Sheelkumar – 20BCP099

ML Project report

**Fake news Detection using Logistic regression**

**1.** **Introduction**

In an age of information overload and social media, the spread of fake news has become a significant problem. The aim of this project is to develop a fake news detection system using logistic regression classification. This system will help in identifying and flagging news articles that are likely to be fake or misleading, thereby promoting the dissemination of accurate information.

**2.** **Data Collection**

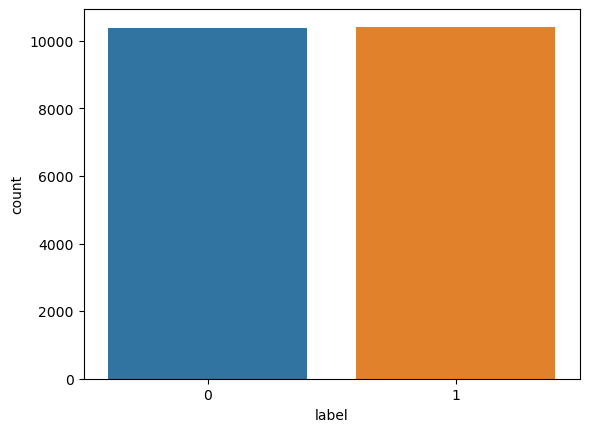
2.1 About Dataset

I collected a dataset of news articles from Kaggle. The dataset consists of both real and fake news which are labelled.

Data set have 20800 rows and 5 columns and following features:

* id: unique id for a news article
* title: the title of a news article
* author: author of the news article
* text: the text of the article; could be incomplete
* label: a label that marks whether the news article is real or fake

both real and fake news count are same



2.2 Data Preprocessing

The data preprocessing steps included:

Text cleaning: Removing punctuation, special characters, replacing null values with whitespaces.

Stop word removal: Eliminating common stop words that don't carry significant information. typically include common words like "the," "and," "is," "in," "to," "it," "a," "an," "of," etc.

Stemming: Reducing words to their base or root form. For example actor, actress, acting --> act

**3.** **Feature Extraction**

3.2 Combining features:

I have combined author name and news article title for simplicity and better results.

3.1 TF-IDF Vectorization

I have used TF-IDF (Term Frequency-Inverse Document Frequency) vectorization to convert the text data into numerical features. TF-IDF assigns a weight to each word in the document, which represents its importance in distinguishing between fake and real news articles.

3.2 Splitting data into test and train

For applying logistic regression we have split the data into test and trail where, test and train ratio is 0.2.

**4.** **Model Building**

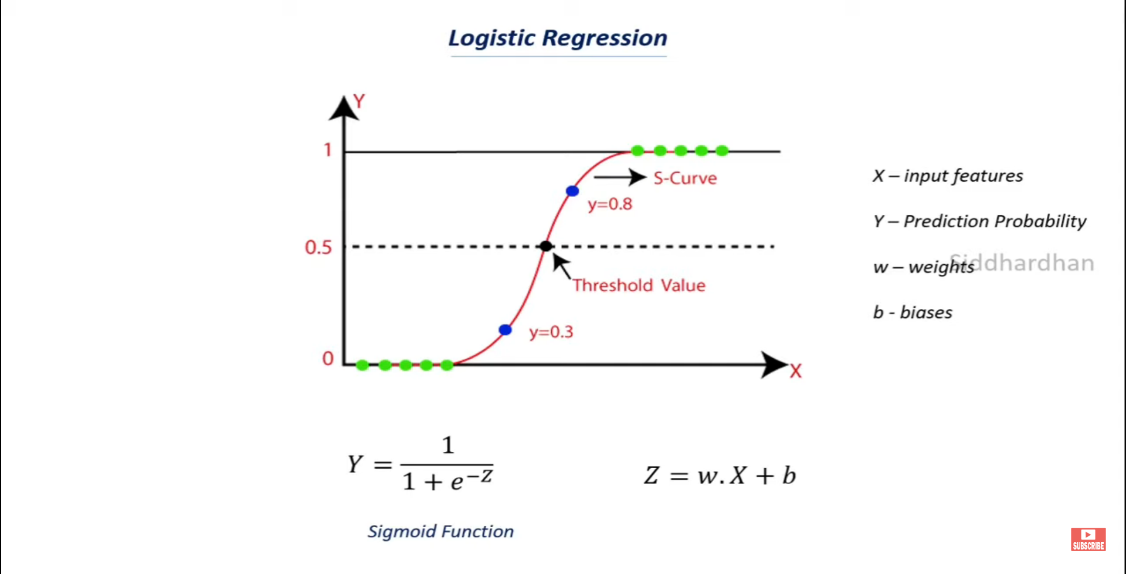
4.1 Logistic Regression

Logistic regression is a simple and effective classification algorithm for binary classification tasks. We trained a logistic regression model using the TF-IDF features to distinguish between real and fake news articles.

Logistic regression uses sigmoid function for classification.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

The curve from the logistic function indicates the likelihood of something



**5.** **Model Evaluation**

5.1 Performance Metrics

We used the following evaluation metrics to assess the model's performance:

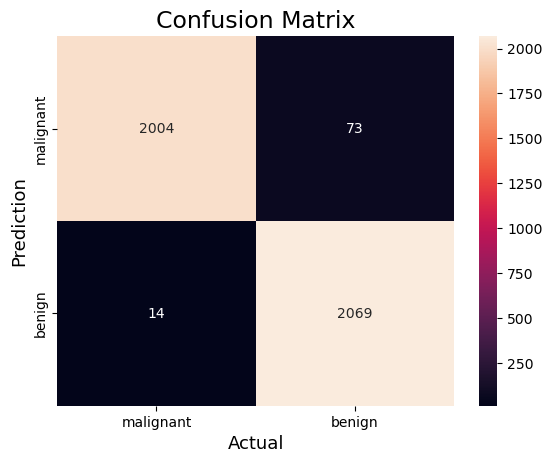
**Accuracy: 0.9790865384615385**

**Precision: 0.9659197012138189**

**Recall: 0.9932789246279404**

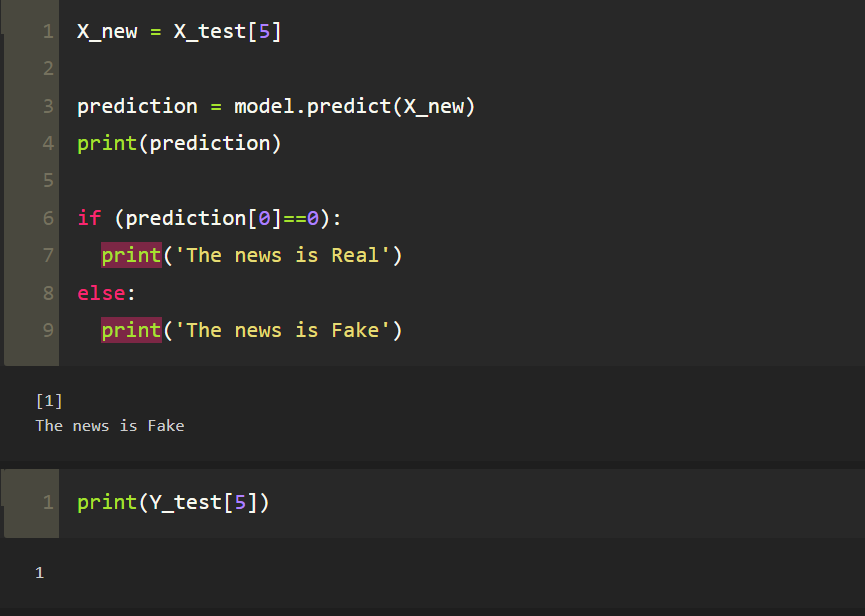
**F1-score: 0.9794082840236686**

**Confusion matrix**



5.2 Results

Sample prediction from test dataset:



**6.** **Conclusion**

In this project, we developed a fake news detection system using logistic regression classification. The system showed promising results in distinguishing between real and fake news articles with high accuracy. It is an essential step towards addressing the issue of fake news and promoting the dissemination of accurate information.