Temporal projection of natural atmospheric events in the United States Regions

Haejeong Choi, Yanyu Long, Seung Ho Woo

1 Introduction

The U.S. experiences a variety of extreme weather events, including hurricanes, floods, blizzards, droughts, and so on.

DISASTER TYPE	EVENTS	PERCENT FREQUENCY	EVENTS/YEAR	COST/YEAR	DEATHS/YEAR
Drought	28	9.6%	0.7	$6.2B	93
Flooding	33	11.3%	0.8	$3.6B	15
Freeze	9	3.1%	0.2	$0.7B	4
Severe Storm	132	45.4%	3.1	$7.0B	42
Tropical Cyclone	52	17.9%	1.2	$24.1B	157
Wildfire	18	6.2%	0.4	$2.5B	9
Winter Storm	19	6.5%	0.5	$1.2B	28
All	**291**	**100.0**	**6.9**	**$45.4B**	**348**

”Billion-dollar” events to affect the United States from 1980 to 2021. Data source: NOAA.

1 Introduction

Study the relationship between weather conditions and the occurrences of extreme weather events in the U.S.

Identify the most relevant indicators and the best-performing classification algorithms to predict natural disasters based on daily weather statistics.

Response

NOAA Storm Event Database

- Extreme weather events such as drought, flood, wildfire, hurricane, etc. at the county level
- **ID: year-month, county**

Predictors

Historical weather data

- Hourly observations of temperature, humidity, pressure, wind direction, wind speed, etc.
- **ID: year-month, city**

Compute the number of episodes observed in each county in each month from 2012 to 2017.

Original data

EVENT_ID	Date	STATE	COUNTY	EVENT_TYPE
678791	20170406	NEW JERSEY	GLOUCESTER	Thunderstorm Wind
679228	20170406	FLORIDA	LEE	Tornado
679268	20170405	OHIO	GREENE	Thunderstorm Wind
682042	20170416	OHIO	CLERMONT	Flood

Reshaped data

state	county	ym	num_episodes
ALABAMA	AUTAUGA	201202	1
		201203	1
		201204	1
		201205	1
		201207	3
		201212	1

Note: An episode may contain several different but related events.

Convert hourly records to monthly summary statistics to match with the storm event dataset and deal with missing values.

Hourly records

datetime	Portland
2012-10-01 13:00:00	81
2012-10-01 14:00:00	80
2012-10-01 15:00:00	80
2012-10-01 16:00:00	80
2012-10-01 17:00:00	79

Daily average

date	city	humidity
2012-10-01	Portland	78.72727
2012-10-02	Portland	65.83333
2012-10-03	Portland	66.20833
2012-10-04	Portland	51.16667
2012-10-05	Portland	40.39130

Monthly mean/sd

ym	city	humidity_avg
2012-10	Portland	72.68403
2012-11	Portland	83.52227
2012-12	Portland	86.07985
2013-01	Portland	81.90679
2013-02	Portland	81.29739

Example: humidity (%)

We match the storm event records to the weather data in the previous month, and obtain a relatively balanced dataset.

ym	county	state	num _episodes	meantemp _avg	difftemp _avg	humidity _avg	...
2012-10	Bernalillo	New Mexico	0	287.0911	13.988084	29.01159	...
2012-10	Fulton	Georgia	0	289.5540	9.677601	71.63047	...
2012-10	Suffolk	Massachusetts	1	286.0031	8.069497	73.28394	...
2012-10	Mecklenburg	North Carolina	1	288.7042	10.650447	70.75719	...
2012-10	Cook	Illinois	1	284.6265	8.173461	62.01605	...
2012-10	Dallas	Texas	0	292.1314	10.156523	61.92640	...

(negative:positive = 948:726)

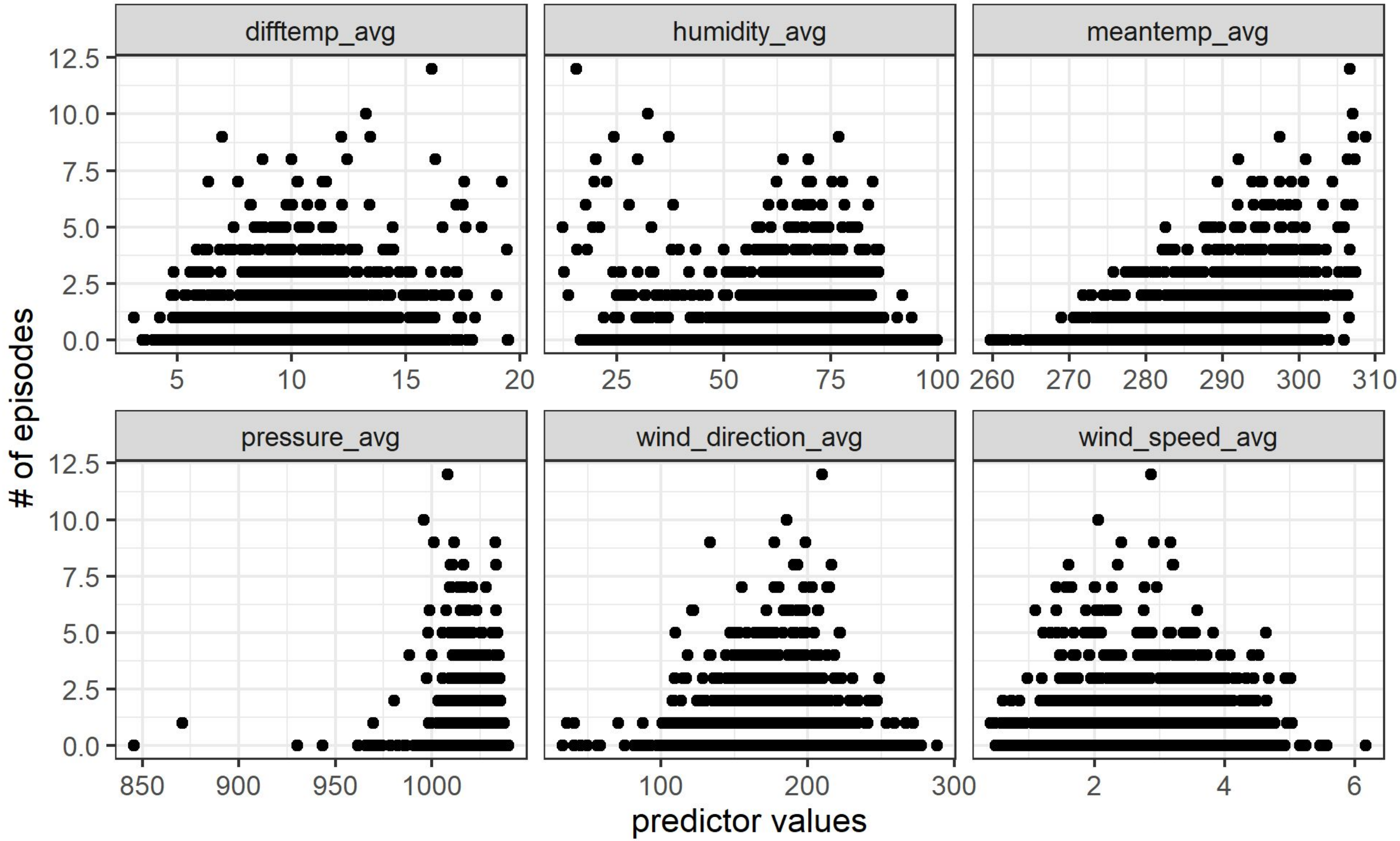
Scale and Encode

- **Categorical Variable(State): LabelEncoder()**
- **Numerical Variable(Others): MinMaxScaler()**

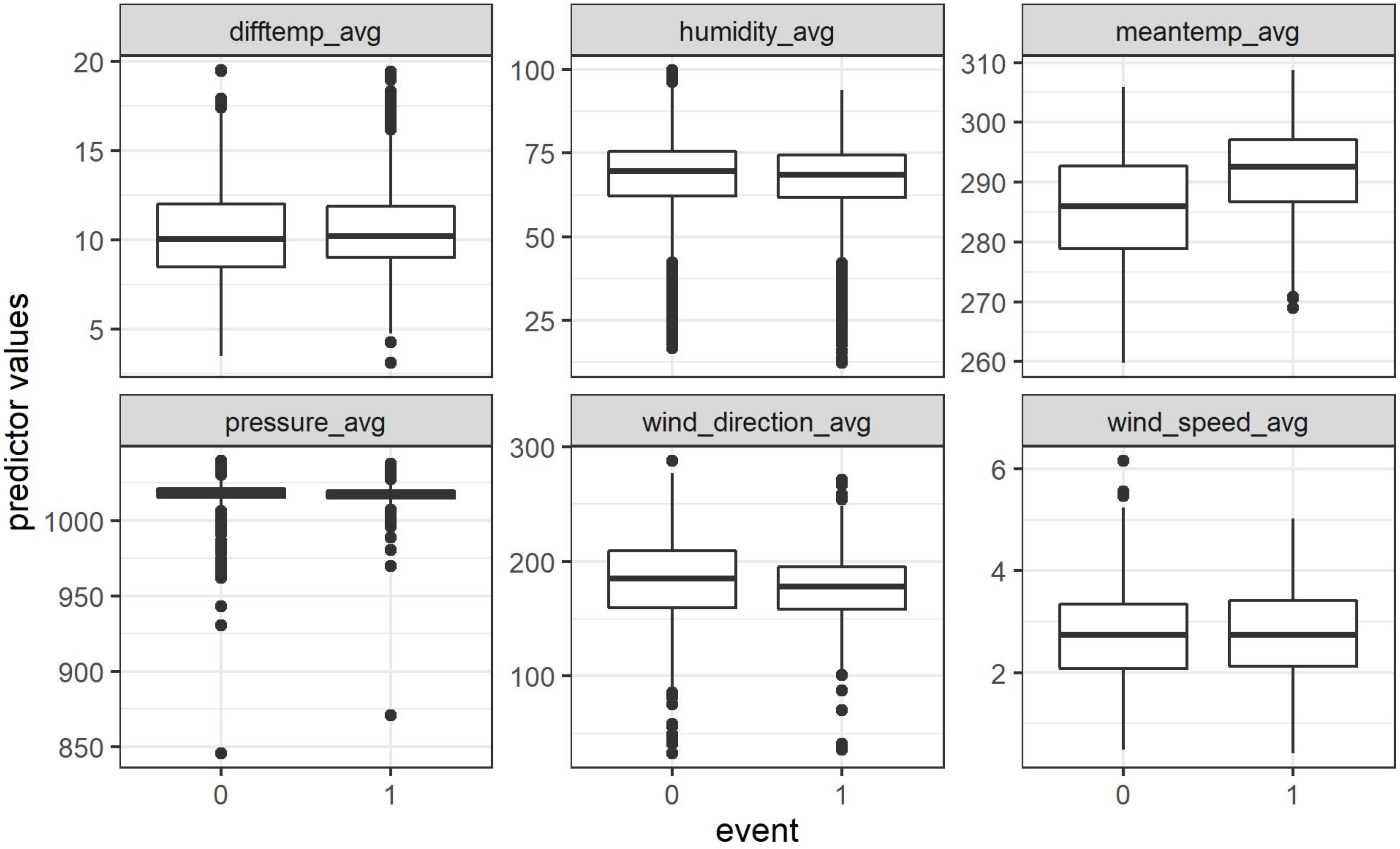
StandardScaler(): worst

Overall RobustScaler() was better than MinMaxScaler() for most of models, but for Gaussian Process Classifier, MinMaxScaler() performed better.

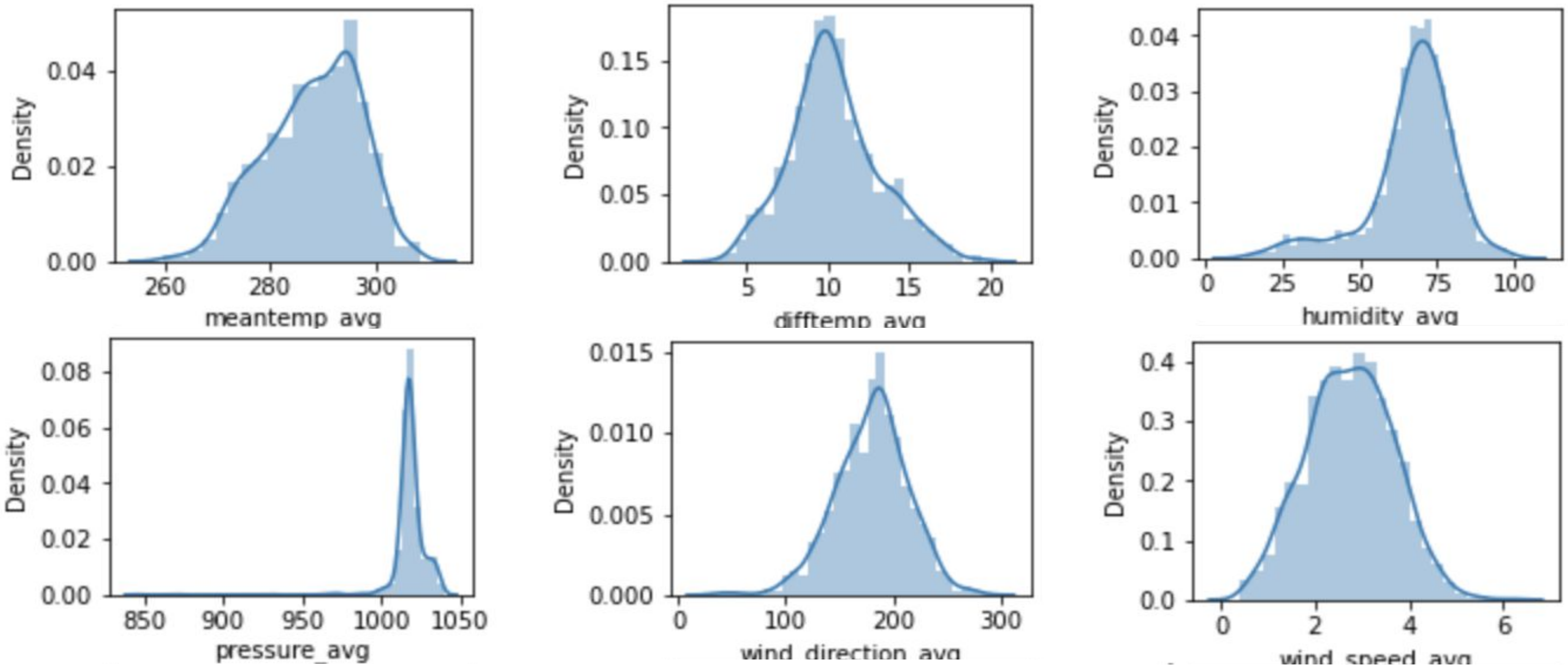
3 Exploratory data analysis



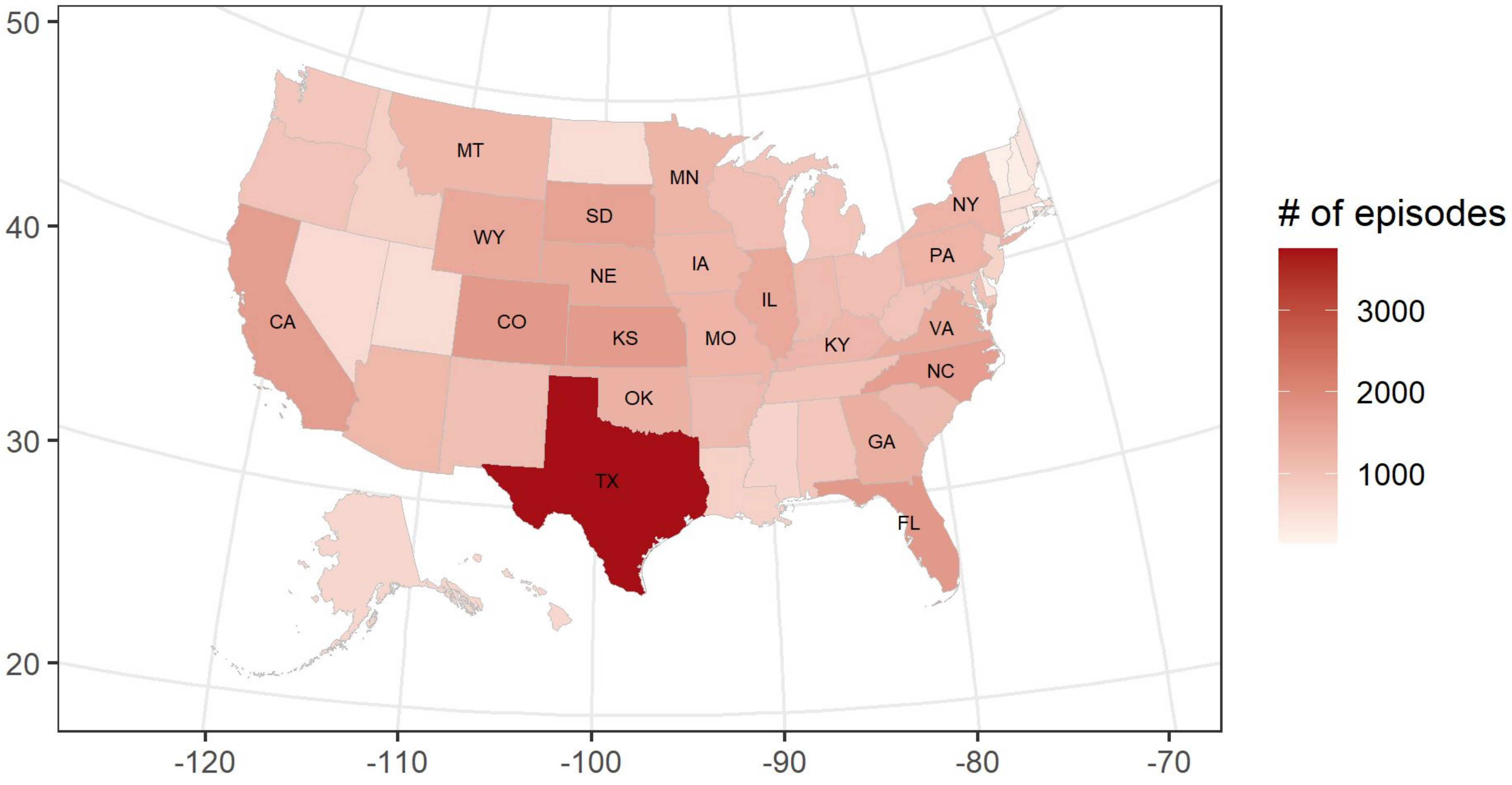
3 Exploratory data analysis



3 Exploratory data analysis



3 Exploratory data analysis



Cumulative number of extreme weather events (episodes) in each state, 2012 - 2017

Models with default parameters(Test Accuracy)

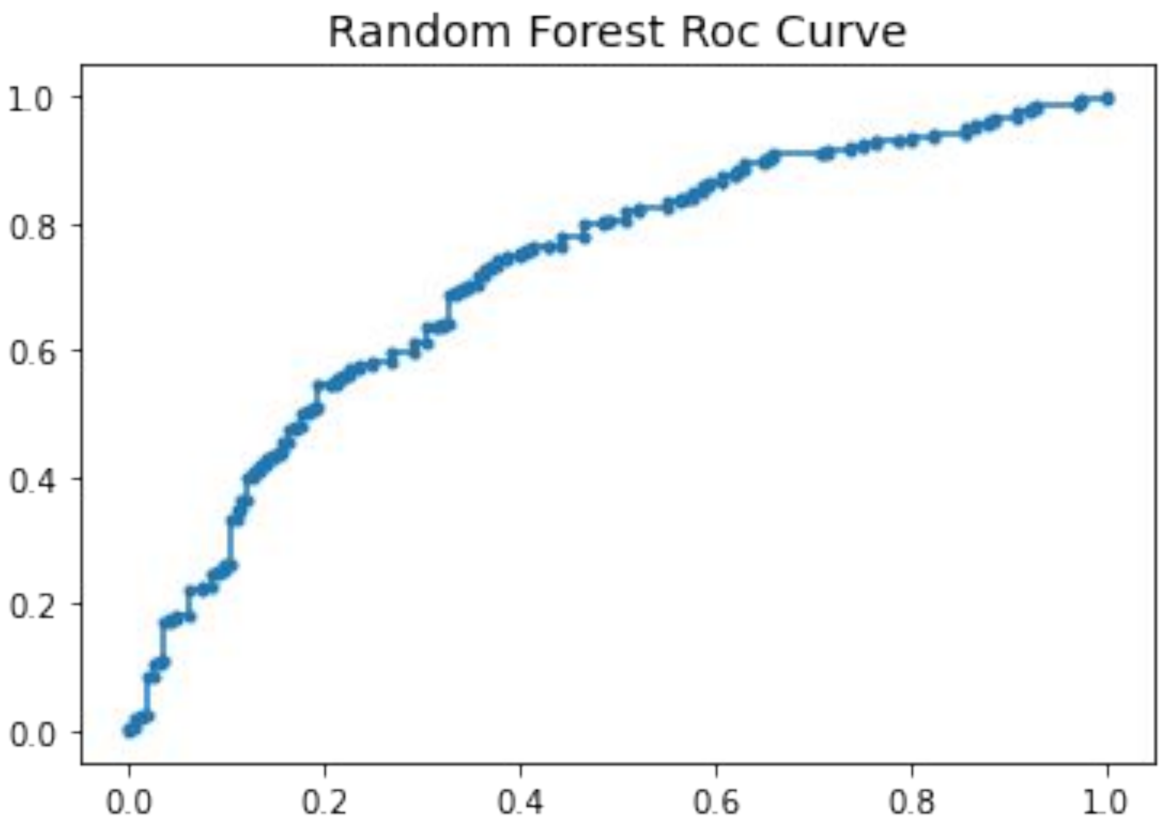
Classifier	Test Accuracy
Gaussian Process Classifier	68.11%
RBF-Support Vector Machine	66.43%
Random Forest	65.95%
Neural Networks	64.75%
Nearest Neighbors	64.51%
Naive Bayes	63.55%
AdaBoost	63.31%
Decision Tree	62.35%
Linear Support Vector Machine	60.43%

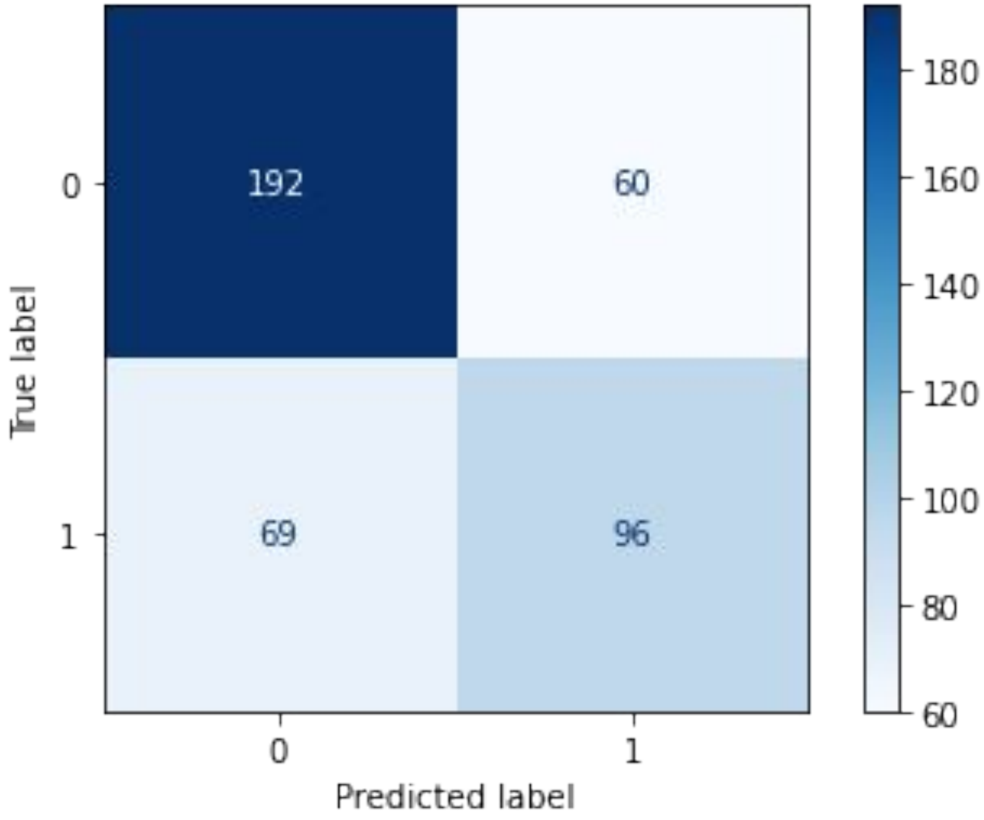
Random Forest Parameter tuning

- RandomizedSearchCV()
- Best Parameters: {'n_estimators': 800, 'min_samples_split': 15, 'min_samples_leaf': 2, 'max_features': 'sqrt', 'max_depth': 40, 'criterion': 'gini', 'bootstrap': True}

Train Accuracy	91.87%
Test Accuracy	69.06%
AUC	0.7230
BER	0.3281

Random Forest - Result



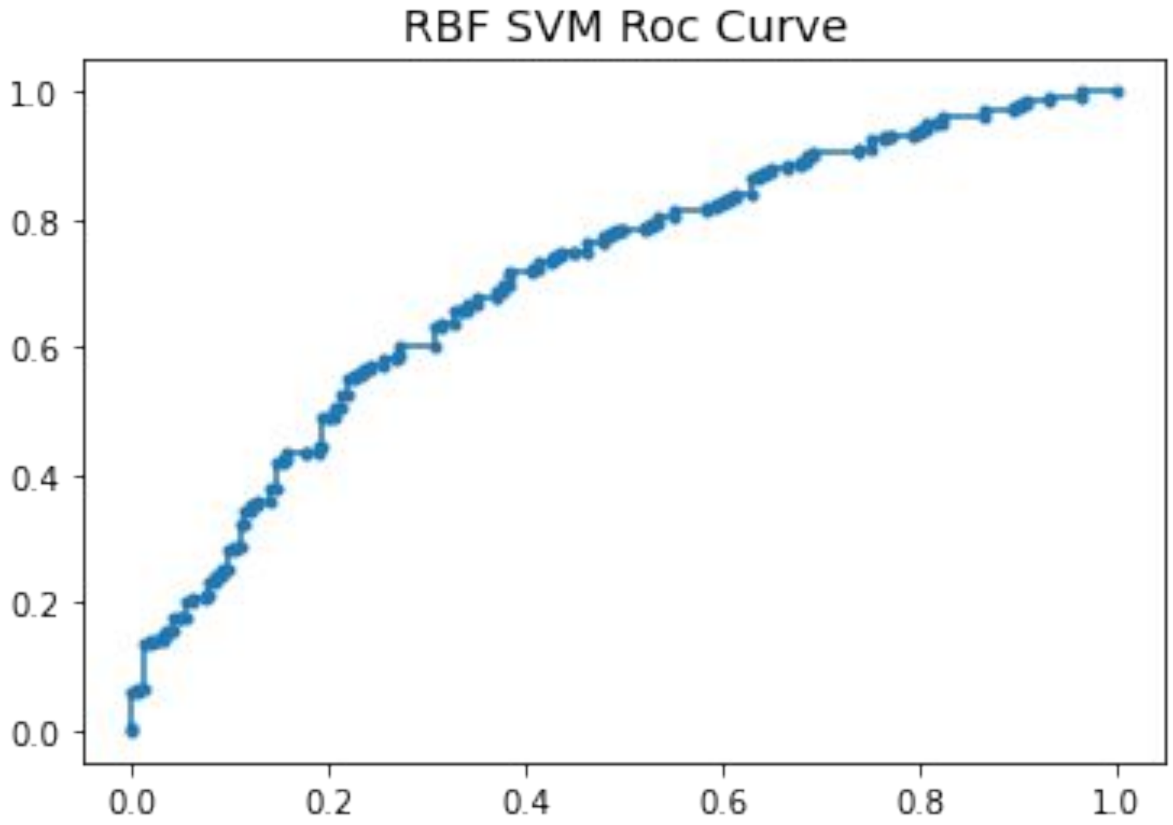


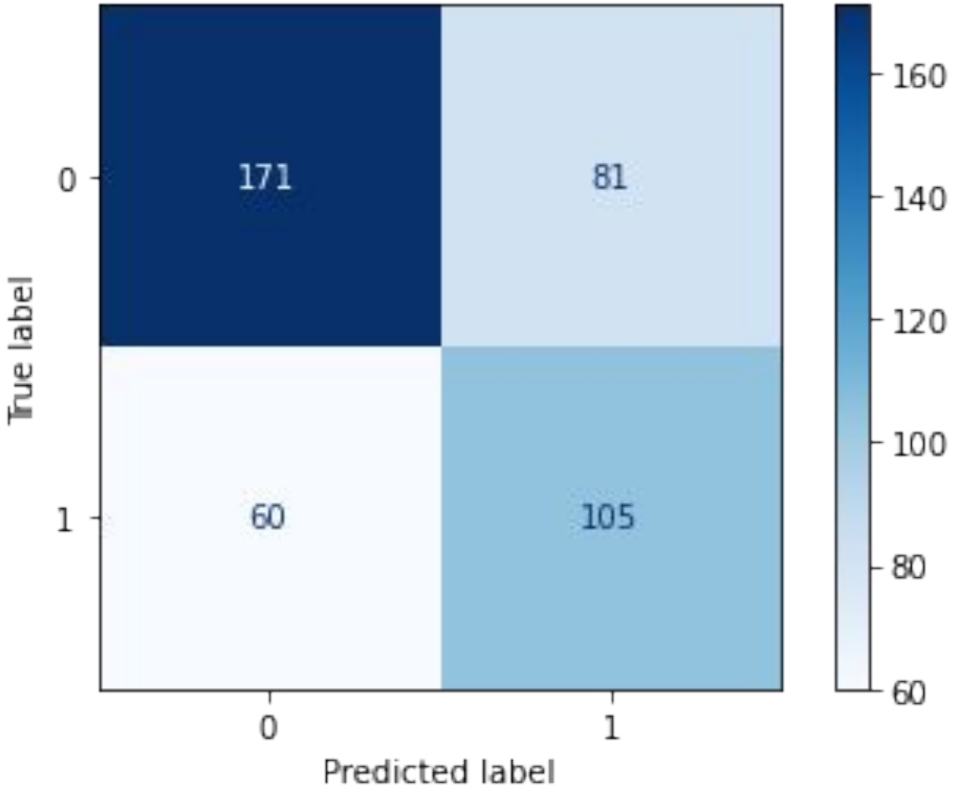
RBF SVM Parameter tuning

- **RandomizedSearchCV()**
- **Best Parameters: (gamma=1, C=5, kernel='rbf')**

Train Accuracy	**73.92%**
Test Accuracy	**66.19%**
AUC	**0.7102**
BER	**0.3425**

RBF SVM - Result



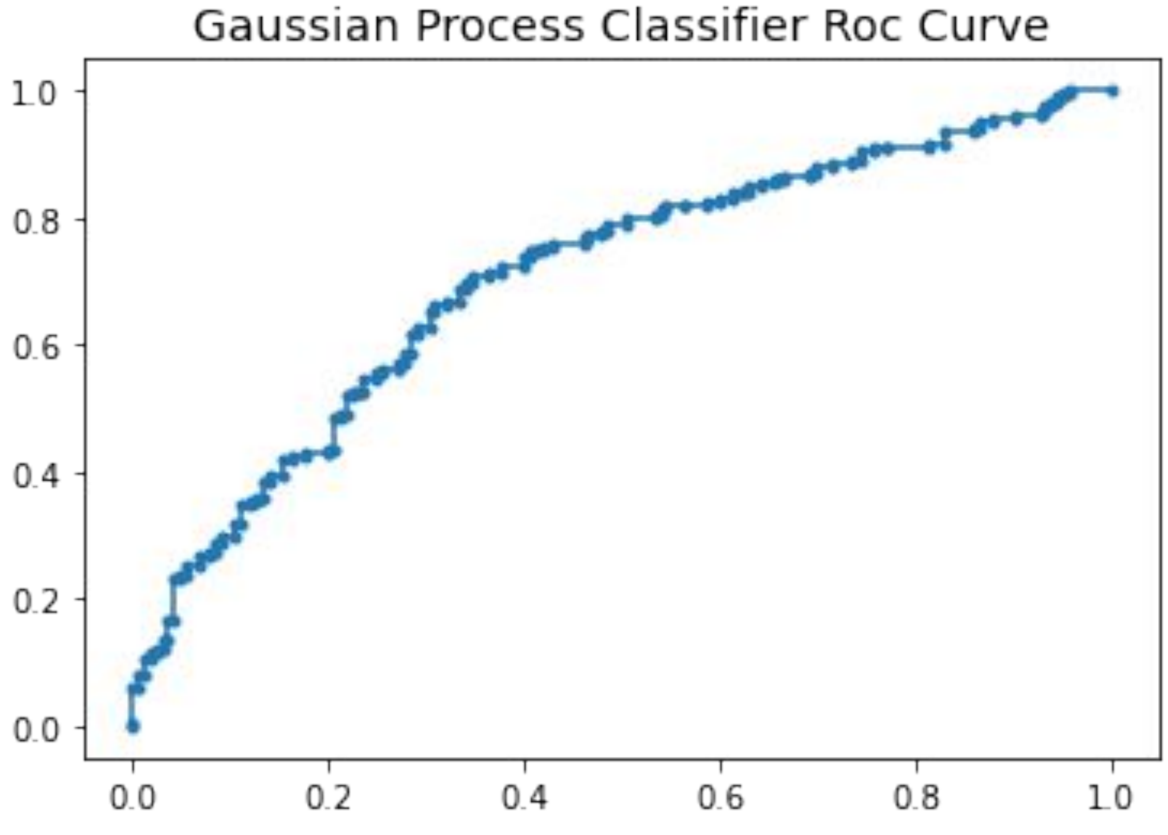


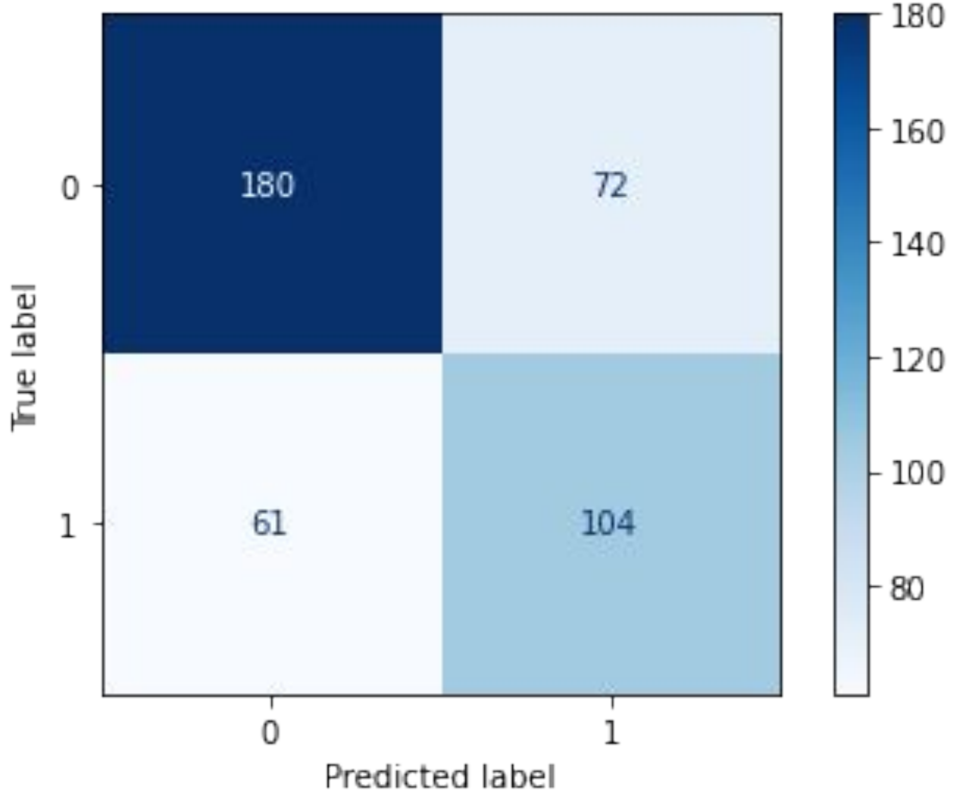
Gaussian Process Classifier Parameter tuning

- **Best Parameters: kernel = 1.0 * RBF(1.0)**
- **Radial Basis Function (RBF) kernel = Gaussian kernel**

Train Accuracy	**73.68%**
Test Accuracy	**68.11%**
AUC	**0.7078**
BER	**0.3277**

Gaussian Process Classifier - Result





Thank you

Haejeong Choi, Yanyu Long, Seung Ho Woo