Project: Suicide Detection

INTRODUCTION

Our project is designed to develop a predictive computational model using features of social media posts and identifying individuals who are at risk of suicide. This is a complex and challenging task, as there is no single factor that can predict whether or not someone will attempt suicide. However, there are a number of warning signs that can be helpful in identifying individuals who may be at risk. This project aims to help identify people who may be feeling really down or struggling with thoughts of hurting themselves by analysing their posts on social media.

NEED/SIGNIFICANCE

Understanding the significance of this project lies in its potential to contribute to mental health awareness by leveraging technology to identify and assist individuals in distress. With the prevalence of social media, this model serves as a proactive tool, offering timely intervention and support to those expressing signs of suicidal thoughts on online platforms.

1. Early identification and intervention
2. Reducing suicide rates and saving lives
3. Enhancing social support
4. Provide access to mental health
5. Provide Vulnerable individuals

METHODS USED

1. Data Collection and Preprocessing
   We utilized web scraping techniques to collect tweets related to suicide. To refine the text data, we applied natural language processing (**NLP**) techniques, including removing punctuation, stop words, and converting text to lowercase. This preprocessing step improved the quality of the data for further analysis.
   We applied **WordCloud** to count the maximum number of terms in tweets.
2. Vectorizer
   These features were then vectorized using **TF-IDF** which is technique for representing text data in a numerical format suitable for machine learning algorithms.
   We also performed Sentiment Analysis using **TextBlob** i.e. a python library for processing textual data for Sentiment analysis.

3. Model Selection

   Machine Learning Models used:

   a. Naive Bayes (Voting Classifier)
      - Gaussian Naive Bayes (GaussianNB)
      - Bernoulli Naive Bayes (BernoulliNB)
      - Multinomial Naive Bayes (MultinomialNB)

We have used these Naïve Bayes Classfiers for text classification and evaluating our model giving the best accuracy.

b. Random Forest- ML algorithm belonging to the ensemble learning method. Widely used for classification and regression tasks handling complex data, reduce overfitting, and provide robust predictions.

c. Decision Tree - A decision tree recursively splits the dataset into subsets based on the most significant attribute at each node resulting in a tree-like structure.

d. Gradient Boosting - Gradient Boosting builds a series of weak learners sequentially, with each new tree correcting the errors of the previous ones.

e. XG Boost - XGBoost stands for Extreme Gradient Boosting, an open-source software library for gradient boosting. It is a popular choice for machine learning tasks because of its high accuracy, speed, and scalability. XGBoost is particularly well-suited for handling large datasets and complex models.

f. K-Nearest Neighbors (KNN) - KNN classifies a new data point based on the majority class of its k-nearest neighbors in the feature space

g. Support Vector Classifier (SVC) - SVC is a supervised learning algorithm that classifies data points into different categories by finding the hyperplane that best separates the classes.

4. Interactive Widget:
We applied an interactive **ipywidget** which act as a live prompt and can respond to input text and giving the output.

DATASET DESCRIPTION

A comprehensive dataset, curated for Twitter tweets sourced from Kaggle and Reddit subreddits, has been compiled and merged into a file named 'Suicide_Detection.csv'.

The dataset, gathered using AI tools, specifically browse.ai, comprises 233,406 entries.

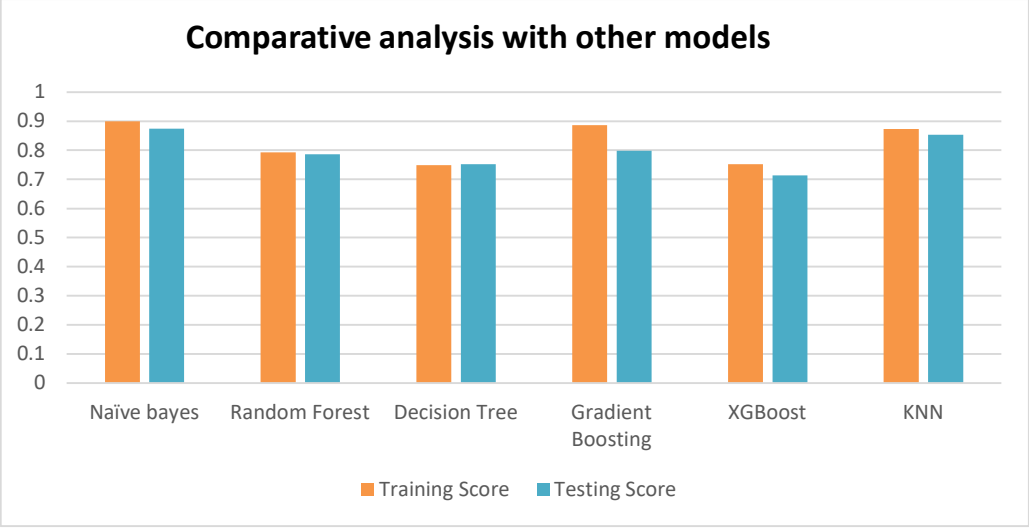
The primary attributes include:

- Text: Represents tweets or subreddit content extracted from individuals' social media posts regarding their emotions and sentimental thinking.

- Class: Indicates whether the content is associated with suicidal or non-suicidal themes.
  It's important to note that the dataset is in its raw and unprocessed form, containing both missing and duplicate values. The dataset can be further divided into training and testing datasets for machine learning purposes.
  Additionally, the dataset offers a unique insight into individuals' emotional and sentimental thinking as expressed through their social media posts.

EXPERIMENATL RESULTS AND COMPARITIVE ANALYSIS WITH OTHER MODELS

Model	Training Score	Testing Score
Naïve bayes	0.8992	0.8753
Random Forest	0.7931	0.7856
Decision Tree	0.7485	0.7526
Gradient Boosting	0.8865	0.7986
XGBoost	0.7522	0.7133
KNN	0.8729	0.8526



In evaluating our suicide detection project, we find that Naïve Bayes demonstrates the highest training and testing scores, leading us to conclude it as the most suitable model.