**Report**

**Twitter airline sentiment**

**Data science Process**

1. **Setting the research goal:** The purpose is to making sure that all stakeholders understand the what , why and how of the project. Our project is about twitter airline sentiment in which we analyze the problem of major US airlines name American, Delta, Southwest, United, US airways and Virgin America. We have twitter datasets of each of these airlines and we have to predict and classify these tweets as positive, negative and neutral tweets, followed by categorizing negative reasons (such as "late flight" or "rude service"). We have to predict that whether the sentiment of the tweets in this dataset was positive, neutral, or negative for six US airlines. Then we create a project charter for the clients where they will know what they are paying for by briefly define the research goal. The project charter also includes how you perform the analysis and what resources will be used so that client can estimate the cost of the project and number of data and people required to make this project a success.
2. **Retrieving data:** The next step in data science process is to retrieve the required data. We have retrieve data from websites like Kaggle which has high quality data freely available for public and commercial use. Data can be stored in many forms, ranging from simple text files to tables in a database. The objective is to acquire all the data we need to analyze them and conclude the results. Data that is acquired may be in raw form that must be polished.
3. **Data preparation:** Data which is retrieved must be prepared for use in modeling phase. Steps in data preparation includes
   1. **Data cleansing:** It is a subprocess which focuses on removing errors from data to make it consistent. Data cleansing process includes errors like
      1. **Missing values:**
      2. **Outliers:**
      3. **Spaces, types:**
   2. **Data transformation:** After data cleansing the next step is to transform the data so that data comes in suitable shape for modeling. Data transformation includes
      1. **Reducing the number of variables:** We need to reduce variables because too many variables in the model make the model difficult to handle.
      2. **Turning variables into dummies:** Variables can be turned into dummy variables.
   3. **Combining data:** We can combine data from different sources which includes
      1. **Merge datasets:**
      2. **Set operators:**
      3. **Creating views:**
4. **Data exploration:** Data is represented using graphical techniques to gain understanding of the data and interactions between variables. We look for patterns, correlations, deviations based on visual and descriptive techniques. In this step we also discover anomalies we missed and try to fix them before exploring the data. Graphical techniques includes
   * 1. **Simple graphs:**
     2. **Combine graphs:**
     3. **Histograms:**
5. **Data Modeling:** Next step is to build the models in order to make better predictions, classifying objects, gain the understanding of the system that is modeling. This phase is much more focus than exploratory analysis step because we know what we are looking for and what you want the outcome to be. Data modeling mainly consists of three steps.
   * 1. **Selection of the modeling technique and variables to enter in the model:**
     2. **Execution of the model:**
     3. **Diagnosis and model comparison:**
6. **Presentation & Automation:** After analyzing data and built a well performed model , we present the findings what we get from the model and explain it to the stakeholders. You need to convince them that your findings will bring an impact in the business process.