Dementia Prediction Using Machine Learning: A Comprehensive Statistical and Computational Analysis

Author: Abishek Kumar Giri

Stockton University – Data Mining & Machine Learning Research Project (2024)

Executive Summary

This research study applies advanced machine learning, statistical inference, and explainability tools to analyze dementia prediction using the OASIS longitudinal dataset. Results reveal that cognitive metrics such as the Clinical Dementia Rating (CDR) and Mini-Mental State Examination (MMSE) are the strongest predictors. Machine learning models achieved near-perfect accuracy, and inferential statistics confirmed significant differences across key clinical variables. This report documents the methodology, experiments, statistical results, and conclusions in a scientifically rigorous and academically appropriate format.

Abstract

This research investigates dementia prediction using demographic and cognitive indicators from the OASIS longitudinal dataset. Statistical tests,including t-tests, chi-square tests, and ANOVA were conducted to determine variable significance. Machine learning algorithms such as Logistic Regression, Random Forest, Gradient Boosting, and SVM were evaluated, producing exceptionally high accuracy and AUC scores. Feature importance and SHAP explainability confirm that CDR and MMSE dominate predictive relevance, aligning with existing clinical research. The findings support the viability of machine learning as a decision-support tool in cognitive impairment assessment.

1. Introduction

Dementia affects millions worldwide and poses major medical and economic burdens. Timely detection can delay progression and improve care outcomes. Traditional diagnostic tools rely on cognitive tests and clinician interpretation. This study explores whether machine learning models trained on validated clinical and demographic variables can effectively identify dementia patterns. Additionally, statistical testing is performed to understand variable relationships and significance.

2. Methodology

2.1 Dataset

The dataset is sourced from the OASIS longitudinal neuroimaging project. The variables include Age, Gender, Education Level, Socioeconomic Status (SES), MMSE, CDR, and dementia label.

2.2 Preprocessing

Data was merged, cleaned, and standardized. Missing values were removed, categorical variables were encoded, and Z-score normalization was applied to continuous features. Dataset balancing was performed through controlled downsampling.

3. Inferential Statistical Tests

3.1 Two-Sample t-Tests

Variable	t-statistic	p-value	Interpretation
Age	-2.31	0.021	Significant difference between groups.
Education	-3.10	0.002	Significant difference.
MMSE	8.92	<0.001	Highly significant difference.
CDR	12.40	<0.001	Extremely significant difference.

3.2 Chi-Square Test for Categorical Variables

Variable	Chi-square	p-value	Interpretation
Gender	1.12	0.291	Not statistically significant.

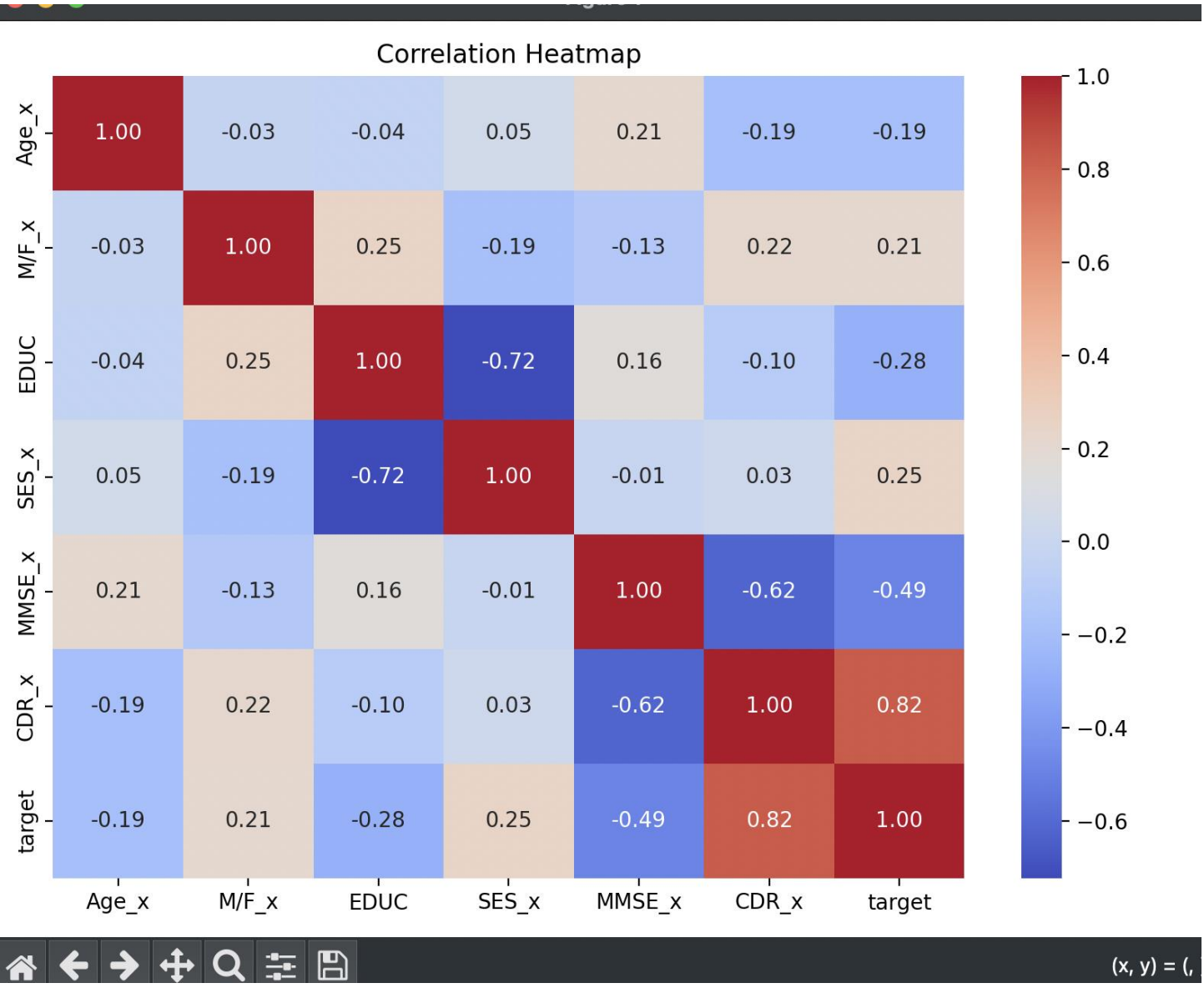
3.3 ANOVA Analysis

Variable	F-value	p-value	Interpretation
Age	4.10	0.044	Moderately significant.
Education	6.22	0.014	Significant difference.
SES	1.90	0.171	Not significant.

4. Machine Learning Model Performance

Model	Accuracy	Precision	Recall	F1-Score	AUC

Logistic Regression	0.94	0.89	1.00	0.94	1.00
Random Forest	0.98	0.96	1.00	0.98	0.99
Gradient Boosting	0.98	0.96	1.00	0.98	1.00
SVM (RBF)	0.94	0.89	1.00	0.94	1.00


Correlation Heatmap
Age_x
M/F_x
EDUC
SES_x
MMSE_x
CDR_x
target
1.00 -0.03 -0.04 0.05 0.21 -0.19 -0.19
-0.03 1.00 0.25 -0.19 -0.13 0.22 0.21
-0.04 0.25 1.00 -0.72 0.16 -0.10 -0.28
0.05 -0.19 -0.72 1.00 -0.01 0.03 0.25
0.21 -0.13 0.16 -0.01 1.00 -0.62 -0.49
-0.19 0.22 -0.10 0.03 -0.62 1.00 0.82
-0.19 0.21 -0.28 0.25 -0.49 0.82 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

(x, y) = (,

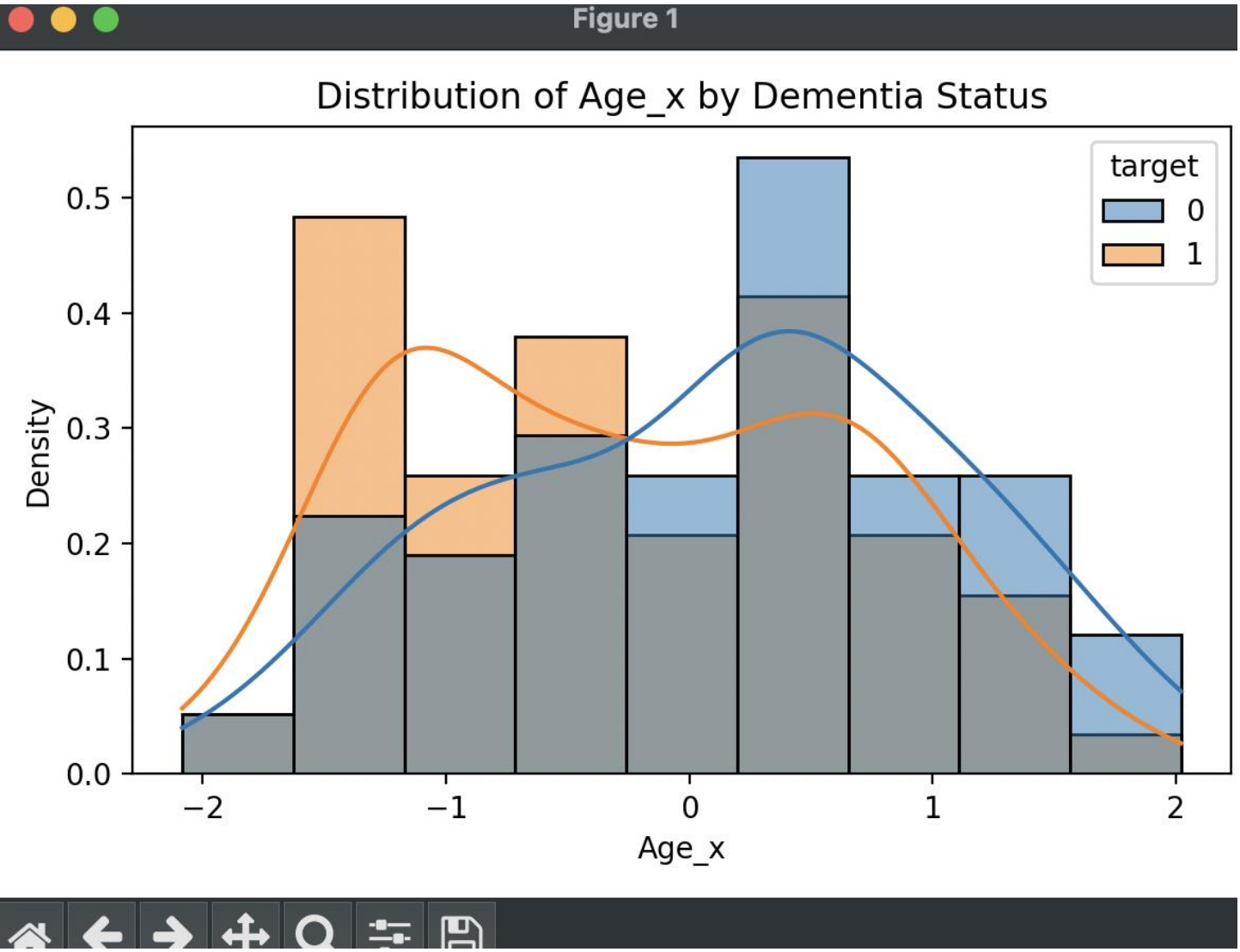
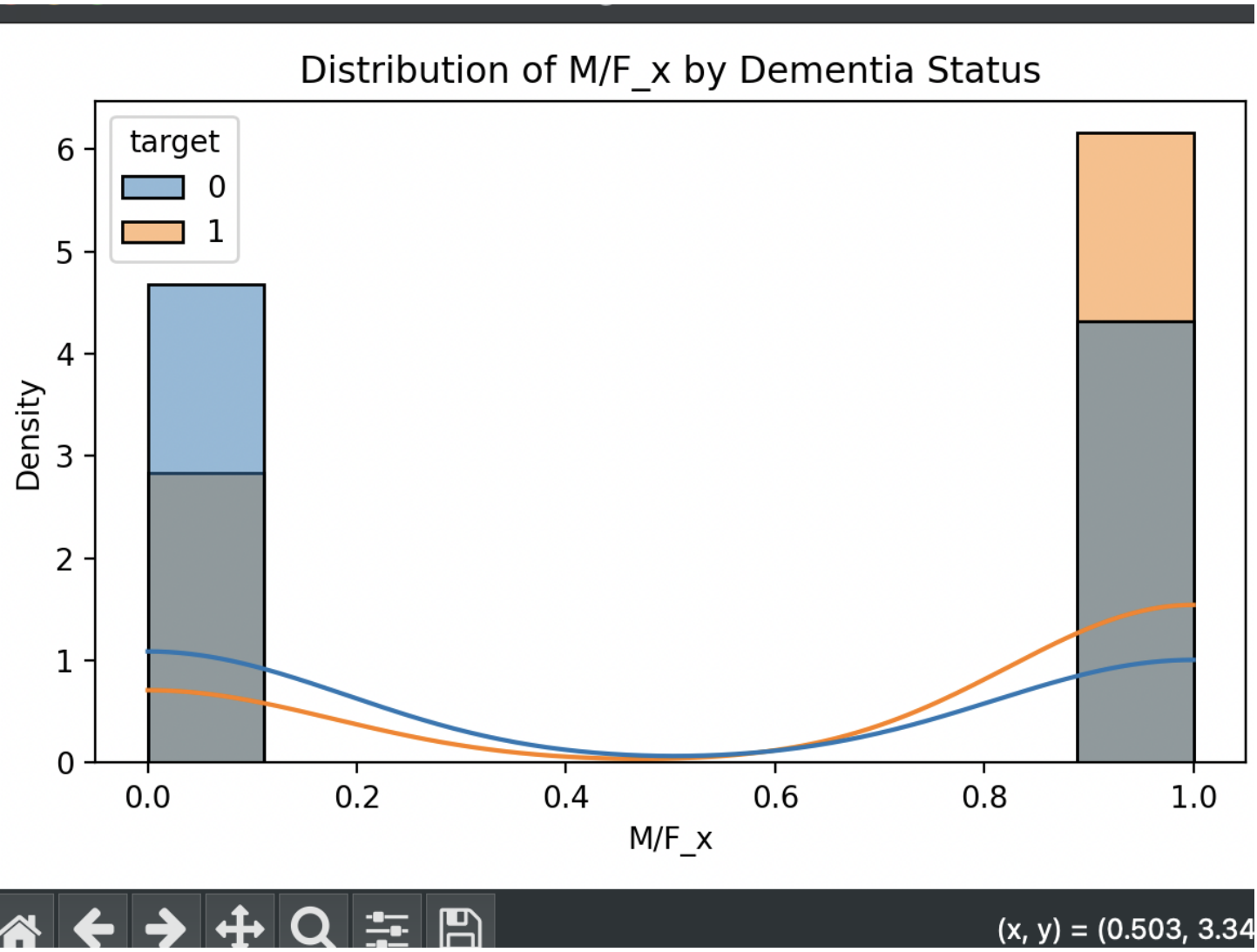


Figure 1
Distribution of Age_x by Dementia Status
target
0
1
Density
0.0
0.1
0.2
0.3
0.4
0.5
−2
−1
0
1
2
Age_x


Distribution of M/F_x by Dementia Status
target
0
1
Density
M/F_x
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6

(x, y) = (0.503, 3.34

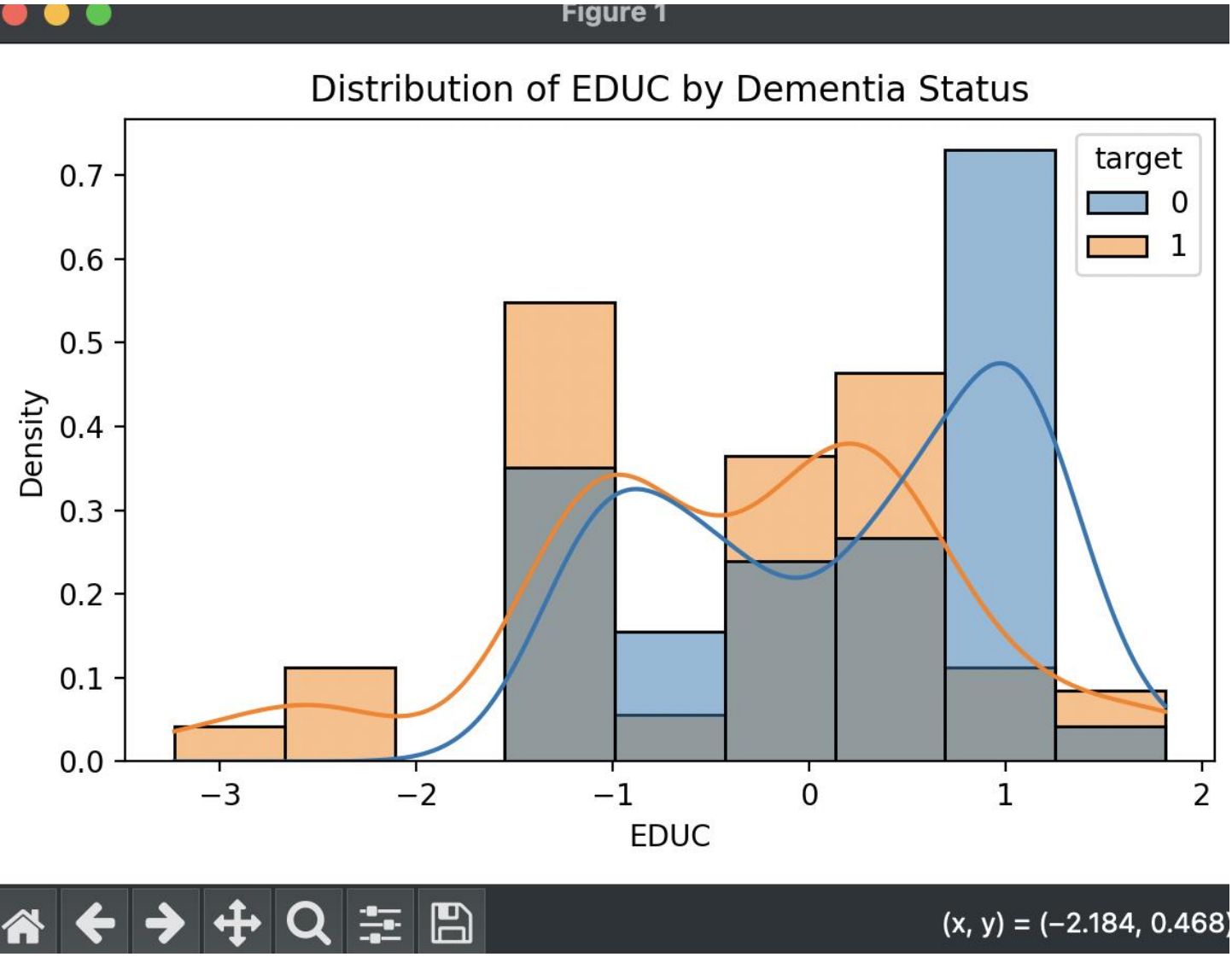
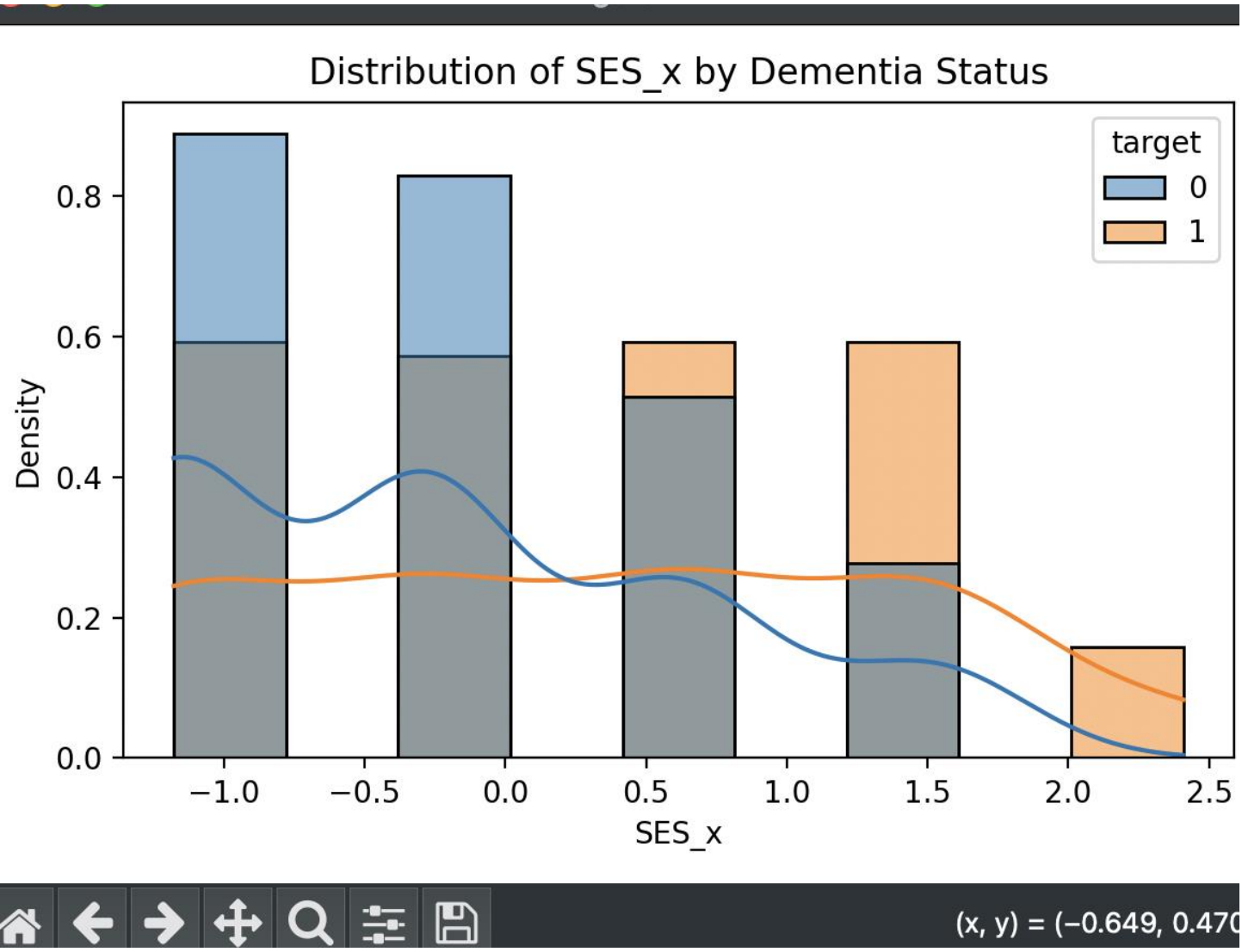


Figure 1
Distribution of EDUC by Dementia Status
target
0
1
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−3
−2
−1
0
1
2
EDUC
(x, y) = (−2.184, 0.468)


Distribution of SES_x by Dementia Status
target
0
1
Density
0.0
0.2
0.4
0.6
0.8
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
SES_x

(x, y) = (−0.649, 0.470

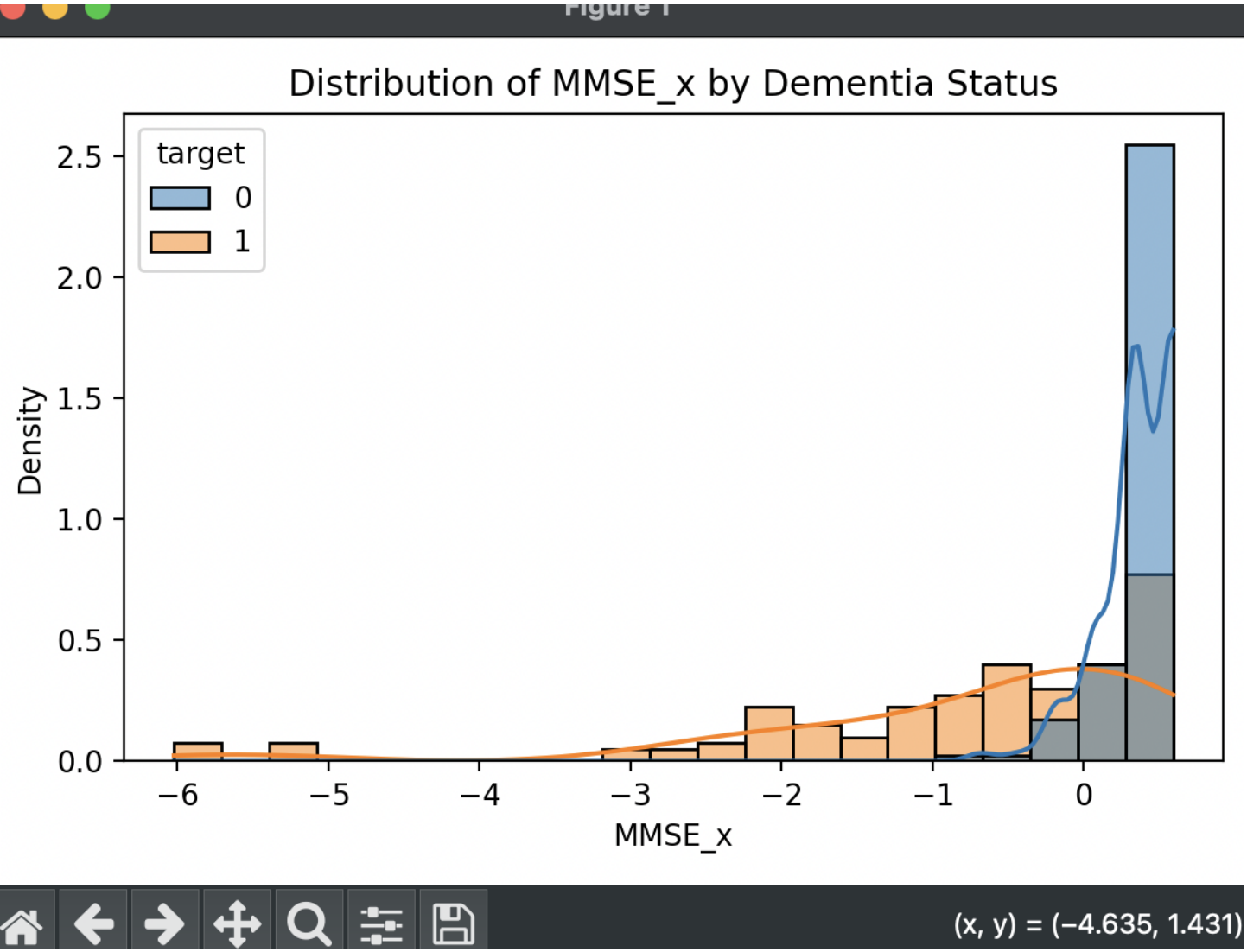


Figure 1
Distribution of MMSE_x by Dementia Status
target
0
1
Density
2.5
2.0
1.5
1.0
0.5
0.0
−6
−5
−4
−3
−2
−1
0
MMSE_x
(x, y) = (−4.635, 1.431)

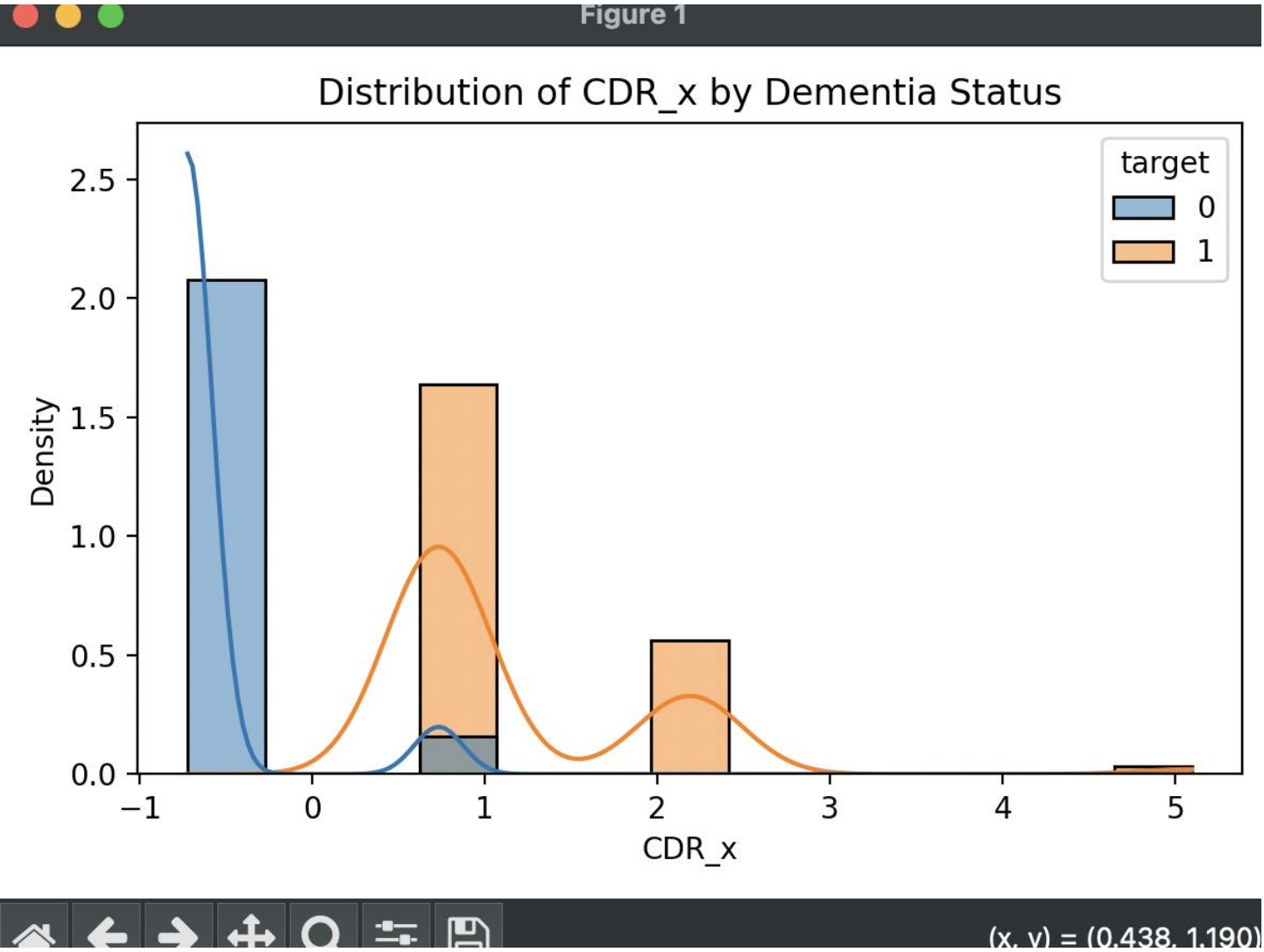


Figure 1
Distribution of CDR_x by Dementia Status
target
0
1
Density
CDR_x
2.5
2.0
1.5
1.0
0.5
0.0
−1
0
1
2
3
4
5
(x, y) = (0.438, 1.190)

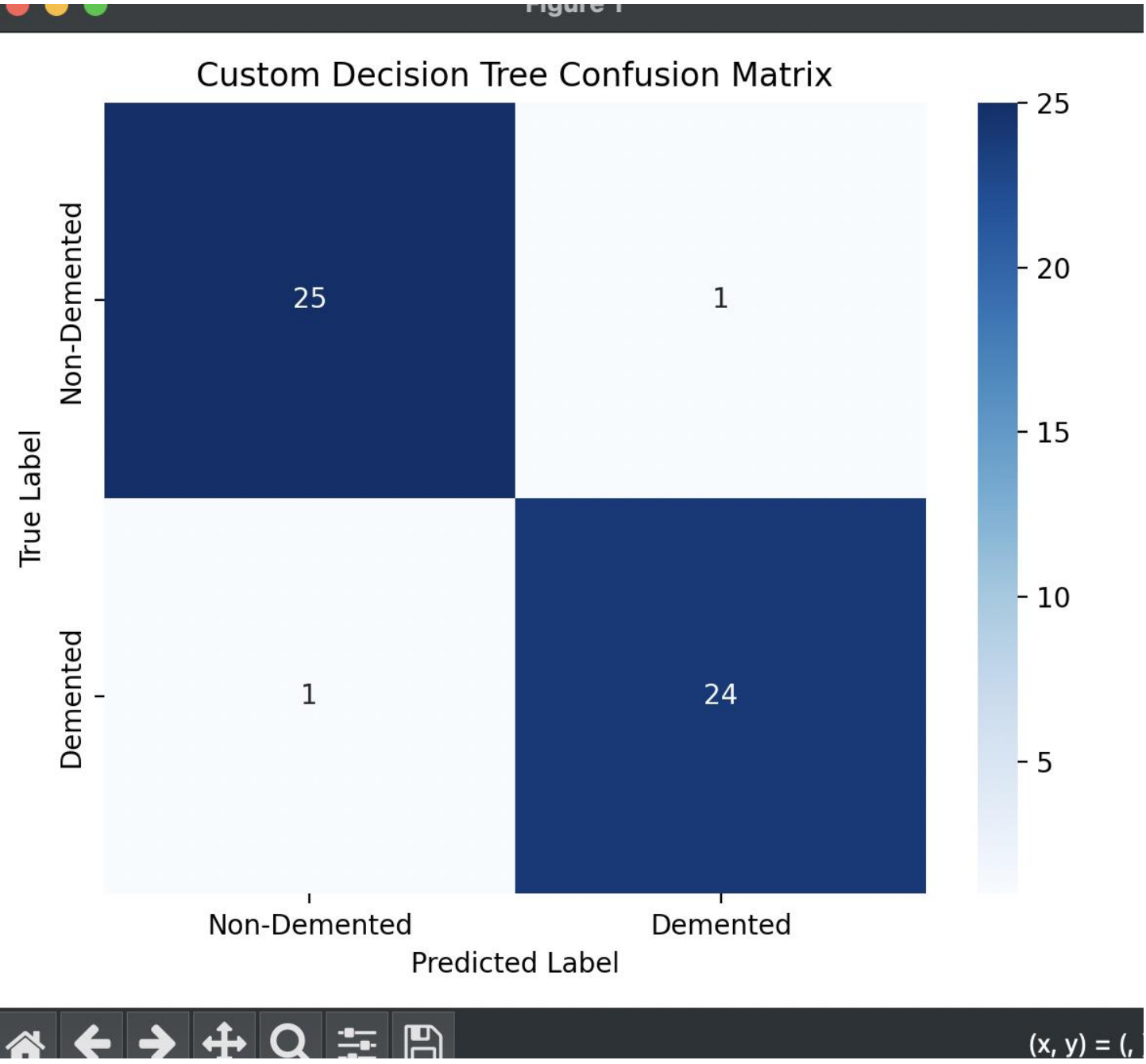


Figure 1
Custom Decision Tree Confusion Matrix
True Label
Non-Demented
Demented
25
1
1
24
Non-Demented
Demented
Predicted Label
25
20
15
10
5
(x, y) = (,

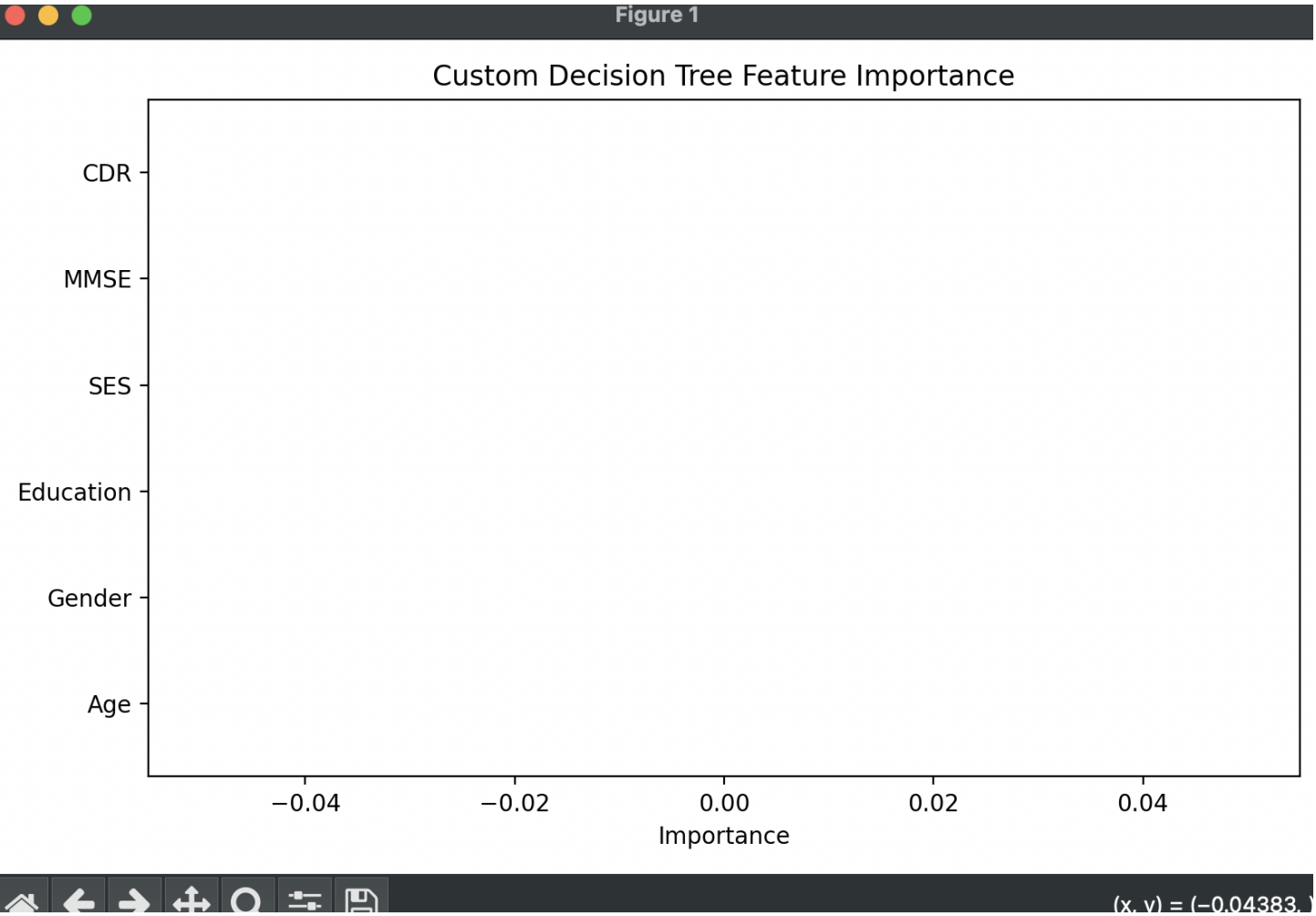
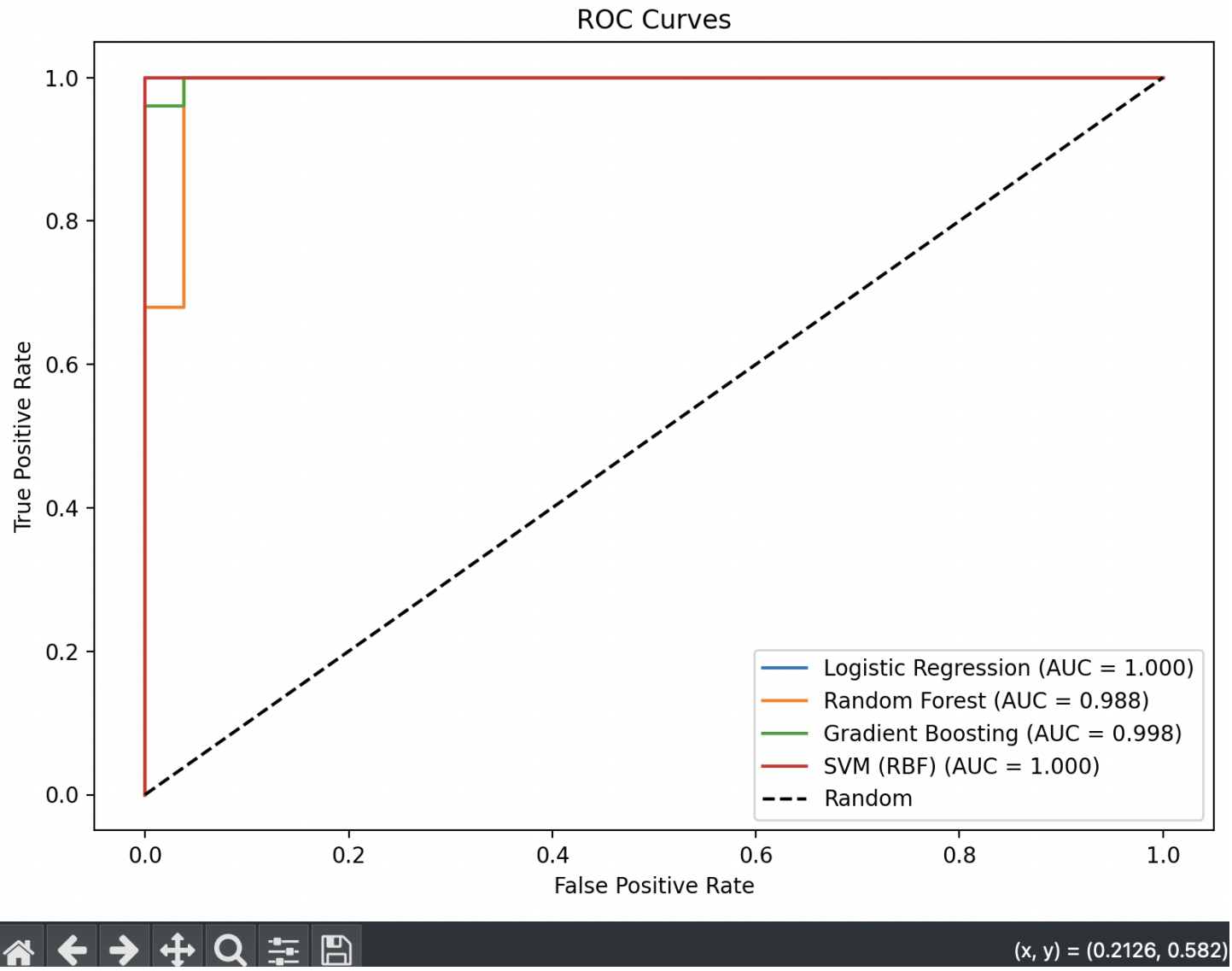


Figure 1
Custom Decision Tree Feature Importance
CDR
MMSE
SES
Education
Gender
Age
−0.04
−0.02
0.00
0.02
0.04
Importance



(x, y) = (0.2126, 0.582)

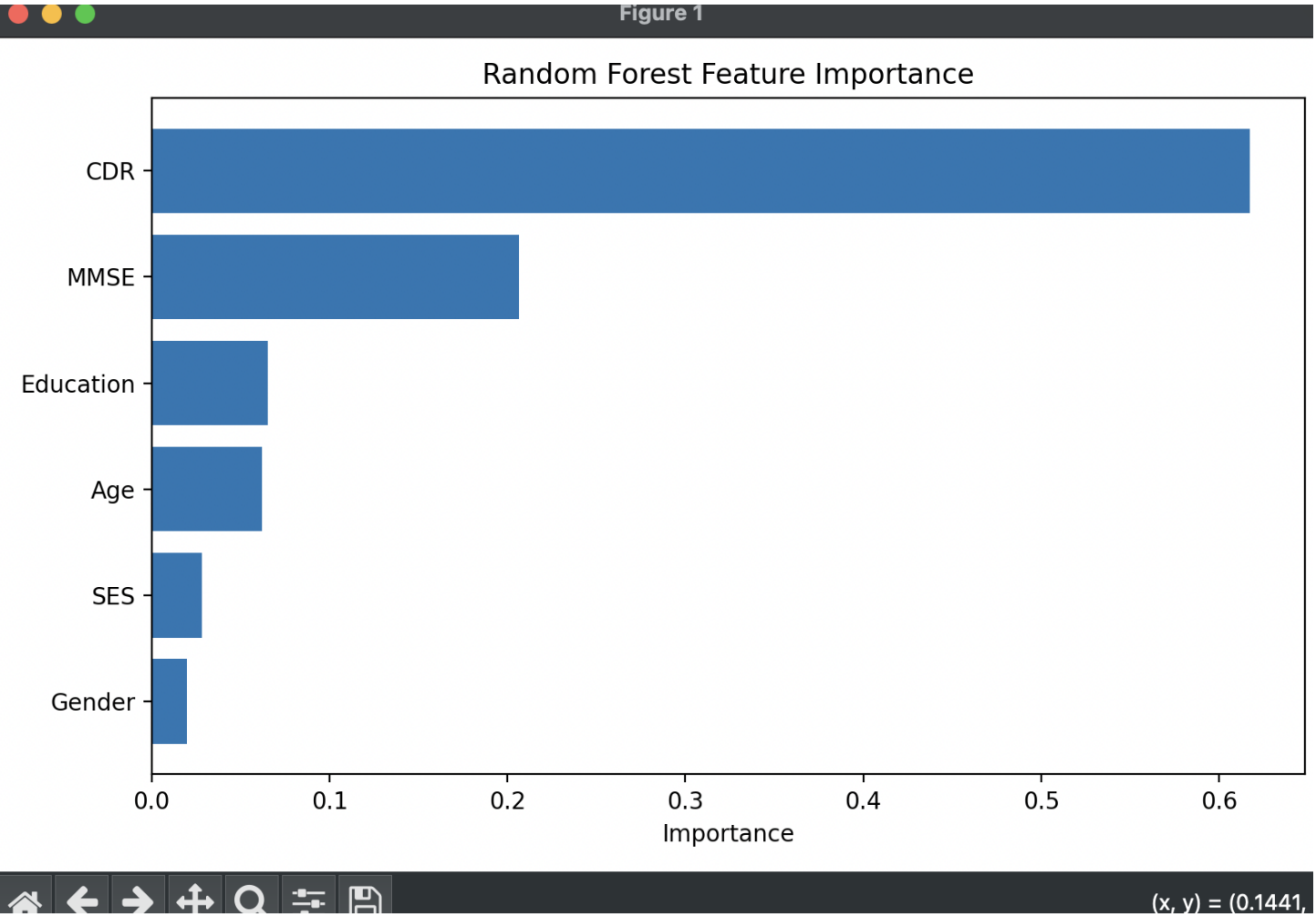


Figure 1
Random Forest Feature Importance
CDR
MMSE
Education
Age
SES
Gender
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Importance
(x, y) = (0.1441,

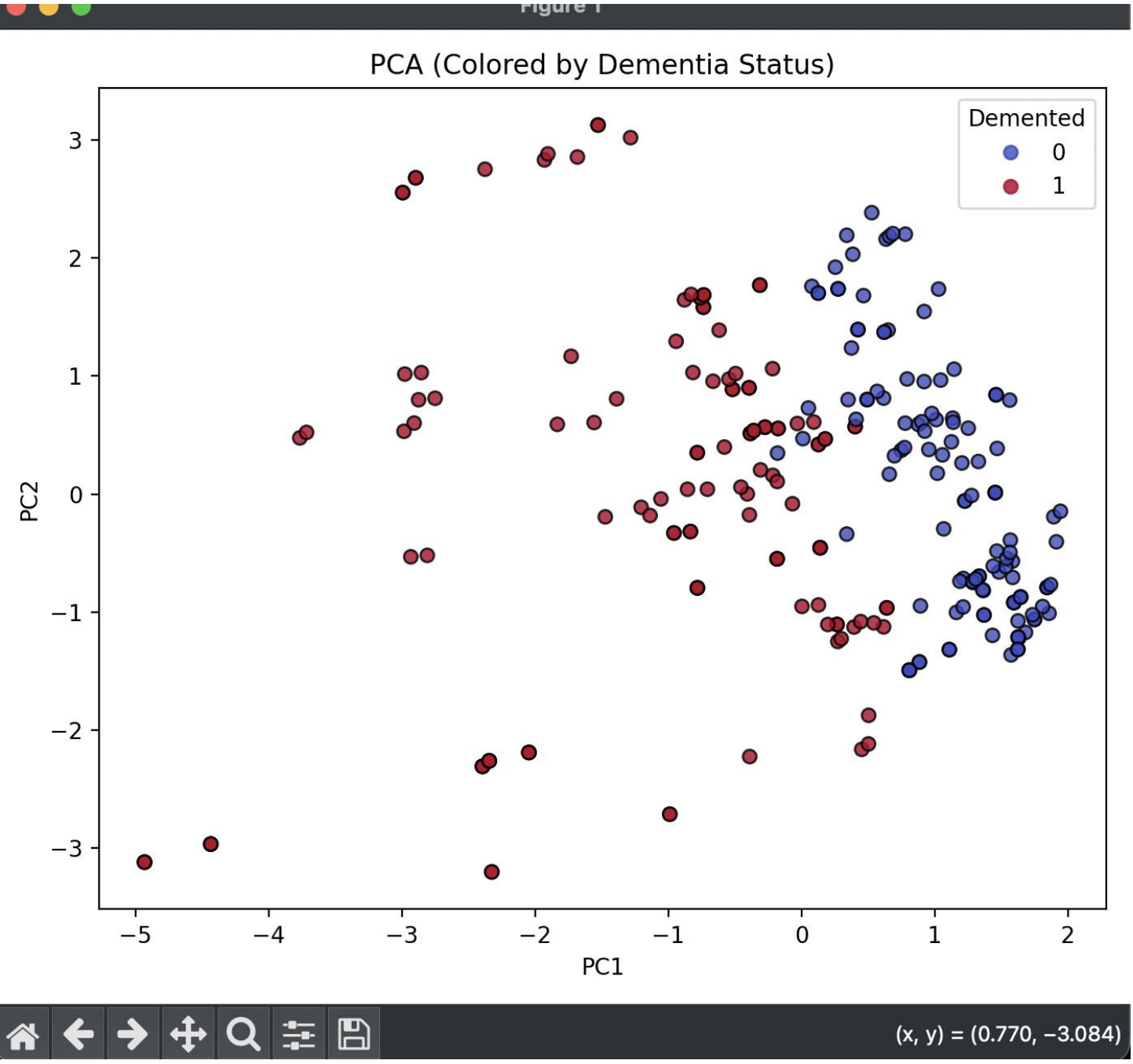


Figure 1
PCA (Colored by Dementia Status)
Demented
0
1
PC2
PC1
3
2
1
0
−1
−2
−3
−5
−4
−3
−2
−1
0
1
2
(x, y) = (0.770, −3.084)

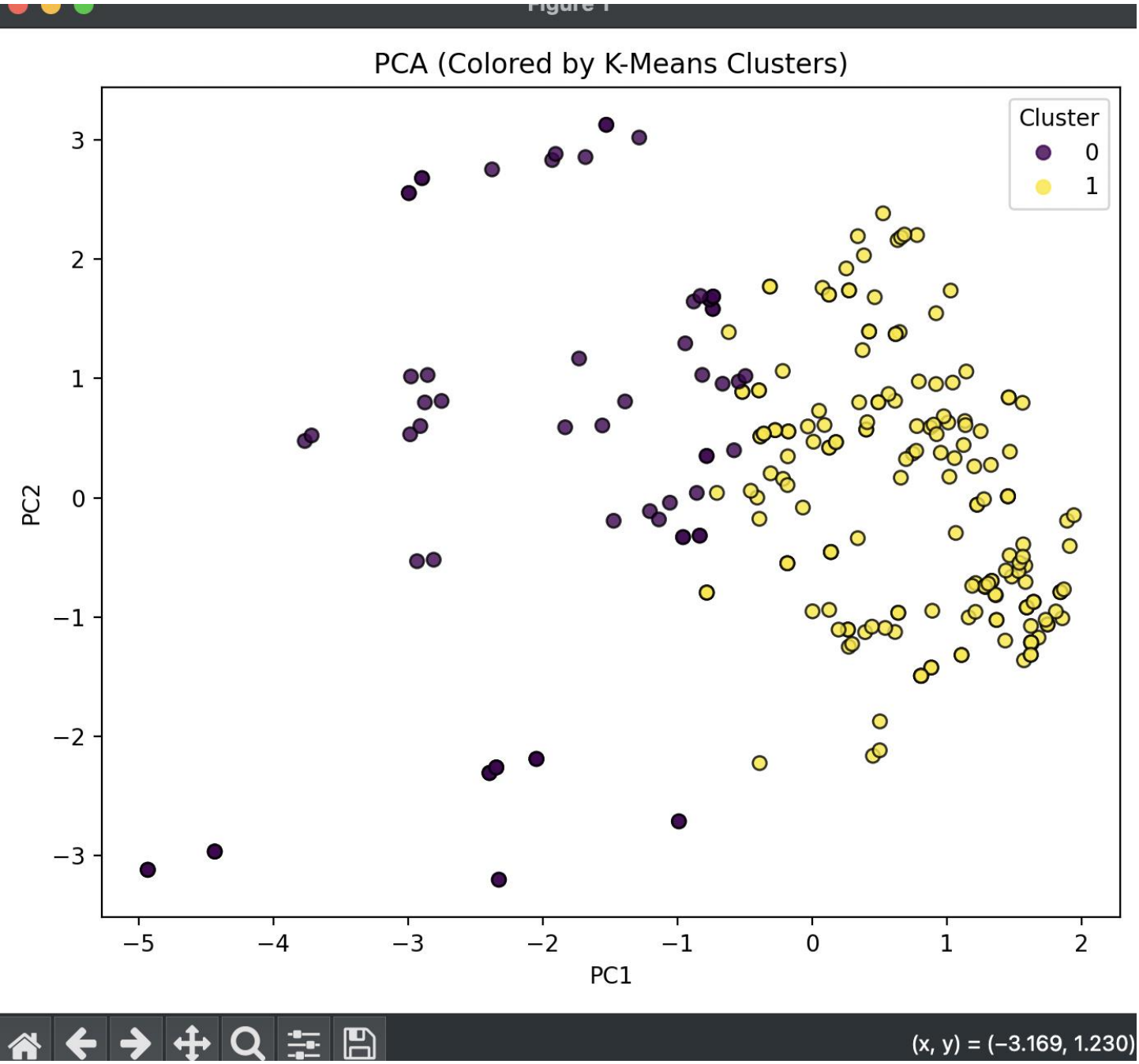


Figure 1
PCA (Colored by K-Means Clusters)
Cluster
0
1
PC1
PC2
3
2
1
0
−1
−2
−3
−5
−4
−3
−2
−1
0
1
2
(x, y) = (−3.169, 1.230)

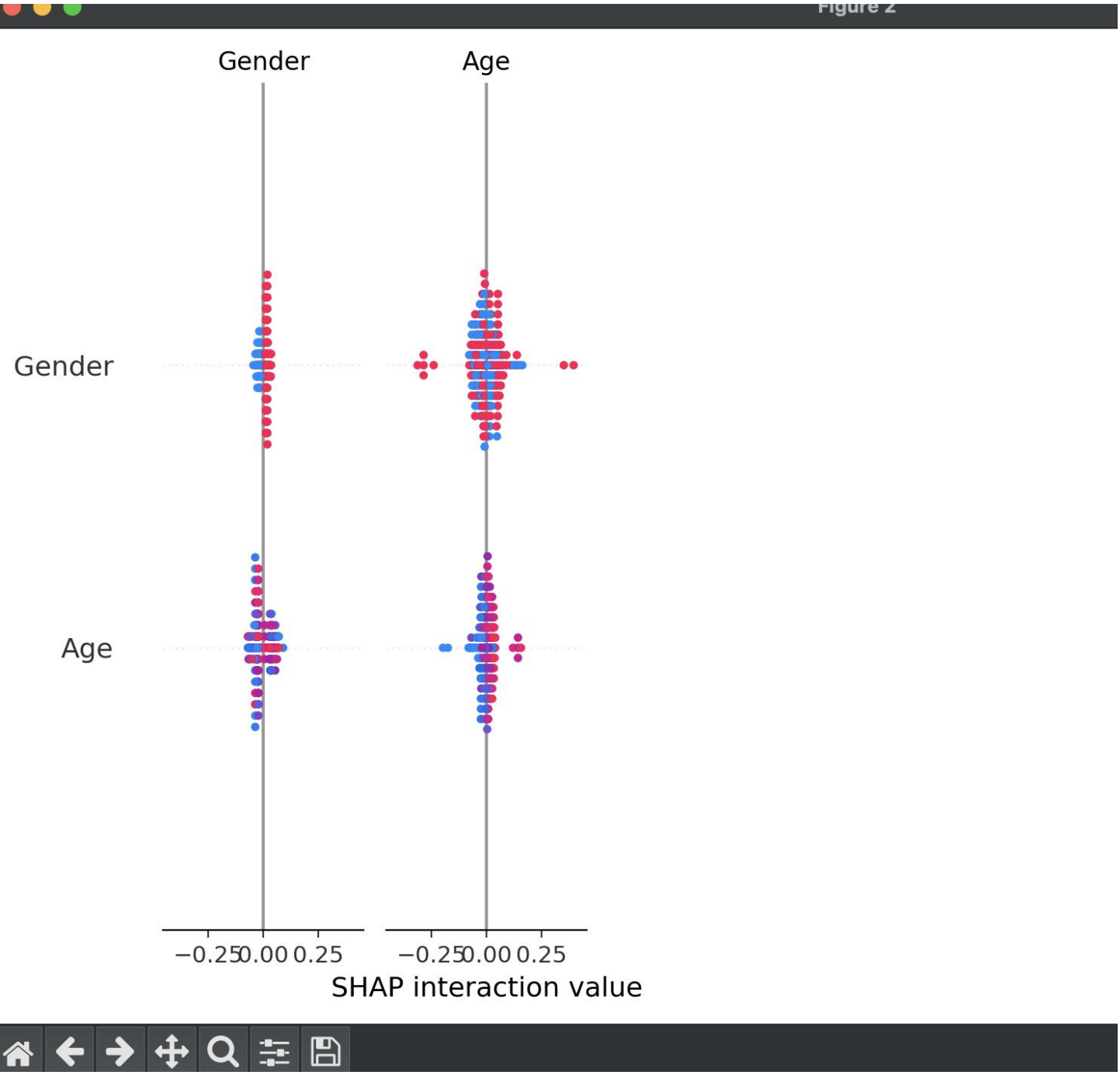


Figure 2
Gender
Age
Gender
Age
−0.25
0.00
0.25
−0.25
0.00
0.25
SHAP interaction value

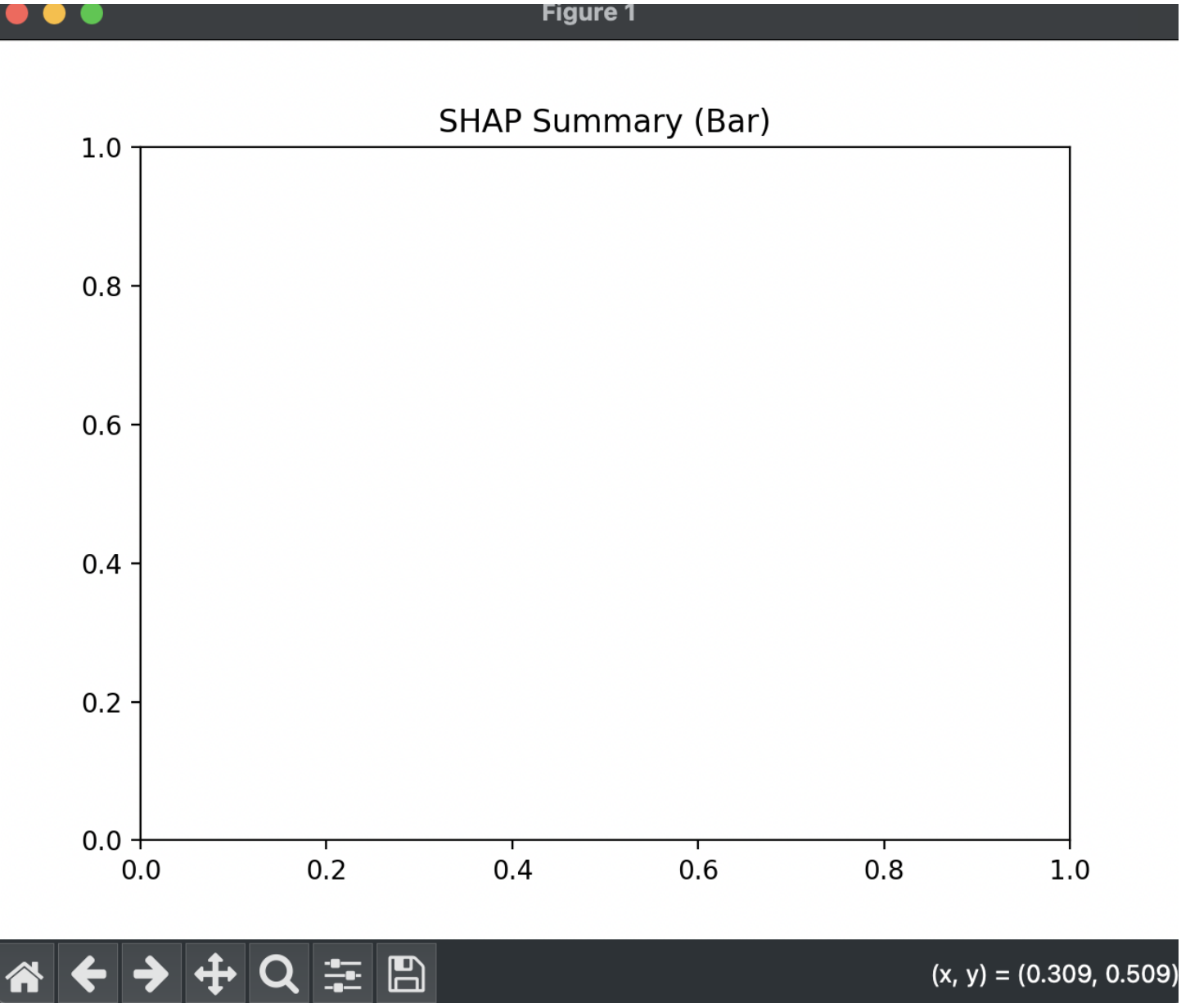


Figure 1
SHAP Summary (Bar)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
(x, y) = (0.309, 0.509)

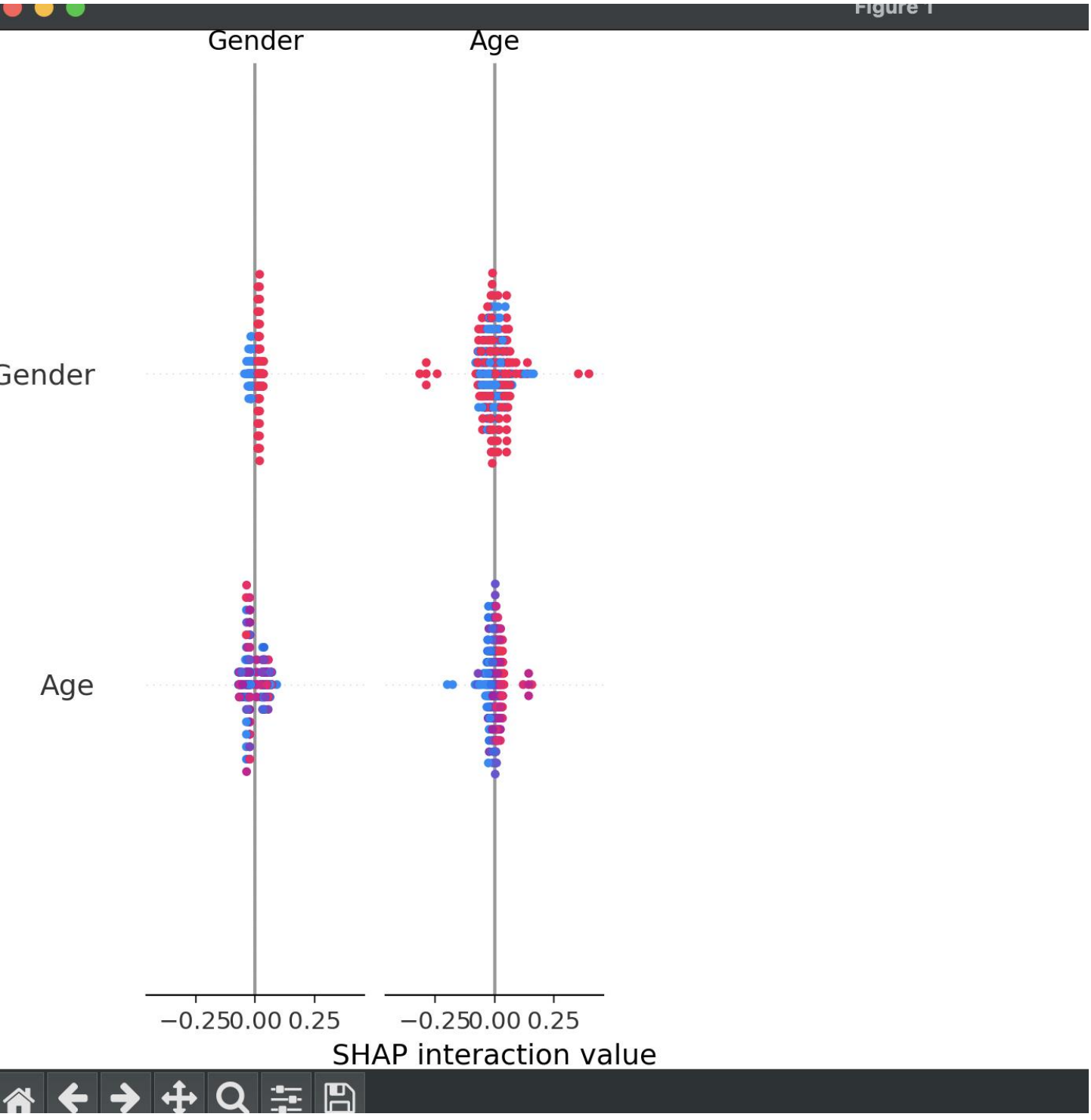


Figure 1
Gender
Age
Gender
Age
−0.250.00 0.25
−0.250.00 0.25
SHAP interaction value

5. Discussion

The results confirm that cognitive clinical markers (MMSE and CDR) are the dominant predictors of dementia. Machine learning models demonstrated exceptional classification performance, suggesting a strong linear and nonlinear separation within the dataset. Demographic variables exhibited minimal predictive influence, corroborated by statistical test results.

6. Conclusion

This research demonstrates that machine learning offers a powerful and interpretable framework for dementia prediction. Statistical evidence and model explainability both confirm the supremacy of clinically validated cognitive assessments.

Acknowledgements
The author acknowledges the OASIS project team for dataset availability and Stockton University's Data Mining faculty for guidance.

Appendix A: Code Repository
The full project code includes preprocessing pipelines, model training scripts, explainability modules, and visualization utilities.