

***Name***: *Chhaya Negi*

***E-mail ID:*** *chhayanegi1224@gmail.com*

***Subject:*** *Machine Learning Major Project*

***Classroom Name:*** *ML09B13*

**MACHINE LEARNING MAJOR PROJECT**

First the necessary python packages are imported. In this project packages used are numpy, pandas, matplotlib.pyplot. The data is read in csv format.

**Question 1**: What are the most common emotions/words used by Males and Females?

**Answer:** Most used word by males is ‘Like’ which was used 350 times and by females is ‘Like’ which was used (449) times.

Steps:

1. Remove all rows having ‘brand’ and ‘unknown’ word in gender columns.
2. Make a new data frame having only texts corresponds to males/females.
3. Make an array containing all the texts of respective genders.
4. Import all the required library eg: nltk.
5. Tokenization: Split the sentence after converting every character in lower case.
6. Stop words: Remove all the stop words.
7. Removing punctuation marks using regular expression.
8. Store every word in dictionary for getting frequencies of every words.
9. Then sort the words according to their frequencies in non-descending order.

**Question 2:** What are the counts of number of females in the dataset?

**Answer:** There are 6700 females in the given data set.

Steps:

1. Introduce a variable and assign it to the length of the gender column where the condition is equal to ‘female

Next is to clean the data. Data cleaning refers to preparing data for analysis by removing or modifying data that is incomplete, irrelevant, duplicated, or improperly formatted. To do so, make a new data frame and assign the given data frame to it. Now in the gender column we have unwanted values like ‘brand’ and ‘unknown’, so we should eliminate them. Next we summarize our data to get a better outlook.

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. We have demonstrated data visualization during feature selection and accuracy calculation.

The next step is Feature selection. Feature selection is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. For this we will be using heatmap and would be focusing on the following types of features: fav\_number, tweet\_count and gender.

Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data. So for this we have used label encoding. Label encoding is simply converting each value in a column to a number.

For the given dataset the independent variables were fav\_number, tweet\_count and the dependent variable was gender.

**Question 3:** Calculate and compare the accuracy using three different algorithms.

**Answer:** Different supervised approaches are used to assess the performance of the proposed features, including:

1. Logistic Regression.
2. Support Vector Machines.
3. K Nearest Neighbors.
4. Logistic Regression:

Steps:

1. Perform a simple feature selection. Assign ‘fav\_number’, ‘tweet\_count’ (independent variable) to X and ‘gender’ (dependent variable) to Y.
2. Import the necessary packages like train\_test\_split, metrics, LogistcRegression, StandardScaler, confusion\_matrix from sklearn.
3. Divide the training and testing set into 80% and 20%
4. Fit and transform the data, fit is used to calculate the mean and variance of each feature in the data. Transform is transforming all the features using the respective mean and variance.
5. Train the logistic regression model on the training data.
6. Print the coefficients and predict.
7. A heatmap is used to visualize the data.
8. K-Nearest Neighbors:

Steps:

1. Perform a simple feature selection. Assign ‘fav\_number’, ‘tweet\_count’ (independent variable) to X and ‘gender’ (dependent variable) to Y.
2. Import the necessary packages like KNeighbhorsClassifier, f1\_score, accuracy\_score, confusion\_matrix from sklearn.
3. Divide the training and testing set into 80% and 20%
4. Fit and transform the data, fit is used to calculate the mean and variance of each feature in the data. Transform is transforming all the features using the respective mean and variance.
5. Train the K-nn model on the training data.
6. Print the coefficients and predict.
7. A heatmap is used to visualize the data.
8. Support Vector Machine

Steps:

1. Perform a simple feature selection. Assign ‘fav\_number’, ‘tweet\_count’ (independent variable) to X and ‘gender’ (dependent variable) to Y.
2. Import the necessary packages like train\_test\_split, SVC, accuracy\_score, confusion\_matrix from sklearn.
3. Divide the training and testing set into 80% and 20%
4. Fit and transform the data, fit is used to calculate the mean and variance of each feature in the data. Transform is transforming all the features using the respective mean and variance.
5. Train the svm model on the training data.
6. Print the coefficients and predict.
7. A heatmap is used to visualize the data.

The accuracy of Logistic Regression was found to be: **51.69%**

The accuracy of K-Nearest Neighbor was found to be: **56.36%**

The accuracy of Support Vector Machine was found to be: **52.60%**

Therefore, the algorithm with the best accuracy was that of ***K-Nearest Neighbor***.

Ensemble Modelling is a process where multiple diverse models are created to predict an outcome, either by using many different modeling algorithms or using different training data sets.

Majority Voting:

1. The first step is to create multiple classification/regression models using some training dataset.
2. Each base model can be created using different splits of the same training dataset and same algorithm, or using the same dataset with different algorithms, or any other method.
3. Every model makes a prediction (votes) for each test instance and the final output prediction is the one that receives more than half of the votes.

Iterative Method:

1. First we train each of the three methods with the same data set.
2. Then compare the output (predicted value) for the given input.
3. That output will be considered which have majority.

**Summary**

In this project we worked on a data set containing twitter profile information. We predicted the gender. The data set contained 20,000 rows and 26 columns. We cleaned the data, performed EDA, visualized the data, performed feature selection and engineering. Besides this ensemble modelling was conducted. Based on the given data set we formed two questions and solved it. Next, we chose three classification algorithms that is logistic regression, k-nearest neighbor and support vector machine, the accuracy of these algorithms were then calculated and compared. On comparison, K-nearest algorithm was found out to be the most accurate.