**Documentation**:

**Methodology:**

1. **Data Collection and Labeling**
   * Four datasets are used: politifact\_real, politifact\_fake, gossipcop\_real, and gossipcop\_fake. These datasets contain articles that are labeled either as real (1) or fake (0).
   * The articles from politifact\_real and gossipcop\_real are labeled as real (1), while the articles from politifact\_fake and gossipcop\_fake are labeled as fake (0).
2. **Data Preprocessing**
   * The datasets are merged into a single DataFrame (df) for further processing.
   * Rows with missing titles are removed.
   * The titles of the articles undergo basic preprocessing, which involves:
     + Removing HTML tags from the text.
     + Eliminating non-alphabetic characters.
     + Converting all text to lowercase to standardize the data.
3. **Feature Engineering**
   * **Named Entity Recognition (NER):** The spaCy library is used to perform NER on the article titles. Three entity types are extracted: Organizations (ORG), Geopolitical Entities (GPE), and Persons (PERSON).
   * **Entity Counts:** The count of each entity type (ORG, GPE, PERSON) is computed and stored for each title.
   * **Additional Features:**
     + The length of the article titles (in terms of word count) is computed and stored as article\_length.
     + The sentiment of each title is analyzed using TextBlob, which generates a sentiment score representing the polarity of the text (positive or negative).
4. **Data Transformation**
   * The entity counts are split into separate columns for each entity type (ORG, GPE, PERSON) to make them ready for modeling.
   * The features for the model consist of:
     + The counts of each entity type (ORG, GPE, PERSON).
     + article\_length, representing the number of words in the title.
     + sentiment, representing the sentiment score of the title.
   * The target variable (y) is the label of the article (1 for real, 0 for fake).
5. **Model Training and Evaluation**
   * The data is divided into training and testing sets using an 80-20 split.
   * A Random Forest Classifier (RandomForestClassifier) is used to train the model. The model is configured with 100 estimators and a maximum depth of 10.
   * After training, the model's performance is evaluated by calculating the accuracy and generating a classification report that includes precision, recall, and F1-score.
6. **Visualization**
   * A **bar chart** is plotted to visualize the frequencies of each entity type (ORG, GPE, PERSON) across the dataset.
   * A **scatter plot** is used to show the relationship between the sentiment of the titles and their length, with different colors representing real and fake articles.
   * A **heatmap** is created to visualize the correlations between the features and the target label (real or fake), highlighting how different features are related to article veracity.

This methodology combines data preprocessing, feature engineering, model training, evaluation, and visualization to build a model that can classify articles as real or fake based on their titles.

**Details about the predictive modelling and performance metrics:-**

**1. Data Splitting**

* The dataset is split into training and testing subsets using train\_test\_split from Scikit-learn. The split is performed with 80% of the data used for training the model and 20% reserved for testing. This ensures that the model is trained on a majority of the data while retaining a separate portion for unbiased evaluation.
* **Training Set:** 80% of the data is used to train the model.
* **Test Set:** 20% of the data is used to evaluate the model's performance.

**2. Model Selection**

* **Random Forest Classifier:** A Random Forest model is chosen due to its effectiveness in handling complex relationships, its ability to perform well on both classification and regression tasks, and its resilience to overfitting. The model consists of multiple decision trees, and each tree makes its own predictions based on random subsets of the data. The final prediction is made by aggregating the results of all individual trees, typically through majority voting for classification tasks.
* **Hyperparameters:**
  + n\_estimators=100: This specifies the number of decision trees in the forest. A higher number of trees can improve performance but at the cost of increased computational resources.
  + max\_depth=10: This limits the depth of the decision trees to prevent overfitting. Deeper trees might capture more noise, so limiting the depth helps the model generalize better.

**3. Model Training**

* The Random Forest Classifier is trained using the training data (X\_train, y\_train). During training, the model learns the relationships between the features (such as entity counts, article length, and sentiment) and the target label (real or fake).

**4. Model Evaluation**

* After training, the model is used to predict the labels of the test data (X\_test). The predicted labels are compared to the actual labels (y\_test). The following performance metrics are computed to evaluate the model's performance.

**5. Performance Metrics**

The model's performance is assessed using several metrics:

* **Accuracy:** Accuracy measures the proportion of correct predictions (both real and fake) out of all predictions. It is calculated by dividing the number of correct predictions by the total number of predictions. While accuracy is a useful metric, it may not be ideal when the classes are imbalanced (i.e., one class occurs much more frequently than the other).
* **Classification Report:** The classification report provides more detailed performance metrics, including:
  + **Precision:** Precision is the proportion of true positive predictions out of all positive predictions (i.e., how many of the predicted real articles are actually real). It is calculated as the ratio of true positives to the sum of true positives and false positives.
  + **Recall (Sensitivity):** Recall is the proportion of true positive predictions out of all actual positive instances (i.e., how many of the real articles were correctly identified as real). It is calculated as the ratio of true positives to the sum of true positives and false negatives.
  + **F1-Score:** The F1-Score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is particularly useful when there is an imbalance between the classes, as it accounts for both false positives and false negatives.
  + **Support:** Support refers to the number of actual occurrences of each class in the dataset.
* **Confusion Matrix:** Although not explicitly generated in the code, a confusion matrix would typically be used to show the counts of true positives, true negatives, false positives, and false negatives. This matrix provides a breakdown of the model's predictions and helps identify where errors occur.

**6. Visualization and Interpretation**

* The accuracy, precision, recall, and F1-score provide a comprehensive understanding of the model's performance. While accuracy is useful for a general overview, precision, recall, and F1-score are especially important when dealing with imbalanced data or when false positives and false negatives carry different consequences.
* **Bar charts** and **scatter plots** are used to visualize the relationships between different features (e.g., entity counts, article length, sentiment) and the classification label (real or fake).
* **Heatmaps** are generated to show the correlation between the features and the target variable (real or fake), helping to identify which features are most predictive of article veracity.

**Insights on how named entities impact article engagement and popularity:-**

**1. Relevance to the Audience**

* **Organizations (ORG):** Articles mentioning well-known organizations (e.g., companies, media outlets, tech giants, etc.) tend to attract more attention due to the audience's familiarity and interest in these organizations. For example, articles about major corporations like Apple or Google are likely to receive more engagement due to the widespread interest in these entities.
* **Geopolitical Entities (GPE):** Mentions of countries, cities, or regions (e.g., the United States, Paris, or the Middle East) can also significantly impact engagement. Articles related to current events in well-known GPEs, especially those with ongoing political or social issues, tend to generate more interest. For example, news about political unrest in a major country like the U.S. or China can attract more readers compared to news about less globally known locations.
* **Persons (PERSON):** Articles mentioning celebrities, politicians, or influential individuals often generate significant engagement, especially if these individuals are currently involved in trending news or controversies. For instance, articles that mention famous personalities like Elon Musk or public figures in politics tend to go viral, driven by the public's fascination with these figures.

**2. Article Trustworthiness and Perception**

* The presence of recognized named entities can sometimes influence the perceived credibility of an article. When articles cite reputable organizations or well-known figures, readers might assume the information is more reliable. This effect can be leveraged in both real and fake news, where well-known entities may lend credibility to a story that could otherwise be questioned.
* Conversely, articles that mention entities associated with controversial or unreliable sources (e.g., lesser-known or discredited organizations) might decrease perceived credibility, impacting reader engagement negatively.

**3. Sentiment and Named Entities**

* The sentiment surrounding named entities can also influence engagement. Positive sentiment around a person or organization can lead to a more favorable reception of the article, attracting more readers. For example, a positive news story about a popular celebrity or a breakthrough by a major company like Tesla can attract large audiences.
* Negative sentiment, on the other hand, particularly when associated with controversial entities (e.g., politicians, CEOs, etc.), might provoke strong reactions and lead to higher engagement in the form of debates, shares, and discussions. Articles involving political figures or global leaders often generate polarized responses, leading to increased popularity and engagement.