**Project Report**

**Models and Results**

Several machine learning models were implemented to analyze and classify the dataset. These models were evaluated based on their predictive accuracy. Below are the results:

* **Support Vector Machine (SVM):** 92.46%
* **Decision Tree:** 93.41%
* **Random Forest:** 93.6%
* **XGBoost:** 93.28%

**Key Observations:**

1. The **Random Forest model** achieved the highest accuracy at **93.6%**, demonstrating its ability to handle complex relationships effectively.
2. **Decision Tree** and **XGBoost** also performed well, suggesting that tree-based models are well-suited for this dataset.
3. **SVM** showed slightly lower accuracy, indicating that it may not fully capture some nonlinear patterns present in the data.

**Attributes Used**

The dataset contains the following features (attributes) used for classification:

1. **IsCaps** - A numerical attribute representing the percentage of uppercase letters in the text.
2. **Line Spacing** - Measures spacing patterns between lines in the text, also numerical.
3. **LessThan5000** - A **binary attribute (0 or 1)** indicating whether the length of the text is less than 5000 characters.

**Data Visualization**

Visual analysis was conducted to understand the relationship between each attribute and the label (**REAL** vs **FAKE**).

**1. IsCaps vs Label**

* A **violin plot** was used to visualize the distribution of uppercase character percentages across labels.
* The plot highlights variations, with **FAKE news** tending to have higher uppercase percentages compared to **REAL news**.

**A graph showing a comparison of a comparison of a product

Description automatically generated with medium confidence**

**2. Line Spacing vs Label**

* Another **violin plot** illustrates the spacing distribution for both labels.
* **REAL news** appears to have more varied spacing patterns compared to **FAKE news**.

**A graph of a plot of spacing

Description automatically generated**

**3. LessThan5000 vs Label**

* A **bar plot** displays the distribution of binary values (0 or 1) for text length.
* Texts shorter than 5000 characters are more likely associated with **REAL news** based on the results.

A graph with colorful squares

Description automatically generated with medium confidence

**Conclusion**

The project successfully tested multiple machine learning models, with **Random Forest** delivering the best performance. However, it was observed that the **LessThan5000** attribute, despite its high predictive power, might lead to **overfitting** due to its binary nature and strong separation between labels.

To address this concern, we incorporated the other two attributes—**IsCaps** and **Line Spacing**—in all models. These features provided additional information and helped improve the models' robustness without relying solely on a single attribute.

The results demonstrated that combining multiple features allowed the models to generalize better, reducing the risk of overfitting while maintaining high accuracy. Visualizations further supported the relationships between attributes and labels, confirming their relevance for classification.