**DATA SCIENCE PROCESS**

**Assignment**

Case Study Topic: Weather Forecasting

Nivetha Jayakumar - RA1811027010013

Kanishk Jain - RA1811027010025

Karmi Makadia - RA1811027010043

1. DISCOVERY: Weather Forecasting is an important and necessary area of investigation in human life. In today’s world we are heavily dependent on the future weather, whether it is from industries to agriculture, travelling to daily commuting, for easy and seamless mobility it is better to predict the weather beforehand. This phase aims at exploring and explaining the need for weather prediction and its effects on a day to day basis. The people involved in this phase include:

* Project Manager
* Project Sponsors

The project/case study looks to predict the weather for a city. This prediction is based on factors like the temperature gradient, humidity, wind speed and more. This will help people of the area to conduct their activities in accordance to the weather thereby reducing the damage otherwise caused.

1. DATA PREPARATION: In the proposed model consists of historical weather data which involves factors that affect the weather, it includes temperature( minimum and maximum ),humidity,wind speed ,precipitation,visibility.It is essential as it allows to manage data in a better manner and operate accordingly. This phase deals with cleaning, preprocessing and filling in empty data. The various processes involved in this phase are:

* Data Extraction: The main motive of this step is to collect data for model building. This data is needed for examination and is taken from the url- *h*[*ttps://www.kaggle.com/selfishgene/historical-hourly-weather-data*](https://www.kaggle.com/selfishgene/historical-hourly-weather-data).

The data for features such as temperature, humidity, pressure, etc. for a selected city will be extracted from individual files and combined according to the date series.

* Data Cleaning: The collected data is then segregated into parts which are in use and which are not of any use in the model. Then the data is passed through a process wherein the missing and the error values in the dataset are replaced by the mean values or the most occurring values in the field.
* Data Transformation: Categorical data, if present, is encoded into decimal values. The cleaned dataset is then divided into two parts: training set and testing set. Training set is to train the model and compute the result, while the testing set is used to find results and compare the actual result with the calculated result. It involves checking correlation between variables,then further identifies and splits them into dependent and independent variables. Finally normalizes the data,if required.

The people contributing in this phase are:

* Database Manager
* Data Scientist/Engineer
* Project Manager
* Business Analyst

1. MODEL PLANNING: In this phase, we explore the various algorithms that can be used to create the model, check for the required libraries to build the model and find the suitable model with respect to various factors such as size of the dataset, relevant features, etc. The model is then fine-tuned according to various statistical indicators, the importance of the different variables is visualised using plots and the weights are changed accordingly. For weather forecasting, features such as temperature, pressure, humidity, rainfall, etc. are used. The algorithms considered here are multiple linear regression and random forest regression.

* Multiple linear regression predicts the dependent variable by establishing a linear weighted relation between the various known features.
* yi​=β0​+β1​xi1​+β2​xi2​+...+βp​xip​+ϵ
* where, for i=n observations, yi​=dependent variable, xi​=explanatory variables, β0​=y-intercept (constant term), βp​=slope coefficients for each explanatory variable, ϵ=the model’s error term (also known as the residuals).
* Random forest regression operates by constructing a multitude of decision trees at training time and outputting the mean prediction (regression) of the individual trees.​
* Data scientists and all aforementioned people but the data engineer, play a key role in this phase.Data scientists develop these models according to the large amount of data received. They must possess clear knowledge with regard to the relations between data elements and the most suitable algorithm that matches the output.

1. MODEL BUILDING: This stage involves trying out various prediction models in order to find the best suited one for the data at hand. The cleaned set is segregated into the training set, validation set and test set in either a 80-10-10 split or 75-12.5-12.5 split, depending on the outcomes for each. The data used in this stage will already have been preprocessed and cleaned. We are going to use multiple linear regression and/or random forest regression algorithms to make our model. Many times it is observed that refining the data preparation process helps in this phase. The more accurate and defined the data the the more accurate the prediction model holds.The people involved in this phase include:

* Data Scientist
* Business Analyst
* Project Manager for supervision

1. OPERATION: In this phase, the model is checked for its accuracy and how well it meets the requirements before being deployed. The conditions set forth by the clients (discovered in phase 1) should be fulfilled by the model in hand. This is done with the help of various statistical and visualisation tools used to determine errors and variations respectively. There can be a sample testing process where external, unseen data can be used to test performance of the model kept in real environmental conditions. Once we are left with the best possible model, it's deployed.

The project manager and project sponsors are sometimes involved for cross-verification, but the data scientists play a major role in this phase.

1. RESULTS: With the algorithm operation and the model set to get the most accurate prediction according to the data available we are only left with the task of explaining the task and showing the results to possible stakeholders and investors. In order to properly communicate the functioning and the results visualization plays an important role. Visualization also helps to obtain the non technical view on the project. The people involved in this are:

* Data Scientists- to explain the visualisation
* Project Manager-to give an overview
* Business Analyst- to explain the business perspective to the project
* Project Sponsors
* Business User