**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| Sachin Yallapurkar  sachin.almabetter@gmail.com  Individual Project |
| **Please paste the GitHub Repo link.** |
| Github Link:- <https://github.com/SachinYallapurkar/Coronavirus_Tweet_Sentiment_Analysis> |
| **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 Corona Virus disease that currently affects the world, cause serious problems to the population at all levels: economic, emotional, status, planning, politics, etc., in addition to the complexity of traditions, ethics, individual psychology and social behavior of people. Therefore, it is required and necessary a people's attitudes analysis when adverse situations arise Identifying people's reaction to this threat can provide important information on how society behaves and reacts to unwanted and unexpected situations, which can be positive or negative, currently the Internet and social networks have become powerful tools to access people’s opinions and comments on various topics  The main objective is to make a predictive model, which could help in predicting the Sentiment of a tweets.  **Approaches involved**   1. Importing Analytical necessary library classes for future analysis. 2. Reading the csv data file from Google drive. 3. Setting figure size for future visualization. 4. Removing future warnings in seaborne plots. 5. Visualizing all the columns of the respective Data frame. 6. Viewing all data information 7. Checking the Unique values in the column (if any) 8. Converting the data types to similar objects as the Analysis Demands. 9. Formatting the “size” column into a single column in the dataset. 10. Eradicating special characters from the dataset columns.   **ALGORITHMS** **Naive Bayes****Stochastic Gradient Descent**  1. **Random Forest** 2. **Support Vector Machine** 3. **Logistic Regression** 4. **Catboost**   **On EDA**   * Original dataset contains 6 columns and 41157 rows. * ‘Location’ column contains approx. 20.87% of Null values. * The columns such as “UserName” and “ScreenName” does not give any meaningful insights for our analysis. * In order to analyse the data we required only two columns "OriginalTweet" & "Sentiment". Hence, to avoid NaN values in "Location" columns we didnot used it further. * There are five types of sentiments- Extremely Negative, Negative, Neutral, Positive and Extremely Positive. So, we merged Extremely Positive with positive and Extremely Negative with Negative. And use encoding with value ‘-1’ for negative, ‘0’ for neutral and ‘1’ for positive. * All tweets data collected between months of March and April 2020 and of around 30 days. * Most of the tweets came from London followed by U.S. * Among top 10 mentions in tweets realDolandTrump was the top mentioned name and "#coronavirus" was most trendiest hashtag that was trending during that period.   **On Model Training**   * At the end we conclude that in our project with 6 models namely Naive Bayes Classifier,Stochastic Gradient Descent, Random Forest Classifier,Support Vector Machine, Logistic Regression and CatBoost. We are getting the highest test accuracy of about 80.98% with Stochastic Gradient Descent. |