**Sentimental Analysis on Twitter Data of Vaccination related tweets Project Report**

Team Number - 56

Student Name 1: Merin Varghese (B1)

Register Number 1: 23MCA1014

Student Name 2: Sarah Maria Toms (B2)

Register Number 2: 23MCA1065

# **Introduction:**

Vaccination, a cornerstone of modern medicine, has historically been vital in preventing infectious diseases and saving lives. However, the development and deployment of COVID-19 vaccines face unique hurdles, from scientific complexities to logistical and ethical issues. Public acceptance of vaccination programs is essential for their success, influenced by factors like trust, scientific knowledge, and social influences.

Social media platforms, notably Twitter, play a significant role in shaping public opinion on vaccination. This paper aims to analyze sentiments expressed on Twitter regarding vaccinations, particularly COVID-19 vaccines. Understanding public sentiment is crucial for tailoring effective communication strategies, identifying emerging trends, and combating misinformation.

Using natural language processing techniques, we analyze a vast dataset of vaccination-related tweets, categorizing them into positive, negative, or neutral sentiments. We explore temporal and geographical variations, identifying key themes and influencers shaping the discourse on Twitter.

The findings have implications for public health communication, vaccine advocacy, and policy interventions. By understanding sentiment dynamics on Twitter, we aim to build trust, combat misinformation, and promote vaccination as a vital tool in fighting COVID-19 and other infectious diseases.

# **Objectives:**

The main objectives of the project are as follows:

1. Conduct sentiment analysis on a dataset comprising tweets from Indian users regarding COVID-19 vaccination.
2. Evaluate the prevailing sentiment (positive, negative, or neutral) towards COVID-19 vaccines among Indian Twitter users.
3. Develop and train machine learning models to accurately predict sentiment in Indian tweets related to COVID-19 vaccination, employing techniques such as Natural Language Processing (NLP) and sentiment analysis algorithms.
4. Analyse the effectiveness of different communication strategies and interventions in shaping public opinion and promoting vaccine acceptance among Indian Twitter users.

# **Dataset:**

The dataset utilized in this study comprises two separate datasets sourced from Kaggle, each containing the same columns related to vaccination discourse on Twitter. Through a process of integration, these datasets were combined to create a unified dataset for analysis. The integration process involved merging the datasets based on common identifiers or attributes, ensuring that each record from both datasets was appropriately matched and consolidated. By combining multiple datasets, we aimed to enhance the comprehensiveness and representativeness of the data, capturing a broader range of perspectives and opinions expressed across different sources on Twitter. This integrated dataset provides a rich and diverse source of information for conducting exploratory data analysis and exploring the dynamics of vaccination-related discussions on Twitter. There are 78,168 rows of data with 16 feature columns. Anyways for sentiment analysis, we are creating a new dataset from this integrated dataset which includes the data related to India only. The new dataset has 9900 rows.

# **Exploratory Data Analysis:**

The exploratory data analysis (EDA) conducted in this study involved a comprehensive examination of various aspects of the integrated dataset containing vaccination-related tweets sourced from Twitter. Through EDA, we sought to gain insights into the characteristics and patterns present in the data, as well as to identify any trends or anomalies that may inform subsequent analysis. The EDA process included descriptive statistics to summarize key features of the dataset, such as the distribution of tweet sentiments, frequencies of different keywords or hashtags, and temporal trends in tweet volume. Visualization techniques, including histograms, box plots, and time series plots, were employed to visually explore the distribution and relationships between variables in the dataset. Additionally, we conducted text mining and sentiment analysis to extract meaningful insights from the textual content of tweets, uncovering prevalent themes, sentiments, and topics within the vaccination discourse on Twitter. Overall, the EDA process provided a foundational understanding of the dataset, guiding further analysis and interpretation of the findings. Figure 1 shows the visualization for count of user names in the integrated dataset. Figure 2 shows the visualization for the count for user locations in the integrated dataset. It is found that most of the tweets i.e. about 25,000 are tweeted from Los Angeles, and India coming after that. Figure 3 shows the number of users created from the year of 2006 to 2022. It is clearly visible that there is a spike in the number of user creation in the years of 2019, 2020 and 2021 which were the years of COVID-19. The figures 4,5 and 6 show the prevalent words through word clouds in different situations.

Fig 1. Count of User names in the integrated dataset

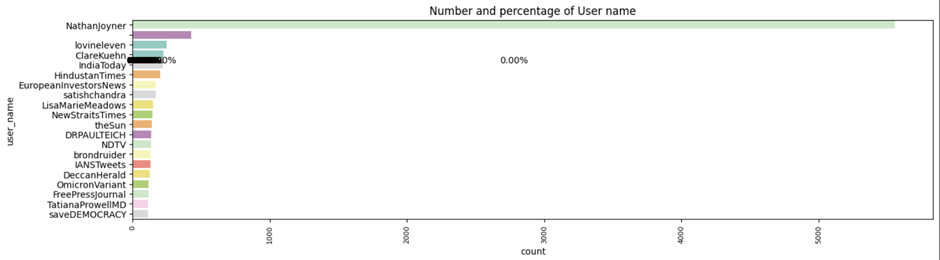


Fig 2. Count of User location in the integrated dataset

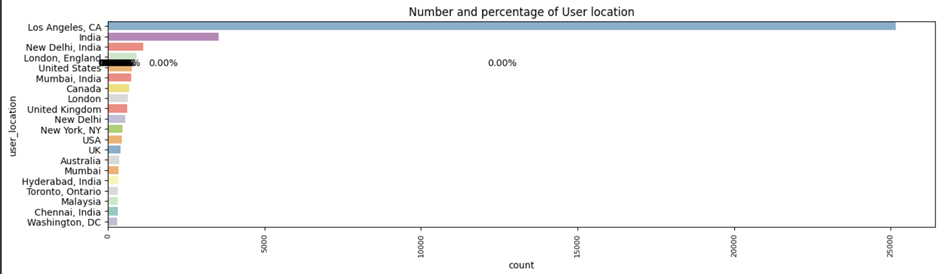


Fig 3. Count of Users created in each year in the integrated dataset

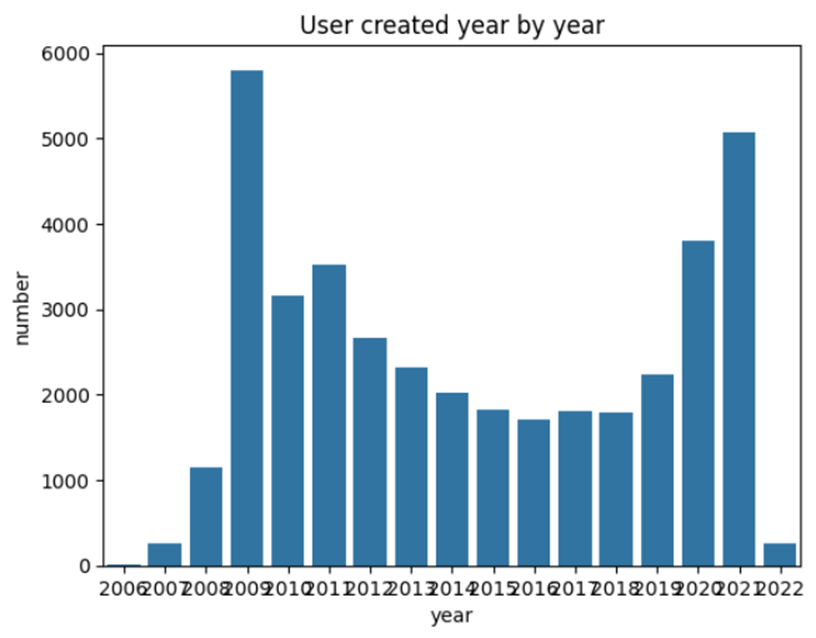


Fig 4. Word cloud to show the prevalent words in the hashtags in the integrated dataset



Fig 5. Word cloud to show the prevalent words in the tweets in the integrated dataset

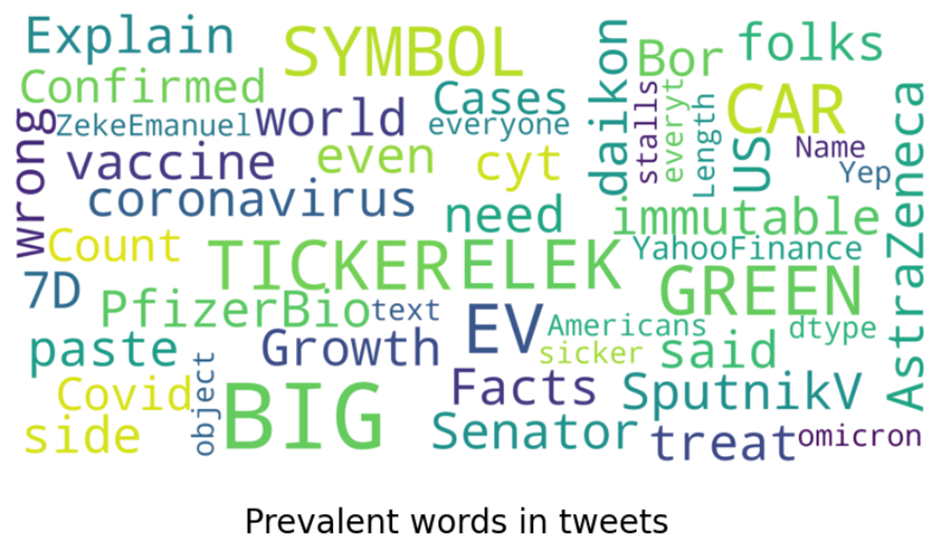
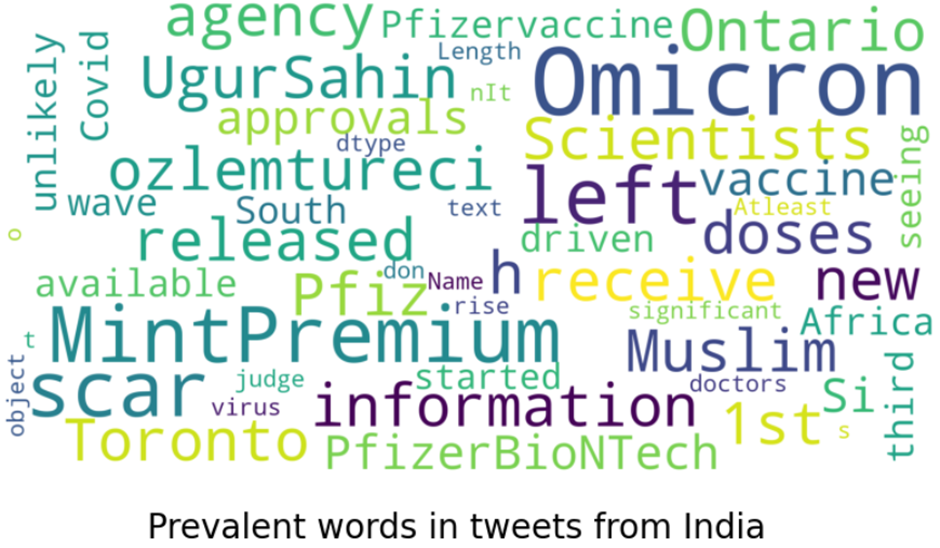


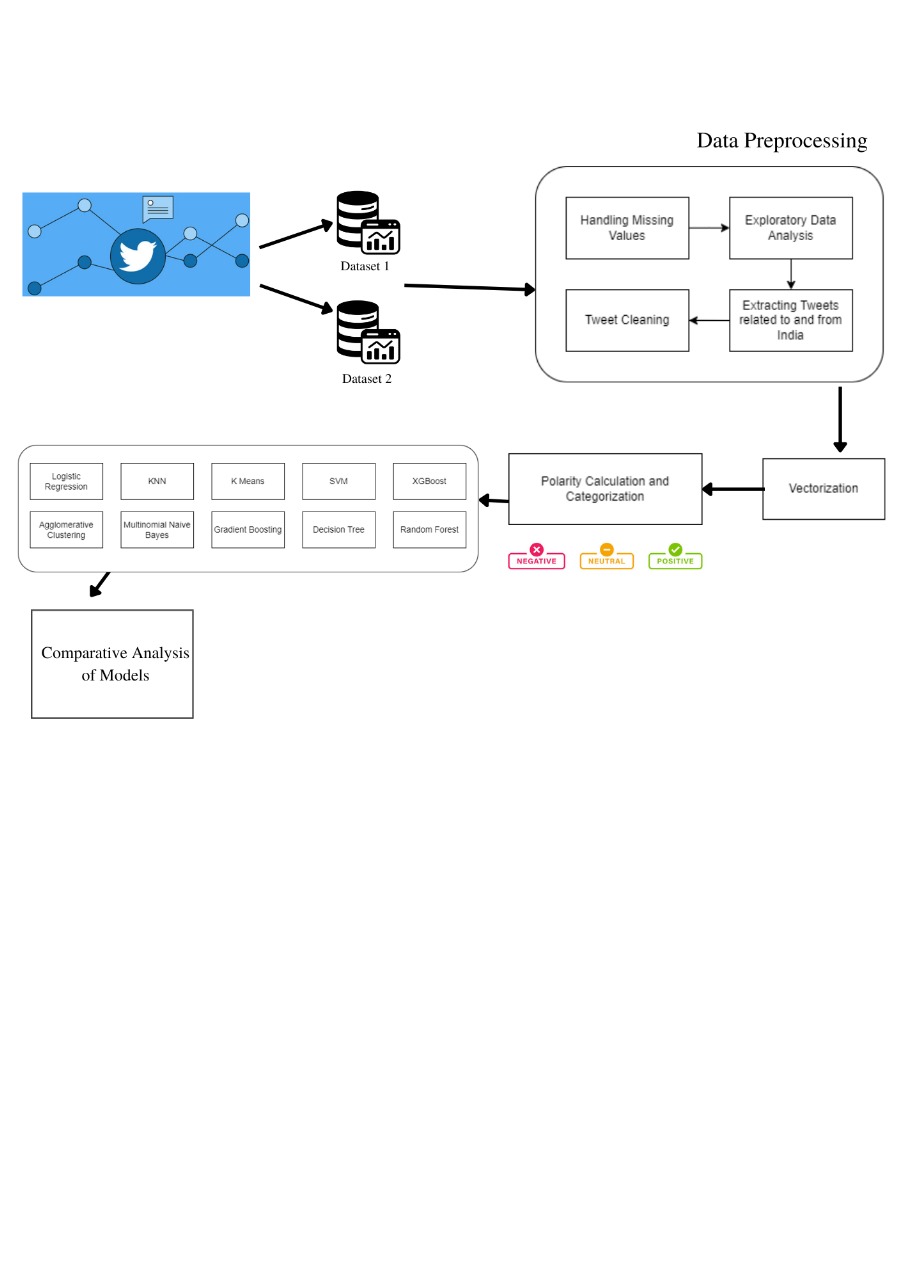
Fig 6. Word cloud to show the prevalent words in the tweets in the Indian dataset



# **Methodology**

Twitter data related to COVID-19 vaccination was collected using the Twitter API, encompassing usernames, locations, descriptions, tweet text, and hashtags. This data underwent preprocessing steps, including handling missing values, cleaning text, and refining features such as user descriptions and hashtags. Techniques such as lemmatization, stopword removal, and regular expressions were utilized for text data cleaning. Exploratory data analysis (EDA) provided insights into features like user locations, names, and tweet lengths. Visualizations such as word clouds were employed to identify prevalent words and topics in tweets and hashtags. Since the project is concentrating only on tweets related to India, a new dataset is created. Sentiment analysis using the VADER sentiment analysis tool determined tweet sentiment (positive, negative, or neutral), with sentiment scores calculated for each tweet and overall sentiment distribution analyzed. Machine learning models including Logistic Regression, Decision Tree, Random Forest, SVM, Multinomial Naive Bayes, Gradient Boosting, KNN, and XGBoost were trained and tested for sentiment prediction, with accuracy and classification reports used for evaluation. Clustering techniques like K-means and hierarchical clustering grouped similar tweets based on content, with evaluation metrics such as silhouette score and inertia assessing clustering performance. Finally, a comparative analysis was conducted on machine learning models, excluding clustering algorithms, to determine the best-performing model. Figure 7 shows the methodology process followed for this project.

Fig 7. Methodology flowchart



# **Implementation**

* Importing CSV Files from Drive: The project begins with importing CSV files containing the Twitter data from Google Drive using the Google Colab environment.
* Data Preprocessing: Data preprocessing steps include handling missing values, cleaning text, and refining features such as user descriptions and hashtags. Techniques like lemmatization, stopword removal, and regular expressions were applied to clean the text data. Also, for the frequent emojis, their expressions or feelings were given as shown in Figure 8.

Fig 8. Emojis and their meanings/feelings



* Feature Exploration: Exploratory data analysis (EDA) was conducted to gain insights into various features such as user locations, user names, and tweet lengths. Visualizations such as word clouds were used to identify prevalent words and topics in tweets and hashtags.
* Sentiment Analysis: Sentiment analysis was performed using the VADER sentiment analysis tool to determine the sentiment of tweets. The sentiment scores were calculated for each tweet, and the overall sentiment distribution was analyzed.
* Model Training and Testing: Several machine learning models were trained and tested on the dataset to predict the sentiment of tweets. Models such as Logistic Regression, Decision Tree, Random Forest, SVM, Multinomial Naive Bayes, Gradient Boosting, KNN, and XGBoost were evaluated based on their accuracy and classification reports. Figures 9.1,9.2,9.3,9.4,9.5,9.6,9.7 and 9.8 shows the code user for all the above models/algorithms.

Fig 9.1 Logistic Regression Code

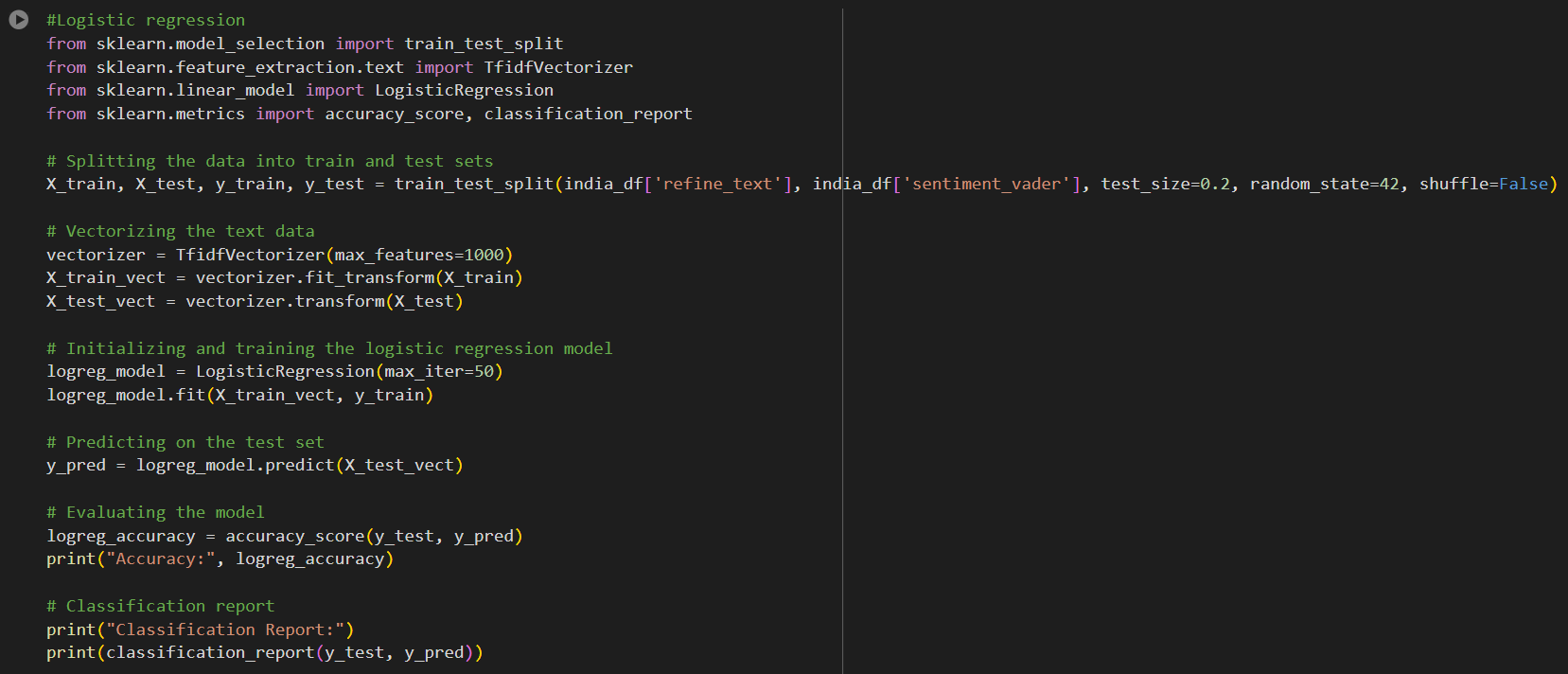


Fig 9.2 Decision Tress Code

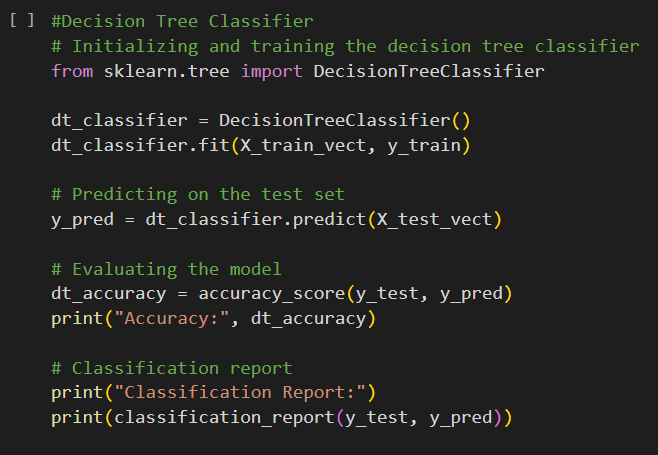


Fig 9.3 Random Forest Code

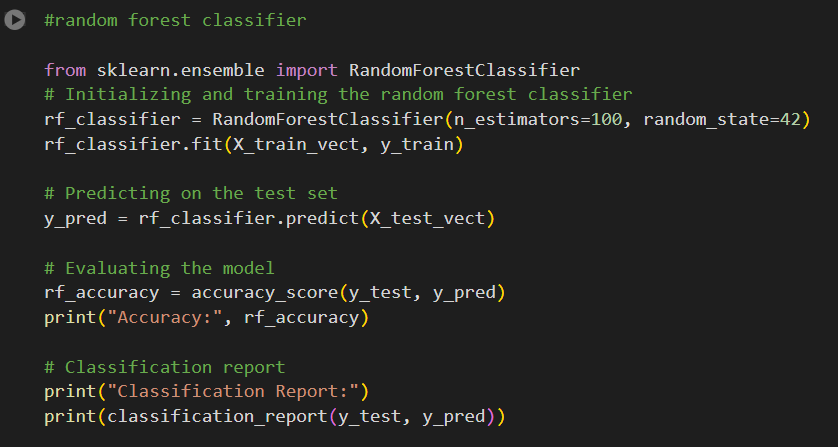


Fig 9.4 SVM Code

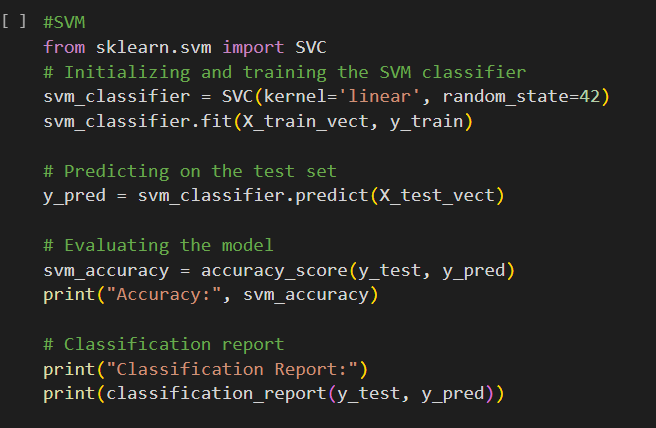


Fig 9.5 Naïve Bayes (multinomial) Code

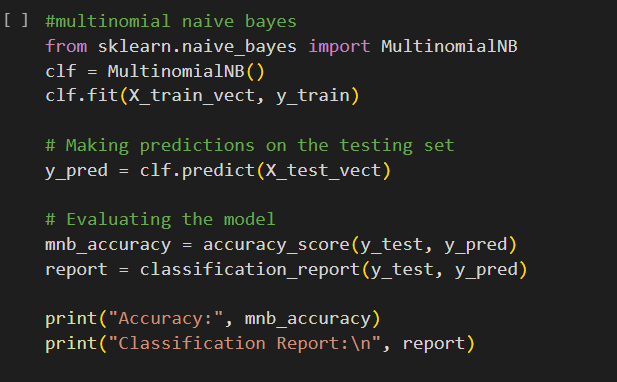


Fig 9.6 Gradient Boosting Code

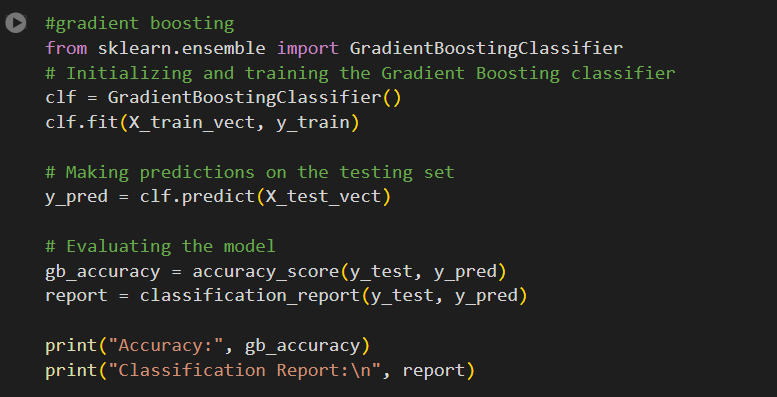


Fig 9.7 KNN Code

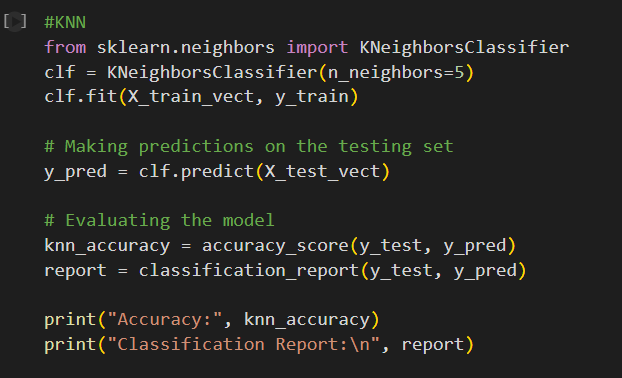
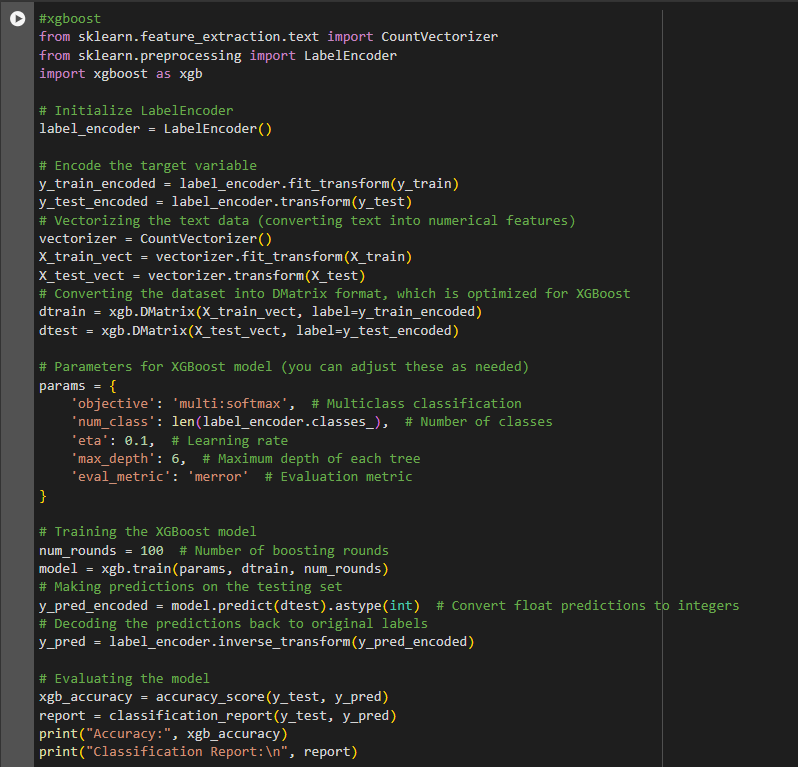


Fig 9.8 XGBoost Code



* Clustering Analysis: Clustering analysis was conducted to group similar tweets together based on their content. K-means clustering and hierarchical clustering techniques were applied to identify clusters within the dataset.

Fig 9.9 KMeans Code

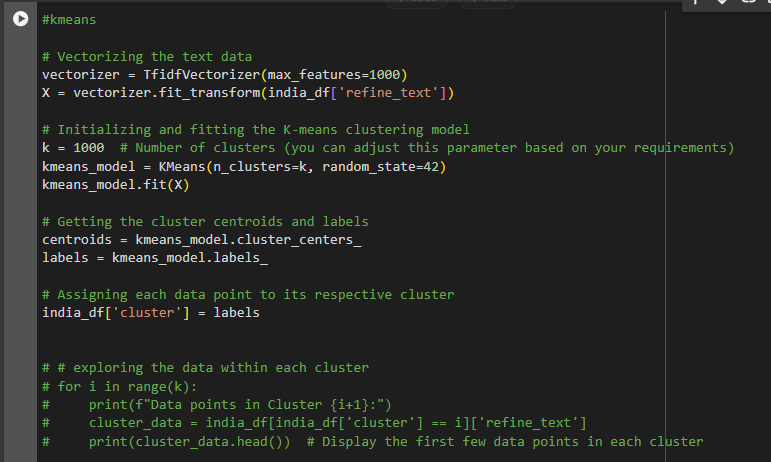


Fig 9.10 Agglomerative Clustering Code

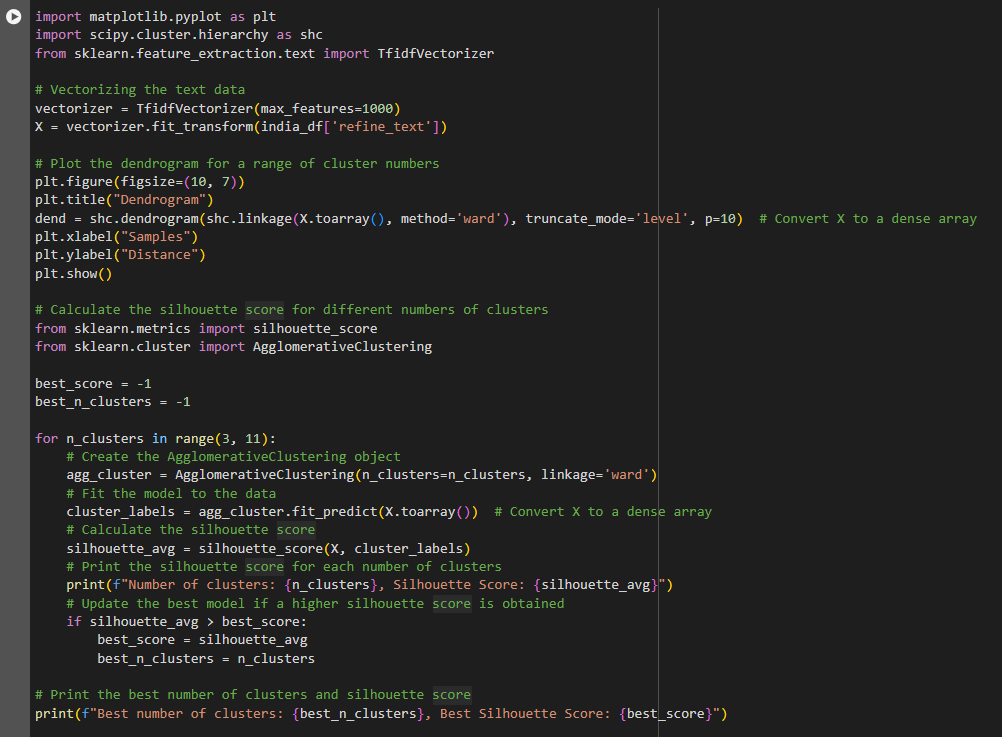
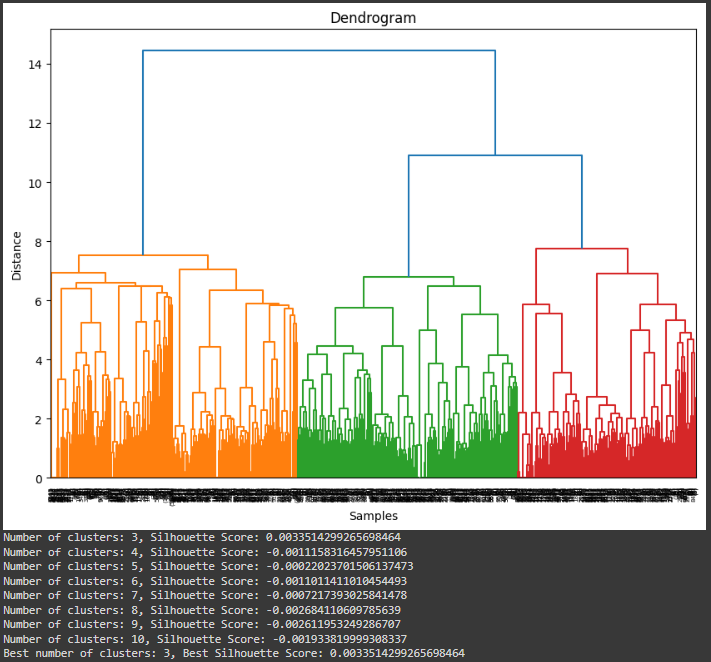


Fig 9.11 Agglomerative Model – Dendrogram



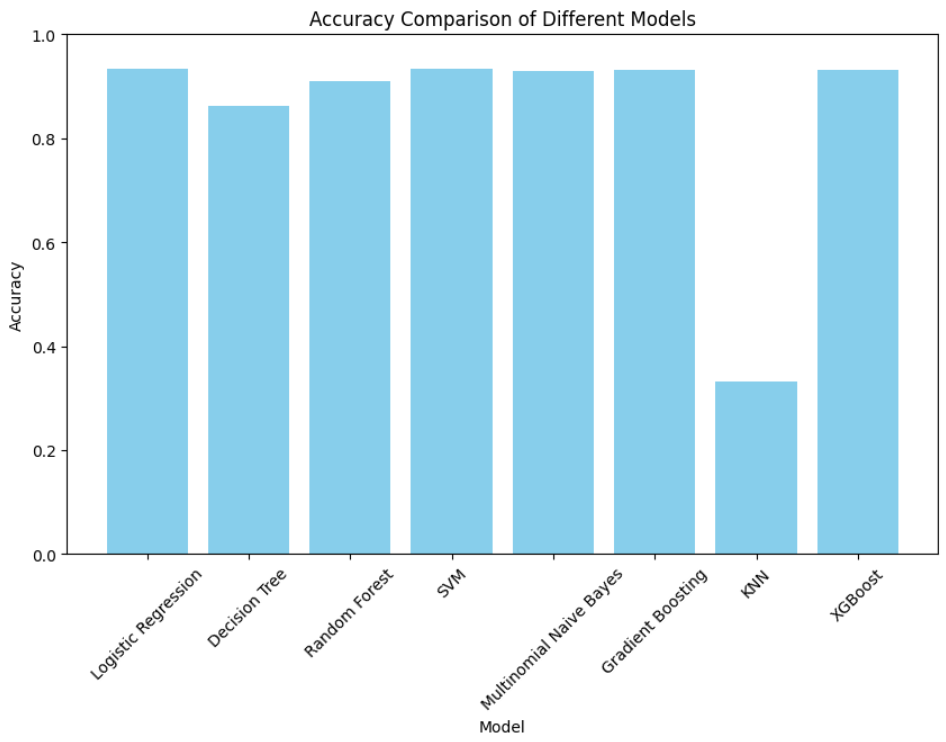
# **Comparative Analysis**

A comparative analysis was performed to compare the performance of different machine learning models in predicting tweet sentiment. The accuracy and classification reports of each model were evaluated to determine the most effective model for sentiment analysis. Table 1 shows the different models used in this project and their respective accuracy scores and silhouette scores. Similarly, Figure 9 shows the visualization for this comparative analysis.

Table 1. Models and their respective accuracy scores and silhouette scores

|  |  |  |  |
| --- | --- | --- | --- |
| ***SI. No*** | ***Algorithm/Model*** | ***Accuracy*** | ***Silhouette Score*** |
| 1. | Logistic Regression | 93.2828 | N/A |
| 2. | Decision Trees | 86.1616 | N/A |
| 3. | KNN | 33.1313 | N/A |
| 4. | SVM | 93.4343 | N/A |
| 5. | Naive Bayes (multinomial) | 92.8787 | N/A |
| 6. | Gradient Boosting | 93.08 | N/A |
| 7. | Random Forest | 90.9596 | N/A |
| 8. | XGBoost | 93.232 | N/A |
| 9. | KMeans | N/A | 0.1244 (for k=1000) |
| 10. | Agglomerative Clustering | N/A | 0.00335 (for k=3) |

Fig 10. Comparative Analysis for the Algorithms/Models



# **Findings:**

The findings of the project are as follows:

1. Sentiment Analysis: The sentiment analysis revealed that the majority of tweets from Indian users regarding COVID-19 vaccination were neutral.
2. Trending Topics: Analysis of trending topics in vaccination-related tweets highlighted key concerns and interests among Indian users. Common topics included vaccine efficacy, side effects, vaccine availability, and government vaccination policies.
3. Influential Users: Certain influential users, including public figures, healthcare professionals, and media outlets, played a significant role in shaping the discourse on COVID-19 vaccination in India. Their endorsements and opinions often influenced public sentiment and behavior towards vaccination.
4. Machine Learning Models: The machine learning models developed for sentiment prediction demonstrated high accuracy in classifying sentiment in Indian tweets. Models such as SVM and Logistic Regression performed particularly well in predicting sentiment.
5. Impact of Communication Strategies: The analysis suggested that effective communication strategies, such as clear and transparent information dissemination, targeted messaging, and engagement with trusted sources, were instrumental in fostering positive sentiment and vaccine acceptance among Indian Twitter users.
6. Policy Implications: The findings provide valuable insights for policymakers and public health authorities to tailor communication strategies and interventions aimed at addressing vaccine hesitancy, enhancing vaccine uptake, and countering misinformation in India. Strategies focusing on community engagement, addressing concerns, and building trust are crucial for promoting vaccine acceptance and achieving vaccination goals.

# **Conclusion:**

The project provides valuable insights into public sentiment regarding COVID-19 vaccination on Twitter. By analyzing tweet content and sentiment, we can better understand public opinions, concerns, and attitudes towards vaccination. Overall, the project contributes to enhancing our understanding of public discourse surrounding vaccination during the COVID-19 pandemic.

# **Recommendations:**

1. Diversify Data Sources: Consider incorporating data from multiple social media platforms beyond Twitter, such as Facebook, Instagram, and Reddit, to obtain a more comprehensive understanding of public sentiment towards COVID-19 vaccination in India.
2. Real-time Monitoring: Establish a real-time monitoring system to continuously track and analyze vaccination-related discourse on social media platforms. This can enable timely identification of emerging issues, misinformation, and sentiment trends, facilitating prompt intervention and response.
3. Collaboration with Health Authorities: Collaborate with public health authorities, research institutions, and non-governmental organizations (NGOs) working in the field of public health communication to leverage their expertise, data resources, and communication channels for disseminating accurate information and promoting vaccine acceptance.

# **References:**

1. <https://www.kaggle.com/datasets/gpreda/omicron-rising>
2. https://www.kaggle.com/datasets/gpreda/pfizer-vaccine-tweets
3. <https://www.projectpro.io/recipes/create-word-cloud-python>
4. https://www.analyticsvidhya.com/blog/2021/06/text-preprocessing-in-nlp-with-python-codes/
5. <https://www.datacamp.com/tutorial/stemming-lemmatization-python>
6. https://medium.com/@rslavanyageetha/vader-a-comprehensive-guide-to-sentiment-analysis-in-python-c4f1868b0d2e
7. <https://www.analyticsvidhya.com/blog/2021/06/vader-for-sentiment-analysis/>
8. https://scikit-learn.org/