CSCI-6409 Process of Data Science
Project Progress Report
Group - 19

Aditya Mahale	**B00867619**
Benny Tharigopala	**B00899629**
Guturu Rama Mohan Vishnu	**B00871849**
Shiva Shankar Pandillapalli	**B00880049**
Smriti Mishra	**B00904799**

Course Instructor: Dr. Axel Soto

1. Problem Description:

Tornadoes are swirling winds that spin at incredible speeds of up to 400 mph, making them one of the most devastating natural phenomena. They, like hurricanes, are created by extreme low-pressure zones that draw in strong winds. Because they have smaller bases, their destructive power is dramatically amplified. Tornadoes are unpredictable and frequently strike without warning.

Tornadoes can range from a few hundred yards to more than a mile in width. These tornadoes can be found throughout the United States, but the Midwest and South regions are particularly more common and vulnerable. When a tornado is approaching, you just have a few moments to make life-or-death decisions. To survive a tornado, you'll need to plan beforehand and react quickly. To deliver such advanced warnings, we aim to use historical data to estimate the possibility of a tornado in a specific region over time. Meteorologists are trying hard to forecast tornadoes with the intent to predict them soon. Hence, we chose to work on this.

2. Datasets:

We were unable to work with the dataset we discovered on NOAA [8] due to insufficient data. As a result, we began searching Kaggle [6] for the dataset. We identified a dataset of tornadoes reported between 1950 and 2015, which included information like time and date, tornado magnitude, tornado starting and ending locations, and tornado width and length. We also have a variety of other features that help anticipate tornadoes and reduce injuries and fatalities. For our work, we also discovered a few supporting datasets [10][11]. As a result, we'd want to use them because the primary dataset has some missing data from 2015 to 2022.

We proposed to use the ensemble technique to apply different machine learning algorithms dependent on the type of feature value we are aiming to predict to predict several features of a tornado. We intend to work on almost all the features because they are all crucial, such as the tornado's magnitude, time, date, and starting and ending places, as well as the tornado's magnitude and loss. As a result, we want to divide all the characteristics among ourselves and, if necessary, develop separate models using the ensemble technique to accurately predict the features.

3. References & Prior Solutions:

To figure out how frequently tornadoes occur, we went to climate.gov in the United States of America [7], which maintains track of weather data for the entire country. We could see that the frequency was extremely high, particularly in a few states in the southern region. Tornedoes were more common in the southern half of the country. The National Severe Storms Laboratory of the National Oceanic and Atmospheric Administration provides all tornado data to Climate.gov. All the states and tornadoes' strike areas, as well as their heatmaps, may be found at spc.noaa.gov/wcm [9], which inspires us to learn more about it in order to forecast these tornadoes.

Tornado predictions and alerts are typically offered on a short-term basis due to the difficulty in forecasting because of quickly developing meteorological subsystems. However, we believe that based on the massive quantity of historical data we got from Kaggle [6], we can forecast these tornado properties, which will aid in the saving of many priceless human lives. Tornado forecasting has become increasingly important in hazard reduction and risk assessment. The intense storms produced by these tornadoes are required to be predicted to preserve important human lives, depending on the tornado's severity.

Data about tornadoes has been studied [13] by several scholars over time. Many difficulties are clearly analyzed in the articles provided at spc.noaa.gov/publications [12]. However, we aim to anticipate the occurrence of a tornado in a certain geo location by developing numerous models to predict multiple tornado properties/attributes [14] in order to accurately predict the tornado's position, time, and day, as well as its magnitude and fatalities. We intend to forecast this based on the existing facts and limited resources.

4. Feature Selection, Pre-processing & Data Science Solutions:

We are working on combing a couple of data sets to get better insights from the data. We will be exploring the features of each of the datasets by plotting a histogram for a continuous feature and a bar plot of the categorical feature. The plots will give us a preliminary understanding of the nature of the feature. We will derive further insights by plotting a scatter plot matrix of the continuous features present in the dataset. While the descriptive statistics and plot will help us in narrowing down the features, we will also study the tornado-related articles to understand the domain better. Once, we identify the target variable, we will select the features that are most important to the target variable by calculating the information gain ratio/Gini index. Further, entropy graphs will help us whether a feature is informative or not.

We will identify data quality issues such as outliers and deal with them by performing imputations. For preprocessing, we will normalize the data for improving the efficiency of training the model. We will use stratified sampling to avoid bias in the dataset. If required, we will also combine oversampling or under sampling along with stratified sampling.

We have decided to utilize supervised learning methods to develop the model for our project. Since we have decided to answer multiple questions in this project, we will use different models for predicting the relevant target variable such as fatalities, magnitude, etc. Ensemble learning model seems like a good option now based on the initial investigation of the dataset due to its ability to predict the outcome with high accuracy. However, we will employ other models as well to test and deploy our model.

5. Project Milestones & Schedules:

Name	Start Date	End Date	Duration	Progress %	Resources
Idea Brainstorm & Idea Pitching	Jun 01, 2022	Jun 03, 2022	3 days	3	Group-19
Literature Survey & Searching Dataset	Jun 04, 2022	Jun 06, 2022	3 days	3	Group-19
Understanding Dataset	Jun 07, 2022	Jun 09, 2022	3 days	4	Group-19
Data Clearning & Preprocessing	Jun 10, 2022	Jun 12, 2022	3 days	10	Group-19
Individual Models and Testing Hypothesis	Jun 13, 2022	Jun 30, 2022	18 days	30	Group-19
Combining All Models and Evaluating	Jul 01, 2022	Jul 16, 2022	16 days	40	Group-19
Final Documentation	Jul 17, 2022	Jul 26, 2022	10 days	10	Group-19

Table 1. Project Schedule

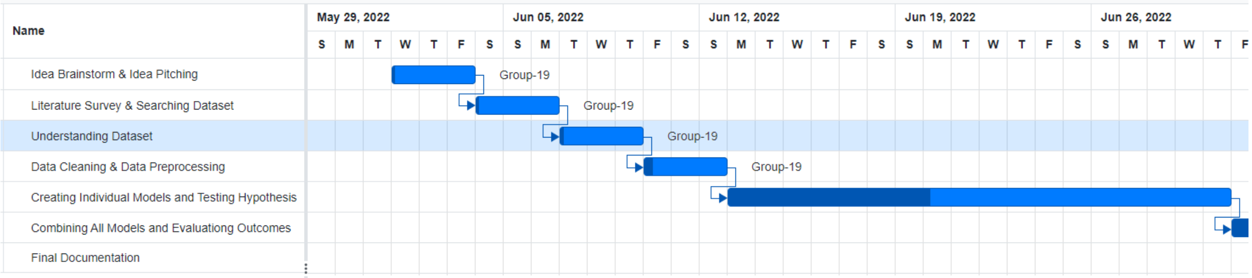


Figure 1. Project Milestone Planning (1)

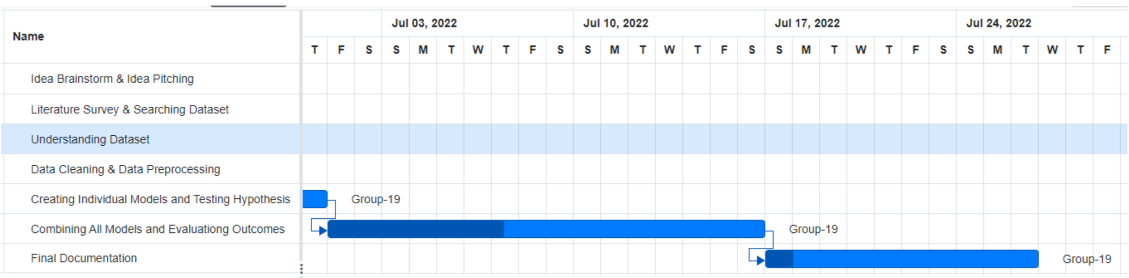


Figure 2. Project Milestone Planning (2)

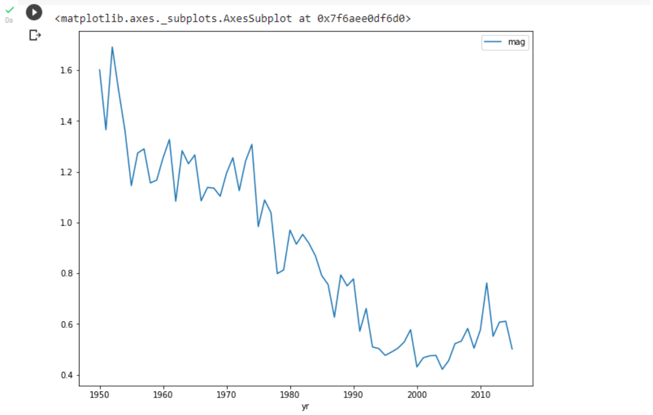
6. Design & Development Process:

Journals and research articles [1-5] relevant to Tornados and Storm prediction centers were analyzed to further the comprehension of the problem at hand and derive an appropriate definition of a solution, which is to predict the frequency of Tornadoes across 52 states in the United States of America. After the dataset for the problem was obtained from Kaggle [6], it was analyzed to determine the target variable and data quality issues through data quality reports. Subsequently, the dataset was transformed and wrangled. Attributes relevant to the problem were then retained, while others were discarded. Erroneous values were handled through approaches such as Imputation, with a measure of the central tendency of the feature.

To approach a feasible solution, several learning algorithms were proposed by the members of the team out of which a candidate list is to be identified and implemented to rank the algorithms based on their performance, viz., the prediction of the target variable. Test and model assessment designs will be performed in parallel to observe the accuracy of each model. We plan to then evaluate the results obtained from testing the optimal model against 'unseen' data and verify if they are substantial to meet the objectives of the solution. Finally, we will summarize our findings and plan to setup a pipeline to facilitate continuous mining and analysis of data from the National Oceanic and Atmospheric Administration's Storm Prediction center to predict Tornado frequencies in the future.

7. Visualization:

This below **Figure 3** represents the relation between the features "mag" and "year" throughout the whole dataset. The feature taken on the x-axis is "year" and the feature taken on y-axis is "average of mag". Each year shows the average magnitude for that one individual year. The observation to be noted from this graph is that, year by year, the average magnitude of the tornadoes keep decreasing and there was never really any spike or contrast in the pattern.



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f6aee0df6d0>
```

Figure 3. Magnitude of Tornadoes in USA from 1950-2015

The following **Figure 4** represents the relation between the geographical map of United States and the Magnitude for all the states from 1950 to 2015. The scale shows the levels of magnitude in 5 levels, each represented by a different color. As we can see, only 2 states out of all the states in US has had an overall magnitude equal to 3 and the rest all states have been showing the overall magnitude of either 1 or 2. The two states which show the magnitude 3 are "AR" and "AL".

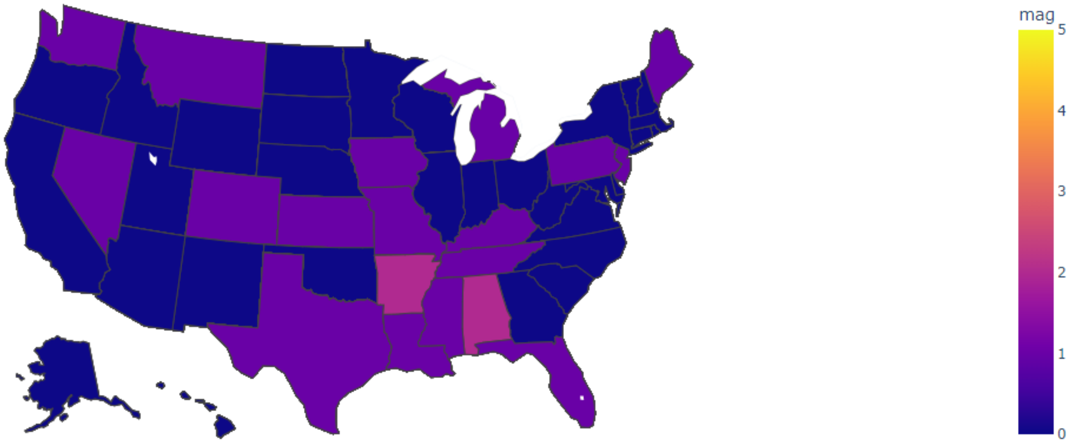
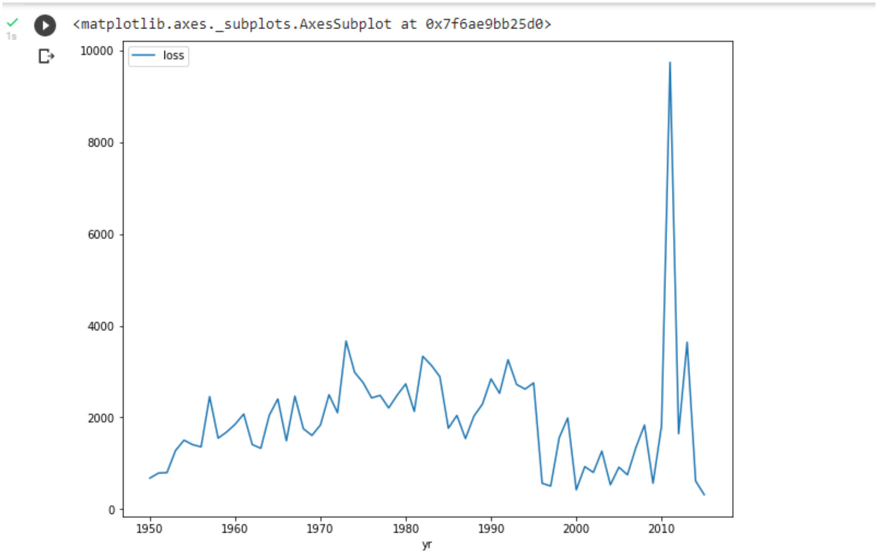


Figure 4. Geographical Representation of Overall Magnitude in USA

This below **Figure 5** represents the relation between the features "loss" and "year" throughout the whole dataset. The feature taken on the x-axis is "year" and the feature taken on y-axis is "loss". The observation to be noted from this graph is that all the years have reported a loss of not more than 4000 but in 2011 alone, there was a huge spike in the loss and fatalities and casualties also which means that there was a major incident contrasting to the other previous years.



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f6ae9bb25d0>
```

Figure 5. Geographical Representation of Overall Magnitude in USA

8. References:

[1] Snook, N., Xue, M. and Jung, Y., 2019. Tornado-Resolving Ensemble and Probabilistic Predictions of the 20 May 2013 Newcastle–Moore EF5 Tornado. *Monthly Weather Review*, 147(4), pp.1215-1235.

[2] F. Aleskerov, S. Demin, and S. Shvydun, “Superposition of Choice Functions and Its Application to Tornado Prediction and Search Problems,” *SN Computer Science*, vol. 1, no. 2, Feb. 2020, doi: 10.1007/s42979-020-0072-2.

[3] J. N. Basalyga, C. A. Barajas, M. K. Gobbert, and J. Wang, “Performance Benchmarking of Parallel Hyperparameter Tuning for Deep Learning Based Tornado Predictions,” *Big Data Research*, vol. 25, p. 100212, Feb. 2021, doi: 10.1016/j.bdr.2021.100212.

[4] M. P. McGuire and T. W. Moore, “Prediction of tornado days in the United States with deep convolutional neural networks,” *Computers & Geosciences*, vol. 159, p. 104990, Feb. 2022, doi: 10.1016/j.cageo.2021.104990.

[5] R. Lagerquist, A. McGovern, C. R. Homeyer, D. J. G. Ii, and T. Smith, “Deep Learning on Three-Dimensional Multiscale Data for Next-Hour Tornado Prediction,” *Monthly Weather Review*, vol. 148, no. 7, pp. 2837–2861, Jun. 2020, doi: 10.1175/MWR-D-19-0372.1.

[6] J. TENNIS, “Storm Prediction Center,” *www.kaggle.com*, Aug. 05, 2016. https://www.kaggle.com/datasets/jtennis/spctornado (accessed Jun. 12, 2022).

[7] gov, C., 2022. Monthly and Annual Numbers of Tornadoes - Graphs and Maps. [online] www.climate.gov. Available at: <https://www.climate.gov/maps-data/dataset/monthly-and-annual-numbers-tornadoes-graphs-and-maps> [Accessed 12 June 2022].

[8] N cei.noaa.gov. 2022. April 2022 Tornadoes Report | National Centers for Environmental Information (NCEI). [online] Available at: <https://www.ncei.noaa.gov/access/monitoring/monthly-report/tornadoes/202204> [Accessed 12 June 2022].

[9] Spc.noaa.gov. 2022. Storm Prediction Center WCM Page. [online] Available at: <https://www.spc.noaa.gov/wcm/> [Accessed 12 June 2022].

[10] world, d., 2022. data.world. [online] Data.world. Available at: <https://data.world/datasets/tornado> [Accessed 12 June 2022].

[11] World, D., 2022. Historical Tornado Tracks - dataset by dhs. [online] Data.world. Available at: <https://data.world/dhs/historical-tornado-tracks> [Accessed 12 June 2022].

[12] Spc.noaa.gov. 2022. SPC Publications. [online] Available at: <https://www.spc.noaa.gov/publications/> [Accessed 12 June 2022].

[13] Pararas, G., 2022. TORNADOES: MODELING AND FORECASTING - George Pararas-Carayannis. [online] Drgeorgepc.com. Available at: <http://drgeorgepc.com/TornadoModelForecast.html> [Accessed 12 June 2022].

[14] Spc.noaa.gov. 2022. SPC Tornado Hail and Wind Database Format Specification. [online] Available at: <https://www.spc.noaa.gov/wcm/data/SPC_severe_database_description.pdf> [Accessed 12 June 2022].