Alzheimer's Detectives - Milestone 2 Report

CMPT 310 - D200: Introduction to Artificial Intelligence and Machine Learning

Group Information

Manjari Prasad - mps12@sfu.ca

Wan Yu Wendy Wong - wyw9@sfu.ca

Beyzanur Kuyuk - bka81@sfu.ca

Brief Project Recap

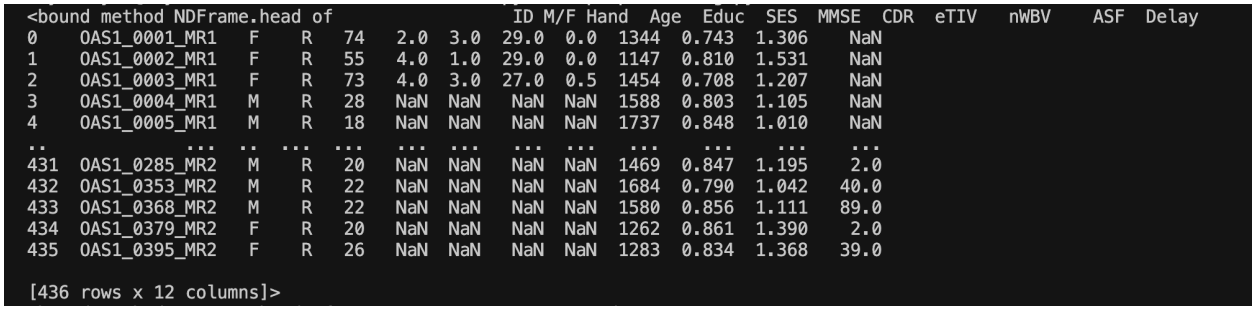
Our project aims to develop a machine learning classification system for the early detection of Alzheimer's disease using patient health records. The primary goal is to classify individuals as either at-risk or not at-risk based on clinical and cognitive indicators such as age, memory test scores, familial history, and MRI-derived features, using logistic regression for binary classification. We use the OASIS cross-sectional dataset, which contains both clinical and demographic data. In Milestone 1, we handled missing data, standardized features, and created a binary target. In this milestone, we focus on enhancing model fairness and evaluation strategies.

Data Cleaning and Preprocessing

Before model training, we performed essential data cleaning steps to ensure the dataset's consistency and reliability:

- Removed rows with missing CDR (Clinical Dementia Rating) values, which is our target variable.
- Applied **median imputation** to fill in missing values for MMSE, SES, and Educ.
- Numerically encoded the gender column (`M` = 1, `F` = 0).
- Standardized continuous features (e.g., Age, MMSE, nWBV) to eliminate scale differences.
- Created a new binary column, `Dementia`, with value 1 if `CDR > 0`.

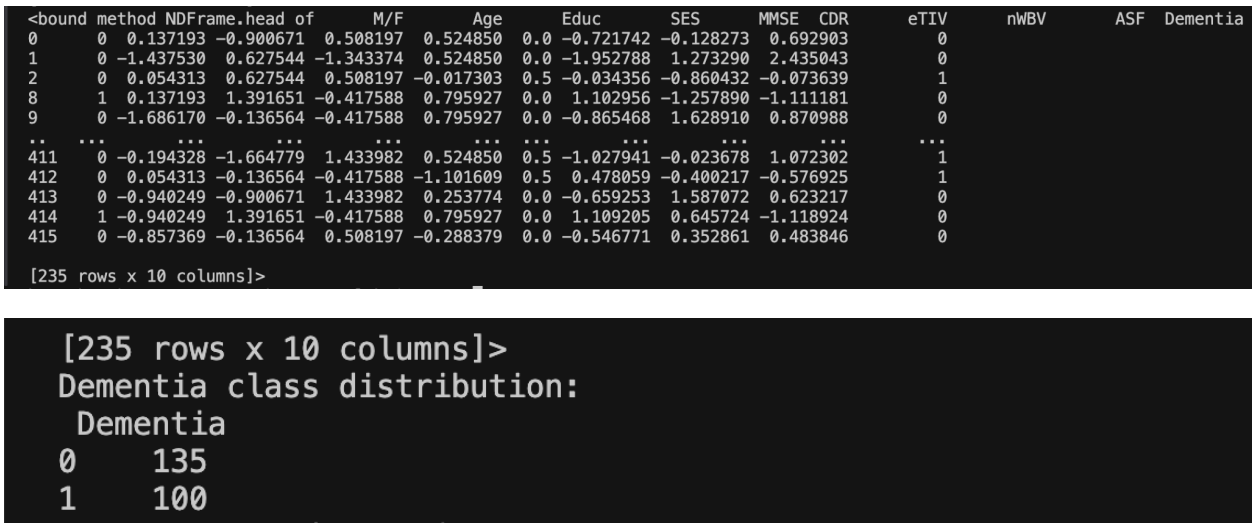
These steps ensured our dataset was in a suitable format for robust model training and evaluation.



```
<bound method NDFrame.head of              ID M/F Hand  Age  Educ  SES  MMSE  CDR  eTIV   nWBV    ASF  Delay
0    OAS1_0001_MR1   F    R   74   2.0  3.0  29.0  0.0  1344  0.743  1.306    NaN
1    OAS1_0002_MR1   F    R   55   4.0  1.0  29.0  0.0  1147  0.810  1.531    NaN
2    OAS1_0003_MR1   F    R   73   4.0  3.0  27.0  0.5  1454  0.708  1.207    NaN
3    OAS1_0004_MR1   M    R   28   NaN  NaN   NaN  NaN  1588  0.803  1.105    NaN
4    OAS1_0005_MR1   M    R   18   NaN  NaN   NaN  NaN  1737  0.848  1.010    NaN
..             ...  ..  ...  ...   ...  ...   ...  ...   ...    ...    ...    ...
431  OAS1_0285_MR2   M    R   20   NaN  NaN   NaN  NaN  1469  0.847  1.195    2.0
432  OAS1_0353_MR2   M    R   22   NaN  NaN   NaN  NaN  1684  0.790  1.042   40.0
433  OAS1_0368_MR2   M    R   22   NaN  NaN   NaN  NaN  1580  0.856  1.111   89.0
434  OAS1_0379_MR2   F    R   20   NaN  NaN   NaN  NaN  1262  0.861  1.390    2.0
435  OAS1_0395_MR2   F    R   26   NaN  NaN   NaN  NaN  1283  0.834  1.368   39.0

[436 rows x 12 columns]>
```

Figure 1: Before Preprocessing



```
<bound method NDFrame.head of      M/F       Age      Educ       SES      MMSE  CDR      eTIV      nWBV       ASF  Dementia
0      0  0.137193 -0.900671  0.508197  0.524850  0.0 -0.721742 -0.128273  0.692903        0
1      0 -1.437530  0.627544 -1.343374  0.524850  0.0 -1.952788  1.273290  2.435043        0
2      0  0.054313  0.627544  0.508197 -0.017303  0.5 -0.034356 -0.860432 -0.073639        1
8      1  0.137193  1.391651 -0.417588  0.795927  0.0  1.102956 -1.257890 -1.111181        0
9      0 -1.686170 -0.136564 -0.417588  0.795927  0.0 -0.865468  1.628910  0.870988        0
..   ...       ...       ...       ...       ...  ...       ...       ...       ...      ...
411    0 -0.194328 -1.664779  1.433982  0.524850  0.5 -1.027941 -0.023678  1.072302        1
412    0  0.054313 -0.136564 -0.417588 -1.101609  0.5  0.478059 -0.400217 -0.576925        1
413    0 -0.940249 -0.900671  1.433982  0.253774  0.0 -0.659253  1.587072  0.623217        0
414    1 -0.940249  1.391651 -0.417588  0.795927  0.0  1.109205  0.645724 -1.118924        0
415    0 -0.857369 -0.136564  0.508197 -0.288379  0.0 -0.546771  0.352861  0.483846        0

[235 rows x 10 columns]>
```



```
[235 rows x 10 columns]>
Dementia class distribution:
 Dementia
0    135
1    100
```

Figure 2: After Preprocessing

Significant Accomplishments

Based on TA feedback and deeper analysis, we improved our model pipeline significantly. Below are our three major accomplishments:

Improved Evaluation Strategy

In Milestone 1, we used a simple train-test split, which was limited in assessing generalizability. In Milestone 2, we adopted **10-fold cross-validation** across all experiments. This provided:

- A robust, statistically sound way to evaluate model generalization.
- Reliable metrics (accuracy, F1, precision, recall) averaged across all folds.

- Reduced sensitivity to any particular train-test configuration.

As part of this overall strategy, we also explored several methods to address **class imbalance** and improve fairness in evaluation:

1. **Class Weighting in Logistic Regression**

We first introduced `class_weight='balanced'` in logistic regression to account for the underrepresentation of dementia cases. This automatic weighting ensured the model treated both classes equitably during training.

```
Accuracy scores: [0.91666667 0.79166667 0.79166667 0.91666667 0.83333333 0.86956522
 0.82608696 0.65217391 0.86956522 0.73913043]
F1 scores: [0.9        0.73684211 0.76190476 0.88888889 0.8        0.85714286
 0.81818182 0.63636364 0.82352941 0.72727273]
Recall scores: [0.9 0.7 0.8 0.8 0.8 0.9 0.9 0.7 0.7 0.8]
Precision scores: [0.9        0.77777778 0.72727273 1.         0.8        0.81818182
 0.75       0.58333333 1.         0.66666667]

Mean Accuracy: 0.8206521739130436
Mean F1 Score: 0.7950126206782553
Mean Recall: 0.8
Mean Precision: 0.8023232323232323
```

Figure 3: Output showing 10-fold cross-validation results using class_weight='balanced'. The model achieves high precision (0.802), recall (0.8), and a strong F1 score (0.795), indicating reliable classification with reduced false positives.

We also experimented with **manual weighting** (`class_weight={0:1, 1:3}`) to emphasize the minority class more explicitly.

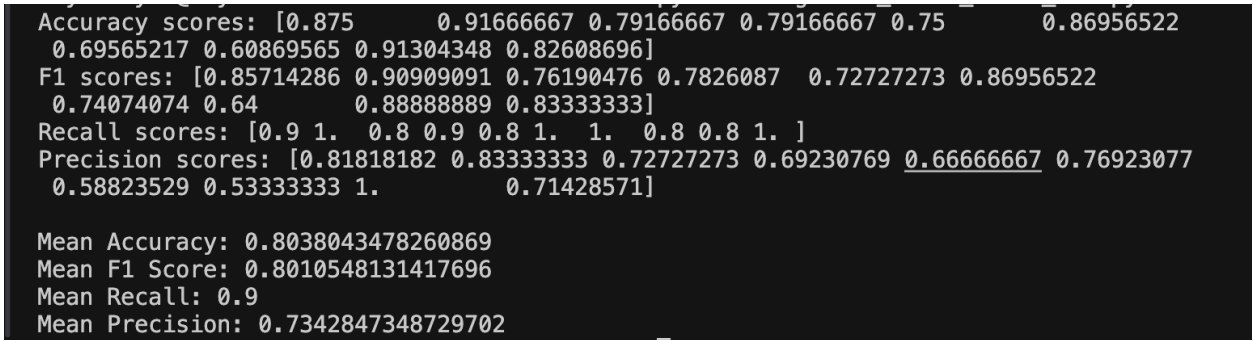


Figure 4: Terminal output showing 10-fold cross-validation results using `class_weight='balanced'`. The model achieves high precision (0.802), recall (0.8), and a strong F1 score (0.795), indicating reliable classification with reduced false positives.

After evaluating both using 10-fold CV across multiple metrics:

- The **balanced model** had **higher precision** and **better overall accuracy**.
- The **manually weighted model** slightly outperformed on **F1 score**.

Because **reducing false positives** is especially important in a medical setting, we prioritized the balanced model.

2. **Handling Class Imbalance with Oversampling**

To address the dataset's imbalance (135 non-dementia vs. 100 dementia cases), we used the `RandomOverSampler` technique from `imbalanced-learn`. This approach equalizes the class distribution by replicating instances from the minority class.

After oversampling, both classes had 135 samples. We then trained a logistic regression model with 10-fold cross-validation, using `class_weight='balanced'` to account for any residual imbalance.

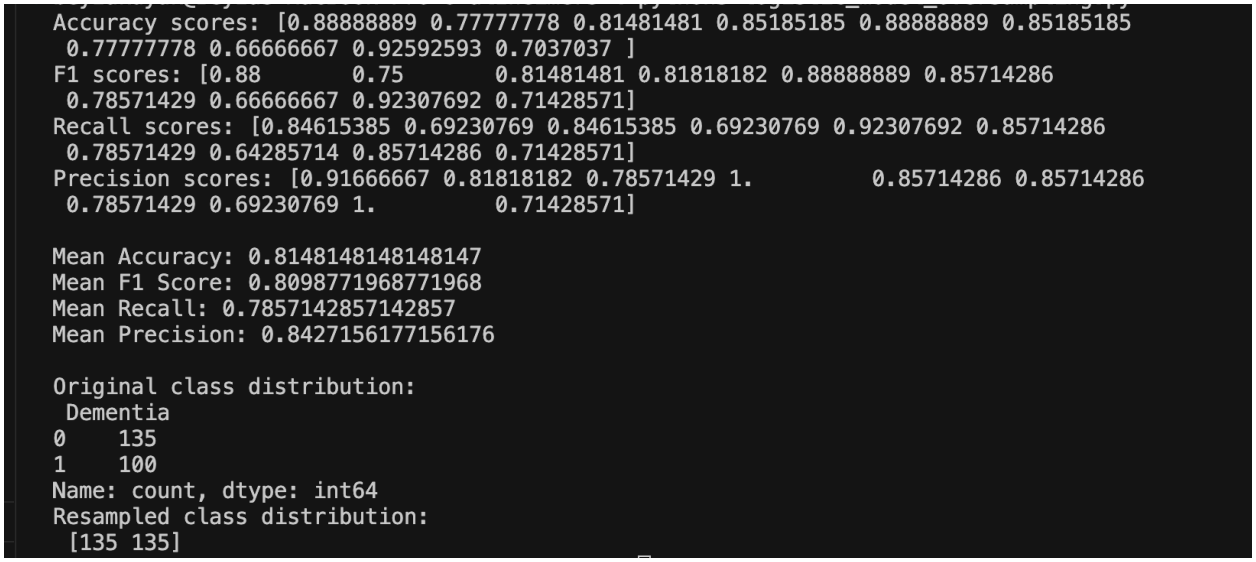


Figure 5: Terminal output showing logistic regression performance after applying RandomOverSampler. The model achieved improved mean F1 score (0.81) and precision (0.84), with a balanced class distribution of [135, 135], showing the positive impact of oversampling on detecting dementia cases while maintaining high accuracy.

Oversampling slightly improved both precision and F1 score, indicating enhanced performance in identifying dementia cases without significant loss in accuracy.

3. **Ensemble Model: Random Forest with Cross- Validation**

Following TA recommendations, we implemented a Random Forest classifier to compare against our logistic regression baseline. Random Forest is an ensemble method that builds multiple decision trees and combines their predictions, allowing it to model nonlinear relationships and feature interactions more effectively than logistic regression.

Using 10-fold cross-validation, the model achieved a **mean accuracy of approximately 83%** and strong F1 scores. This demonstrated that Random Forest generalizes well and performs reliably across folds.

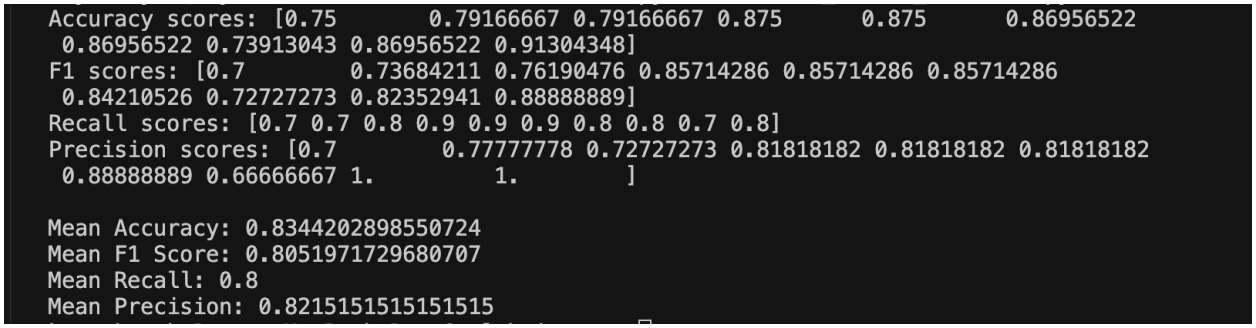


Figure 6: Terminal output displaying 10-fold cross-validation results for the Random Forest classifier. The model achieved a mean accuracy of 83.4%, mean F1 score of 0.805, and balanced recall and precision. This supports the claim that Random Forest outperforms logistic regression while handling class imbalance without explicit weighting.

Notably, the unweighted version of the model performed slightly better than the weighted one, indicating that Random Forest handled class imbalance effectively on its own. Compared to logistic regression, it achieved slightly higher mean accuracy and F1 scores, making it a strong candidate for the final model.

Proof of Accomplishment

The following plots provide visual evidence for the improvements described above:

The **boxplot** below illustrates how accuracy, F1 score, recall, and precision varied across the ten folds of cross-validation for the logistic regression model. The consistent spread of results across folds demonstrates that our evaluation is not dependent on a single train-test split. This confirms the reliability and generalizability of our model's performance and reflects the impact of the adjustments made, such as class weighting and oversampling.

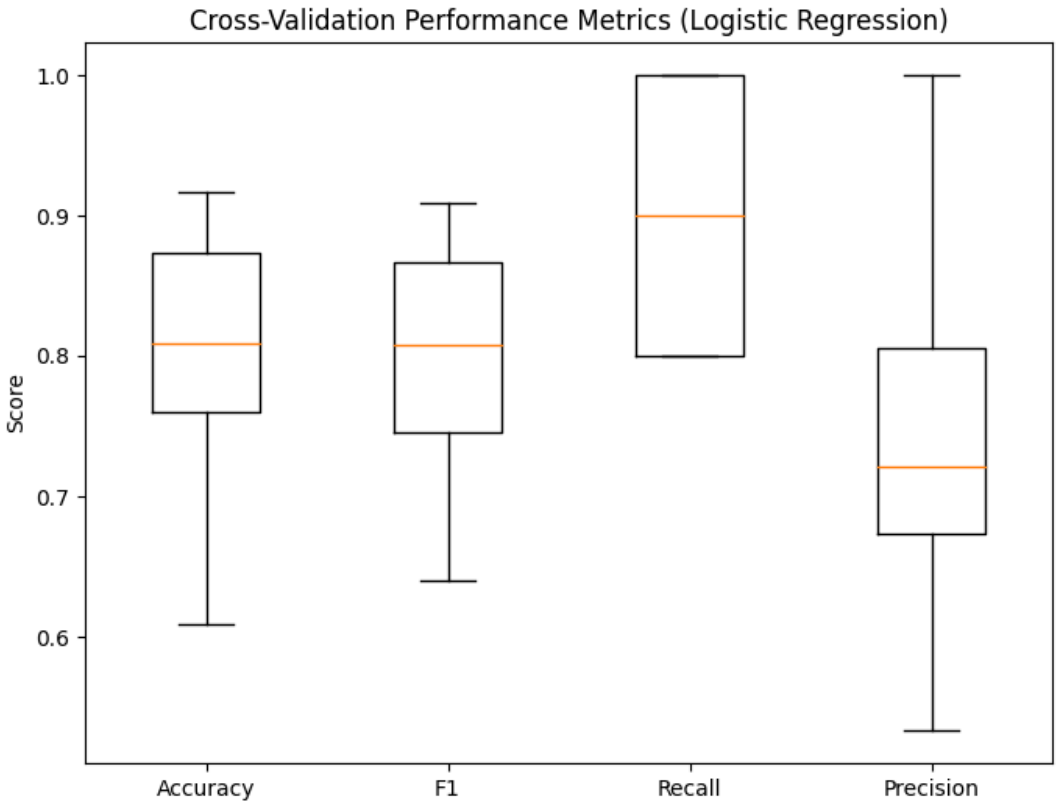


Figure 7: Boxplot showing how accuracy, F1 score, recall, and precision varied across the 10 folds of cross-validation for logistic regression

The **confusion matrix** provides a detailed view of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for the logistic regression model evaluated on a held-out test set. This plot is particularly useful for assessing how class weighting and oversampling affected the model's behavior. Specifically, it shows whether the model is effectively reducing false positives and better identifying the minority (dementia) class, both of which are important in a medical context.

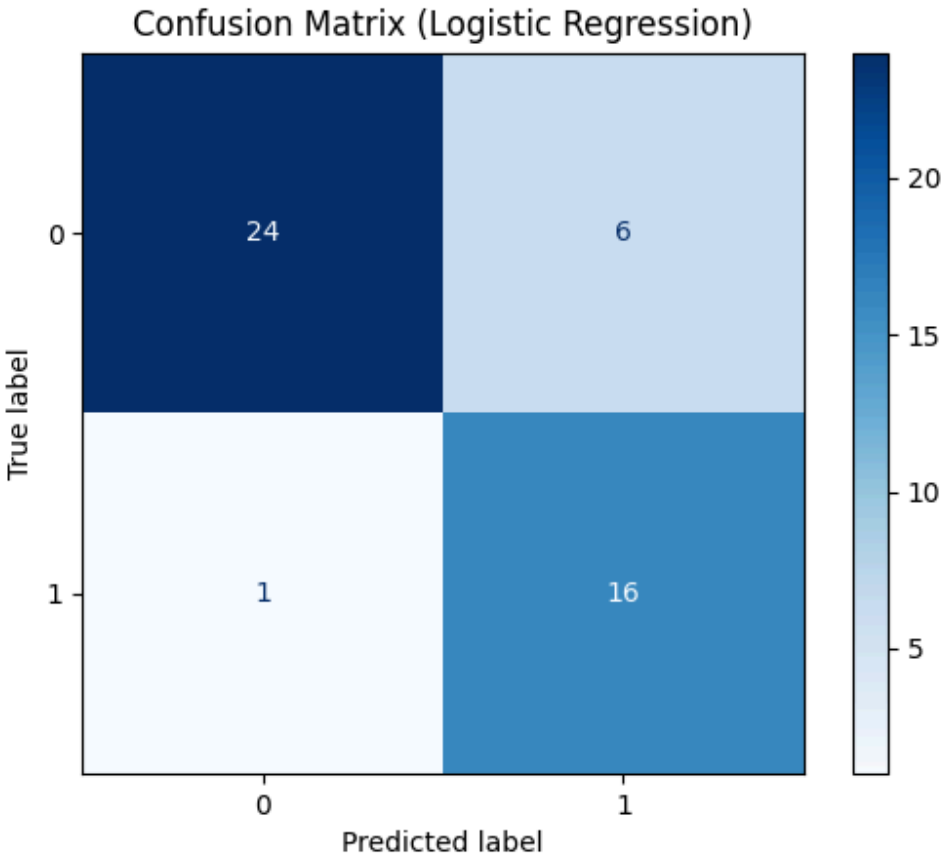


Figure 8: Confusion matrix for logistic regression model, showing true positives, false positives, true negatives, and false negatives for dementia classification.

The **ROC curve** (Receiver Operating Characteristic) captures the model's sensitivity (true positive rate) versus its specificity (1 - false positive rate) across different thresholds. The AUC (Area Under the Curve) quantifies how well the model distinguishes between dementia and non-dementia classes. This visualization is especially valuable in healthcare, where balancing false positives and false negatives can directly impact screening outcomes and patient care. Additionally, it allows us to evaluate how well the different balancing techniques generalized to unseen data.

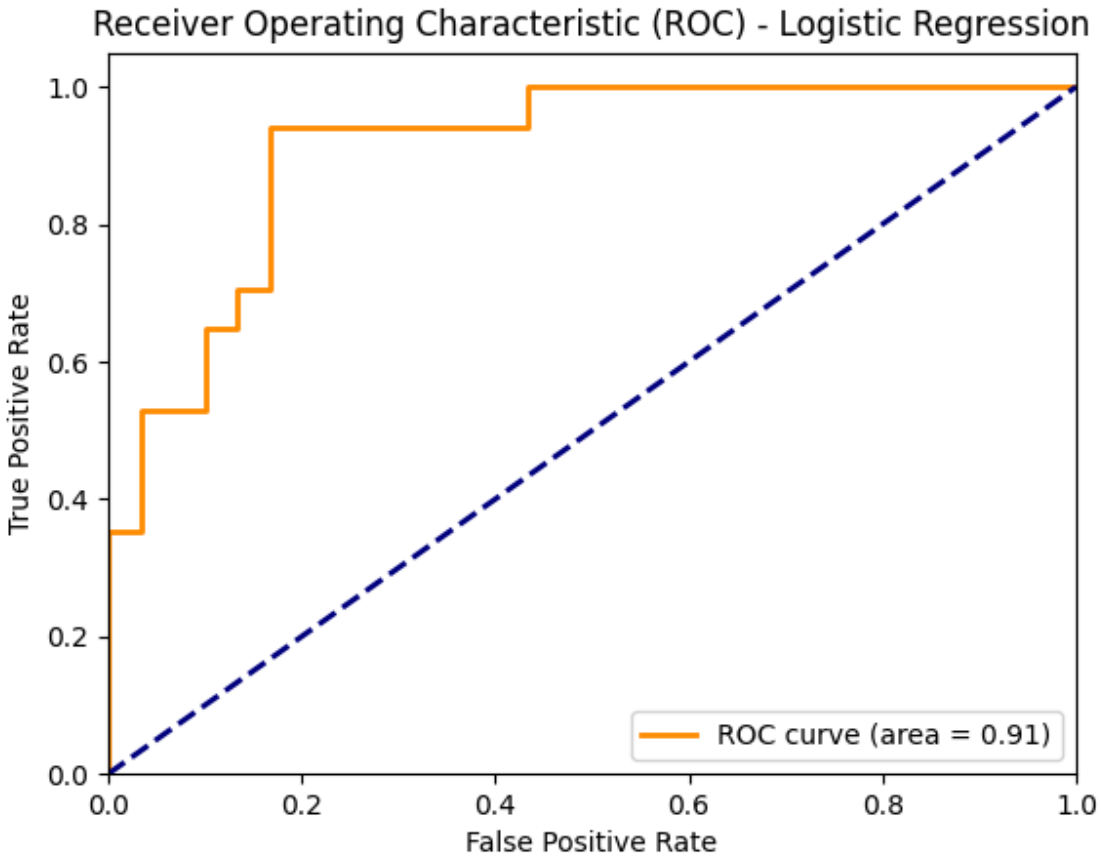


Figure 9: Receiver Operating Characteristic (ROC) curve for logistic regression, with AUC indicating the model's discrimination power between dementia and non-dementia classes.

These plots serve as strong visual evidence for the effectiveness of our model improvements. Together, they demonstrate that the use of class weighting and oversampling has led to:

- A more stable and generalizable evaluation (boxplot),
- A more sensitive classification toward dementia cases (confusion matrix),
- And a better balance between sensitivity and specificity (ROC/AUC).

They confirm the fairness and completeness of our evaluation methodology and provide confidence in the model's ability to aid in early-stage dementia detection.

Challenges or Roadblocks

We encountered the following challenges during this phase of the project:

- **Class Imbalance:** The original dataset had significantly more non-dementia cases, making it difficult for models to correctly identify dementia cases. This issue was particularly noticeable in early logistic regression runs, which showed poor recall.

- **Evaluation Limitations in Milestone 1:** We initially relied on a single train-test split, which may have led to overestimated performance. Switching to 10-fold CV in M2 significantly improved evaluation reliability.

- **Trade-offs in Balancing Methods:** We tested both weighting and oversampling. While both improved fairness, gains varied across metrics, requiring nuanced decisions during model selection.

Changes from Original Plan

- **Expanded Model Scope:** Originally, we planned to use only logistic regression. Based on TA feedback and performance results, we added Random Forest and briefly tested Gradient Boosting, which did not yield successful results and was excluded.

- **Shift to Cross-Validation:** We moved from a single train-test split to **10-fold cross-validation** for all experiments to gain more reliable performance estimates.

- **Balancing Strategy Added:** Addressing class imbalance wasn't part of our original proposal. However, after reviewing the dataset and receiving feedback, we introduced class weighting and oversampling to improve fairness and model sensitivity.