**Binary Classification Tree based machine learning model to analyze the associations between adults with ADHD and co-occurring conditions, specifically anxiety and substance abuse.**

**Abstract**

Attention-Deficit/Hyperactivity Disorder (ADHD) is a diverse chronic condition affecting nearly three percent of the adult population in the world. The disorder often severely impacts social and occupational functioning, as quality of life for those affected. Some of the most common symptoms are hyperactivity, impulsivity, and inattention, which are present during social, academic, or occupational activities. Adults with ADHD may display symptoms of impulsivity and hyperactivity, including talkativeness, restlessness, and a proclivity for making impulsive decisions without considering potential consequences.

[HYPERAKTIV](https://osf.io/2tk4r) is a public dataset containing health, activity, and heart rate data from adult patients diagnosed with attention deficit hyperactivity disorder. The dataset consists of data collected from 51 adult patients with ADHD and 52 adults for clinical-controls. A total of 103 adult patients.

As a result, using Python for Data Science and Machine Learning, this project Binary Classification Tree model, presents an analysis of the associations between adults with ADHD and co-occurring conditions such as anxiety and substance abuse. Understanding the age range of the individuals in the dataset to gain insights into their interrelationships and potential predictive factors.

More feature engineering and research for predictive features can have a positive impact on model quality, in the future iterations of this project. It is advised to perform cross-validation to evaluate the models' generalization capability and prevent overfitting. Both the Random Forest and AdaBoost models show limitations in generalizing to new data, indicating potential overfitting and underfitting. To improve the models' performance in predicting anxiety based on age and ADHD, further refinement through feature engineering, hyperparameter tuning, or additional data variables is recommended.

Additionally, increasing the size and diversity of the training data and exploring different models or ensemble methods may be necessary to enhance the models' predictive power. Findings at a high-level include:

Two models were created:

* Model 1

|  |  |
| --- | --- |
| **X = Predictor variables** | **Y = variables being predicted, TARGET VARIABLE** |
| **AGE**: Participant ages are presented in four groups, where (1) = 17-29 years, (2) = 30-39 years, (3) = 40-49 years and (4) = 50-67 years. (AGE Qualitative>Ordinal variable) | **SUBSTANCE**: drug, alcohol, addictions. not present (0), present (1), unknown (9), or control (2). (Qualitative>Nominal variable) |
| **ADHD**: General presence of ADHD n=51. not present (0), present (1), or unknown (9). (ADHD Qualitative>Nominal>Categorical variable) |

* Model 2

|  |  |
| --- | --- |
| **X = Predictor variables** | **Y = variables being predicted, TARGET VARIABLE** |
| **AGE**: Participant ages are presented in four groups, where (1) = 17-29 years, (2) = 30-39 years, (3) = 40-49 years and (4) = 50-67 years. (AGE Qualitative>Ordinal variable) | **ANXIETY**: not present (0), present (1), unknown (9), or control (2). (Qualitative>Nominal variable). |
| **ADHD**: General presence of ADHD n=51. not present (0), present (1), or unknown (9). (ADHD Qualitative>Nominal>Categorical variable) |

For SUBSTANCE ABUSE, both the AdaBoostClassifier and RandomForestClassifier models achieved an accuracy of 81.48% and F1 Score of 73.1%. Further model refinement, feature selection, hyperparameter tuning, and exploring ensemble methods could be beneficial to address misclassifications.

For ANXIETY the model's F1 Score (weighted average) of 0.5213 suggests a moderate overall predictive performance. Accuracy, 59.2%. However, the low F1-scores for both training and testing data indicate a need for improvement.

| **Baseline Model Measurements – Decision Tree Classifier Binary**  **ADHD and AGE** | |
| --- | --- |
| **Model 1 Substance** | **Model 2 Anxiety** |
| Training Time: 0.0030031204223632812 seconds  Prediction Time: 0.000995635986328125 seconds | Training Time: 0.004212379455566406 seconds  Prediction Time: 0.001508951187133789 seconds |
| Accuracy: 0.8148148148148148 | Accuracy: 0.5925925925925926 |
| Confusion Matrix: [[22 0] [ 5 0]] | Confusion Matrix: [[13 0] [11 3]] |
| F1 Score (weighted average): 0.7316704459561602  F1-score training is 0.0  F1-score testing is 0.0 | F1 Score (weighted average): 0.5213448742860508  F1-score training is 0.19230769230769232  F1-score testing is 0.35294117647058826 |

|  |  |
| --- | --- |
| **Updated Model Measurements – Decision Tree Classifier Binary**  **ADHD and AGE** | |
| **Model 1 Substance** | **Model 2 Anxiety** |
| Adaboost Classifier  Accuracy: 81.48%  Confusion Matrix: [[22 0] [ 5 0]]  Best Parameters: {'learning\_rate': 0.01, 'n\_estimators': 10}  Best Estimator: AdaBoostClassifier (learning\_rate=0.01, n\_estimators=10) | No further algorithms were applied to ANXIETY. The model's F1 Score (weighted average) of 0.5213 suggests a moderate overall predictive performance. However, the low F1-scores for both training and testing data indicate a need for improvement. |
| Random Forest  Accuracy: 81.48%  Confusion Matrix: [[22 0] [ 5 0]]  Best Parameters: {'max\_depth': 10, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 100}  Best Estimator: RandomForestClassifier(max\_depth=10) |  |

**Introduction**

Adult ADHD is a neuropsychiatric disorder characterized by symptoms of hyperactivity, impulsivity, and inattention, which must persist over time and impact social, academic, or occupational functioning for diagnosis. In adults, inattention, disorganization, distraction, and impulsivity are prominent features, often leading to difficulties in staying focused and making impulsive decisions. Unlike in children, ADHD in adults shows a more equal gender distribution. Stimulants are the primary medications for ADHD, but they carry a risk of misuse and dependency. Additionally, adult ADHD is associated with an increased risk of substance use disorders and often coexists with other psychiatric conditions like mood and anxiety disorders, as well as sleeplessness.

The HYPERAKTIV dataset used for this analysis is open source and contains health, activity, and heart rate data from 51 adult patients with ADHD and 52 clinical controls, totaling 103 adult patients. ADHD affects nearly five percent of the adult population, significantly impacting social and occupational functioning and quality of life. Currently, ADHD diagnosis relies on subjective evaluation and clinical observations, highlighting the need for more objective methods.

For the purpose of this project a Binary Classification Tree based machine learning model was created to analyze the associations between adults with ADHD and co-occurring conditions, specifically anxiety and substance abuse, while seeking to understand the age range of the individuals in the dataset to gain insights into their interrelationships and potential predictive factors. More feature engineering and research for predictive features can have a positive impact on model quality in the future iterations of this project.

Previous studies from 2020 to present have indicated a high comorbidity of adult ADHD with anxiety, major depressive disorder, bipolar disorder, and substance use disorders, highlighting the absence of gender-specific modifications and emphasizing the prevalence of comorbid conditions in both men and women with ADHD.

The application of algorithms, including AdaBoostClassifier and RandomForestClassifier, provided similar results to the original baseline model, with an accuracy of 81.48% and the same confusion matrix [[22 0] [ 5 0]]. Further model refinement, feature selection, hyperparameter tuning, and exploring ensemble methods could be beneficial to address misclassifications and improve predictive capabilities.

Data clean-up efforts involved handling unknown and control values, resulting in a dataset of 133 patients.

These findings have implications for guiding further research, intervention, and support strategies for individuals affected by ADHD, anxiety, and substance abuse.

**Related Work**

Based on the provided web search results, the research conducted from 2020 to the present on ADHD, anxiety, and substance abuse it has led to significant findings.

The studies have revealed a high comorbidity of adult ADHD with anxiety, major depressive disorder, bipolar disorder, and substance use disorders. The research has also highlighted the absence of gender-specific modifications, emphasizing the prevalence of comorbid conditions in both men and women with ADHD. Additionally, the effectiveness of ADHD treatment in reducing the risk of comorbidity has been suggested as a potential area for further investigation. The papers contributing to this research include:

1. "Adult ADHD: a new disease?" by Zalsman G, Shilton T. Int J Psychiatry Clin Pract. 2016
2. "ADHD in children and young people: prevalence, care pathways, and service provision" by Sayal K, Prasad V, Daley D, Ford T, Coghill D. Lancet Psychiatry. 2018
3. "Attention-deficit/hyperactivity disorder: diagnostic criteria, epidemiology, risk factors and evaluation in youth" by Cabral MD, Liu S, Soares N. Transl Pediatr. 2020
4. "Trends in the prevalence and incidence of attention-deficit/hyperactivity disorder among adults and children of different racial and ethnic groups" by Chung W, Jiang SF, Paksarian D, Nikolaidis A, Castellanos FX, Merikangas KR, Milham MP. JAMA Netw Open. 2019
5. "ADHD: current concepts and treatments in children and adolescents" by Drechsler R, Brem S, Brandeis D, Grunblatt E, Berger G, Walitza S. Neuropediatrics. 2020

**Methodology**

**Dataset**

HYPERAKTIV is a public dataset located in the [Kaggle](https://www.kaggle.com/datasets/arashnic/adhd-diagnosis-data) website containing health, activity, and heart rate data from adult patients diagnosed with attention deficit hyperactivity disorder, better known as ADHD. A total of 103 patients were recruited, 51 of which were diagnosed with ADHD, and 52 with other diagnoses (clinical controls). The Norwegian Regional Medical Research Ethics Committee West approved the original protocol for the data collection, and all processes were in accordance with the Helsinki Declaration of 1975.

The data collected were recordings of motor activity and heart rate, the output of a computerized test of attention-related problems, as well as various diagnostic and clinical assessments. A total of thirty-one data points or variables are collected into four main files:

1) Activity data contains the activity measurements from all participants, each organized into separate files with metadata at the beginning.

2)HRV\_data holds the heart rate data from each participant, similarly organized into individual files with metadata preceding the data.

3) The hyperaktiv\_with\_controls data is also present.

4) CPT\_II\_ConnersContinuousPerformanceTest.csv file contains individual responses to CPT-II test trials, and the file names features.csv contains pre-extracted features for the experiments.

5) patient\_info.csv file includes various participant attributes such as age, sex, and mental state information, along with output data from a neuropsychological test.

For the purpose of this project file number 5) was used to apply Python for Data Science and Machine Learning, selecting the following variables to be applied to two models:

* Model 1

|  |  |
| --- | --- |
| X = Predictor variables | Y = variables being predicted, TARGET VARIABLE |
| AGE: Participant ages are presented in four groups, where (1) = 17-29 years, (2) = 30-39 years, (3) = 40-49 years and (4) = 50-67 years. (AGE Qualitative>Ordinal variable) | SUBSTANCE: drug, alcohol, addictions. not present (0), present (1), unknown (9), or control (2). (Qualitative>Nominal variable) |
| ADHD: General presence of ADHD n=51. not present (0), present (1), or unknown (9). (ADHD Qualitative>Nominal>Categorical variable) |

* Model 2

|  |  |
| --- | --- |
| X = Predictor variables | Y = variables being predicted, TARGET VARIABLE |
| AGE: Participant ages are presented in four groups, where (1) = 17-29 years, (2) = 30-39 years, (3) = 40-49 years and (4) = 50-67 years. (AGE Qualitative>Ordinal variable) | ANXIETY: not present (0), present (1), unknown (9), or control (2). (Qualitative>Nominal variable). |
| ADHD: General presence of ADHD n=51. not present (0), present (1), or unknown (9). (ADHD Qualitative>Nominal>Categorical variable) |

Data clean up was executed to transfer the (9) unknown values from the variables, ADHD, SUBSTANCE and ANXIETY to (0) not present. Also, (2) which are control values were transferred to (0) for not present ADHD, SUBATANCE or ANXIETY.

Data clean-up efforts involved handling unknown (9) and control values (0), resulting in a dataset of 133 patients.

**Exploratory Data Analysis (EDA)**

* ADHD and AGE

From 133 individuals in the dataset, a total of 82 individuals do not have a diagnosis of ADHD, and 51 do have ADHD.

The ADHD mean is 0.383459, indicating that 38% of the individuals in the data set have ADHD in the age group (2) = 30-39 years old.

From 133 individuals, the great majority of individuals are in the group (1) and (2) with 74 in total.

1. 17-29 years = 38
2. 30-39 years = 37
3. 40-49 years = 35
4. 50-67 years = 24

* SUBSTANCE and ANXIETY

From 133 individuals, ANXIETY is not present in 76. In contrast, 57 individuals have ANXIETY. SUBSTANCE is not present in 109 individuals, but it is present in 24 individuals.

If combining the results 81 individuals have ANXIETY and SUBSTANCE.

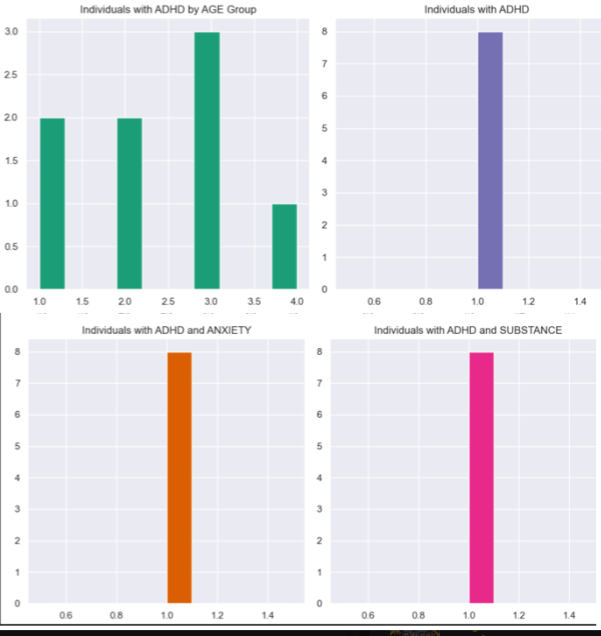
The mean for ANXIETY is 0.428571, indicating that 42% of individuals have anxiety and 18% have substance abuse. However, it is not necessarily, because of a diagnosis of ADHD.

In executing the Chi-Square for categorical variables, in this case: (AGE and ANXIETY), (AGE and SUBSTANCE) the following findings were relevant:

| **AGE and ANXIETY** | **AGE and SUBSTANCE** |
| --- | --- |
| **Ho = There is no significant association between adults AGE and ANXIETY in adults.** | **Ho = There is no significant association between AGE and SUBSTANCE ABUSE in adults.** |
| Chi-squared statistic: 0.9069256756756758 indicates is no significant association between the AGE and ANXIETY.  P-value: 0.8237561705191967 suggests that the observed data is likely under the assumption of the null hypothesis, indicating a lack of evidence against the null hypothesis.  The results indicate that the variables are likely independent of each other for AGE and ANXIETY.  Degrees of freedom: 3 number of categories minus 1. In this case, it suggests that there are 3 degrees of freedom is no significant association between the AGE and ANXIETY, as suggested by the high p-value and low chi-squared statistic. | Chi-squared statistic: 3.490771361958289 indicates a moderate magnitude of association  P-value: 0.32196101817316936 fail to reject the null hypothesis, 0.32 is greater than the typical significance level of 0.05. This suggests that there is not enough evidence to reject the null hypothesis, indicating a lack of association between the variables.  Degrees of freedom: 3 number of categories minus 1. In this case, it suggests that there are 3 degrees of freedom is no significant association between the AGE and SUBSTANCE. |

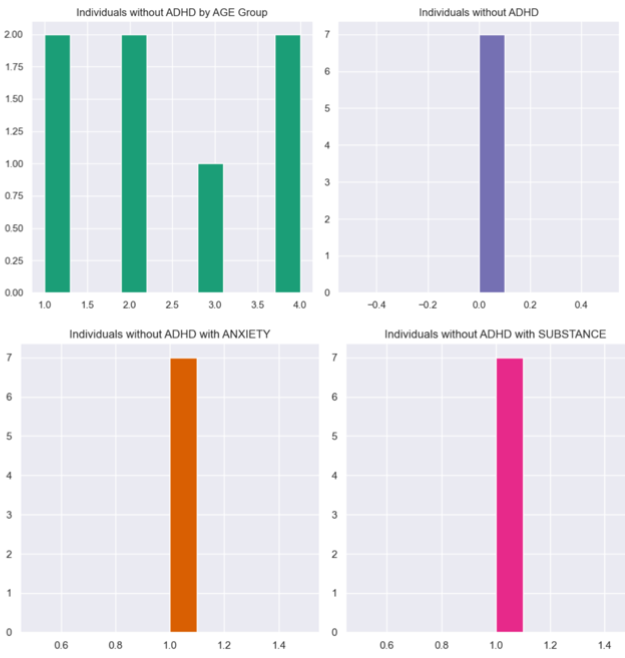
The histogram reveals that from a sample of 133 patients, 51 patients have ADHD, 8 has ANXIETY with 15.68% and 8 has a SUBSTANCE = 15.68%. As a result, 31.37% of individuals with ADHD have ANXIETY and SUBSTANCE abuse.

The age distribution is higher in group (3) with (40 to 49) years old.



From 133 sample, 82 individuals do not have ADHD, 7 has ANXIETY with 8.53% and 7 has SUBSTANCE abuse with 8.53%. As a result, individuals with out ADHD have a 17.07% of ANXIETY and SUBSTANCE.

The age distribution is higher in groups (1) with (17 to 29) years old (2) with (30 to 39) years old, (4) with (50 to 67 years old) years old.



It may be implied that there could be potential correlations or comorbidities between ADHD, and other variables that should be analyzed in future studies. These findings could warrant further investigation into the potential interrelationships and impacts of these conditions on the affected individuals.

These conclusions could guide further research, intervention, and support strategies for individuals affected by ADHD, ANXIETY, and SUBSTANCE abuse.

**Baseline Model**

The Train Test Split was created to evaluate the simulation of the data set. A total of 133 individuals exists in the cleaned data frame. A total of 106 individuals equivalent to 80% was allocated to the X\_Train and Y\_Train variables. In addition, 20% was allocated to the X\_Test and Y\_Test variables with 27 individuals.

|  |  |  |
| --- | --- | --- |
| **X\_Train** | **X\_Test** | **Total** |
| 106 | 27 | 133 |
| **Y\_Train** | **Y\_Test** | **Total** |
| 106 | 27 | 133 |

* Model 1 ADHD, AGE and SUBSTANCE

-Training Time: The result of 0.0030031204223632812 seconds indicates that the machine learning model is very efficient and can quickly learn from the given training data. It is possible that it is influence by the size of the data set for 133 samples.

-Prediction Time: The result of 0.000995635986328125 seconds indicates that the model is highly efficient, capable of quick training, and suitable for real-time or resource-constrained applications.

-Accuracy: is an intuitive and widely used metric, suitable for evaluating the overall performance of the benchmark model. The result was 0.8148148148 or 81.48% indicating that the base model predicts the Y target variable of SUBSTANCE.

-F1 Score: considers both precision and recall, providing a single value that balances both false positives and false negatives. The based model has a F1Score of 0.7316704459561602 or 73% indicates that the model's overall performance for precision and recall is reasonably good. The training and testing were 0.0 indicating that potential issues happen for overfitting.

-Precision and Recall: provide insights into the benchmark model's ability to correctly identify positive instances and avoid false positives.

-The confusion matrix describes the model's performance in terms of:

True positive (TP) 22 the model predicted 22 instances as positive

False Negative (FN) 0 There are no instances

False Positive (FP) 5 The model incorrectly predicted 5 instances as positive when they were actually negative.

True Negative (TN) 0 There are no instances

The model's performance seems to be heavily skewed towards predicting positive instances

Confusion Matrix

|  |  |  |
| --- | --- | --- |
|  | Condition Positive | Condition Negative |
| Test Outcome | True Positive  22 | False Negative  0 |
| False Negative  5 | True Negative  0 |

* Model 2 ADHD, AGE and ANXIETY

-Training Time: The result of 0.004212379455566406 seconds indicates that the machine learning model is very efficient and can quickly learn from the given training data.

-Prediction Time: The result of 0.001508951187133789 seconds indicates that the model is highly efficient, capable of quick training, and suitable for real-time or resource-constrained applications.

-Accuracy: the accuracy result was 0.5925925925925926 indicates that the model predicts the target variable approximately 59.25% of the time.

-F1 Score: the result was 0.5213448742860508 equivalent to 52% model's overall predictive performance is moderate. The training result was 19% indicating model's ability to predict anxiety based on AGE and ADHD in the training dataset is low. For the testing score of 35% indicates slightly improved, but still relatively low performance on the testing dataset.

-The confusion matrix describes the model's performance in terms of: true positive, true negative, false positive, and false negative predictions. The results are as following:

True Positive (TP): 13 - The number of instances that are actually positive (belong to the positive class) and are predicted as positive by the model.

False Negative (FN): 0 - The number of instances

False Positive (FP): 11 - The number of instances that are actually negative but are predicted as positive by the model.

True Negative (TN): 3 - The number of instances that are actually negative and are predicted as negative by the model.

Confusion Matrix

|  |  |  |
| --- | --- | --- |
|  | Condition Positive | Condition Negative |
| Test Outcome | True Positive  13 | False Negative  0 |
| False Negative  11 | True Negative  3 |

| **Baseline Model Measurements – Decision Tree Classifier Binary**  **ADHD and AGE** | |
| --- | --- |
| **Model 1 Substance** | **Model 2 Anxiety** |
| Training Time: 0.0030031204223632812 seconds  Prediction Time: 0.000995635986328125 seconds | Training Time: 0.004212379455566406 seconds  Prediction Time: 0.001508951187133789 seconds |
| Accuracy: | Accuracy: 0.5925925925925926 |
| Confusion Matrix: [[22 0] [ 5 0]] | Confusion Matrix: [[13 0] [11 3]] |
| F1 Score (weighted average): 0.7316704459561602  F1-score training is 0.0  F1-score testing is 0.0 | F1 Score (weighted average): 0.5213448742860508  F1-score training is 0.19230769230769232  F1-score testing is 0.35294117647058826 |

**Algorithms**

The following machine learning algorithms were used, to evaluate the model predictions for these variables:

* Model 1

|  |  |
| --- | --- |
| **X = Predictor variables** | **Y = variables being predicted, TARGET VARIABLE** |
| AGE: Participant ages are presented in four groups, where (1) = 17-29 years, (2) = 30-39 years, (3) = 40-49 years and (4) = 50-67 years. (AGE Qualitative>Ordinal variable | SUBSTANCE: drug, alcohol, addictions. not present (0), present (1), unknown (9), or control (2). (Qualitative>Nominal variable) |
| ADHD: General presence of ADHD n=51. not present (0), present (1), or unknown (9). (ADHD Qualitative>Nominal>Categorical variable) |

* Model 2

|  |  |
| --- | --- |
| **X = Predictor variables** | **Y = variables being predicted, TARGET VARIABLE** |
| AGE: Participant ages are presented in four groups, where (1) = 17-29 years, (2) = 30-39 years, (3) = 40-49 years and (4) = 50-67 years. (AGE Qualitative>Ordinal variable | ANXIETY: not present (0), present (1), unknown (9), or control (2). (Qualitative>Nominal variable) |
| ADHD: General presence of ADHD n=51. not present (0), present (1), or unknown (9). (ADHD Qualitative>Nominal>Categorical variable) |

* Algorithms

F1 Score: Used to balances precision and recall, providing a single measure of the model's accuracy in predicting positive cases.

F1-score Training and Testing: Assess the model's ability to correctly predict positive cases, providing insights into its performance on the training and testing datasets.

Accuracy: Measures the overall correctness of the model's predictions, offering an intuitive understanding of its correctness.

Confusion Matrix: Provides a detailed breakdown of the model's predictions and misclassifications, offering insights into the types of errors the model is making.

AdaBoostClassifier: Used to improve the performance of a weak learner (e.g., a decision tree) by combining multiple instances of it

RandomForestClassifier: Builds multiple decision trees and merges them together to get a more accurate and stable prediction.

|  |  |
| --- | --- |
| **Updated Model Measurements – Decision Tree Classifier Binary**  **ADHD and AGE** | |
| **Model 1 Substance** | **Model 2 Anxiety** |
| Adaboost Classifier  Accuracy: 81.48%  Confusion Matrix: [[22 0] [ 5 0]]  Best Parameters: {'learning\_rate': 0.01, 'n\_estimators': 10}  Best Estimator: AdaBoostClassifier (learning\_rate=0.01, n\_estimators=10) | No further algorithms were applied to ANXIETY. The model's F1 Score (weighted average) of 0.5213 suggests a moderate overall predictive performance. However, the low F1-scores for both training and testing data indicate a need for improvement. |
| Random Forest  Accuracy: 81.48%  Confusion Matrix: [[22 0] [ 5 0]]  Best Parameters: {'max\_depth': 10, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 100}  Best Estimator: RandomForestClassifier(max\_depth=10) |  |

From the previous model baseline the F1-score was 0.7316 and with the application of algorithms the model remained the same. For SUBSTANCE abuse, both the AdaBoostClassifier and RandomForestClassifier models achieved an accuracy of 81.48%, similar to the original base model.

The confusion matrix returned the same results, after applying the algorithms, as the original base model [[22 0] [ 5 0]].

Further model refinement, feature selection, hyperparameter tuning, and exploring ensemble methods could be beneficial to address misclassifications.

**Metrics**

As the project is Supervised, Classification, Decision Tree binary, the metrics used are the standard as:

-Prediction time and training time: Are important in machine learning to identify efficiency and practicality of deploying machine learning models. If the result leads to a faster time, it will allow quicker model deployment.

-F1 Score: Given the nominal nature of the ANXIETY and SUBSTANCE ABUSE variables, the F1 score is crucial as it balances precision and recall, providing a good measure of the model's predictive power, while considering false positives and false negatives.

-Precision and Recall: Precision is the ratio of correctly predicted positive observations to the total predicted positives, while recall is the ratio of correctly predicted positive observations to all actual positives.

-Accuracy: This metric provides a simple and intuitive measure of the model's overall correctness and is valuable for understanding the proportion of correct predictions.

These metrics are crucial for evaluating the model's performance in identifying positive cases (ANXIETY and SUBSTANCE presence) accurately and with minimal false positives or false negative.

**Experiments**

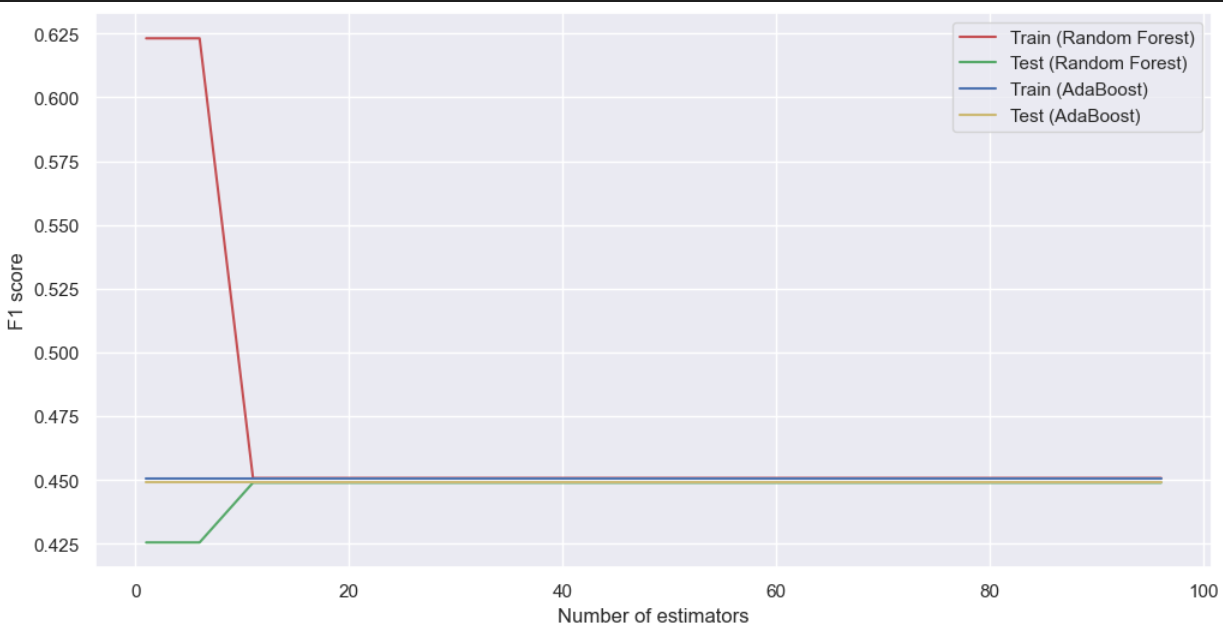
At least two experiments were carried out for each set of variables (SUBSTANCE and ANXIETY) to evaluate the base line and updated models.

However, n=estimator for Y = SUSBTANCE was equal to 10, as an enough score for achieving the highest test set performance.

AdaBoost Best Parameters: {'learning\_rate': 0.01, 'n\_estimators': 10}

AdaBoost Best Estimator: AdaBoostClassifier (learning\_rate=0.01, n\_estimators=10)

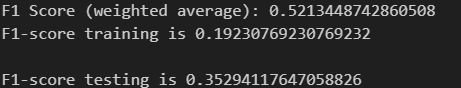
Y = SUSBTANCE



In contrast for Y = ANXIETY the model's performance in predicting ANXIETY based on AGE and ADHD was not ideal for the general population, due to the F1Sore of 53%.

As a result, the model needs to improve, before running experiments for ANXIETY.

Y = ANXIETY



**Results and Analysis**

AGE, ADHD, and SUBSTANCE: Both the AdaBoostClassifier and RandomForestClassifier models achieved an accuracy of 81.48%. Further model refinement, feature selection, hyperparameter tuning, and exploring ensemble methods could be beneficial to address misclassifications.

However, this indicates that the train F1 Score of 0.625 Random Forest model performed reasonably well on the data it was trained on for precision and recall and it has good generalization performance. The test F1 Score has a score of 0.45 meaning the model performance decreased suggesting potential overfitting. As a result, training data can be increased by adding more variables to the analysis e.g., ADHD and AGE with Bipolarity, ASRS which evaluate symptoms of ADHD (0-72 SUM) if score is high, then more severe symptoms of impulsivity, hyperactivity, inattention will occur. Also, to include MADRS: montgomery and asher depression rating, severity of ongoing depression (0-60) < 10 = no depression.

In contrast, the Train and Test F1 Score of 0.45 for Adaboots Classifier shows underfitting and this model can not be generalized to the population, using the data set.

AGE, ADHD, and ANXIETY: The model's F1 Score (weighted average) of 0.5213 suggests a moderate overall predictive performance. However, the low F1-scores for both training and testing data indicate a need for improvement.

| **Updated Model Measurements – Decision Tree Classifier Binary**  **ADHD and AGE** | |
| --- | --- |
| **Model 1 Substance** | **Model 2 Anxiety** |
| Adaboost Classifier: 81.48% accuracy  Confusion Matrix: [[22 0] [ 5 0]]  Best Parameters: {'learning\_rate': 0.01, 'n\_estimators': 10}  Best Estimator: AdaBoostClassifier (learning\_rate=0.01, n\_estimators=10) |  |
| Random Forest: 81.48%  Confusion Matrix: [[22 0] [ 5 0]]  Best Parameters: {'max\_depth': 10, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 100}  Best Estimator: RandomForestClassifier(max\_depth=10) |  |
| For SUBSTANCE ABUSE, both the AdaBoostClassifier and RandomForestClassifier models achieved an accuracy of 81.48%. Further model refinement, feature selection, hyperparameter tuning, and exploring ensemble methods could be beneficial to address misclassifications. | The model's F1 Score (weighted average) of 0.5213 suggests a moderate overall predictive performance. However, the low F1-scores for both training and testing data indicate a need for improvement. |

**Conclusion**

The analysis of the patient sample revealed significant insights into the prevalence of comorbid conditions among individuals with and without ADHD. The findings indicated that a high proportion of individuals with ADHD experience both anxiety and substance abuse issues, bringing the importance of holistic care and interventions for this population.

Moreover, the observed age distribution highlighted specific age groups, such as the 40-49 range, as being particularly affected by ADHD, anxiety, and substance abuse. Additionally, the prevalence of comorbid anxiety and substance abuse among individuals without ADHD underscored the need for comprehensive mental health assessments and interventions across various age groups.

Further research and targeted interventions are warranted to better understand and address the needs of individuals with these comorbid conditions.

In addition, for the Machine Learning techniques proposed in this data set are but not limit to:

Feature Selection: Analyzing additional features that have a stronger influence on predicting ANXIETY, SUBSTANCE ABUSE based on AGE and ADHD. Also, move from a binary decision tree to a multiclass decision tree iteration, will be recommended for the future to apply more variables that can explain the variation in the Y Target Variable.

Hyperparameter Tuning: Applying estimators as AdaBoostClassifier, and parameters like max depth and number of estimators in RandomForestClassifier, to improve predictive performance.

Ensemble Methods: Exploring the use of ensemble methods like bagging and boosting to combine the predictions of multiple models.

In future studies, will be recommended to analyze more variables associated with ADHD and AGE as Bipolarity, ASRS which evaluate symptoms of ADHD (0-72 SUM) if score is high, then more severe symptoms of impulsivity, hyperactivity, inattention will occur. Also, to include MADRS: montgomery and asher depression rating, severity of ongoing depression (0-60) < 10 = no depression.

**Appendix**

By following this end-to-end workflow, the project aims to develop a robust binary decision tree model to analyze the associations between adults with ADHD and co-occurring conditions, providing valuable insights into their interrelationships and potential predictive factors.

