Hyperparameter Tuning and Model Evaluation Report

1. Introduction

This report outlines the hyperparameter tuning process and evaluation metrics for different machine learning models applied to misinformation detection. The models used include K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), K-Means Clustering, Neural Networks (MLP), and Gradient Boosting Classifier.

2. Data Preprocessing

- The dataset was split into training (80%), validation (10%), and testing (10%) using stratified sampling to ensure balanced class distribution.
- Text preprocessing steps included:
  - Converting text to lowercase
  - Removing URLs
  - Converting emojis to text representation
  - Removing extra spaces
- TF-IDF vectorization was applied with a maximum feature size of 5000.

3. Hyperparameter Tuning

3.1 K-Nearest Neighbors (KNN)

- **Hyperparameters tuned:**
  - Number of neighbors: **5**
  - Distance metric: **Euclidean**
- **Final Model Performance:**
  - Accuracy: **92%**
  - F1-score: **0.92**

3.2 Logistic Regression

- **Hyperparameters tuned:**
    - Maximum iterations: **500**
    - Solver: **lbfgs**
- **Final Model Performance:**
    - Accuracy: **93%**
    - F1-score: **0.93**

3.3 Support Vector Machine (SVM)

- **Hyperparameters tuned:**
    - Kernel: **Linear**
    - Regularization parameter (C): **1.0**
- **Final Model Performance:**
    - Accuracy: **95%**
    - F1-score: **0.95**

3.4 K-Means Clustering

- **Hyperparameters tuned:**
    - Number of clusters: **2**
    - Random state: **42**
- **Final Model Performance:**
    - Accuracy: **37%**
    - F1-score: **0.29**
- K-Means clustering did not perform well due to the unsupervised nature of the model and lack of proper class separation in feature space.

3.5 Neural Network (MLPClassifier)

- **Hyperparameters tuned:**
    - Hidden layer sizes: **(100,)**
    - Activation function: **ReLU**
    - Solver: **Adam**
    - Maximum iterations: **300**
- **Final Model Performance:**
    - Accuracy: **94%**

    - F1-score: **0.94**

3.6 Gradient Boosting Classifier

- **Hyperparameters tuned:**
    - Number of estimators: **100**
    - Learning rate: **0.1**
    - Max depth: **3**
- **Final Model Performance:**
    - Accuracy: **88%**
    - F1-score: **0.88**

4. Conclusion

- The **SVM model** performed the best with **95% accuracy** and **0.95 F1-score**, making it the most suitable model for misinformation classification.
- **Neural Networks (MLP) and Logistic Regression** also showed high performance with **94% and 93% accuracy, respectively**.
- **K-Means clustering was the worst-performing model** since it is an unsupervised technique that struggled to distinguish between fake and real labels.
- **Gradient Boosting performed reasonably well but was outperformed by SVM and MLP.**

5. Final Model Selection

The **Support Vector Machine (SVM)** is chosen as the final model for misinformation detection due to its high accuracy and reliability in handling text-based classification tasks.