F21DL –

DATA MINING AND MACHINE LEARNING

PREDICTING AND ANALYSING DAMAGE FROM HURRICANES AND TYPHOONS

Presented By:

Joseph William Abdo

Thomson Thomas

Lubna Gulnaar

Irin Varughese

Joepaul Ettumanookaran

OBJECTIVES

01 Prediction:

Analysing numerical data to predict the formation and trajectory of hurricanes and typhoons, focusing on wind speeds, directions, and pressure

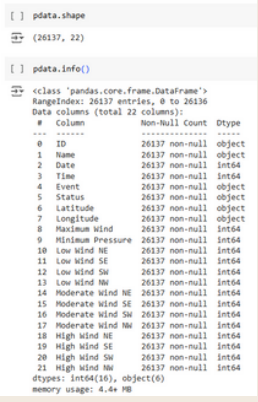
02 Damage Analysis:

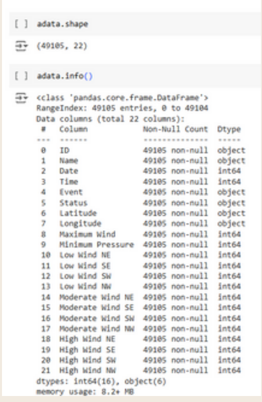
Using satellite images to classify areas as "damage" or "no damage" to assess the extent of destruction to buildings, infrastructure, and landscapes

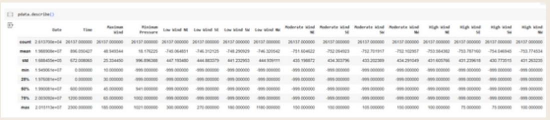
EXPLORATORY DATA ANALYSIS

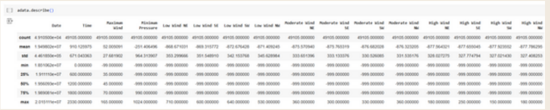
- The numerical dataset (loaded as a .CSV file) was analyzed using .info(), .describe(), and .shape() to examine its structure, data types, and key statistics such as value ranges and missing data
- Rows with missing values were either removed or replaced to ensure data completeness and maintain dataset quality

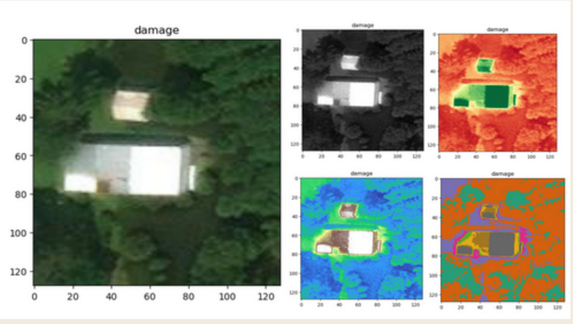
- The image dataset was preprocessed by normalizing and resizing the images
- Random images from both the "damage" and "no damage" sets were plotted to verify the quality and correctness of the processed data.
- Colormap application was tested to enhance feature visibility

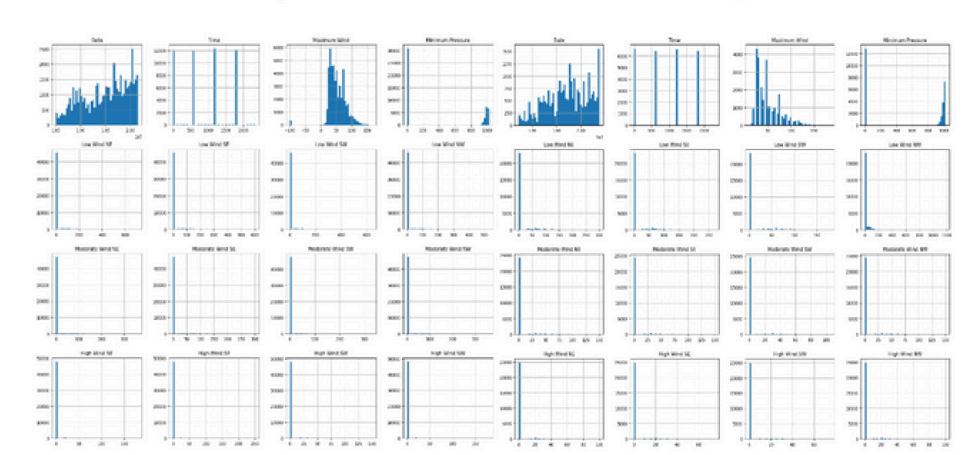




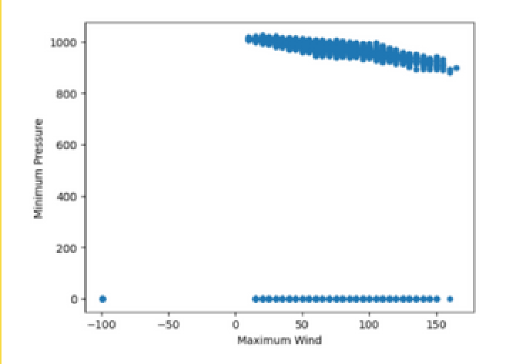


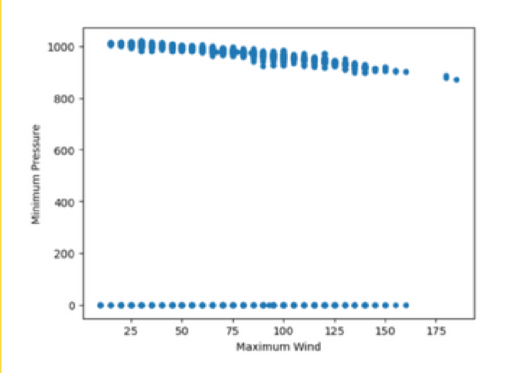




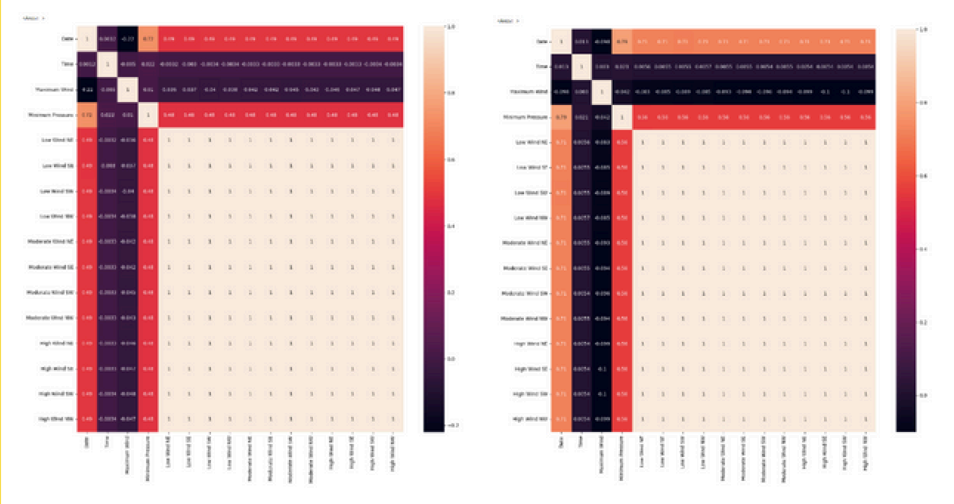


HISTOGRAMS

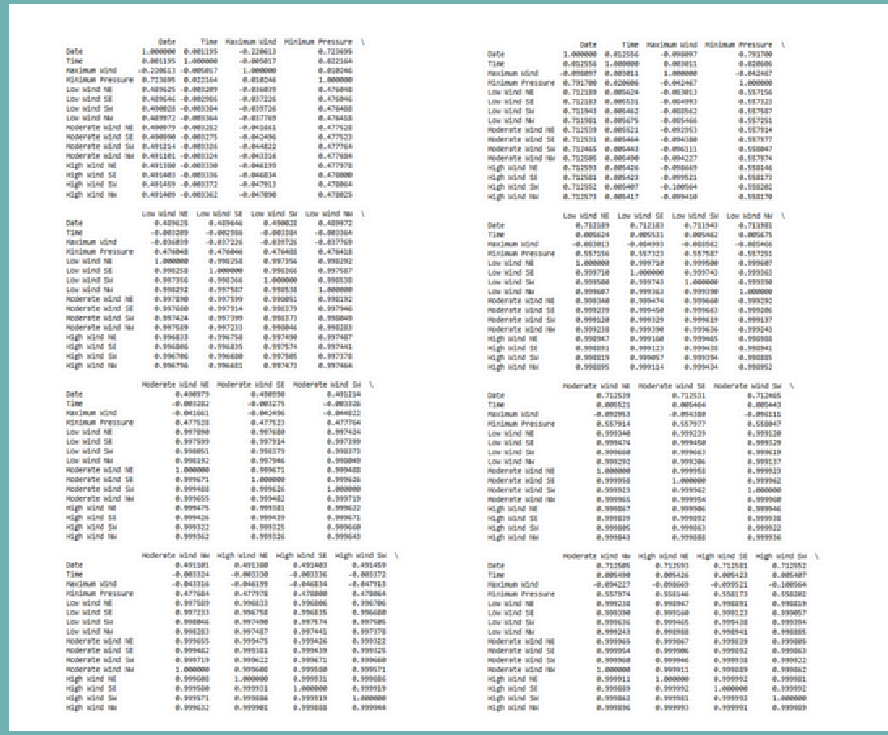




SCATTER PLOTS



HEAT MAPS

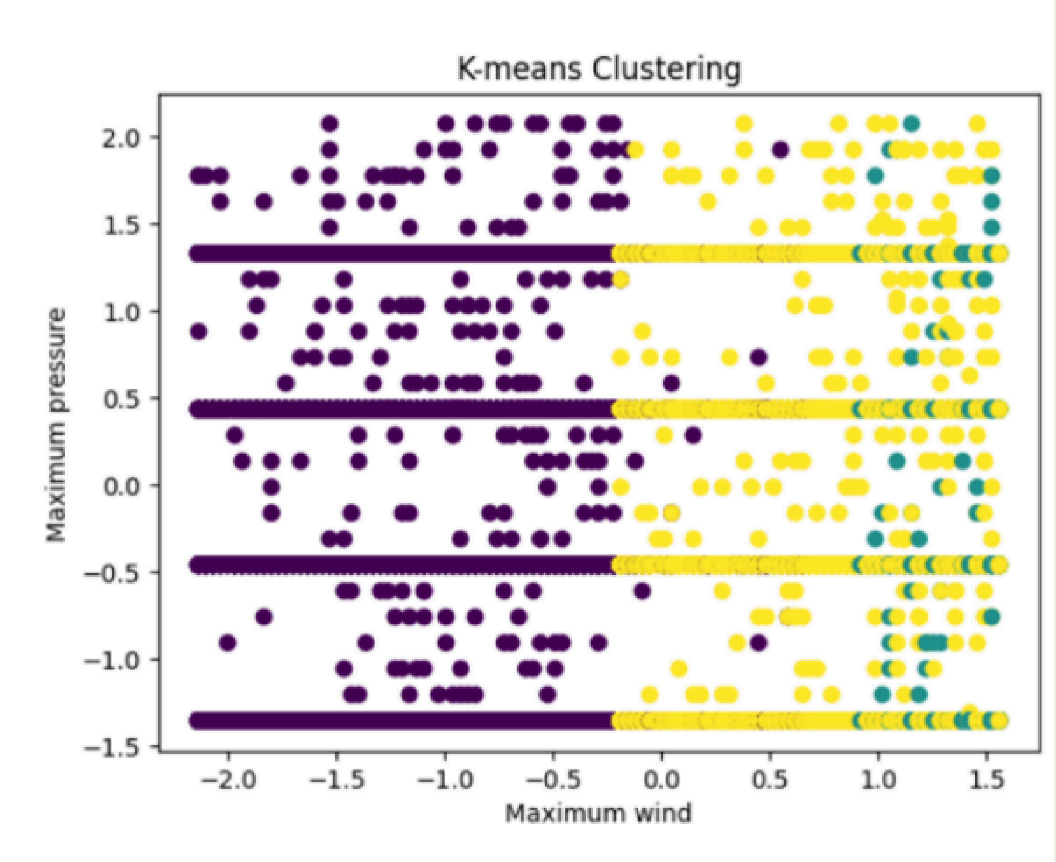


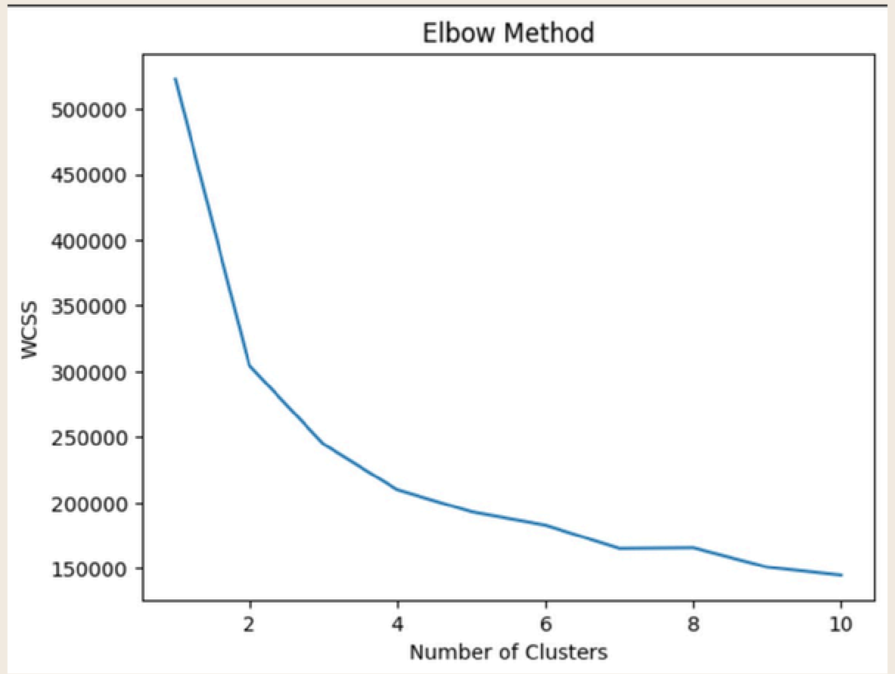
CORRELATION MATRIX

CLUSTERING USING K-MEANS

For clustering the K-Means algorithm with k=3 was used. The clusters showed distinct groupings representing low-intensity, moderate-intensity, and high-intensity storms.

- The ones marked in purple represent hurricanes with moderate wind speeds and pressure. These likely represent average hurricanes.
- Yellow represents hurricanes with higher wind speeds and lower pressure, which likely represents high intensity hurricanes.
- Blue represents hurricanes with lower wind speeds and higher pressures, which are likely to represent low intensity hurricanes.





- Choosing 3 clusters (K=3) was based on the analysis of the Elbow Method , which indicated that three distinct groups best represent the data's structure. This clustering helps in capturing the complexity of hurricane categories without overfitting.
- The clustering shows clear groupings, which could indicate distinct categories or intensities of hurricanes.

BASELINE TRAINING & NEURAL NETWORK

NAIVE BAYES

- The dataset was tested using two configurations—50-50 and 70-30 splits.
- GNB calculated mislabelled points by comparing predicted labels (y_pred) with actual labels (y_test).
- A bar graph was used to compare results from both splits, assessing model performance.

CNN

- A CNN was chosen for image processing because it recognizes patterns effectively
- The dataset was split & the model was trained on segmented images
- Images were processed through multiple layers, extracting key features and connecting them via fully connected layers
- The process was repeated to enhance performance

KNN

- KNN classified hurricane data into Atlantic (0) and Pacific (1) regions, using a 70/30 train-test split
- The KNN classifier was then trained on the selected features, including Maximum Wind, Minimum Pressure, and Low Wind NE, with the number of neighbors set to 3
- The model was evaluated by calculating its accuracy score on the test set and making predictions using new data points

LINEAR REGRESSION

- Linear Regression analyzed the relationship between Minimum Pressure (feature) and Maximum Wind (target)
- The date column was converted to datetime and set as the index for time-based analysis
- Minimum Pressure was considered as the independent variable (feature) and Maximum Wind as the dependent variable (target)

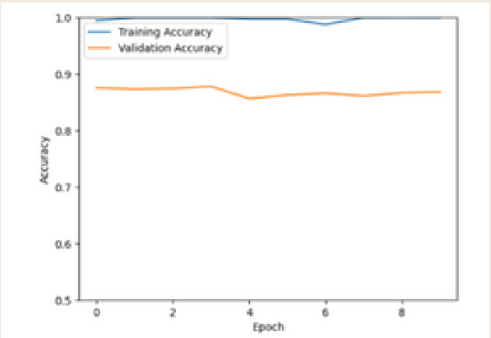
DECISION TREES

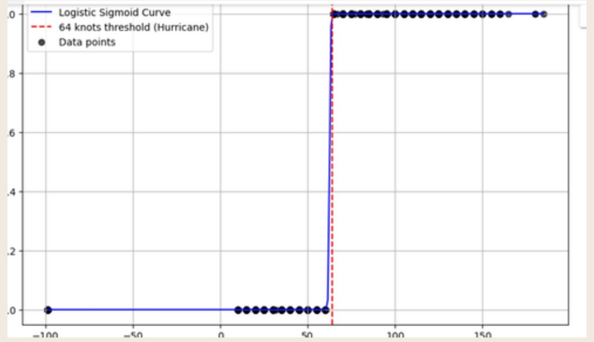
- Used for predicting storm regions (Atlantic/Pacific) and storm types (11 categories) using Maximum Wind and Minimum Pressure.
- The tree structure (depth 2) was visualized to confirm accuracy in storm region prediction. Atlantic data had a broader Minimum Pressure range, while Pacific data showed extended Maximum Wind range.
- Overlapping data between regions suggested similar storm characteristics. Improved classification (depth 1) for storm types reduced overlap, distinguishing Atlantic (higher pressure, lower wind) and Pacific (lower pressure, higher wind)

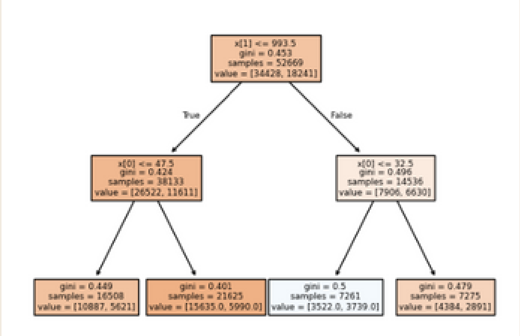
LOGISTIC REGRESSION

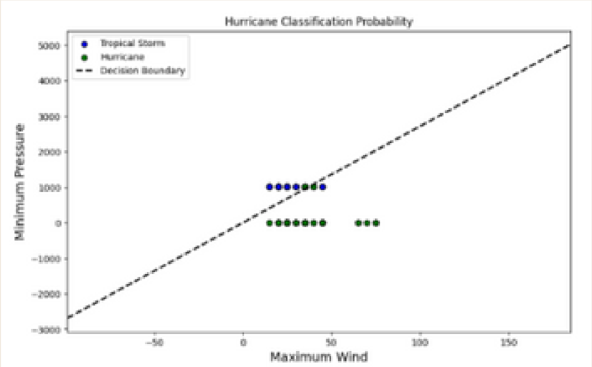
- To predict hurricane occurrence using Maximum Wind and Minimum Pressure as features. Categorical target variables were encoded with LabelEncoder for compatibility
- The dataset was split 70-30 for training and testing. The model's performance was assessed using a classification report with metrics such as accuracy, precision, recall, and F1-score.
- Probability contours and decision boundaries were plotted

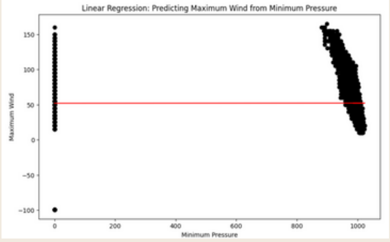
GRAPHS AND PLOTS

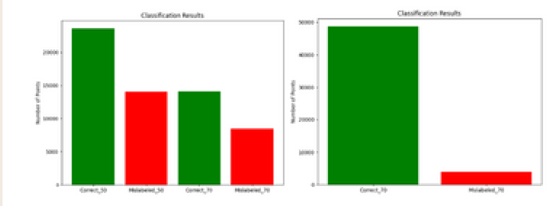


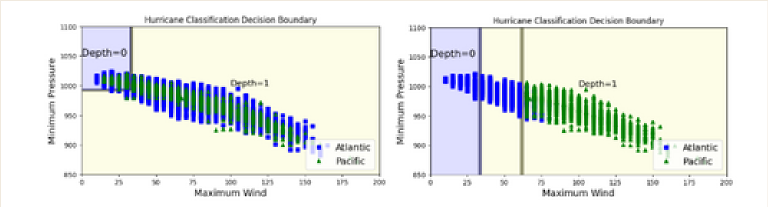












COMPARING THE PREDICTIVE MODELS PERFORMANCES

BY_Pred

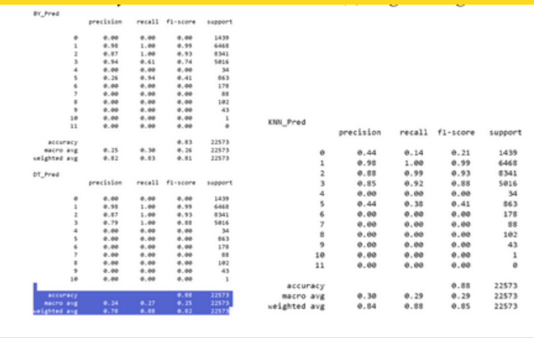
	precision	recall	f1-score	support
Atlantic	0.68	0.79	0.73	14677
Pacific	0.45	0.31	0.37	7896
accuracy			0.62	22573
macro avg	0.56	0.55	0.55	22573
weighted avg	0.60	0.62	0.60	22573

DT_Pred

	precision	recall	f1-score	support
0	0.68	0.90	0.77	14677
1	0.52	0.20	0.29	7896
accuracy			0.66	22573
macro avg	0.60	0.55	0.53	22573
weighted avg	0.62	0.66	0.60	22573

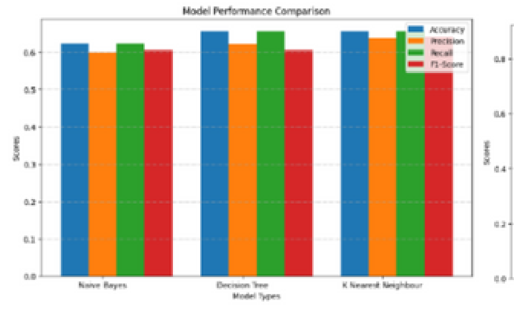
KNN_Pred

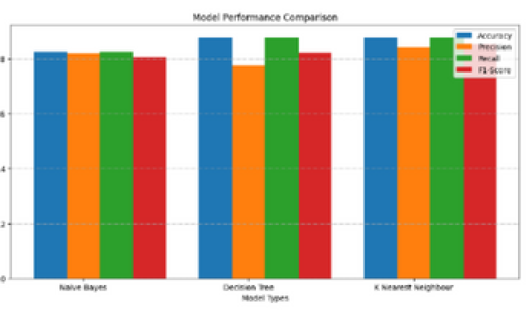
	precision	recall	f1-score	support
0	0.71	0.81	0.75	14677
1	0.51	0.37	0.43	7896
accuracy			0.66	22573
macro avg	0.61	0.59	0.59	22573
weighted avg	0.64	0.66	0.64	22573

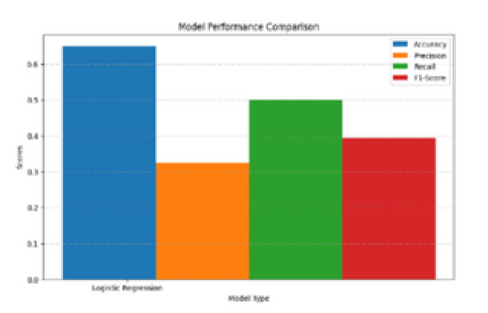


BY_Pred

	precision	recall	f1-score	support
0	0.65	1.00	0.79	14677
1	0.00	0.00	0.00	7896
accuracy			0.65	22573
macro avg	0.33	0.50	0.39	22573
weighted avg	0.42	0.65	0.51	22573







PREDICTIVE MODEL RESULTS & CONCLUSION

```
[ ] k_nearest_neighbour.predict([[20, 1002]]) # This represents a Atlantic prediction as the output came as 0

array([0])

[ ] k_nearest_neighbour.predict([[25, 1009]]) # This represents a Pacific prediction as the output came as 1

array([1])
```

Output predicts whether the storm is Atlantic or Pacific based on the values

```
      Actual Maximum Wind  Predicted Maximum Wind      Error
0                      80               51.927395  28.072605
1                      80               51.927395  28.072605
2                      80               51.927395  28.072605
3                      80               51.927395  28.072605
4                      80               51.927395  28.072605
...                   ...                     ...        ...
49100                  55               52.131997   2.868003
49101                  55               52.132414   2.867586
49102                  50               52.132831  -2.132831
49103                  45               52.132831  -7.132831
49104                  45               52.133249  -7.133249

[49105 rows x 3 columns]
```

(a)

```
      Actual Maximum Wind  Predicted Maximum Wind      Error
0                      45               50.295555  -5.295555
1                      45               50.295555  -5.295555
2                      45               50.295555  -5.295555
3                      45               50.295555  -5.295555
4                      45               50.295555  -5.295555
...                   ...                     ...        ...
26132                  35               47.637965 -12.637965
26133                  30               47.624704 -17.624704
26134                  30               47.622052 -17.622052
26135                  25               47.619399 -22.619399
26136                  20               47.616747 -27.616747
```

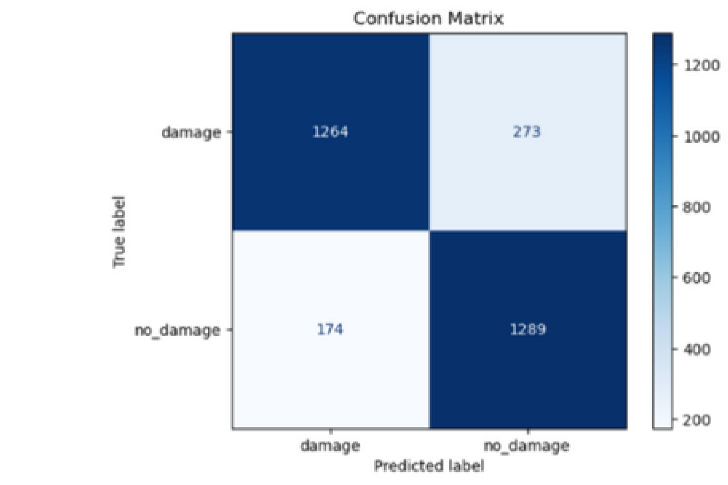
(b)

Actual Wind values, Predicted Wind values and Prediction Errors
(a) Atlantic (b)Pacific

```
Linear Regression Score (R^2): 1.3113275888154696e-05
Intercept: 51.92739459193165
Coefficient: [0.00020857]
Mean Squared Error: 766.2620365261442
```

```
Linear Regression Score (R^2): 0.0027135141849548017
Intercept: 50.295555340485386
Coefficient: [-0.00265229]
Mean Squared Error: 640.0682472016902
```

Values for the intercept, coefficient, R^2, and Mean Squared Error (MSE)



Confusion Matrix of Image Datasets

```
[ ] tree_hurricane_identifier.predict([[20, 1002]])# based off the dataset it shows that its a Tropical Depression

array([3])

[ ] tree_hurricane_identifier.predict([[80, 0]]) #Based off the dataset from these values the array of 1 shows its a hurricane

array([1])
```

Results of using Decision Tree Classifier to Predict the Hurricane Type based on Wind and Pressure values

```
new_data = pd.DataFrame([[80, 950]], columns=['Maximum Wind', 'Minimum Pressure'])

# Prediction
prediction = model.predict(new_data)
probability = model.predict_proba(new_data)

print("Prediction:", "Hurricane" if prediction[0] == 1 else "No Hurricane")
print("Probability of Hurricane:", probability[0][1])

Prediction: Hurricane
Probability of Hurricane: 0.9492472314724114
```

Results of using Logistic Regression classifier to predict hurricane type based on Wind and Pressure values

```
print("Number of mislabeled points out of a total %d points : %d"
      % (X_test2.shape[0], (y_test2 != y_pred2).sum()))

Number of mislabeled points out of a total 22573 points : 8496

print("Number of mislabeled points out of a total %d points : %d"
      % (X_test.shape[0], (y_test != y_pred).sum()))
```

Output showing Number of mislabelled points

```
Storm_Prediction: [3]

       Latitude Longitude
22332     32.9N     86.5W
31699     25.2N     98.7W
31700     25.3N     99.0W
40358     31.7N     96.9W
43815     35.3N     89.1W
45622     32.7N     88.6W
45679     40.2N     88.7W
45680     42.2N     86.5W
46735     17.5N     92.8W
```

	Latitude	Longitude
22332	32.9N	86.5W
31699	25.2N	98.7W
31700	25.3N	99.0W
40358	31.7N	96.9W
43815	35.3N	89.1W
45622	32.7N	88.6W
45679	40.2N	88.7W
45680	42.2N	86.5W
46735	17.5N	92.8W

```
Storm_Prediction: [1]

      Latitude Longitude
0        28.0N     94.8W
1        28.0N     95.4W
2        28.0N     96.0W
3        28.1N     96.5W
4        28.2N     96.8W
...        ...       ...
15815    25.8N    173.6E
15853    32.7N    177.4W
15889    16.5N    147.0W
17059    29.9N    129.7E
17343     6.7N    169.8E

[1542 rows x 2 columns]
```

	Latitude	Longitude
0	28.0N	94.8W
1	28.0N	95.4W
2	28.0N	96.0W
3	28.1N	96.5W
4	28.2N	96.8W
...	...	...
15815	25.8N	173.6E
15853	32.7N	177.4W
15889	16.5N	147.0W
17059	29.9N	129.7E
17343	6.7N	169.8E

1542 rows × 2 columns

Predicating the location of the storm

RESULTS DISCUSSION:

Based on the results, the project demonstrated that with a large numerical dataset, data may overlap in some cases, and using different methods can improve some model's scores and may improve the accuracy and information of the data. Furthermore, different data models show that they may have better accuracy but not as much precision and other such things as other data

CONCLUSION:

This project aimed at providing accurate results, being able to categorize not only the type of storm based on numerical data, but allowing the assessment of damage with satellite images of general areas. This allows for application in international storm watching and prediction use cases.