**Project Proposal:**

**Sentiment Analysis for Early Detection of Mental Health Status**

**Introduction**

The increasing prevalence of mental health disorders necessitates innovative approaches for early detection and timely intervention. Leveraging advancements in Natural Language Processing (NLP), this project aims to develop a multi-class classification model to identify various mental health statuses from textual data. By analyzing statements labeled with mental health conditions, the model will assist healthcare providers in recognizing signs of mental health issues, facilitating proactive measures for patient care.

**Problem Statement**

Accurate and early detection of mental health conditions is critical for effective treatment and intervention. Traditional methods of diagnosis can be time-consuming and subjective, often relying heavily on in-person assessments. By utilizing a comprehensive dataset of over 50,000 statements, this sentiment analysis aims to enable healthcare providers to recognize potential mental health issues and take appropriate actions as early as possible.

**Objectives**

1. **Develop a Multi-Class Classification Model**: Build and implement a predictive model to accurately classify various mental health statuses, such as "Normal," "Depression," "Suicidal," "Anxiety," "Bipolar," "Stress," and "Personality disorder," based on textual statements.
2. **Explore and Evaluate Machine Learning Algorithms**: Investigate and compare the performance of Bernoulli Naive Bayes, Logistic Regression, and Decision Tree Classifier to identify the most effective model for sentiment analysis in mental health.
3. **Facilitate Early Detection of Mental Health Issues**: Create a prototype tool that enables healthcare providers to monitor and trigger timely interventions based on predicted mental health statuses, ultimately improving patient care outcomes.

**Dataset Overview**

The dataset comprises over 50,000 statements derived from various sources and labeled with distinct mental health statuses. It includes the following features:

**unique\_id**: A unique identifier for each entry.

**statement**: The textual data or post representing an individual's mental health status.

**mental\_health\_status**: Categorical labels indicating the mental health condition, including:

i. **Normal**, ii. **Depression**, iii. **Suicidal**, iv. **Anxiety**, v. **Bipolar**, vi. **Stress**, vii. **Personality Disorder**

The dataset has been compiled by **Suchintika Sarkar** and integrates information from multiple Kaggle datasets, ensuring a rich and diverse resource for analysis.

**Methodology**

1. **Data Acquisition**:
   * Obtain the comprehensive dataset of over 50,000 statements tagged with various mental health statuses, as compiled by **Suchintika Sarkar** from multiple sources on Kaggle.
2. **Data Preprocessing**:
   * **Data Cleaning**: Implement a custom function to clean the text data by:
     + Converting text to lowercase.
     + Removing HTML tags and URLs.
     + Eliminating special characters, punctuation, and numeric words.
     + Reducing sequences of repeated characters.
     + Expanding contractions for better readability.
     + Removing newline characters and extra spaces.
   * **Stop word Removal**: Remove common stop words to focus on significant terms that contribute to sentiment.
3. **Feature Engineering**:
   * Extract and create additional features to enhance the dataset:
     + **Character Count**: Count of characters in each statement.
     + **Word Count**: Number of words in each statement.
     + **Sentence Count**: Total number of sentences in each statement using sentence tokenization.
     + **Average Character per Word**: Calculate the average number of characters per word.
     + **Average Characters per Sentence**: Calculate the average number of characters per sentence.
     + **Average Words per Sentence**: Calculate the average number of words per sentence.
4. **Data Tokenization and Stemming**:
   * Perform tokenization to convert the cleaned text into individual tokens (words).
   * Implement stemming to reduce words to their base or root form.
5. **Model Development**:
   * **Split Dataset**: Divide the dataset into training and testing sets to evaluate model performance.
   * **Implement Machine Learning Models**: Train and evaluate the classifiers:
     + **Bernoulli Naive Bayes:** A probabilistic model effective for binary and multi-class classification.
     + **Logistic Regression:** A straightforward model that provides interpretability and serves as a baseline for comparison.
     + **Decision Tree Classifier:** A model that captures non-linear relationships and allows for easy visualization of decision-making processes.
6. **Model Evaluation**:
   * Use accuracy as the primary evaluation metric to assess model performance.
   * Present confusion matrices to illustrate model performance in classifying mental health statuses.
   * Compare the accuracy of the different models to determine the most effective classifier for identifying mental health statuses.
7. **Implementation of Early Detection Tool**:
   * Develop a prototype tool that healthcare providers can use to monitor patients’ mental health statuses based on model predictions, facilitating timely interventions.
8. **Documentation and Reporting**:
   * Document methodologies, results, and insights gained throughout the project.
   * Prepare a comprehensive report and presentation to share findings with stakeholders, contributing to ongoing discussions about mental health awareness and intervention strategies.
9. **Expected Outcomes:**

* A robust and reliable multi-class classification model capable of accurately identifying various mental health statuses from textual statements.
* A tool that can aid healthcare providers in the early detection of mental health conditions, facilitating timely interventions and support.

**Conclusion**

This project aims to enhance mental health awareness and early intervention strategies by leveraging machine learning techniques to analyze and classify textual data. By developing a robust sentiment analysis model, the project not only addresses a critical healthcare challenge but also represents an essential step in my transition to a data science career.