

# Machine Learning Course - Projecte.

## Part 1: Supervised Learning - Classifying Text or Images

### Option A: Classify News Headlines into Categories

#### Introduction

Explain that the goal is to classify news headlines into categories (World, Sports, Business, Science/Technology) using supervised learning techniques.

Dataset: AG News Dataset from Hugging Face (wangrongsheng/ag\_news).

Multiple machine learning models were trained and evaluated, including Logistic Regression, Decision Tree, Gradient Boosting, and K-Nearest Neighbors.

#### Exploratory Data Analysis

The dataset contains 120,000 samples, split into 90,000 training samples and 30,000 testing samples. The dataset is well-balanced across the four categories, with approximately 30,000 samples per category as visualized in the bar chart.

A balanced dataset helps avoid bias towards any particular class during model training.

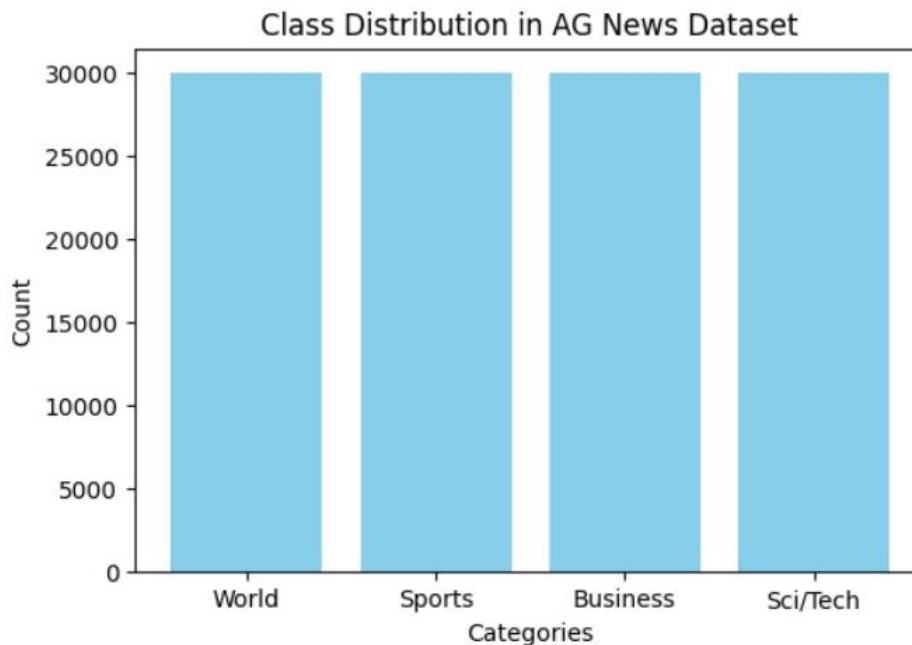


Figure 1 - visualized using a bar chart to ensure data balance

The resulting chart confirmed a nearly equal distribution of samples among all four categories, indicating a balanced dataset.

#### Preprocessing the Text Data

The news headlines were preprocessed using the TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer.

This approach converts textual data into numerical features, which are required for machine learning models.

The TF-IDF vectorizer was configured to remove English stopwords and to limit the feature set to 5000 most important words.

So the dataset was split into training (80%) and testing (20%) subsets. And a stratified split was used to maintain the class distribution in both sets.

## Model Training

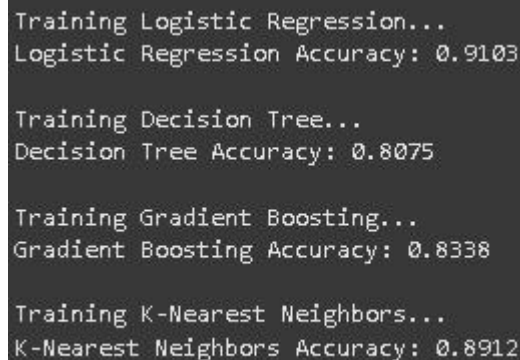
**Logistic Regression:** Suitable for multiclass classification problems and performs well with high-dimensional data.

**Decision Tree:** Provides interpretability and handles non-linear relationships.

**Gradient Boosting:** An ensemble method that improves accuracy by combining weak learners.

**K-Nearest Neighbors (KNN):** Classifies data points based on the closest labeled examples.

Each model was trained using the training data and evaluated on the test set. Accuracy scores were calculated for each model to compare their performance.



```
Training Logistic Regression...
Logistic Regression Accuracy: 0.9103

Training Decision Tree...
Decision Tree Accuracy: 0.8075

Training Gradient Boosting...
Gradient Boosting Accuracy: 0.8338

Training K-Nearest Neighbors...
K-Nearest Neighbors Accuracy: 0.8912
```

Figure 2 - Accuracy scores of four different ML models

- **Logistic Regression:** Achieved the highest accuracy of **0.91**.
- **Decision Tree:** Performed moderately with **0.80** accuracy.
- **Gradient Boosting:** Achieved **0.83** accuracy, showing stability but not outperforming Logistic Regression.
- **K-Nearest Neighbors:** Competitive accuracy of **0.89**, suggesting that it is a strong model for this dataset.

Overall comparison:

**Best Model:** Logistic Regression with 91.03% accuracy.

**Runner-Up:** K-Nearest Neighbors with 89.12% accuracy, showing strong performance but potentially higher computational costs.

**Moderate Performance:** Gradient Boosting at 83.38%, with potential for improvement through hyperparameter tuning.

**Lowest Accuracy:** Decision Tree with 80.75%, possibly due to overfitting or difficulty in handling high-dimensional sparse data.

## Model Evaluation Using Precision, Recall, and F1-Score

In this we evaluated the performance of four machine learning models: **Logistic Regression, Decision Tree, Gradient Boosting, and K-Nearest Neighbors**. The evaluation focused on classification metrics including:

**Precision:** Measures how many of the predicted positive instances are actually positive.

**Recall:** Measures how many of the actual positive instances were correctly predicted.

**F1-Score:** The harmonic mean of precision and recall, providing a balance between them.

**Support:** The number of actual occurrences of each class in the test dataset.

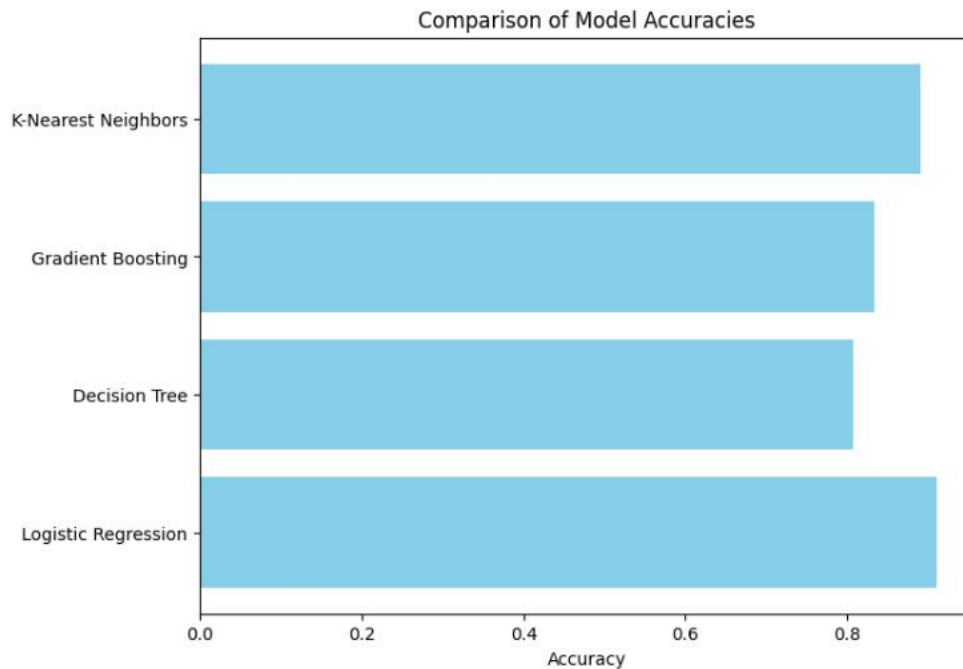


Figure 3 - Model Accuracy Comparison Chart

The comparison chart illustrates the accuracy of four machine learning models—**Logistic Regression**, **Decision Tree**, **Gradient Boosting**, and **K-Nearest Neighbors (KNN)**—in classifying news headlines. **Logistic Regression** achieved the highest accuracy 91%, demonstrating strong performance with high-dimensional text data. **K-Nearest Neighbors** closely followed 89%, showing competitive results but with higher computational costs. **Gradient Boosting** achieved moderate accuracy 83%, suggesting potential for improvement through hyperparameter tuning. The **Decision Tree** model had the lowest accuracy 81%, likely due to overfitting or challenges with sparse TF-IDF features. Overall, **Logistic Regression** is recommended for its balance of accuracy, efficiency, and interpretability.