1. INTRODUCTION

Hurricanes, typhoons and cyclones are nothing but types of tropical storms. Tropical storm is a system formed of rapidly flowing air in a circular formation, with a low-pressure center and overall low-level pressures. This is also usually accompanied by spiral thunderclouds and strong winds which also leads to heavy rains and squalls. The intensity of a storm is measured using various scales of classification depending on the agency keeping track of the storm. For our application, the Saffir–Simpson scale is used which categorizes the storms into five increasing categories depending on their wind speeds. The tracking and monitoring of these natural phenomenon is very important for the safety of people world-wide. And the use of machine learning to predict the behavior of these winds will provide a massive leap in the preparedness in such events. The data used in constructing the algorithms is taken from the Atlantic, Northeast and North Central Pacific hurricane databases. Comparisons are made of the models trained to determine the most appropriate model for our application.

2. LITERATURE SURVEY

On average, every year ten storms enter land from the sea across the globe. And about 4 of these storms are classified as a potential threat to the safety of the people. An estimated 60 billion dollars is lost every year in damage to physical property and relief initiatives. These numbers keep going up every year, majorly due to climate change. The warming of the entire globe has led to more storms forming and has also increased the chance that the storm will grow in magnitude to become a bigger phenomenon. The lasting damage that the natural disaster leaves in its trail is something that communities recover from too slowly. Hence, prediction systems are put into place to reduce the losses gained by us. Efforts are continuously made to improve these systems, but errors are still made as we can never predict the nature of these storms completely accurately. This only means that the systems will continue to be developed till they give negligible error. The machine learning model built by us has a low bias-variance trade-off and thus the problem of multicollinearity is eradicated. The best performing machine learning model is Random Forest and the marqee point is that it is not affected whether the categorical variables are oversampled and undersampled leading to removal of overfitting and underfitting. On Microsoft Azure ML Studio, we applied 2 class Decision Jungle algorithm which is a cherry on cream as it is a noise resistance algorithm and a super set of Random Forest leading to approximately 99% of accuracy without overfitting. The concept has been implemented in the past using image processing rather than using signal acquisition. And further, we see the superiority of our implementation over theirs.

3. METHODOLOGY

3.1. Data Acquisition and Exploratory Data Analysis

The National Hurricane Centre (NHC) analyses the aftermath of each tropical storm in the Atlantic basin and the North Pacific Ocean to obtain official storm surveys history. NHC publishes a hurricane database in a form known as HURDAT, short for Hurricane Database. In addition, the NHC conducts continuous reviews of any retrospective storm analysis submitted and regularly updates the historical record to reflect changes made. First and foremost, we have two datasets comprising of different types of hurricanes and typhoons in the corresponding oceans. We will be operating on two different datasets simultaneously to find any similarities or differences. The first stage of any machine learning pipeline is exploratory data analysis and here it is done by removing the null values and outliers present in both the datasets. As date comes out to be the most important columns, we have to convert into HHMM format for better forecasting. The target column in our scenario is Status column comprises of Categorical column as thus we change the type of column. The below figures Fig 1 and 2 shows the top 10 hurricanes and typhoons by frequency.

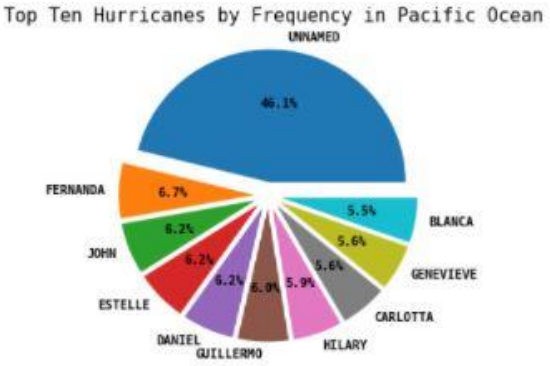

Fig. 1. Pie Chart representing Top 10 Hurricanes by Frequency in Pacific Ocean

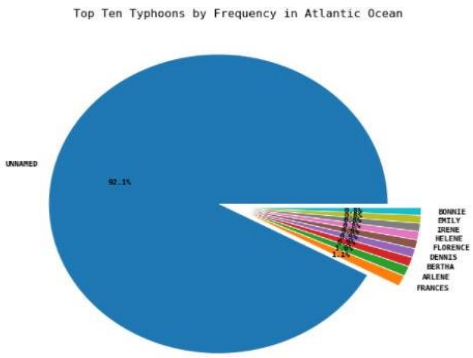

Fig. 2. Pie Chart representing Top 10 Hurricanes by Frequency in Atlantic Ocean

The next step is to use to converted date to see the trend over the last 100 years over the rise of hurricanes and typhoons which is depicted by figures Fig 3 and Fig 4.

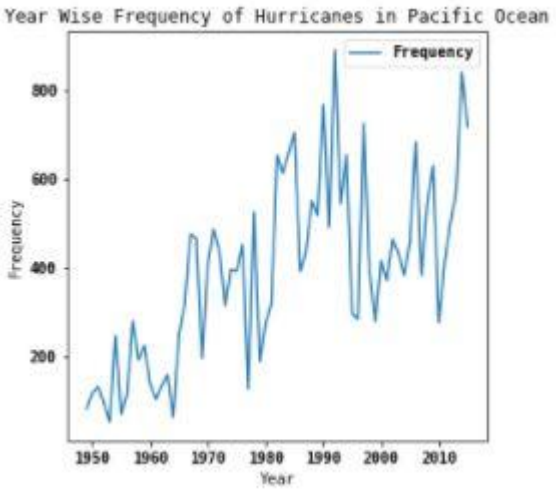

Fig. 3. Line Chart representing Year Wise Frequency of Hurricanes in Pacific Ocean

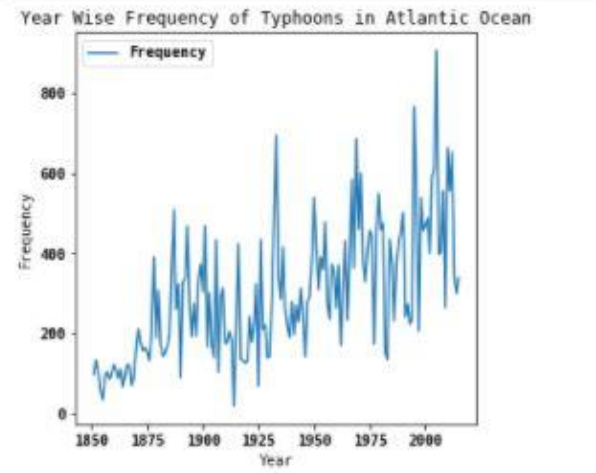

Fig. 4. Line Chart representing Year Wise Frequency of Hurricanes in Atlantic Ocean

3.2. Feature Engineering and Application of Machine Learning Algorithms

Feature engineering is the process of using domain information to extract features (features, properties, attributes) from raw data. A feature is an asset shared by independent units in which analysis or forecasting will be performed. Features are used for speculation models. The metric used to find the feature importance in our scenario is done by concept of eigen value which is present in Principal Component Analysis. After doing feature selection, the major factor influencing the results was Maximum wind and thus we have plotted the density plot comprising of maximum wind and the frequency of different types of cyclones using seaborn library. The density plots are shown in figures Fig 5 and Fig 6.

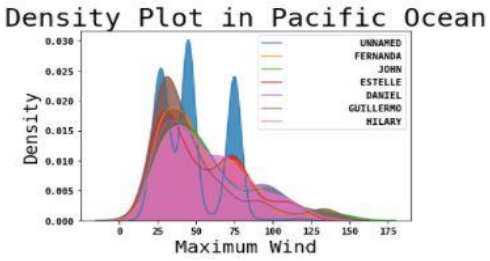

Fig. 5. Density Plot representing frequency of hurricanes w.r.t Maximum Wind in Pacific Ocean

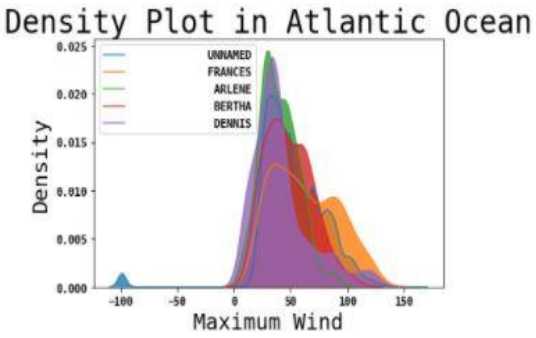

Fig. 6. Density Plot representing frequency of hurricanes w.r.t Maximum Wind in Atlantic Ocean

- Logistic Regression: Logistic Regression produces Results in a binary format that means it is used when value which need to be predicted is categorical. Sigmoid function (S-curve) is used to predict the output of the dataset. Sigmoid curve shrinks the value from (-infinity, +infinity) to discrete Values between 0 and 1 and a threshold Value is kept to differentiate it as 0 or 1. When we applied logistic regression to both the datasets, we got to know that Extratropical Cyclone is also a frequent in Atlantic Ocean. The Tropical Cyclone of tropical depression intensity and tropical cyclone of hurricane intensity both goes hand in hand and has a precision, recall and f1 score the same for Pacific Ocean and its approximately equal to 1. When we come to Atlantic Ocean, there is a slight dip in the accuracy metrics as there is more presence of categorical columns that means different types of cyclones. The logistic regression gives best output for Tropical Cyclone of hurricane intensity (>64 knots), its precision, recall and f1 score in approximately near to 0.98.

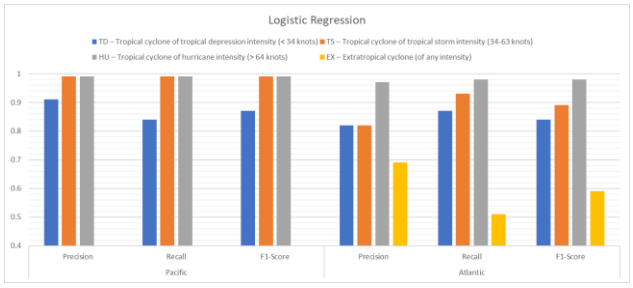

Fig. 7. Output of Statistical Measures of Logistic Regression of both Pacific and Atlantic Ocean

- Decision Tree Classifier: A Decision Tree is a graphical representation of all the possible solutions to a decision based on certain conditions. We Start with the root Node also called as base node & entire data is given to it. Each node will have a condition about one of the features depending on which the data is diffused into subsets such as root node is selected on the basis of Information gain. The goal is to integrate the labels as we proceed down or in other words to produce purest possible distribution at each node. The decision tree classifier successfully classifier four types of hurricanes/ typhoons for both Atlantic and Pacific Ocean. The dis-advantage of the algorithm is that it gives a low precision, recall and f1 score for extratropical cyclone (of any intensity). The marquee point is that the tree algorithm gives all the accuracy metrics approximately nearly equal to 1 except the for extratropical cyclone (of any intensity). The precision, recall and f1 score gradually increases keeping the base as precision for all the types of cyclones in both the oceans.

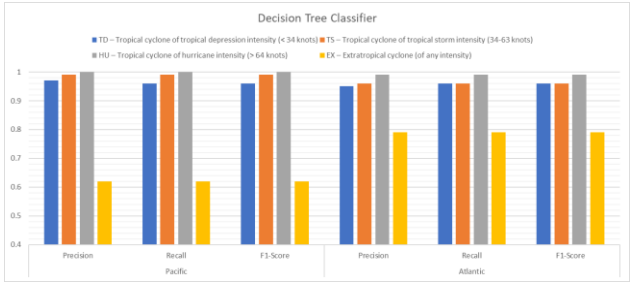

Fig. 8. Output of Statistical Measures of Decision Tree Classifier of both Pacific and Atlantic Ocean

- Random Forest Classifier: Random Forest is a Supervised machine learning algorithm. It can be used for both classification and Regression problems. It operates by constructing multiple decision tree during the training phase, while splitting a node it selects random feature, rather than the most important feature. Thus, it has a wide diversity which results in a better machine learning model. The decision of the majority of the trees is selected by the random forest. Due to presence of multiple trees, it works good with missing date. The Random Forest classifier is the best performing algorithm in the entire scenario and it gives an accuracy of 0.98. All the hurricanes/typhoons are successfully classified for both the oceans especially Tropical Cyclone of tropical depression intensity (<34 knots), tropical cyclone of tropical storm intensity (34-63 knots). The precision recall and f1 score are more or less same for all the three types of classified cyclones except for extratropical cyclone (of any intensity).

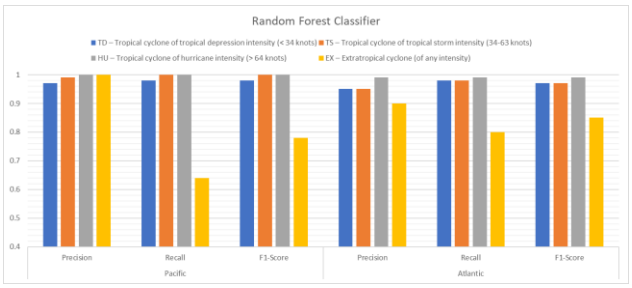

Fig. 9. Output of Statistical Measures of Random Forest Classifier of both Pacific and Atlantic Ocean

- Support Vector Classifier: Support Vector Classifier is a classification algorithm which segregates the data in best possible way. It follows an approach to select a hyper plane with maximum possible margin between the 2 Support Vectors in the given dataset. The non-linear data points have more dimensions which are added with the help of kernel function. The support vector machine supports 3 different types of kernel functions. The support vector classifier is the least performing machine learning model out of all the models. The precision level which is the most important factor is the missing factor here. The accuracy attained here is between approximately 60-75% which is approximately 22% less as compared to other models. Though the SVC model performs averagely in the Atlantic Ocean dataset but it performs poorly on the pacific dataset due to a smaller number of datapoints. The average precision value of precision in Pacific Ocean is approximately near to 0.25. The Support vector classifier fails to identify the different types of cyclones and thus it is not preferred

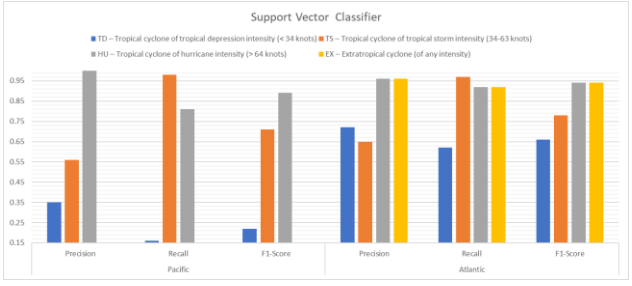

Fig. 10. Output of Statistical Measures of Support Vector Machine Classifier of both Pacific and Atlantic Ocean

- Gradient Boosting Classifier: Gradient boosting classifier creates a model that adds in a high-quality way; allows for the performance of lost and divisive functions. In each phase the n_classess regression trees are equal to the negative gradient of the binomial or multinomial loss function. Binary separation is a special case when a single regression tree is found. Features are always allowed at each break. Therefore, the best available divisions may differ, even with the same training data and max_feature = n_feature, if the improvement in determination is similar to the several cracks calculated during the search for the best divisions. To get timely determination, random_state must be configured. In our circumstances the gradient boosting algorithm performs excellently well as it gives an accuracy of 97%. It keeps in the computational power of the device at the back of mind and produces the corresponding results. The precision, recall and f1 score are approximately near to 0.99 except for extratropical cyclone (of any intensity). The reason of not getting accurate results for extratropical cyclone (of any intensity) is that the occurrence of such cyclone is very rare as shown in the density plot. Hence the statistical measures fail for extratropical cyclone (of any intensity).

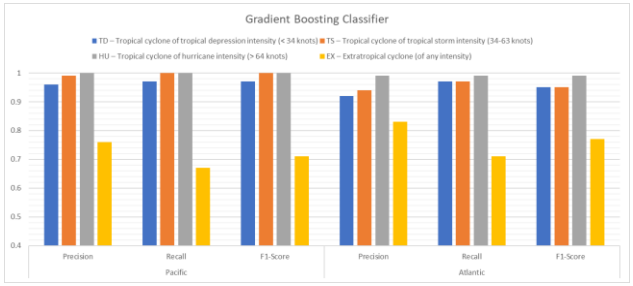

Fig. 11. Output of Statistical Measures of Gradient Boosting Classifier of both Pacific and Atlantic Ocean

3.3. Computing the entire ML Model on Microsoft Azure for Web Service

Azure Machine Learning can be used for any type of machine learning, from classical ml to in-depth, supervised, and transfer learning. Whether we choose to write code in Python or R with the SDK or work with non-coded / low-coded studio options, we can create, train, and track machine learning and in-depth learning models in the Azure Machine Learning Workspace. Azure Machine Learning Studio is a web site on Azure Machine Learning for low-cost options and model training codes, shipping, and asset management. The studio merges with the Azure Machine Learning SDK for a seamless experience. In studio we applied two class decision jungle machine learning algorithms. The corresponding algorithm allows tree branches to come together, the directed acyclic graphs (DAG) decision often has lower memory and better performance to do than a decision tree, even if it costs a certain amount of training longer. Decision jungles are non-parasitic species that can govern the boundaries of off-line decisions. They make the selection and classification of features integrated and are able to be strong when there are sound features. The accuracy achieved was approximately 99% and to be precise the model is not overfitting. The model was converted into a forecasting experiment which helps to deploy the model using webservice and thus it can be called by REST API and HTTP call. The below figure shows the demonstration of entire web service on azure ML studio.

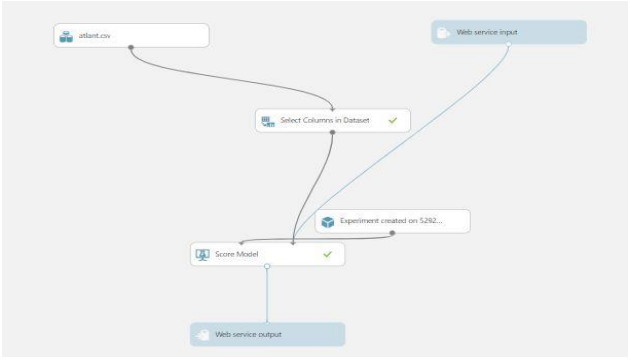

Fig. 12. Flow Model of Microsoft ML Azure for Web Service

3.4. Real Time predicting the severity of natural disaster using Folium library

Folium library builds on the competing power of Python system data and the power of the leaflet.js library map. It generates our data in Python, and visualize it on a Leaflet map with folium. Folium makes it easy to visualize data used in Python on a map of interactive tracts. It enables data binding on choropleth view map and transfers rich vector / raster / HTML views as map markers. The library has many built-in tile-sets from OpenStreetMap, Map-Box, and Stamen, and supports custom tile-sets with Map-Box or Cloud-made API keys. Folium supports both photo, video, GeoJSON and TopoJSON overlay. In our scenario we created a small subset with real time latitudes and longitudes of costal area of United States of America. The frequency and severity of cyclones were crucial parts in that dataset. Seeing the severity, we marked the landfall areas according to the colour scheme as shown in below figure Fig 13.

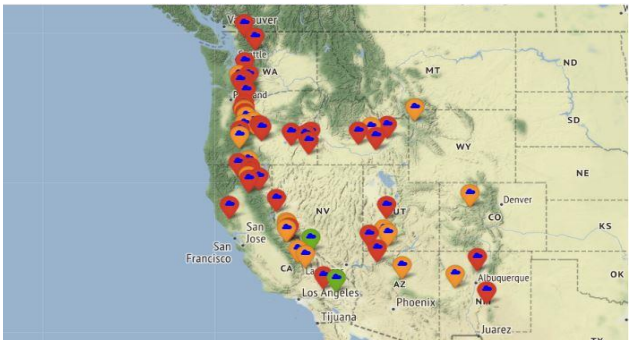

Fig. 13. Live Map representing the severity of cyclone

3.5. Results and Discussions

The entire research paper consists of an entire machine learning pipeline and it was implemented on

two different datasets of Hurricanes and Cyclone of Atlantic and Pacific Oceans. When we applied machine learning algorithm on Atlantic Ocean the result came in favour of random forest algorithm with a whooping accuracy of 96% and the least was obtained by Support Vector Machine of 75% as it considered all the datapoints with the disturbances. The graphical representation is shown below in Fig 14. On the counterpart when we applied machine learning algorithm to pacific ocean , the content here remains the same and thus the performance is much or more related to Atlantic ocean and thus random forest becomes the best performing algorithm with an accuracy of 98% and lowest is given by Support Vector Machine by 62% as shown in Fig 15.

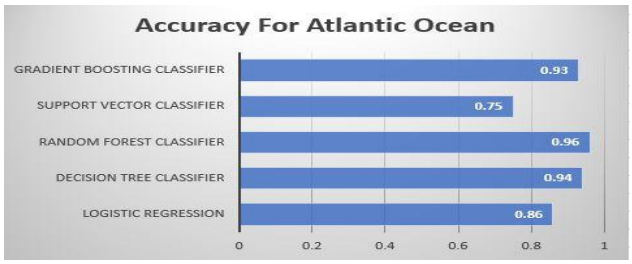

Fig. 14. Horizontal Bar Graph representing Accuracy for Atlantic Ocean

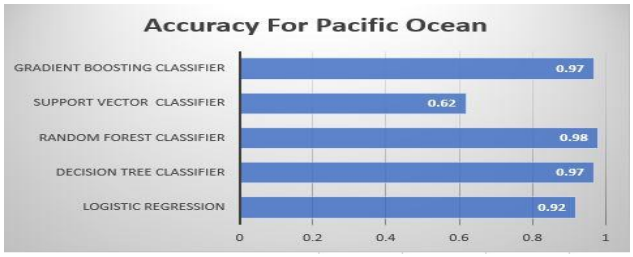

Fig. 15. Horizontal Bar Graph representing Accuracy for Pacific Ocean

4. CONCLUSION

In the entire research paper, we have done machine learning classification on NHC's Hurricanes and Typhoons dataset, starting with exploratory data analysis where we found out the pattern and frequency of occurrence of hurricanes and cyclones in the respective oceans. Then we applied principal component analysis and found the eigen values and the most important factor influencing the results and we plotted a density plot with respect to the number of occurrences of the corresponding cyclone. The next step was to apply machine learning algorithm to the processed data and thus we applied five classification supervised machine learning algorithms. The statistical metrics we used to differentiate the five machine learning algorithms were precision, recall, F1 score and accuracy. The best performing algorithm was Random Forest Classifier. Then the entire pipeline was shifted to Microsoft azure studio and then we applied two-class decision jungle and the accuracy increase by approximately 1-1.5 %. To add further into the project, we used Folium library and its ability to mark the markers on real time map with latitudes and longitudes. We marked the points of high severity of cyclone on a small subset of dataset. The complete pipeline was converted to a predictive webservice where it can be called by an API or REST API which eventually leads to deployment.

5. REFERENCES:

[1] R. Berg, "Tropical cyclone report: Hurricane Ike (AL092008) 1–14 September 2008", 2009

[2] R. Nateghi, S. D. Guikema and S. M. Quiring, "Comparison and validation of statistical methods for predicting power outage durations in the event of hurricanes", Risk Anal., vol. 31, no. 12, pp. 1897-1906, 2011.

[3] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: A methodology review", J. Biomed. Informat., vol. 35, no. 5, pp. 352-359, 2002.

[4] T. Shelton, A. Poorthuis, M. Graham and M. Zook, "Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data'", Geoforum, vol. 52, pp. 167-179, 2014.

[5] Conoscenti C, Angileri S, Cappadonia C, Rotigliano E, Agnesi V, Märker M (2014) Gully erosion susceptibility assessment by means of GIS-based logistic regression: a case of Sicily (Italy). Geomorphology 204:399–411

[6] Dan Li, K. D. Wong, Yu Hen Hu and A. M. Sayeed, "Detection classification and tracking of targets", IEEE Signal Processing Magazine, vol. 19, no. 2, pp. 17-29, Mar 2002.

[7] S. Meckelnburg, A. Jurczyk, J. Szturc and K. Osrodka, "Quantitative Precipitation Forecasts (QPF) Based on Radar Data for Hydrological Models", cost action, 2002