MAJOR PROJECT SUMMARY

Batch 3
Class: ML063B5

Description:- We were provided a dataset called Information.csv which contained information about profiles of twitter users. We were asked to use “gender” as the dependent variable and we removed the unused features.

This project has been done in two ways- first, using only numerical values and second using only text.

Data plays a vital role in the project. proper data cleansing can either make or break the project.

Larger portion of our time was invested in cleaning the data. In data cleaning, the null values have been removed and feature is selected in a way that improves the quality of data.

Label encoding was performed on the columns gender, profile_yn and new columns were created as gender1 and profile_yn1 respectively, so that it is easier for modeling and performing data visualisation.

Data visualization has been performed for better understanding of the data.

The algorithms which were chosen for ensemble learning was based on their accuracies.

* The selected features are _unit_id, _golden, gender, gender:confidence, profile_yn, profile_yn:confidence, description, fav_number, name,retweet_count, text, tweet_count.

* we've balanced the data set and plotted the outliers. we've performed natural language processing using the NLTK to clean the tweet set.

* Algorithms were trained in two ways
    * Text based
    * Numerical based

Questions framed are:-

1. What are the most common emotions/words used by Males and Females?

2. Which gender prefers prime numbers as their favourite number?

Algorithms trained using numericals :-

1. Logistic Regression
2. Decision Tree Classifier
3. Gaussian Naive - Bayes
4. Random Forest Classifier
5. K-Nearest Neighbours
6. voting classifier for ensemble modelling

Voting classifier for ensemble modelling :-

voting classifier is trained using 3 algorithms.

1. Logistic Regression
   Accuracy:- 52%

2. Random Forest Classifier
   Accuracy:- 58%

3. Decision Tree Classifier
   Accuracy:- 53%

TOTAL ACCURACY:- 57%

Algorithms trained using Text :-

1. Multinomial Naive-Bayes
2. Decision Tree Classifier
3. Random Forest Classifier
4. Gaussian Naive - Bayes
5. Logistic Regression
6. voting classifier for ensemble modelling

Voting classifier for ensemble modelling :-

voting classifier is trained using 3 algorithms.

1. Logistic Regression
   Accuracy:- 58%

2. Random Forest Classifier
   Accuracy:- 56%

3. Multinominal Naive-Bayes
   Accuracy:- 59%

TOTAL ACCURACY:- 58%

Q1. What are the most common emotions/words used by Males and Females?
ANS:-

The most common word used by male:
LIKE
GET
LOVE
DAY

The most common words used by female:
LIKE
GET
ONE
TIME
GO

Q2. Which gender prefers prime numbers as their favourite number?
ANS:-

The output shows that males prefer prime numbers as their favourite number when compared to females.