CLASSIFICATION OF CYCLONES USING PACIFIC DATASET

*A project report submitted to ICT Academy of Kerala*

*in partial fulfillment of the requirements*

*for the certification of*

**CERTIFIED SPECIALIST**

**IN**

**DATA SCIENCE & ANALYTICS**

submitted by

**Team -13**

**Members**

**Harish K S**

**KarunniaMenon M**

**Linto Jose D**

**Liya Chacko Panakkal**

**Saheer Mohammed**



**ICT ACADEMY OF KERALA**

**THIRUVANANTHAPURAM, KERALA, INDIA**

**Feb 2022**

**Table of Contents**

1. **Abstract………………………………………………………………….…….3**
2. **Problem definition…………………………………………………………….4**
   1. **Overview……………………………………………………………………..4**
   2. **Problem statement………………………………………………………….4**
3. **Introduction…………………………………………………………….……...5**
4. **Literature Survey……………………………………………………….……..6**
5. **Result…………………………………………………………………….……..7**
6. **Conclusion……………………………………………………………………...9**

**Abstract**

CLASSIFICATION OF CYCLONES USING PACIFIC DATASET

A Tropical cyclone, also called typhoon or hurricane, is an intense circular storm that originates over warm tropical oceans and is characterized by low atmospheric pressure, high winds, and heavy rain. The National Hurricane Center (NHC) and the Central Pacific Hurricane Center (CPHC) conducts post-storm analyses of each tropical cyclone in their respective areas of responsibility over the North Pacific Ocean to determine the official assessment of the cyclone's history.The dataset which we are going to use is from the revised Northeast and North Central Pacific hurricane database (HURDAT2) - Chris Landsea, James Franklin, Eric Blake, and Ray Tanabe – February 2016.

In this project we aim to make a Machine Learning model to classify the various types of cyclones. i.e. Classification into various Hurricanes or Typhoons using Logistic Regression, Decision Tree, Random Forest , Naive Bayes and SVM. Also, we will analyse what all factors contribute to the classification of cyclones into various categories.

The Target Variable for this Model will be the “Status” column.

**1. Problem Definition**

**1.1 Overview**

In this project, the dataset which we have taken has missing values, outliers etc. so in order to do the project we have to do data understanding, exploratory data analysis, data pre-processing then modelling of the data according to the target feature “Status”.

**1.2 Problem Statement**

Classification of cyclones mentioned in the dataset according to their intensity of occurrence.

**2. Introduction**

The project which we are dealing is basically a natural disaster domain and we have selected cyclones as our topic. As mentioned in the abstract, our project title is “Classification Of Cyclones In The Pacific Dataset”, we classify cyclone according to their status. We went through the dataset and analysed the necessary steps which we have to perform, in order to classify the data.

Initially we conducted the literature survey, then tried to understand our data and its characteristics. For that we did univariate and bivariate exploratory data analysis such as bar graph , donut graph, pie-chart, pair plot etc. Since our project is on the first stage, we have not started with data pre-processing etc.

In the second week of our project, we did data preprocessing such as find missing values and removing them, then feature engineering, made copy of the dataset then did label encoding finally, started with the modelling of the dataset which will be completed in later stage.

In the third week of our project, we made different classification models to find out which suits best for our cyclone prediction problem and also we tried make some modifications in that model too.

**3. Literature Survey**

We basically started with searching about cyclones and their formations and their characteristics changing with variations in wind pressure, intensity etc. Since the dataset includes oceanographic details, we need to use python libraries which usually does not use.

We took the dataset from Kaggle website so we analyzed the note given in that website also which gave relevant information about the dataset and its features descriptions.

**4. Result**

After doing the initial steps, we found out the following:-

* We analyzed the data using python
* There are initially 26137 rows & 22 columns in our dataset
* There are missing values in the dataset.
* We have identified 13 columns having Missing values in it.
* From the bar plot, we found that the month of **August** is prone to maximum number of cyclone formation.
* A vast Majority of the Cyclones were **Unnamed** (i.e., 59.5%).
* The **TS (tropical storm)** is the most occurring type of Cyclone.
* The cyclones in the category of **HU** and **ST** have a wind speed of above 100 Knots.
* Cyclones in the categories: LO, DB, SD, SS, PT occur at higher pressure levels compared to the rest.
* At higher pressure’s, the wind speed is lower.
* The Wind Speed and pressure have a negative correlation.

In the second stage, we did data pre-processing and received the following:-

* Dropped ‘ID’, ‘Event’ columns and sorted out values above zero.
* As the part of feature engineering, created columns Latitude Hemisphere and Longitude Hemisphere and converted the latitude and longitude Column to numeric type.
* Created a copy of the data for doing data modelling with and without label encoding.
* Label encoded the column ‘Name’, which is categorical in nature.
* Finally started doing data modelling, which will be completed in next stage.

In the third stage of our project, we did data modelling using both encoded and non-encoded data and used different classification algorithms as mentioned below:

* Firstly started with logistic regression algorithm , we got the scores as following:
* ..For encoded data:-
  + Accuracy of logistic regression classifier on test set: 0.80
  + precision is : 0.4318677081291625
  + recall is : 0.381928207660089
  + F1 score is : 0.3841900447170007
* ….For non-encoded data:-
  + The accuracy score of LR model is : 0.7570444583594239
  + The F1 score of LR model is : 0.7570444583594239
* Then for KNN-classifer we got :
* ….For encoded data:-
  + Accuracy is : 0.9229805886036319
  + Precision is : 0.813370253861397
  + Recall is : 0.7148704478428995
  + F1 score is : 0.733551813573818
* ….For non-encoded data:
  + Accuracy is : 0.9286161552911709
  + Precision is : 0.5839895600223628
  + Recall is : 0.6155082987131248
  + F1 score is : 0.5978264555489499
* For DECISION TREE classifier , we got
* …For encoded data:
  + Accuracy is : 0.9304946775203506
  + Precision is : 0.9305700764235455
  + Recall is : 0.9304946775203506
  + F1 score is : 0.9302285004483205
* …For non-encoded data:
  + Accuracy is : 0.9348778960551033
  + Precision is : 0.934972313510038
  + Recall is : 0.9348778960551033
  + F1 score is : 0.9348975777079942
* For RANDOM FOREST classifier, we got:
* …For encoded data:
  + Accuracy is : 0.9530369442705072
  + Precision is : 0.9527731364577577
  + Recall is : 0.9530369442705072
  + F1 score is : 0.9525114984114019
* …For non-encoded data:
  + Accuracy of the Model = 0.9555416405760802
  + Precision is : 0.9548719509566814
  + Recall is : 0.9555416405760802
  + F1 score is : 0.9550938295857359
* For GRADIENT BOOST classifier, we got:
* ….For encoded data:
  + Accuracy is : 0.9355040701314965
  + Precision is : 0.936909802973596
  + Recall is : 0.9355040701314965
  + F1 score is : 0.9359759728546262
* …For non-encoded data:
  + Accuracy is : 0.9229805886036319
  + Precision is : 0.9227358032048711
  + Recall is : 0.9229805886036319
  + F1 score is : 0.9225444773703255

After doing hyper-tuning of parameters in gradient boost algorithm we received:

**5. Conclusion**

We have performed the Exploratory Data Analysis on the dataset we have.

We observed that various parameters like Wind speed, Pressure etc. play important role in the classification of cyclones into various categories**.**

Find missing value is the most important part of a data pre-processing and we completed that task very efficiently. Also, since we are dealing with oceanographic data , we didn’t do much oh outliers handling because in cyclones occurrences , there can be abnormalities which can increase the intensity of the cyclone and make them more disasterous.