1. What is the main goal of this project?

The goal is to automatically classify BBC news articles into predefined categories like business, politics, sport, entertainment, and tech using machine learning. This helps improve content organization, user experience, and personalized recommendations.

2. Why is article classification important for BBC News?

With a growing volume of digital content, it becomes challenging to manually sort articles. Classification helps users find relevant news faster, enables better search optimization, and makes content delivery more efficient.

3. What kind of data was used for this analysis?

We used a dataset containing BBC news articles. Each article had a unique ID, the full text, a target category, the character length, a cleaned version of the text, and a numerical label for machine learning models.

4. How were the articles prepared before feeding into the models?

The articles underwent preprocessing, which included removing stop words, punctuation, and applying stemming. This made the text cleaner and more standardized for model training.

5. What models were used and which one performed the best?

We tested Logistic Regression, Naive Bayes, Random Forest, and XGBoost. Random Forest performed the best, with the highest classification accuracy and consistency across all categories.

6. Why did Naive Bayes perform slightly worse than other models?

Naive Bayes assumes that all features (words) are independent, which often isn't true in natural language. It’s fast and effective for simple tasks but struggles when text contains complex language patterns or subtle context differences.

7. How does the imbalance in category frequency affect the results?

Some categories, like Sport and Business, had more articles than Tech. This imbalance can bias models to favor the dominant categories. We noted this challenge and recommended future strategies like oversampling or synthetic data generation.

8. Could the same article belong to more than one category?

In the current setup, each article belongs to only one category. However, in real life, some articles might overlap categories. Multi-label classification can be explored in future work.

9. How can this system be used in real-time news classification?

Once trained, the best model (Random Forest) can be deployed via an API. As new articles are published, they can be automatically processed and categorized in real-time, improving content delivery speed and personalization.

10. What ethical concerns did you consider?

We aimed to ensure fairness and transparency. This included addressing data imbalance to avoid biased predictions and protecting sensitive information in the dataset. Transparency in how the model makes decisions is also essential for public trust.