FINALS: Simulate Dataset

Title: " Fake and Real News Dataset"

Description

This analysis explores the Fake and Real News Dataset to identify patterns that differentiate fake news from real news using three (3) different algorithms. By analyzing key attributes such as title, text, subject, date, and label, we aim to develop a predictive model that can classify news articles accurately. The methodology involves data preprocessing, feature extraction, model training, and evaluation, with a focus on metrics like accuracy and F1-score. The ultimate goal is to create a reliable model for distinguishing news authenticity, providing a valuable tool to combat misinformation.

STEP-BY-STEP PROCESS:

1. Data Simulation

Given the attributes, a simulated dataset was created using CSV. The dataset includes attributes such as title, text, subject, date, and label.

2. Setup the Environment

- Install VSCode: Download and install Visual Studio Code from here.
- Install Python: Ensure you have Python installed. You can download it from here.
- Install Extensions: In VSCode, install the Python extension for syntax highlighting, debugging, and other features.

Install Required Libraries

- Open a terminal in VSCode (Ctrl+`). Then type “python -m venv”
- Create a virtual environment (optional but recommended):

Load and Preprocess the Dataset

- Download the Dataset: Ensure you have the Fake and Real News Dataset in a CSV format.
- Create a new Python script: In VSCode, create a new file, e.g., news_analysis.py.
- Write the code to load and preprocess the dataset:

Run the Simulation

This approach provides a basic framework for simulating the dataset and building a predictive model using Python in VSCode. Adjust the preprocessing steps and model parameters as needed for more sophisticated analysis.

3. Insights

a. Gain insights into combinations of attributes such as class, age, sex, and fare that are associated with fake and real news outcomes in the dataset.

Support Vector Machine (SVM)

(Best Algorithm)

Insights

Accuracy

The overall accuracy of the SVM classifier is **99.43%**. This indicates that 99.43% of the instances in the dataset were correctly classified by the model.

Confusion Matrix

The confusion matrix provides a detailed breakdown of the model's performance by showing the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions.

- The top-left cell (4661) represents the number of true negatives (actual negatives correctly classified).
- The top-right cell (31) represents the number of false positives (actual negatives incorrectly classified as positives).
- The bottom-left cell (20) represents the number of false negatives (actual positives incorrectly classified as negatives).
- The bottom-right cell (4268) represents the number of true positives (actual positives correctly classified).

Classification Report

The classification report provides a summary of evaluation metrics for each class (0 and 1) in the dataset.

- **Precision:** Precision is the ratio of true positives to the total number of predicted positives. For class 0, precision is 1.00, and for class 1, precision is 0.99.

- **Recall:** Recall (also known as sensitivity) is the ratio of true positives to the total number of actual positives. For class 0, recall is 0.99, and for class 1, recall is 1.00.
- **F1-score:** F1-score is the harmonic mean of precision and recall. It provides a balance between precision and recall. For both classes, the F1-score is 0.99.
- **Support:** Support is the number of actual occurrences of each class in the dataset.

Macro Avg and Weighted Avg:

These are the average scores of precision, recall, and F1-score across all classes. Macro average calculates the unweighted mean of the scores, while weighted average calculates the mean weighted by the support of each class. In this case, both macro and weighted averages are 0.99 for precision, recall, and F1-score.

Recommendation

- Deploy the trained SVM classifier for real-world use, leveraging its high accuracy and robust performance.
- Implement monitoring mechanisms to regularly evaluate the model's performance over time, tracking metrics such as accuracy, precision, recall, and F1-score.
- Periodically update the SVM classifier with new data to improve its accuracy and generalization capabilities.
- Conduct thorough data quality assessments to ensure that the input data remains representative and free from biases or anomalies.

Random Forest (2nd)

Insights

Accuracy

This indicates the proportion of correctly classified instances out of the total instances. In this case, the accuracy is approximately 0.9886, which means the model correctly classified about 98.86% of the instances.

Confusion Matrix

This is a table that shows the number of true positives, true negatives, false positives, and false negatives. It helps you understand the performance of your classification model. In this confusion matrix:

- True Positives (TP): 4234 instances were correctly predicted as class 1.
- False Positives (FP): 48 instances were incorrectly predicted as class 1.
- True Negatives (TN): 4644 instances were correctly predicted as class 0.
- False Negatives (FN): 54 instances were incorrectly predicted as class 0.

Classification Report

- **Precision:** The proportion of true positive predictions out of all positive predictions. High precision indicates low false positive rate. Precision for class 0 and class 1 is both approximately 0.99, indicating that the model's predictions are precise for both classes.
- **Recall (also called Sensitivity or True Positive Rate):** The proportion of true positive predictions out of all actual positives. High recall indicates low false negative rate. Recall for class 0 and class 1 is both approximately 0.99, indicating that the model captures most of the positive instances for both classes.
- **F1-score:** The harmonic mean of precision and recall. It provides a balance between precision and recall. F1-score for class 0 and class 1 is both approximately 0.99.
- **Macro avg:** The average of the precision, recall, and F1-score for both classes, giving each class equal weight.
- **Weighted avg:** The weighted average of precision, recall, and F1-score, where each class's score is weighted by its support (the number of true instances).

Recommendation

- Review your dataset to see if there are any features that aren't contributing much to the model's performance. Removing or simplifying these less important features could improve efficiency and potentially accuracy.
- Experiment with different settings for the Random Forest model's parameters. Tweaking parameters like the number of trees in the forest or the depth of each tree can help find the best configuration for your specific dataset, potentially boosting accuracy.

Naive Bayes
(3rd)

Insights

Accuracy

Accuracy measures the proportion of correctly classified instances out of the total instances. In your case, the accuracy is approximately 93.61%, indicating that the model correctly classifies about 93.61% of the instances.

Confusion Matrix

This is a table that describes the performance of a classification model. It presents a summary of the correct and incorrect classifications done by the classifier. In your confusion matrix:

- True Positives (TP): 4433 instances were correctly classified as class 0.
- False Positives (FP): 259 instances were incorrectly classified as class 1.
- False Negatives (FN): 315 instances were incorrectly classified as class 0.
- True Negatives (TN): 3973 instances were correctly classified as class 1.

Precision

Precision is the ratio of correctly predicted positive observations to the total predicted positives. It measures the accuracy of positive predictions. Your precision for class 0 (0) is 0.93, and for class 1 (1) is 0.94.

Recall (Sensitivity)

Recall is the ratio of correctly predicted positive observations to the all observations in actual class. It measures the ability of the classifier to find all the positive samples. Your recall for class 0 (0) is 0.94, and for class 1 (1) is 0.93.

F1-score

F1-score is the harmonic mean of precision and recall. It provides a balance between precision and recall. Your F1-score for class 0 (0) is 0.94, and for class 1 (1) is 0.93.

Recommendation

- Since both precision and recall are high for both classes, it indicates that your model is effectively identifying instances of both classes without too many false positives or false negatives. This suggests that the model is doing well in distinguishing between the two classes.
- Even though the overall performance is good, it's still essential to review the instances that were misclassified (false positives and false negatives). Investigating these instances can provide insights into why the model made errors and potentially help improve its performance further.
- While Naive Bayes is performing well, it's worth experimenting with other classification algorithms to see if they can achieve even better performance on your dataset. Techniques like ensemble methods or more complex models might uncover patterns that Naive Bayes cannot capture.

Submitted by: ENDcoders

LEADER: Bernal, Christian Dave

MEMBERS:

Arendayen, Ajhay

Custodio, Melvin

Galanaga, Desiree

Molina, Ricky James

Adviser:

Prof. Cherry Rose Concha