**MACHINE LEARNING MINI PROJECT**

**TITLE:** PREDICTING DISEASE OUTBREAKS FROM SOCIAL MEDIA

**ABSTRACT:**

Through the examination of social media data, we hope to use machine learning techniques in this research to forecast disease outbreaks. Since the introduction of social media platforms, a great amount of information—including talks about health issues—has been exchanged in real time. We want to use this data to build a predictive model that can show possible disease outbreaks in particular regions before they happen. In our method, tweets from Twitter are gathered and processed, relevant attributes are extracted, prediction models are trained, and the anticipated outbreak zones are visualized.

**OBJECTIVES:**

* collect and prepare information from social media about various illnesses.
* Apply feature engineering to the text data to draw out important information.
* Create a machine learning model that uses social media usage to forecast illness outbreaks.
* Utilize the right metrics to assess the performance of the model.
* Make visual representations of expected outbreak locations to contrast with actual reported occurrences.

**REFERENCES:**

1. Nagar R., Yuan Q., Freifeld C.C., Santillana M., Nojima A., Chunara R., Brownstein J.S. (2014). A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives. Journal of Medical Internet Research, 16(10), e236.

2. Chew C., Eysenbach G. (2010). Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. PLoS ONE, 5(11), e14118.

3. Paul M.J., Dredze M. (2011). You are what you tweet: Analysing Twitter for public health. Proceedings of the Fifth International AAAI Conference on Weblogs and social media.

4. Culotta A. (2010). Towards Detecting Influenza Epidemics by Analysing Twitter Messages. Proceedings of the First Workshop on Social Media Analytics.

**SOURCES:**

The Twitter Disease Surveillance Dataset utilized for this project was compiled from a number of publicly available datasets and research initiatives on disease surveillance that included tweets about health and diseases. The dataset, which may be acquired through APIs and data-sharing sites, consists of tweets, metadata, and timestamps.

**GROUP INFORMATION:**

NAME: *Shreyas Sudeer*

REGISTRATION NUMBER: *210962001*

NAME: *Bhavana Kedari*

REGISTRATION NUMBER: *210962077*

CLASS: CSE AI & ML