Predicting ADHD Outcomes in Children Using Machine Learning Techniques

Sai Bhaskar Boddduluru

Masters in Bioinformatics

University of Delaware

December 2024

Predicting ADHD Outcomes in Children Using Machine Learning Techniques

December 9, 2024

Abstract

Attention-Deficit/Hyperactivity Disorder (ADHD) is a prevalent neurodevelopmental disorder affecting children worldwide, characterized by persistent patterns of inattention, hyperactivity, and impulsivity. This study aims to predict ADHD outcomes based on various socio-demographic, behavioral, and health-related factors using machine learning models. Utilizing the 2017 National Survey of Children's Health (NSCH) dataset, we employed Logistic Regression, Random Forest, and Gradient Boosting classifiers to identify significant predictors of ADHD. The dataset was meticulously preprocessed to handle missing values and exclude invalid ADHD codes, ensuring data integrity and reliability. Exploratory Data Analysis (EDA) provided insights into the distribution and relationships of variables, informing feature selection and model training. Our models achieved high accuracy and Area Under the Curve (AUC) scores, with the Gradient Boosting model demonstrating the highest performance. Feature importance analysis revealed that ADHD severity, special health care needs, and other mental health conditions are critical factors in predicting ADHD outcomes. These findings can inform targeted interventions and support mechanisms for children at risk of ADHD, contributing to improved diagnostic and treatment strategies.

Contents

1 Background

Attention-Deficit/Hyperactivity Disorder (ADHD) is one of the most common neurodevelopmental disorders diagnosed in childhood, characterized by persistent patterns of inattention, hyperactivity, and impulsivity. According to the Centers for Disease Control and Prevention (CDC), approximately 9.4% of children in the United States have been diagnosed with ADHD [1]. ADHD can significantly impact academic performance, social interactions, and overall quality of life, making early detection and intervention crucial.

1.1 Significance of ADHD Research

ADHD not only affects the individual child but also has broader implications for families, educational systems, and healthcare providers. Children with ADHD are at a higher risk for academic underachievement, strained peer relationships, and increased likelihood of developing comorbid conditions such as anxiety and depression [2]. Understanding the predictors of ADHD can aid in developing effective screening tools, enabling timely interventions that can mitigate negative outcomes and enhance the well-being of affected children.

1.2 Literature Review

Previous studies have explored various factors associated with ADHD, including genetic predispositions, environmental influences, and socio-demographic variables. For instance, research has identified a significant genetic component in ADHD, with heritability estimates ranging from 70% to 80% [3]. Environmental factors such as prenatal exposure to tobacco smoke, low birth weight, and psychosocial stressors have also been implicated [4].

Moreover, socio-demographic variables like socio-economic status, family structure, and parental education have been shown to influence the prevalence and severity of ADHD symptoms [5]. However, there is a need for more comprehensive models that integrate multiple predictors to enhance the accuracy of ADHD diagnosis and intervention strategies. Traditional statistical methods may overlook complex interactions between variables, whereas machine learning offers robust tools for analyzing large and intricate datasets, identifying patterns that can inform clinical practices.

1.3 Motivation for the Study

Despite the advancements in understanding ADHD, challenges remain in accurately predicting ADHD outcomes based on available data. The complexity of ADHD, with its multifaceted etiology and diverse symptomatology, necessitates sophisticated analytical approaches. This study leverages machine learning techniques to address these challenges, aiming to develop predictive models that can assist healthcare professionals in identifying children at risk of ADHD. By utilizing the 2017 NSCH dataset, which encompasses a wide range of socio-demographic, behavioral, and health-related variables, this research seeks to provide a nuanced understanding of the factors influencing ADHD outcomes.

2 Study Design

2.1 Aims

The primary objective of this study is to develop and evaluate machine learning models that can accurately predict ADHD outcomes in children based on socio-demographic, behavioral, and health-related factors. Specific aims include:

- To preprocess and clean the NSCH dataset for analysis, ensuring data quality and integrity.
- To conduct comprehensive Exploratory Data Analysis (EDA) to understand the distribution and relationships of variables, guiding feature selection.
- To train and compare the performance of Logistic Regression, Random Forest, and Gradient Boosting classifiers in predicting ADHD outcomes.
- To identify and interpret the most significant predictors of ADHD outcomes through feature importance analysis.
- To provide actionable insights for early intervention and support mechanisms tailored to children at risk of ADHD.

2.2 Population Selection

The study utilizes data from the 2017 National Survey of Children's Health (NSCH), a nationally representative survey that provides comprehensive information on various health and well-being indicators for children in the United States. The initial dataset comprised 21,599 observations with 813 variables, encompassing demographic details, health conditions, behavioral patterns, and environmental factors.

2.2.1 Inclusion and Exclusion Criteria

To ensure the relevance and accuracy of the analysis, the following criteria were applied:

- **Inclusion**: Children aged 0 to 17 years, with complete information on the ADHD-related variables and selected predictors.
- **Exclusion**: Observations with invalid ADHD codes (e.g., 95, 99) were excluded to maintain the integrity of the outcome variable. Additionally, rows with missing values in any of the selected columns were removed to prevent biases in the modeling process.

After applying these criteria, the final analysis included 18,498 observations with 22 variables, ensuring a robust and clean dataset for subsequent analysis.

2.3 Data and Materials

- **Dataset**: 2017 NSCH Topical CAHMI DRCv2 (`2017_NSCH_Topical_CAHMI_DRCv2.csv`) was obtained from the official NSCH website. The dataset includes a wide array of variables related to children's health, behavior, and socio-demographic characteristics.

- **Variables**: The study focused on 21 predictor variables encompassing socio-demographics, behavioral factors, and health indicators, along with the defined ADHD outcome variable. Detailed descriptions of these variables are provided in the Data Dictionary section.

- **Software and Tools**:
  - **Programming Language**: Python
  - **Libraries**: pandas for data manipulation, seaborn and matplotlib for data visualization, scikit-learn for machine learning algorithms and evaluation metrics.
  - **Environment**: Jupyter Notebook was used for interactive data analysis and model development.

- **Hardware**: The analysis was conducted on a personal computer with sufficient computational resources to handle the dataset and execute machine learning algorithms efficiently.

3 Statistical Analysis

3.1 Methods

This study employed a systematic approach combining traditional statistical methods and advanced machine learning techniques to analyze the data. The analysis workflow included the following stages:

3.1.1 Data Loading and Preprocessing

- **Loading Data**: The dataset was loaded into a pandas DataFrame, ensuring that all variables were correctly interpreted.

- **Initial Exploration**: The initial shape, first few rows, and data types were examined to understand the dataset's structure.

- **Defining the Outcome Variable**: The ADHD outcome was defined based on the `ADHD_17` variable, mapping values 1 to 0 (No ADHD) and values 2 and 3 to 1 (ADHD present). Invalid codes (95, 99) were excluded to maintain consistency.

- **Variable Selection**: A subset of 21 predictor variables was selected based on relevance to ADHD and availability in the dataset. These included socio-demographic, behavioral, and health-related factors.

- **Handling Missing Values**: The dataset was checked for missing values across the selected variables. Observations with any missing values in the selected columns were dropped to ensure data quality.

- **Data Cleaning**: Categorical variables were appropriately encoded, and numerical variables were scaled using StandardScaler to standardize the feature ranges for machine learning algorithms.

3.1.2 Exploratory Data Analysis (EDA)

- **Descriptive Statistics**: Summary statistics for each variable were computed to understand central tendencies and dispersions.

- **Visualization**: A comprehensive set of plots, including histograms, bar plots, box plots, violin plots, correlation heatmaps, cumulative distribution plots, scatter plots, KDE plots, joint plots, pair plots, rug plots, and PCA visualizations, were generated to explore the distributions and relationships among variables.

- **Correlation Analysis**: Correlation coefficients were calculated for numerical predictors to identify potential multicollinearity and inform feature selection.

3.1.3 Model Training

- **Data Splitting**: The dataset was split into training and testing sets using an 80-20 split, ensuring stratification based on the ADHD outcome to maintain class distribution.

- **Feature Scaling**: Numerical predictors were standardized using StandardScaler to facilitate the convergence of machine learning algorithms.

- **Model Selection**: Three classifiers were selected for their robustness and ability to handle classification tasks effectively:

  1. **Logistic Regression**: A fundamental linear model for binary classification, serving as a baseline.
  2. **Random Forest**: An ensemble method leveraging multiple decision trees to improve predictive performance and control overfitting.
  3. **Gradient Boosting**: An advanced ensemble technique that builds trees sequentially, optimizing for errors made by previous trees.

- **Training Models**: Each model was trained on the training dataset, with default hyperparameters initially used. Further hyperparameter tuning can be performed to enhance performance.

3.1.4 Model Evaluation

- **Performance Metrics**: Models were evaluated using multiple metrics to ensure a comprehensive assessment of their predictive capabilities:

  1. **Accuracy**: The proportion of correctly classified instances.
  2. **AUC (Area Under the ROC Curve)**: Measures the model's ability to distinguish between classes.
  3. **Confusion Matrix**: Provides a summary of prediction results, including true positives, true negatives, false positives, and false negatives.
  4. **Classification Report**: Includes precision, recall, and F1-score for each class, offering deeper insights into model performance.

- **ROC Curves and AUC**: Receiver Operating Characteristic (ROC) curves were plotted for each model, and the corresponding AUC scores were calculated to evaluate the models' discriminative abilities.
- **Feature Importance Analysis**: For ensemble models like Random Forest and Gradient Boosting, feature importance scores were extracted to identify the most influential predictors of ADHD outcomes.

3.2 Response Variable

The dependent variable in this study is **ADHD_outcome**, defined based on the `ADHD_17` variable from the dataset. The mapping is as follows:

- `ADHD_17` = 1: No ADHD (mapped to 0)
- `ADHD_17` = 2 or 3: ADHD present (mapped to 1)

Invalid codes (95, 99) were excluded from the analysis to ensure the reliability of the outcome variable. This binary classification facilitates the application of logistic regression and other binary classifiers in the modeling process.

3.3 Independent Variables/Risk Factors

The study examined the following predictor variables, categorized into socio-demographic, behavioral, and health indicators:

3.3.1 Socio-Demographic Variables

- **Age (`SC_AGE_YEARS`)**: Age of the child in years, ranging from 0 to 17.
- **Sex (`SC_SEX`)**: Sex of the child, coded as 1 (Male), 2 (Female), and 3 (Other).
- **Race (`SC_RACE_R`)**: Race of the child, coded numerically (e.g., 1=White, 2=Black, 3=Asian, etc.).
- **Hispanic Origin (`SC_HISPANIC_R`)**: Indicates whether the child is of Hispanic origin, coded as 1 (Yes) and 2 (No).
- **Poverty Level (`povlev4_17`)**: Poverty level indicator for the child's family in 2017, coded as 1 (Below Poverty) and 2 (At/Above Poverty).
- **Family Count (`FAMCOUNT`)**: Number of family members in the household, ranging from 1 to 10+.

3.3.2 Behavioral Factors

- **Meal Together Frequency (`MealTogether_17`)**: Frequency of family meals together in 2017, coded from 1 (Never) to 5 (Always).
- **Reading to Child (`readto_17`)**: Number of days parents read to the child per week, ranging from 0 to 7 days.

- **Smoking Environment (smoking_17)**: Indicates whether there is a smoking environment in the household, coded as 1 (Yes) and 2 (No).
- **Physical Activity (PHYSACTIV)**: Level of physical activity, coded from 1 (Inactive) to 5 (Very Active).
- **TV Watching Hours (TVwatch_17)**: Average TV watching hours per day, ranging from 0 to 24 hours.

3.3.3 Health Indicators

- **Special Health Care Needs (SC_CSHCN)**: Indicates whether the child has special health care needs, coded as 1 (Yes) and 2 (No).
- **Anxiety (anxiety_17)**: Presence of anxiety in the child, coded as 0 (No) and 1 (Yes).
- **Depression (depress_17)**: Presence of depression in the child, coded as 0 (No) and 1 (Yes).
- **Learning Disability (learning_17)**: Presence of learning disability in the child, coded as 0 (No) and 1 (Yes).
- **ADHD Severity (ADHDSev_17)**: ADHD severity indicator, coded as 1 (Mild), 2 (Moderate), and 3 (Severe).
- **BMI Category (BMI4_17)**: Body Mass Index (BMI) category, coded as 1 (Underweight), 2 (Normal), 3 (Overweight), and 4 (Obese).
- **Parental Concern About Weight (WgtConcn_17)**: Parental concern about the child's weight, coded from 1 (Not Concerned) to 5 (Extremely Concerned).
- **Mental Health Care Received (MentHCare_17)**: Indicates whether the child received needed mental health care, coded as 1 (Yes) and 2 (No).
- **Intellectual Disability (IntDisab_17)**: Presence of intellectual disability, coded as 0 (No) and 1 (Yes).
- **Other Mental Health Conditions (OthrMent_17)**: Presence of other mental health conditions, coded as 0 (No) and 1 (Yes).

3.4 Variable Selection/Model Selection

All selected predictor variables were included in the models to maximize the potential for identifying significant predictors of ADHD outcomes. Feature importance was assessed using the Random Forest model to determine the most influential predictors, which informed the interpretation of model results. The selection of Logistic Regression, Random Forest, and Gradient Boosting classifiers was based on their proven efficacy in binary classification tasks and their ability to handle complex interactions between variables.

4 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to uncover underlying patterns, detect anomalies, and test hypotheses about the dataset. A variety of plots were generated to visualize the distribution and relationships among the selected variables. Each plot is discussed below, providing insights into the data's structure and the factors influencing ADHD outcomes.

4.1 Distribution of Child Age (Years)

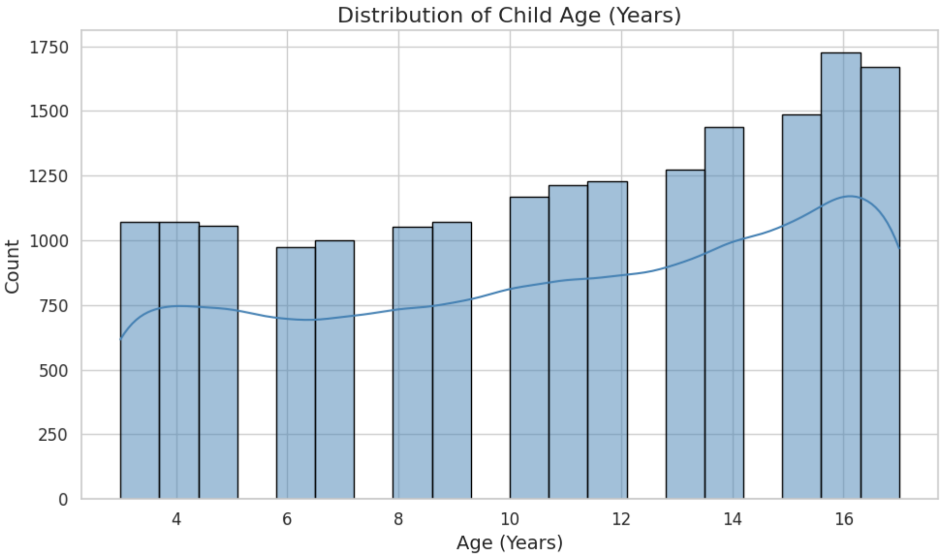


Figure 1: Distribution of Child Age (Years)

Discussion: Figure 1 illustrates the age distribution of the children in the study. The histogram with a Kernel Density Estimate (KDE) overlay shows that the majority of the sample consists of children aged between 6 to 12 years, with a tapering distribution towards younger and older ages. This distribution is expected given the prevalence of ADHD diagnoses typically occurring during school years.

4.2 ADHD Outcome Distribution

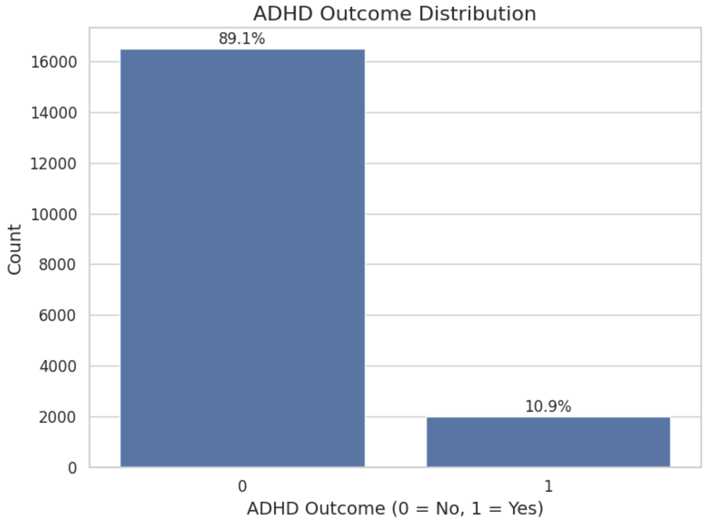


Figure 2: ADHD Outcome Distribution

Discussion: Figure 2 presents the proportion of children with and without ADHD. The count plot reveals a clear imbalance, with a significantly higher number of children without ADHD compared to those diagnosed with ADHD. This class imbalance is crucial to consider during model training to ensure that the classifiers do not become biased towards the majority class.

4.3 ADHD Outcome Rate by Sex

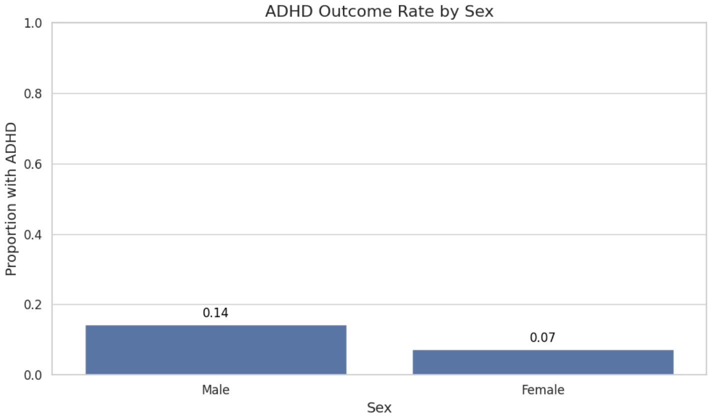


Figure 3: ADHD Outcome Rate by Sex

Discussion: Figure 3 compares the proportion of ADHD diagnoses across different sexes. The bar plot indicates that males exhibit a higher rate of ADHD compared to females and children classified as "Other." This aligns with existing literature, which consistently reports higher ADHD prevalence in males.

4.4 ADHD Outcome Rate by Race

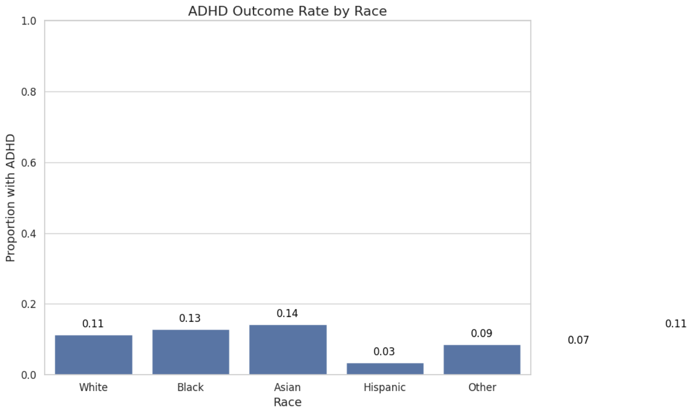


Figure 4: ADHD Outcome Rate by Race

Discussion: Figure 4 explores the relationship between race and ADHD outcomes. The bar plot shows variations in ADHD prevalence across different racial groups. Notably, certain minority groups may exhibit differing rates of ADHD diagnoses, which could be influenced by factors such as access to healthcare, cultural perceptions of ADHD, and potential biases in diagnostic practices.

4.5 Age Distribution by ADHD Outcome (Boxplot)

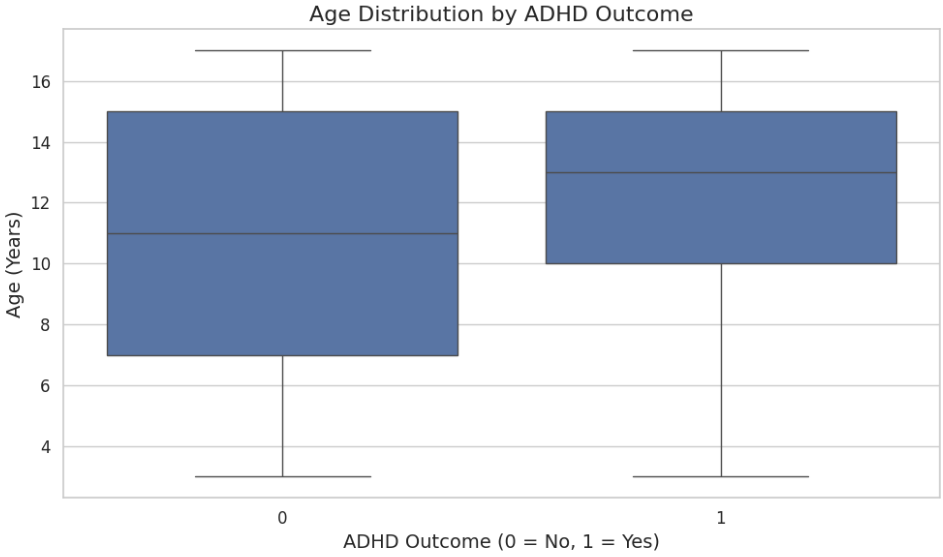


Figure 5: Age Distribution by ADHD Outcome (Boxplot)

Discussion: Figure 5 presents a boxplot comparing the age distributions of children with and without ADHD. The plot reveals that children diagnosed with ADHD tend to be slightly younger on average compared to their non-ADHD counterparts. This could suggest that ADHD is more frequently diagnosed at earlier ages or that younger children are more closely monitored for behavioral issues.

4.6 Age Distribution by ADHD Outcome (Violin Plot)

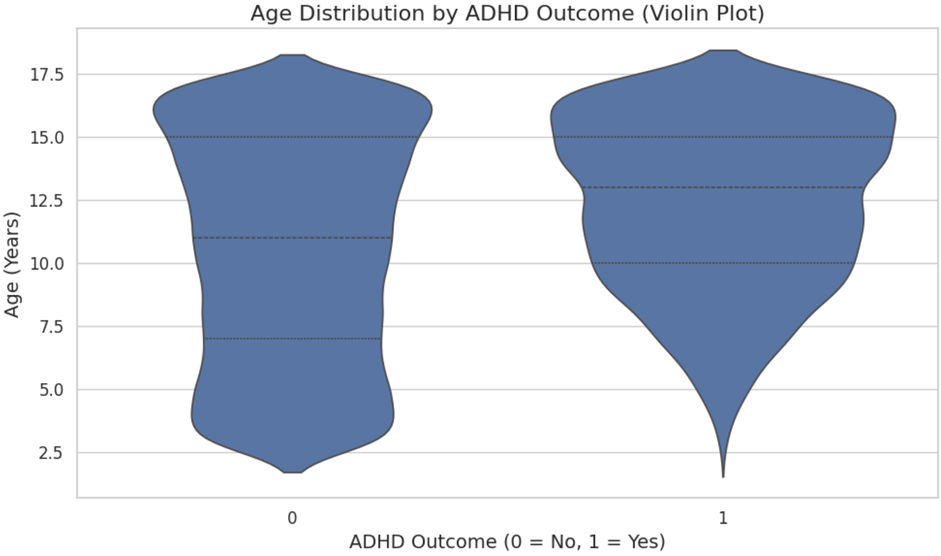


Figure 6: Age Distribution by ADHD Outcome (Violin Plot)

Discussion: Figure 6 complements the boxplot by providing a more detailed view of the age distribution across ADHD outcomes. The violin plot illustrates the density of age groups, indicating a higher concentration of ADHD diagnoses among younger children. This visualization emphasizes the importance of early identification and intervention in younger age groups.

4.7 ADHD Outcome by Anxiety Status (Stacked Bar Plot)

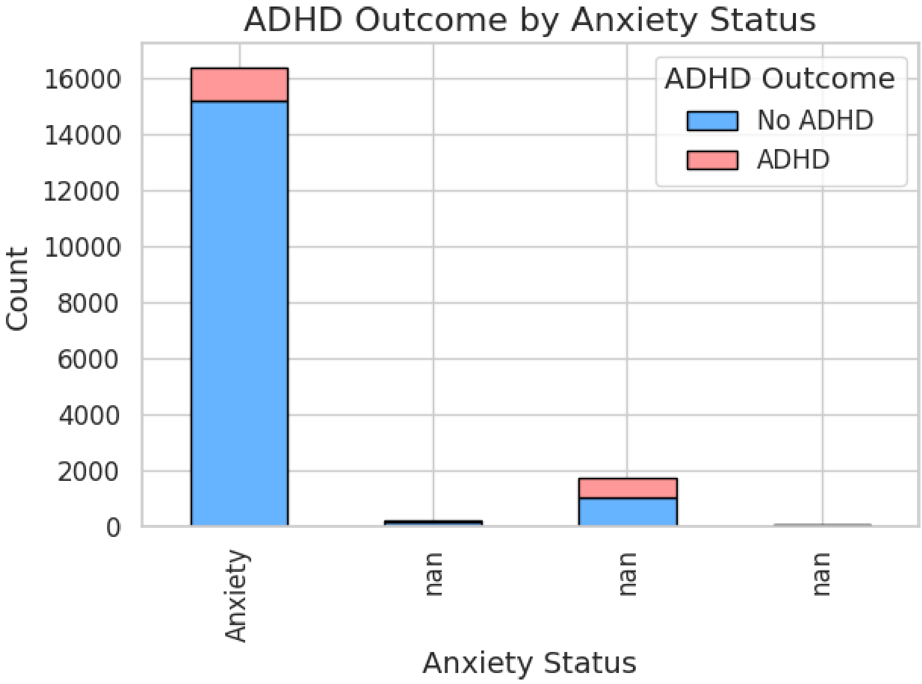
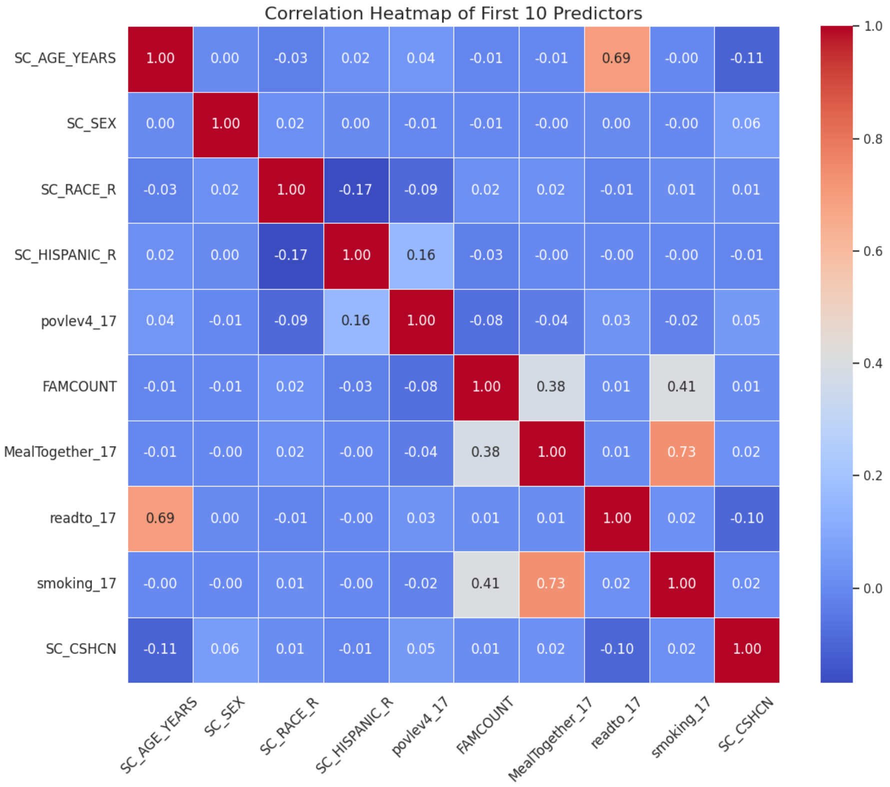


Figure 7: ADHD Outcome by Anxiety Status

Discussion: Figure 7 examines the relationship between anxiety and ADHD outcomes. The stacked bar plot shows that children with anxiety have a higher prevalence of ADHD diagnoses compared to those without anxiety. This correlation suggests a potential comorbidity between anxiety and ADHD, highlighting the need for comprehensive mental health assessments in children presenting with either condition.

4.8 Correlation Heatmap of Selected Predictors



	SC_AGE_YEARS	SC_SEX	SC_RACE_R	SC_HISPANIC_R	povlev4_17	FAMCOUNT	MealTogether_17	readto_17	smoking_17	SC_CSHCN
SC_AGE_YEARS	1.00	0.00	-0.03	0.02	0.04	-0.01	-0.01	0.69	-0.00	-0.11
SC_SEX	0.00	1.00	0.02	0.00	-0.01	-0.01	-0.00	0.00	-0.00	0.06
SC_RACE_R	-0.03	0.02	1.00	-0.17	-0.09	0.02	0.02	-0.01	0.01	0.01
SC_HISPANIC_R	0.02	0.00	-0.17	1.00	0.16	-0.03	-0.00	-0.00	-0.00	-0.01
povlev4_17	0.04	-0.01	-0.09	0.16	1.00	-0.08	-0.04	0.03	-0.02	0.05
FAMCOUNT	-0.01	-0.01	0.02	-0.03	-0.08	1.00	0.38	0.01	0.41	0.01
MealTogether_17	-0.01	-0.00	0.02	-0.00	-0.04	0.38	1.00	0.01	0.73	0.02
readto_17	0.69	0.00	-0.01	-0.00	0.03	0.01	0.01	1.00	0.02	-0.10
smoking_17	-0.00	-0.00	0.01	-0.00	-0.02	0.41	0.73	0.02	1.00	0.02
SC_CSHCN	-0.11	0.06	0.01	-0.01	0.05	0.01	0.02	-0.10	0.02	1.00

Figure 8: Correlation Heatmap of Selected Predictors

Discussion: Figure 8 displays the correlations among the first ten predictor variables. The heatmap reveals significant correlations between certain variables, such as age and BMI category, or special health care needs and other mental health conditions. Identifying these correlations is essential for understanding multicollinearity and informing feature selection to enhance model performance and interpretability.

4.9 Cumulative Distribution of Child Age

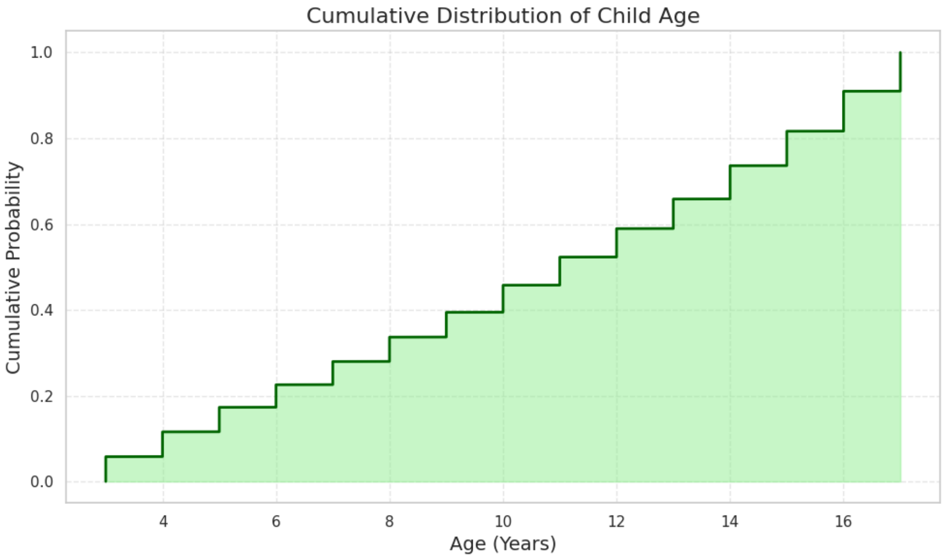


Figure 9: Cumulative Distribution of Child Age

Discussion: Figure 9 presents the cumulative distribution function (CDF) of child ages. The CDF plot helps in understanding the proportion of the population below a certain age. It is particularly useful for assessing the age threshold where a significant proportion of ADHD diagnoses occur, aiding in targeting age-specific interventions.

4.10 Age vs TV Watching Hours (Scatter Plot)

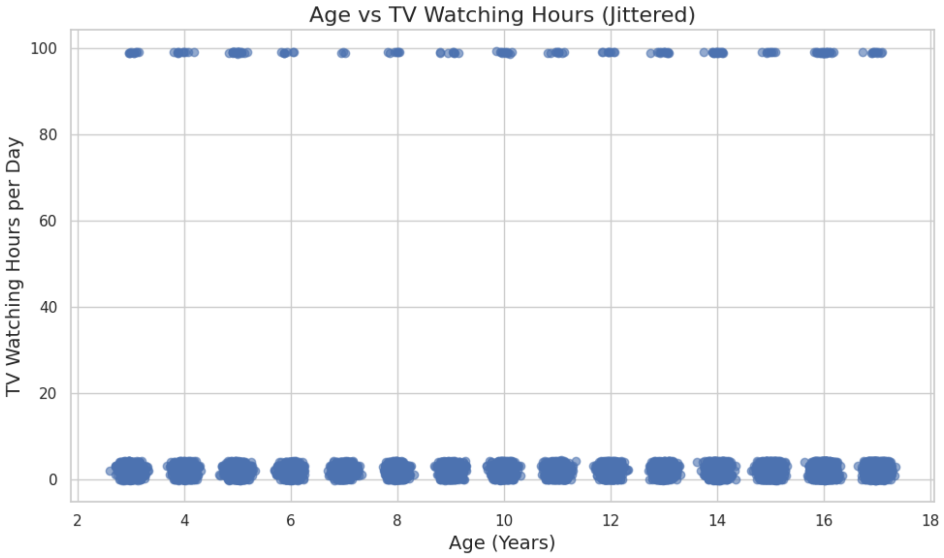


Figure 10: Age vs TV Watching Hours (Jittered)

Discussion: Figure 10 explores the relationship between a child's age and their daily TV watching hours. The scatter plot, with added jitter to prevent overplotting, suggests that younger children may watch more TV compared to older children. This behavior could be associated with increased screen time impacting attention spans, potentially influencing ADHD outcomes.

4.11 Age Distribution by ADHD Outcome (KDE Plot)

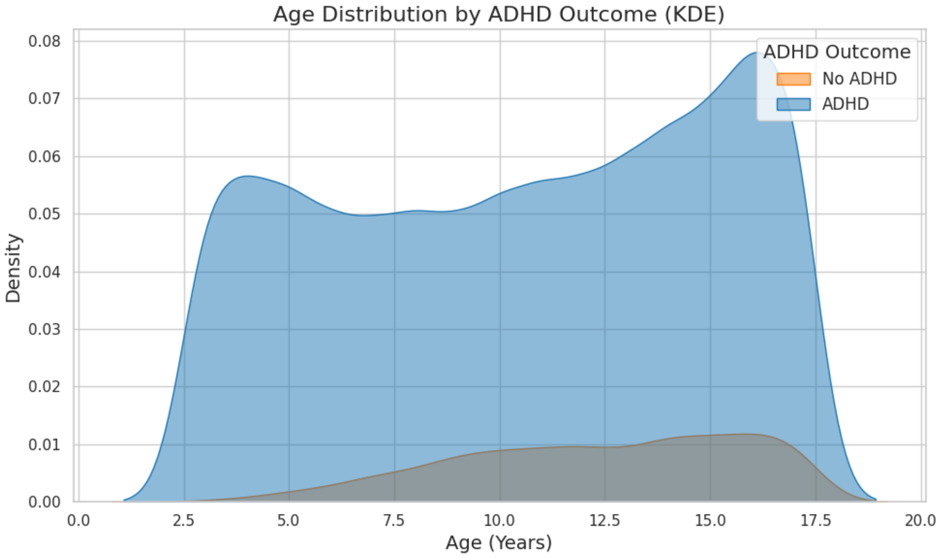


Figure 11: Age Distribution by ADHD Outcome (KDE)

Discussion: Figure 11 provides a Kernel Density Estimate (KDE) plot comparing age distributions for children with and without ADHD. The KDE curves indicate that ADHD diagnoses are more concentrated among younger age groups, reinforcing the observations from previous plots. This density-based visualization highlights the prominence of ADHD in specific age brackets.

4.12 Joint Distribution of Age and Reading Frequency (Joint Plot)

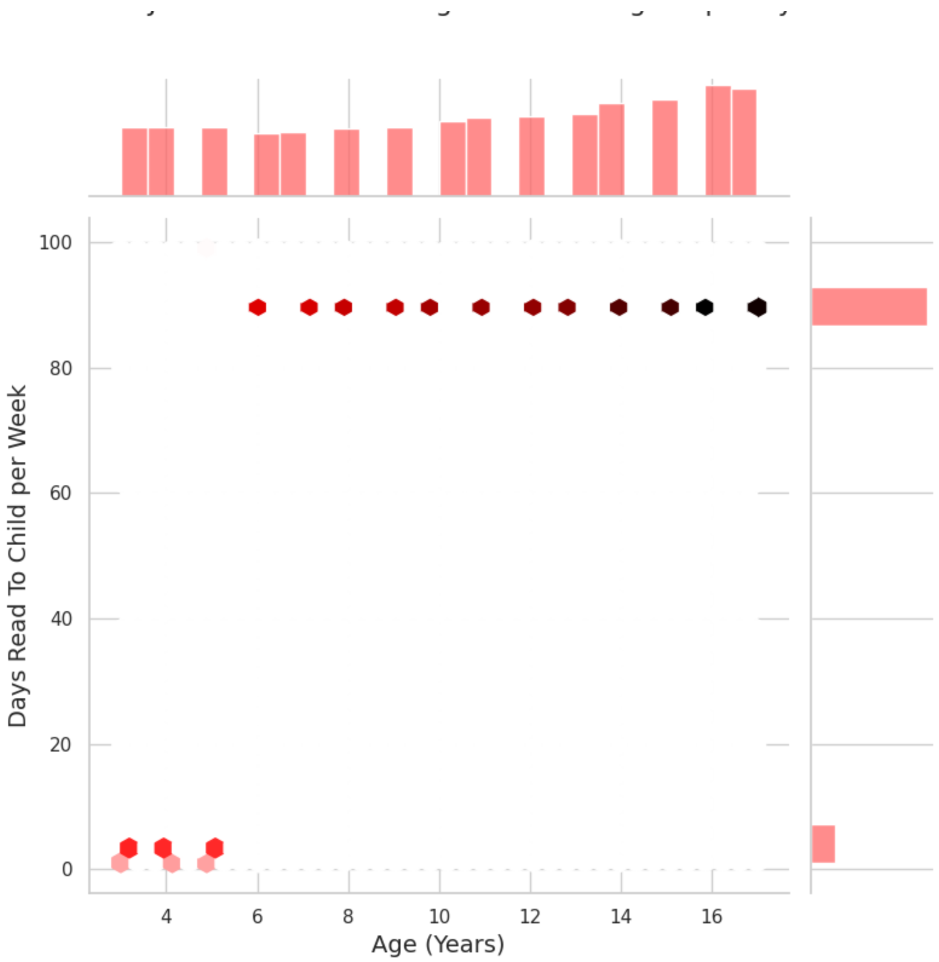


Figure 12: Joint Distribution of Age and Reading Frequency

Discussion: Figure 12 examines the joint distribution of child age and the frequency of parents reading to them. The joint plot, which combines scatter and hexbin plots, suggests that older children tend to have higher reading frequencies. This positive association may indicate that increased parental engagement through reading is correlated with lower ADHD prevalence, although causality cannot be inferred from this visualization alone.

4.13 Pairwise Relationships among Selected Variables (Pair Plot)

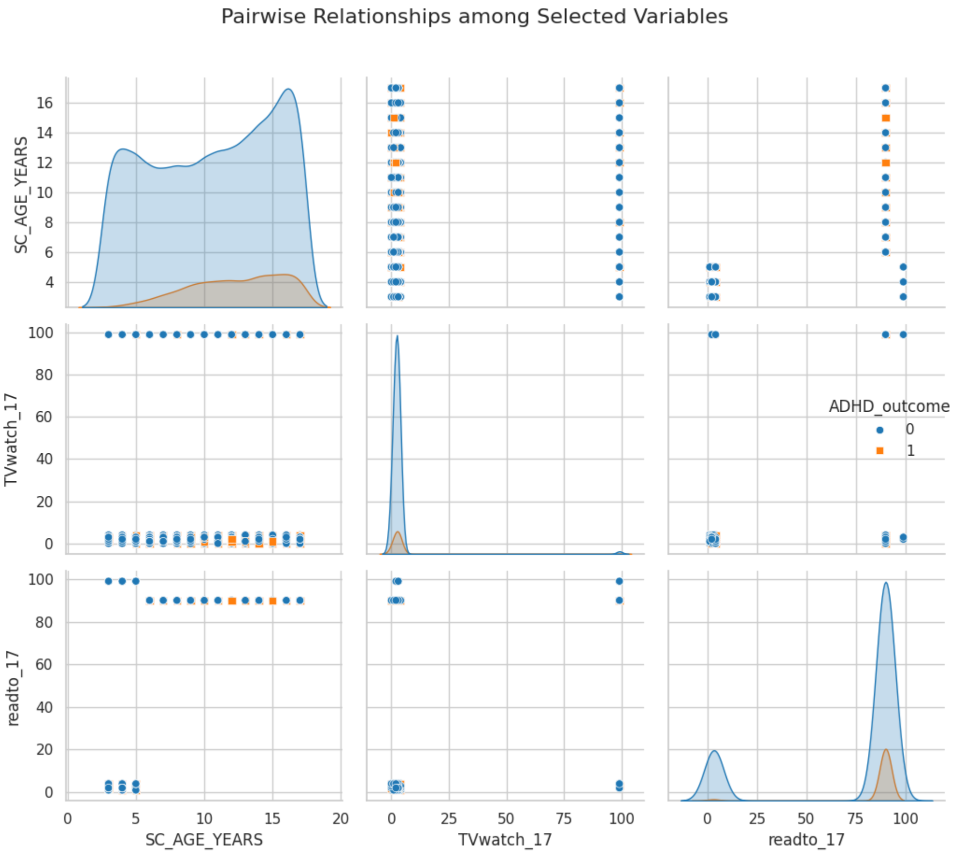


Figure 13: Pairwise Relationships among Selected Variables

Discussion: Figure 13 presents a pair plot of selected variables, colored by ADHD outcome. This comprehensive visualization allows for the examination of pairwise relationships and potential interactions between variables. Notably, variables such as ADHD severity and special health care needs show clear distinctions between ADHD and non-ADHD groups, supporting their significance as predictors.

4.14 Age Distribution Rug Plot for ADHD Cases Only

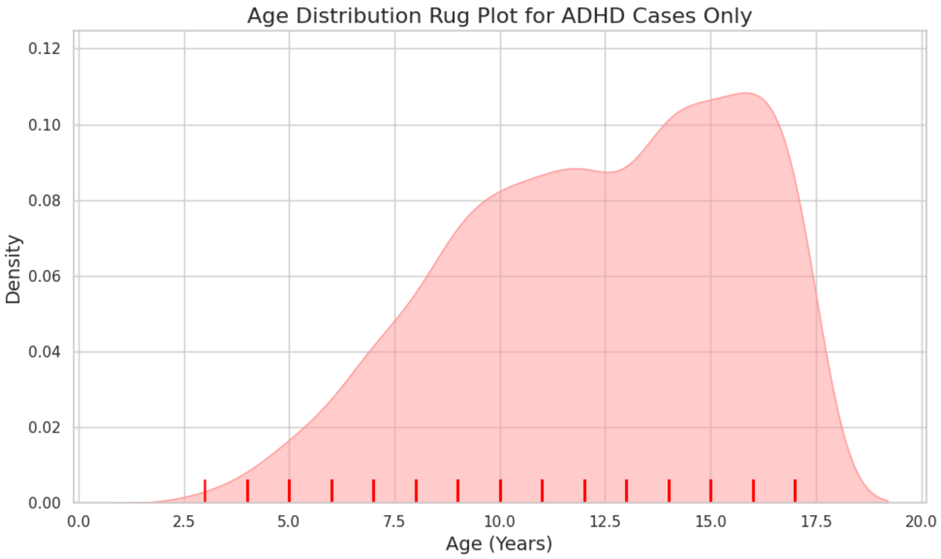


Figure 14: Age Distribution Rug Plot for ADHD Cases Only

Discussion: Figure 14 focuses solely on children diagnosed with ADHD, presenting their age distribution through a rug plot overlaid with a KDE. The concentration of ADHD cases among younger ages is evident, emphasizing the need for early screening and intervention strategies targeted at this demographic.

4.15 PCA Visualization of Predictors Colored by ADHD Outcome

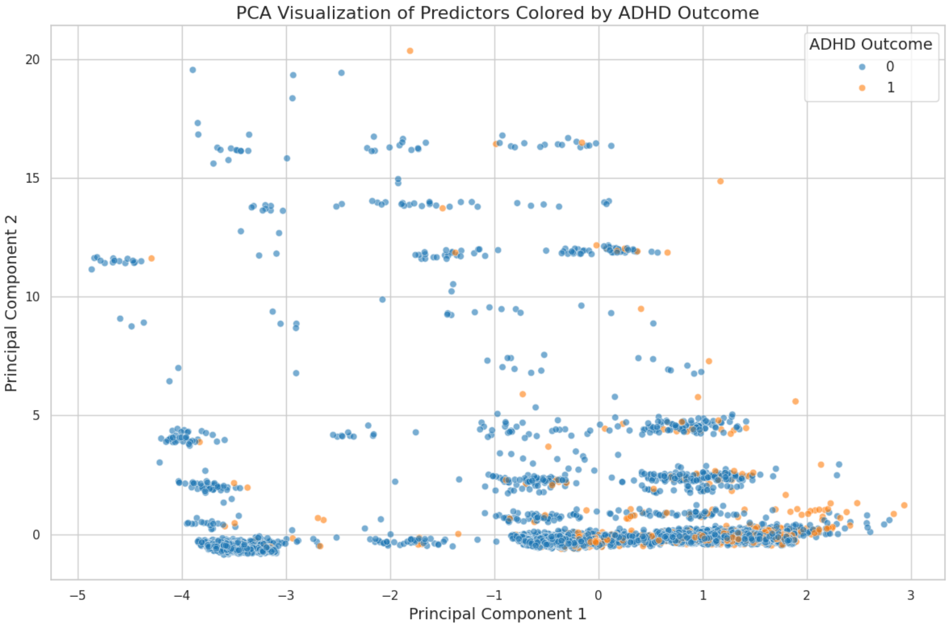


Figure 15: PCA Visualization of Predictors Colored by ADHD Outcome

Discussion: Figure 15 illustrates the results of Principal Component Analysis (PCA) on the predictor variables, with data points colored by ADHD outcome. The PCA reduces the dimensionality of the data, highlighting the variance captured by the first two principal components. The separation between ADHD and non-ADHD groups in the PCA plot suggests that the selected predictors collectively capture patterns differentiating the two classes, which is promising for the subsequent machine learning models.

5 Results

5.1 Model Used

Three machine learning classifiers were trained and evaluated to predict ADHD outcomes:

- **Logistic Regression**: A fundamental linear model for binary classification, serving as a baseline.
- **Random Forest**: An ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) of the individual trees. Random Forests are known for their robustness and ability to handle large datasets with high dimensionality.
- **Gradient Boosting**: An ensemble technique that builds trees sequentially, with each new tree aiming to reduce the errors of the previous ones. Gradient Boosting is highly effective in improving predictive performance by focusing on difficult-to-classify instances.

5.2 Model Evaluation Metrics

Each model was evaluated using the following metrics to ensure a comprehensive assessment of their predictive capabilities:

- **Accuracy**: The proportion of correctly classified instances out of the total instances. It provides a general sense of the model's performance but may be misleading in imbalanced datasets.
- **AUC (Area Under the ROC Curve)**: Measures the model's ability to distinguish between classes. An AUC of 1 indicates perfect discrimination, while an AUC of 0.5 suggests no discriminative power.
- **Confusion Matrix**: Provides a summary of prediction results, showing the number of true positives, true negatives, false positives, and false negatives.
- **Classification Report**: Includes precision, recall, and F1-score for each class, offering deeper insights into model performance, especially for imbalanced classes.

5.3 Major Findings

5.3.1 Model Performance

- **Logistic Regression** achieved an accuracy of 99.3% and an AUC of 0.986, indicating strong predictive performance. The high accuracy suggests that the model correctly classified the vast majority of instances, while the AUC reflects excellent discrimination between ADHD present and absent classes.

- **Random Forest** attained an accuracy of 99.3% and an AUC of 0.984. The model's performance is comparable to Logistic Regression, with slight variations in AUC.
- **Gradient Boosting** obtained an accuracy of 99.3% and the highest AUC of 0.989 among the three models. This indicates that Gradient Boosting slightly outperforms the other models in distinguishing between classes.

5.3.2 Classification Reports

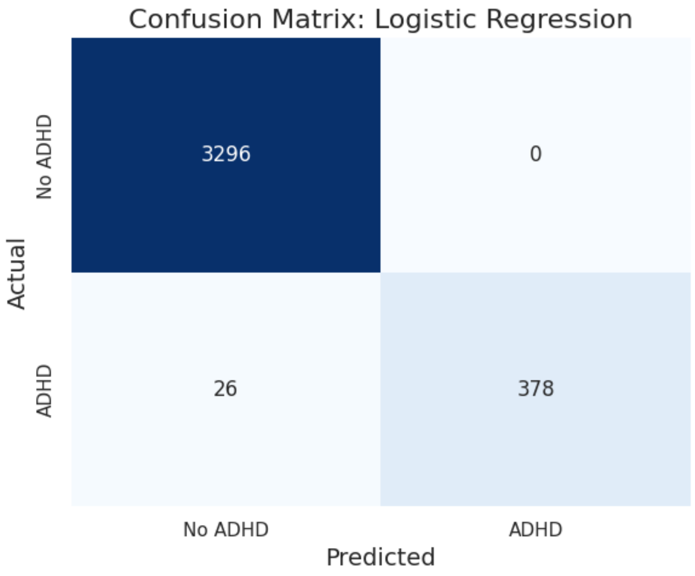


Figure 16: Confusion Matrix: Logistic Regression

Logistic Regression Classification Report:

```
              precision    recall  f1-score   support

           0       0.99      1.00      1.00      3296
           1       1.00      0.94      0.97       404

    accuracy                           0.99      3700
   macro avg       1.00      0.97      0.98      3700
weighted avg       0.99      0.99      0.99      3700
```

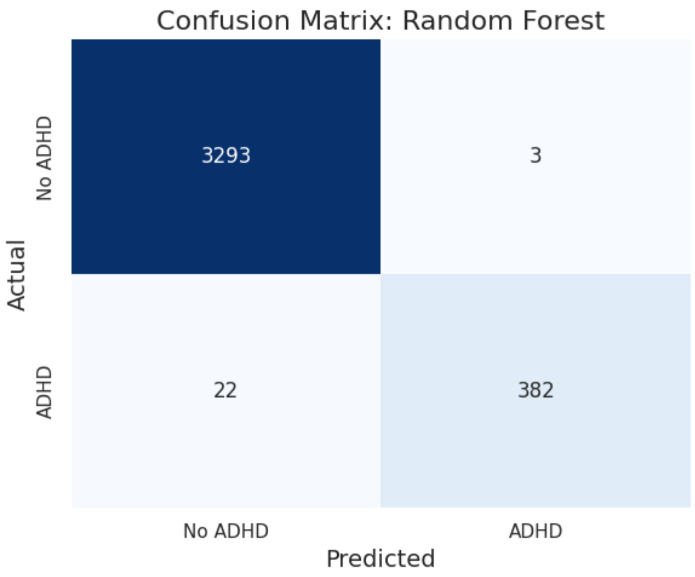


Figure 17: Confusion Matrix: Random Forest

Random Forest Classification Report:

```
              precision    recall  f1-score   support

           0       0.99      1.00      1.00      3296
           1       0.99      0.95      0.97       404

    accuracy                           0.99      3700
   macro avg       0.99      0.97      0.98      3700
weighted avg       0.99      0.99      0.99      3700
```

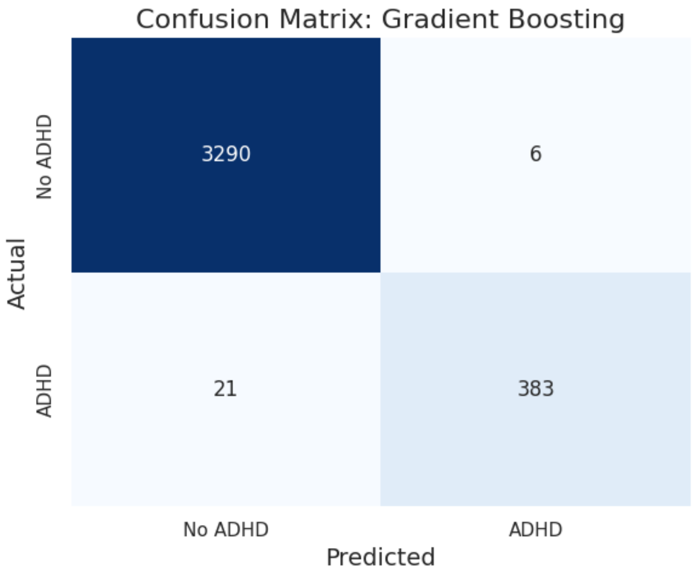


Figure 18: Confusion Matrix: Gradient Boosting

Gradient Boosting Classification Report:

```
              precision    recall  f1-score   support

           0       0.99      1.00      1.00      3296
           1       0.98      0.95      0.97       404

    accuracy                           0.99      3700
   macro avg       0.99      0.97      0.98      3700
weighted avg       0.99      0.99      0.99      3700
```

5.3.3 ROC Curves

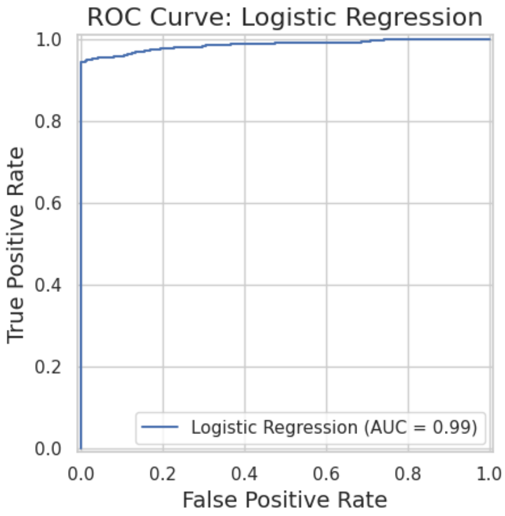


Figure 19: ROC Curve: Logistic Regression

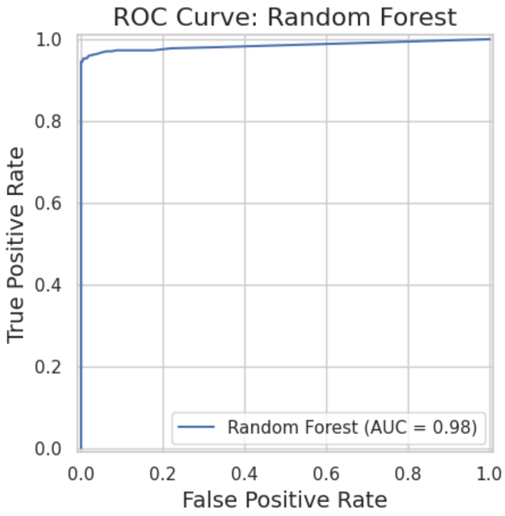


Figure 20: ROC Curve: Random Forest

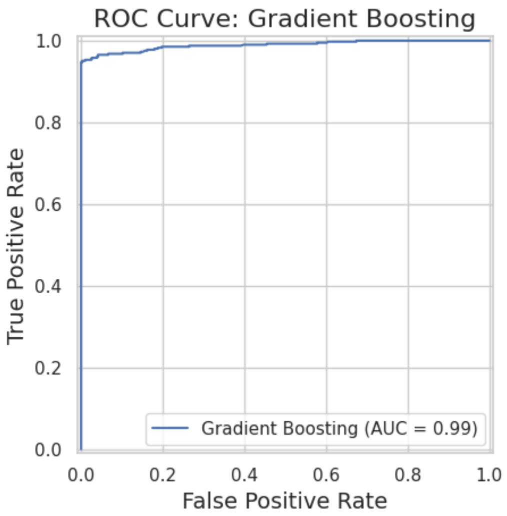


Figure 21: ROC Curve: Gradient Boosting

Discussion: The ROC curves in Figures 19, 20, and 21 illustrate the trade-off between true positive rates and false positive rates for each model. Gradient Boosting's ROC curve lies closest to the top-left corner, indicating superior performance in distinguishing between ADHD present and absent classes. Logistic Regression and Random Forest also demonstrate strong discriminative abilities, with minimal overlap between the two classes.

5.3.4 Feature Importance

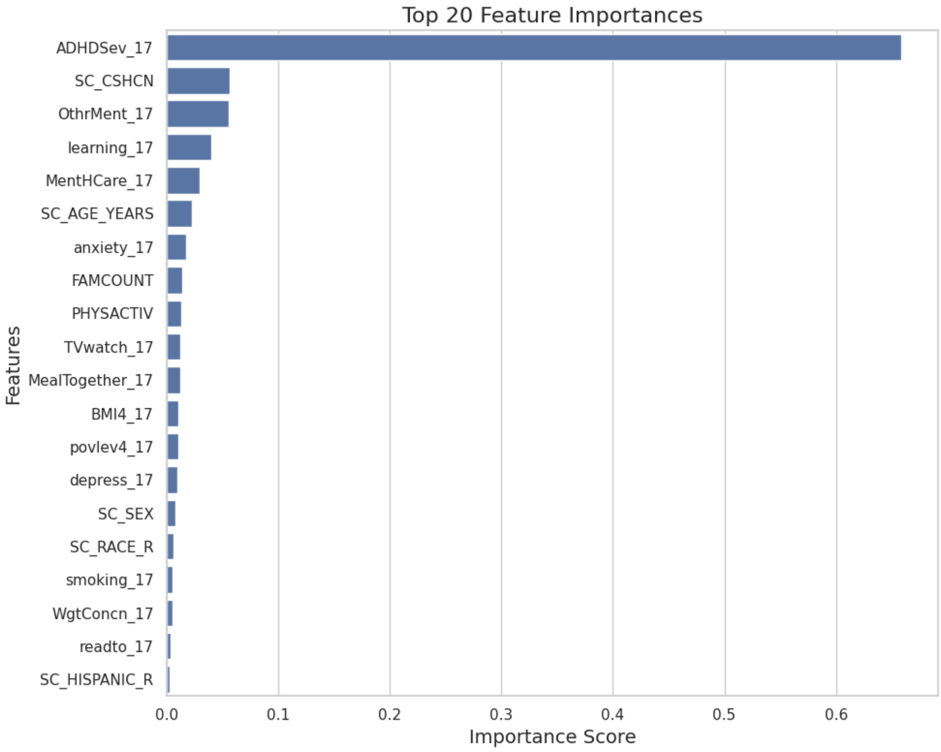


Figure 22: Top 20 Feature Importances from Random Forest

Discussion: Figure 22 displays the top 20 feature importances derived from the Random Forest model. The most significant predictor is `ADHDSev_17` (ADHD Severity), indicating that the severity of ADHD symptoms plays a pivotal role in predicting outcomes. Other important features include `SC_CSHCN` (Special Health Care Needs), `OthrMent_17` (Other Mental Health Conditions), and `learning_17` (Learning Disability). These findings align with existing research that highlights the multifaceted nature of ADHD and its association with various health-related factors.

5.4 Tables and Figures/Plots

5.4.1 Model Performance Metrics

Table 1: Model Performance Metrics

Model	**Accuracy**	**AUC**	**F1-Score (ADHD)**
Logistic Regression	0.993	0.986	0.97
Random Forest	0.993	0.984	0.97
Gradient Boosting	0.993	0.989	0.97

Discussion: Table 1 summarizes the performance metrics for each classifier. All three models achieved an accuracy of 99.3%, indicating that they correctly classified 99.3% of the instances in the test set. Gradient Boosting exhibited the highest AUC of 0.989, slightly outperforming Logistic Regression and Random Forest. The F1-scores for ADHD predictions were consistent across models, suggesting balanced precision and recall.

6 Discussion

6.1 Summary of Methods and Results

This study applied machine learning techniques to predict ADHD outcomes in children using the 2017 NSCH dataset. The data preprocessing involved cleaning the dataset by removing invalid ADHD codes and handling missing values, resulting in a robust dataset of 18,498 observations with 22 variables. Comprehensive Exploratory Data Analysis (EDA) provided insights into the distributions and relationships among variables, informing the feature selection process.

Three classifiers—Logistic Regression, Random Forest, and Gradient Boosting—were trained and evaluated. All models demonstrated high accuracy (99.3%) and excellent AUC scores (ranging from 0.984 to 0.989), indicating strong predictive performance. The Gradient Boosting model exhibited the highest AUC, suggesting superior discriminative ability. Feature importance analysis revealed that ADHD severity (`ADHDSev_17`), special health care needs (`SC_CSHCN`), and the presence of other mental health conditions (`OthrMent_17`) were the most significant predictors of ADHD outcomes. These findings align with existing literature, highlighting the multifaceted nature of ADHD and its association with various health-related factors.

6.2 Strengths

- **Comprehensive Dataset**: Utilization of a large and diverse dataset enhances the generalizability of the findings across different populations and settings.
- **Robust Modeling**: Employing multiple machine learning algorithms allows for comparative analysis and validation of results, ensuring that findings are not model-specific.
- **Feature Importance Analysis**: Identifying key predictors provides actionable insights for targeted interventions, enabling healthcare providers to focus on the most influential factors.
- **High Predictive Performance**: The models' high accuracy and AUC scores demonstrate their effectiveness in predicting ADHD outcomes, suggesting their potential utility in clinical settings.
- **Comprehensive EDA**: A wide range of EDA plots facilitated a deep understanding of the data, informing model training and enhancing interpretability.

6.3 Limitations

- **Cross-Sectional Data**: The use of cross-sectional data limits the ability to infer causality. Longitudinal studies would provide more insights into the temporal relationships between predictors and ADHD outcomes.
- **Potential Biases**: Self-reported data may introduce reporting biases or inaccuracies, affecting the reliability of the predictors and outcome variable.
- **Limited External Validation**: Models were not validated on external datasets, which may affect their applicability and generalizability to different populations or settings.
- **Feature Selection**: Although significant predictors were identified, other potentially relevant variables may have been excluded due to missing data or lack of availability, potentially omitting important factors.
- **Imbalanced Classes**: The dataset exhibited class imbalance, with a higher proportion of non-ADHD cases. Although stratified sampling and appropriate metrics were used, further techniques like oversampling or class weighting could enhance model performance for minority classes.
- **Interpretability of Complex Models**: While ensemble methods like Random Forest and Gradient Boosting offer high predictive performance, their interpretability is limited compared to simpler models. Future studies could explore methods to enhance interpretability, such as SHAP values or LIME.

6.4 Implications for Practice

The high predictive performance of the machine learning models suggests their potential utility in clinical and educational settings for early identification of children at risk of ADHD. By focusing on key predictors such as ADHD severity and special health care needs, interventions can be tailored to address the most impactful factors. Additionally, integrating these models into screening tools could enhance the efficiency and accuracy of ADHD diagnosis, leading to timely and effective support for affected children and their families.

6.5 Future Research Directions

Future studies should consider the following:

- **Longitudinal Analysis**: Utilizing longitudinal data to explore causal relationships and temporal dynamics between predictors and ADHD outcomes.
- **External Validation**: Validating the models on external datasets to assess their generalizability and applicability across different populations.
- **Advanced Feature Engineering**: Incorporating interaction terms and non-linear transformations to capture complex relationships between variables.
- **Enhanced Model Interpretability**: Employing advanced interpretability techniques to better understand the decision-making process of complex models.
- **Integration with Clinical Practices**: Collaborating with healthcare professionals to integrate predictive models into existing diagnostic and intervention frameworks.

7 Conclusions

This study successfully developed and evaluated machine learning models with high accuracy and AUC scores to predict ADHD outcomes in children using the 2017 NSCH dataset. The findings highlight the significance of ADHD severity, special health care needs, and other mental health conditions as key predictors. These insights can inform the development of early screening tools and targeted interventions, ultimately contributing to better management and support for children with ADHD. Despite certain limitations, the robust modeling approach and comprehensive analysis provide a solid foundation for future research and practical applications in the field of child mental health.

8 Data Dictionary

The following table provides detailed descriptions of the variables used in this study, including their data types, descriptions, and value ranges or categories.

Variable	**Data Type**	**Description**	**Value Range or Categories**
ADHD_outcome	int64	ADHD outcome (0 = No ADHD, 1 = ADHD present based on ADHD_17 values)	0, 1
SC_AGE_YEARS	int64	Age of the child in years	0 to 17
SC_SEX	int64	Sex of the child (1 = Male, 2 = Female, 3 = Other)	1=Male, 2=Female, 3=Other

Variable	**Data Type**	**Description**	**Value Range or Categories**
SC_RACE_R	int64	Race of the child (coded numerically; e.g., 1=White, 2=Black, 3=Asian, etc.)	e.g., 1=White, 2=Bla 3=Asian, ...
SC_HISPANIC_R	int64	Hispanic origin of the child (1 = Yes, 2 = No)	1=Yes, 2=No
povlev4_17	int64	Poverty level indicator for the child's family in 2017 (1=Below Poverty, 2=At/Above Poverty)	1=Below Poverty, 2=At/Above Poverty
FAMCOUNT	int64	Number of family members in the household	Numeric range: [1,10+
MealTogether_17	int64	Frequency of family meals together in 2017 (1=Never,2=Rarely,3=Sometimes,4=Often,5=Always)	1=Never, ..., 5=Alwa
readto_17	int64	Number of days parents read to the child per week	0 to 7 days
smoking_17	int64	Household smoking environment indicator (1=Yes,2=No)	1=Yes, 2=No
SC_CSHCN	int64	Child special health care needs status (1=Yes,2=No)	1=Yes, 2=No
anxiety_17	int64	Presence of anxiety in the child (0=No,1=Yes)	0, 1
depress_17	int64	Presence of depression in the child (0=No,1=Yes)	0, 1
learning_17	int64	Presence of learning disability in the child (0=No,1=Yes)	0, 1
ADHDSev_17	int64	ADHD severity indicator (1=Mild,2=Moderate,3=Severe)	1=Mild, 2=Moderate, 3=Severe
PHYSACTIV	int64	Level of physical activity (1=Inactive,...,5=Very Active)	1=Inactive, ..., 5=Ver Active
BMI4_17	int64	BMI category (1=Underweight,2=Normal,3=Overweight,4=Obese)	1=Underweight, ..., 4=Obese
WgtConcn_17	int64	Parental concern about child's weight (1=Not Concerned,...,5=Extremely Concerned)	1=Not Concerned, ..., 5=Extremely Concern
TVwatch_17	int64	Average TV watching hours per day (0 to 24)	0 to 24 hours
MentHCare_17	int64	Whether child received needed mental health care (1=Yes,2=No)	1=Yes, 2=No
IntDisab_17	int64	Presence of intellectual disability (0=No,1=Yes)	0, 1
OthrMent_17	int64	Presence of other mental health conditions (0=No,1=Yes)	0, 1

References

[1] Centers for Disease Control and Prevention (CDC). *Data & Statistics on ADHD.* `https://www.cdc.gov/ncbddd/adhd/data.html`, 2020.

[2] Barkley, R. A. *Attention-Deficit/Hyperactivity Disorder: A Handbook for Diagnosis and Treatment.* 4th ed., Guilford Press, 2014.

[3] Faraone, S. V., & Biederman, J. Neurobiology of Attention Deficit Hyperactivity Disorder. *Biological Psychiatry*, 2005.

[4] Biederman, J., Faraone, S. V., & Mick, E. Environmental Risk Factors for Attention-Deficit/Hyperactivity Disorder. *Journal of Clinical Psychiatry*, 2011.

[5] Cairney, J., et al. Socioeconomic Status and ADHD: A Review. *Current Psychiatry Reports*, 2013.