**IEOR 242 Project Plan**

**Team members:** Siyi Weng, Jing Yu, Tianyi Xu, Chengjie Xia, Zhibo Tang

* **Motivation:** 
  + Problem: Using climatic factors such as wind direction and speed, pressure, and humidity, etc. on the previous day to predict if a location (in Australia) will rain the second day.
  + Why:
    - Daily use: As a reference for people to dress tomorrow. Help people to prepare for the need of bringing umbrellas, rain coats etc.
    - Business: Help businesses to make better decisions, such as the airline company, so they can schedule their plan in advance.
    - Farming: Help farmers to get prepared for weather hazards, intense fall/rainstorms to avoid the risk of loss.
* **Impact:** 
  + Potential impact of Data:
    - Nature of data:
      * This dataset contains about 10 years of daily weather observations from many locations across Australia.
      * `RainTomorrow` is the target variable to predict. It means -- Will it rain the next day, Yes or No? (This column is Yes if the rain for that day was 1mm or more.)
      * ‘Location’ is the name of the location of the weather station.
      * ‘MinTemp’ is the minimum temperature in degrees celsius.
      * ‘MaxTemp’ is the maximum temperature in degrees celsius.
      * ‘Rainfall’ is the amount of rainfall recorded for the day in mm.
      * ‘Evaporation’ is the amount of evaporation (mm) in the 24 hours to 9am.
      * ‘Sunshine’ is the number of hours of bright sunshine in the day.
      * ‘WindGustDir’ is the direction of the strongest wind gust in the 24 hours to midnight.
      * ‘WindGustSpeed’ is the speed (km/h) of the strongest wind gusts in the 24 hours to midnight.
      * ‘WindDir9am’ is the direction of the wind at 9 am.
    - Data collection and processing：
      * Collection:
        + <https://www.kaggle.com/jsphyg/weather-dataset-rattle-package>
        + Source: Observations were drawn from numerous weather stations. The daily observations are available from: <http://www.bom.gov.au/climate/data>.
      * Processing:
        + Exploratory Data Analysis:

Data visualization(Histograms, scatter plots, box plots etc.) to get a better understanding of the structure and information of the data.

Draw correlation plots and do feature selection by deleting highly-correlated features and unimportant features.

* + - * + Data cleaning:

Step 1:Remove abnormal values generated during the data collection.

Step 2: Put ‘NA’ to columns of missing values. Apply specific methods (for those features having over 50% missing values, drop the column; for discrete features, use mode to fill in the missing values; for continuous features, use mean/median to fill in the missing values; also try out random forest algorithm to fill in the missing values and cross validate the effect on validation set) to handle the missing values.

Step 3: Apply method of normalization to speed up the convergence of gradient descent when using unit-sensitive algorithms (e.g. support vector machine)

Step 4: Deal with the tricky variable, location. Collect more information via web crawlers and fully utilize the location information.

* + Analytics models:
    - Methods: CART Decision tree, logistic regression, random forest, support vector machine, XGBoost etc.
    - Results: XGBoost is expected to achieve the best results.
    - Confidence: XGBoost is expected to achieve the best results no matter how the raw dataset is splitted into the training set and test set.
  + **Impacts:**

**People:**

Prediction on whether or not it will rain tomorrow helps people adjust their ways to work and trip plans.

**Industries:**

* **Agriculture:**

Farm operation is highly dependent on the weather conditions. Continuous rainy days may lead to crop harvest and livestock’s living standards. The prediction helps farm owners make necessary adjustments to the schedule to avoid economic loss.

* **Insurance:**

Insurance companies care about if it rains tomorrow because many natural hazards such as floods may damage the construction, increase the possibility of car incidents, etc.

* **Commercial Fishing:**

Commercial fishing companies depend on weather prediction to decide whether or not they will go fishing on that day. The accurate prediction could potentially avoid economic and livelihood loss.

* **Tourism:**

Prediction helps tourism companies to better plan the trip and serve customers.