DSTI School of Engineering

Applied MSc in Data Analytics
Applied MSc in Data Science & Artificial Intelligence
Applied MSc in Data Engineering & Artificial Intelligence

Project : Tropical cyclone severity prediction

Instructor : Pauline Salis, PhD

Group members :

- Joanna De Azevedo
- Oksana Denisova
- Julian Lilas
- Ruth Mpozagara
- Zakaria Sidi Bachir
- Benjamin Uzan

Introduction

Cyclones are powerful atmospheric phenomena that can cause severe destruction, loss of life, and significant economic damage. Accurate predictions of cyclone intensity are crucial because they help:

- Warn populations in advance, reducing casualties and property damage.
- Optimize emergency response measures, such as evacuations and infrastructure protection.
- Support insurance companies in risk assessment and disaster planning.
- Assist governments and organizations in making strategic decisions.

IBTrACS (International Best Track Archive for Climate Stewardship) is a global collection of tropical cyclones available. It integrates tropical cyclones data from different agencies all around the world in order to provide a unified dataset and enhance inter-agency comparisons. Developed in collaboration with all World Meteorological Organization (WMO) Regional Specialized Meteorological Centers and other organizations and experts, IBTrACS is a resource for cyclone research and analysis.

The 3rd project is based on the IBTrACS dataset and is composed of 174 columns and 297,099 rows. Its aim is to use this dataset to train a model that predicts the severity of a tropical cyclone based on geographical input data. The predicted value is the TD9636_STAGE column.

1. Understanding the dataset

1.1. General description of the dataset

Each row represents a specific geographic location (latitude and longitude) where storm data has been recorded at a given time. The dataset provides interpolated storm data at 3-hour intervals.

The large number of columns results from the inclusion of various storm-related variables (such as type, wind speed, pressure, and maximum wind radii) reported by 15 agencies worldwide.

1.2. Pre-selection of relevant columns

The project presents a challenge not only due to its subject matter but also because of the large number of columns and rows to manage. To address this issue, a significant portion of the exploratory data analysis (EDA) was dedicated to reviewing the default documentation provided with the dataset, which describes the columns. Additionally, following the instructor's recommendation, we sought more comprehensive documentation.

By examining these two key documents:

- Technical Documentation: Caveats, usage information, and other details about how to use and apply IBTrACS data.
- V04r01 Column Documentation: Description of each CSV and shapefile data column

We have been able to process the preselection of the columns that seem to be relevant for the EDA and the model.

2. Data preparation and preprocessing

2.1. Explanation of the amount of missing data in the pre-selected columns

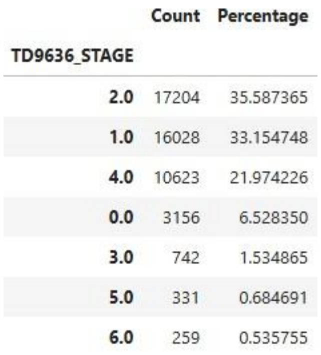
After the changeset of type for the string, numerical and time columns, we have been able to count the number of missing values in our dataset.

At this stage, the IBTrACS dataset includes 50 columns, with missing values ranging from 0% to 100%, which is expected due to its integration of historical and recent storm data from multiple global agencies. These agencies cover different regions, time periods, and storm types. The dataset spans from 1980 to 2024 (2025 for seasonal years), with some agencies providing older data dating back to the 20th century. Data from TapeDeck9636_STAGE covers 1980 to 1989.

Each row represents a storm point with attributes like longitude, latitude, and varying time intervals (mostly 6 hours, but some agencies provide 3-hour or event-specific data), as said before and this variability in data intervals contributes to the missing values.

2.2. Removal of missing values in the target column, consideration of the track type and initial distribution

Given the high percentage (83.7%) of missing data in the target column, rows with missing values will be removed for the EDA phase. This helps achieve clearer statistical summaries. Data augmentation will be addressed later in the project. Only the "main" storm type will be retained, as provisional data and differing agency reports lead to inconsistencies. After removing N/A values and non-main tracks, the dataset contains 48,343 rows and 48 columns, representing 783 storms. The TD9636_STAGE column is an ordinal variable related to storm intensity, with stages 1 and 2 (depression and storm) being the most frequent.



	Count	Percentage
TD9636_STAGE		
2.0	17204	35.587365
1.0	16028	33.154748
4.0	10623	21.974226
0.0	3156	6.528350
3.0	742	1.534865
5.0	331	0.684691
6.0	259	0.535755

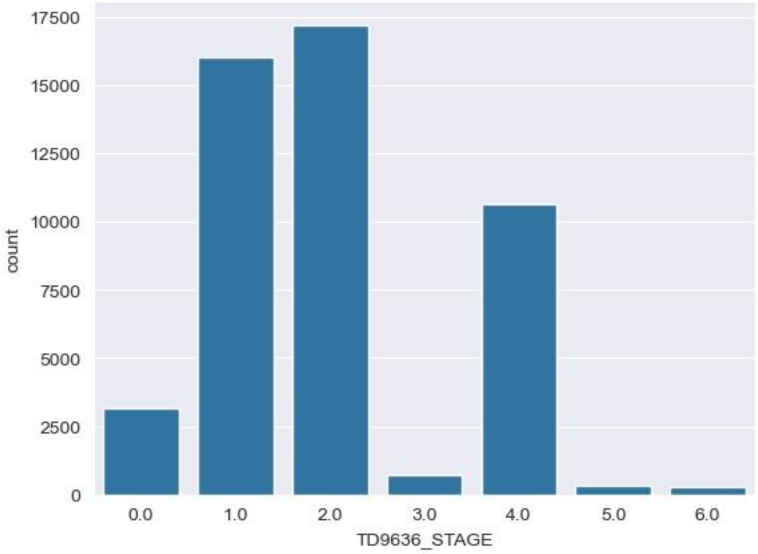


Figure 1. Distribution of the target column

2.3. ISO_TIME column

After removing completely empty columns from the dataset, we focused on the ISO_TIME column. The ISO_TIME column, providing timestamps in UTC, was used to extract year and month for seasonality analysis. The SEASON and ISO_TIME columns were then dropped to avoid redundancy.

2.4. Descriptive statistics of the pre-selected columns

Maximum wind speed columns

The maximum wind speed column represents the highest surface wind speeds recorded at a height of 10 meters. These values vary significantly across different agencies due to differences in measurement techniques and missing data. Most agencies report an average wind speed of around 50 knots, with notable exceptions, such as TOKYO_WIND, which reports higher values.

Distribution: The data is positively skewed, with the mean wind speed higher than the median, suggesting a small number of extreme storms with much higher wind speeds. The KDE (Kernel Density Estimation) plots show that wind speeds are typically concentrated between 35-45 knots, but there is a long tail extending toward higher wind speeds, reflecting fewer occurrences of extreme storms.

Minimum central pressure columns

The minimum central pressure is the lowest recorded pressure at the center of a tropical cyclone, generally measured at sea level. This is a key indicator of storm intensity, with lower pressures typically signifying more intense storms. However, differences in observational methods between agencies lead to some discrepancies in pressure readings.

Distribution: The pressure data is negatively skewed, as the mean pressure is lower than the median, suggesting that fewer storms experience very low pressures. The KDE plots reveal a left tail, indicating that although lower pressures are less common, they are present in the dataset. Most of the data is concentrated between 980 mb and 1000 mb, and lower pressures generally correlate with more intense storms.

Proximity to land columns

The DIST2LAND variable measures the distance from a storm to the nearest landmass, taking into account all continents and islands larger than 1400 km^2. The LANDFALL variable represents the nearest land location within the next 6 hours, but it was deemed redundant and removed in favor of DIST2LAND, which is more reflective of the storm's real-time position.

Distribution: The DIST2LAND data is positively skewed, meaning most storms are located near land. The KDE plots show a high density around 0 km, indicating that many storms are close to land. However, there is also a long tail extending toward larger distances, suggesting that some storms, although less common, occur far from any landmass.

Storm occurrence and localization

As a reminder, the TD9636 dataset is used in IBTrACS outside of the Noth Atlantic.

- The Western Pacific (WP) is the most active storm region (48.1%), and the South Indian Ocean (26%) with the South Pacific Ocean (14.7%) together account for 40 % of storms, indicating a significant storm activity in the Southern Hemisphere.
- Most storms (77%) occur in the MM, which means no subbasins are provided for WP and SI. This suggest that Western Pacific and South Indian dominate storm activity

Storm occurrence and type

Tropical storms dominate the dataset (94.23%), which was expected since most storms originate as tropical.

STORM_SPEED column

STORM_SPEED represents the translation speed of a cyclone, measured in knots, and is derived from changes in latitude and longitude over time. The data indicates an average speed of 9.51 knots, with values ranging from 0 to 69 knots and a standard deviation of 5.57 knots, reflecting moderate variability. The median speed of 9 knots suggests a right-skewed distribution, meaning that most storms travel at relatively low speeds, while a few reach significantly higher velocities.

Distribution: The KDE plot confirms this pattern, with a peak around 6-10 knots and a long tail extending towards higher speeds, indicating that extreme storm speeds are rare but possible.

STORM_DIR column

STORM_DIR represents the translation direction of a cyclone, measured in degrees from 0° to 360°, calculated based on latitude and longitude changes. The data has an average direction of 224.54° and a standard deviation of 96.09°, showing high variability in storm movement. The median of 265° suggests that most storms travel westward.

Distribution: The KDE plot reveals a multimodal distribution, meaning storms follow distinct directional patterns rather than being uniformly spread. A dominant peak between 270°–300° suggests that most storms move westward, while smaller peaks around 50°–100° and 150°–200° indicate secondary preferred storm paths.

2.5. Filling Missing Values in WMO_WIND and WMO_PRES

WMO_WIND & WMO_PRES have a lot missing values. We attempted to fill these columns as we believe they are major factors in explaining the strength of a storm and therefore would be more than useful for the ML algorithm as explanatory variables.
3 main steps have been done:

- The first one, to fill the WMO_AGENCY column, to know for each storm which agency is supposed to have the results
- Then we filled the WMO_WIND column based on the WMO_AGENCY one
- Finally, we filled the WMO_PRES column following the same methodology

3. Exploratory Data Analysis

3.1. Correlation between Features

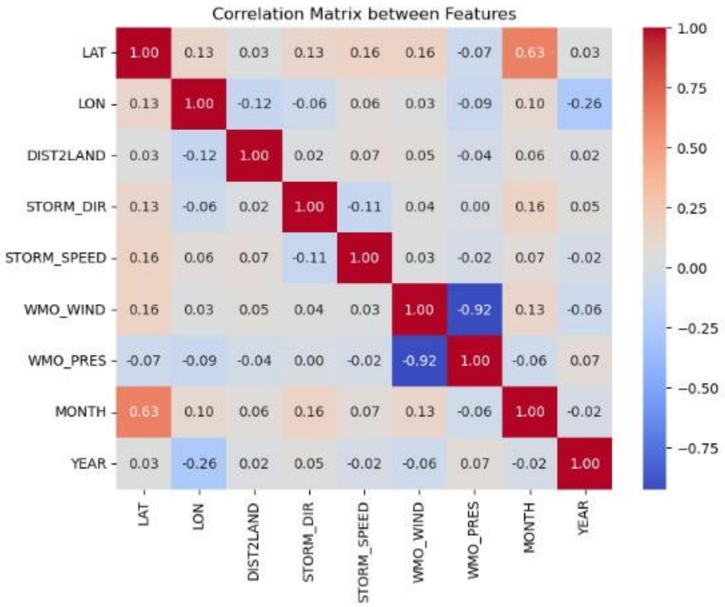


Figure 2. Correlation Matrix between Features

Despite the high correlation between WHO_WIND and WHO_PRES, both features are retained for the following reasons:

- Unique Information: WHO_WIND provides wind speed, while WHO_PRES captures pressure. Both are important for understanding storm intensity and removing either could result in the loss of valuable data.
- Tree-Based Models: Unlike linear regression, tree-based models (like decision trees or random forests) handle correlated features well, so both can remain useful.

Additionally, we notice a relationship between month and latitude, suggesting possible seasonality in the data, which we will explore further in the next section. Other features are uncorrelated.

3.2. Correlation with Target Variable and ANOVA test

Since the Spearman and Kendall correlations showed an insignificant or weak correlation with the target variable, we will conduct an additional ANOVA test to check whether there are statistically significant differences between groups:

- Null Hypothesis: There are no significant differences in the means of the groups for the TD9636_STAGE.
- Alternative Hypothesis: At least one group mean is significantly different from the others for the given feature.

A p-value of 0.0000 indicates strong evidence against the null hypothesis, meaning:

- The feature is statistically significant: The predictor (e.g., latitude, longitude, storm direction, etc.) does have a significant effect on the storm stage.
- Reject the null hypothesis: Since the p-value is much smaller than the significance level of 0.05, we reject the null hypothesis and conclude that the means of the groups are different for each of these predictors.
- Strong influence: The predictors (LAT, LON, DIST2LAND, STORM_DIR, STORM_SPEED, WMO_WIND, WMO_PRES, MONTH, YEAR) are likely to be important factors in determining the storm stage.

This contradicts the results of the correlation tests (Spearman and Kendall), which showed a weak or insignificant relationship. This means that although the variables may have a weak monotonic dependence on the target variable, they can still influence its values, causing significant differences between groups.

3.3. Storm Distribution Over Time and Visualization

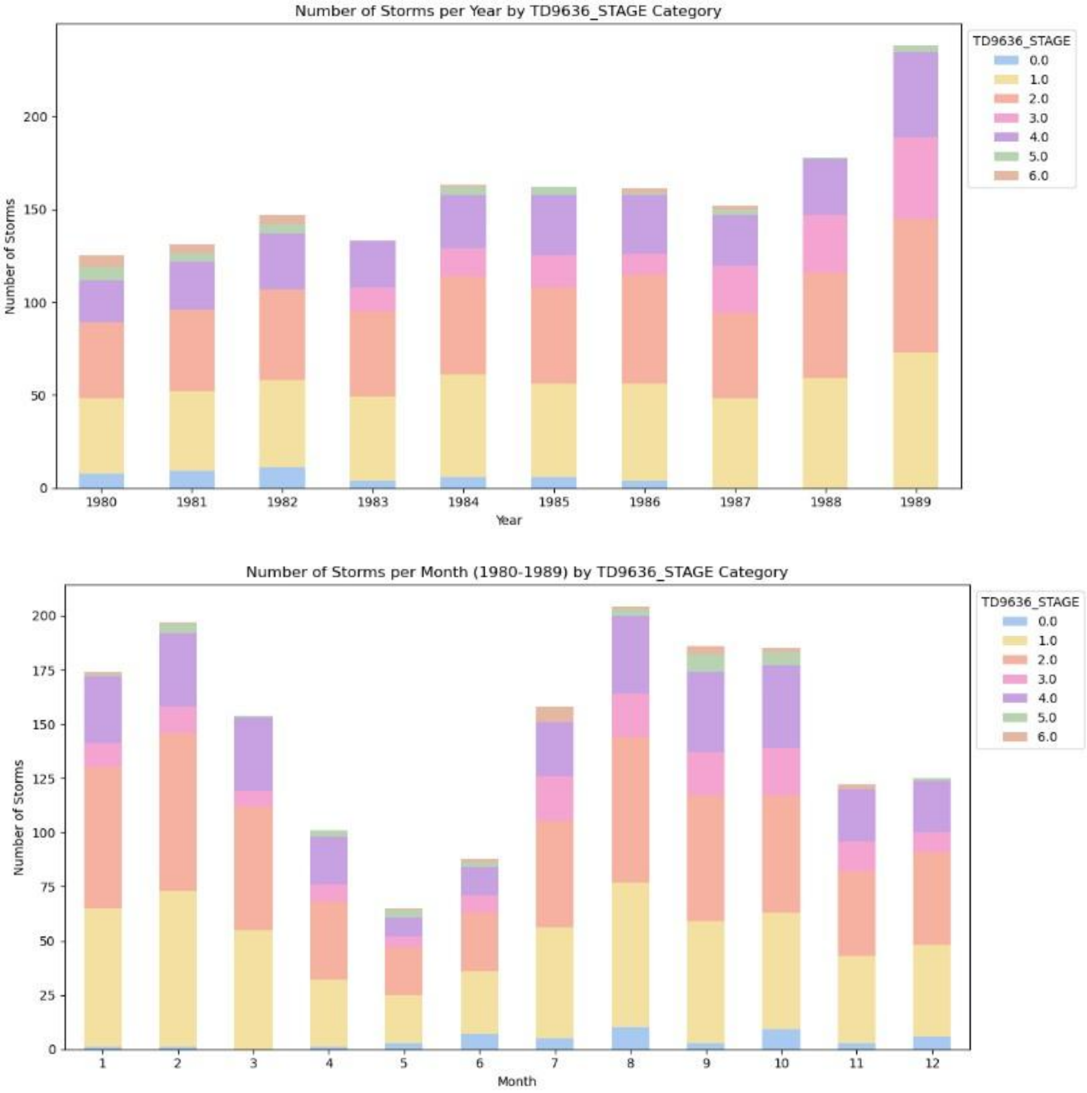


Figure 3. Number of storms per year and month by TD9636_STAGE category

We have data on storms from 1980 to 1989, and the number of storms gradually increased throughout this period, reaching its peak in 1989. We can clearly see those categories 0, 5, and 6 are significantly lower than the others. Also, after 1986, there are no storms of category 0, and storms of category 3 appear only from 1983 onwards.

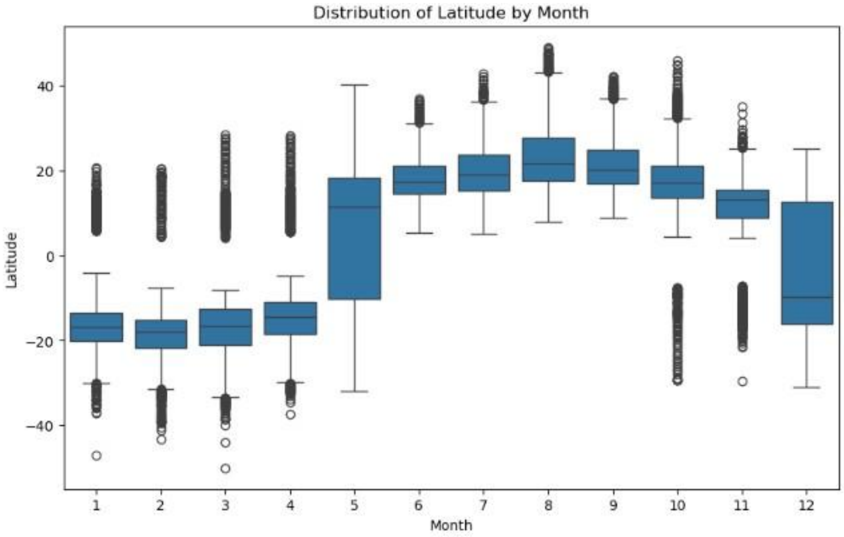


Figure 4. Distribution of Latitude by month

We can see that from May to November, the median is close to 20, while from December to April, the median is closer to -20. This means that from May to November, storms mostly occurred in the Northern Hemisphere, while from December to April, they occurred in the Southern Hemisphere. Additionally, these values indicate that the storms were relatively close to the equator (latitude equals zero).

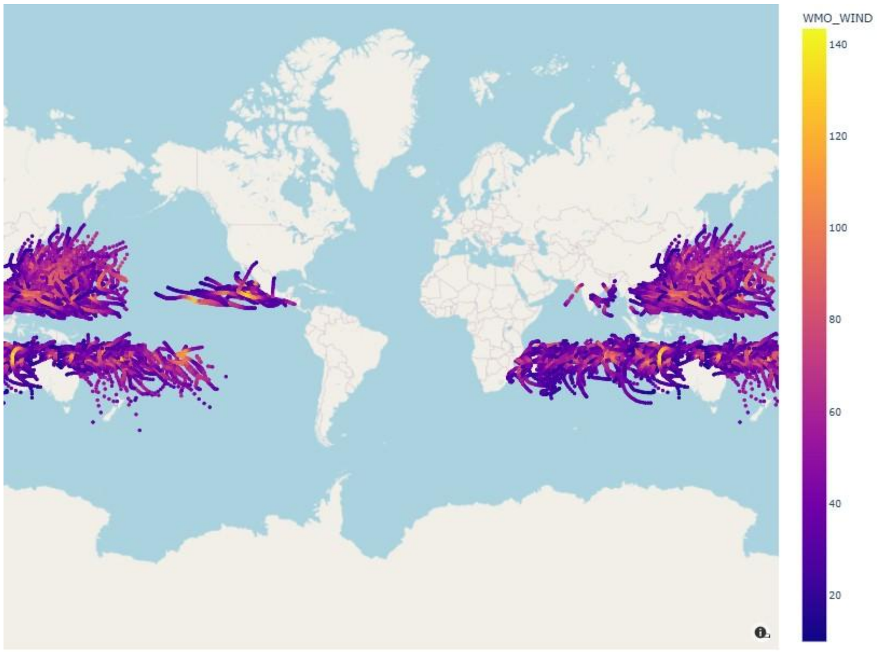


Figure 5. Map showing the locations of storm points based on maximum wind speed.

4. Feature Engineering & Selection

Regarding the feature engineering part we decided to keep the columns:

- Basin
- Subbasin
- USA_SSHS
- Month
- Nature
- Lat
- Lon
- WMO_WIND
- WMO_PRES
- Dist2land

We believe that each of these features provide very important information regarding what a disturbance could be, whether it is a simple depression or a hurricane. For instance, the month gives the information of what time of the year are we in and for example hurricanes do more often happened during specifics months of the year. The basin and subbasin give information about the localization of the disturbance. Since storms and hurricanes usually happen in specifics area of the world, this is as well a valuable information to input in the model. Latitude and Longitude provide information as well regarding the positioning of the storm, same as Dist2land but in a different way than Basin and Subbasin are.

USA_SSHS tells us the hurricane scale, which in some way is similar to our target column TD9636_STAGE, though it is much more precise and subtle than the target.

Finally, WMO_WIND and WMO_PRES provide recorded wind and pressure information about the storm. We believe these two features to be the most important in our dataset as ultimately a storm will be classified as being a depression, storm or hurricane mostly based on these two features that record the “intensity” of the storm.

Among those, Basin, Subbasin, USA_SSHS, Month and Nature are all categorical features, which requires to be encoded to be used by the model. The other features, WMO_WIND, WMO_PRES, lat, lon and Dist2land are all numerical features for which no encoding is necessary.

4.1. Model Selection, Comparison and Evaluation

First, as explained in our notebook, we're aiming at predicting 7 categories for our target column:

- 0 - tropical disturbance,
- 1 - depression with winds < 34knots,
- 2 - storms with winds between 34 to 63 knots,
- 3 - points where wind reached 64 knots,
- 4 - hurricane with winds > 64 knots,
- 5 - Extratropical storms,
- 6 - Dissipating Storm.

First thing first, regarding our dataset, categories 0, 3, 5 and 6 are much less represented than the others. They each represent respectively 0.8%, 2%, 0.7% and 0.2% of the data set. Most of our data regards depression, storms and hurricanes (category 1, 2 and 4).

It is understandable because categories 1, 2 and 4 are all "big" categories that encompass a bigger range in terms of criteria compared to the other categories. It therefore makes sense to have more storms (data) falling in these categories than the others.

4.1.1. Evaluation

How to evaluate our model? Based on those categories, we can already see that some outcomes could be worse than others.

Indeed, categories 0, 1, 2 and 5 represent disturbances, storms and depression with low to mild dangers regarding human life compared to category 3 or 4 which represent the typing point between which a storm switch to hurricane and actual hurricanes.

Regarding the first 3 categories, we believe the worst possible outcome for our model would be to predict that it is either one, while in fact it is a much worse storm like a hurricane. For example, if we predict a simple depression and warn people that a depression is on its way, they will not prepare to potential catastrophic damage, and if that depression turns to be a hurricane, the consequences would be dire. Hence, for category 0, 1 and 2 we will focus on the precision since as explained, we believe that false positives are worse than false negatives. Indeed, if we flag a hurricane, warn people so they would get ready, and the hurricane turns out to be a simple depression, even if not ideal the danger to human life would be much less significant than for false positives.

About category 4, it is the opposite as for category 0, 1 and 2, because here the worst-case scenario are false negatives rather than false positives. Indeed, flagging a hurricane as being a depression for instance (false negative) would have dire consequences as people would not be aware a hurricane is about to hit land. Recall for this category seems to be more important than precision.

Category 3 is kind of in the middle of both category 2 and 4. They represent both big range of storms and hurricanes, while category 3's criterion is a point where wind reached 64knots, which is much more subtle and precise than category 2 or 4. Because of that, we're not so sure about precision or recall being more important than the other. Therefore, we believe that both measures are as important as the others.

Category 6, the least represented set in our data. Representing dissipating storms, the worst for us would be false positives. Indeed, flagging a storm as dissipating while it is not at all. Therefore, precision would be the most important measure regarding this category's prediction.

4.1.2. Model Selection

Regarding the machine learning part, we had in mind first try out a linear model to see if our problem could be solved by something simple. This is why we started first with a logistic regression model.

We had in mind to try out non-linear models as well, to see if they would fit more our problem. We wanted to use a decision tree as well as random forest algorithms. Our problem being a categorical problem, decision tree and random forest seemed to be suited to us.

4.1.3. Comparison

After modelling all three models it seems that the Random forest model isthe most suited regarding our problem. It has the best accuracy and precision as well as recall for our categories regarding our problem.

Yet, we found that the model was very likely to overfit the dataset, same as the decision tree classifier.

Given our problem, our aim, which is to predict the severity of a storm (likely to warn people in advance), and our data (which is old – the data used to train the model range only in the 1980's), we believe overfitting to be critical. It is possible, given the time range of our dataset, that recent data could have subtle change and be different than the one in 1980s. Especially if we take into account the global warming and change in global temperatures, which have a direct impact on the creation of storms worldwide. Hence, we believe that keeping flexibility (less overfitting) at the cost of lower performance is a good choice as it should be better at finding patterns in new and different data than the one it trained on.

This is why we decided to stay with the tuned random forest classifier in our notebook.

Conclusion

Most important features – WMO_PRES, WMO_WIND, LAT, NATURE_TS, NATURE_MX, and DIST2LAND. This means that pressure, wind speed, latitude, and storm characteristics have the strongest influence on the predictions.

Least important features – BASIN_EP, SUBBASIN_AS, MONTH_6, SUBBASIN_BB, MONTH_11. These variables contribute very little to the model.

Region-related categories (BASIN, SUBBASIN, USA_SSHS) are less important compared to numerical storm characteristics like wind speed and pressure.

Time-related variables (MONTH_X) have a relatively low impact, suggesting that the time of year does not strongly affect the predictions.

Longitude (LON) is less important than latitude (LAT), meaning storms might be more influenced by their position relative to the equator rather than east-west location.

So Overall Summary:

- The model focuses on physical storm characteristics (pressure, wind, coordinates).
- Time and region have a weak influence.

Regarding the model performance, the random forest seems to provides the best results regarding this project. Nonetheless the model did have a tendency to overfitt which, as we explained in the notebook, is not a good thing at all, especially given our situation.

The main issue we found all came from the fact that we didn’t have enough data for specifics class, given that we removed most of our dataset due to nulls vales in the target values. Because of this, we end up with underrepresented classes and our model is less good at predicting them. Thoughts about further improving the model would be to try out further data augmentation or to change the target variable to another one, like USA_SSHS, that has much less nulls values. In that way we would have much more data, more recent data and that would definitely be helpful in the training of the machine learning algorithm.