UNIVERISTY OF PIRAEUS - DEPARTMENT OF INFORMATICS

MSc «Informatics»

<u>MSc Thesis</u>

Thesis Title: Τίτλος Διατριβής:	**Cloud Web Application Tool for collecting, storing and processing data for Machine Learning Operations** Cloud Web Application Tool για τη συλλογή, αποθήκευση και επεξεργασία δεδομένων για λειτουργίες Μηχανικής Εκμάθησης
Student's name-surname:	**Anargyros Douladiris**
Father's name:	**IRAKLIS**
Student's ID No:	ΜΠΠΛ19014
Supervisor:	**Alepis Efthimios, Associate Professor**

November 2023/ Νοέμβριος 2023

3-Member Examination Committee

Virvou Maria	**Alepis Efthimios**	**Patsakis Constantinos**
Professor	**Associate Professor**	**Associate Professor**

Contents

1. Abstract

The purpose of the thesis is to create a Web Application tool which collects data, stores them in database, performs processing and machine learning algorithms, and finally displays the results in charts. The core functionalities of processing and machine learning algorithms are to perform correlation analysis between Covid-19 new cases and Weather, Citizen and Vaccination data. Then, create a NARX (NAR-DY variant) Univariate neural network model for the time series Covid-19 new cases and compare its performance against a NARX (NARX-DY variant) Multivariate neural network model to identify if exogenous correlated inputs are affecting the model predictions (estimations).

Ο σκοπός της διπλωματικής εργασίας είναι να δημιουργήσει ένα εργαλείο διαδικτυακής εφαρμογής που συλλέγει δεδομένα, τα αποθηκεύει σε βάση δεδομένων, εκτελεί επεξεργασία και αλγόριθμους μηχανικής μάθησης και τέλος εμφανίζει τα αποτελέσματα σε γραφήματα. Οι βασικές λειτουργίες επεξεργασίας και μηχανικής μάθησης είναι να εκτελέσουν ανάλυση συσχέτισης μεταξύ των νέων κρουσμάτων Covid-19 και των δεδομένων Καιρού, Πολιτών και Εμβολιασμού. Στη συνέχεια, να δημιουργήσουν ένα μοντέλο νευρωνικού δικτύου Univariate NARX (NAR-DY variant) για τη χρονοσειρά των νέων περιπτώσεων Covid-19 και να συγκρίνουν την απόδοσή του με ένα μοντέλο νευρωνικού δικτύου Multivariate NARX (NARX-DY variant) για να προσδιορίσουν εάν οι εξωγενείς συσχετισμένες είσοδοι επηρεάζουν τις προβλέψεις (εκτιμήσεις) του μοντέλου.

2. Introduction

The coronavirus outbreak in Greece after February 2020 had a drastic impact on the lives of the citizens. Covid-19 is a virus that affects the respiratory system and spreads through droplets when coughing or sneezing. The social life of the people depended on the number of new cases in different Prefectures in the following months. The expert committee tried to assess whether the new cases of Covid-19 would increase or decrease in the next weeks to help the government impose stricter measures. One of the factors that influenced the new cases was the seasons. The measures were less strict in the summer because people were more likely to be outdoors and the new cases were fewer, while the measures were tighter in the winter because people were more likely to be indoors due to the cold. The relationship between weather conditions and Covid-19 transmission was studied, and it was found that the virus spread more in crowded events. By collecting data on the virus transmission, it might be possible to model the data related to Covid-19 and estimate the actual values of new cases. Data will be collected using Dotnet framework, calculations will be done in MATLAB R2022a and charts will be drawn with Google Charts. This process will be deployed on Azure in a form of a multi-container web application, where each container has its own role.

3. Technology stack

- MS SQL Server 2022 - Ubuntu 20.04.5 LTS (Developer Edition)
- AspNet 5.2.9 - Asp Net Core 2.2.7 - Debian GNU/Linux 11 (bullseye)
- Entity Framework Core 7.0.0
- Python 3.10
- MATLAB R2022a – (Ubuntu 20.04.4 LTS)
- Angular 13.3.0 - Node Server Alpine Linux v3.7

- Docker Compose v3.7

4. Architecture

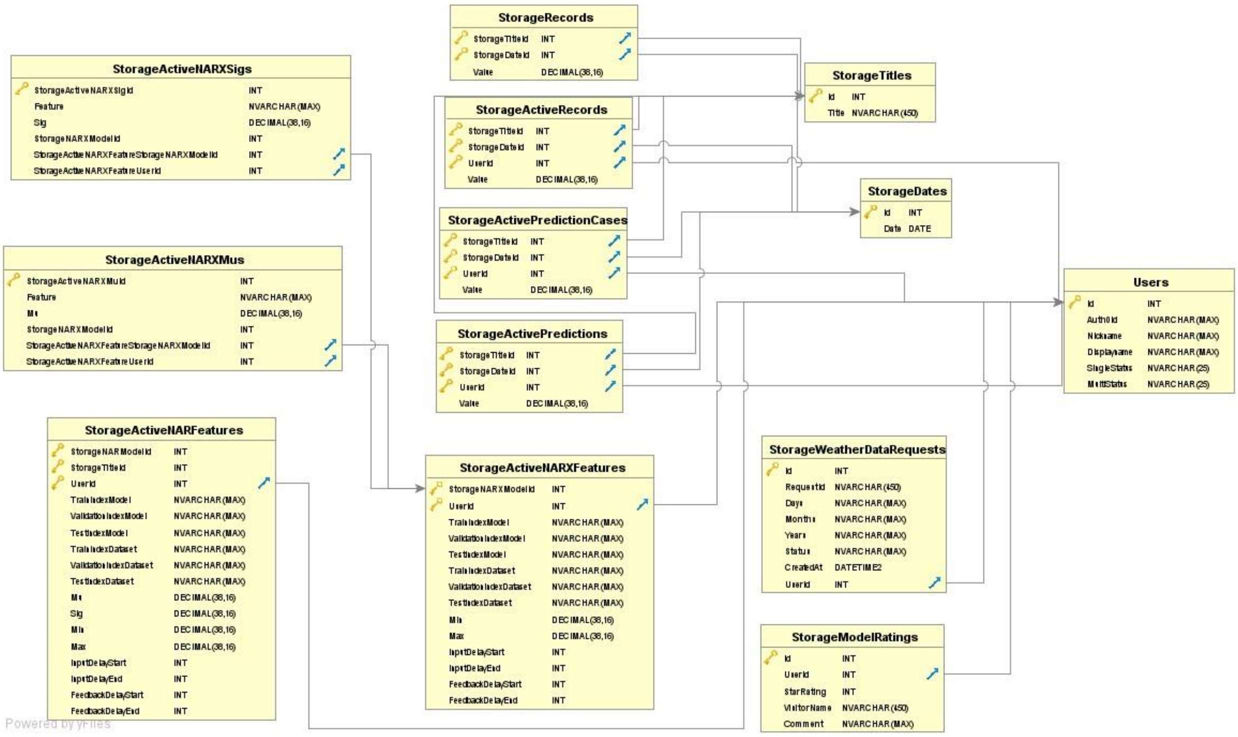
The Web Application is consisted of five containers with different tasks.

- Database Server
- Retrieve Data Server
- MatLab Server
- Machine Learning Server
- Angular Frontend Server

All containers are deployed in the same network.

1. Database Server

The database is an SQL server. The table’s structure is satisfying up to 4th Normal Form, where tables with 3 or more columns do not have any Multi-valued Dependency.



StorageRecords	
StorageTitleId	INT
StorageDateId	INT
Value	DECIMAL(38,16)

StorageTitles	
Id	INT
Title	NVARCHAR(450)

StorageActiveNARXSigs	
StorageActiveNARXSigId	INT
Feature	NVARCHAR(MAX)
Sig	DECIMAL(38,16)
StorageNARXModelId	INT
StorageActiveNARXFeatureStorageNARXModelId	INT
StorageActiveNARXFeatureUserId	INT

StorageActiveRecords	
StorageTitleId	INT
StorageDateId	INT
UserId	INT
Value	DECIMAL(38,16)

StorageDates	
Id	INT
Date	DATE

StorageActivePredictionCases	
StorageTitleId	INT
StorageDateId	INT
UserId	INT
Value	DECIMAL(38,16)

StorageActiveNARXMus	
StorageActiveNARXMuId	INT
Feature	NVARCHAR(MAX)
Mu	DECIMAL(38,16)
StorageNARXModelId	INT
StorageActiveNARXFeatureStorageNARXModelId	INT
StorageActiveNARXFeatureUserId	INT

Users	
Id	INT
Auth0Id	NVARCHAR(MAX)
Nickname	NVARCHAR(MAX)
Displayname	NVARCHAR(MAX)
SingleStatus	NVARCHAR(25)
MultiStatus	NVARCHAR(25)

StorageActivePredictions	
StorageTitleId	INT
StorageDateId	INT
UserId	INT
Value	DECIMAL(38,16)

StorageActiveNARFeatures	
StorageNARModelId	INT
StorageTitleId	INT
UserId	INT
TrainIndexModel	NVARCHAR(MAX)
ValidationIndexModel	NVARCHAR(MAX)
TestIndexModel	NVARCHAR(MAX)
TrainIndexDataset	NVARCHAR(MAX)
ValidationIndexDataset	NVARCHAR(MAX)
TestIndexDataset	NVARCHAR(MAX)
Mu	DECIMAL(38,16)
Sig	DECIMAL(38,16)
Min	DECIMAL(38,16)
Max	DECIMAL(38,16)
InputDelayStart	INT
InputDelayEnd	INT
FeedbackDelayStart	INT
FeedbackDelayEnd	INT

StorageActiveNARXFeatures	
StorageNARXModelId	INT
UserId	INT
TrainIndexModel	NVARCHAR(MAX)
ValidationIndexModel	NVARCHAR(MAX)
TestIndexModel	NVARCHAR(MAX)
TrainIndexDataset	NVARCHAR(MAX)
ValidationIndexDataset	NVARCHAR(MAX)
TestIndexDataset	NVARCHAR(MAX)
Min	DECIMAL(38,16)
Max	DECIMAL(38,16)
InputDelayStart	INT
InputDelayEnd	INT
FeedbackDelayStart	INT
FeedbackDelayEnd	INT

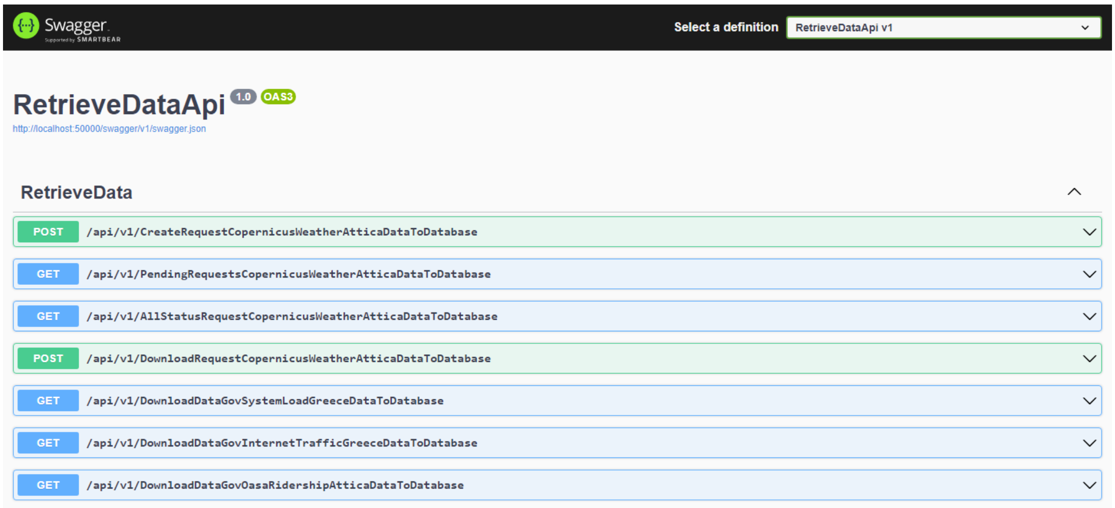
StorageWeatherDataRequests	
Id	INT
RequestId	NVARCHAR(450)
Days	NVARCHAR(MAX)
Months	NVARCHAR(MAX)
Years	NVARCHAR(MAX)
Status	NVARCHAR(MAX)
CreatedAt	DATETIME2
UserId	INT

StorageModelRatings	
Id	INT
UserId	INT
StarRating	INT
VisitorName	NVARCHAR(450)
Comment	NVARCHAR(MAX)

2. Retrieve Data Server

Retrieve Data server is responsible for collecting data. Server’s functionalities are exposed via API (following figure from Swagger UI).

http://localhost:50000/swagger/index.html



RetrieveDataApi 1.0 OAS3

http://localhost:50000/swagger/v1/swagger.json

RetrieveData

POST /api/v1/CreateRequestCopernicusWeatherAtticaDataToDatabase

GET /api/v1/PendingRequestsCopernicusWeatherAtticaDataToDatabase

GET /api/v1/AllStatusRequestCopernicusWeatherAtticaDataToDatabase

POST /api/v1/DownloadRequestCopernicusWeatherAtticaDataToDatabase

GET /api/v1/DownloadDataGovSystemLoadGreeceDataToDatabase

GET /api/v1/DownloadDataGovInternetTrafficGreeceDataToDatabase

GET /api/v1/DownloadDataGovOasaRidershipAtticaDataToDatabase

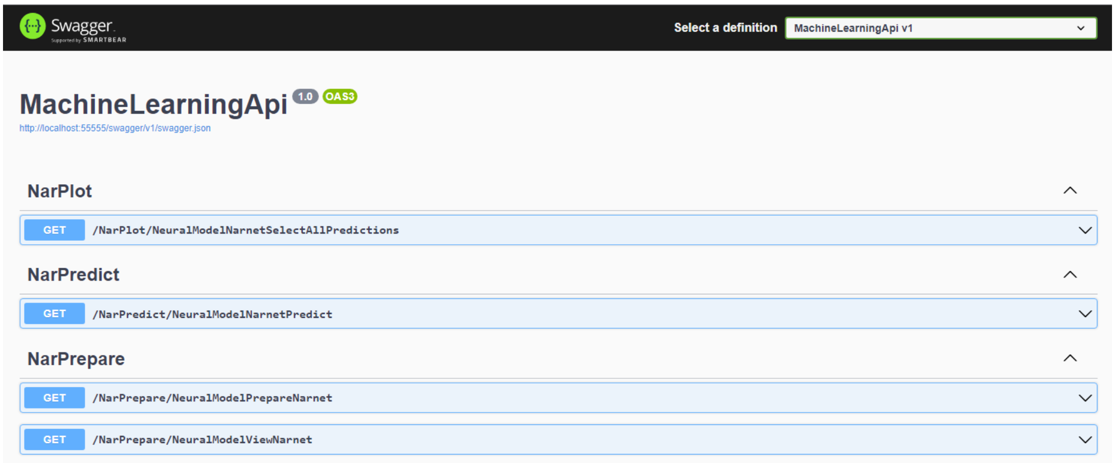
The server is retrieving data by sending HTTP requests to external servers. After the responses are received and parsed, the desired data are extracted, mapped in corresponding class objects [1, p. 3] and then transform in Entity classes. Then, they are sent again with HTTP in the database server. To parse the response from Copernicus external API, the Retrieve Data server is using Python scripts because Asp Net Core library for parsing this specific type of data, which are in GRIB format, is only for Windows OS and the container is using Linux OS. Also, the Covid cases data, which is retrieved from EODY website, is in PDF format and after parsing the PDF file, regular expressions are used to extract the desired data. The rest of the external server responses are in JSON format which do not need any additional handling.

3. Machine Learning Server

This server is responsible for retrieving data via HTTP requests and sending them to the MatLab server for further calculations. The requests for retrieving data are passed to the Retrieve Data server, which is the responsible one to make database calls and, after fetching, the data are sent back to the Machine Learning server.

Machine Learning server is also responsible to make calls to the MatLab server. Every response from the MatLab server is sent back to the Retrieve Data server to be stored in database via Machine Learning server. Server functionalities are exposed via its API (following figure from Swagger UI).

http://localhost:55555/swagger/index.html



MachineLearningApi 1.0 OAS3

http://localhost:55555/swagger/v1/swagger.json

NarPlot

GET /NarPlot/NeuralModelNarnetSelectAllPredictions

NarPredict

GET /NarPredict/NeuralModelNarnetPredict

NarPrepare

GET /NarPrepare/NeuralModelPrepareNarnet

GET /NarPrepare/NeuralModelViewNarnet

4. Retrieve Data and Machine Learning Servers architectures

Both servers are using the five SOLID principles:

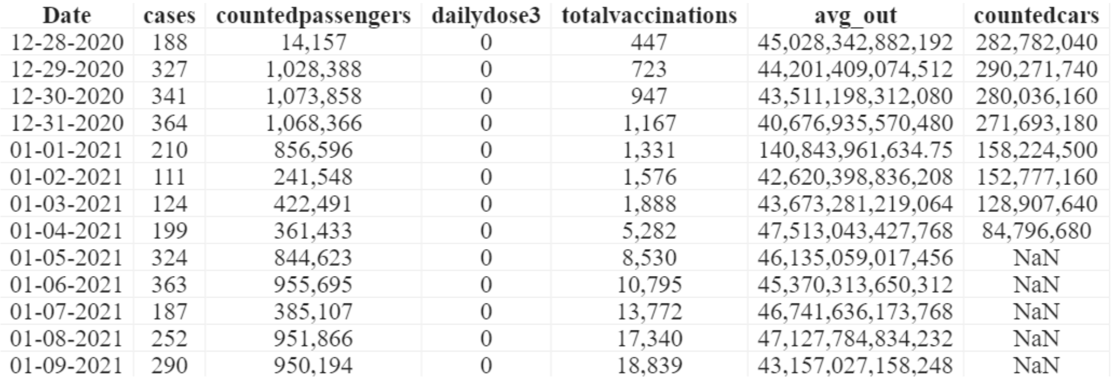
- Single-responsibility principle, where a class has a single responsibility.
- Open-close principle, where a class is open for extension, but closed for modification.
- Interface segregation principle, where each class only implements small usable interfaces.
- Substitution principle, where each subclass could replace its superclass without affecting functionalities of the program.
- Dependency Inversion principle, where high-level classes and low-level classes depend on abstractions, such as interfaces or abstract classes.

Additionally, three programming techniques are implemented:

- Dependency injection, where an object receives its dependencies from external sources rather than creating them internally.
- Background tasks with hosted services, where long-running tasks are running in the background of an Asp.Net Core web application, such as sequentially parsing multiple asynchronous HTTP responses from a queue.
- HTTP Async parsing, where an HTTP response's payload asynchronous parsing is starting immediately after HTTP headers are received and not waiting for the complete payload to be received.

5. MatLab

This server contains functions implemented in MatLab. JSON format is used for both request and response. The functions are directly exposed via API when the server is deployed in a container. All data are parsed in the form of a datatable which contains rows as daily observations and columns as feature names as illustrated on the below figure.



Date	cases	countedpassengers	dailydose3	totalvaccinations	avg_out	countedcars
12-28-2020	188	14,157	0	447	45,028,342,882,192	282,782,040
12-29-2020	327	1,028,388	0	723	44,201,409,074,512	290,271,740
12-30-2020	341	1,073,858	0	947	43,511,198,312,080	280,036,160
12-31-2020	364	1,068,366	0	1,167	40,676,935,570,480	271,693,180
01-01-2021	210	856,596	0	1,331	140,843,961,634.75	158,224,500
01-02-2021	111	241,548	0	1,576	42,620,398,836,208	152,777,160
01-03-2021	124	422,491	0	1,888	43,673,281,219,064	128,907,640
01-04-2021	199	361,433	0	5,282	47,513,043,427,768	84,796,680
01-05-2021	324	844,623	0	8,530	46,135,059,017,456	NaN
01-06-2021	363	955,695	0	10,795	45,370,313,650,312	NaN
01-07-2021	187	385,107	0	13,772	46,741,636,173,768	NaN
01-08-2021	252	951,866	0	17,340	47,127,784,834,232	NaN
01-09-2021	290	950,194	0	18,839	43,157,027,158,248	NaN

To perform correlation analysis various algorithms are executed, like the following:

- Pearson Correlation
- Cross-Correlation
- Autocorrelation

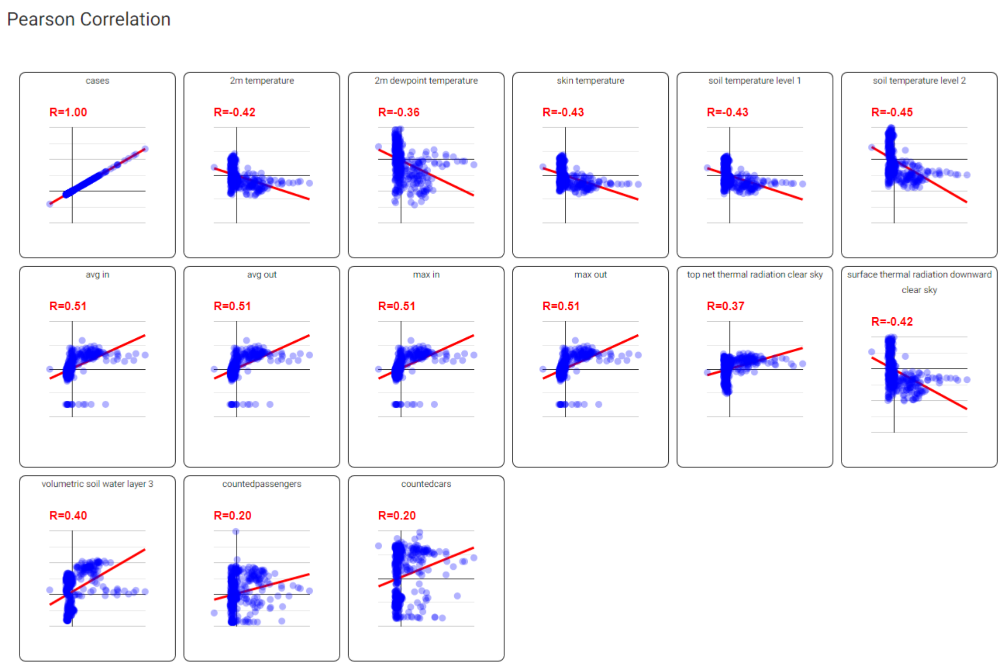
Additionally, preprocess functionality to fill missing data can be used to manipulate the data. Finally, neural models can also be created and stored in database. Only two types of models can be created:

- NAR-DY (Univariate model)
- NARX-DY (Multivariate model)

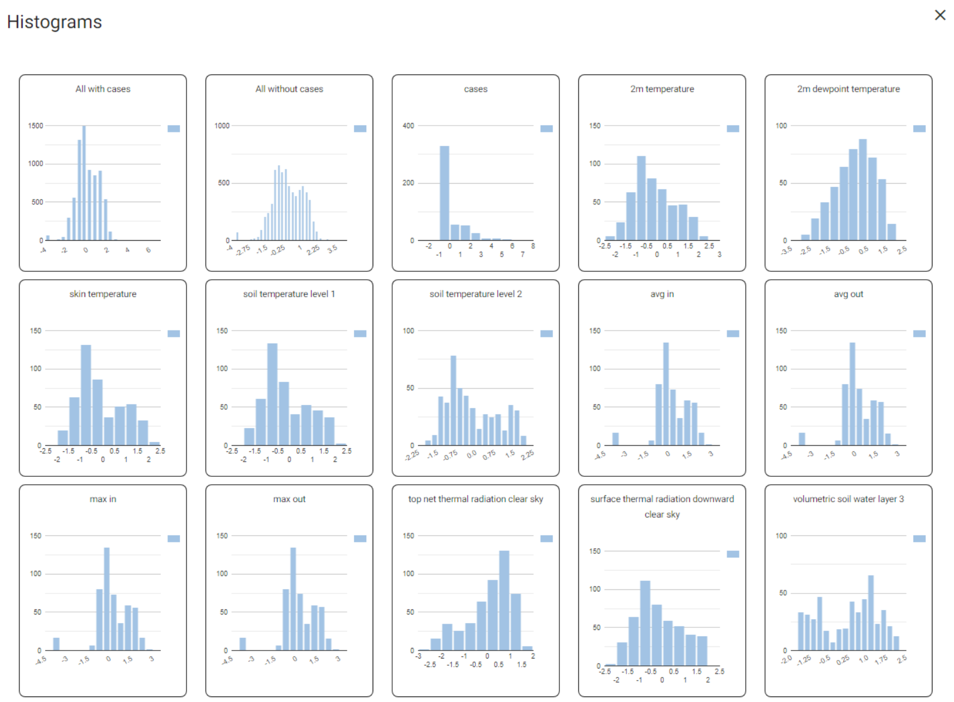
Models are serialized in the form of strings in the MatLab server, before sending them to database. After models are created, predictions can be made only for the target data (Covid cases).

6. Angular Frontend

This server holds the security for the entire web application. It utilizes user methods for login, logout and sign-up, which are provided by the external identity platform Auth0. It implements Single Sign-on (SSO) feature where a user needs to login only one time. It is also a Single Page Application in the form of tabs to avoid additional server calls and increase responsiveness. Apart from the previous functionalities, it provides a rating system for visitors to rate each user's model's predictions. Finally, the frontend visualizes data utilizing Google Charts. Below is a preview:

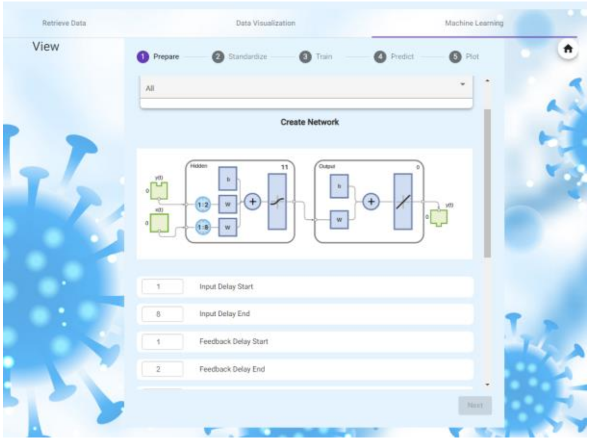


Pearson Correlation



For neural model creation, training and prediction, a wizard with steps was implemented.

Create



All

Create Network

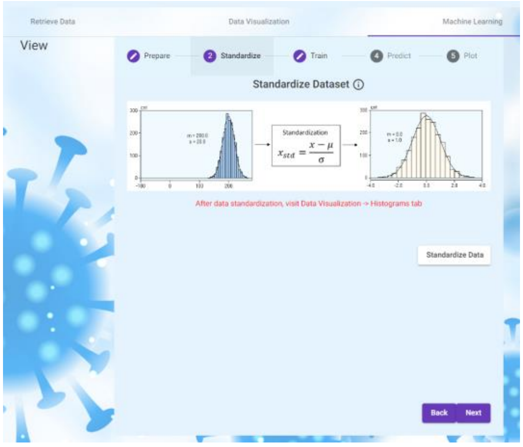
1 Input Delay Start

8 Input Delay End

1 Feedback Delay Start

2 Feedback Delay End

Standardize



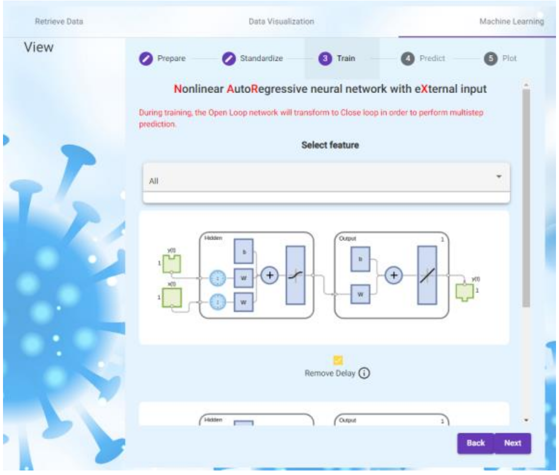
Standardize Dataset

$x_{std} = \frac{x - \mu}{\sigma}$

After data standardization, visit Data Visualization -> Histograms tab

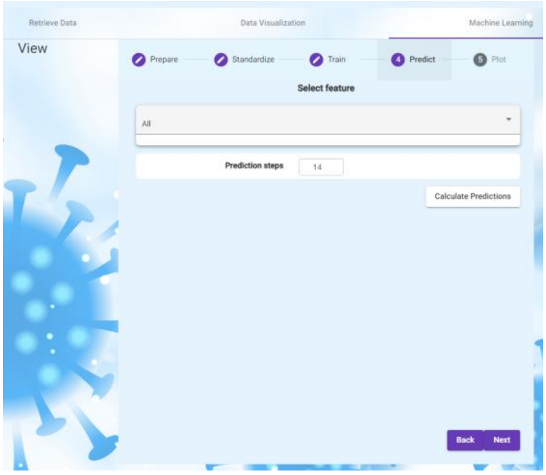
Train

Predict

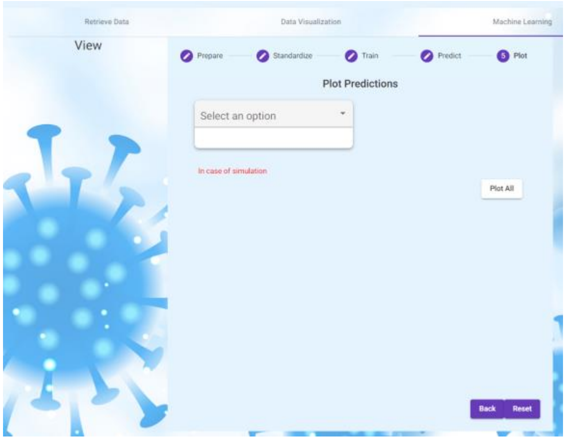


Nonlinear AutoRegressive neural network with eXternal input

During training, the Open Loop network will transform to Close loop in order to perform multistep prediction.

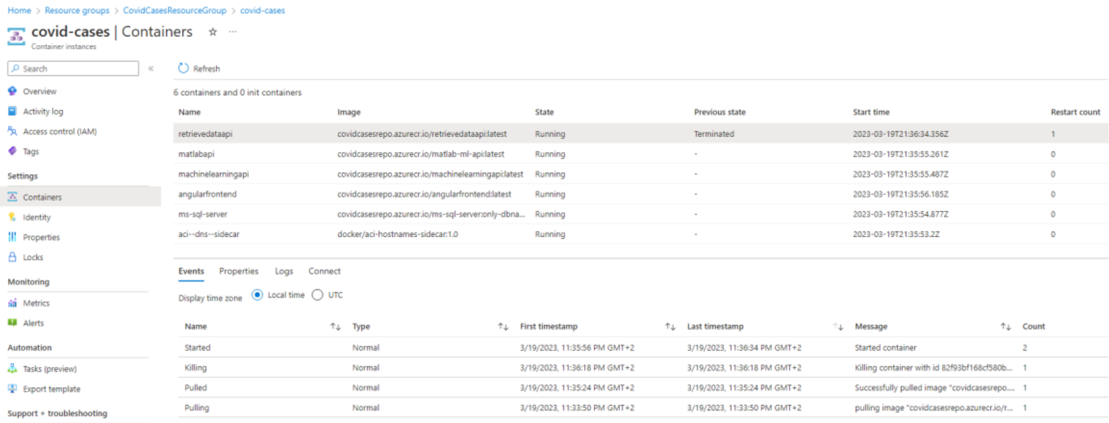


Plot



5. Cloud deployment

All containers are deployed in a single network on Azure utilizing docker-compose for Azure.



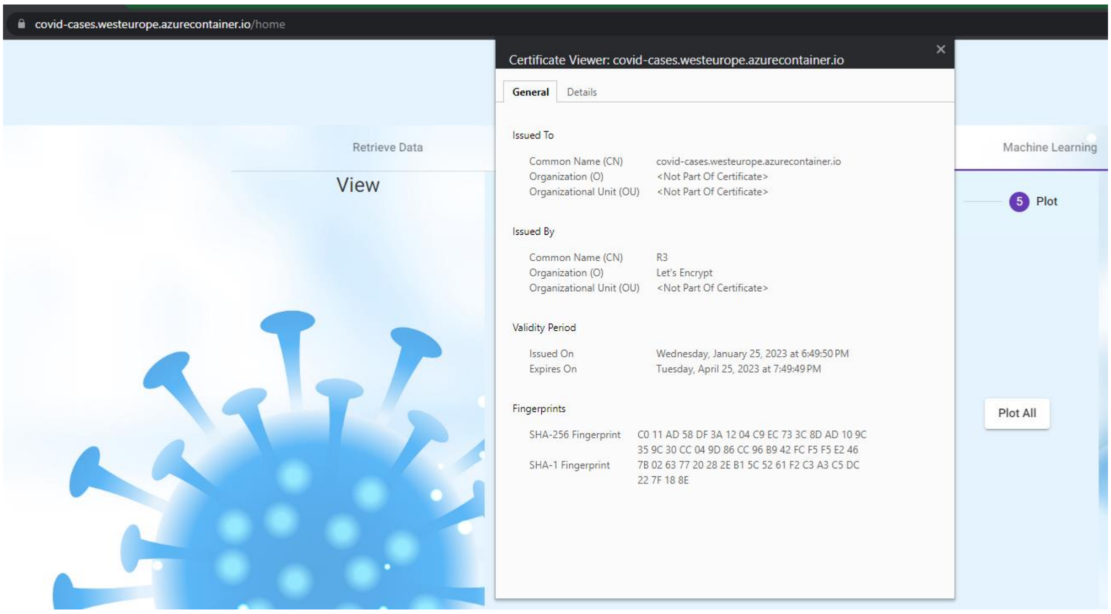
6 containers and 0 init containers

Name	Image	State	Previous state	Start time	Restart count
retrievedataapi	covidcasesrepo.azurecr.io/retrievedataapi:latest	Running	Terminated	2023-03-19T21:36:34.356Z	1
matlabapi	covidcasesrepo.azurecr.io/matlab-ml-api:latest	Running	-	2023-03-19T21:35:55.261Z	0
machinelearningapi	covidcasesrepo.azurecr.io/machinelearningapi:latest	Running	-	2023-03-19T21:35:55.487Z	0
angularfrontend	covidcasesrepo.azurecr.io/angularfrontend:latest	Running	-	2023-03-19T21:35:56.185Z	0
ms-sql-server	covidcasesrepo.azurecr.io/ms-sql-server:only-dbna...	Running	-	2023-03-19T21:35:54.877Z	0
aci--dns--sidecar	docker/aci-hostnames-sidecar:1.0	Running	-	2023-03-19T21:35:53.2Z	0

Display time zone ● Local time ○ UTC

Name	Type	First timestamp	Last timestamp	Message	Count
Started	Normal	3/19/2023, 11:35:56 PM GMT+2	3/19/2023, 11:36:34 PM GMT+2	Started container	2
Killing	Normal	3/19/2023, 11:36:18 PM GMT+2	3/19/2023, 11:36:18 PM GMT+2	Killing container with id 82f93bf168cf580b...	1
Pulled	Normal	3/19/2023, 11:35:24 PM GMT+2	3/19/2023, 11:35:24 PM GMT+2	Successfully pulled image "covidcasesrepo....	1
Pulling	Normal	3/19/2023, 11:33:50 PM GMT+2	3/19/2023, 11:33:50 PM GMT+2	pulling image "covidcasesrepo.azurecr.io/r...	1

It has its own free SSL certificate generated from LetsEncrypt.



Certificate Viewer: covid-cases.westeurope.azurecontainer.io

General Details

Issued To

Common Name (CN)	covid-cases.westeurope.azurecontainer.io
Organization (O)	<Not Part Of Certificate>
Organizational Unit (OU)	<Not Part Of Certificate>

Issued By

Common Name (CN)	R3
Organization (O)	Let's Encrypt
Organizational Unit (OU)	<Not Part Of Certificate>

Validity Period

Issued On	Wednesday, January 25, 2023 at 6:49:50 PM
Expires On	Tuesday, April 25, 2023 at 7:49:49 PM

Fingerprints

SHA-256 Fingerprint	C0 11 AD 58 DF 3A 12 04 C9 EC 73 3C 8D AD 10 9C 35 9C 30 CC 04 9D 86 CC 96 B9 42 FC F5 F5 E2 46
SHA-1 Fingerprint	7B 02 63 77 20 28 2E B1 5C 52 61 F2 C3 A3 C5 DC 22 7F 18 8E

Also, in this network, a reverse-proxy, such as Nginx is used, which is running in the Frontend server. A reverse-proxy redirects the HTTP requests from the Frontend server, which is exposed to public, to the other network internal containers by mapping a server method to the corresponding server port.

6. Publication

The core logic [2], which is implemented in MatLab, was presented at IISA 2023 conference as a paper and it will be publish in the future.

Covid-19 New Cases Correlation Analysis: Weather Conditions, Citizen Traffic and Vaccination Statistics impact in NARX Estimated Regressions in Attica, Greece

***Abstract* —** The purpose of this paper is to perform correlation analysis between Covid-19 new cases and Weather, Citizen and Vaccination data. Then, create a NARX (NAR-DY variant) Univariate model for the time-series Covid-19 new cases and compare its performance against a NARX (NARX-DY variant) Multivariate model to identify if exogenous correlated inputs are affecting the model predictions (estimations).

Keywords—Pearson Correlation, Autocorrelation, Cross Correlation, NARX Neural Networks

1. Introduction

After February 2020 due to Coronavirus breakout in Greece, citizens lives changed drastically. Covid-19 is a virus which infects respiratory system and it spreads through liquid particles when coughing or sneezing [3]. The following months, people social life was depending on whether the number of new cases in most Prefectures was low or high. The committee of experts was trying to evaluate whether the number of Covid-19 new cases will increase or not the following weeks to assist government to take stricter measures. One of those indicators to identify whether rise or fall of new cases was seasons. During summer the measures were not so strict because fewer new cases were reported due to people tended to be outdoors and during winter measures were stricter because people tended to be inside building due to cold. Correlations between Weather conditions and Covid-19 transmission were investigated [4] where virus spreading was increasing in crowded events. By collecting information regarding the virus transmission, it might be possible to model Covid-19 correlated data and estimate real values of Covid-19 new cases. All calculations will be performed in MATLAB R2022a [5] and charts will be plotted with Google Charts [6].

2. Data collection

Related data were collected from open sources available to public.

1.1.1 Weather data

Climate data were collected from the European website Copernicus [7]. Data are extracted from ERA5-Land [8], which is a reanalysis dataset providing a consistent view of the evolution of land variables over several decades. Data are retrieved in the format of GRIB files. This format is designed for storing and distributing weather data. GRIB files are widely used in meteorological applications [9]. To extract data from GRIB files, python packages cfgrib [10] and Pandas [11] were utilized to convert GRIB data to datatables. Sampling rate is daily, calculating the mean value (average) for every hour. Coordinates are the center of Attica for all weather data.

1.1.2 Citizen data

Citizen traffic data were collected from Greece Open Data [12] in JSON format.

- Road traffic*:* Road traffic of Attica region is extracted from various roads of Attica daily. All cars which passed from those roads are summed to a total and this total is the daily sample.
- OASA ridership: OASA ridership of Attica region is extracted from various train stations of Attica. All passengers passing through stations are summed to a total and this total is the daily sample.
- Internet traffic: Internet traffic of Greece with daily sampling.
- MWH of Greece*:* Megawatt-per-hour of energy consumption in Greece.

1.1.3 Vaccination data

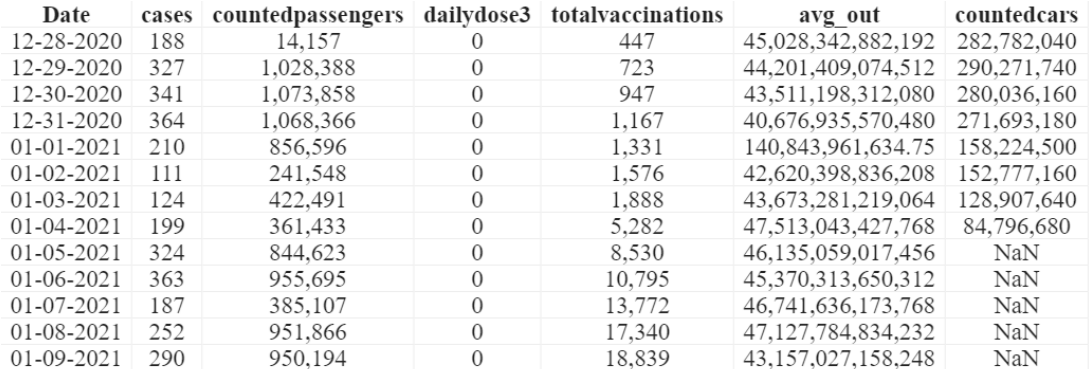
Citizen vaccination statistics were also collected from Greece Open Data [12] in JSON format. Data are extracted for Attica regions only (regions which are part of Attica). Each region statistics are summed to a total and this total is the daily sample.

1.1.4 Covid-19 data

This paper target data will be the daily Covid-19 new cases. Data were extracted from EODY [13] for Attica regions only (regions which are part of Attica) in PDF format. Data are extracted from parsed PDF files to a datatable with daily sample.

3. Data sampling

A time series is a set of observations x_t, each one being recorded at specific time t [14]. Because features observations start at various time points, the first date is going to be the first common date for all data categories. As a result, dates will start from 12-28-2020 and will end to 7-10-2022. The end date is when our daily observations for daily Covid-19 new cases are not available anymore in Greece (daily Covid-19 new cases were replaced with weekly Covid-19 new cases). Every daily observation was collected over time for each feature as illustrated in **Error! Reference source not found.**.



Date	cases	countedpassengers	dailydose3	totalvaccinations	avg_out	countedcars
12-28-2020	188	14,157	0	447	45,028,342,882,192	282,782,040
12-29-2020	327	1,028,388	0	723	44,201,409,074,512	290,271,740
12-30-2020	341	1,073,858	0	947	43,511,198,312,080	280,036,160
12-31-2020	364	1,068,366	0	1,167	40,676,935,570,480	271,693,180
01-01-2021	210	856,596	0	1,331	140,843,961,634.75	158,224,500
01-02-2021	111	241,548	0	1,576	42,620,398,836,208	152,777,160
01-03-2021	124	422,491	0	1,888	43,673,281,219,064	128,907,640
01-04-2021	199	361,433	0	5,282	47,513,043,427,768	84,796,680
01-05-2021	324	844,623	0	8,530	46,135,059,017,456	NaN
01-06-2021	363	955,695	0	10,795	45,370,313,650,312	NaN
01-07-2021	187	385,107	0	13,772	46,741,636,173,768	NaN
01-08-2021	252	951,866	0	17,340	47,127,784,834,232	NaN
01-09-2021	290	950,194	0	18,839	43,157,027,158,248	NaN

Figure 1. Visualized features in a datatable

4. Data analysis

Data collection contains 79 features, which are going to be used as predictors. Data with constant values, without any fluctuation, are discarded. Before proceeding, correlation analysis needs to be done to evaluate which features are suitable.

1.1.5 Pearson Correlation

The product-moment correlation coefficient (Pearson correlation) measures the tendency of two variables to change in value together. The formula for the correlation coefficient is [15, p. 176]:

$$r_{coeff} = \frac{\sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{S_x} \right) \left(\frac{y_i - \bar{y}}{S_y} \right)}{N-1} \tag{1}$$

where *N* are the total observations. $\bar{x}, \bar{y}$ are the mean of the two time-series. The mean is calculated as follows:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} A_i \tag{2}$$

S_x, S_y are the standard deviation of the two time-series.

$$S = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \left| A_i - \mu \right|^2} \tag{3}$$

y will always be the Covid-19 new cases and x will be each feature. Pearson correlation results contain features which have linear relationship with Covid-19 new cases as illustrated in **Error! Reference source not found.**. But it is not known which precedes the other. Feature x or feature y ?

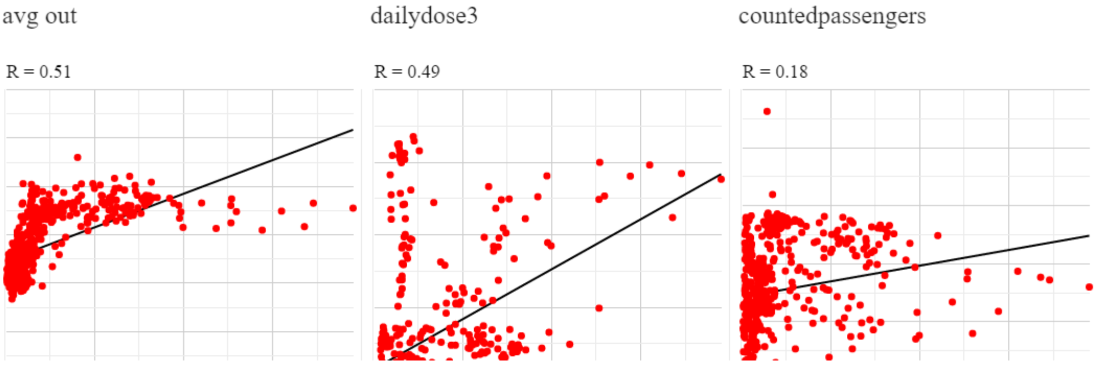


Figure 2. Features Pearson Correlation with descending R

1.1.6 Cross Correlation

To identify which feature precedes Covid-19 new cases, this paper will use the Cross-Correlation between each feature against new cases. According to [16, p. 193] (7.4), Cross-Correlation is the Pearson product-moment correlation coefficient. The difference is that the y variable of the x and y time-series is going to be calculated for a selected number of lags. The formula is the following:

$$r_{coeff}(lag) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_{i-lag} - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_{i-lag} - \bar{y})^2}} \quad (4)$$

where x is the feature series and y is the Covid-19 new cases series. In the plot of each feature cross-correlation, this paper will try to identify a peak in the negative side of *x-axis*. That will indicate that a similarity of the two time-series is going to be apparent in the lag where the peak occurs. That means that the feature (x) will precede Covid-19 new cases (y) and not the opposite. Positive peaks in the *y-axis* indicate positive correlation and negative peaks indicate negative correlation. The following features in Table I and **Error! Reference source not found.** were selected based on their negative peaks at *x-axis* and lags lower than the Autocorrelation lag which is calculated in the next section.

TABLE I. CROSS-CORRELATION X-AXIS NEGATIVE PEAKS.

Cross-Correlation with Covid-19 new cases	
Features	***Negative Lags***
dailyDose3	-28
energyMwh	-14
avgOut	-20
10mUcomponentOfWind	-29
totalPrecipitation	-21
totalCloudCover	-55
countedPassengers	-14
countedCars	-35

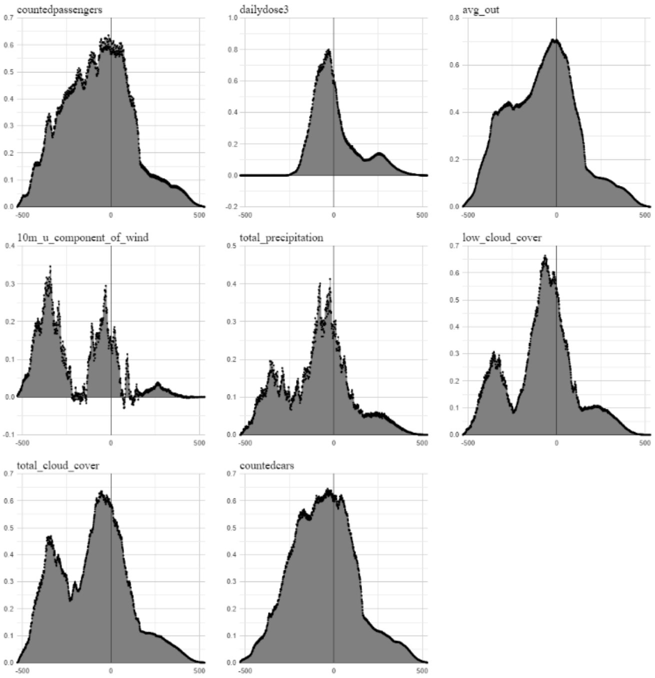


Figure 3. Cross-Correlation x-axis negative peaks

1.1.7 Autocorrelation

To identify if Autocorrelation exists in the Covid-19 new cases time series, the following formula [17, p. 26] and MATLAB [18] equation will be used:

$$r_{lag} = \frac{\sum_{i=1}^{N-lag} \left(Y_i - \bar{Y}\right)\left(Y_{i+lag} - \bar{Y}\right)}{\sigma_Y^2} \qquad (5)$$

where Y is the Covid-19 new cases, N is the total number of observations and σ_Y^2 the sample variance. This method will take a copy of itself and calculate the correlation coefficient r_{lag} between y_i and y_{i+lag} for all observations, where lags range will be from 0 to the total number of observations minus 1. At **Error! Reference source not found.** the horizontal red lines are the calculated 95% confidence bounds upper and lower limits respectively. All r_{lag} coefficients outside the 95% confidence bounds are significant lags, which means autocorrelation exists. The point before crossing the upper confidence bound is going to be used, which is 95.

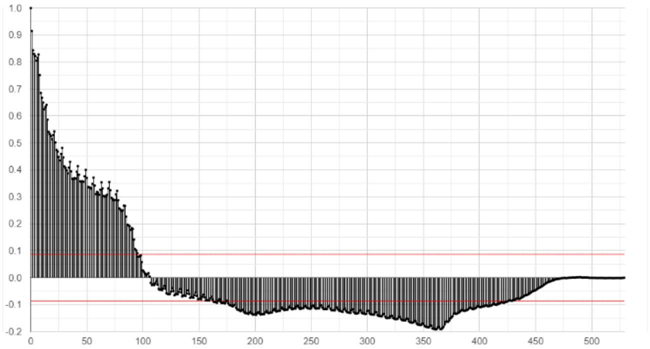


Figure 4. Autocorrelation Covid-19 new cases

5. Neural Network selection

This paper will use a Non-Linear Autoregressive model with exogenous inputs (NARX). This model is a recurrent dynamic network, with feedback connections enclosing several layers of the network [19]. Such network can be algebraically stated as:

$$y_t = F\left(y_{t-1}, y_{t-2}, \ldots, y_{t-n_y}, u_{t-1}, u_{t-2}, \ldots, u_{t-n_u}\right) \tag{6}$$

where y_t will be the Covid-19 new cases and u_t the exogenous inputs, like weather or citizen traffic data. Estimated outputs from y_{t-1} to y_{t-n_y}, where n_y to be the total number of y past values, will be fed back to the input of the network along with exogenous inputs u_t from u_{t-1} to u_{t-n_u}, where n_u to be the total number of past values for the exogenous input. n_y is the Input Delay and n_u is the Feedback Delay of the network. **Error! Reference source not found.** is depicting the NARX network, where the vectors of past values are shown as Tapped Delay Lines (TDL).

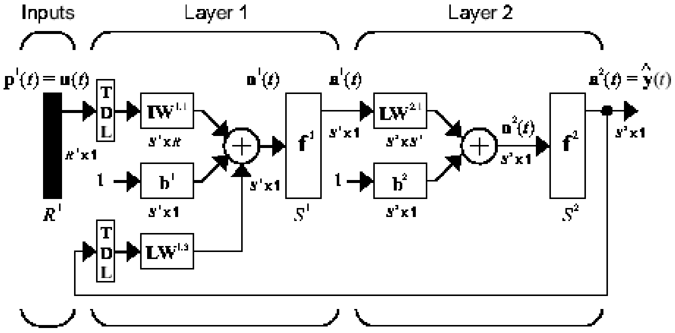


Figure 5. NARX network [17]

According to paper [20, p. 11], several variations of NARX can be utilized. Two of them were selected and described below.

1.1.1 NAR-DY

This variant is using delayed targets and predictions (estimated outputs) feedback but without exogenous inputs. This means that the (6) is modified as below:

$$y_t = F\left(d_{t-1}, d_{t-2}, ..., d_{t-ID}, y_{t-1}, y_{t-2}, ..., y_{t-FD}\right) \tag{7}$$

where d is the target past values with ID the total number of the input delays and y is the past predictions with FD the total number of feedback delays. This Univariate model variant will be used to predict the target d , which is the Covid-19 new cases.

1.1.2 NARX-DY

This variant is using delayed targets, predictions feedback and exogenous inputs. The (6) is modified as below:

$$y_t = F\left(u_t, u_{t-1}, ..., u_{t-ID}, d_{t-1}, ..., d_{t-ID}, y_{t-1}, ..., y_{t-FD}\right) \tag{8}$$

where u is the exogenous input of current and past values with ID the total number of the input delays. u and d total number of delays is the same. This Multivariate model variant will be used to predict the Covid-19 new cases using the exogenous and the target time-series as inputs.

6. Preprocess data

1.1.3 Fill missing data

For data which contain NaN values, the moving window of 14 values will be used filling NaN values. This method will create a window of 14 consecutive values and if a NaN value exists in the window, it will be replaced by the mean of the existing values of the window.

1.1.4 Split dataset

The target is to predict the last 30 days of the dataset, which are from 6-11-2022 to 7-10-2022. The dataset to train the model contains 530 observations, from 12-28-2022 to 6-10-2022. For the Test set, the last 30 days will be used. For the Validation set, the previous 90 days will be used and for the Training set the rest of the observations will be used. This means that the dataset will be split along its time index. The first 77% of the observations will be used for Training, the next 17.5% for Validation and the next 5.5% for Testing.

1.1.5 Standardize dataset

To standardize the data, the dataset will be split again to Training and Validation as one set and Test as another set. The mean (***μ***) from (2) and the standard deviation (***S***) from (3) of the Training and Validation sets will be extracted and standardize the values of all the datasets. For each observation A_i , standardization is going to be the following:

$$Astd_i = \frac{A_i - \mu}{S} \tag{9}$$

7. Training function

Both models will be trained using the Scaled Conjugate Gradient algorithm [21]. This training function does not depend in any user parameters. It solves function approximation problems and it is faster for large networks with many weights.

8. Performance function

Both models will use the Mean Squared Error performance function [22] with default parameters to measure the performance of the regression models.

9. Hidden layers and number of Neurons

Two hidden layers will be used. The number of neurons at each layer will be the result of Bayesian Optimization algorithm using the network as a black box with input parameters the number of neurons of both layers and output the performance metrics of the estimated values against the true values. This process is executed for 200 iterations.

10. Training Network

During training, the output of the model will be fed back to the input. But during training true output is available, so the model will be trained at first in an Open loop (**Series-Parallel**) architecture **Error! Reference source not found.** where true outputs will be fed instead of estimated outputs [19]. Then, the model will be converted to Close Loop (**Parallel**) architecture **Error! Reference source not found.** and the model will be trained again.

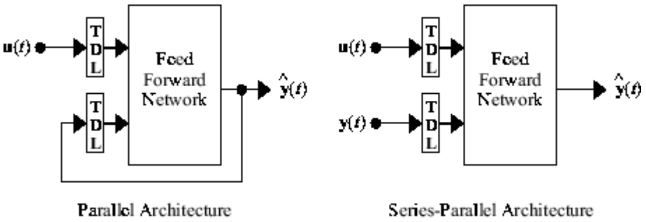


Figure 6. NARX architectures

11. Univariate NAR-DY Network

To create the Univariate model, the autocorrelation significant lags from **Error! Reference source not found.**, which is 95, will be used for both input and feedback delays. As mentioned before, predictions will be from 6-11-2022 to 7-10-2022. To evaluate model predictions against true values, the Root Mean Squared Error (RMSE) is calculated with the following formula:

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{\left(\hat{y}_i - y_i\right)^2}{N}} \tag{10}$$

where $\hat{y}$ is the estimated output (prediction) de-standardized and y is the true output.

12. Multivariate NARX-DY Network

To create the Multivariate model, the autocorrelation significant lags from **Error! Reference source not found.** will be used again as feedback and target input delays. The input delays from exogenous inputs should be same as the significant lags from Cross-Correlation Table I of the corresponding feature. The (8) should convert to the below:

$$y_t = F\left(u_t, u_{t-1}, ..., u_{t-ID_{XCORR}}, d_{t-1}, ..., d_{t-ID_{ACORR}}, y_{t-1}, ..., y_{t-FD}\right) \qquad (11)$$

where $ID_{XCORR} < ID_{ACORR}$ and $ID_{ACORR} = FD$. But implementation of MATLAB NARX model allows only the same total number of sequences as inputs, where $ID_{XCORR} = ID_{ACORR}$. This paper will test 3 cases as the total number of input delays.

1. Cross-Correlation lag.
2. Bayesian Optimization algorithm parameter with range from Cross-Correlation lag to Autocorrelation lag.
3. Autocorrelation lag of Covid-19 new cases.

13. Conclusion

From **Error! Reference source not found.** as depicted in **Error! Reference source not found.**, Univariate model resulted in an RMSE of 2229. From **Error! Reference source not found.** collected results of the Multivariate models in **Error! Reference source not found.**, **Error! Reference source not found.** and **Error! Reference source not found.**, the lowest RMSE (Minimum Error) of the 3 of the input delays cases will be compared as presented in Table VI. It appears that the feature with the minimum error is the total number of passengers of OASA daily ridership (countedpassengers) with an RMSE of 1775. The feature total number of the daily 3rd dose of Covid-19 vaccine (dailydose3), even though it has a higher Pearson Correlation R from countedpassengers, resulted in RMSE of 1998 which is greater than 1775. The feature Total Cloud Cover through the atmosphere (total_cloud_cover) has the highest RMSE of 4125. The rest of the features resulted approximately in an RMSE of 2600, which is still greater that the Univariate model RMSE of 2229. Furthermore, a combination of the features with the lowest RMSE could be made. After applying the Bayesian Optimization algorithm to identify the Input Delays which performed best, both features countedpassengers and dailydose3 as exogenous inputs resulted in an RMSE of 1541 Table VII and **Error! Reference source not found.**. As a conclusion, it seems that the total number of passengers of OASA daily ridership and the total number of the daily 3rd dose of Covid-19 vaccine seem to have the strongest correlation with the daily Covid-19 new cases.

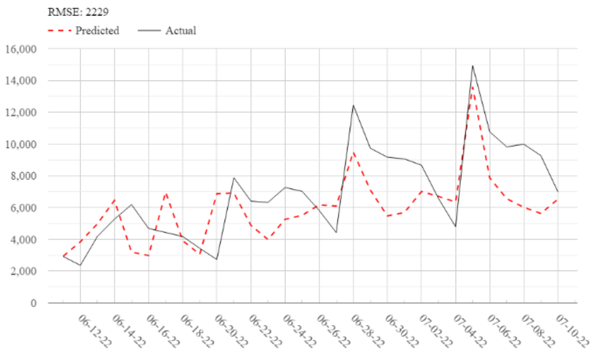


Figure 7. Attica Covid-19 new cases predictions from 06-11-2022 to 7-10-2-22 with RMSE = 2229

TABLE II. UNIVARIATE MODEL PREDICTION

Feature	*RMSE*	*Input Delays*	*Feedback Delays*	*1st layer Neurons*	*2nd layer Neurons*
Covid-19 new cases	2229	95	95	6	22

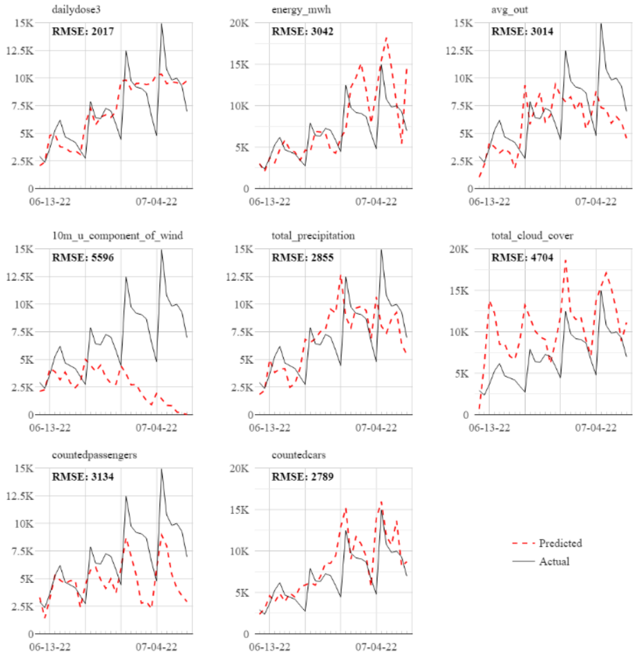


Figure 8. Input Delays RMSE Case 1 from 06-11-22 to 07-10-22

TABLE III. INPUT DELAYS RMSE CASE 1

Feature	*RMSE*	*Input Delays*	*Feedback Delays*	*1st layer Neurons*	*2nd layer Neurons*
dailyDose3	2017	28	95	21	15
energyMwh	3042	14	95	36	32
avgOut	3014	20	95	36	40
10mUcomponentOfWind	5596	29	95	14	20
totalPrecipitation	2855	21	95	27	38
totalCloudCover	4704	55	95	40	27
countedPassengers	3134	14	95	31	29
countedCars	2789	35	95	30	24

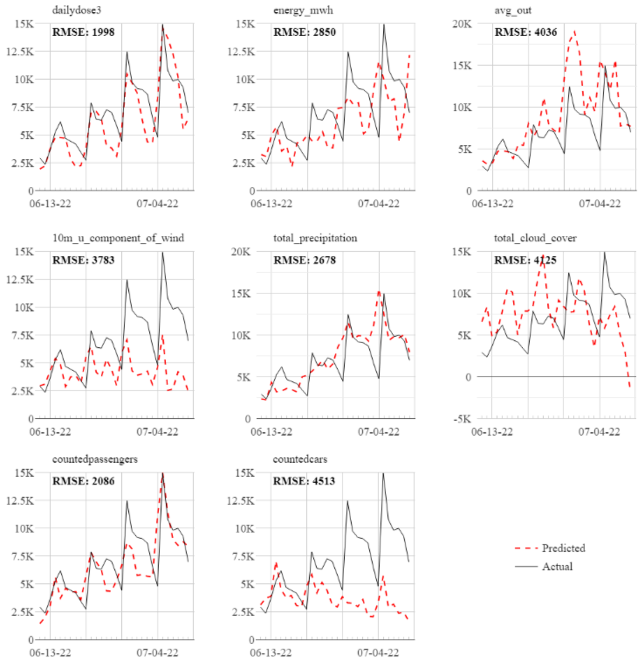


Figure 9. Input Delays RMSE Case 2 from 06-11-22 to 07-10-22

TABLE IV. INPUT DELAYS RMSE CASE 2

Feature	*RMSE*	*Input Delays*	*Feedback Delays*	*1st layer Neurons*	*2nd layer Neurons*
dailyDose3	1998	51	95	26	20
energyMwh	2850	14	95	40	34
avgOut	4036	89	95	37	17
10mUcomponentOfWind	3783	62	95	40	40
totalPrecipitation	2678	25	95	36	36
totalCloudCover	4125	55	95	20	31
countedPassengers	2086	71	95	33	36
countedCars	4513	67	95	22	31

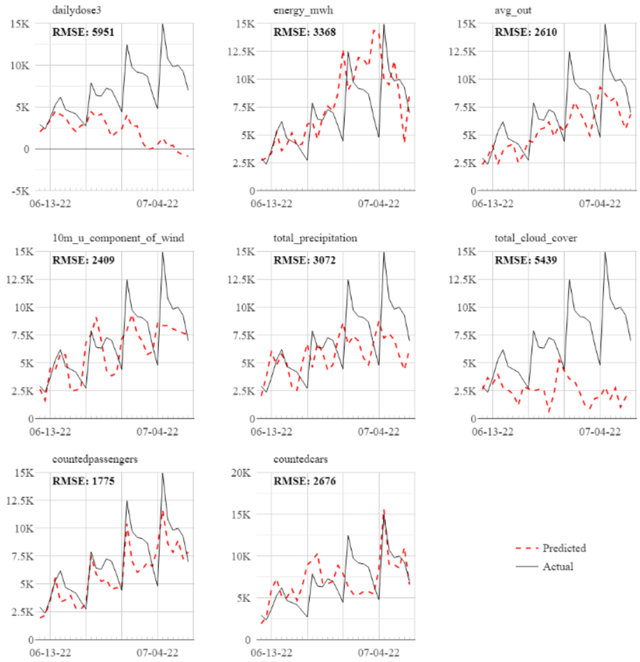


Figure 10. Input Delays RMSE Case 3 from 06-11-22 to 07-10-22

TABLE V. INPUT DELAYS RMSE CASE 3

Feature	*RMSE*	*Input Delays*	*Feedback Delays*	*1st layer Neurons*	*2nd layer Neurons*
dailyDose3	5951	95	95	35	21
energyMwh	3368	95	95	24	16
avgOut	2610	95	95	38	10
10mUcomponentOfWind	2409	95	95	15	40
totalPrecipitation	3072	95	95	40	23
totalCloudCover	5439	95	95	21	29
countedPassengers	1775	95	95	30	32
countedCars	2676	95	95	12	32

TABLE VI. SUMMARY

Feature	*Minimum RMSE*	*Pearson Correlation*
dailyDose3	1998	0.49
energyMwh	2850	0.05
avgOut	2610	0.51
10mUcomponentOfWind	2409	-0.02
totalPrecipitation	2678	0.11
totalCloudCover	4125	0.19
countedPassengers	1775	0.18

Feature	*Minimum RMSE*	*Pearson Correlation*
countedCars	2676	0.19

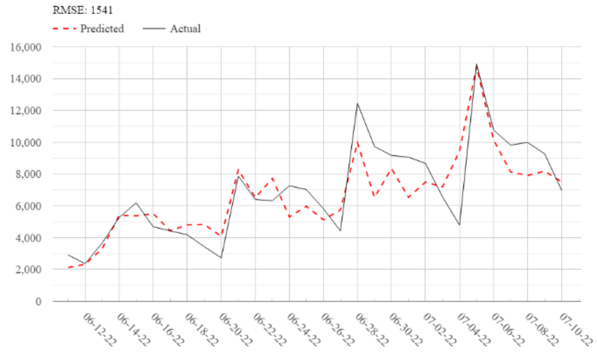


Figure 11. Attica Covid-19 new cases predictions **from 06-11-2022 to 7-10-2-22 with RMSE = 1541**

TABLE VII. COUNTEDPASSENGERS AND DAILYDOSE3 RMSE

Feature	*RMSE*	*Input Delays*	*Feedback Delays*	*1st layer Neurons*	*2nd layer Neurons*
dailyDose3 + countedPassengers	1541	92	95	13	10

7. Conclusions

For creating this web application tool, multiple technologies were used. MatLab provided complete services for implementing machine learning techniques and data processing with quick container deployment. Asp.Net Core provided complete solutions for handling data structures along with Entity Framework to quickly mapping classes to SQL tables. Angular implements so many modules to quickly built a complete user interface. Finally, docker-compose, by incorporating Azure dedicated commands, made the cloud deployment feasible. By using this stack, which offers a wide range of services, this web application tool building was speed up by avoiding writing boilerplate code and focusing more on implementing ideas.

8 Bibliography

[1] V. M. Alepis Efthymios, «Object oriented architecture for affective multimodal e-learning interfaces».

[2] E. A. G. A. T. L. C. J. Maria Virvou, Machine Learning Paradigms: Advances in Learning Analytics.

[3] «Coronavirus disease (COVID-19),» [Ηλεκτρονικό]. Available: https://www.who.int/health-topics/coronavirus#tab=tab_1.

[4] C. C. M. L. R. N. Q. L. D. S. S. L. T. Y. Y. L. B. D. a. X. W. W. Cao, «Important factors affecting COVID-19 transmission and fatality in,» *Elsevier - PMC COVID-19 Collection.*

[5] «MATLAB,» [Ηλεκτρονικό]. Available: https://www.mathworks.com/products/matlab.html.

[6] «Google Charts,» [Ηλεκτρονικό]. Available: https://developers.google.com/chart.

[7] «Copernicus Climate Change Service,» [Ηλεκτρονικό]. Available: https://climate.copernicus.eu/.

[8] «ERA5-Land hourly data from 1950 to present,» [Ηλεκτρονικό]. Available: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview.

[9] «What are GRIB files and how can I read them,» [Ηλεκτρονικό]. Available: https://confluence.ecmwf.int/display/CKB/What+are+GRIB+files+and+how+can+I+read+them.

[10] «cfgrib: A Python interface to map GRIB files to the NetCDF Common Data Model following the CF Convention using ecCodes,» [Ηλεκτρονικό]. Available: https://pypi.org/project/cfgrib/.

[11] «pandas: powerful Python data analysis toolkit,» [Ηλεκτρονικό]. Available: https://pypi.org/project/pandas/.

[12] «Find open data,» [Ηλεκτρονικό]. Available: https://data.gov.gr/.

[13] «Εκθέσεις επιδημιολογικής επιτήρησης λοίμωξης από τον SARS-CoV-2,» [Ηλεκτρονικό]. Available: https://eody.gov.gr/epidimiologika-statistika-dedomena/ektheseis-epidimiologikis-epitirisis-loimoxis-apo-ton-sars-cov-2/.

[14] R. A. D. Peter J. Brockwell, Introduction to Time Series and Forecasting, Second Edition.

[15] S. Boslaugh, Statistics in a Nutshell, 2nd Edition.

[16] P. a. J. M. Timothy R.Derrick, «Time Series Analysis: The Cross-Correlation Function».

[17] G. M. J. G. R. George Box, «Time series analysis_ Forecasting and control-Prentice Hall (1994),» 1994.

[18] «Sample autocorrelation,» [Ηλεκτρονικό]. Available: https://www.mathworks.com/help/econ/autocorr.html.

[19] «Design Time Series NARX Feedback Neural Networks,» [Ηλεκτρονικό]. Available: https://www.mathworks.com/help/deeplearning/ug/design-time-series-narx-feedback-neural-networks.html.

[20] E. Hristev, «NARX neural networks for sequence processing tasks».

[21] M. F. MEILLER, «A Scaled Conjugate Gradient Algorithm».

[22] «Mean squared normalized error performance function,» [Ηλεκτρονικό]. Available: https://www.mathworks.com/help/deeplearning/ref/mse.html.