Machine Learning

Project 2

Submitted to: Dr. Saira Osama

Group Members

|  |  |
| --- | --- |
| Group Member ID | Group member Name |
| Muhammad Ali | F2020376004 |
| Danish Jameel | F2020376076 |
| Muhammad Faizan | F2020376009 |

# Problem formulation

The problem involves developing a fake news detection model with two output classes: "Real" and "Fake". The objective is to train the model using various algorithms, including SVM (Support Vector Machines), PCA (Principal Component Analysis), decision trees, k-means clustering, logistic regression, and linear regression.

Each news article in the dataset is labeled as either "Real" or "Fake". The model aims to learn patterns and features from the text content of the news articles to accurately classify them into their respective classes.

The SVM algorithm can be used to create a decision boundary that separates the real and fake news articles in a high-dimensional feature space. It maximizes the margin between the classes, making it effective for binary classification tasks.

PCA can be employed as a dimensionality reduction technique to transform the high-dimensional feature vectors of the news articles into a lower-dimensional space. This can help in reducing noise and improving the efficiency of subsequent algorithms.

Decision trees are a popular choice for classification tasks. They partition the feature space based on different features to create a tree-like structure. Each leaf node represents a class label, and the model can predict the class of a new news article by traversing the tree based on its features.

K-means clustering can be used to group similar news articles together based on their features. By assigning labels to the clusters, the model can classify new articles based on their proximity to existing clusters.

Logistic regression is a statistical model that can be used to estimate the probability of a news article belonging to a particular class. It uses a logistic function to model the relationship between the features and the class probabilities.

Linear regression can also be employed to predict the class probabilities of the news articles based on their features. It models the relationship between the features and the class labels using a linear equation.

By training the model using these algorithms, it aims to achieve accurate classification of news articles into "Real" or "Fake" classes. The performance of each algorithm can be evaluated using appropriate metrics such as accuracy, precision, recall, and F1 score, to assess their effectiveness in fake news detection.

# Dataset

[Dataset of Fake and Real News<-------](https://drive.google.com/file/d/1gw6Ej0XsaEZiM5kReRDT_e9ueHJq8JzL/view?usp=sharing) (The dataset was quite big so we uploaded on our drive)

The dataset consists of over 57,000 news articles primarily sourced from Pakistani media. These articles cover a diverse range of topics and domains, including politics, current affairs, economy, sports, entertainment, and more.

Since the dataset is predominantly composed of news articles from Pakistani media, it reflects the specific context, perspectives, and characteristics of the Pakistani media landscape. The articles are likely to cover local news events, regional politics, societal issues, and cultural developments relevant to Pakistan.

The dataset may encompass articles from various Pakistani news outlets, including newspapers, online news platforms, and television channels. It is expected to cover a broad timeframe, potentially spanning several years, capturing news articles published over an extended period.

The articles within the dataset are represented as text documents, typically containing the headline, the main body of the article, and potentially additional metadata such as publication date, author, or source.

To facilitate analysis and modeling, the dataset is likely to be preprocessed. Text preprocessing techniques such as tokenization, lowercasing, removal of stop words, and stemming/lemmatization may have been applied to standardize the text data and remove noise or irrelevant information.

Given the size and diversity of the dataset, it provides a substantial corpus for training and evaluating machine learning models for fake news detection specifically within the context of Pakistani media.

# Results

Results for all algorithms with their train and test out of 1/100% are down below:

## 1. Train and Test for Logistic Regression Model:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Train Set** | Accuracy | F1 score | precision | recall | ROC AUC |
| Fold 1 | 0.898 | 0.9 | 0.89 | 0.9 | 0.963 |
| Fold 2 | 0.898 | 0.9 | 0.9 | 0.9 | 0.962 |
| Fold3 | 0.897 | 0.89 | 0.9 | 0.9 | 0.962 |
| Average | 0.8977 | 0.897 | 0.897 | 0.9 | 0.962 |
| **Test Set** | Accuracy | F1 score | precision | recall | ROC AUC |
| Fold 1 | 0.872 | 0.87 | 0.87 | 0.87 | 0.946 |
| Fold 2 | 0.869 | 0.87 | 0.87 | 0.87 | 0.947 |
| Fold 3 | 0.877 | 0.88 | 0.88 | 0.88 | 0.949 |
| Average | 0.873 | 0.873 | 0.873 | 0.873 | 0.947 |

## 2. Train and Test for Linear Regression Model:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Train Set** | Accuracy | F1 score | precision | recall | mse |
| Fold 1 | 0.91 | 0.91 | 0.91 | 0.91 | 0.09 |
| Fold 2 | 0.909 | 0.91 | 0.91 | 0.91 | 0.091 |
| Fold3 | 0.91 | 0.91 | 0.91 | 0.91 | 0.09 |
| Average | 0.91 | 0.91 | 0.91 | 0.91 | 0.09 |
| **Test Set** | Accuracy | F1 score | precision | recall | mse |
| Fold 1 | 0.86 | 0.86 | 0.86 | 0.86 | 0.139 |
| Fold 2 | 0.859 | 0.86 | 0.86 | 0.86 | 0.141 |
| Fold 3 | 0.86 | 0.86 | 0.86 | 0.86 | 0.139 |
| Average | 0.86 | 0.86 | 0.86 | 0.86 | 0.14 |

## 3. Train and Test for PCA Model:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Train Set** | Accuracy | F1 score | precision | recall | ROC AUC |
| Fold 1 | 0.877 | 0.88 | 0.88 | 0.88 | 0.877 |
| Fold 2 | 0.876 | 0.88 | 0.88 | 0.88 | 0.8 |
| Fold3 | 0.873 | 0.87 | 0.87 | 0.87 | 0.87 |
| Average | 0.875 | 0.88 | 0.88 | 0.88 | 0.849 |
| **Test Set** | Accuracy | F1 score | precision | recall | ROC AUC |
| Fold 1 | 0.863 | 0.86 | 0.86 | 0.86 | 0.86 |
| Fold 2 | 0.86 | 0.86 | 0.86 | 0.86 | 0.86 |
| Fold 3 | 0.864 | 0.86 | 0.86 | 0.86 | 0.86 |
| Average | 0.842 | 0.86 | 0.86 | 0.86 | 0.86 |

## 4. Train and Test for K-Means Model

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Train Set** | Accuracy | F1 score | precision | recall | ROC AUC |
| Fold 1 | 0.47 | 0.44 | 0.45 | 0.47 | 0.46 |
| Fold 2 | 0.528 | 0.5 | 0.55 | 0.53 | 0.54 |
| Fold3 | 0.526 | 0.5 | 0.55 | 0.53 | 0.54 |
| Average | 0.508 | 0.48 | 0.52 | 0.51 | 0.51 |
| **Test Set** | Accuracy | F1 score | precision | recall | ROC AUC |
| Fold 1 | 0.48 | 0.45 | 0.46 | 0.48 | 0.47 |
| Fold 2 | 0.53 | 0.5 | 0.55 | 0.53 | 0.54 |
| Fold 3 | 0.53 | 0.51 | 0.55 | 0.53 | 0.54 |
| Average | 0.51 | 0.49 | 0.52 | 0.52 | 0.51 |

## 5. Train and Test for Decision Tree Model:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Train Set** | Accuracy | F1 score | precision | recall | ROC AUC |
| Fold 1 | 0.996 | 1 | 1 | 1 | 0.99 |
| Fold 2 | 0.996 | 1 | 1 | 1 | 0.99 |
| Fold3 | 0.996 | 1 | 1 | 1 | 0.99 |
| Average | 0.996 | 1 | 1 | 1 | 0.99 |
| **Test Set** | Accuracy | F1 score | precision | recall | ROC AUC |
| Fold 1 | 0.827 | 0.83 | 0.83 | 0.83 | 0.83 |
| Fold 2 | 0.83 | 0.83 | 0.83 | 0.83 | 0.83 |
| Fold 3 | 0.827 | 0.83 | 0.83 | 0.83 | 0.83 |
| Average | 0.828 | 0.83 | 0.83 | 0.83 | 0.83 |

## 6. Train and Test for SVM Model:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Train Set** | Accuracy | F1 score | precision | recall | ROC AUC |
| Fold 1 | 0.986 | 0.99 | 0.99 | 0.99 | 0.98 |
| Fold 2 | 0.988 | 0.99 | 0.99 | 0.99 | 0.98 |
| Fold3 | 0.986 | 0.99 | 0.99 | 0.99 | 0.98 |
| Average | 0.987 | 0.99 | 0.99 | 0.99 | 0.98 |
| **Test Set** | Accuracy | F1 score | precision | recall | ROC AUC |
| Fold 1 | 0.859 | 0.86 | 0.86 | 0.86 | 0.85 |
| Fold 2 | 0.865 | 0.87 | 0.87 | 0.86 | 0.84 |
| Fold 3 | 0.863 | 0.86 | 0.86 | 0.86 | 0.83 |
| Average | 0.862 | 0.86 | 0.86 | 0.86 | 0.84 |