**Capstone Proposal**

Nomadic Coder

Today, Right Now.

**Domain Background**

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.

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

"ADHD: current concepts and treatments in children and adolescents" by Drechsler R, Brem S, Brandeis D, Grunblatt E, Berger G, Walitza S. Neuropediatrics.

**Problem Statement**

Develop a Classification Decision Tree Binary machine learning model 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.

Develop a Binary Classification Tree based machine learning model

The predictive model can have an accuracy of 70% as a baseline and it will be evaluated after executing a detail analysis using Python for Data Science and Machine learning algorithms.

**Datasets and Inputs**

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.

Source: https://www.kaggle.com/datasets/arashnic/adhd-diagnosis-data, https://dl.acm.org/doi/pdf/10.1145/3458305.3478454 (white paper)

https://github.com/simula/hyperaktiv (GitHub)

https://osf.io/2tk4r (Data set view)

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 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) | 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) |  |

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) |  |
| ADHD: General presence of ADHD n=51. not present (0), present (1), or unknown (9). (ADHD Qualitative>Nominal>Categorical variable) | SUBSTANCE: drug, alcohol, addictions. not present (0), present (1), unknown (9), or control (2). (Qualitative>Nominal variable) |

**Solution Statement**

Performing exploratory data analysis and machine learning methodologies will help to evaluate the models' generalization capability. Using, Random Forest and AdaBoost models could describe the generalization of data, provide information for overfitting or underfitting to improve the model performance; in predicting anxiety based on age and ADHD> Further refinement through hyperparameter tuning, or additional data variables will be executed.

**Benchmark Model**

An estimate of a baseline for accuracy is 0.7000000. Ideally by applying a machine learning model the accuracy can increase.

**Evaluation 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

**Project Design**

In this final section, summarize a workflow for your capstone. List out in steps (diagrams encouraged!) what you need to do, end-to-end to complete your capstone. Start at the beginning (the data) and work your way through until the end results. Try to be as detailed as possible. This planning session will greatly help you implement your project.

**Step 1: Data Collection and Preprocessing**

Data Collection: Obtain access to the HYPERAKTIV dataset to identify demographic and diagnostic information of adults, including ADHD diagnosis, anxiety status, substance abuse, and age range.

Data Cleaning: Address any missing or inconsistent data, handle unknown (9) and control values (2), and ensure data quality and consistency, to reflect only (0) not present or (1) present for ADHD, SUBSTNACE and ANXIETY.

Data Exploration: Conduct exploratory data analysis to understand the distribution and relationships within the dataset, and visualize the age range and frequency of co-occurring conditions. Keeping in mind that the variable types:

* X= AGE Qualitative>Ordinal variable
* X =ADHD Qualitative>Nominal>Categorical variable
* Y = ANXIETY Qualitative>Nominal variable
* Y = SUBSTANCE Qualitative>Nominal variable

**Step 2: Feature Engineering and Selection**

Determine relevant features that can explain the variation in anxiety and substance.

**Step 3: Model Development and Training**

Decision Tree Model Selection: from a list of suitable algorithms as Adaboost, Random Forest or XGBoost.

Data Splitting: Split the dataset into training and testing sets to enable model evaluation. For example,

X = Predictor variables (ADHD, AGE)

* X\_train:80%
* X\_test:20%

Y = variables being predicted (SUBSTANCE, ANXIETY) TARGET VARIABLE

* Y\_train:80%
* Y\_test: 20%

Model Training: Train the decision tree model using the HYPERAKTIV dataset, including the features related to ADHD, anxiety, substance abuse, and age as input variables.

**Step 4: Model Evaluation and Optimization**

Performance Metrics: Evaluate the model's performance using metrics such as accuracy, precision, recall, and F1 score.

Cross-Validation: Implement cross-validation techniques to ensure robust model evaluation and mitigate overfitting.

Hyperparameter Tuning: Optimize the decision tree model's hyperparameters to enhance its predictive capabilities and generalization.

**Step 5: Interpretation and Insights**

Model Interpretation: Analyze the decision tree model's structure and rules to understand the associations and predictive factors related to ADHD, anxiety, substance abuse, and age.

**Step 6: Visualization and Reporting**

Results Visualization: Visualize the decision tree model's structure, feature importances, and key insights using appropriate visualization techniques.

Report Generation: Summarize the project findings, including the associations between ADHD and co-occurring conditions, the impact of age range, and potential predictive factors.

**Step 7: Deployment and Application**

Model Deployment: Prepare the decision tree model for deployment in GitHub

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.



