Predicting ADHD from fMRI Connectomes & Socio-Demographics

Project Outline

Background

ADHD classification enables earlier intervention

Data & Methods

Multi-modal data with PCA for connectomes

Goals

Identify best model and key correlating factors

CATEGORICAL Dataset 1: Overview

Participant Identification

Unique ID, enrollment year, study site

Demographics

Ethnicity, race, MRI scan location

Parental Background

Education and occupation data

Key Variables in ADHD Prediction Study

Participant Identification

Unique ID, enrollment year, study site

participant_id, Basic_Demos_Enroll_Year, Basic_Demos_Study_Site,

Demographics

Ethnicity, race

PreInt_Demos_Fam_Child_Ethnicity, PreInt_Demos_Fam_Child_Child_Race

Parental Background

Education, occupation

Barratt_Barratt_P1_Edu,

Barratt_Barratt_P1_Occ,

""P2_Edu, ""P2_Occ.

MRI Scan Location

Location of MRI scan

MRI_Track_Scan_Location

QUANTITATIVE Dataset 2 : Overview

Emotional Health

EHQ total score and ColorVision test

Alabama Parenting

Measures parenting style across multiple dimensions

Strengths & Difficulties

Behavioral profiles with composite scores

Age at Scan

Controls for developmental stage

Key Variables in ADHD Prediction Study

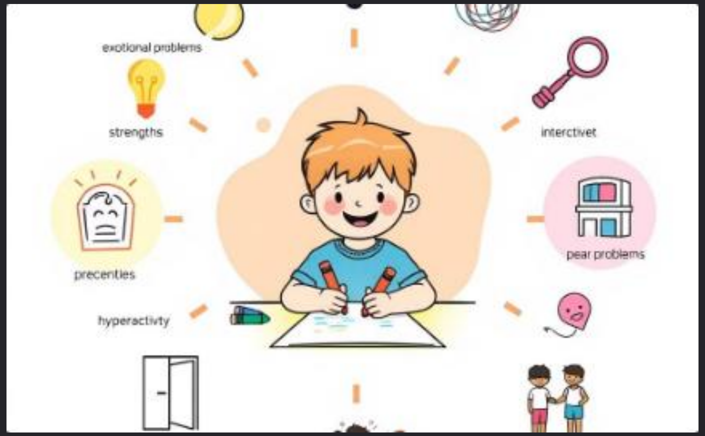
Participant ID

Unique identifier for each participant.

EHQ Total Score

Total score on the Emotional Health Questionnaire.(for 100)

EHQ_EHQ_Total,



SDQ Scores

Strengths and Difficulties Questionnaire. Respective Students Score

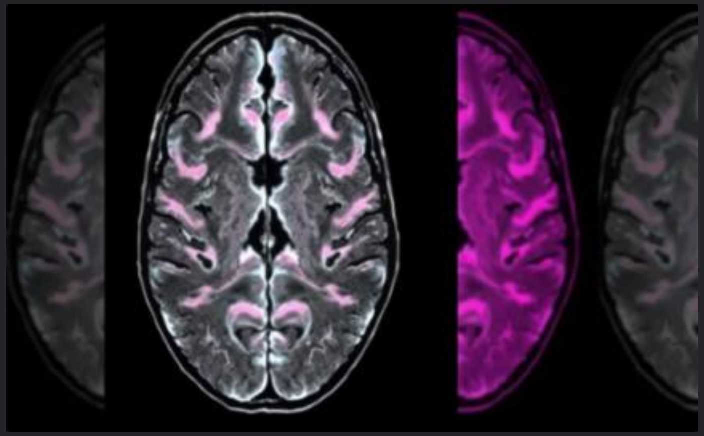
SDQ_SDQ_Conduct Problems, "_Difficulties_Total, "_Emotional_Problems, "Externalizing, "Generating_Impact, "Hyperactivity, "Internalizing, "PeerProblems, "Prosocial

APQ Child Problems

Alabama Parental Questionnaire. Parent-reported Child Problems subscale score.

APQ_P_APQ_P_CP, "_ID,

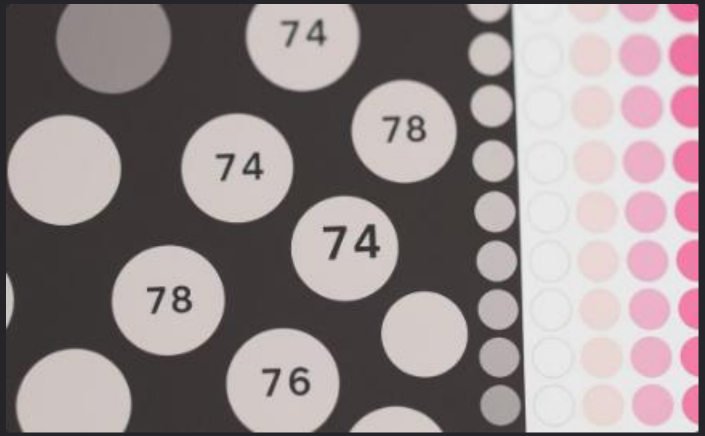
"_INV, "_OPD, "PM, "PP,



Age at Scan

Participant's age at the time of the MRI scan.

MRI_Track_Age_at_Scan



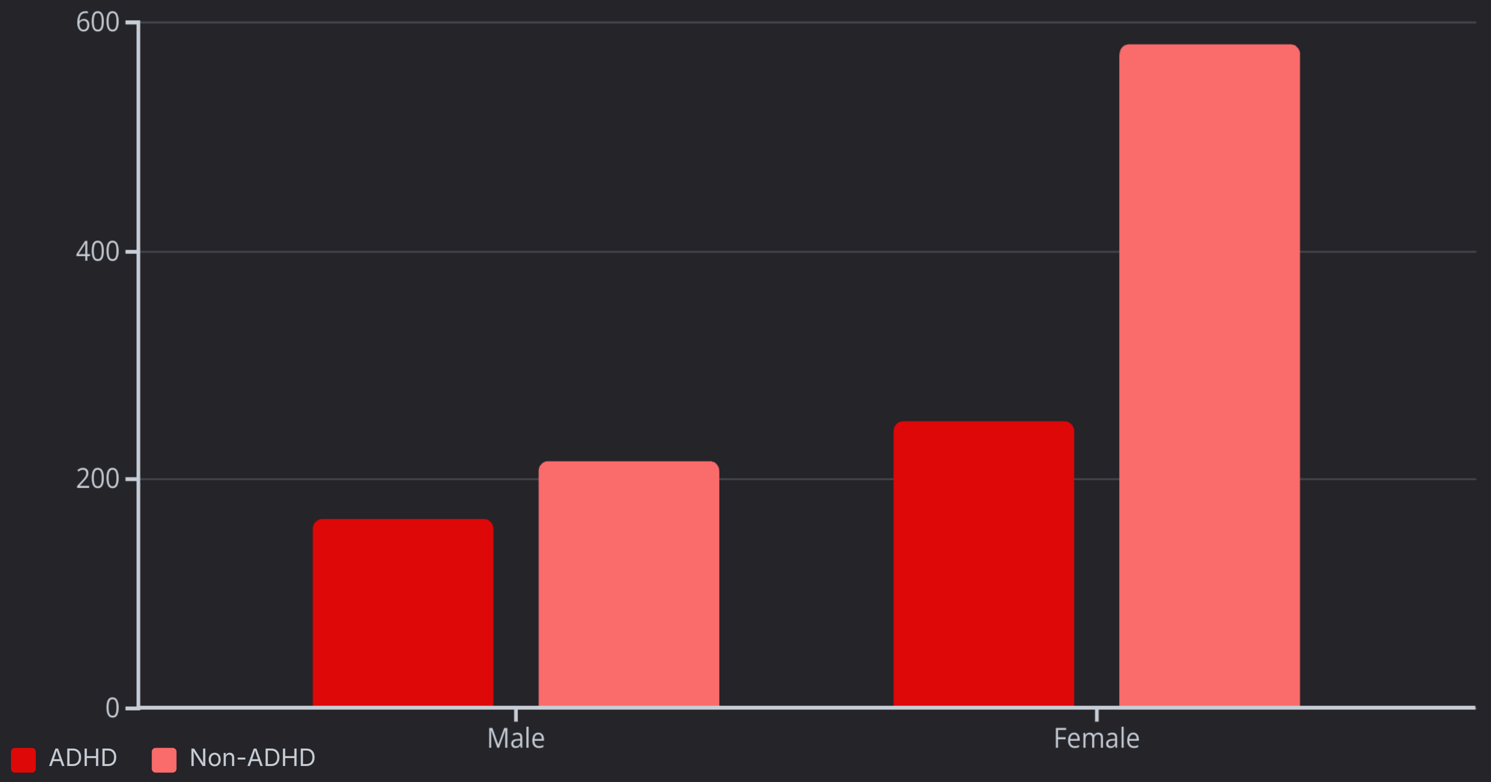
Color Vision Score

Score achieved on a color vision test. (for 14)

ColorVision_CV_Score

Basic EDA for responses:

Sex & ADHD Distribution

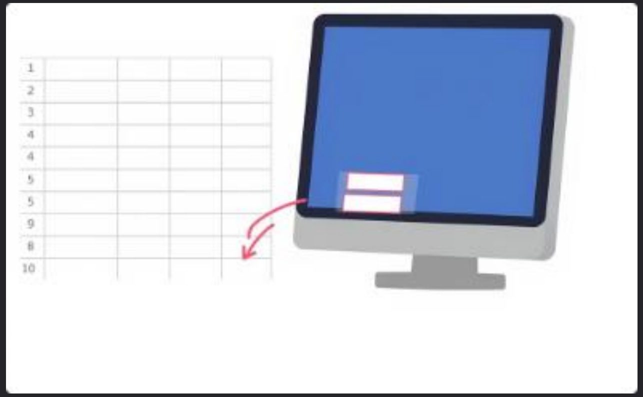


Males more prone to ADHD.
Imbalanced data distribution for female so, the aim of this is to explain about the "ADHD in Female".
For the purpose of this presentation ADHD_Outcome as **sole target variable.**

Data Cleaning & Imputation

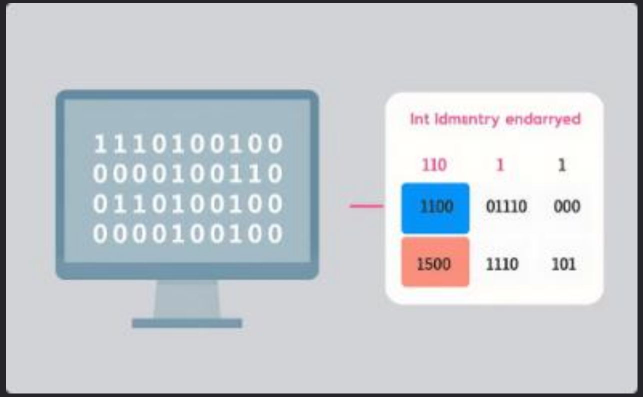
Identifying Missing Values

Checked variables for missing entries in demographic fields and fMRI metadata.



Median Imputation

Imputed missing numeric values using the median of the feature to handle outliers.



Dummy Encoding

Created dummy variables for categorical data to allow algorithms to handle them.

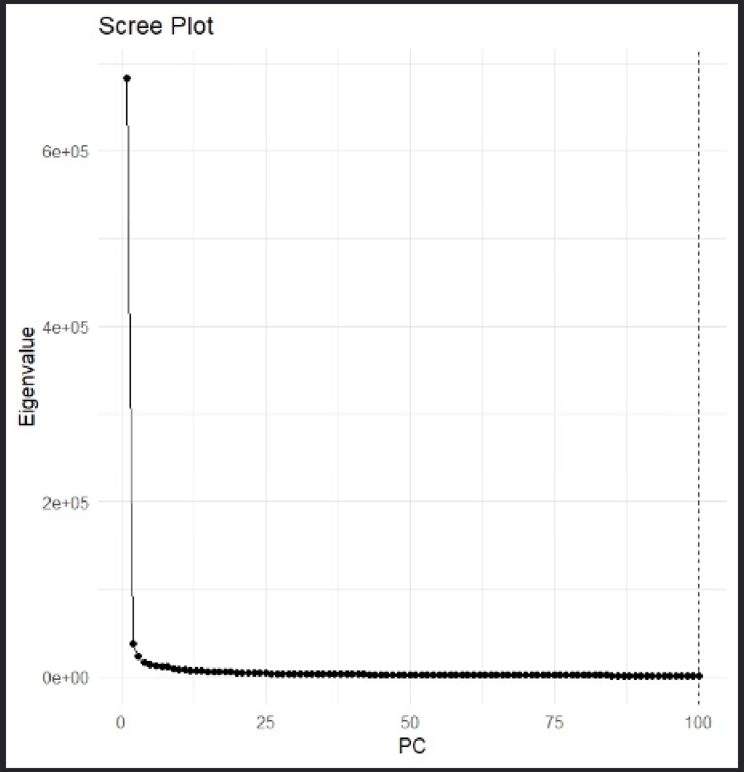
Data Integrity Checks

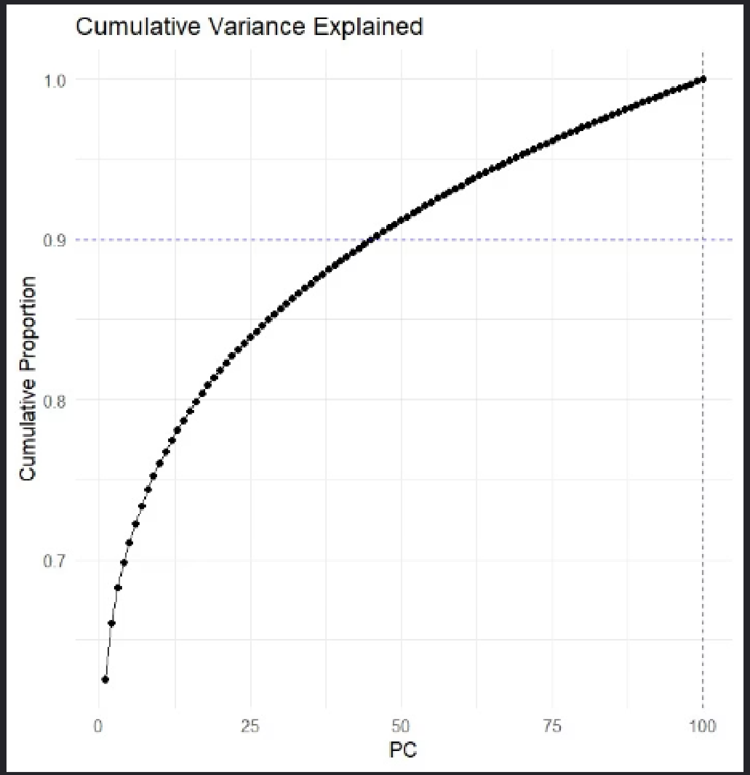
Verified no new missing values were introduced and participant data aligned correctly.

This process ensures a consistent and complete dataset, ready for dimensionality reduction and modeling.

PCA Dimensionality Reduction:

Raw fMRI connectomes contained ~19,900 edges per participant. Principal Component Analysis (PCA) reduced dimensionality to 60 components, informed by scree plot and cumulative variance.





1. **Rationale**: PCA reduces overfitting and improves efficiency by extracting components capturing maximum variance.
2. **Implementation**: Retained 60 principal components, explaining >90% cumulative variance.
3. **Outcome**: Reduced feature set to 60 PCA-based features, merged with cleaned demographic and questionnaire data.

This dimensionality reduction step was crucial for managing fMRI data complexity.

Model Construction

Response

ADHD_Outcome (binary)

Dataset1 + Dataset2 + (Dataset Connectomes 19000 → 60)

Preprocessing

Imputation, encoding, scaling

Train/Test

70/30 split, stratified

Classification Methods

K-Nearest Neighbors

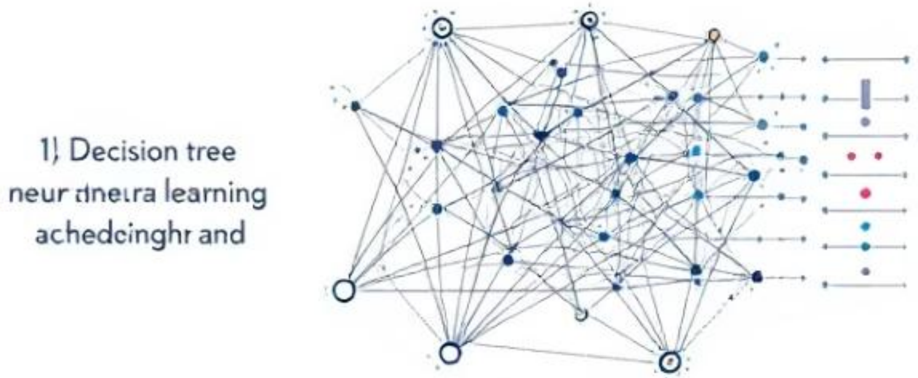
5-fold CV determined k=17

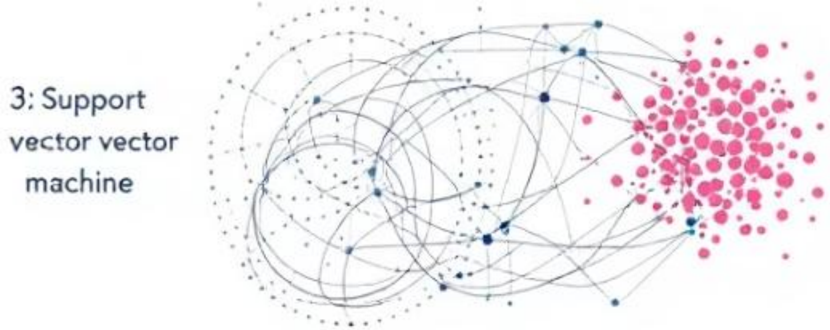
Logistic Regression

Linear approach with interpretable coefficients

Random Forest

Ensemble method robust to interactions





Results: Accuracy, F1, AUC

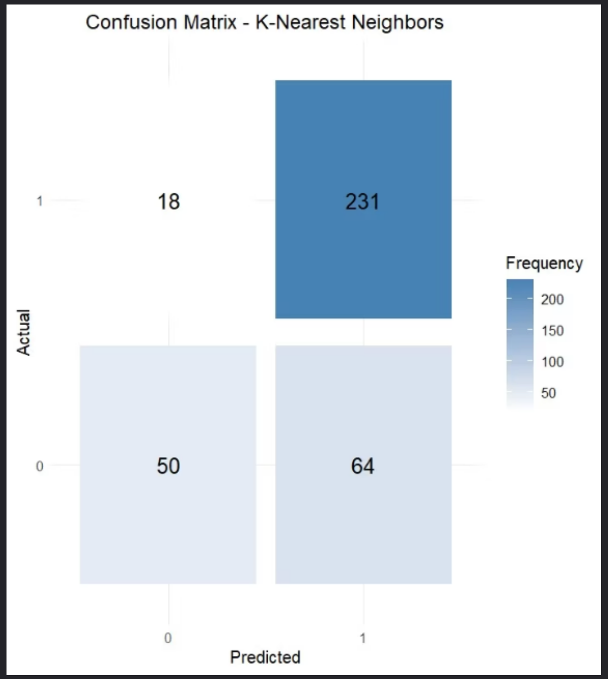
Model	Accuracy	F1	AUC
Logistic Regression	0.7741	0.8386	0.8269
Random Forest	0.7879	0.8571	0.8165
K-Nearest Neighbors	0.7741	0.8493	0.8093

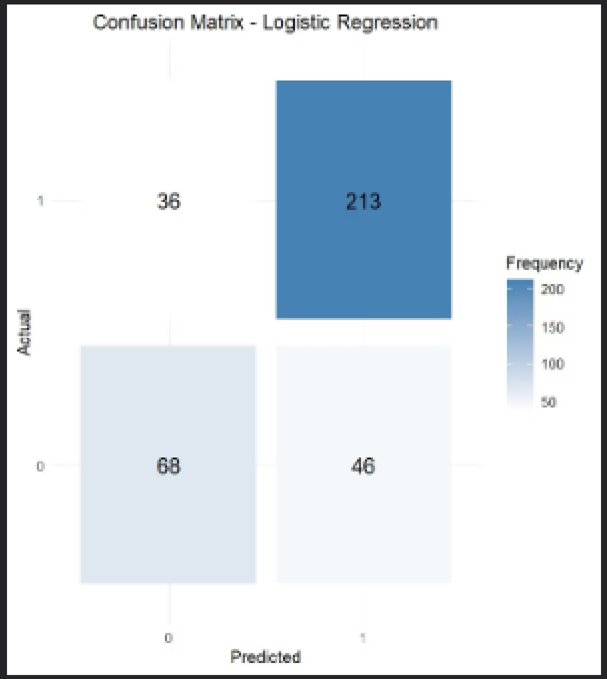
RF leads on Accuracy & F1

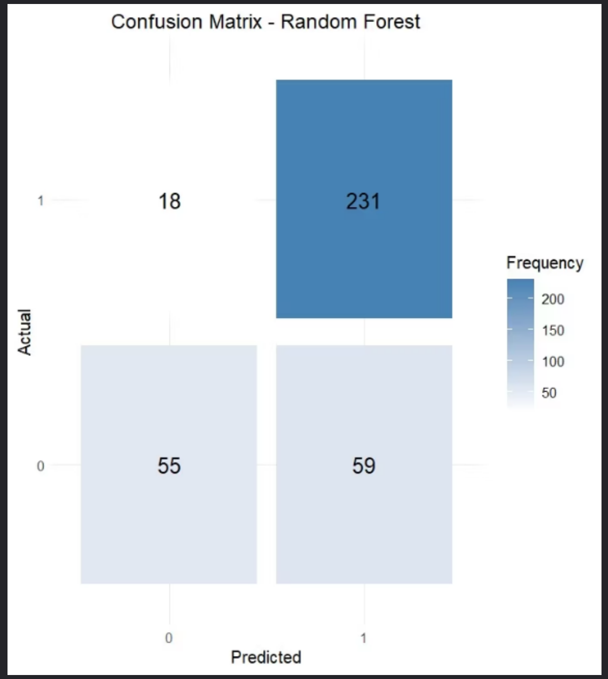
Logistic best on AUC → More interpretable

KNN competitive on F1

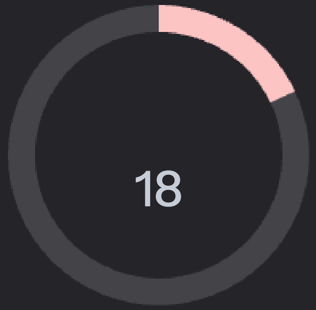
Confusion Matrices & Inference



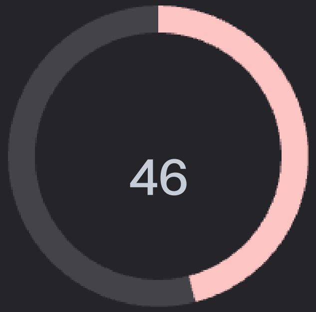




Logistic FN
Missed ADHD cases



RF/KNN FN
Better at finding real cases

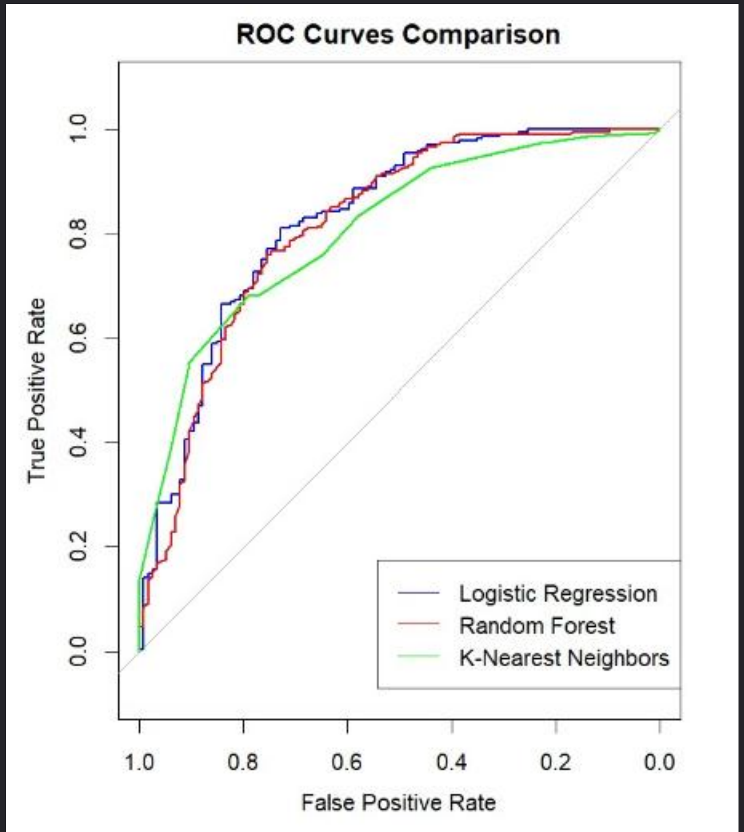


Logistic FP
Fewer false positives

RF/KNN FP
More false positives

If missing ADHD cases is worse, Random Forest or KNN might be better (fewer FN). However, they both have more FP compared to Logistic. As false alarms are more problematic, then Logistic is preferable—though it does miss more ADHD cases.

ROC Curves & Model Preference



Reviewing ROC curves:

- **Logistic Regression**: Highest AUC (0.827), best at ranking positives vs. negatives.
- **Random Forest**: AUC ~0.816, close behind Logistic. KNN: AUC ~0.809.

Final Preference:

- Maximize sensitivity (catch more true ADHD cases): Favor **Random Forest** or **KNN** (minimize false negatives).
- Value fewer false positives, prefer interpretability: Choose **Logistic Regression** (higher AUC, lower FP).

The best model depends on whether minimizing false negatives or false positives is more critical. Random Forest balances performance; Logistic offers interpretability.

Conclusions & Next Steps

Key Observations

1. **PCA** effectively handled the high-dimensional fMRI data.
2. **Random Forest** stands out with strong accuracy and low FN, while **Logistic** has the highest AUC and fewer FP.
3. **KNN** performs moderately well, matching Random Forest's low FN but having even more FP. Future improvements might include hyperparameter tuning, class imbalance techniques, or reintroducing `Sex_F` into a multi-output classification model.

Future Work

- Model tuning for imbalance
- Deeper interpretability
- Sex as predictor variable