**MAJOR PROJECT**

# Ensemble Learning

**Description of dataset**: it was gender classifier dataset with initially 20050 rows and 26 columns.

**Getting relevant information and understanding the datset**: first step was to import data set using pandas library and then using functions in pandas getting information about various columns and description about various columns

**LIBRARIES USED**: pandas, numpy, matplotlib, sklearn.

## DATA CLEANING:

1. **cleaning gender column:** removing the categorical values of gender column having "brand" and "unknown" as labels then checking if there is any null values and filling it with fillna function with method of ffill. Finally we get gender column with 12991 rows.

2. **Cleaning gender: confidence column:** removing the null values and filling it with method of ffill.

3. **Cleaning _last_judgement_at column:** removing the null values and filling it with method of ffill.

4. **Removing the gender_gold column :** as the column contains very few values therefore it is best to remove it.

5. **Removing the profile_yn_gold column:** as the column contains very few values therefore it is best to remove it.

6. **Removing th twee_coord column:** as the column contains very few values therefore it is best to remove it.

7. **cleaning the tweet_location column:** removing the null values and filling it with method of ffill.

8. **cleaning the user_timezone column:** removing the null values and filling it with method of ffill.

## EXPLORATORY DATA ANALYSIS (EDA):

**Q1).Which is the most occurring _trusted_judgements along with visualization?**

ANS).3 was the most occurring _trusted_judgements which is obtained by using pandas library along with bar plot showing the peak of most common _trusted_judgements.

**Q2).Which gender has the third highest tweet_count and also visualization of total tweet_count of female and male?**

ANS).”male” has the third highest tweet_count which is obtained by applying pandas library . Visulaization of total tweet_count gender wise is done by bar plot using matplotlib library.

## ML MODELS AND DATA PREPROCESSING:

## CLASSIFICATION ALGORITHMS :

**DATA PREPROCESSING :**

IT IS DONE ALGORITHM TO ALGORITHM WHICH INCLUDES FEATURE SELECTION AND REMOVING THE OUTLIERS WHERE IT IS NECESSARY. DIFFERENT ALGORITHM HAVE DIFFERENT DATA PREPROCESSING.

**1.LOGISTIC REGRESSION:**

a).Started with converting categorical columns into continuous column using label encoding. Converted 12 columns from categorical column to continuous column. List of columns are ["_golden", "_unit_state", "profile_yn", "link_color", "sidebar_color", "tweet_location", "user_timezone", "name", "_last_judgment_at", "created", "profileimage", "tweet_created"].

b**).Independent attributes are**: '_unit_state', '_trusted_judgments', '_last_judgment_at', 'gender:confidence', 'profile_yn:confidence', 'created', 'fav_number', 'link_color', 'name', 'sidebar_color', "profileimage", "tweet_created", "tweet_count", "tweet_id", 'user_timezone'.

c).**Dependent attribute is:** ”gender”.

d**).Gender column:** doesn’t get biased so I deleted some rows with “female” as category to keep ratio of female to male good for fitting and training the model.

e).**ACCURACY:** is approximately ~71%.

**2.SUPPORT VECTOR MACHINE(SVM):**

**a). Gender column:** doesn’t get biased so I deleted some rows with “female” as category to keep ratio of female to male good for fitting and training the model.

**b).Independent attributes**: unit_state', '_trusted_judgments', '_last_judgment_at', 'gender:confidence', 'profile_yn:confidence', 'created', 'fav_number', 'link_color', 'name', 'sidebar_color', "user_timezone","tweet_id","profileimage","tweet_created","tweet_count".

**c).Dependent attributes**: ”gender”.

**d).ACCURACY**: is approximately~72%.

**3.RANDOM FOREST:**

a).first thing is done to get the correlation between various categories using pandas library.

b).**Independent variables:** : unit_state', '_trusted_judgments', '_last_judgment_at', 'gender:confidence', 'profile_yn:confidence', 'created', 'fav_number', 'link_color', 'name', 'profileimage', 'retweet_count', 'sidebar_color', 'tweet_count', 'tweet_created','tweet_location', 'user_timezone'.

c**).dependent variables:** ”gender”

d**).ACCCURACY(n_estimators=300)=** ~78%

e).TUNING THE HYPERPARAMETER (n_estimators) and getting accuracy score with different no of estimators and storing accuracy in a dictionary.

**4. KNN CLASSIFIER:**

a).**REMOVING THE OUTLIERS:**First step was to remove the outlier from fav_number and tweet_count as much it is necessary.

b).changing the hyperparameters such as n_neighbors and metric to get better result in accuracy.

c).**ACCURACY (n_neighbors=4, metric=”Euclidean”)** =~73.2%.

d).generating accuracy with different no of neighbors and storing accuracy in dictionary.

**CONCLUSION:**

BEST ML CLASSIFIER WHICH SUITS THIS PARTICULAR DATASET IS **RANDOM FOREST** HAVING ACCURACY BETWEEN **78%-80%.**

***NOTE:** IN THIS PARTICULAR DATASET OUTLIERS ACTUALLY HELP TO IMPROVE THE ACCURACY OF THE MODEL .