MAJOR PROJECT SUMMARY :

- The major project is performed on the provided data set **'information.csv'** .
- The data set has the following features :

'_unit_id'
'_golden'
'_unit_state'
'_trusted_judgments'
'_last_judgment_at'
'gender'
'gender:confidence'
'profile_yn'
'profile_yn:confidence'
'created'
'Description'
'Fav_number'
'Gender_gold'
'Link_color'
'name',
'profile_yn_gold',
'profileimage'
'retweet_count'
'sidebar_color'

'text'
'tweet_coord'
'tweet_count'
'tweet_created'
'tweet_id'
‘tweet_location'
'user_timezone'.

From the above features I created new features:

- ‘clean_text’
- ‘clean_description’
- 'gender_num'
- ‘all_features’

<u>Accuracy Selection:</u>

- Independent Variables for this set are **'all_features'** which is the new feature of concatenation of 'text' and 'description' ; 'gender:confidence' and 'tweet_count' .

- Dependent variable is **'gender_num'** which is the label encoding of 'gender' .

- Here is a simple table showing the accuracies of each algorithm used.

S.No	Algorithm	Accuracy (%)
1	**K-Nearest Neighbor**	44.57
2	**Random Forest**	57.44
3	**Multinomial Naive Bayes**	63.63

- Multinomial Naive Bayes , KNN and Random Forest are Classification Algorithms.

- On performing Ensemble Learning **61.03%** accuracy has been found.

The best algorithm for this data set is - **MultinomialNaive Bayes.**

Questions asked on the dataset are :

Q1) What are the most common emotions/words used by Males and Females?
Ans) love, like, get, one, life are some of the common words used by males and females.

Q2) What is the time when most of the tweets are created by Males and Females?
Ans) 10/26/15 12:40.