DS-GA 1017: Responsible Data Science
Spring 2025 Final Project Report

Name: Cissy Xie, Kangrui Yu
NYU ID: lx2154 ky2390

August 3, 2025

1 Background

The Automated Decision System (ADS) developed for the WiDS Datathon 2025 aims to support early identification of Attention Deficit Hyperactivity Disorder (ADHD) and to classify biological sex based on neuroimaging and behavioral data. The stated purpose of this system is to assist researchers and clinicians in uncovering patterns in brain function, emotional responses, and demographic characteristics that are predictive of ADHD and sex classification. The dataset provided includes a diverse cohort of children and adolescents, featuring functional MRI data, self-reported emotional health assessments, and demographic variables. These insights could contribute to improving diagnostic practices and understanding the neurobiological basis of ADHD.

The ADS has two primary goals: (1) predicting ADHD status, and (2) predicting biological sex. While these are distinct classification tasks, they are performed using the same data inputs. There may be trade-offs in model performance if the features predictive of one target interfere with the other (e.g., emotional regulation patterns that differ by sex may confound ADHD signals). Overall, the system is designed to balance accuracy across both objectives.

2 Input and Output

2.1 Input

The input features include 19900 FCM matrix features, 18 quantitative features, and 9 categorical features. For the sake of analysis, we decide to only analyze a subset of categorical and quantitative features that we find are most relevant to the task.

2.1.1 Categorical Features

We selected two categorical features for profiling:

- `PreInt_Demos_Fam_Child_Ethnicity`
  Description: Child's reported ethnicity
  Data type: Integer
  Null count: 11 missing values out of 970 training samples
  Distribution: See Figure 1
- `PreInt_Demos_Fam_Child_Race`
  Description: Child's reported race
  Data type: Integer
  Null count: No missing values in 970 training samples
  Distribution: See Figure 1

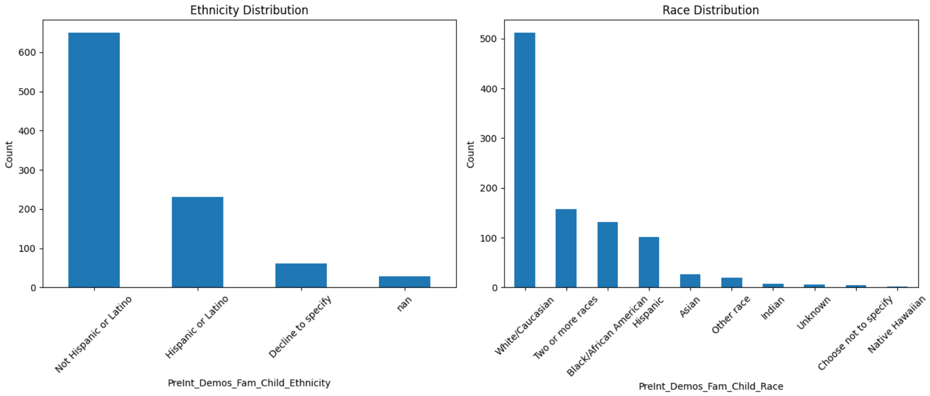

Figure 1: Distributions of Race and Ethnicity categorical features

2.1.2 Quantitative Features

We profile 16 quantitative features derived from three validated instruments: the Edinburgh Handedness Questionnaire (EHQ), the Alabama Parenting Questionnaire (APQ), and the Strength and Difficulties Questionnaire (SDQ).

- `EHQ_EHQ_Total` — Laterality Index (LI) Score
  Description: Assesses an individual's hand preference. Higher values indicate stronger right-hand dominance.
  Interpretation:
  - `-100`: Extreme left-handed
  - `-28` to `48`: Ambidextrous (middle range)
  - `+100`: Extreme right-handed

  Data type: Float
  Null count: No missing values
  Distribution: See Figure 2

- `Alabama Parenting Questionnaire (APQ)`
  Description: Behavioral scores assessing parenting styles. Higher scores reflect more frequent occurrence of the corresponding behavior. The features include:
  - `APQ_P_APQ_P_CP` — Corporal Punishment
  - `APQ_P_APQ_P_ID` — Inconsistent Discipline
  - `APQ_P_APQ_P_INV` — Involvement
  - `APQ_P_APQ_P_OPD` — Other Discipline Practices
  - `APQ_P_APQ_P_PM` — Poor Monitoring
  - `APQ_P_APQ_P_PP` — Positive Parenting

  Data type: Integer for each score
  Null count: No missing values
  Distribution: See Figure 3

- `Strength and Difficulties Questionnaire (SDQ)`
  Description: A set of behavioral and emotional scales measuring children's psychological attributes. Higher scores typically indicate greater intensity of the measured trait (except for the prosocial scale where higher is positive). The features include:

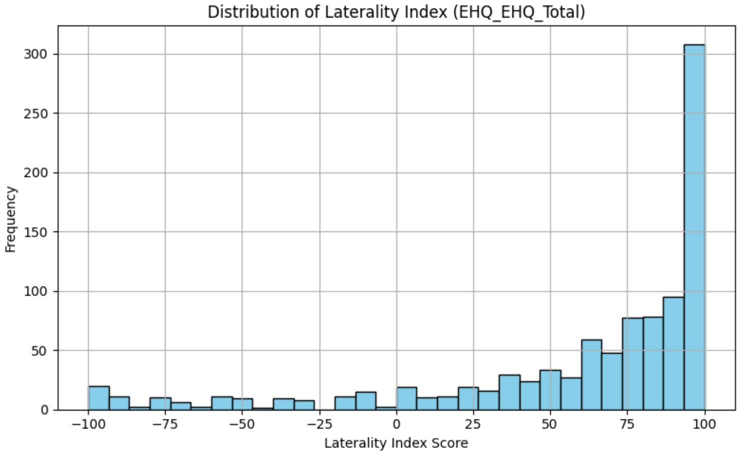

Figure 2: Distribution of Laterality Index (LI) scores from the Edinburgh Handedness Questionnaire.

- `SDQ_SDQ_Conduct_Problems` — Conduct Problems Scale
- `SDQ_SDQ_Difficulties_Total` — Total Difficulties Score
- `SDQ_SDQ_Emotional_Problems` — Emotional Problems Scale
- `SDQ_SDQ_Externalizing` — Externalizing Score
- `SDQ_SDQ_Generating_Impact` — Generating Impact Score
- `SDQ_SDQ_Hyperactivity` — Hyperactivity Scale
- `SDQ_SDQ_Internalizing` — Internalizing Score
- `SDQ_SDQ_Peer_Problems` — Peer Problems Scale
- `SDQ_SDQ_Prosocial` — Prosocial Scale

Data type: Integer for each score
Null count: No missing values across all SDQ scores
Distribution: See Figure 4

- `Matrix Features`
  The dataset may include matrix-type data such as fMRI scans or connectivity matrices. However, such features are difficult to interpret individually and are excluded from detailed analysis in this report.

We also plot the pairwise correlation heatmap (Figure 5) to examine relationships between features from the Alabama Parenting Questionnaire (APQ) and the Strength and Difficulties Questionnaire (SDQ).
The SDQ scores exhibit strong internal correlations, especially between emotional, internalizing, and peer problems, as well as between conduct, hyperactivity, and externalizing behaviors. This suggests that children experiencing one type of psychological difficulty often experience others as well. Similarly, moderate correlations are observed within APQ scores—most notably between positive parenting and involvement—indicating some consistency across reported parenting practices. In contrast, cross-questionnaire correlations (between APQ and SDQ scores) are generally weak, suggesting limited direct linear relationships between parenting behaviors and reported child difficulties in this dataset.

2.2 Output

- Predict sex: 0 for male, 1 for female
- Predict ADHD diagnosis: 0 for negative, 1 for positive

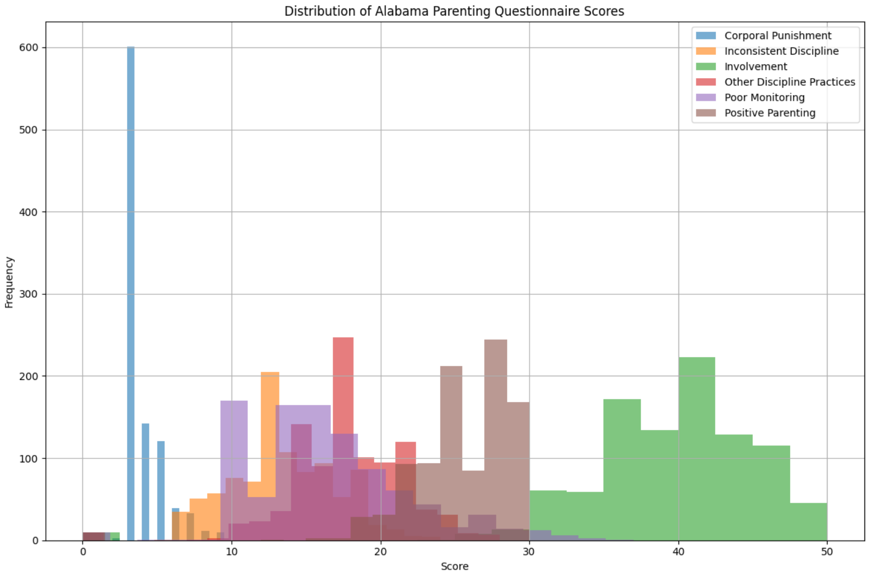

Figure 3: Distribution of six behavioral scores from the Alabama Parenting Questionnaire (APQ). Higher scores indicate more frequent corresponding behaviors.

3 Implementation and Validation

3.1 Implementation

We audit a publicly available baseline solution[1], which is built in Python and leverages `scikit-learn` and `LightGBM`. The main steps are:

1. **Data Loading.**
   - Quantitative and categorical metadata are read from the provided Excel files.
   - Functional connectome matrices are loaded from CSV.
2. **Preprocessing.**
   - *Missing values* in numeric features are imputed with the median; categorical features use the most frequent category.
   - *Standard Scaling* is performed via on all numeric features.
   - *Categorical encoding* is handled with one-hot encoding, ignoring unseen categories in test data.
3. **Model Training.**
   - Separate `LGBMClassifier` models are trained for the two targets (ADHD and Sex)(We focus on ADHD prediction in this report).
   - Hyperparameters such as `num_leaves`, `learning_rate`, and `n_estimators` are set to 63 and 1000.
   - Class imbalance is addressed by computing `scale_pos_weight` from class frequencies.
   - **Early stopping** (50 rounds) is applied based on the validation F1 score to prevent over-fitting.

3.2 Validation

To gauge generalization performance, we adopted an 80/20 stratified train/validation split on the original training set (stratifying on `ADHD_Outcome`). After fitting the preprocessing pipeline on the training fold, we evaluated both classifiers on the held-out 20%.
Results. On the validation set, the baseline models achieved:

[1] `https://www.kaggle.com/code/olaflundstrom/wids-datathon-2025-adhd-analysis-notebook`

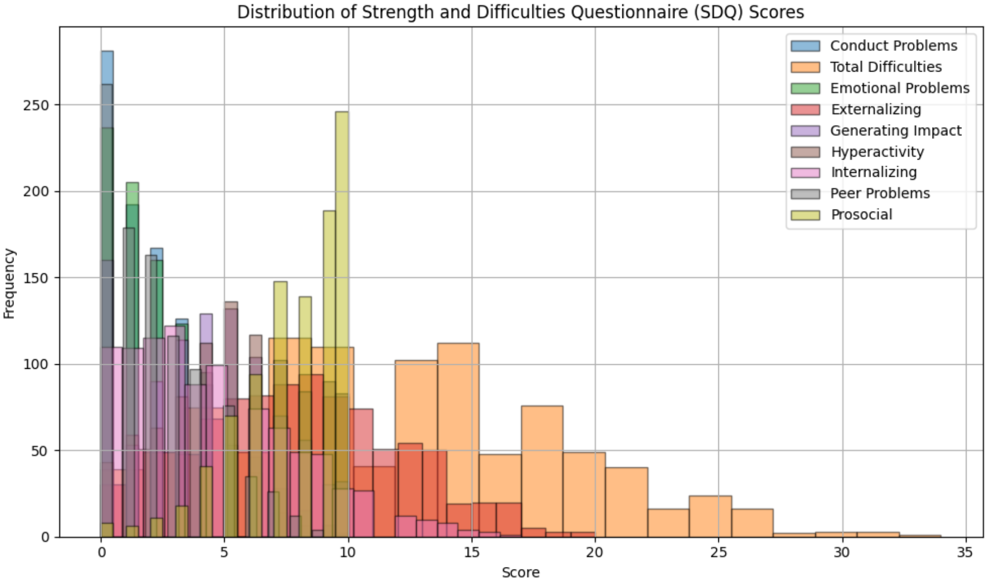

Figure 4: Distribution of nine behavioral and emotional scores from the Strength and Difficulties Questionnaire (SDQ).

- ADHD prediction: F1 = 0.8499

4 Outcomes

4.0.1 Fairness

In the context of ADHD screening the cost of a missed diagnosis is high, so we place primary emphasis on:

- **Recall** ($\frac{\text{TP}}{\text{TP}+\text{FN}}$) — the fraction of true ADHD cases correctly identified.
- **False Negative Rate** ($\frac{\text{FN}}{\text{TP}+\text{FN}} = 1 - \text{Recall}$) — the fraction of ADHD cases the model fails to detect.

By maximizing recall (minimizing false negatives), we reduce the risk of under-diagnosing ADHD and the avoid consequences for those individuals.

Because both `Ethnicity` and `Race` comprise more than two subgroups, all disparity metrics below are computed on a *max–min* basis—that is, the difference between the subgroup with the highest metric value and the subgroup with the lowest.

	FNR_diff	FPR_diff	Demographic_Parity_ratio	Selection_Rate_diff
Ethnicity	0.4565	0.5000	0.4088	0.4821
Race	1.0000	0.6429	0.0000	0.8889

Table 1: Fairness metrics by group (computed on a max–min basis for multi-class subgroups).

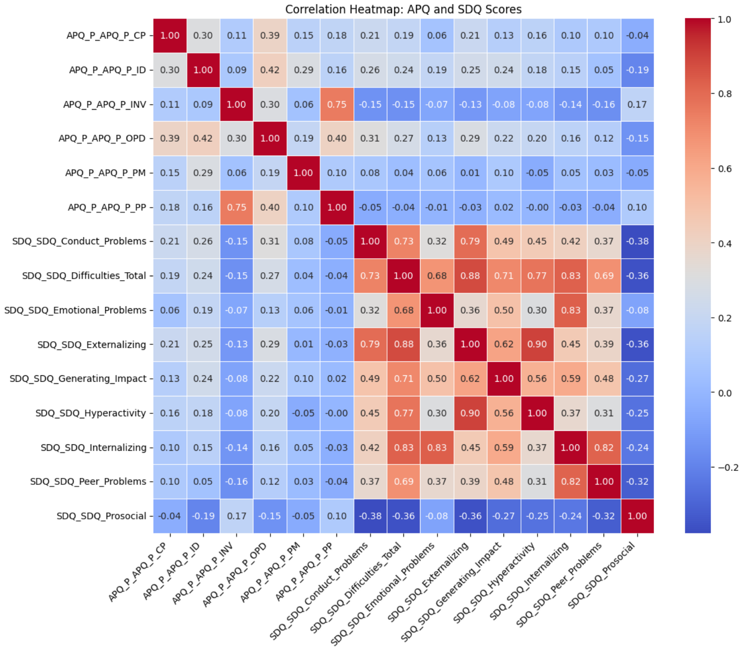

Figure 5: Pairwise correlation heatmap for all features in the APQ and SDQ questionnaires.

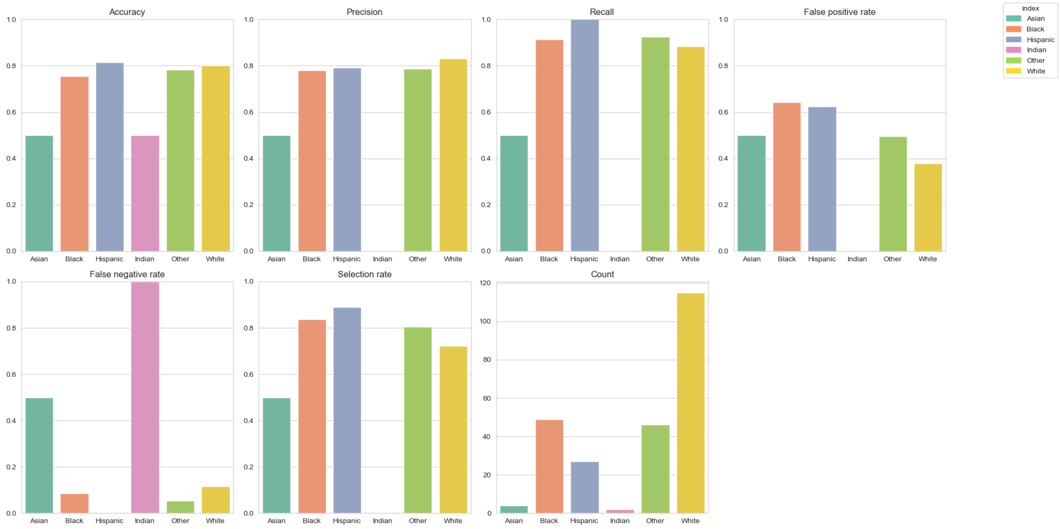

Figure 6: Group-wise performance for `PreInt_Demos_Fam_Child_Ethnicity`

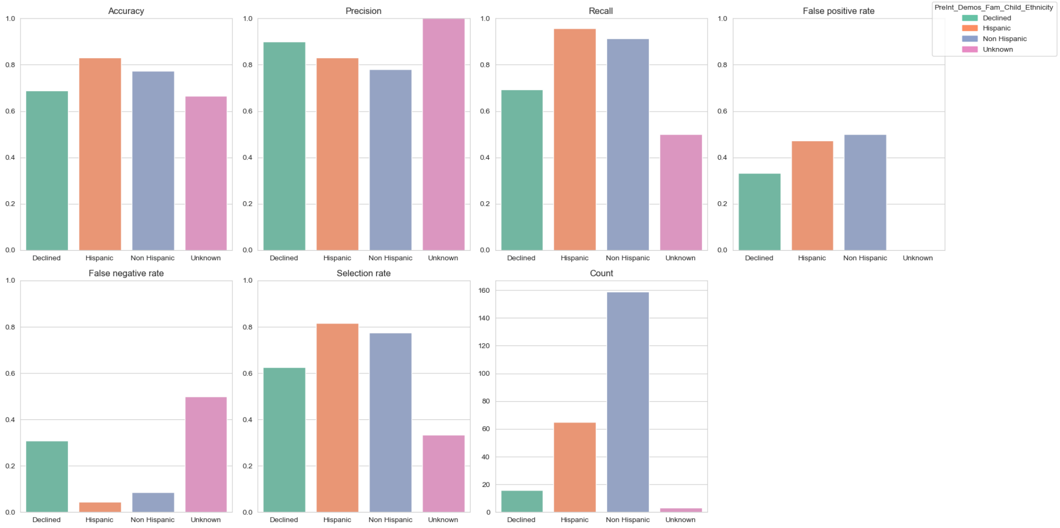

Figure 7: Group-wise performance for `PreInt_Demos_Fam_Child_Race`

- **False Negative Rate Disparity.**
  - Ethnicity gap of 0.4565 implies some ethnic subgroups are over 45 % more likely to be missed.
  - Race gap of 1.0000 reflect the racial group ”Indian” has 0% recall and 100% FNR.
- **Demographic Parity Ratio.**
  - For both groups, the demographic parity ration is lower than common threshold 0.8. Indicating a highly unbalanced performance across different race and ethnicity groups.

Conclusion. The model exhibits substantial unfairness across both ethnicity and race. Extreme max–min gaps in FNR and FPR, plus zero demographic parity in Race, indicate that some subgroups are either never flagged for ADHD (risking under-treatment) or are over-flagged (risking over-treatment).

4.0.2 Shap

To interpret the ADHD classification model, we applied SHAP (SHapley Additive explanations), an approach to quantify each feature's contribution to individual predictions.
1. Overall Feature Importance. We computed SHAP values on the validation set and visualized global feature importance using a beeswarm plot (Figure 8). The most influential features included `SDQ_SDQ_Hyperactivity` and `SDQ_SDQ_Externalizing`, highlighting the central role of behavioral attributes in ADHD prediction. High hyperactivity scores strongly increased the model's confidence in predicting ADHD, while low scores significantly reduced the likelihood. Externalizing scores exhibited a similar, though less pronounced, pattern: higher values tended to increase the predicted probability of ADHD, whereas lower values modestly decreased it.
2. Feature-wise Analysis. We further explored how individual features contribute to model predictions by examining SHAP dependency plots (Figure 9 and Figure 10). For `SDQ_SDQ_Hyperactivity`, higher values consistently pushed the model toward predicting ADHD, reflecting a strong and direct association between hyperactivity and the condition. Similarly, higher values in `SDQ_SDQ_Externalizing` generally increased the predicted probability of ADHD, though with more moderate impact. These patterns indicate that behavioral hyperactivity and externalizing behaviors—are key drivers in the model's decision-making process.
3. Individual Prediction Decomposition. To better understand how the model makes decisions at the individual level, we analyzed SHAP force plots for one high-confidence positive prediction and one high-confidence negative prediction (Figure 11 and Figure 12). In the positive case, high values for both `SDQ_SDQ_Hyperactivity` and `SDQ_SDQ_Externalizing` strongly pushed the model toward predicting ADHD, as indicated by their prominent positive SHAP values. In contrast, in the negative case, low values for

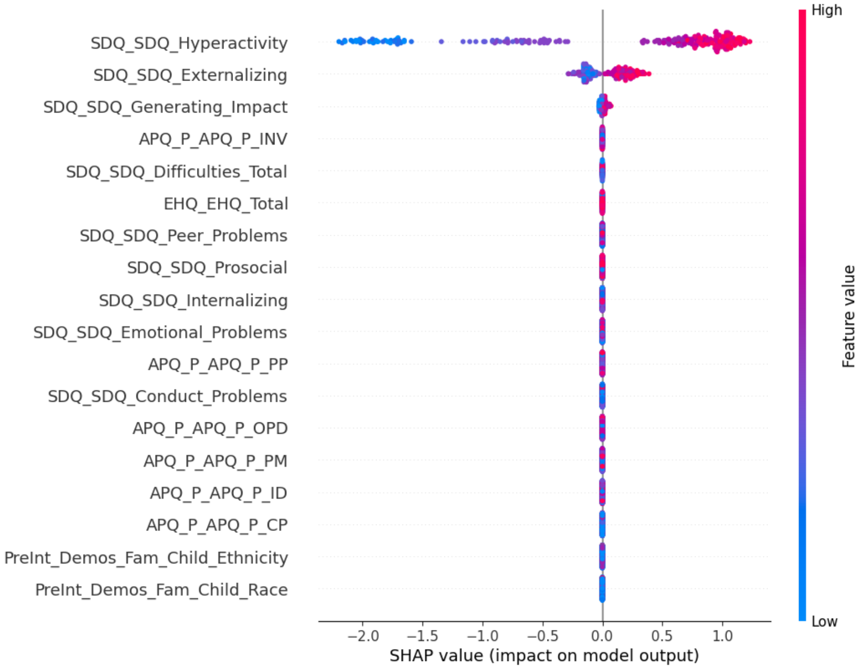

Figure 8: SHAP beeswarm plot showing global feature importance for ADHD classification. Each point represents a sample; color indicates the feature value, and horizontal position reflects impact on model output.

the same features had a strong suppressive effect on the ADHD prediction, pulling the output well below the base value. These examples further demonstrate how hyperactivity and externalizing behaviors operate symmetrically in the model: higher scores raise the predicted ADHD probability, while lower scores reduce it. Overall, SHAP analysis provides meaningful insights into how the model weighs different aspects of child behavior in predicting ADHD outcomes.

5 Summary

In this project, we audited a publicly available LightGBM–based ADHD screening model on the WiDS Datathon 2025 dataset, focusing on both overall performance and subgroup fairness. Our key findings are:

- **Fairness concerns.** Fairness metrics computed on a max–min basis revealed substantial disparities across both ethnicity and race groups.
- **Model interpretability.** SHAP analysis highlighted behavioral features—particularly `SDQ_Hyperactivity` and `SDQ_Externalizing`—as the most influential drivers of ADHD predictions.
- **Possible Mitigation Methods.** Possible mitigation strategies include assigning weights to different subgroups as in-processing and adding threshold optimizer as post-processing.

6 Contribution

Cissy Xie is responsible for the section: Background1, Input and Output2, Shap Analysis4.0.2

Kangrui Yu is responsible for the section: Implementation and Validation3, Fairness4.0.1, Summary5

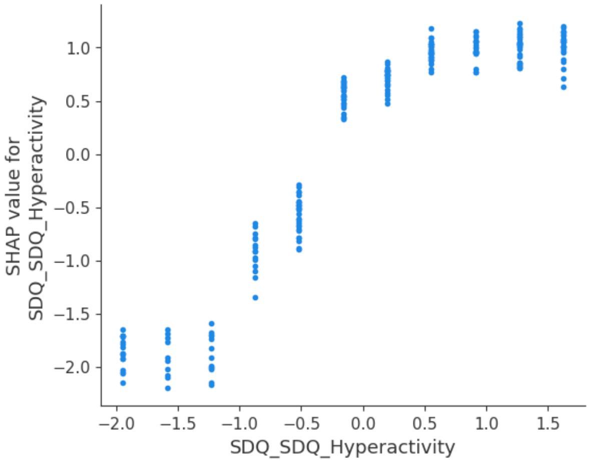

Figure 9: SHAP dependency plot for SDQ_SDQ_Hyperactivity

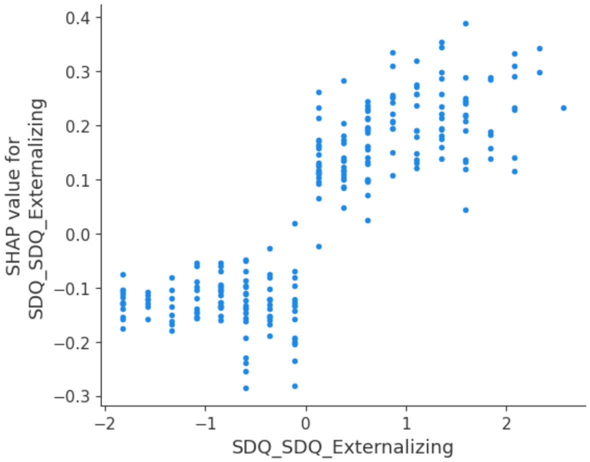

Figure 10: SHAP dependency plot for APQ_P_APQ_P_INV

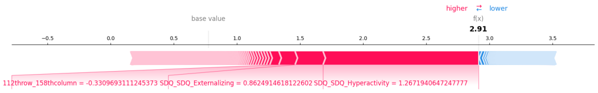

Figure 11: SHAP force plot for a positive ADHD prediction

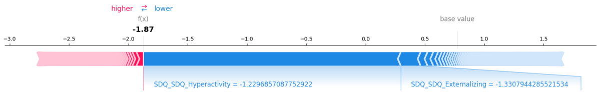

Figure 12: SHAP force plot for a negative ADHD prediction

References

[1] Kaggle. *WiDS Datathon 2025*. Available at: `https://www.kaggle.com/competitions/widsdatathon2025/overview`