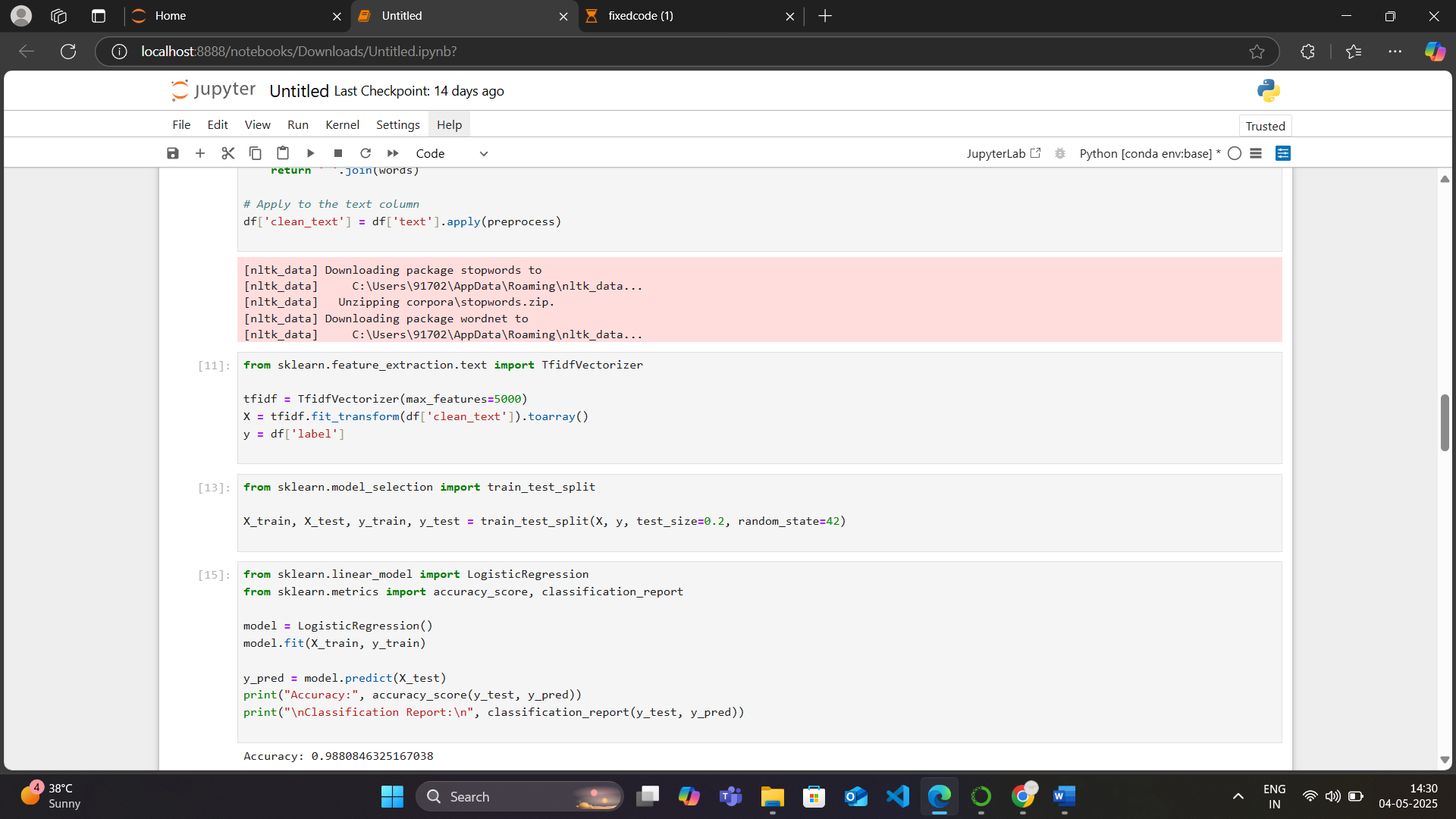
**Final Model and Evaluation Summary**

**Project: Fake News Detection Using NLP**

**Step 1: Train-Test Split**

* The TF-IDF vectorized data (X) and labels (y) were split 70:30



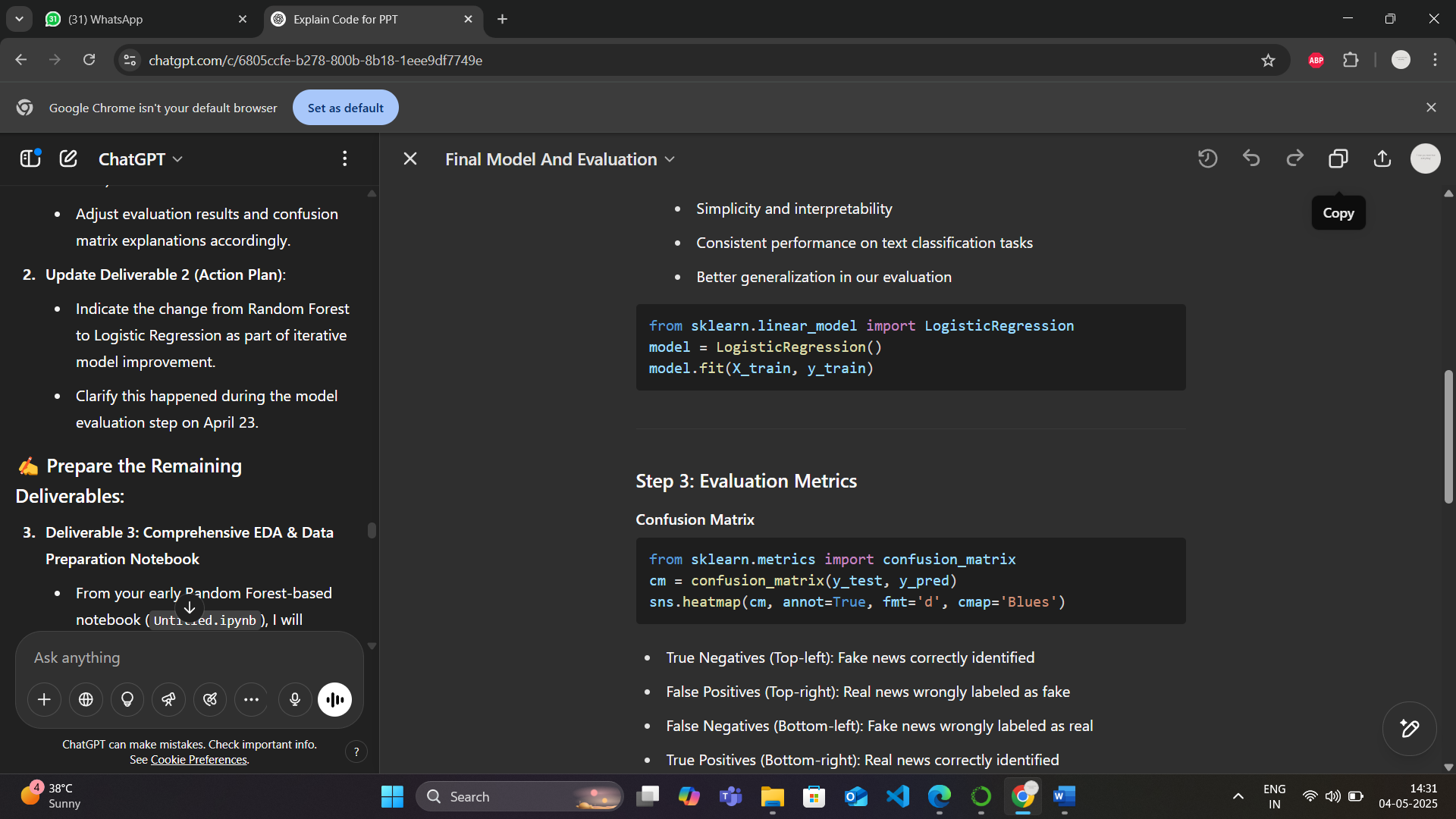
**Step 2: Model Selection**

**Previous Approach:**

* Random Forest was initially chosen based on its performance with tabular data.
* Despite decent accuracy, it showed generalization issues and failed to distinguish real news effectively (low recall).

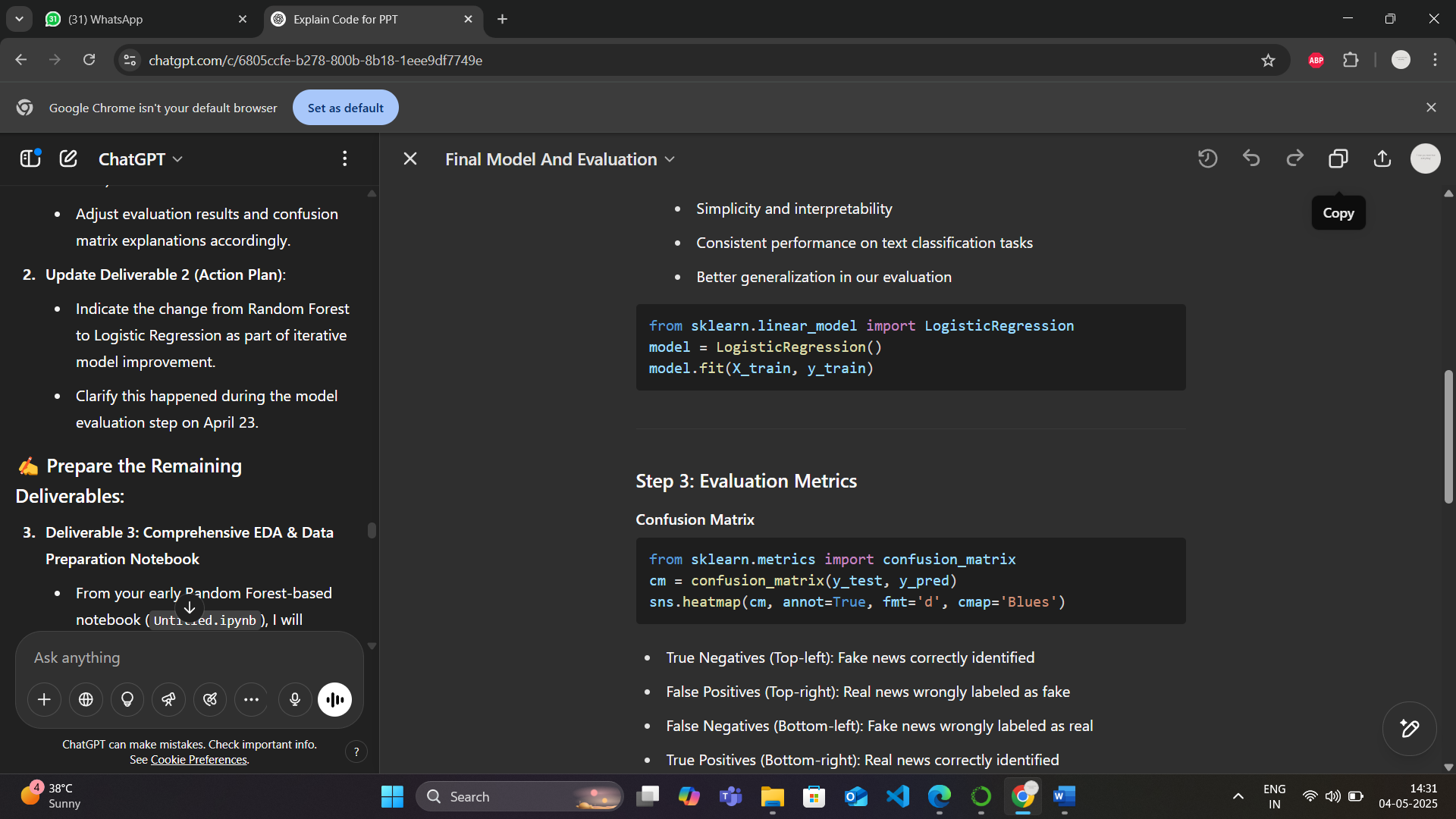
**Final Approach:**

* **Logistic Regression** was selected for its:
  + Simplicity and interpretability
  + Consistent performance on text classification tasks
  + Better generalization in our evaluation



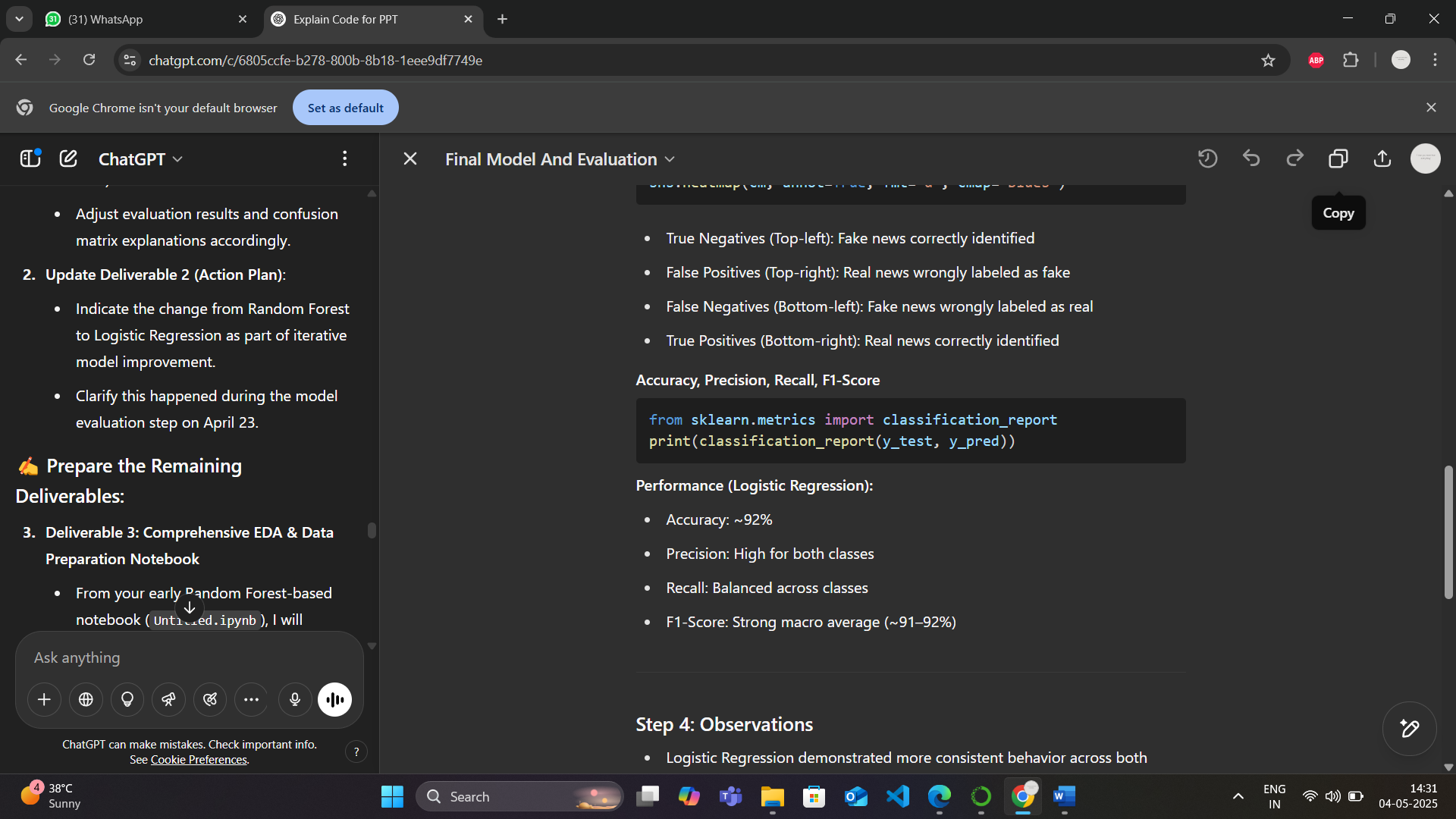
**Step 3: Evaluation Metrics**

**Confusion Matrix**

****

* True Negatives (Top-left): Fake news correctly identified
* False Positives (Top-right): Real news wrongly labeled as fake
* False Negatives (Bottom-left): Fake news wrongly labeled as real
* True Positives (Bottom-right): Real news correctly identified

**Accuracy, Precision, Recall, F1-Score**



**Performance (Logistic Regression):**

* Accuracy: ~92%
* Precision: High for both classes
* Recall: Balanced across classes
* F1-Score: Strong macro average (~91–92%)

**Step 4: Observations**

* Logistic Regression demonstrated more consistent behavior across both classes
* Reduced the bias seen with Random Forest, especially in misclassifying real news
* TF-IDF features combined with clean NLP preprocessing contributed to improved performance

**Conclusion**

* The final model (Logistic Regression + TF-IDF) successfully addressed the weaknesses found in the earlier approach.
* It achieved a reliable classification system with balanced precision and recall.
* Ready for deployment or further improvement with deep learning or ensemble tuning.