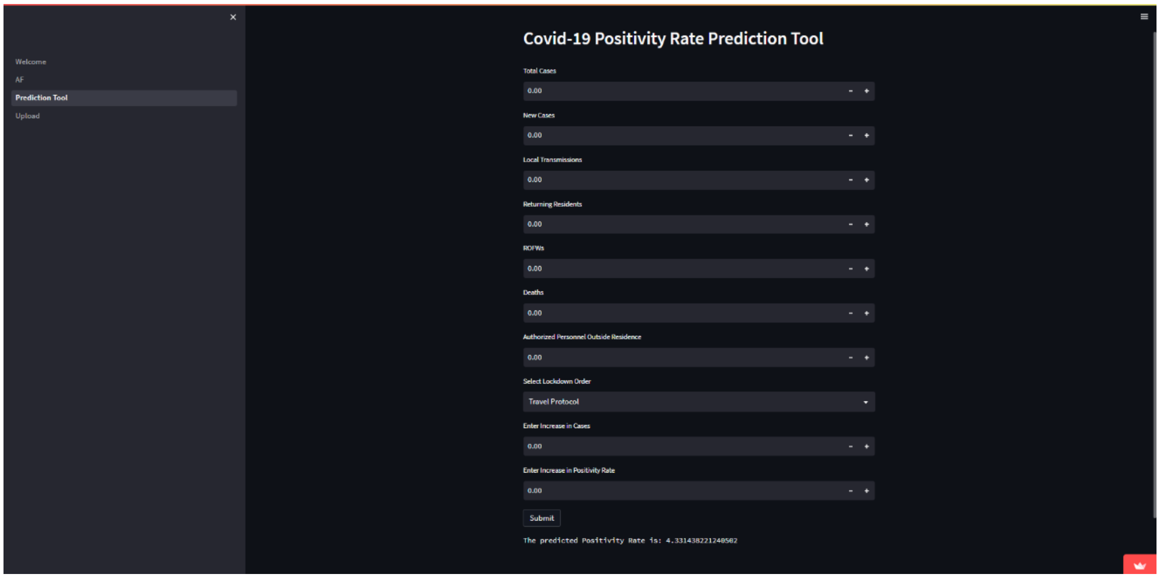


Figure 4.6 Prediction Tool Window

Figure 4.6 shows a window for the prediction tool. Prediction result shows when number of corresponding data is entered.

Results and Interpretation

Initially, after preprocessing the data and using Multiple Linear Regression, the study finds that there is a massive spike or variance in the data that would skew results as different precautions were being given at a certain period. From the graph in Figure 4.7, this becomes apparent during the month of November 2021 up until it eventually decreases in early May 2022. This is the portion of the data that the researchers wanted to show significance in and test if there was a decision to be made that would alter such a large change.

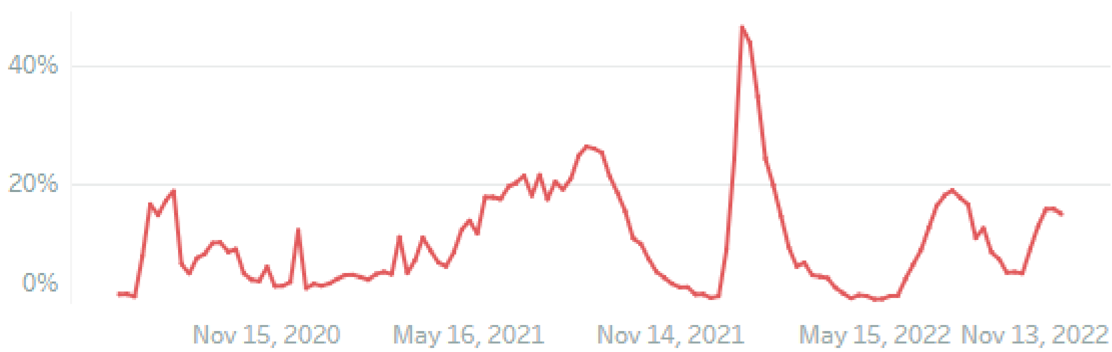


Figure 4.7 Weekly Positivity Rate

In Figure 4.7, with the use of Multiple Linear Regression, the researchers first processed the data to be used that can be interpreted by the algorithm. Afterward, the researchers split the data to train and test, wherein the researchers used 80% of the data to be trained, the remaining as the test raw data. Results of running Multiple Linear Regression with the training data are as follows:

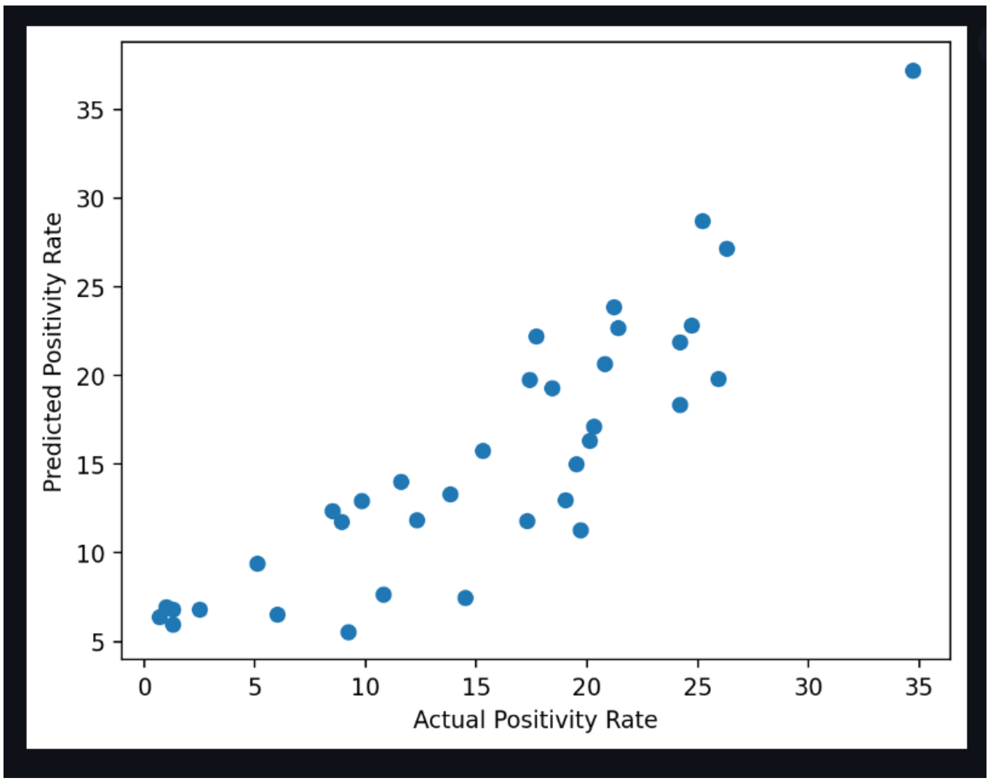


Figure 4.8 Train data

Figure 4.8 shows that the usage of Multiple Linear Regression in the study is still quite effective and the points are still close enough to be significant to each other as opposed to being far apart. The r2 score in the study also showed a value of 0.7698900132832466.

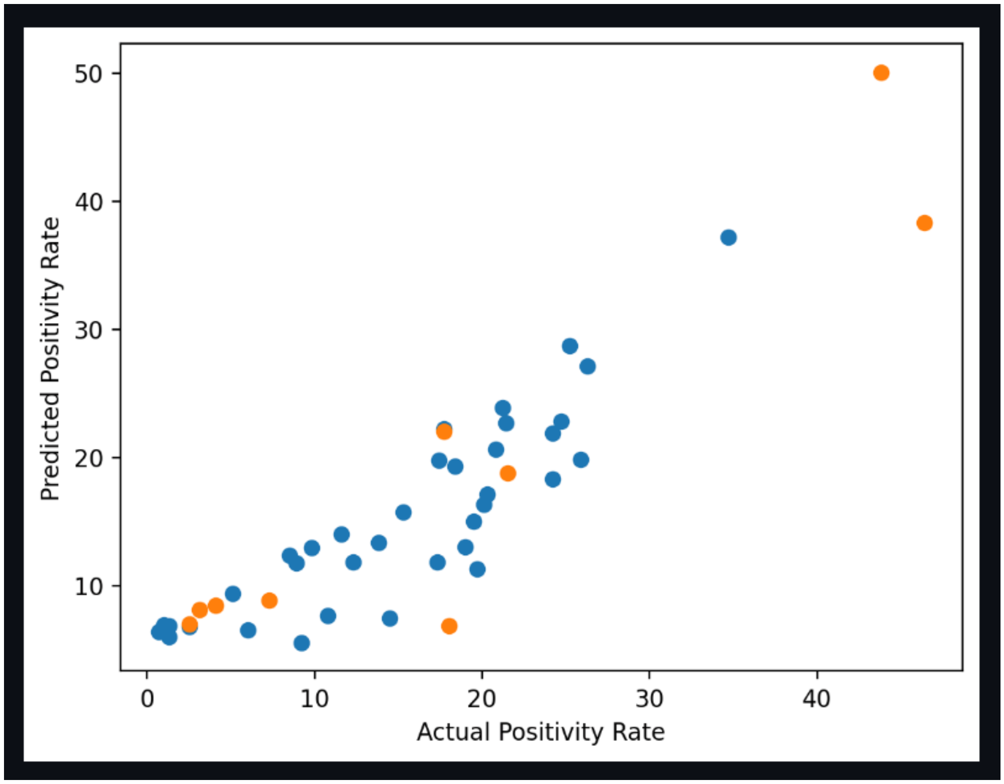


Figure 4.9 Test Data

Figure 4.9 shows the test data, although the researchers still have similarly close points near lower values of the Positivity Rate, the researchers have points that are far from each other as this might have been caused by the extreme variance in data that the researchers saw from the previous figure indicating a spike in cases. The r2 for testing data is 0.8570249367325329. Lastly, in Figure 4.10 is a generated heatmap that contains features correlating with the Positivity rate can give us a brief view of which variables are directly contributing to the positivity rate here in Iloilo City.

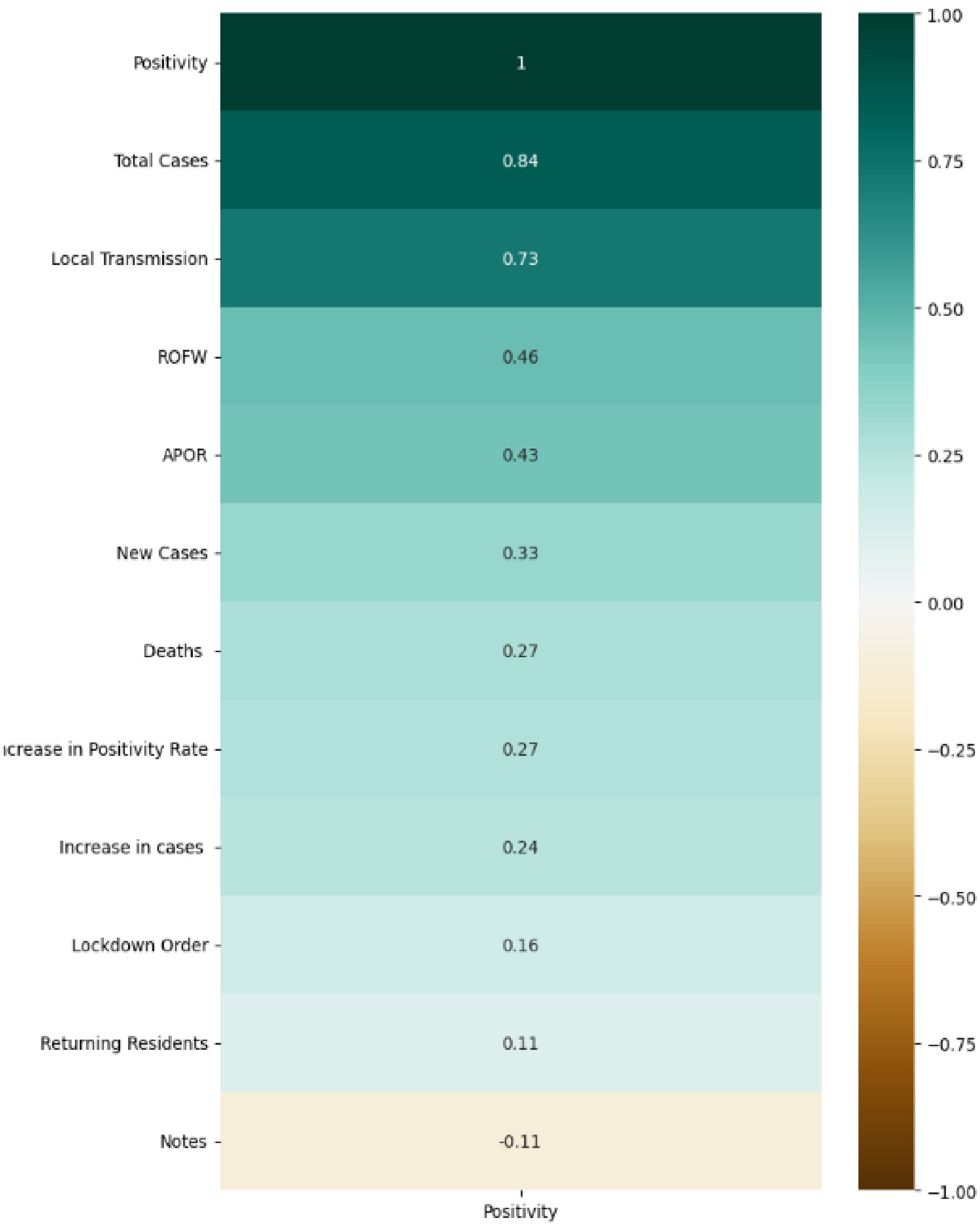


Fig. 4.10 Feature correlating with Positivity Rate

Figure 4.10 at first glance shows that the researchers can see the Local Transmissions are very close to the positivity rate but should also take "Lockdown Order" into consideration as well. This variable is a placeholder for general guidelines or restrictions being given in an area and the reason being that it is believed it is inversely related to the Positivity rate is that changes to this variable do not immediately affect the Positivity Rate but may eventually reflect in the week that comes later. This is to acknowledge the fact that Covid-19 may linger in an individual's system for up to 1-2 weeks and a change in guidelines that restrict local movement may not directly affect these people and may not reflect in the data instantly.

With the answer of which variables should be given importance to with Multiple Linear Regression, the results of using Python Automatic Forecasting as a method to forecast positivity rate are as follows:

Table 3: Positivity Rate results using Python AF

	Month	Positivity	Positivity_Forecast
60	2021-08-29	26.3	25.071735
61	2021-09-05	25.9	27.168256
62	2021-09-12	25.2	26.834857
63	2021-09-19	21.2	25.523572
64	2021-09-26	18.4	20.936300
65	2021-10-03	15.3	17.542963
66	2021-10-10	10.8	13.994918
67	2021-10-17	9.8	9.614961
68	2021-10-24	7.3	7.631216
69	2021-10-31	5.1	5.172719
70	2021-11-07	4.1	2.580440
71	2021-11-14	3.1	1.537333
72	2021-11-21	2.5	1.076102
73	2021-11-28	2.5	1.143891
74	2021-12-05	1.3	1.692638
75	2021-12-12	1.3	1.076536
76	2021-12-19	0.7	1.025120
77	2021-12-26	1.0	0.572461
78	2022-01-02	NaN	1.008557
79	2022-01-09	NaN	1.141886
80	2022-01-16	NaN	1.424335
81	2022-01-23	NaN	1.667756
82	2022-01-30	NaN	1.824182
83	2022-02-06	NaN	2.026915
84	2022-02-13	NaN	2.251130
85	2022-02-20	NaN	2.503398
86	2022-02-27	NaN	2.814940
87	2022-03-06	NaN	3.066776
88	2022-03-13	NaN	3.345265
89	2022-03-20	NaN	3.563282

Table 3 simulates that data was only gathered before the actual spike that happened in November to better see if the algorithm was able to still predict a spike with the given data alone.

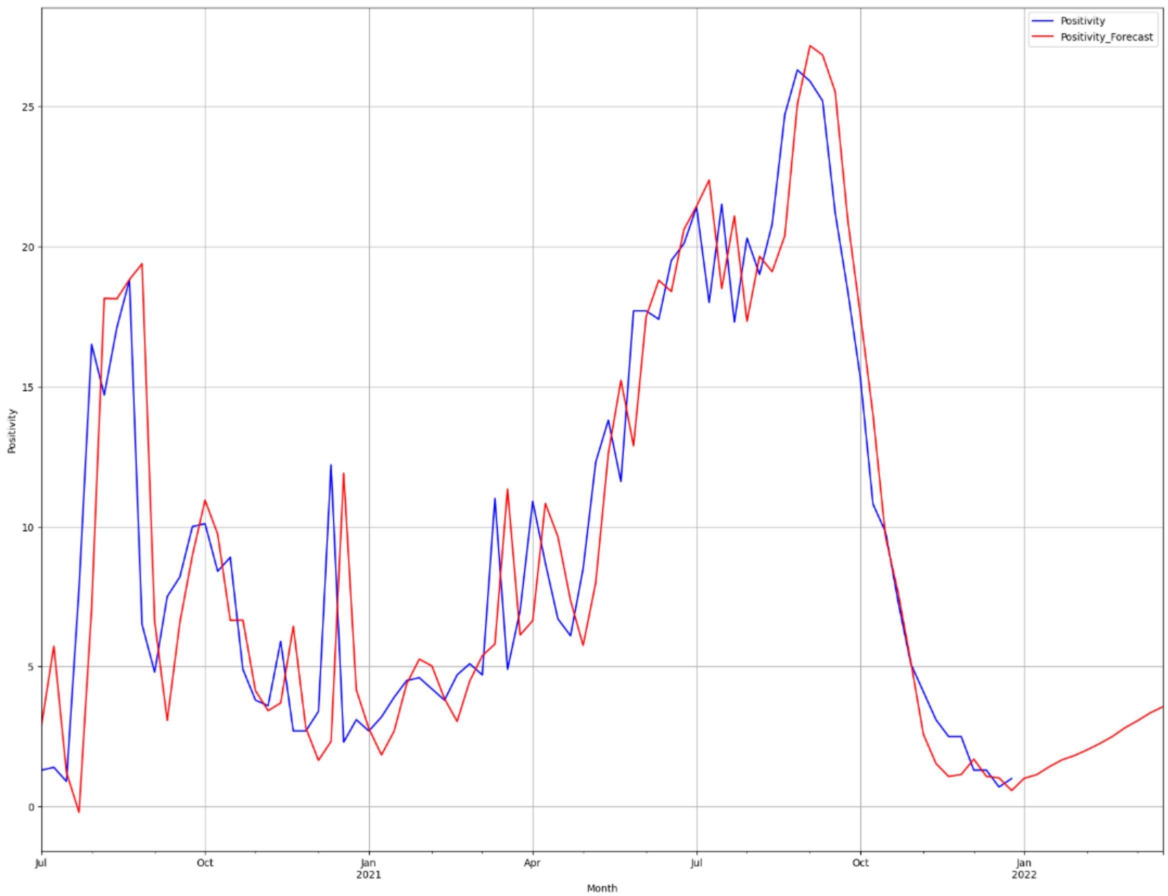


Figure 4.11 Predicted Results

From the graph in Figure 4.12, study finds that the algorithm was still able to predict a spike in cases although not as aggressively as the actual spike that occurred during the time.

Multiple Linear Regression data has 45 weeks of data, this is due to the fact that Iloilo City Covid 19 Emergency Operations Center stopped logging many of the data we wanted such as APORs, OFWs, and returning residents.

Researchers use 80/20 split since it is a commonly used ratio for training/test data and we may opt for 70/30 since some studies also use this but researchers want to be as aggressively accurate with what portion of our data we use

MAPE_Fit=0.1733 MAPE_Forecast=0.1733 MAPE_Test=0.1733
Mean Absolute Percentage Error - 17% (Considered good)

The mean of all absolute % errors between the expected and actual values is known as MAPE (Mean Absolute % Error). It is a well-liked statistic since it shows error percentages, which makes it easy for end users to compare model accuracy across datasets and use cases.

Good MAPE score is depending on the use case, a value that is good for time series forecasting is one with a MAPE lower than 20%. This indicate that, on average, the forecasts were less than 20% off the actual values throughout the entire time period.

According to Table 4 which is the MAPE graph shown above this result is considered good as the MAPE value of less than 20% is generally regarded as a good value. The MAPE metric is a percentage of errors whose value represents the average amount of error in predictions. As a result, a lower MAPE is preferable, as the model's accuracy increases with the value.

MAPE	**Interpretation**
<10%	Very good
10% - 20%	Good
20% - 50%	OK
>50%	Not Good

Table 4: MAPE Score

Systems Evaluation and Results

In the initial round, 20 randomly selected respondents—10 users, 5 IT specialists, and 5 medical professionals—were contacted as part of a random strategy to assess their opinions on the overall ratings and effectiveness of the proposed solutions. The ISO 25010, or Software Quality Product, served as the foundation for the questionnaire. The questionnaire utilized a rating scale; With a rating of '5' being excellent, '4' being very good, '3' being satisfactory, '2' being fair, and '1' being poor.

Table 5: The various numerical numbers and their equivalent ratings or interpretations

5 – EXCELLENT
4 - VERY GOOD
3 – SATISFACTORY
2 – FAIR
1 - POOR

Table 6: The baseline scale for the calculated means

Rating	Mean Range	Descriptive Rating
5	4.21-5.00	Excellent
4	3.31-4.20	Very Good
3	2.61-3.30	Satisfactory
2	1.81-2.60	Fair
1	1.00-1.80	Poor

In the Table 6, the researchers conducted a basic mean evaluation of the responses received on various parameters of the ISO-based questionnaire in the sections that follow. The Scale scores and accompanying mean ranges are shown in the table, which will often be used as a baseline for the means obtained from various ISO indicators.

After the tabulation for the averages on different indicators, it was indicated that in fact, the majority of the indicators fall under the 'Very Good' rating as shown in Table 6. Among the weighted means, the usability indicator has the highest rating, while the security aspect has the lowest – the aspect that the proposed system needs room for improvement.

Table 7: Mean scores for various metrics and their corresponding ratings were tabulated

Standard Basis	Weighted Mean	Descriptive Rating
Usability Functionality	3.74	Very Good
Performance Efficiency	3.68	Very Good
Compatibility	3.72	Very Good
Usability	3.79	Very Good
Reliability	3.71	Very Good
Security	3.30	Satisfactory
Maintainability	3.68	Very Good
Portability	3.73	Very Good
Overall Mean	3.26	Satisfactory

The accuracy and efficiency of the system were evaluated using the amount of data that was available. The advantage of this system lies in its ability to forecast and visualize uploaded data, in contrast to other systems that only focused on prediction. Combining these abilities results in a complete structure that not only forecasts future results but also delivers visual representations for easier comprehension. Users can obtain insights, spot trends, find anomalies, and make wise decisions with the aid of the generated visuals. By providing both precise predictions and cutting-edge data visualization capabilities, this characterizes the system.

CHAPTER 5 SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

Summary of Proposed Study and Research Design

The main objective of the study is to make a prediction model based on the positivity rates around Iloilo city using PythonAF and Multiple Linear Regression. This study aims to see the significance relation of dependent variables using MLR and comparing the result to the prediction of PythonAF. The PythonAF and MLR will assist in predicting and analyzing data on future COVID-19 and other infectious disease positivity rates around Iloilo City.

Summary of Findings

The proposed system was developed using data collection and manual coding. The system was provided to the pertinent evaluators (Information Technology experts, Users, and Medical Professionals) after it had been partially created so they could evaluate the most recent version of the created decision support system. The 'Usability' features of the system received the greatest mean scores from the evaluators while ‘Security’ received low mean scores among the indicators, and these mean scores were tallied by the researchers. The develop system is an open site that can be manipulated by the users. In order to use it, users must upload data to show and predict the results of test data.

Study of Design and Implementations

This study uses Steamlit for its visualization of the gathered data. It also provides information on how the data is being processed. A prediction model is created to determine the future positivity rates of the gathered data from the group. This study consists of different variables readily extracted from the Iloilo City Covid-19 Emergency Operator Center. This study uses algorithms such as

PythonAF and Multiple Linear Regression. This algorithm intends to understand the correlation of the variables and the impact positivity rates. By this the authorities can make an inference based on the data collected using the stated algorithms. As in the result this system model is effective and useful in predicting the positivity rates particularly here in Iloilo City.

Conclusion

The researchers came to the conclusion that the proposed system had achieved the set of goals that were set forth in the study's initial phase.

1. There is a significant relationship with the dependent variable (Positivity Rate) along with the independent variables that was determined using Multiple Linear Regression.

2. The PynthonAF and Multiple Linear Regression algorithms were able to predict a spike or variance in the data that would skew the results as different precautions were being given at a certain period.

3. The algorithms were able to analyze the time which has a direct correlation with rising cases of Covid-19.

4. The results were able to give a precaution to aid medical personnel or the local government on what future decisions to make.

Recommendations

The primary goal of this study was to develop a system that is able to utilize practical algorithms that can interpret data regarding Covid-19 and make inferences to assist local government units better and health workers make timely decisions to lessen the spread or how to better address Covid-19 in general. In light of the findings, the researchers discovered during the study, the researchers summarize the recommendations through the following:

1. Although Multiple Linear Regression is a fairly old algorithm as mentioned by the panelists, it was still pivotal in how the researchers were able to discover which datasets the researchers could use in tandem with the newer algorithm which is Python Automatic Forecasting. With this, the findings also suggest utilizing newer algorithms to forecast available data regarding Covid-19 or infectious diseases and test their efficiency.
2. Although the algorithm with Python Automatic Forecasting was considered acceptable through the range of values, the researchers had hoped more exogenous variables extracted from Multiple Linear Regression were available to use, with this in mind, the researchers suggest that future

researchers of this study are either able to provide more data regarding Covid-19 or Local Government Units to make data readily available to the public to better improve the prediction methods of algorithms.

3. In regards to the system that was developed, the researchers recommend that more features for ease of use should be developed and considered. This is to not only focus on the system being available only towards users that are interested in visualization and scientific data but also towards users that may not know how to interpret the data that is displayed and will generally be used by the public.
4. The researchers also acknowledge that the system does not automate how the data being processed and collected is gathered and therefore suggest that should the system be improved, the researchers should focus on methods to allow live gathering of data from data sources as well as processing the data for the algorithm in real-time.
5. The researchers strongly suggest that future studies or other studies to find means to provide such data that would impact or affect the findings of this study which is the positivity rate.

6. As the researcher's data from Multiple Linear Regression suggests, although the researchers were able to extract the total cases and new cases that were directly correlated to the positivity rate and that was indirectly correlated, that being the lockdown order, the researchers are advised to find more methods to prove this correlation either through the use of alternative algorithms or similar algorithms that serve the same purpose.

References

Alali, Yasminah, et al. "A Proficient Approach to Forecast COVID-19 Spread via Optimized Dynamic Machine Learning Models." Scientific Reports, vol. 12, no. 1, Springer Science and Business Media LLC, Feb. 2022. Crossref, https://doi.org/10.1038/s41598-022-06218-3.

Alenezi, Mohammed N., et al. "Building a Sensible SIR Estimation Model for COVID-19 Outspread in Kuwait." Alexandria Engineering Journal, vol. 60, no. 3, Elsevier BV, June 2021, pp. 3161–75. Crossref, https://doi.org/10.1016/j.aej.2021.01.025.

Allwright, S. (2022, December 6). *What is a good MAPE score? (simply explained)*. Stephen Allwright. https://stephenallwright.com/good-mape-score/

Almalki, A., Gokaraju, B., Acquaah, Y., & Turlapaty, A. C. (2022). Regression analysis for COVID-19 infections and deaths based on food access and health issues. *Healthcare*, *10*(2), 324. https://doi.org/10.3390/healthcare10020324

Antoinecarme. (n.d.). *GitHub - antoinecarme/pyaf: PyAF is an Open Source Python library for Automatic Time Series Forecasting built on top of popular pydata modules.* GitHub. https://github.com/antoinecarme/pyaf

Bi, Y., Hu, W., Liu, H., Xiao, Y., Lee, H., Chen, S., Zhao, L., & Tong, S. (2012). Can slide positivity rates predict malaria transmission? *Malaria Journal*, *11*(1). https://doi.org/10.1186/1475-2875-11-117

Binti Hamzah FA, Lau C, Nazri H, Ligot DV, Lee G, Tan CL, et al. CoronaTracker: Worldwide COVID-19 Outbreak Data Analysis and Prediction. [Preprint]. Bull World Health Organ.pub: 19 March 2020. doi: http://dx.doi.org/10.2471/BLT.20.255695

Chae, S., Kwon, S., & Lee, D. (2018). Predicting infectious disease using deep learning and big data. *International Journal of Environmental Research and Public Health*, *15*(8), 1596. https://doi.org/10.3390/ijerph15081596

Cîtu, C., Gorun, F., Motoc, A. G. M., Sas, I., Gorun, O. M., Burlea, B., Sas, I. T., Tomescu, L., Neamtu, R., Malița, D., & Citu, I. M. (2022). The predictive role of NLR, D-NLR, MLR, and SIRI in COVID-19 mortality. *Diagnostics*, *12*(1), 122. https://doi.org/10.3390/diagnostics12010122

Coronavirus disease (COVID-19) pandemic. (2024, April 22). https://www.who.int/europe/emergencies/situations/covid-19

Dallal, A. A., AlDallal, U., & Dallal, J. A. (2021). Positivity rate: an indicator for the spread of COVID-19. *Current Medical Research and Opinion*, *37*(12), 2067–2076. https://doi.org/10.1080/03007995.2021.1980868

Dash, Satyabrata, et al. "COVID-19 Outbreak in Orissa: MLR and H-SVR-based Modelling and Forecasting." International Journal of Computer Applications in Technology, vol. 66, no. 3/4, Inderscience Publishers, 2021, p. 401. Crossref, https://doi.org/10.1504/ijcat.2021.120468.

Dhir, H. S. . V. (2019). *Effectiveness of Python Libraries in Machine Learning: A* *review*. https://webology.revistaclinicapsicologica.com/abstract.php?id=4911

Hamzah, F. B., Lau, C., Nazri, H., Ligot, D. V., Lee, G., Tan, C. L., ... & Chung, M. H. (2020). CoronaTracker: worldwide COVID-19 outbreak data analysis and prediction. Bull World Health Organ, 1(32), 1-32.

Hirose, H., & Wang, L. (2012). *Prediction of infectious disease spread using Twitter: A case of influenza*. https://www.semanticscholar.org/paper/Prediction-of-Infectious-Disease-Spread-Using-A-of-Hirose-Wang/704287c6be85c3a2e8f4b5476ff86a7ca90fc70e

Humayun, Mamoona & Alsayat, Ahmed. (2021). Prediction Model for Coronavirus Pandemic Using Deep Learning. Computer Systems Science and Engineering. 40. 947-961. 10.32604/CSSE.2022.019288.

Infectious diseases - Symptoms & causes - Mayo Clinic. (2022, February 18). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/infectious-diseases/symptoms-causes/syc-20351173.

Jajosky, R. A., & Groseclose, S. L. (2004). Evaluation of reporting timeliness of public health surveillance systems for infectious diseases. *BMC Public Health*, *4*(1). https://doi.org/10.1186/1471-2458-4-29

Jm, A. M., Je, S., De Castro, R., Caoili, S. E. C., & Boñgolan, V. P. (2020). Protection after Quarantine: Insights from a Q-SEIR Model with Nonlinear Incidence Rates Applied to COVID-19. *medRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2020.06.06.20124388

Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining*, *15*(4), 531–538. https://doi.org/10.1002/sam.11583

Kiernan, D. (n.d.). *Chapter 7: Correlation and Simple Linear Regression | Natural Resources Biometrics*. https://courses.lumenlearning.com/suny-natural-resources-biometrics/chapter/chapter-7-correlation-and-simple-linear-regression/

Kumar, R., Al-Turjman, F., Srinivas, L.N.B. et al. ANFIS for prediction of epidemic peak and infected cases for COVID-19 in India. Neural Comput & Applic (2021). https://doi.org/10.1007/s00521-021-06412-w.

Kumari, Rajani, et al. "Analysis and Predictions of Spread, Recovery, and Death Caused by COVID-19 in India." Big Data Mining and Analytics, vol. 4, no. 2, Tsinghua UP, June 2021, pp. 65–75. Crossref, https://doi.org/10.26599/bdma.2020.9020013.

Lawton, G., Carew, J. M., & Burns, E. (2022, January 21). *predictive modeling*. Enterprise AI. https://www.techtarget.com/searchenterpriseai/definition/predictive-modeling

Mohan, S., Solanki, A. K., Taluja, H. K., Anuradha, & Singh, A. (2022). Predicting the impact of the third wave of COVID-19 in India using hybrid statistical machine learning models: A time series forecasting and sentiment analysis approach. *Computers in Biology and Medicine*, *144*, 105354. https://doi.org/10.1016/j.compbiomed.2022.105354

Osborne, J. W. (2019). Prediction in multiple regression. *Practical Assessment, Research, and Evaluation*, 7(1), 2.

Paul, M., & Held, L. (2011). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Statistics in Medicine*, *30*(10), 1118–1136. https://doi.org/10.1002/sim.4177

Rath, S., Tripathy, A., & Tripathy, A. (2020). Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. Diabetes & Metabolic Syndrome, 14(5), 1467–1474. https://doi.org/10.1016/j.dsx.2020.07.045

Suganya, R., R.Arunadevi, & Buhari, S. M. (2020). COVID-19 Forecasting using Multivariate Linear Regression. *Research Square (Research Square)*. https://doi.org/10.21203/rs.3.rs-71963/v1

Sah, Sweeti, et al. "Forecasting COVID-19 Pandemic Using Prophet, ARIMA, and Hybrid Stacked LSTM-GRU Models in India." Computational and Mathematical Methods in Medicine, edited by Andrei Korobeinikov, vol. 2022, Hindawi Limited, May 2022, pp. 1–19. Crossref, https://doi.org/10.1155/2022/1556025.

Salgotra, Rohit, et al. "Time Series Analysis and Forecast of the COVID-19 Pandemic in India Using Genetic Programming." Chaos, Solitons & Fractals, vol. 138, Elsevier BV, Sept. 2020, p. 109945. Crossref, https://doi.org/10.1016/j.chaos.2020.109945.

Talabis, D. a. S. J., Babierra, A. L., Buhat, C. a. H., Lutero, D. S., Quindala, K. M., & Rabajante, J. F. (2021b). Local government responses for COVID-19 management in the Philippines. BMC Public Health, 21(1). https://doi.org/10.1186/s12889-021-11746-0

Tulshyan, V., Sharma, D., & Mittal, M. (2022). An Eye on the Future of COVID-19: Prediction of Likely Positive Cases and Fatality in India over a 30-Day Horizon Using the

Prophet Model. Disaster Medicine and Public Health Preparedness, 16(3), 980-986. doi:10.1017/dmp.2020.444

"What Is Positive and Negative Correlation?" What Is Positive & Negative Correlation? | Tastytrade, 9 Nov. 2022, www.tastytrade.com/concepts-strategies/correlation.

Wiguna, Wildan, and Dwiza Riana. "DIAGNOSIS OF CORONAVIRUS DISEASE 2019 (COVID-19) SURVEILLANCE USING C4.5 ALGORITHM." Jurnal Pilar Nusa Mandiri, vol. 16, no. 1, PPPM STMIK Nusa Mandiri, Mar. 2020, pp. 71–80. Crossref, https://doi.org/10.33480/pilar.v16i1.1293.

Yadav, S. K., & Akhter, Y. (2021). Statistical modeling for the prediction of infectious disease dissemination with special reference to COVID-19 spread. Frontiers in public health, 680.

Zhang, H., Huang, J., He, X., Lv, P., & Da-Long, Q. (2007). An analysis of the current status of hospital emergency preparedness for infectious disease outbreaks in Beijing, China. *American Journal of Infection Control*, *35*(1), 62–67. https://doi.org/10.1016/j.ajic.2006.03.014

Zhao, D., Zhang, R., Zhang, H., & He, S. (2022). Prediction of global omicron pandemic using ARIMA, MLR, and Prophet models. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-23154-4

Zhou, T. (2022). A linear regression prediction model of infectious disease spread based on Baidu migration and effective distance. *Computational and Mathematical Methods in Medicine*, *2022*, 1–7. https://doi.org/10.1155/2022/9554057

Appendices

Appendix A

Letter to the Adviser

January 21, 2022

Dr. Ma. Luche P. Sabayle
Department Head , CS, CHED & ISO Liaison Officer, Assistant Professor III
College of Information Communication and Technology
Luna St. La Paz, Iloilo City

Dear Ma'am,

The undersigned are BS in Computer Science Research 1/Thesis 1 students of CICT, this university. Our thesis/capstone project title is "Modeling and Prediction of Infectious Disease Cases in Iloilo City using Multiple Linear Regression".

Knowing of your expertise in research and on the subject matter, we would like to request you to be our ADVISER.

We are positively hoping for your acceptance. Kindly check the corresponding box and affix your signature in the space provided. Thank you very much.

Respectfully yours,

Arle Kai Franco Gorriceta
Hiro Parcon
Joan Marie Arsenio
April Joy Gallano
Regina Flor Tonogbanua

Note by:

Dr. Ma. Luche P. Sabayle
Research Adviser

Appendix B

Letter to the Editor

Respectfully endorsed to the Technical Editor, the attached manuscript of the thesis entitled:

MODELING AND PREDICTION OF INFECTIOUS DISEASE CASES IN ILOILO CITY USING MULTIPLE LINEAR REGRESSION

Said manuscript has been presented to me for preliminary evaluation and guidance, and after a series of corrections/directions given which was implemented by the proponents whose names are listed hereunder and their thorough research, we have come to its completion.

Now therefore, I hereby ENDORSE the said thesis manuscript to the Technical Editor for TECHNICAL EDITING.

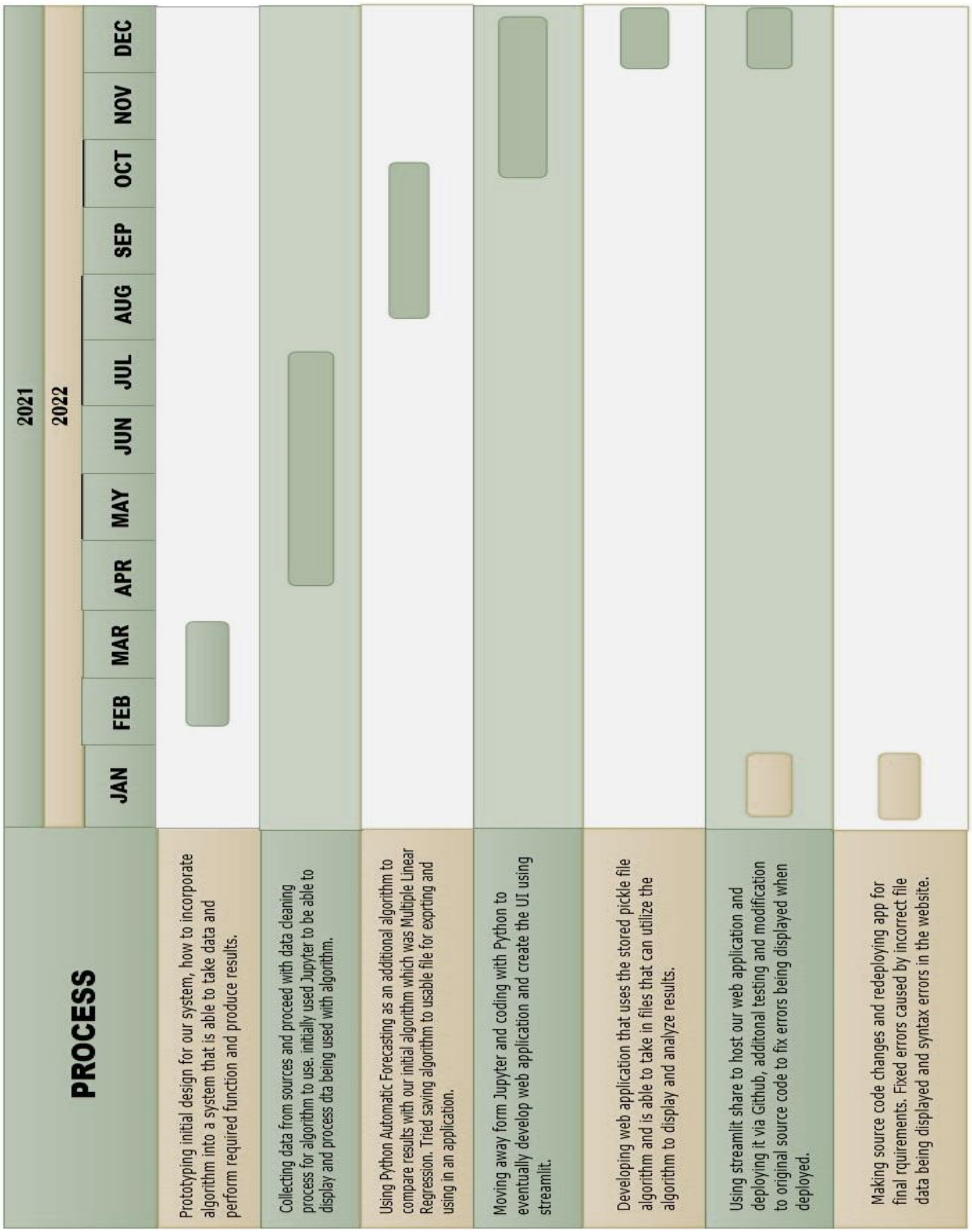
Dr. Ma. Luche P. Sabayle

Date: May 18, 2023

Group Members:
1. Joan Marie Arsenio
2. April Joy Gallano
3. Arle Kai Franco Gorriceta
4. Hiro Parcon
5. Regina Flor Tonogbanua

Appendix C

Gantt Chart



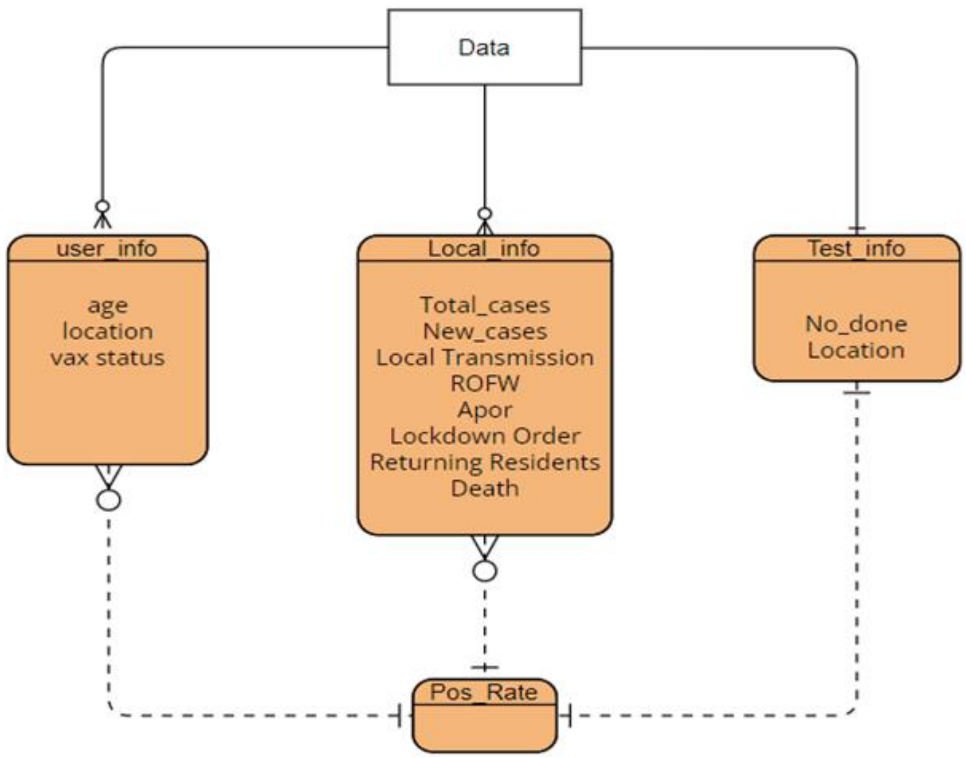
Appendix D

Data Dictionary

Field	Database Field	Data type	Description
Age	Age	int	person's age
Location	Location		person's location
Vaccination Status	Vax_status		status of person's vaccination
Total number of cases	Total_cases		Number of total cases per day
New number of cases	New_cases		Recorded number of cases per day
Local Transmission	local transmission		Recorded number of local transmission per day
ROFW	ROFW		number of returning OFW recorded in Iloilo City
APOR	APOR		Authorize person to go out in residents such as front liners
Lockdown Order	lockdown order		restrictions given in an area for certainperiod of time
Returning Residents	returning residents		recorded number of persons returning from outside the province
Death	death		recorded number of deaths caused by covid 19
Number of tests done	no_done		recorded number of test done per day
Positivity Rate	pos_rate		total rate of positive disease

Appendix E

Entity-Relationship Diagram / Script



Appendix F

Sample Program Codes / Project Charter

Welcome.py

```
import pandas as pd #visualization
from sklearn.model_selection import train_test_split
#Train/Test/Split
from sklearn.linear_model import LinearRegression #Actual
Algorithm
import matplotlib.pyplot as plt #To plot data
from sklearn.metrics import r2_score, mean_squared_error
import streamlit as st
import datetime
import seaborn as sn
from math import sqrt

lr = LinearRegression()
df = pd.read_csv('Covid.csv')
df.head()

df['Month'] = df['Month'].apply(lambda x :
datetime.datetime.strptime(x, "%Y-%m-%d"))
df = df.drop('Month', axis=1)
```

```
st.title("Development of Multiple Linear Regression with
Covid-19 data")
st.markdown("""
     The data set gathered contains important data gathered
from the official Department of Health
     Website as well as other variables used from Iloilo
City Covid-19 Emergency Operations Center Page.
     The data we have gathered here contains 45 weeks of
data due to the fact many of these datasets were
     omitted over time. The data sets gathered here are as
follows:
     ### Data Description
     -   Positivity (Retrieved from DOH)
     -   Total Cases
     -   New Cases
     -   Local Transmissions
     -   Returning Residents
     -   ROFW (Returning Overseas Filipino Workers)
     -   Deaths
     -   APOR (Authorized Personnel Outside Residence)
     -   Lockdown Order
     -   Notes (Types of Lockdowns)
```

```
    -   Increase in Cases (Compared to previous day's
data)
    -   Increase in Positivity Rates (Compared to previous
day's data)
""")
st.sidebar.title("Dataset Options for Viewing")


q1 = st.sidebar.checkbox("Show Raw Data", False)
q2 = st.sidebar.checkbox("Show Positivity Rates Only",
False)
q3 = st.sidebar.checkbox("Show Intercept", False)
q4 = st.sidebar.checkbox("Show Coefficients", False)
q5 = st.sidebar.checkbox("Show Trained Data Prediction",
False)
q6 = st.sidebar.checkbox("Show Trained Data r2 score",
False)
q7 = st.sidebar.checkbox("Show Test Data Prediction",
False)
q8 = st.sidebar.checkbox("Show Test Data r2 score", False)
q9 = st.sidebar.checkbox("Show Heatmap of Correlating
Variables", False)
q10 = st.sidebar.checkbox("Show RMSE and MSE Scores",
False)
```

```
DATA_URL = ('Covid.csv')
def load_data(nrows):
    data = pd.read_csv(DATA_URL, nrows=nrows)
    lowercase = lambda x: str(x).lower()
    data.rename(lowercase, axis='columns', inplace=True)
    return data
data = load_data(10000)


if q1:
    st.subheader('Raw data')
    st.write(data)


x = df[["TotalCases", "NewCases", "LocalTransmission",
"ReturningResidents", "ROFW", "Deaths", "APOR", "LDOrder",
"Increaseincases", "IncreaseinPositivityRate"]]
y = df["Positivity"]


if q2:
    st.subheader("Positivity Rates Only")
    st.write(y)


x_train, X_test, y_train, y_test = train_test_split(x,y,
test_size = 0.2, random_state = 0) #80% to be trained
```

```
lr.fit(x_train, y_train)
LinearRegression()
c = lr.intercept_


if q3:
    st.subheader("Intercept")
    st.write(c)


m =  lr.coef_
print(m)


if q4:
    st.subheader("Coefficients")
    st.write(m)


y_pred_train = lr.predict(x_train)
if q5:
    st.subheader('Trained Data')
    plt.scatter(y_train, y_pred_train)
    plt.xlabel('Actual Positivity Rate')
    plt.ylabel('Predicted Positivity Rate')
    st.pyplot(plt)
```

```
if q6:
    st.subheader('Trained Data r2 score')
    score = r2_score(y_train, y_pred_train)
    st.write(score)

y_pred_test = lr.predict(X_test)

if q7:
    st.subheader('Test Data')
    plt.scatter(y_test, y_pred_test)
    plt.xlabel('Actual Positivity Rate')
    plt.ylabel('Predicted Positivity Rate')
    st.pyplot(plt)

if q8:
    st.subheader('Test Data r2 score')
    testscore = r2_score(y_test, y_pred_test)
    st.write(testscore)

if q9:
    st.subheader('Heatmap of Correlating Variables')
    fig, ax = plt.subplots()
```

```
sn.heatmap(df.corr()[['Positivity']].sort_values(by='Positi
vity', ascending=False), vmin=-1, vmax=1, annot=True,
cmap='BrBG')
    st.write(fig)

if q10:
    st.subheader("MSE and RMSE score of Trained Data")
    predtrain = lr.predict(x_train)
    msetrain = mean_squared_error(y_train, predtrain)
    rmsetrain = sqrt(msetrain)
    st.write("MSE Score")
    msetrain
    st.write("RMSE Score")
    rmsetrain

    st.subheader("MSE and RMSE score of Test Data")
    predtest = lr.predict(X_test)
    msetest = mean_squared_error(y_test, predtest)
    rmsetest = sqrt(msetest)
    st.write("MSE Score")
    msetest
    st.write("RMSE Score")
```

```
    rmsetest


Model.py
import pandas as pd #visualization
from sklearn.model_selection import train_test_split
#Train/Test/Split
from sklearn.linear_model import LinearRegression #Actual
Algorithm
import joblib
from sklearn.metrics import r2_score, mean_squared_error
from math import sqrt

df = pd.read_csv('Test.csv')
df.head()

x = df[["TotalCases", "NewCases", "LocalTransmission",
"ReturningResidents", "ROFW", "Deaths", "APOR", "LDOrder",
"Increaseincases", "IncreaseinPositivityRate"]]
y = df["Positivity"]

lr = LinearRegression()
LinearRegression()
```

```
x_train, X_test, y_train, y_test = train_test_split(x,y,
test_size = 0.3, random_state = 0) #80% to be trained
lr.fit(x_train, y_train)

y_pred_test = lr.predict(X_test)
print(y_pred_test)

predtest = lr.predict(X_test)
msetest = mean_squared_error(y_test, predtest)
rmsetest = sqrt(msetest)
print(msetest)

joblib.dump(lr, "clf.pkl")
```

AF.py

```
import streamlit as st
import pandas as pd
from streamlit_pandas_profiling import st_profile_report
import os
import matplotlib.pyplot as plt
import datetime
import pandas_profiling
import pyaf.ForecastEngine as autof
```

```
import numpy as np


if os.path.exists('./dataset.csv'):
    df = pd.read_csv('dataset.csv', index_col=None)


with st.sidebar:
    st.title("Upload Own Files for Prediction (Please Use
CSV File)")
    choice = st.radio("Navigation",
["Upload","Profiling","Modelling"])
    st.info("Please make sure categorical data is changed
to numerical data beforehand")


if choice == "Upload":
    st.title("Upload Your Dataset")
    file = st.file_uploader("Upload Your Dataset")
    if file:
        df = pd.read_csv(file, index_col=None)
        df.to_csv('dataset.csv', index=None)
        st.dataframe(df)


if choice == "Profiling":
    st.title("Exploratory Data Analysis")
```

```
    profile_df = df.profile_report()
    st_profile_report(profile_df)

if choice == "Modelling":
    chosen_target = st.selectbox('Choose Your Dependent
Variable', df.columns)
    x = df.drop([chosen_target],axis = 1).values
    convert_date = st.selectbox('Choose Column with Date
Configuration',df.columns)
    df[convert_date] = df[convert_date].apply(lambda x :
datetime.datetime.strptime(x, "%Y-%m-%d"))

    if st.button('Run Modelling'):
        lEngine = autof.cForecastEngine()
        lEngine.train(df, convert_date, chosen_target, 12);
        lEngine.getModelInfo()
        df = lEngine.forecast(df, 12);
        df2 = df.fillna(0) #Changing NAN values to 0
        st.write(df2)

        forecast = df[[convert_date , chosen_target ,
'Positivity_Forecast', "Positivity_Forecast_Upper_Bound",
"Positivity_Forecast_Lower_Bound"]]
```

```
        df3 = forecast.fillna(0) #Changing NAN values to 0
        st.write(df3)

        st.line_chart(forecast, x =convert_date,
y=[chosen_target, "Positivity_Forecast",
"Positivity_Forecast_Upper_Bound",
"Positivity_Forecast_Lower_Bound"])
```

Prediction.py

```
import streamlit as st
import pandas as pd
import joblib
st.header("Covid-19 Positivity Rate Prediction Tool")

#"TotalCases", "NewCases", "LocalTransmission",
"ReturningResidents", "ROFW", "Deaths", "APOR", "LDOrder",
"Increaseincases", "IncreaseinPositivityRate"

# Input bar 1
TotalCases = st.number_input("Total Cases",)

# Input bar 2
```

```
NewCases = st.number_input("New Cases")


# Input bar 2

LocalTransmission = st.number_input("Local Transmissions")

# Input bar 2

ReturningResidents = st.number_input("Returning Residents")

# Input bar 2

ROFW = st.number_input("ROFWs")

# Input bar 2

Deaths = st.number_input("Deaths")

# Input bar 2

APOR = st.number_input("Authorized Personnel Outside
Residence")

# Input bar 2

LDOrder = st.selectbox("Select Lockdown Order", ("Travel
Protocol", "General", "Modified", "Enhanced"))

# Input bar 2

Increaseincases = st.number_input("Enter Increase in
Cases")

# Input bar 2

IncreaseinPositivityRate = st.number_input("Enter Increase
in Positivity Rate")

# If button is pressed
```

```
if st.button("Submit"):
    # Unpickle classifier
    clf = joblib.load("clf.pkl")

    # Store inputs into dataframe
    x = pd.DataFrame([[TotalCases, NewCases,
LocalTransmission, ReturningResidents, ROFW, Deaths, APOR,
LDOrder, Increaseincases, IncreaseinPositivityRate]],
                     columns=["TotalCases", "NewCases",
"LocalTransmission", "ReturningResidents", "ROFW",
"Deaths", "APOR", "LDOrder", "Increaseincases",
"IncreaseinPositivityRate"])
    x = x.replace(["Travel Protocol", "General",
"Modified", "Enhanced"], [0, 1, 2, 3])
    # Get prediction
    prediction = clf.predict(x)[0]
    # Output prediction
    st.text(f"The predicted Positivity Rate is:
{prediction}")
```

Upload.py

```
import streamlit as st
import pandas as pd
```

```
from streamlit_pandas_profiling import st_profile_report
import os
from sklearn.metrics import r2_score, mean_squared_error
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
from math import sqrt
import pandas_profiling

if os.path.exists('./dataset.csv'):
    df = pd.read_csv('dataset.csv', index_col=None)

with st.sidebar:
    st.title("Upload Own Files for Prediction")
    choice = st.radio("Navigation",
["Upload","Profiling","Modelling"])
    st.info("Please make sure categorical data is changed
to numerical data beforehand")

if choice == "Upload":
    st.title("Upload Your Dataset")
    file = st.file_uploader("Upload Your Dataset (Please
Use CSV Format)")
```

```
    if file:
        df = pd.read_csv(file, index_col=None)
        df.to_csv('dataset.csv', index=None)
        st.dataframe(df)

if choice == "Profiling":
    st.title("Exploratory Data Analysis")
    profile_df = df.profile_report()
    st_profile_report(profile_df)

if choice == "Modelling":
    chosen_target = st.selectbox('Choose Your Dependent
Variable', df.columns)
    x = df.drop([chosen_target],axis = 1).values
    chosen_size = st.slider('Select Test Size, (Normal Test
Size is 0.2)', 0.0, 1.0)
    if st.button('Run Modelling'):
        y = df[chosen_target]

        #Initializing Algorithm
        lr = LinearRegression()
```

```
        x_train, X_test, y_train, y_test =
train_test_split(x, y, test_size=chosen_size,
random_state=0)
        lr.fit(x_train, y_train)
        y_pred = lr.predict(X_test)
        y_pred_train = lr.predict(x_train)
        y_pred_test = lr.predict(X_test)

        #TRAINING DATA DETAILS
        st.subheader('Trained Data')
        plt.scatter(y_train, y_pred_train)
        plt.xlabel('Actual Positivity Rate')
        plt.ylabel('Predicted Positivity Rate')
        st.pyplot(plt)
        st.subheader('Details:')
        score = r2_score(y_train, y_pred_train)
        st.write("r2 score:", score)
        predtrain = lr.predict(x_train)
        msetrain = mean_squared_error(y_train, predtrain)
        rmsetrain = sqrt(msetrain)
        st.write("MSE score:", msetrain)
        st.write("RMSE score:", rmsetrain)
```

```
#TEST DATA DETAILS
st.subheader('Test Data')
plt.scatter(y_test, y_pred_test)
plt.xlabel('Actual Positivity Rate')
plt.ylabel('Predicted Positivity Rate')
st.pyplot(plt)
st.subheader('Details:')
testscore = r2_score(y_test, y_pred_test)
st.write("r2 score:", testscore)
predtest = lr.predict(X_test)
msetest = mean_squared_error(y_test, predtest)
rmsetest = sqrt(msetest)
st.write("MSE score:", msetest)
st.write("RMSE score:", rmsetest)

#PREDICTIONS
st.subheader('Predictions:')
comparison = pd.DataFrame({"Actual value": y_test,
"Predicted Value": y_pred, "Difference": y_test - y_pred})
st.write(comparison)
```

Appendix G
Sample Questionnaires/Interview Forms / PERT CPM

ISO 25010 QUESTIONNAIRE (SOFTWARE QUALITY STANDARDS)

This form will be composed of two (2) parts: (1) filling up your personal information, and (2) rating the proposed system under the ISO 25010 using the Good to Bad likert scale on the succeeding indicators.

PART I. FILLING UP PERSONAL INFORMATION
please fill up the following fields.
NAME:___________________________________
AGE: ___________________________________
SEX: ____Male ____Female ____LGBTQ+ ____Prefer Not To Say
OCCUPATION: ____________________________
YEARS IN SERVICE:_______________________

PART II. RATING THE PROPOSED SYSTEM
Using the scale below and on the succeeding pages, evaluate the system features of the system choosing on your inferred degree of compliance on a specific metric:

Numeric Value Equivalent Rating:

5 - Excellent

4 - Very Good

3 - Satisfactory

2 - Fair

1 – Poor

	5	4	3	2	1
A. USABILITY FUNCTION					
Completeness The set of functions covers all the specified tasks and user objectives.					
Correctness A product or system provides the correct results with the needed degree of precision.					
Appropriateness The functions facilitate the accomplishment of specified tasks and objectives.					
B. PERFORMANCE EFFICIENCY					
Time-Behavior The response and processing times and throughput rates of a product or system, when performing its functions, meet requirements.					
Resource Utilization The amounts and types of resources used by a product or system, when performing its functions, meet requirements.					
Capacity					

The maximum limits of a product or system parameter meet requirements					
C. COMPATIBILITY					
Co-existence A product can perform its required functions efficiently while sharing a common environment and resources with other products, without detrimental impact on any other product.					
Interoperability Two or more systems, products, or components can exchange information and use the information that has been exchanged.					
D. Usability					
Appropriateness Recognizability Users can recognize whether a product or system is appropriate for their needs.					
Learnability A product or system can be used by specified users to achieve specified goals of learning to use the product or					

system with effectiveness, efficiency, freedom from risk, and satisfaction in a specified context of use.					
Operability A product or system has attributes that make it easy to operate and control.					
User Error Protection A system protects users against making errors.					
User Interaction Aesthetics A user interface enables pleasing and satisfying interaction for the user.					
Accessibility A product or system can be used by people with the widest range of characteristics and capabilities to achieve a specified goal in a specified context of use.					
E. Reliability					
Maturity A system, product, or component meets the needs for reliability under normal operation.					
Availability					

A system, product, or component is operational and accessible when required for use.					
Fault Tolerance A system, product, or component operates as intended despite the presence of hardware or software faults.					
Recoverability In the event of an interruption or a failure, a product or system can recover the data directly affected and re-establish the desired state of the system.					
F. Security					
Confidentiality A product or system ensures that data are accessible only to those authorized to have access.					
Integrity A system, product, or component prevents unauthorized access to, or modification of, computer programs or data.					

Non-repudiation Actions or events can be proven to have taken place so that the events or actions cannot be repudiated later.					
Accountability The actions of an entity can be traced uniquely to the entity.					
Authenticity The identity of a subject or resource can be proved to be the one claimed.					
G. Maintainability					
Modularity A system or computer program is composed of discrete components such that a change to one component has minimal impact on other components.					
Reusability An asset can be used in more than one system, or in building other assets.					
Analysability It is possible to assess the impact on a product or system of an intended change to one or more of its parts, or to					

diagnose a product for deficiencies or causes of failures, or to identify parts to be modified.					
Modifiability A product or system can be effectively and efficiently modified without introducing defects or degrading existing product quality.					
Testability Test criteria can be established for a system, product, or component and tests can be performed to determine whether those criteria have been met.					
H. Portability					
Adaptability A product or system can effectively and efficiently be adapted for different or evolving hardware, software, or other operational or usage environments.					
Installability A product or system can be successfully installed and/or					

uninstalled in a specified environment.					
Replaceability A product can replace another specified software product for the same purpose in the same environment.					

Feedback:

____

____

____

____

ISO 25010 QUESTIONNAIRE (SOFTWARE QUALITY STANDARDS)

This form will be composed of two (2) parts: (1) filling up your personal information, and (2) rating the proposed system under the ISO 25010 using the Good to Bad likert scale on the succeeding indicators.

PART I. FILLING UP PERSONAL INFORMATION

please fill up the following fields.

NAME:_____Dale Alvin Coo______________________________

AGE: ______28____________________________

SEX: __✓_Male ____Female ____LGBTQ+ ____Prefer Not To Say

OCCUPATION: _Nurse_______________________

YEARS IN SERVICE:____1 year______________

PART II. RATING THE PROPOSED SYSTEM

Using the scale below and on the succeeding pages, evaluate the system features of the system choosing on your inferred degree of compliance on a specific metric:

Numeric Value Equivalent Rating:

5 – Excellent

4 - Very Good

3 – Satisfactory

2 – Fair

1 – Poor

	5	4	3	2	1
A. USABILITY FUNCTION					
Completeness The set of functions covers all the specified tasks and user objectives.		✓			
Correctness A product or system provides the correct results with the needed degree of precision.		✓			
Appropriateness The functions facilitate the accomplishment of specified tasks and objectives.	✓				
B. PERFORMANCE EFFICIENCY					
Time-Behavior The response and processing times and throughput rates of a product or system, when performing its functions, meet requirements.	✓				
Resource Utilization The amounts and types of resources used by a product or system, when performing its functions, meet requirements.		✓			
Capacity	✓				

The maximum limits of a product or system parameter meet requirements					
C. COMPATIBILITY					
Co-existence A product can perform its required functions efficiently while sharing a common environment and resources with other products, without detrimental impact on any other product.		✓			
Interoperability Two or more systems, products, or components can exchange information and use the information that has been exchanged.		✓			
D. Usability					
Appropriateness Recognizability Users can recognize whether a product or system is appropriate for their needs.		✓			
Learnability A product or system can be used by specified users to achieve specified goals of learning to use the product		✓			

or system with effectiveness, efficiency, freedom from risk, and satisfaction in a specified context of use.					
Operability A product or system has attributes that make it easy to operate and control.	✓				
User Error Protection A system protects users against making errors.		✓			
User Interaction Aesthetics A user interface enables pleasing and satisfying interaction for the user.		✓			
Accessibility A product or system can be used by people with the widest range of characteristics and capabilities to achieve a specified goal in a specified context of use.		✓			
E. Reliability					
Maturity A system, product, or component meets the needs for reliability under normal operation.	✓				

Availability A system, product, or component is operational and accessible when required for use.	✓				
Fault Tolerance A system, product, or component operates as intended despite the presence of hardware or software faults.		✓			
Recoverability In the event of an interruption or a failure, a product or system can recover the data directly affected and re-establish the desired state of the system.		✓			
F. Security					
Confidentiality A product or system ensures that data are accessible only to those authorized to have access.		✓			
Integrity A system, product, or component prevents unauthorized access to, or		✓			

modification of, computer programs or data.					
Non-repudiation Actions or events can be proven to have taken place so that the events or actions cannot be repudiated later.		✓			
Accountability The actions of an entity can be traced uniquely to the entity.		✓			
Authenticity The identity of a subject or resource can be proved to be the one claimed.		✓			
G. Maintainability					
Modularity A system or computer program is composed of discrete components such that a change to one component has minimal impact on other components.	✓				
Reusability An asset can be used in more than one system, or in building other assets.		✓			
Analysability	✓				

It is possible to assess the impact on a product or system of an intended change to one or more of its parts, or to diagnose a product for deficiencies or causes of failures, or to identify parts to be modified.					
Modifiability A product or system can be effectively and efficiently modified without introducing defects or degrading existing product quality.	✓				
Testability Test criteria can be established for a system, product, or component and tests can be performed to determine whether those criteria have been met.		✓			
H. Portability					
Adaptability A product or system can effectively and efficiently be adapted for different or evolving hardware, software, or other operational or usage environments.		✓			

Installability A product or system can be successfully installed and/or uninstalled in a specified environment.	✓				
Replaceability A product can replace another specified software product for the same purpose in the same environment.			✓		

Feedback:

____

____

____

____

Appendix H

Letter to the Domain Expert

____

____

____

Dear Ma'am/Sir:

Greetings!

We are currently conducting research entitled "Modelling and Prediction of Infectious Disease Cases in Iloilo City Using Multiple Linear Regression" in partial fulfillment of the requirements for the degree, Bachelor of Science in Computer Science. I have come across your impressive work and would greatly value your expertise for our study.

We are reaching out to kindly request your involvement as a domain expert for our research. Your insights would greatly contribute to the depth and quality of our study.
Please accept our deepest gratitude regarding this matter.

Very truly yours,

JOAN MARIE D. ARSENIO

APRIL JOY N. GALLANO

ARLE KAI FRANCO E. GORRICETA

HIRO G. PARCON

REGINA FLOR P. TONOGBANUA

Student Researchers

Noted:

DR. MA. LUCHE P. SABAYLE

Research Adviser

January 8, 2023

MR. DALE ALVIN COO

Nurse

Dear Ma'am/Sir:

We are currently conducting research entitled "Modelling and Prediction of Infectious Disease Cases in Iloilo City Using Multiple Linear Regression" in partial fulfillment of the requirements for the degree, Bachelor of Science in Computer Science. I have come across your impressive work and would greatly value your expertise for our study.

We are reaching out to kindly request your involvement as a domain expert for our research. Your insights would greatly contribute to the depth and quality of our study.

Please accept our deepest gratitude regarding this matter.

Very truly yours,

JOAN MARIE D. ARSENIO

APRIL JOY N. GALLANO

ARLE KAI FRANCO E. GORRICETA

HIRO G. PARCON

REGINA FLOR P. TONOGBANUA

Student Researchers

Noted:

DR. MA. LUCHE P. SABAYLE

Research Adviser

Action Taken:	DALE ALVIN COO
☑ I Accept. ○ Sorry. I don't accept.	Signature over printed name

Appendix I

Disclaimer

This software project and its corresponding documentation entitled "Modeling And Prediction Of Infectious Disease Cases In Iloilo City Using Multiple Linear Regression" is submitted to the College of Information and Communications Technology, West Visayas State University, in partial fulfillment of the requirements for the degree, Bachelor of Science in Computer Science. It is the product of our own work, except where indicated text.

We hereby grant the College of Information and Communications Technology permission to freely use, publish in local or international journal/conferences, reproduce, or distribute publicly the paper and electronic copies of this software project and its corresponding documentation in whole or in part, provided that we are acknowledged.

Joan Marie D. Arsenio
April Joy N. Gallano
Arle Kai Franco E. Gorriceta
Hiro G. Parcon
Regina Flor P. Tonogbanua

June 2023