Project Title: Fake News Detection Model using Machine Learning

Technology Platform: Python on Anaconda Jupyter

Technical Domain: Machine learning

Business Domain: Media

Done by: Mandadi Harshitha Reddy

Suggestions and Findings:

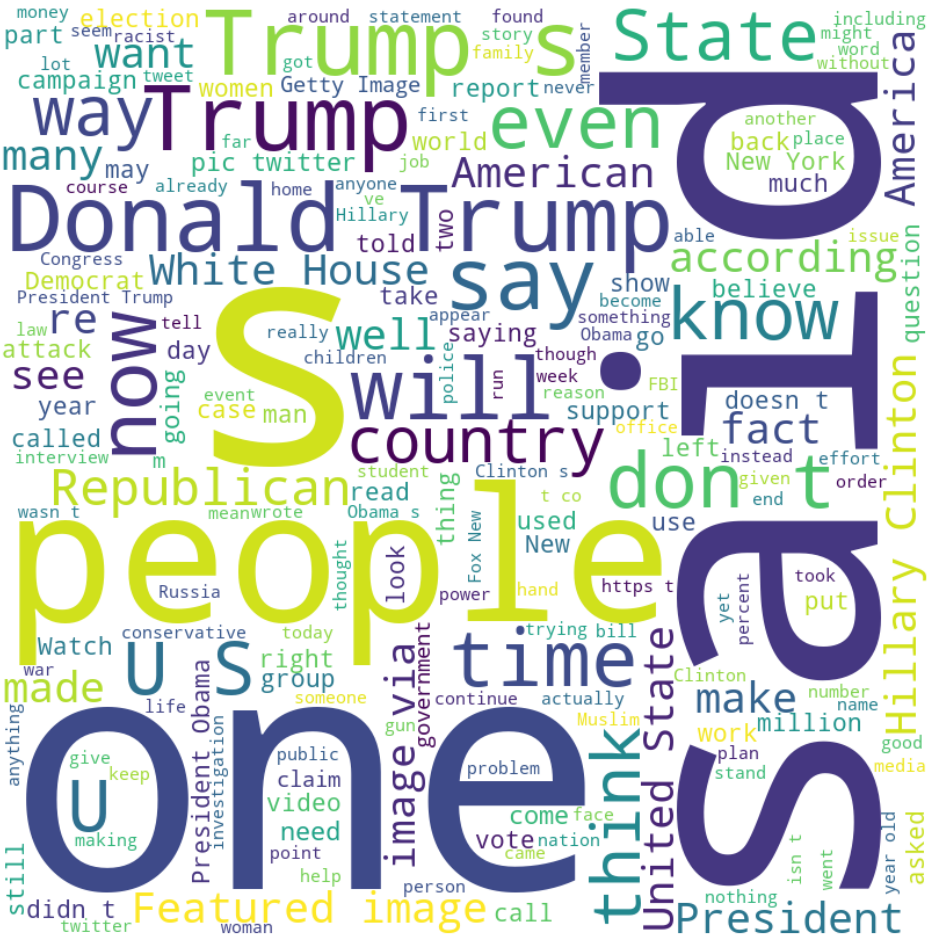
- Based on the evaluations, the Decision Tree Classifier proves to be the most effective model for fake news detection, offering the highest accuracy, precision, recall, and F1-scores. However, caution should be exercised regarding potential overfitting. It is recommended to validate the Decision Tree's performance on an independent dataset to ensure it generalizes well to unseen data.
- Logistic Regression and Random Forest Classifiers also demonstrated strong performance with high accuracy and balanced metrics. These models serve as robust alternatives to the Decision Tree, providing reliable results for fake news detection. Their consistent performance suggests they can be considered for scenarios where model interpretability or stability is crucial.
- The Naive Bayes model, while still effective, shows lower accuracy and performance metrics compared to the other models. Its performance suggests that it may be less suitable for this particular task compared to Decision Trees, Logistic Regression, or Random Forest.
- The data analysis revealed key patterns in both fake and true news articles. For fake news, terms related to politics and multimedia elements were prominent, while true news articles frequently included specific dates, quantitative information, and references to established news agencies. This insight can guide further feature engineering and model refinement.
- The word clouds from the datasets highlight the different focuses of fake and true news articles. Incorporating additional features based on these observations—such as political terms or specific day references—might enhance model performance by capturing more nuanced patterns in the data. This approach can help fine-tune the models and potentially enhance their accuracy and robustness.

- Given the high accuracy of the Decision Tree model, it's vital to check for overfitting. Techniques such as pruning, setting constraints on the tree depth, and using cross-validation can help mitigate overfitting and improve the model's generalization.

DATA ANALYSIS

Following is the data analysis done in Jupyter notebook.

1. **Word cloud from the text data in the 'fake' dataset**



Observation:

- These terms are very large and central, indicating that a significant portion of the fake news articles in the dataset involve or mention Donald Trump.
- **"People",** this word is also very large, suggesting that many fake news articles refer to people or involve general statements about people.

- **"State", "Country", "Republican", "U.S.", "President"**, these terms indicate that fake news articles often discuss topics related to politics, government, and national affairs.
- Words like **"Trump," "President," "Republican," "Democrat," "Congress,"** and **"White House"** show that fake news articles frequently cover political topics.
- Words like **"one," "people," "time," "know,"** and **"say"** are common, indicating that fake news articles often use general and conversational language.
- Words like **"say," "will," "make," "called," "attack,"** and **"going"** suggest that the articles often discuss events and actions.
- **"Image", "video", "read", "reporter",** these terms indicate that fake news articles may include multimedia elements or discuss media coverage.

2. Word cloud from the text data in the 'true' dataset

Observation:

- **"Said"** word is the most prominent, indicating that legitimate news articles frequently quote people, providing attributions and statements.
- **"Trump"** and **"Donald Trump"**, these terms appear prominently, suggesting that Donald Trump is a significant focus in many legitimate news articles.
- **"U.S."** and **"United States"**, these terms are large and central, indicating that many articles discuss topics related to the United States.
- Words like **"Trump," "President," "government," "White House," "Republican," "Democrat,"** and **"Congress"** show that legitimate news articles heavily cover political topics.
- Frequent words like **"statement," "according," "said," "called," "meeting," "plan,"** and **"support"** suggest that the articles often discuss events, actions, and statements made by officials.
- Words like **"Thursday," "Friday," "Monday," "Tuesday,"** and **"Wednesday"** indicate that legitimate news articles often reference specific days of the week, likely related to events or statements made on those days.
- Words like **"China," "Russia," "North Korea," "Saudi Arabia,"** and **"Turkey"** suggest that legitimate news articles frequently cover international topics and geopolitical issues.
- References to places like **"New York"** and **"Washington"** indicate a focus on key locations within the United States.
- Words like **"million," "percent,"** and **"billion"** imply that legitimate news articles often include quantitative information and statistics.
- Words like **"Reuters"** show that legitimate news articles often reference established news agencies as sources of information.

MODEL EVALUATION

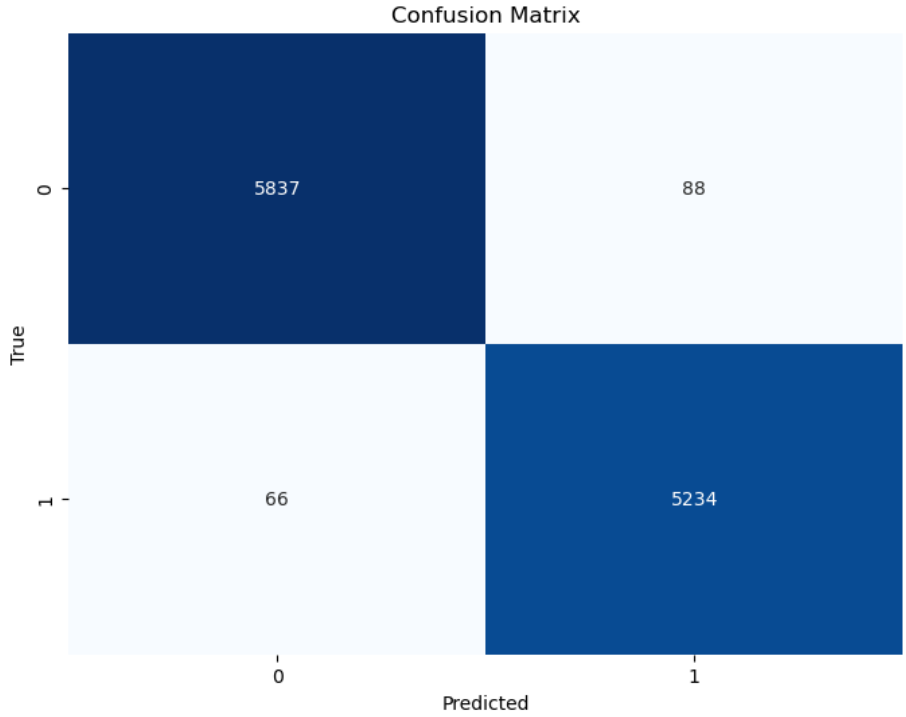
To detect fake news, I have built four models:

- Logistic Regression,
- Naive Bayes,
- Decision Tree Classifier,
- Random Forest Classifier

3. Evaluation of Fake news using Logistic Regression Model

```
Logistic Regression accuracy: 0.9862806236080178
```

	precision	recall	f1-score	support
0	0.99	0.99	0.99	5925
1	0.98	0.99	0.99	5300
accuracy			0.99	11225
macro avg	0.99	0.99	0.99	11225
weighted avg	0.99	0.99	0.99	11225



Observation:

- The model has a very high overall accuracy of 98.63%, indicating that it correctly classifies 98.63% of the news articles as either fake or real.
- Precision measures the accuracy of the positive predictions. A high precision for both classes (0.99 for class 0 and 0.98 for class 1) means that when the model predicts an article as fake or real, it is very likely to be correct.
- Recall measures the ability of the model to find all relevant instances in the dataset. The recall for both classes (0.99 for both class 0 and class 1) is also very high, indicating that the model successfully identifies most of the fake and real news articles.
- F1-Score is the harmonic mean of precision and recall. High F1-scores (0.99 for both classes) indicate a good balance between precision and recall.
- The support values (5925 for class 0 and 5300 for class 1) indicate the number of instances for each class in the test set. The model performs well across both classes, suggesting that it is not biased towards either class.
- The macro average of precision, recall, and F1-score is 0.99, indicating that the model performs well across both classes without taking the imbalance into account.
- The weighted average is also 0.99, which takes into account the number of instances in each class, confirming that the model performs well overall.

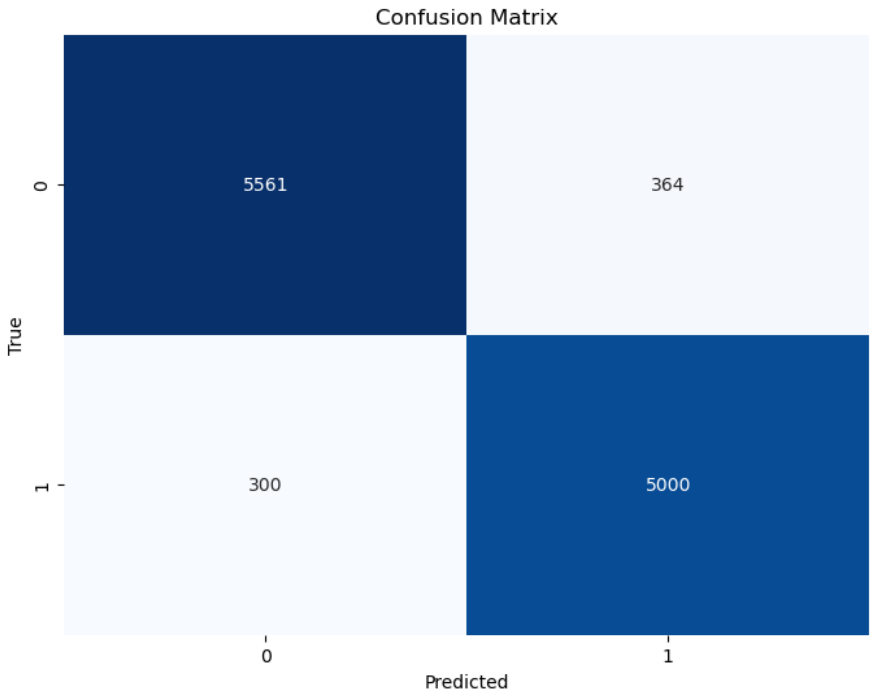
4. Evaluation of Fake news using Naïve Bayes Model

```
Naive Bayes Accuarcy: 0.9408463251670378

              precision    recall  f1-score   support

           0       0.95      0.94      0.94      5925
           1       0.93      0.94      0.94      5300

    accuracy                           0.94     11225
   macro avg       0.94      0.94      0.94     11225
weighted avg       0.94      0.94      0.94     11225
```



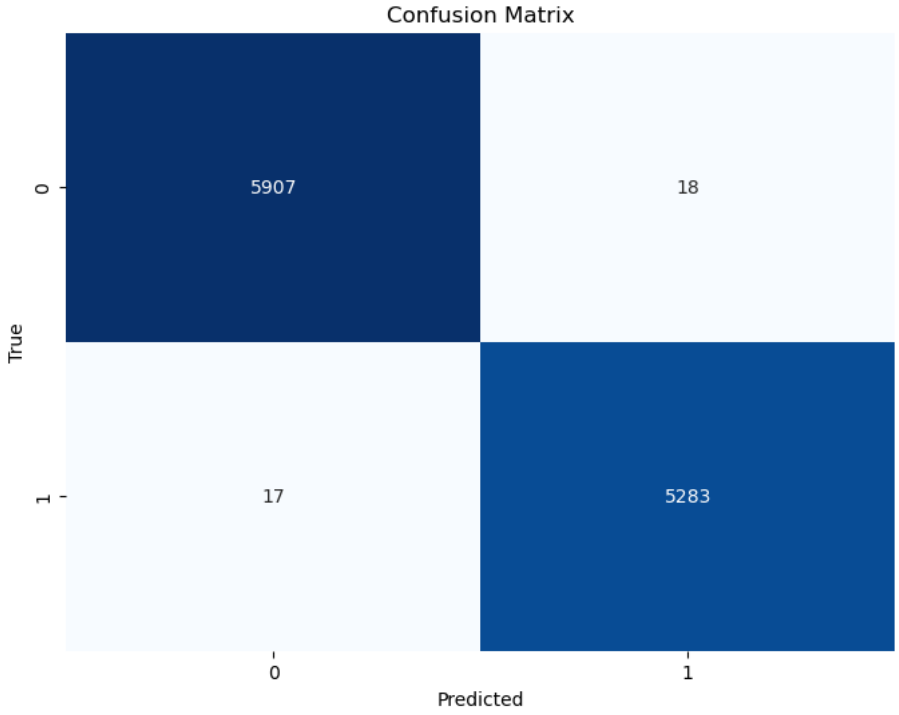
Observation:

- The model has a high overall accuracy of 94.08%, indicating that it correctly classifies 94.08% of the news articles as either fake or real.
- The precision for class 0 (fake news) is 0.95, and for class 1 (real news) is 0.93. This means that when the model predicts an article as fake, it is correct 95% of the time, and when it predicts an article as real, it is correct 93% of the time.
- The recall for both classes is 0.94. This indicates that the model correctly identifies 94% of both fake and real news articles.
- The F1-score, which is the harmonic mean of precision and recall, is 0.94 for both classes, indicating a good balance between precision and recall.
- The support values (5925 for class 0 and 5300 for class 1) indicate the number of instances for each class in the test set. The model performs consistently across both classes.
- The macro average of precision, recall, and F1-score is 0.94, showing that the model performs well across both classes without considering the imbalance.
- The weighted average is also 0.94, which accounts for the number of instances in each class, confirming the model's overall performance.

5. Evaluation of Fake news using Decision Tree Model

Decision tree classifier accuracy: 0.9968819599109131

	precision	recall	f1-score	support
0	1.00	1.00	1.00	5925
1	1.00	1.00	1.00	5300
accuracy			1.00	11225
macro avg	1.00	1.00	1.00	11225
weighted avg	1.00	1.00	1.00	11225



Observation:

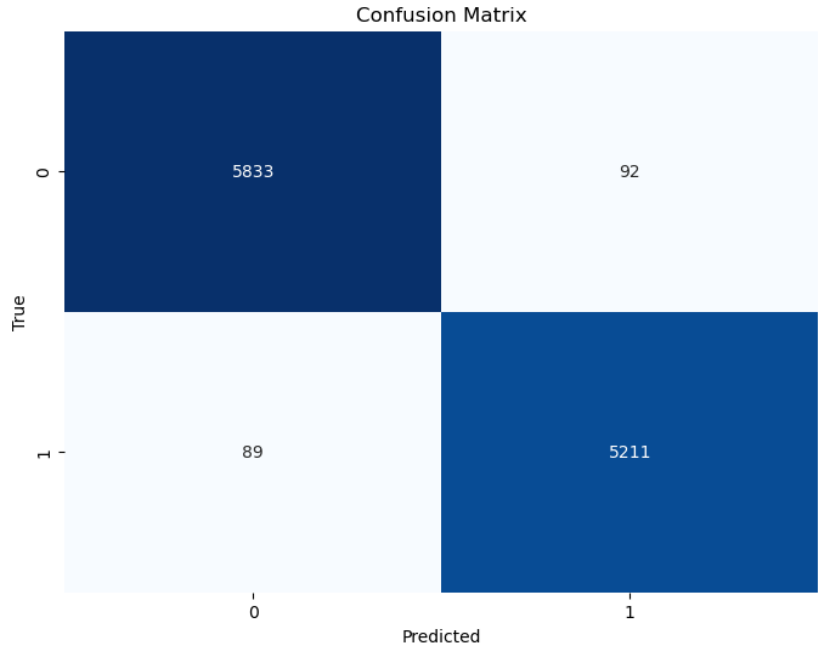
- The model has an extremely high overall accuracy of 99.69%, indicating that it correctly classifies 99.69% of the news articles as either fake or real.
- The precision for both class 0 (fake news) and class 1 (real news) is 1.00. This means that when the model predicts an article as fake or real, it is almost always correct.
- The recall for both classes is also 1.00, indicating that the model correctly identifies almost all instances of both fake and real news articles.

- The F1-score for both classes is 1.00, demonstrating a perfect balance between precision and recall.
- The support values (5925 for class 0 and 5300 for class 1) indicate the number of instances for each class in the test set. The model performs perfectly across both classes.
- The macro average of precision, recall, and F1-score is 1.00, showing that the model performs perfectly across both classes without considering the imbalance.
- The weighted average is also 1.00, which accounts for the number of instances in each class, confirming the model's exceptional overall performance.

6. Evaluation of Fake news using Random Forest Classifier Model

```
Random Forest Classifier accuarcy: 0.9838752783964365
```

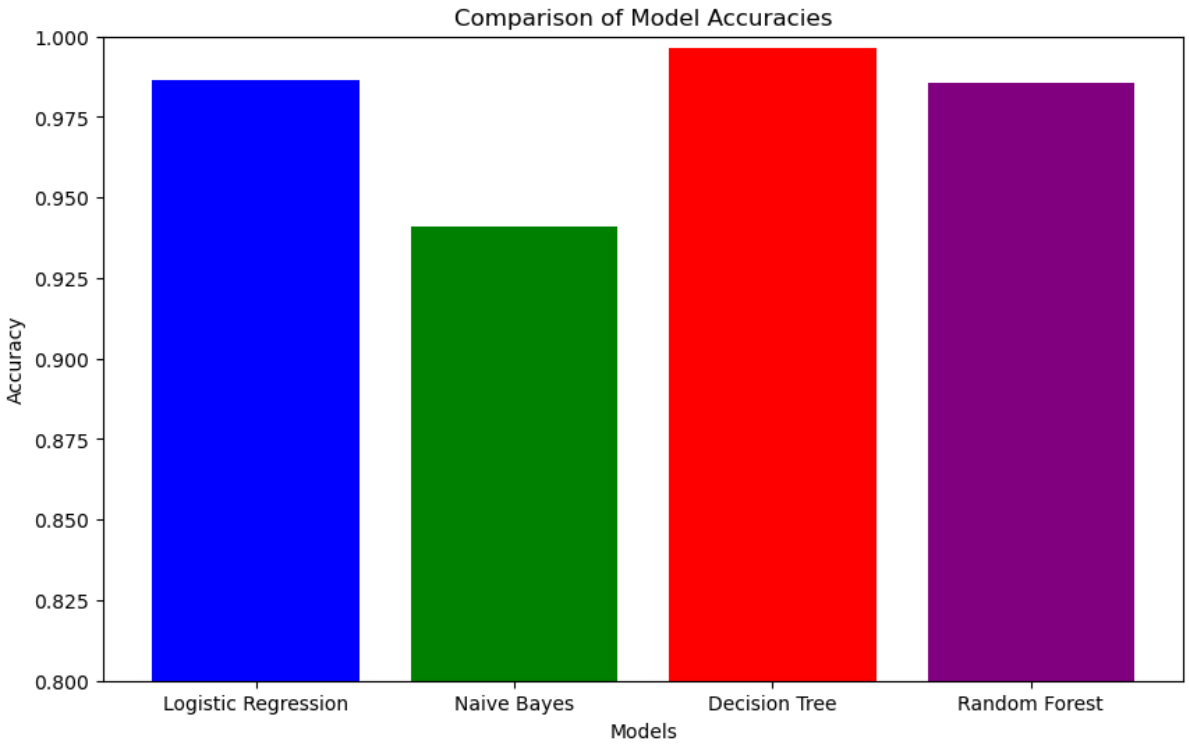
	precision	recall	f1-score	support
0	0.98	0.98	0.98	5925
1	0.98	0.98	0.98	5300
accuracy			0.98	11225
macro avg	0.98	0.98	0.98	11225
weighted avg	0.98	0.98	0.98	11225



Observation:

- The Random Forest Classifier performed exceptionally well on your dataset with an accuracy of approximately 98.39%. Here's a breakdown of the classification report:
- The precision for both classes (0 and 1), the precision is 0.98. This means that when the model predicts a class, it's correct 98% of the time. Precision is particularly important in contexts where false positives are costly.
- The recall is also 0.98 for both classes. This indicates that the model successfully identifies 98% of the actual instances of each class. Recall is crucial when you want to ensure you don't miss any positive cases.
- The F1-score, which balances precision and recall, is 0.98 for both classes. This suggests that the model has a good balance between precision and recall, making it robust in terms of both metrics.
- Support vectors reflect the number of actual occurrences of each class in the test set. Both classes are well-represented with 5925 instances of class 0 and 5300 instances of class 1.
- This is the average performance of the model across classes without considering their support. Here, the macro average is 0.98 for precision, recall, and F1-score, indicating consistent performance across both classes.
- Weighted Average accounts for the class imbalance by weighting the metrics according to the number of true instances for each class. The weighted averages are also 0.98 for all metrics, suggesting that the model's performance is balanced despite the class sizes.

7. Comparison of different models in accuracy and other parameters



- The Decision Tree has the highest accuracy (99.63%), followed by Logistic Regression (98.63%), Random Forest (98.55%), and Naive Bayes (94.08%).
- The Decision Tree also has the highest precision, recall, and F1-score for both classes (1.00 for all metrics). Logistic Regression and Random Forest have very similar metrics, which are slightly lower than those of the Decision Tree. Naive Bayes has the lowest performance across these metrics.
- Decision Tree is the best model for detecting fake news based on this comparison. It provides the highest accuracy and perfect precision, recall, and F1-scores for both classes. However, it's worth noting that Decision Trees can sometimes overfit the data, but in this case, given the perfect scores, it's crucial to ensure that the model is not overfitting to the training data.
- Logistic Regression and Random Forest are also strong contenders, showing high performance across the metrics. Naive Bayes is less effective in this case, as it has lower accuracy and lower metrics across the board.

CONCLUSION

Based on the evaluation of various machine learning models for detecting fake news, the Decision Tree Classifier emerges as the most effective choice. It achieved an exceptional accuracy of 99.69%, with perfect precision, recall, and F1-scores for both fake and real news categories. This indicates that the Decision Tree accurately classifies news articles with a high degree of reliability and consistency. However, it's important to consider that Decision Trees can sometimes overfit the training data, so it's essential to ensure that the model generalizes well to unseen data.

While the Decision Tree Classifier stands out, Logistic Regression and Random Forest Classifiers also demonstrated strong performance. Logistic Regression achieved an accuracy of 98.63% and exhibited a good balance between precision, recall, and F1-scores, making it a robust alternative. The Random Forest Classifier, with an accuracy of 98.39%, provided similar high performance and was only slightly less accurate than the Decision Tree. On the other hand, the Naive Bayes model, despite being effective, showed lower accuracy and performance metrics compared to the other models.

In conclusion, the Decision Tree Classifier is the best option for fake news detection based on its superior accuracy and balanced performance metrics. Nonetheless, validating the model's performance on new data is crucial to ensure it does not overfit. Logistic Regression and Random Forest are also highly effective models, offering reliable alternatives with strong performance across the board.