Public Policy & Opinion Cell
Indian Institute of Technology, Kanpur

Summer Project'24

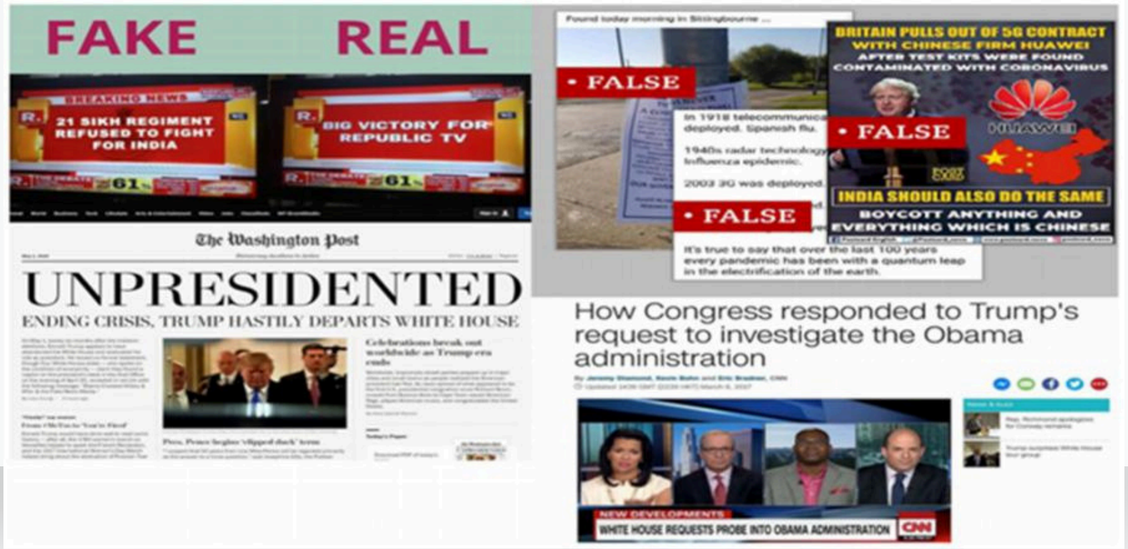
FACT FILTER

End-Term Evaluation

CONTENTS

INTRODUCTION

- With the proliferation of misinformation online, the ability to automatically identify fake news is crucial for maintaining the integrity of information dissemination.
- The principal aim of this project was to develop a fact filtering machine learning model that can reliably differentiate between authentic and fraudulent news stories.
- The model will be trained on datasets of both real and fake news articles, utilizing both machine learning and natural language processing (NLP) techniques for feature extraction and classification of news articles as fake or real.



Public Policy
& Opinion Cell
Indian Institute of Technology, Kanpur

PROJECT TIMELINE

Public Policy & Opinion Cell
Indian Institute of Technology, Kanpur

WEEK 1

Introduction to Machine learning and Python Libraries

WEEK 2

Data collection and Preprocessing

WEEK 3

Feature Extraction

WEEK 4

Machine learning models and evaluation metrics

WEEK 5

Final Model

WEEK-WISE ACCOMPLISHMENT

Public Policy & Opinion Cell
Indian Institute of Technology, Kanpur

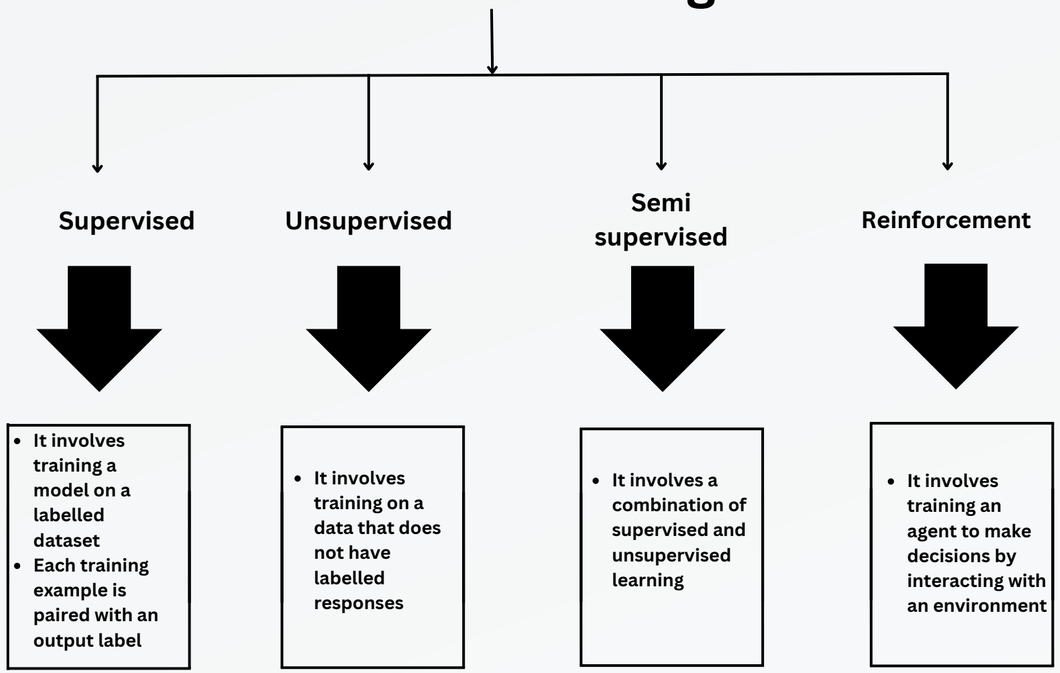
WEEK-1

Public Policy & Opinion Cell
Indian Institute of Technology, Kanpur

Introduction to Machine Learning and Python Libraries

ML is basically a subset of AI that focuses on developing algorithms and statistical models by analyzing data and identifying patterns

Machine Learning



Used MACHINE Learning Techniques

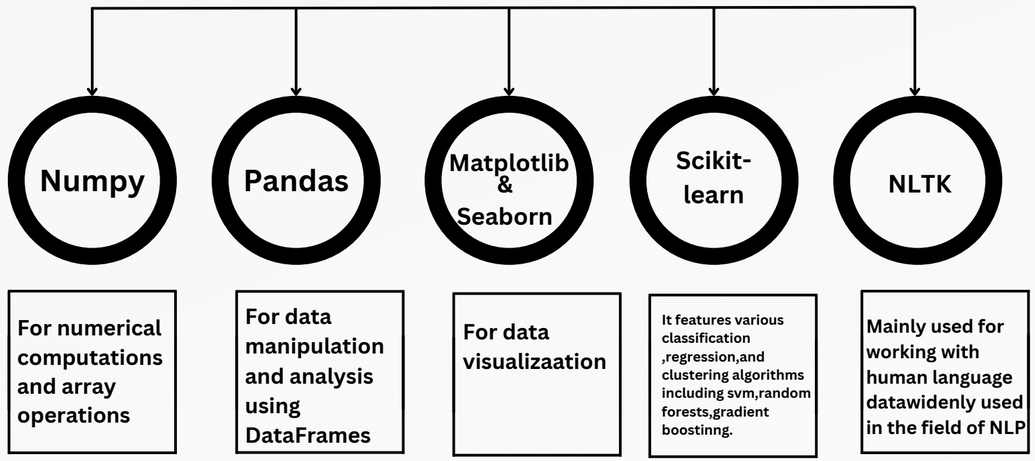
Logistic Regression
Use for binary classification problems
It predicts the probability of a binary outcome

Decision Trees
A decision tree in machine learning splits data into branches based on feature values to predict outcomes and make decisions.
Used for both classification and regression tasks

SVMs
Support Vector Machines (SVMs) are supervised learning models that classify data by finding the optimal hyperplane that separates classes in feature space.

Random forest
Random forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and control overfitting.

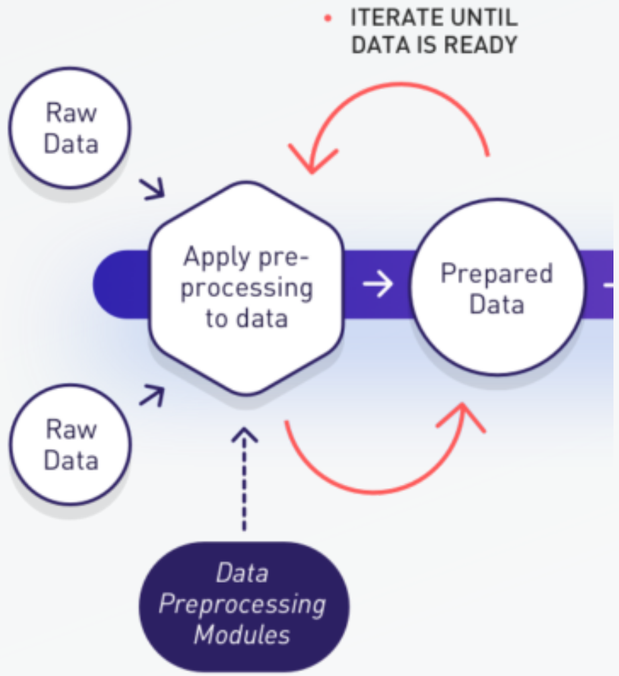
Python libraries



WEEK-2

Data collection

Collecting data for training the ML model is the basic step in the machine learning pipeline. The predictions made by ML systems can only be as good as the data on which they have been trained.



Data Preprocessing

Real-world raw data and images are often incomplete, inconsistent and lacking in certain behaviors or trends. They are also likely to contain many errors. So, once collected, they are pre-processed into a format the machine learning algorithm can use for the model.

Preprocessing techniques

Data Cleaning	Data Imputations	Oversampling	Data Integration	Data Normalization
These techniques, manual and automated, remove data incorrectly added or classified.	It includes imputing missing values with standard deviation, mean, median and k nearest neighbors (k-NN) of the data in the given field.	It increases the number of minority class samples in an imbalanced dataset to improve machine learning model performance.	Combining multiple datasets to get a large corpus can overcome incompleteness in a single dataset.	Normalization reduces the size by reducing the order and magnitude of data.

WEEK-3

Public Policy & Opinion Cell
Indian Institute of Technology, Kanpur

Feature Extraction

process of transforming raw data into numerical features that can be used by machine learning algorithms.

One Hot Encoding

Converts categorical data (words) into a binary format that machine learning algorithms can process.

Word Embedding

Represents words in continuous vector space where semantically similar words are closer together.

WordtoVec

It is a popular technique in NLP that converts words into numerical vectors, capturing semantic meanings

Regex

- Extracts relevant text patterns, like dates, emails, and keywords, from raw data.
- Removes unwanted text, like special characters or noise, to prepare data for analysis.
- Converts unstructured text into structured data by capturing specific information, enhancing data processing and analysis.

Bag of words

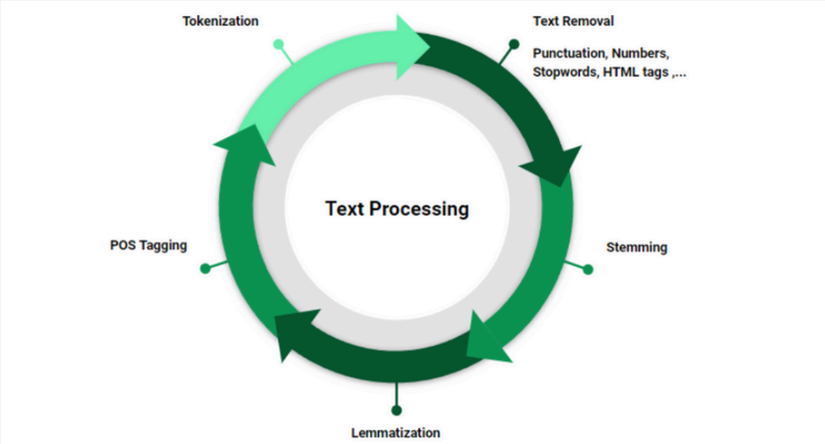
Represents text data by the frequency (or presence) of words, ignoring grammar and word order.

TF-IDF

Enhances the Bag of Words model by reducing the weight of commonly occurring words and emphasizing rarer but potentially more informative words.

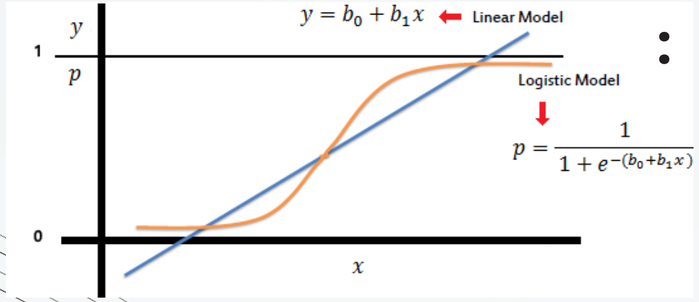
WEEK-4

Public Policy & Opinion Cell
Indian Institute of Technology, Kanpur

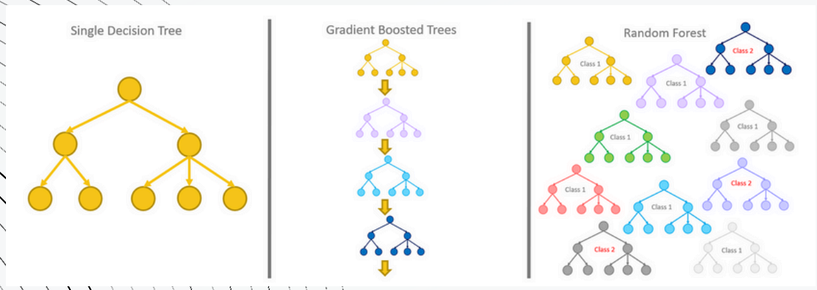


Text cleaning Techniques

Suitable models that can be used for our project

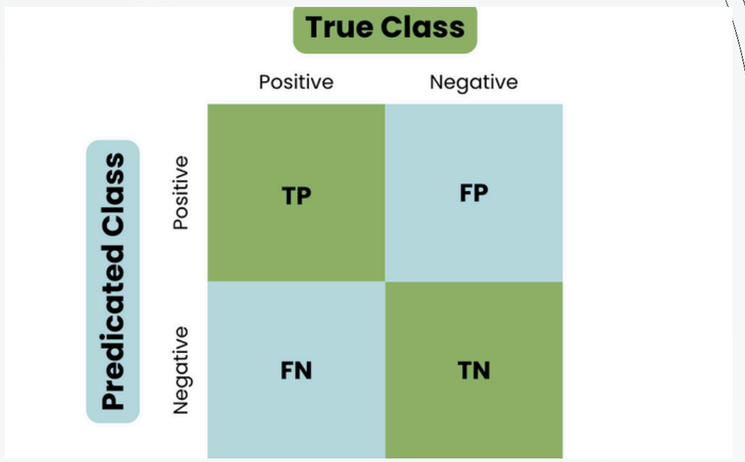


Logistic Regression

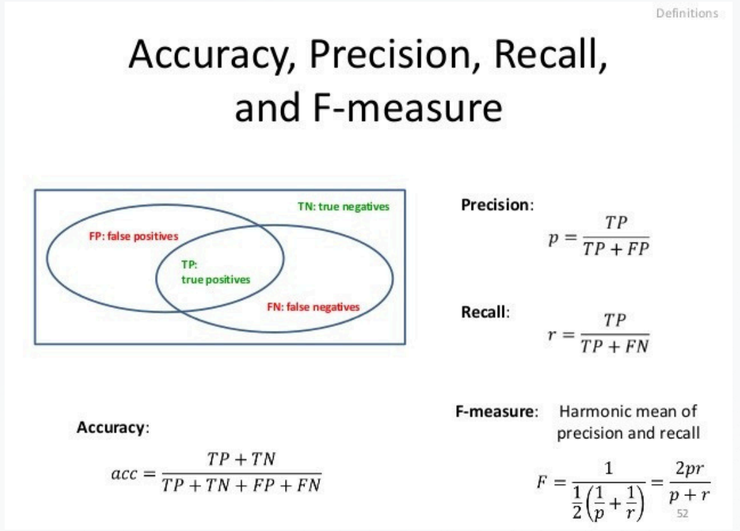


Decision Tree
Gradient Boosting
Random Forest Classifier
and their main similarities differences

Evaluation metrics



Confusion Matrix



Evaluation metric formulae

WEEK-5

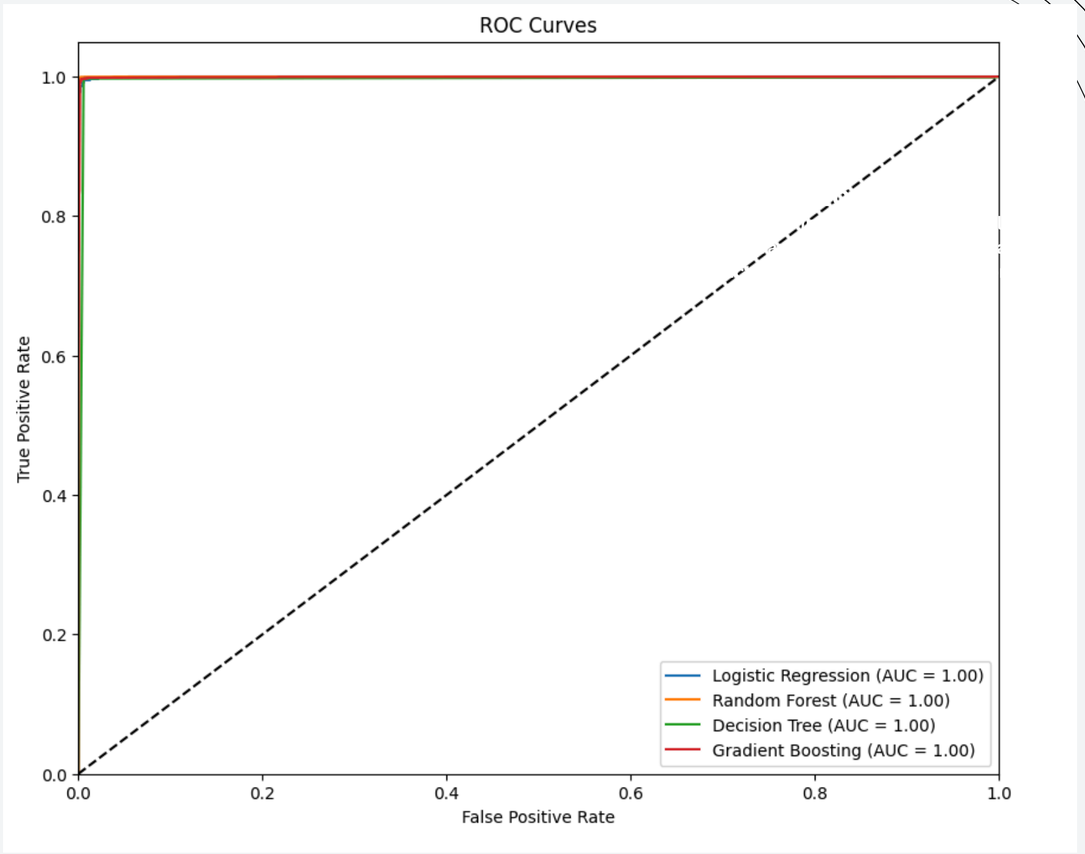
Public Policy & Opinion Cell
Indian Institute of Technology, Kanpur

Model Evaluation and Comparison:

Model evaluation assesses a model's performance using metrics like accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix to ensure it generalizes well to new data.

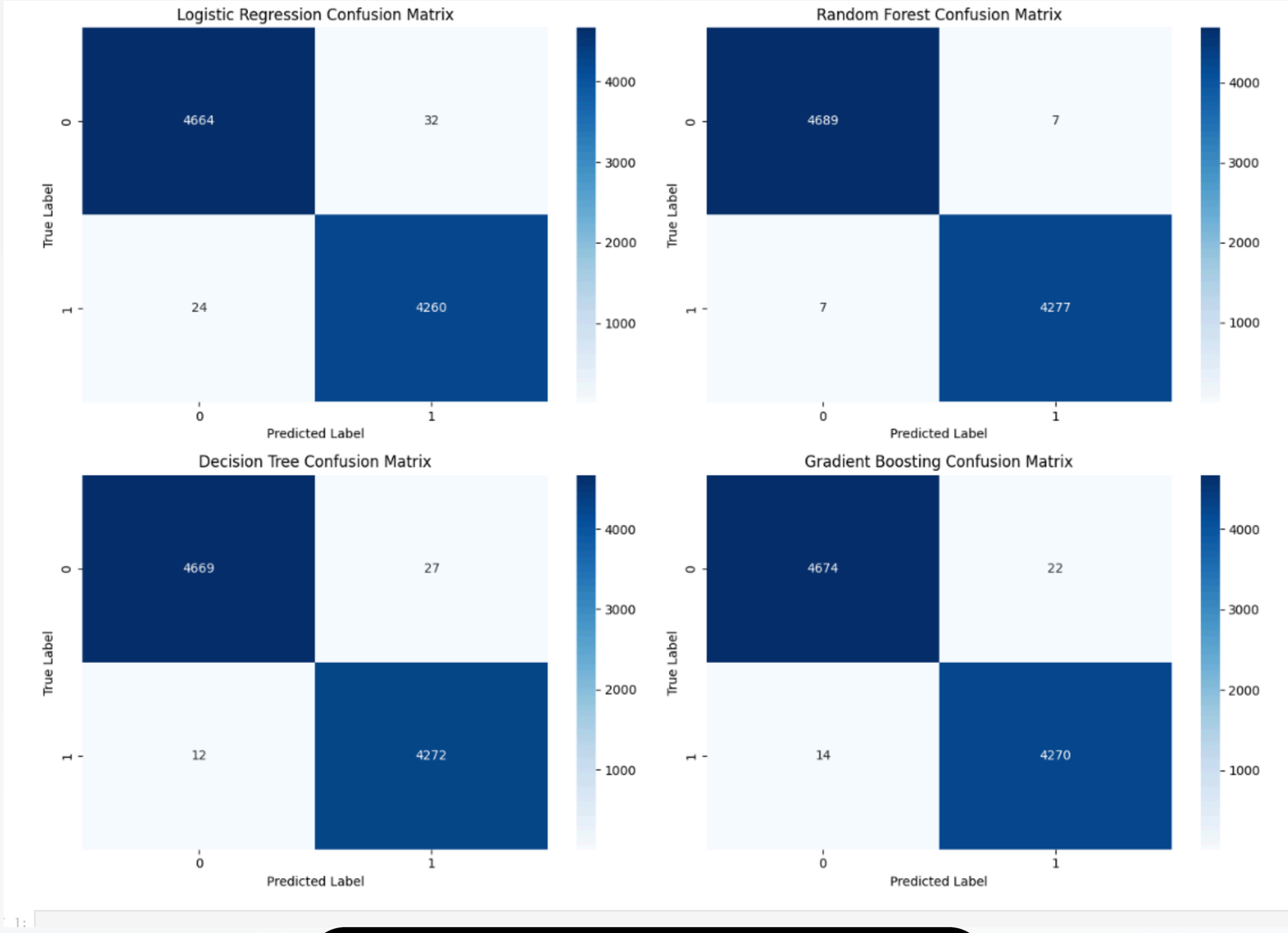
Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.9938	0.9925	0.9944	0.9935	0.9998
Random Forest	0.9984	0.9984	0.9984	0.9984	0.9999
Decision Tree	0.9957	0.9937	0.9972	0.9955	0.9952
Gradient Boosting	0.996	0.9949	0.9967	0.9958	1.0

Model comparison involves evaluating multiple models on the same dataset to determine the best-performing one, using key metrics to guide the selection process.



This ROC curve demonstrates that all four models (Logistic Regression, Random Forest, Decision Tree, and Gradient Boosting) are performing perfectly on the given dataset, with an AUC of 1.00, indicating that they all have perfect classification abilities on this specific dataset. This could be a sign of overfitting, where the models perform exceptionally well on the training/testing data but may not generalize to new, unseen data.

WEEK-5



Each models Confusion Matrix

Public Policy & Opinion Cell
Indian Institute of Technology, Kanpur

THANK YOU!