**Capstone Project Submission**

| **Team Member’s Name, Email and Contribution:** |
| --- |
| 1. Anjali Tidke ( [anjalitidke123@gmail.com](mailto:anjalitidke123@gmail.com) )  * Importing Libraries * Loading Dataset * Data Analysis using EDA * Model Development * Logistic Regression * Random Forest Classifier * Model Selection * Conclusions  1. Shubham Dukare ([shubhamalmabetter@gmail.com](mailto:shubhamalmabetter@gmail.com))  * Importing Libraries * Loading datasets * Feature Engineering * Model Development * Decision Tree Classifier * Gradient Boosting regressor * Observations * Conclusions |
| **Please paste the GitHub Repo link.** |
| Github Link(Shubham Dukare):-  <https://github.com/shubham4955/Coronavirus-tweet-sentiment-analysis>  Github link(Anjali Tidke):-  <https://github.com/AnjaliTidke/Coronavirus_Tweets>  Anjali Tidke Drive Link:-  <https://drive.google.com/drive/folders/1_PQxnerdWoj8_pkn4c7iyplGy2M0EEaL?usp=sharing>  Shubham Drive Link:-  <https://drive.google.com/drive/folders/120LaBO9_pzk1iJtEoNuLoXtiM_lbwVua?usp=sharing> |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)**   | **Problem Statement**: The problem statement is to build a classification model to predict the sentiment of COVID-19 tweets.  **Introduction**: COVID-19 originally known as Coronavirus Disease of 2019, has been declared as a pandemic by the World Health Organization (WHO) on 11th March 2020.    Sentiment Analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic is positive, negative, or neutral  Project work flow:   1. Importing Libraries 2. Loading the Dataset 3. EDA on features 4. Feature Engineering 5. Model Training 6. Performance metrics and training 7. Conclusion   EDA work Result:   * Dataset have total 41157 rows and 6 columns. * We dropped the null values from location column. * "UserName" and "ScreenName" have only integer values. Which does not provide any meaningful information. * For modeling, we required only two columns "OriginalTweet" and "Sentiment". * We removed specil characters from original tweets column. * After dropping and adding a new column, now we have 7 columns and 32567 rows. * There are five types of sentiments - Extremely Positive, Positive, Extremely Negative, Negative and Neutral. * We replaced Extremely positive to positive and Extremely negative to negative column. * The graphical representation of top 10 locations shows us that most of the tweets came from London followed by United States.   Total Regression we have used four models here –   1. Logistic Regression Model 2. Decision Tree Classifier 3. Random Forest Classifier 4. Gradient Boosting Classifier.   We have built models for multiclass classification and binary class classification and in binary class classification models perform very well compared to the multiclass classification model. Out of four models, the best model for this dataset We can see the better accuracy is with Logistic Regression that is 77%.  . | | --- | |