West Nile Virus
Prediction

Predict West Nile virus in mosquitoes across the city of Chicago

DSI03 : 28 Nov 2023

Introduction

West Nile virus (WNV) is the leading cause of mosquito-borne disease in the continental United States. It is most commonly spread to people by the bite of an infected mosquito.

There are no vaccines to prevent or medications to treat WNV in people.

Agenda

01

Problem Declaration

02

Data Process

03

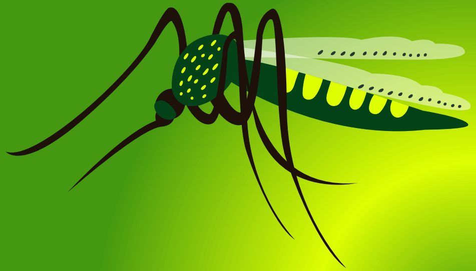
Overview & Data Cleansing

04

Modeling & Evaluation

05

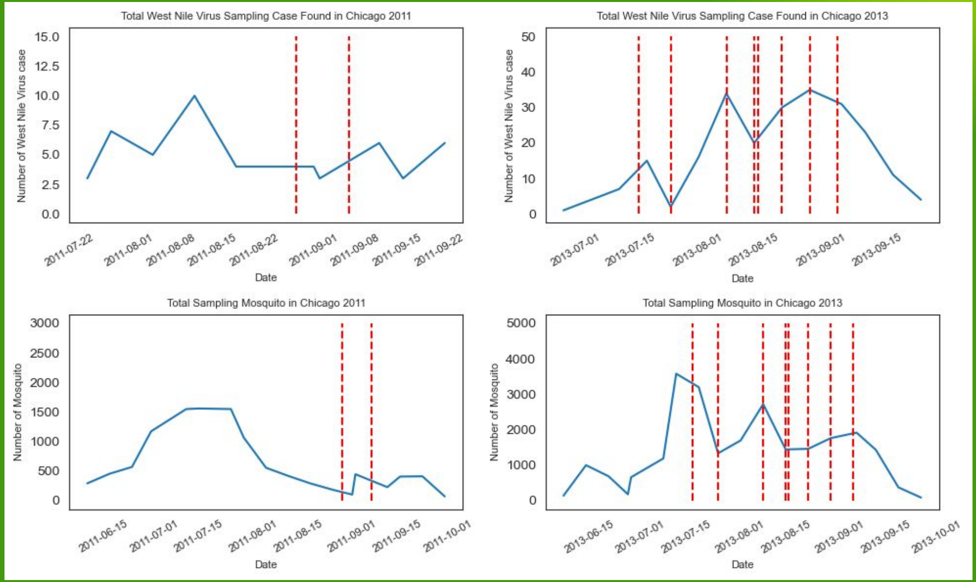
Conclusions & Recommendations



Problem Declaration

“..In 2011 and 2013, in order to prevent West Nile Virus; WNV Infection in Chicago, insecticide sprayings are conducted in some specific areas to reduce the number of mosquitoes.

Anyways, after gaining the number of mosquitoes caught in traps around Chicago city on the spraying days, Government of Illinois are suspected if the Chicago City Council were doing the right way as the number of mosquitoes and the number of West Nile Virus sampling cases still increased even after spraying dates..

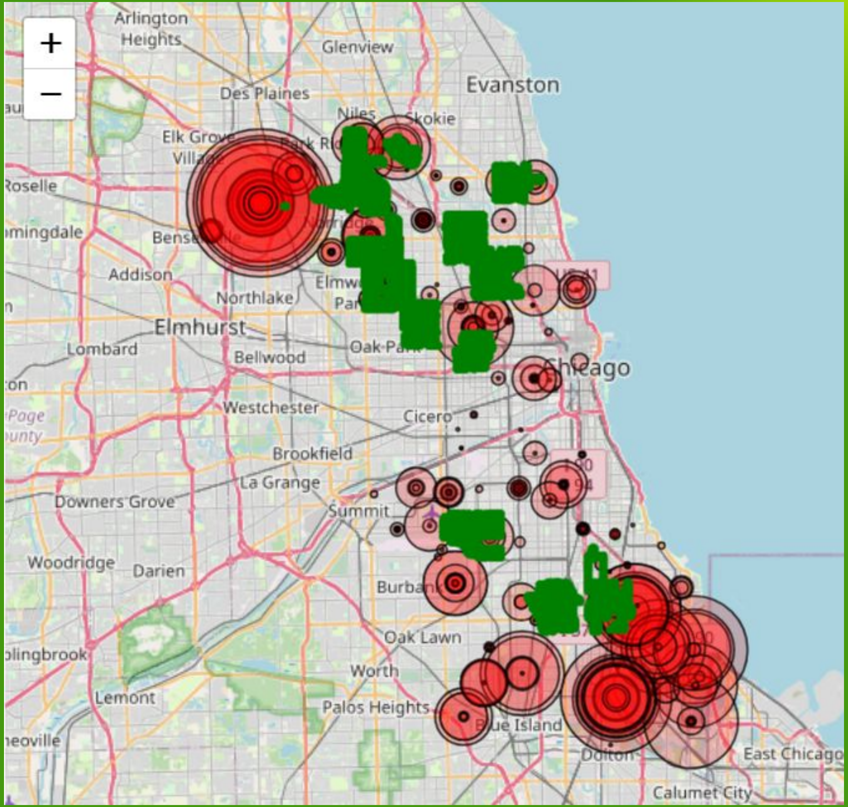


Problem Declaration (Cont.)

Head of Data Science of Government of Illinois found that the spraying location were not matched with the major outbreaking areas so they were curious on how they determined the spraying locations.

After meeting with Chicago City Council, Government of Illinois requested the Head of Data Science and his team to develop a classification model that can notify the risk of West Nile Virus outbreaks in specific geographic regions in Chicago.

In the Data Science team splint planning, the team decided to use daily local weather data to create the classification model after studying the West Nile Virus..”



Data Process

Cleaning Data

Check null and duplicate
data , fix missing value

Working on Features

Create and extract
columns in need

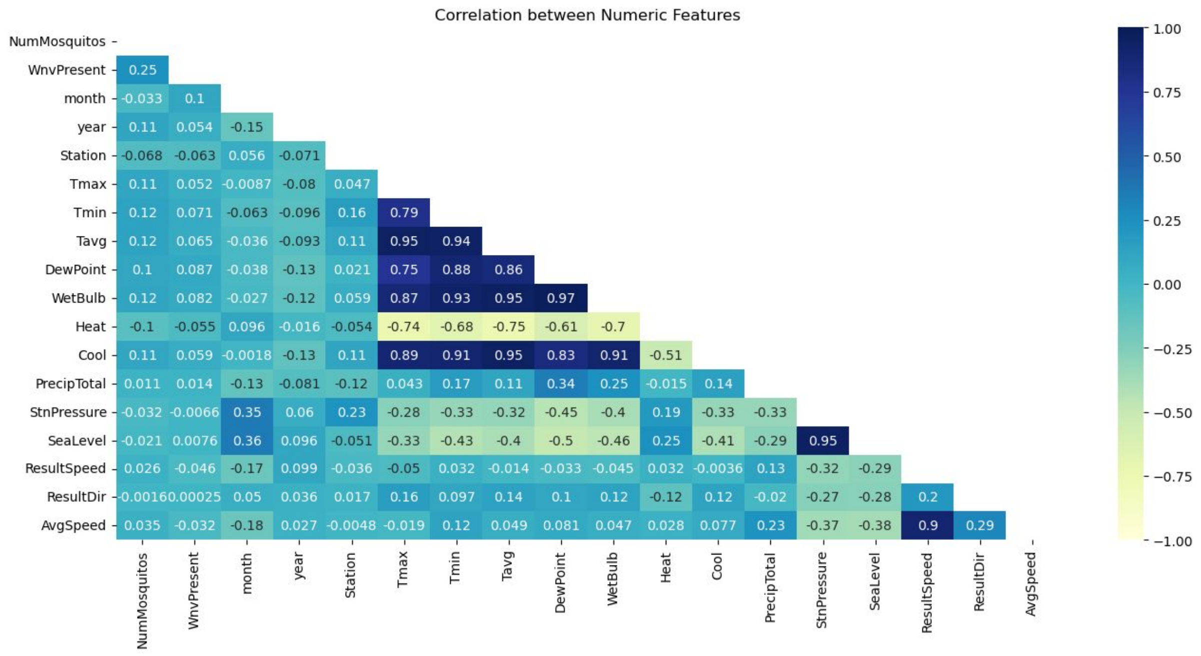
Create Dataset

Selected features for for
Train & Test data

Overview & Data Cleansing

- Observations were collected in 2007, 2009, 2011 and 2013 in aspect of geographical features, presence of West Nile Virus, recording date related to Mosquito trap ID.
- 9,663 as total number of observations after duplicated value removal.
- Merging the above DataFrame with local daily weather dataset by closest climate station and climate recording date as joining keys.
- SMOTE technique is applied to the dataset which haves imbalance of target variable classes (5% - NW Virus positive, 95% - NW Virus negative) when splitting the datasets

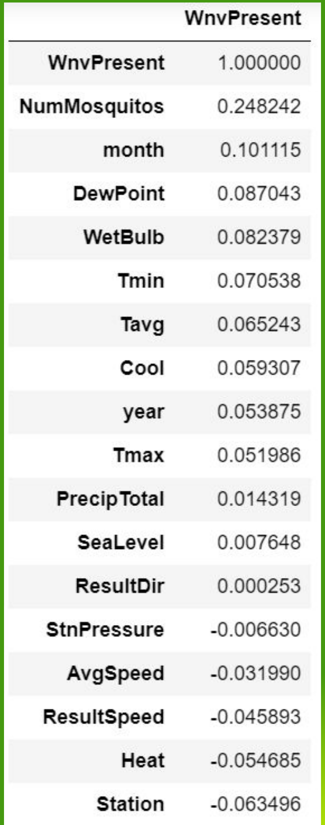
The number of trapped mosquitoes seems to have the strongest positive correlation with the presence of West Nile Virus



Correlation between Numeric Features

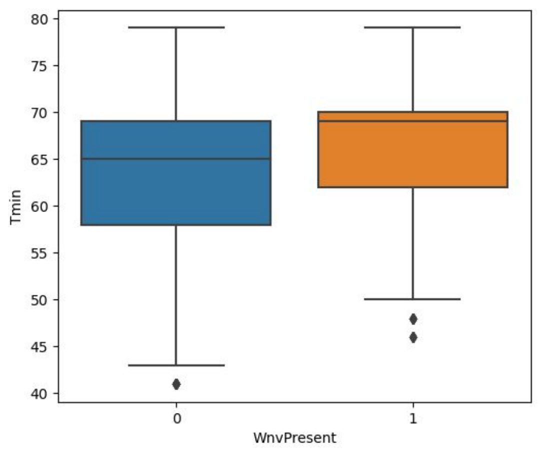
	NumMosquitos	WnvPresent	month	year	Station	Tmax	Tmin	Tavg	DewPoint	WetBulb	Heat	Cool	PrecipTotal	StnPressure	SeaLevel	ResultSpeed	ResultDir	AvgSpeed
NumMosquitos																		
WnvPresent	0.25																	
month	-0.033	0.1																
year	0.11	0.054	-0.15															
Station	-0.068	-0.063	0.056	-0.071														
Tmax	0.11	0.052	-0.0087	-0.08	0.047													
Tmin	0.12	0.071	-0.063	-0.096	0.16	0.79												
Tavg	0.12	0.065	-0.036	-0.093	0.11	0.95	0.94											
DewPoint	0.1	0.087	-0.038	-0.13	0.021	0.75	0.88	0.86										
WetBulb	0.12	0.082	-0.027	-0.12	0.059	0.87	0.93	0.95	0.97									
Heat	-0.1	-0.055	0.096	-0.016	-0.054	-0.74	-0.68	-0.75	-0.61	-0.7								
Cool	0.11	0.059	-0.0018	-0.13	0.11	0.89	0.91	0.95	0.83	0.91	-0.51							
PrecipTotal	0.011	0.014	-0.13	-0.081	-0.12	0.043	0.17	0.11	0.34	0.25	-0.015	0.14						
StnPressure	-0.032	-0.0066	0.35	0.06	0.23	-0.28	-0.33	-0.32	-0.45	-0.4	0.19	-0.33	-0.33					
SeaLevel	-0.021	0.0076	0.36	0.096	-0.051	-0.33	-0.43	-0.4	-0.5	-0.46	0.25	-0.41	-0.29	0.95				
ResultSpeed	0.026	-0.046	-0.17	0.099	-0.036	-0.05	0.032	-0.014	-0.033	-0.045	0.032	-0.0036	0.13	-0.32	-0.29			
ResultDir	-0.0016	0.00025	0.05	0.036	0.017	0.16	0.097	0.14	0.1	0.12	-0.12	0.12	-0.02	-0.27	-0.28	0.2		
AvgSpeed	0.035	-0.032	-0.18	0.027	-0.0048	-0.019	0.12	0.049	0.081	0.047	0.028	0.077	0.23	-0.37	-0.38	0.9	0.29	

1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

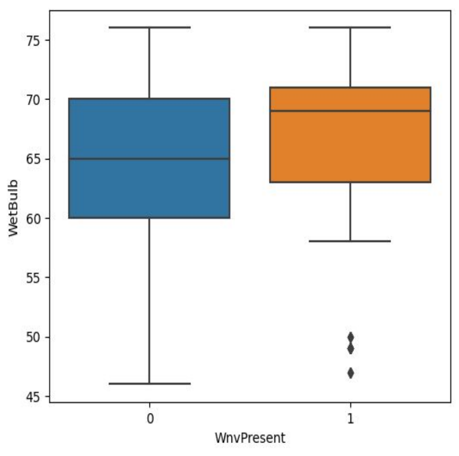


	WnvPresent
WnvPresent	1.000000
NumMosquitos	0.248242
month	0.101115
DewPoint	0.087043
WetBulb	0.082379
Tmin	0.070538
Tavg	0.065243
Cool	0.059307
year	0.053875
Tmax	0.051986
PrecipTotal	0.014319
SeaLevel	0.007648
ResultDir	0.000253
StnPressure	-0.006630
AvgSpeed	-0.031990
ResultSpeed	-0.045893
Heat	-0.054685
Station	-0.063496

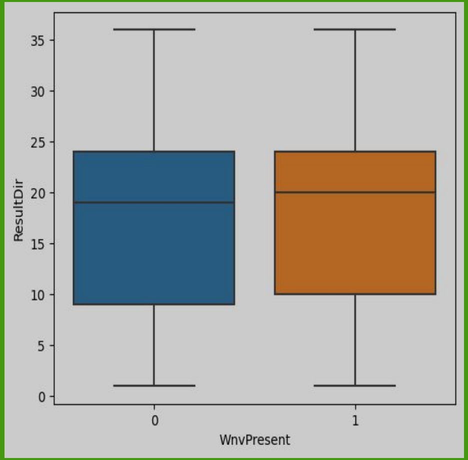
With consideration of the correlation with the presence of the virus, If the distribution of features by the presence of the virus not different, It will not be selected for the classification model.



Selected

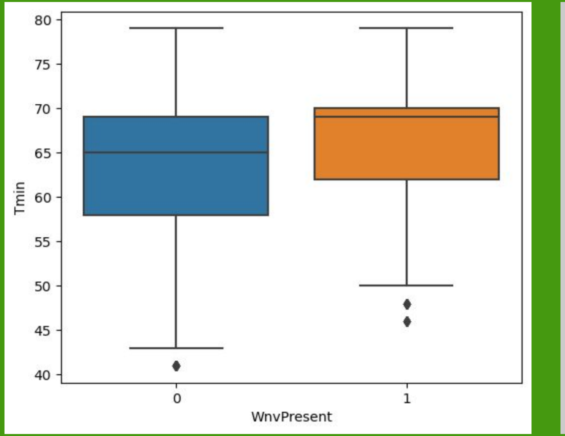


Selected

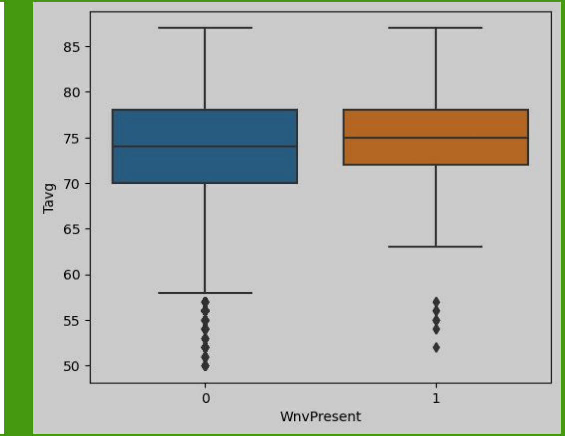


Dropped

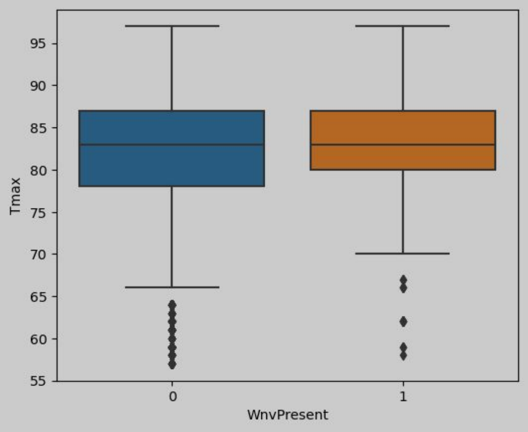
In order the prevent multicollinearity of features, ‘Tmin’ was selected for creating model as stronger correlation with the presence of the virus than ‘Tavg’ and ‘Tmax’,



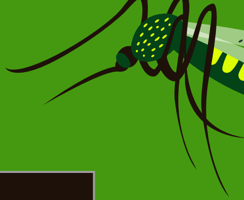
Selected



Dropped



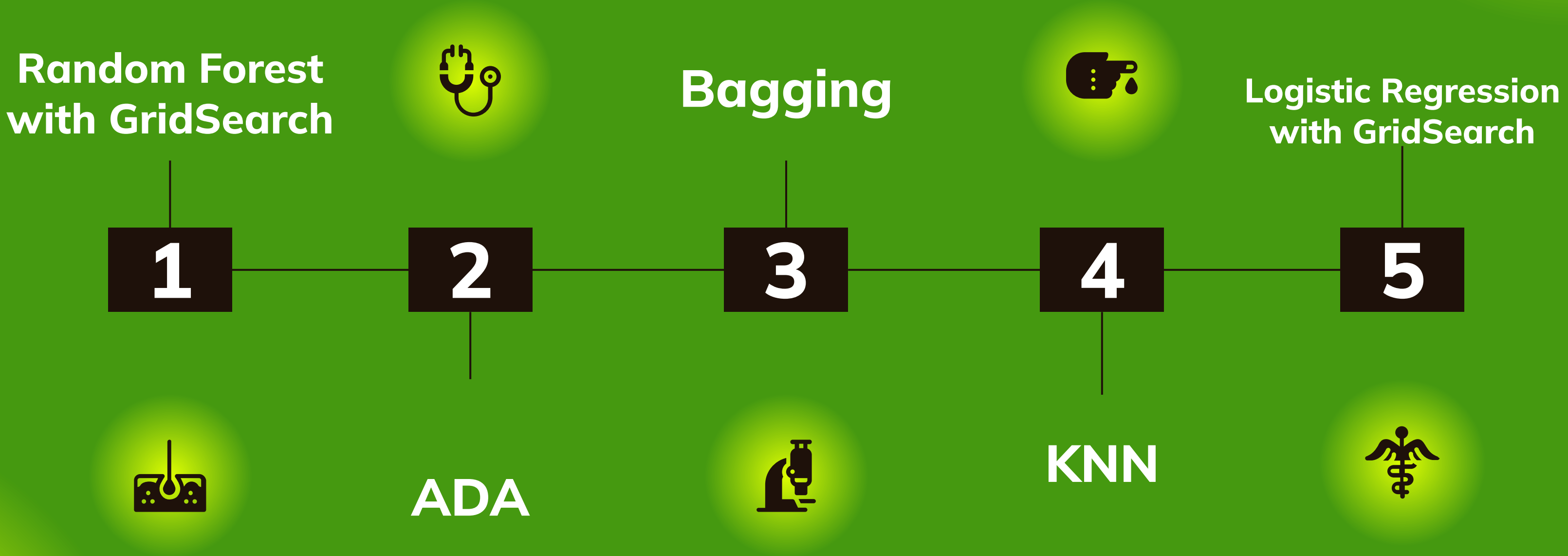
Dropped



Features used for creating classification models.

Features	Description	Data type	Source (dataset)
Species	Species of the trapped mosquitoes	string	train
Trap	Mosquitoes Trap ID	string	train
month	Recording month	int	train
year	Recording year	int	train
Tmin	Minimum temperature of the date	int	weather
Dewpoint	Daily average Dewpoint	int	weather
WetBulb	Daily average Wet Bulb	int	weather
Heat	Heating (season begins with July)	int	weather

Modeling & Evaluation



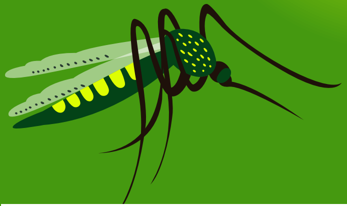
1 Random Forest with GridSearch

2 ADA

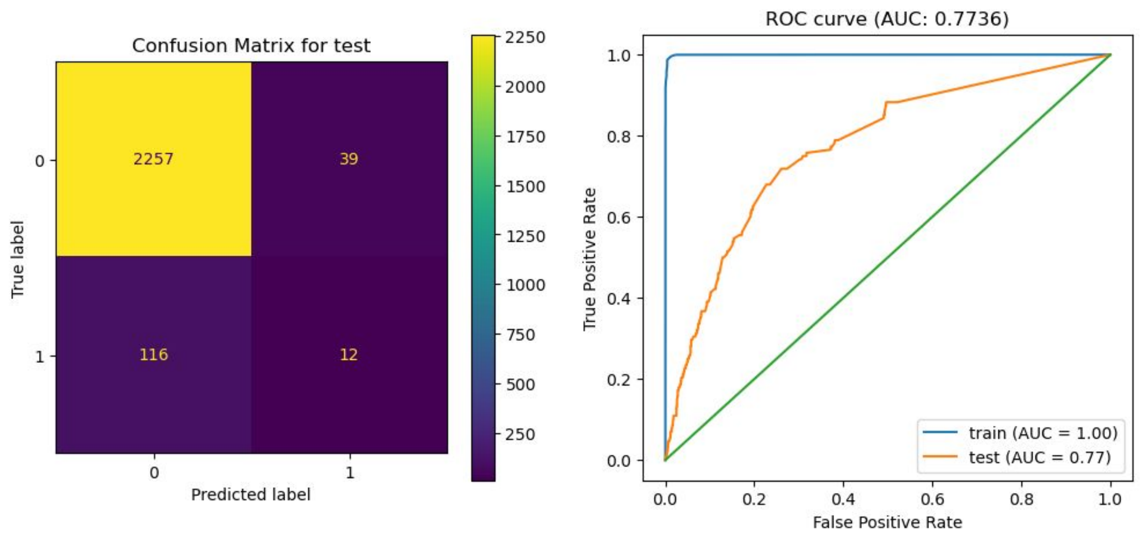
3 Bagging

4 KNN

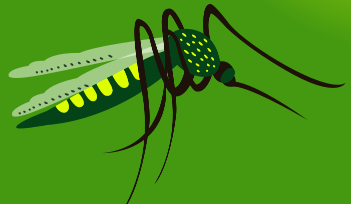
5 Logistic Regression with GridSearch



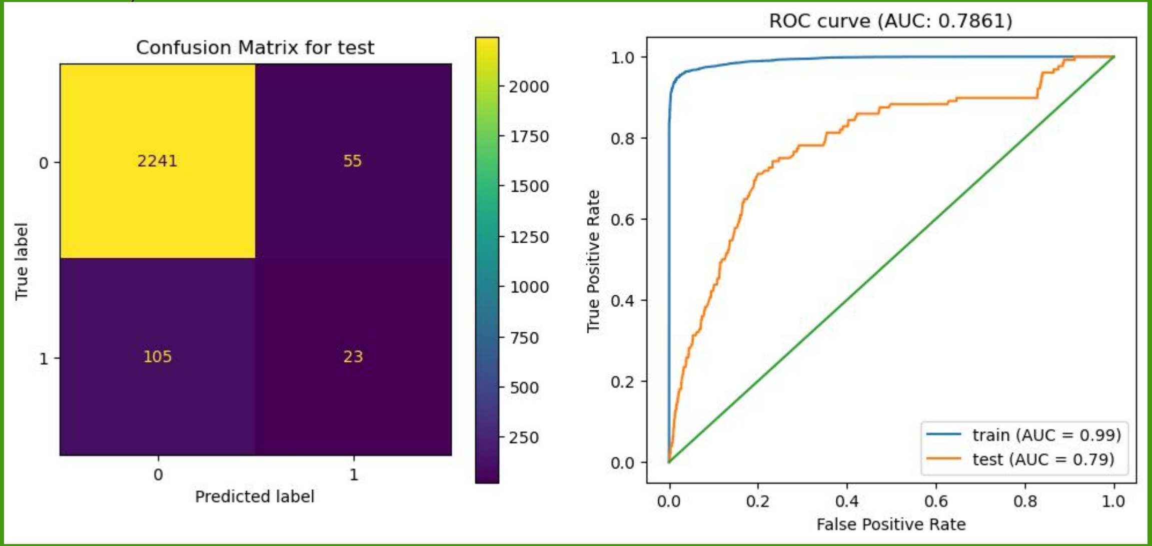
Random Forest with GridSearchCV



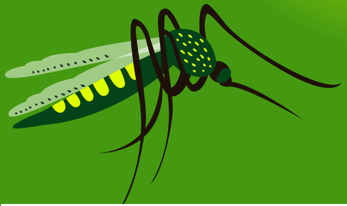
Dataset	Accuracy	Precision	F1	Recall
Train	0.99	0.99	0.99	0.99
Test	0.93	0.24	0.13	0.09



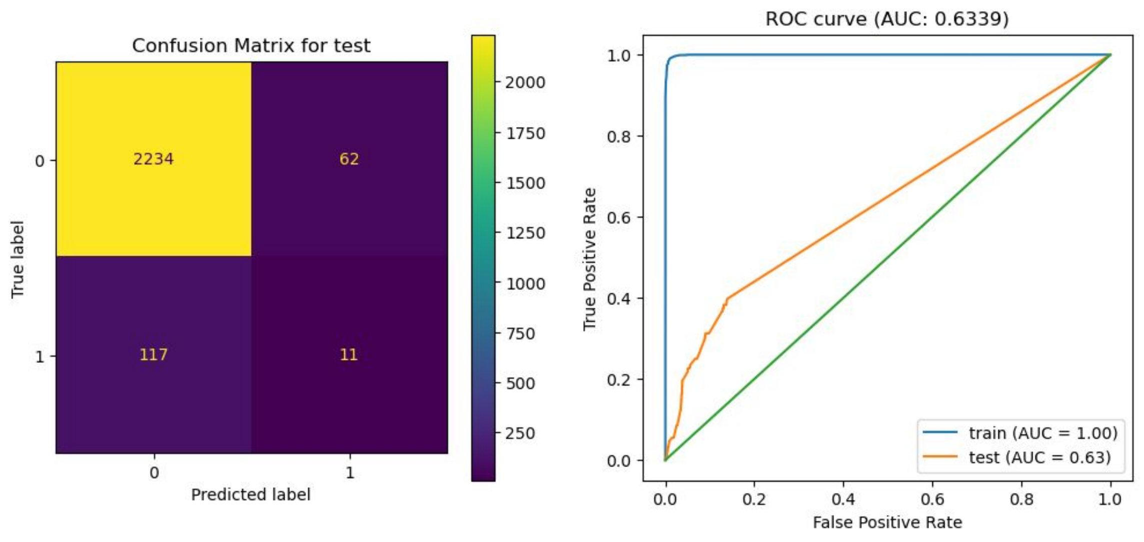
ADA



Dataset	Accuracy	Precision	F1	Recall
Train	0.96	0.97	0.96	0.95
Test	0.93	0.29	0.22	0.18

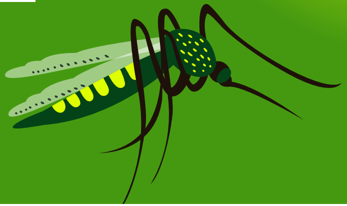


Bagging

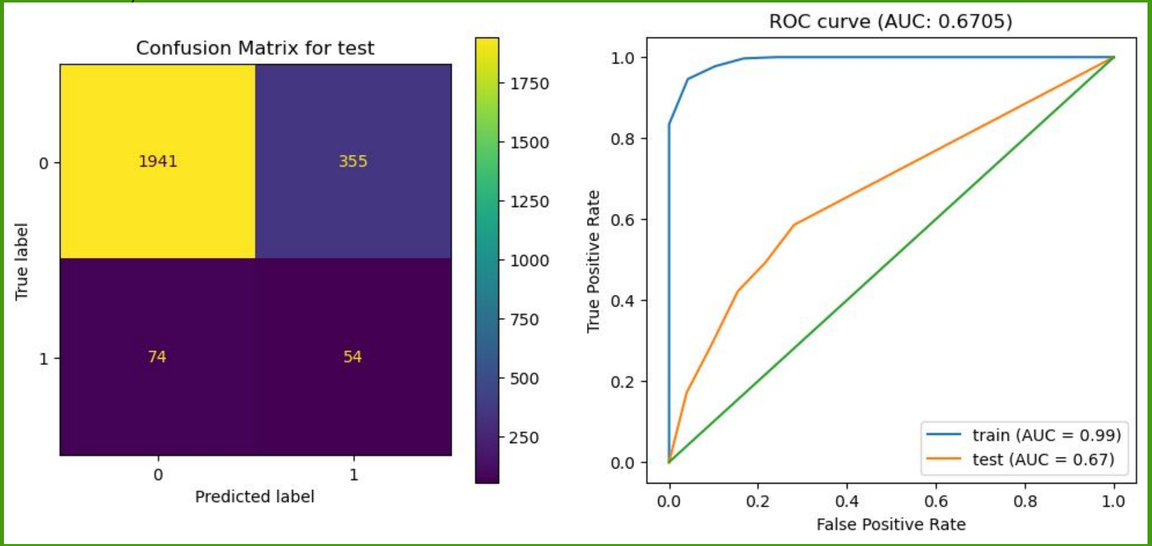


Dataset	Accuracy	Precision	F1	Recall
Train	0.99	0.99	0.99	0.99
Test	0.92	0.15	0.11	0.09

0.98



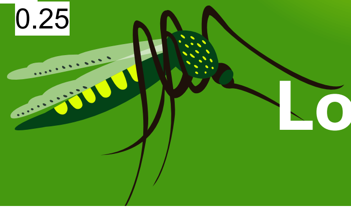
KNN



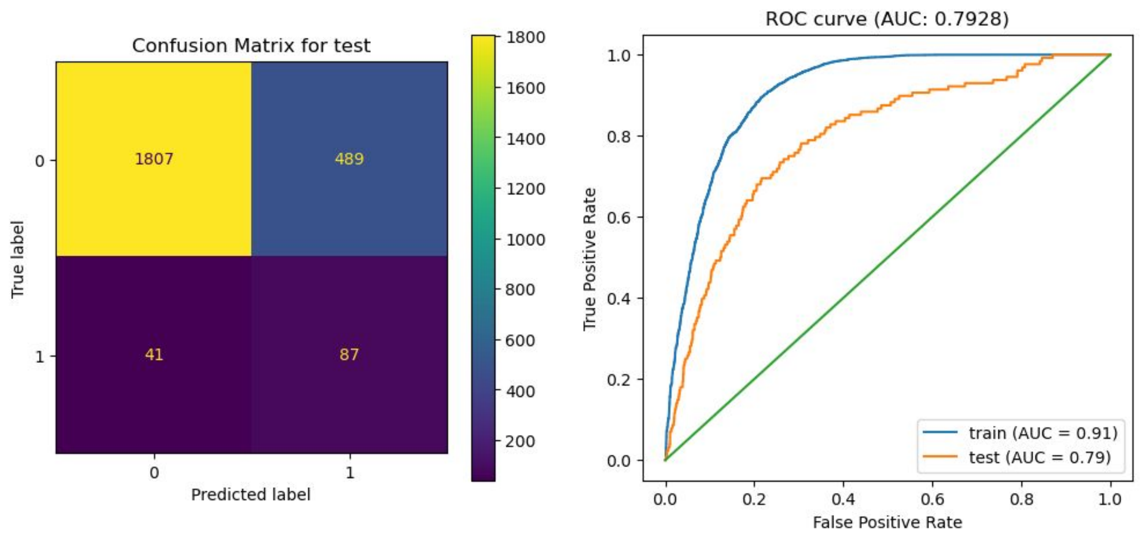
Dataset	Accuracy	Precision	F1	Recall
Train	0.93	0.90	0.94	0.98
Test	0.82	0.13	0.20	0.42

0.25

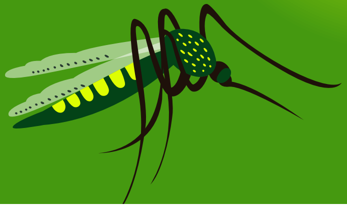
0.25



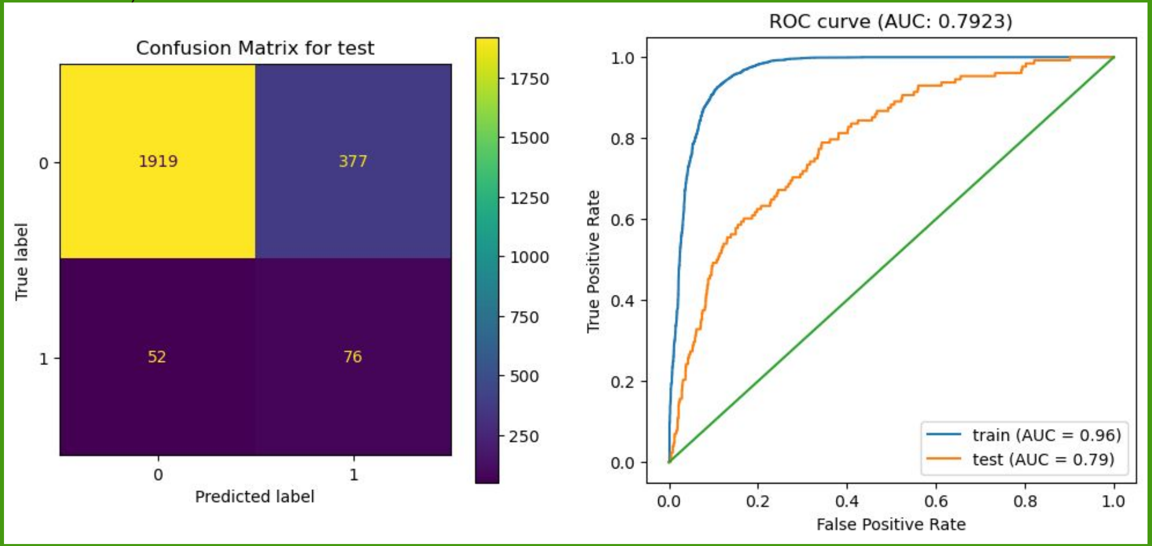
Logistic Regression with GridsearchCV



Dataset	Accuracy	Precision	F1	Recall
Train	0.90	0.81	0.85	0.89
Test	0.79	0.15	0.25	0.68



SVM



Dataset	Accuracy	Precision	F1	Recall
Train	0.90	0.85	0.91	0.97
Test	0.82	0.17	0.26	0.59

Model Evaluation Comparison

Model	Recall score	Precision score	AUC
Random Forest with GridSearchCV	0.09	0.24	0.77
ADA	0.18	0.29	0.79
Bagging	0.09	0.15	0.63
KNN	0.42	0.13	0.67
Logistic Regression with GridsearchCV	0.68	0.15	0.79
SVM	0.59	0.17	0.79

Conclusions & Recommendations

After creating 6 models, the data science team decided to deliver Logistic Regression with GridSearchCV Model, as a first proposal to Government of Illinois, which is able to detect 68% of total West Nile Virus presence on unseen dataset. In addition, from the predicted results of the model, only 15% of the predicted presence of West Nile Virus is actually existed, which might lead to over-budgeting on hospitality expense in Chicago. If Government of Illinois approved to use this model, Chicago City Council also needs to find countermeasures for operation cost-saving.

As for spray that not effective might change to use Aerial spraying that is process by Airplanes and helicopters to treat very large areas with larvicides that kills mosquito larvae that hatch from eggs or adulticides to quickly kills flying mosquitoes.Both larvicides and adulticides can temporarily help reduce the number of mosquitoes in an area.

Cost - Benefit

Early intervention can help prevent the spread of the virus, reducing the overall cost of responding to a widespread outbreak and lowers the burden on healthcare systems.This can lead to sustained cost savings in public health expenditures related to WNV surveillance, treatment, and response in the future.

Data Team at Government of Illinois :

Ms.Plaii
Analytics & Visualisation

Mr.Mhor
Code Hacker

Mr.Win
Scrum Master