Research Log - ML SEM-4 PROJECT

—

Avantika Agarwal

TOPIC: Multimodal ML models integrating Genomic and medical information for ASD detection

EXPERIMENT-1

DATE: 23/01/24

Model Description and Input

The model used was Logistic Regression and the input was a Breast cancer dataset present as a default dataset in scikit learn (for practice purposes). In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function. In this code, we are carrying out Binary Logistic Regression, meaning, there would be two possible outcomes of the prediction (that is, it predicts the presence or absence of breast cancer - 0/1).

Results

PARAMETER	EXPLANATION	VALUE
Accuracy score (normalized)	The accuracy_score function computes the accuracy→ the fraction (default) in this case	0.951048951048951
Accuracy score (no normalization)	The accuracy_score function computes the accuracy→ the count (normalize=False) of correct predictions.	136
Area under the curve (AUC)	AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Compute Area Under the Curve (AUC) using the trapezoidal rule.	0.9937106918238994

	This is a general function, given points on a curve	
Sensitivity	Sensitivity = TP / (TP + FN)	0.9444444444444444
Specificity	Specificity = TN / (TN + FP)	0.9622641509433962
Precision	The precision is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative. The best value is 1 and the worst value is 0.	0.9770114942528736
F1 score	F1 score, also known as balanced F-score or F-measure, can be interpreted as a harmonic mean of precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is: F1 = 2 * TP / (2 * TP + FN + FP) Where “TP” is the number of true positives, “FN” is the number of false negatives, and “FP” is the number of false positives. F1 is by default calculated as 0.0 when there are no true positives, false negatives, nor false positives.	0.96045197740113

Interpretations

Values close to 1 (>0.95) were obtained for all the parameters, indicating that the Logistic Regression Model is a good classification model for the breast cancer dataset

EXPERIMENT-2

DATE: 23/01/24

Model Description and Input

The models used were:

1) Random Forests
2) Decision Trees
3) Support Vector Machine
4) XGBoost

The input was a Breast cancer dataset present as a default dataset in scikit learn (for practice purposes). The same steps as that of logistic regression were implemented in all the above 4 models and the various results obtained were compared.

Results

PARAMETER	RANDOM FOREST	DECISION TREE	SVM	XGBOOST
Accuracy score (normalized)	0.972027972027972	0.8811188811188811	0.9370629370629371	0.972027972027972
Accuracy score (no normalization)	139	126	134	139
Area under the curve (AUC)	0.9969601677148846	0.8939203354297695	0.9842767295597485	0.9987421383647799
Sensitivity	0.9666666666666667	0.8444444444444444	0.9888888888888889	0.9666666666666667
Specificity	0.9811320754716981	0.9433962264150944	0.8490566037735849	0.9811320754716981

F1 score	0.9775280898876404	0.8994082840236687	0.9518716577540107	0.9775280898876404
Precision	0.9886363636363636	0.9620253164556962	0.9175257731958762	0.9886363636363636

Interpretation

- It was observed that the runtime of the random forests code was greater compared to the others
- Even xgboost had a slightly greater runtime relative to decision trees and svm
- Decision trees gave a lower accuracy, AUC, Sensitivity and F1 score relative to the other models indicating that this model is not that suitable for predicting breast cancer
- Random Forests and XGBoost models gave exactly same values for all the metric (accuracy, precision, sensitivity etc) when the data was split into train and test sets using random state = 0 , changing the value of random state (say 40 for both models) gave different values.

EXPERIMENT-3

DATE: 24/01/24

Model Description and Input

The input was a Breast cancer dataset present as a default dataset in scikit learn (for practice purposes). The logistic regression code of Experiment 1 was modified to scale data and carry out 10 fold cross validation on the same as part of the pipeline. The pipeline (Pipeline from scikit-learn) includes two steps: standardization (StandardScaler) and logistic regression (LogisticRegression). StandardScaler transform was applied to standardize the input variables. The pipeline was further evaluated using a 10 fold cross validation - RepeatedStratifiedKFold. Each fold was trained on the scaled data

n_splits=10: Specifies the number of folds in the cross-validation. In this case, it's set to 10, meaning the dataset will be split into 10 parts, and the model will be trained and evaluated 10 times, each time using a different fold as the test set and the remaining data as the training set.

n_repeats=3: Specifies the number of times the cross-validation process will be repeated. It repeats the entire cross-validation process three times with different random splits.

random_state=1: Provides a seed for the random number generator. Setting a seed ensures reproducibility, meaning if you run the code with the same seed, you should get the same random splits each time.

Results

```
Accuracy of pipeline: 0.977 (0.018)
Accuracy Score - default: 0.958041958041958
Accuracy Score - no normalization: 137
AUC: 0.9914046121593292
Sensitivity: 0.9666666666666667
Specificity: 0.9433962264150944
F1 score: 0.9666666666666667
Precision: 0.9666666666666667
              precision    recall  f1-score   support

           0       0.94      0.94      0.94        53
           1       0.97      0.97      0.97        90

    accuracy                           0.96       143
   macro avg       0.96      0.96      0.96       143
weighted avg       0.96      0.96      0.96       143
```

```
Accuracy Score - default: 0.951048951048951
Accuracy Score - no normalization: 136
AUC: 0.9937106918238994
Sensitivity: 0.9444444444444444
Specificity: 0.9622641509433962
F1 score: 0.96045197740113
Precision: 0.9770114942528736
              precision    recall  f1-score   support

           0       0.91      0.96      0.94        53
           1       0.98      0.94      0.96        90

    accuracy                           0.95       143
   macro avg       0.94      0.95      0.95       143
weighted avg       0.95      0.95      0.95       143
```

LHS -> Metrics obtained after standardization and cross validation

RHS -> Metrics obtained after experiment 1

Interpretation

The pipeline accuracy was close to 1 indicating that the standardization done was proper. Slight increase in accuracy and sensitivity was observed. However, overall the model was not quite affected by the standardization, indicated by the similar values of the metrics pre and post standardization

EXPERIMENT-4

DATE: 25/01/24

Model Description and Input

Implemented 10 fold cross validation(without using the default function in sci-kit learn) and standardization on all the 5 classifiers and printed metrics for each fold, finally printing the mean and standard deviation of each of the 6 metrics. Input used was the breast cancer dataset present by default in sci-kit learn.

Results

Mean and Standard deviation of Metrics over 10 fold cross validation for the 5 classifiers

LOGISTIC REGRESSION

Metric	Mean	Standard Deviation
Accuracy	0.865972	0.274663
AUC	0.902248	0.285399
Sensitivity	0.883766	0.280223
Specificity	0.836678	0.271012
F1 score	0.875243	0.277244
Precision	0.867883	0.276302

RANDOM FORESTS

Metric	Mean	Standard Deviation
Accuracy	0.870785	0.276297
AUC	0.898192	0.284401
Sensitivity	0.883622	0.280178
Specificity	0.849075	0.276065
F1 score	0.878971	0.278497
Precision	0.875487	0.279092

DECISION TREES

Metric	Mean	Standard Deviation
Accuracy	0.821172	0.261122
AUC	0.817454	0.260927
Sensitivity	0.832468	0.266304
Specificity	0.80244	0.267702

| F1 score | 0.838499 | 0.266066 |

| Precision | 0.847754 | 0.272047 |

SUPPORT VECTOR MACHINE

Metric	Mean	Standard Deviation
Accuracy	0.830799	0.264151
AUC	0.887194	0.280768
Sensitivity	0.888817	0.281768
Specificity	0.734357	0.243995
F1 score	0.849673	0.269425
Precision	0.815038	0.260516

XGBOOST

Metric	Mean	Standard Deviation
Accuracy	0.881949	0.279702
AUC	0.903572	0.285805
Sensitivity	0.888672	0.281725

Specificity	0.870523	0.277784
F1 score	0.887495	0.281156
Precision	0.886626	0.28121

Interpretation

Each of the models gave similar results for the metrics, which might have been due to the near ideal nature of the dataset

EXPERIMENT-5

DATE: 30/01/24

Description and Input

This experiment involved analysis of the different datasets in SPARK (.csv files) in R for detection of the number of individuals affected by ASD and the number of unaffected individuals.

Results

File	Total Individuals	ASD Affected	Non-ASD
core_descriptive_variables-2023-07-21.csv	132368	132368	0
basic_medical_screening-2023-07-21.csv	192557	79631	112926
cbcl_1_5-2023-07-21.csv	2835	2835	0
cbcl_6_18-2023-07-21.csv	10290	10290	0
dcdq-2023-07-21.csv	35097	35097	0
individuals_registration-2023-07-21.csv	328974	132139	196835

predicted_iq_experimental-2023-07-21.csv	132368	132368	0
rbsr-2023-07-21.csv	46517	46517	0
roles_2023-07-17.csv	132368	132368	0
scq-2023-07-21.csv	89678	63491	26187
srs-2_adult_self-2023-07-21.csv	1798	0	1798
vineland-3-2023-07-21.csv	23160	23159	1
background_history_sibling-2023-07-21.csv	19700	21	19679

Filename: asd_noasd_num_basicmedical.r and asd_noasd_num_multipledatasets.r

Interpretation

Only 3 datasets : individual registration, basic medical screening and scq has relevant number of both ASD affected and unaffected individuals, indicating that these datasets can probably used for future experiments

Recommendations for Improvement

1) The above 3 mentioned datasets can be used as primary datasets for building model
2) For the subject IDs mentioned in these 3 datasets, relevant information (columns) can be extracted from the other datasets and added to a single data frame
3) Datasets can be merged on the basis of unique subject ID

EXPERIMENT-6

DATE: 01/02/24

Description and Input

This experiment involved analysis of the basic medical screening dataset in SPARK (.csv file) in R for detection of the number of affected and unaffected individuals for the following disorders:

1) ASD
2) ADHD
3) ODD
4) OCD
5) Schizophrenia

Results

TOTAL NUMBER OF INDIVIDUALS -> 192557 , Filename : asd_comorbidities.r

Disease	Non-Affected Individuals	Affected Individuals
ASD	112926	79631
ADHD	148304	44253
ODD	185436	7121
OCD	179556	13001
Schizophrenia	190794	1763

EXPERIMENT-7

DATE: 02/02/24

Description and Input

This experiment involved analysis of the basic medical screening dataset in SPARK (.csv file) in R for detection of the following conditions pertaining ASD and ADHD:

1) Individuals affected by ASD only and no other comorbidities
2) Individuals affected by ADHD only and no other comorbidities
3) Individuals affected by both ASD and ADHD
4) Individuals unaffected by any of the 2 disorders

Results

TOTAL NUMBER OF INDIVIDUALS -> 192557

ASD only	ADHD only	Both ASD and ADHD	Unaffected by any of the 2 disorders
49804	14426	29827	98500

Filename: asd_adhd_numbers.r

EXPERIMENT-8

DATE: 03/02/24

Description and Input

This experiment involved analysis of the basic medical screening dataset in SPARK (.csv file) and filtering out the columns that will be used to develop the ML model (based on certain criteria inferred from literature).

Results

TOTAL NUMBER OF INDIVIDUALS -> 192557

IGNORED COLUMNS:

- Ignored the columns containing various SPARK ids because the dataset will be used as a whole for model building and not use individual ids
    - subject_sp_id
    - respondent_sp_id
    - family_sf_id
    - biomother_sp_id
    - biofather_sp_id
    - current_depend_adult
- Ignored the columns concerning other comorbidities:
    - behav_conduct
    - behav_intermitt_explos
    - behav_odd
- Current_depend_adult
- Flag
- The following columns are broad columns having sub-columns (branching qs) :- They are probably placing 1/ null on the basis of answers to all the sub questions but it is vague what is the basis of choosing the final value
    - Attn_behav
    - Birth_def_cns -> Branching Question: Brain and spinal cord birth defects
    - Birth_def_bone
    - birth_def_fac
    - Birth_def_gastro

    - Birth_def_thorac (heart or lung)
    - Birth_def_urogen
    - Dev_lang
    - Gen_test
    - med_cond_birth
    - med_cond_birth_def
    - med_cond_growth
    - med_cond_neuro
    - Med_cond_visaud
    - mood_ocd
- Birth defects: (defects not impacting asd/adhd)
    - Birth_def_bone_club
    - birth_def_bone_miss
    - birth_def_bone_polydact
    - birth_def_bone_spine
    - birth_def_cleft_lip
    - Birth_def_cleft_palate
    - Birth_def_cns_myelo
    - birth_def_gi_esoph_atres
    - birth_def_gi_hirschprung
    - birth_def_gi_intest_malrot
    - Birth_def_gi_pylor_sten
    - birth_def_thorac_heart
    - Birth_def_thorac_lung
    - birth_def_urogen_hypospad
    - birth_def_urogen_renal
    - birth_def_urogen_renal_agen
    - birth_def_urogen_uter_agen
    - Birth_def_oth_calc
    - birth_ivh
    - Birth_oth_calc
- Eating_probs - Not professionally diagnosed so probably not a definitive measure
- Eating_disorder - Not directly associated with ASD
- Etoh_subst
- gen_dx_oth_calc_self_report
- Genetic testing:
    - gen_test_cgh_cma
    - gen_test_chrom_karyo

    - gen_test_ep
    - gen_test_fish_angel
    - gen_test_fish_digeorge
    - gen_test_fish_williams
    - gen_test_fish_oth
    - gen_test_frax
    - gen_test_id
    - gen_test_mecp2
    - gen_test_nf1
    - gen_test_noonan
    - gen_test_pten
    - gen_test_tsc
    - gen_test_unknown
    - gen_test_wes
    - gen_test_wgs
    - Gen_test_oth_calc
- Calculated variable: individual has entered in text field an "Other" growth condition not represented in current coding
- growth_low_wt
- growth_macroceph
- growth_microceph
- growth_obes
- growth_short
- Growth_oth_calc
- prev_study_calc
- Eval_year
- neuro_inf
- Neuro_lead
- neuro_sz
- neuro_tbi

- Neuro_oth_calc
- pers_dis
- prev_study_oth_calc
- psych_oth_calc
- Schiz
- visaud_blind
- visaud_catar
- visaud_deaf
- Visaud_strab
- sleep_dx
- sleep_eat_toilet
- sleep_probs
- tics

CHOSEN COLUMNS:

1) Sex -> To know which gender is more susceptible to a particular disorder
2) Asd -> Professional diagnosis results
3) Age_at_eval_months -> To know the age distribution of each disorder
4) Age_at_eval_years -> To know the age distribution of each disorder (we can use any one of 3 or 4 -
5) Behav_adhd -> professional diagnosis of adhd
6) Birth defects:
    a) Birth_def_cns_brain -> Brain malformation/abnormality (shown on MRI)
    b) Birth_def_thorac_cdh -> Congenital diaphragmatic hernia - Increased Risk: Studies have shown that children with CDH have a higher risk of developing Autism Spectrum Disorder (ASD) compared to the general population. The estimated prevalence of ASD in CDH survivors is around 9-14%, compared to 1.5% in the general population.Although there is research on the increased risk of ASD in CDH, there is currently no established direct link between CDH and Attention Deficit Hyperactivity Disorder (ADHD).
    c) Birth_etoh_subst -> Fetal Alcohol Syndrome, alcohol or drug exposure in mother's pregnancy

d) Birth_oxygen - > Insufficient oxygen at birth with NICU stay - Studies suggest a higher risk of ASD and ADHD in children who experienced HIE compared to the general population.
e) Birth_pg_inf -> Serious prenatal infection (for example, German measles) - Studies suggest a modest increase in the risk of ASD in children whose mothers contracted certain severe infections during pregnancy
f) Birth_prem -> Premature birth (delivery before 37 weeks) - Premature birth, delivering before 37 weeks of pregnancy, is indeed linked to an increased risk of Autism Spectrum Disorder (ASD) and Attention Deficit Hyperactivity Disorder (ADHD). The earlier the birth, the higher the risk, according to multiple studies

7) cog_med- Cognitive delays or impairment due to another medical condition or exposure (For example, brain injury, stroke, lead poisoning, FAS, HIV, radiation, hydrocephalus, brain tumor, drug effects, etc.) —> Too many factors so might make it confusing
8) Developmental properties:
    a) Dev_id - Intellectual disability, cognitive impairment, global developmental delay, or borderline intellectual functioning
    b) Dev_lang_dis - Language difficulties are core features of ASD for many individuals.While not a core symptom of ADHD, individuals with ADHD can sometimes experience language difficulties.
    c) Dev_ld - Learning disability (LD, learning disorder, including reading, written expression, math, or NVLD (Nonverbal learning disability))
    d) Dev_motor - Motor delay (e.g., delay in walking) or developmental coordination disorder - Both motor delays and DCD can co-occur with ASD and ADHD.
    e) Dev_mutism - Mutism, particularly SM, can co-occur with ASD and ADHD.
    f) Dev_soc_prag - Social (Pragmatic) Communication Disorder - SCD can co-occur with both ASD and ADHD.
    g) Dev_speech - Speech articulation problems are more common in individuals with ASD compared to the general population. Not much in ADHD
9) Encopres - Encopresis (has bowel accidents beyond age expected).Studies suggest children with ASD and ADHD might be at higher risk for encopresis compared to the general population.
10) Enures - Enuresis (wets self beyond age expected). Children with ASD are 2-3 times more likely to experience enuresis compared to typically developing children

11) Feeding_dx - Feeding/eating problems - Children with ASD are five times more likely to experience feeding difficulties compared to typically developing peers.

12) Parent-reported genetic testing:

    a) Gen_test_aut_dd - Parent reported Genetic testing for autism or developmental delay genes

13) Gest_age - How many weeks (gestational age) was a child/dependent when he/she was born? - Premature birth (born before 37 weeks) is associated with a slightly increased risk of developing both ASD and ADHD.

14) Mood disorders: - usually associated with ASD / ADHD

    a) Mood_anx - Anxiety disorder, such as panic, phobia, agoraphobia, or generalized anxiety disorder (GAD) except for social anxiety
    b) Mood_bipol - Bipolar (Manic-Depressive) Disorder
    c) Mood_dep - Depression or dysthymia
    d) Mood_dmd - Disruptive Mood Dysregulation Disorder
    e) Mood_hoard - Hoarding
    f) Mood_or_anx - Obsessive-Compulsive Disorder
    g) Mood_sep_anx - Separation Anxiety
    h) Mood_soc_anx - Social Anxiety Disorder/Social Phobia

prev_study_agre

prev_study_asc

prev_study_charge

prev_study_earli

prev_study_marbles

prev_study_mssng

prev_study_seed

prev_study_ssc

prev_study_vip

EXPERIMENT-9

DATE: 17/02/24

Description and Input

We developed a XGBoost Classifier on the basic medical screening dataset to make a 3-way classification model that classifies patients into individuals with ASD only, ADHD only or both.

COLUMNS CHOSEN FOR FEATURES ARE:

Categorical columns : ['sex']

Numerical columns : ['age_at_eval_months', 'age_at_eval_years', 'birth_def_cns_brain', 'birth_def_thorac_cdh', 'birth_etoh_subst', 'birth_oxygen', 'birth_pg_inf', 'birth_prem', 'cog_med', 'dev_id', 'dev_lang_dis', 'dev_ld', 'dev_motor', 'dev_mutism', 'dev_soc_prag', 'dev_speech', 'eating_probs', 'eating_disorder', 'encopres', 'enures', 'feeding_dx', 'gen_test_aut_dd', 'gest_age', 'mood_anx', 'mood_bipol', 'mood_dep', 'mood_dmd', 'mood_hoard', 'mood_or_anx', 'mood_sep_anx', 'mood_soc_anx', 'prev_study_agre', 'prev_study_asc', 'prev_study_charge', 'prev_study_earli', 'prev_study_marbles', 'prev_study_mssng', 'prev_study_seed', 'prev_study_ssc', 'prev_study_vip', 'sleep_dx', 'sleep_eat_toilet', 'sleep_probs']

Results

Df size after removing unaffected people = 94057

Number of X features = 44

MEAN AND STANDARD DEVIATION OF VARIOUS METRICS

MEAN AND STANDARD DEVIATION OF VARIOUS METRICS

```
+-------------+----------+----------------------+
| Metric      |     Mean |   Standard Deviation |
+=============+==========+======================+
```

| Accuracy | 0.686094 | 0.00584678 |

+----------+----------+----------+

| Sensitivity | 0.671497 | 0.0134625 |

+----------+----------+----------+

| Specificity | 0.939345 | 0.00220678 |

+----------+----------+----------+

| F1 score | 0.682309 | 0.00617388 |

+----------+----------+----------+

| Precision | 0.681323 | 0.00621106 |

+----------+----------+----------+

Filename: model_3way_basicmedical_1.py

Interpretation

The classification of the numerical and categorical columns was incorrect as all the columns with Boolean values became a part of the Numerical columns while they should have been a part of the categorical columns. The ROC-AUC value is not being calculated for a 3 way classification model.

Recommendations for Improvement

- The classification of the numerical and categorical columns should be improved - by manually observing the numerical and categorical columns
- Modifying the chosen features - some features are important but not part of the model

EXPERIMENT-10

DATE: 20/02/24

Description and Input

We modified the above XGBoost Classifier on the basic medical screening dataset to make a 3-way classification model that classifies patients into individuals with ASD only, ADHD only or both by making changes in the features being used, also modifying the columns (categorical and numerical) used by the column transformer

COLUMNS CHOSEN FOR FEATURES ARE:

Categorical columns : ['birth_def_bone_club', 'birth_def_bone_miss', 'birth_def_bone_polydact', 'birth_def_bone_spine', 'birth_def_cleft_lip', 'birth_def_cleft_palate', 'birth_def_cns_brain', 'birth_def_cns_myelo', 'birth_def_gi_esoph_atres', 'birth_def_gi_hirschprung', 'birth_def_gi_intest_malrot', 'birth_def_gi_pylor_sten', 'birth_def_thorac_cdh', 'birth_def_thorac_heart', 'birth_def_thorac_lung', 'birth_def_urogen_hypospad', 'birth_def_urogen_renal', 'birth_def_urogen_renal_agen', 'birth_def_urogen_uter_agen', 'birth_etoh_subst', 'birth_ivh', 'birth_oxygen', 'birth_pg_inf', 'birth_prem', 'cog_med', 'dev_id', 'dev_lang_dis', 'dev_ld', 'dev_motor', 'dev_mutism', 'dev_soc_prag', 'dev_speech', 'eating_probs', 'eating_disorder', 'encopres', 'enures', 'feeding_dx', 'growth_low_wt', 'growth_macroceph', 'growth_microceph', 'growth_obes', 'growth_short', 'mood_anx', 'mood_bipol', 'mood_dep', 'mood_dmd', 'mood_hoard', 'mood_or_anx', 'mood_sep_anx', 'mood_soc_anx', 'neuro_inf', 'neuro_lead', 'neuro_sz', 'neuro_tbi', 'sleep_dx', 'sleep_eat_toilet', 'sleep_probs', 'visaud_blind', 'visaud_catar', 'visaud_deaf', 'visaud_strab']

Numerical columns : ['gest_age']

Results

Df size after removing unaffected people = 94057

Number of X features = 62

MEAN AND STANDARD DEVIATION OF VARIOUS METRICS

MEAN AND STANDARD DEVIATION OF VARIOUS METRICS

Metric	Mean	Standard Deviation
Accuracy	0.650914	0.00487393
Sensitivity	0.538473	0.00984392
Specificity	0.928407	0.00376441
F1 score	0.641011	0.00519251
Precision	0.643825	0.00509689

Filename: model_3way_basicmedical_2.py

Interpretation

The low accuracy but high sensitivity might indicate that the features being used are good for predicting ASD only → ADHD is not being predicted well.

Recommendations for Improvement

1) Reading up about the features being used and finding out what can be eliminated - features should be decreased
2) Try ordinal encoding instead of one hot on all the categorical columns
3) Try recursive feature elimination to work on features
4) Gestational age column - replace NAN with 40
5) EDA on categorical columns - as a bar plot
6) To make it inclusive of normal individuals - include bg datasets - bg child, sibling etc.

EXPERIMENT-11

DATE: 27/02/24

Description and Input

I carried out Exploratory data analysis on the chosen features to understand which of the features are truly relevant. I plotted a histogram for the numerical column (gestational age) and created a table for the categorical ones.

COLUMNS CHOSEN FOR FEATURES ARE:

Categorical columns : ['birth_def_bone_club', 'birth_def_bone_miss', 'birth_def_bone_polydact', 'birth_def_bone_spine', 'birth_def_cleft_lip', 'birth_def_cleft_palate', 'birth_def_cns_brain', 'birth_def_cns_myelo', 'birth_def_gi_esoph_atres', 'birth_def_gi_hirschprung', 'birth_def_gi_intest_malrot', 'birth_def_gi_pylor_sten', 'birth_def_thorac_cdh', 'birth_def_thorac_heart', 'birth_def_thorac_lung', 'birth_def_urogen_hypospad', 'birth_def_urogen_renal', 'birth_def_urogen_renal_agen', 'birth_def_urogen_uter_agen', 'birth_etoh_subst', 'birth_ivh', 'birth_oxygen', 'birth_pg_inf', 'birth_prem', 'cog_med', 'dev_id', 'dev_lang_dis', 'dev_ld', 'dev_motor', 'dev_mutism', 'dev_soc_prag', 'dev_speech', 'eating_probs', 'eating_disorder', 'encopres', 'enures', 'feeding_dx', 'growth_low_wt', 'growth_macroceph', 'growth_microceph', 'growth_obes', 'growth_short', 'mood_anx', 'mood_bipol', 'mood_dep', 'mood_dmd', 'mood_hoard', 'mood_or_anx', 'mood_sep_anx', 'mood_soc_anx', 'neuro_inf', 'neuro_lead', 'neuro_sz', 'neuro_tbi', 'sleep_dx', 'sleep_eat_toilet', 'sleep_probs', 'visaud_blind', 'visaud_catar', 'visaud_deaf', 'visaud_strab']

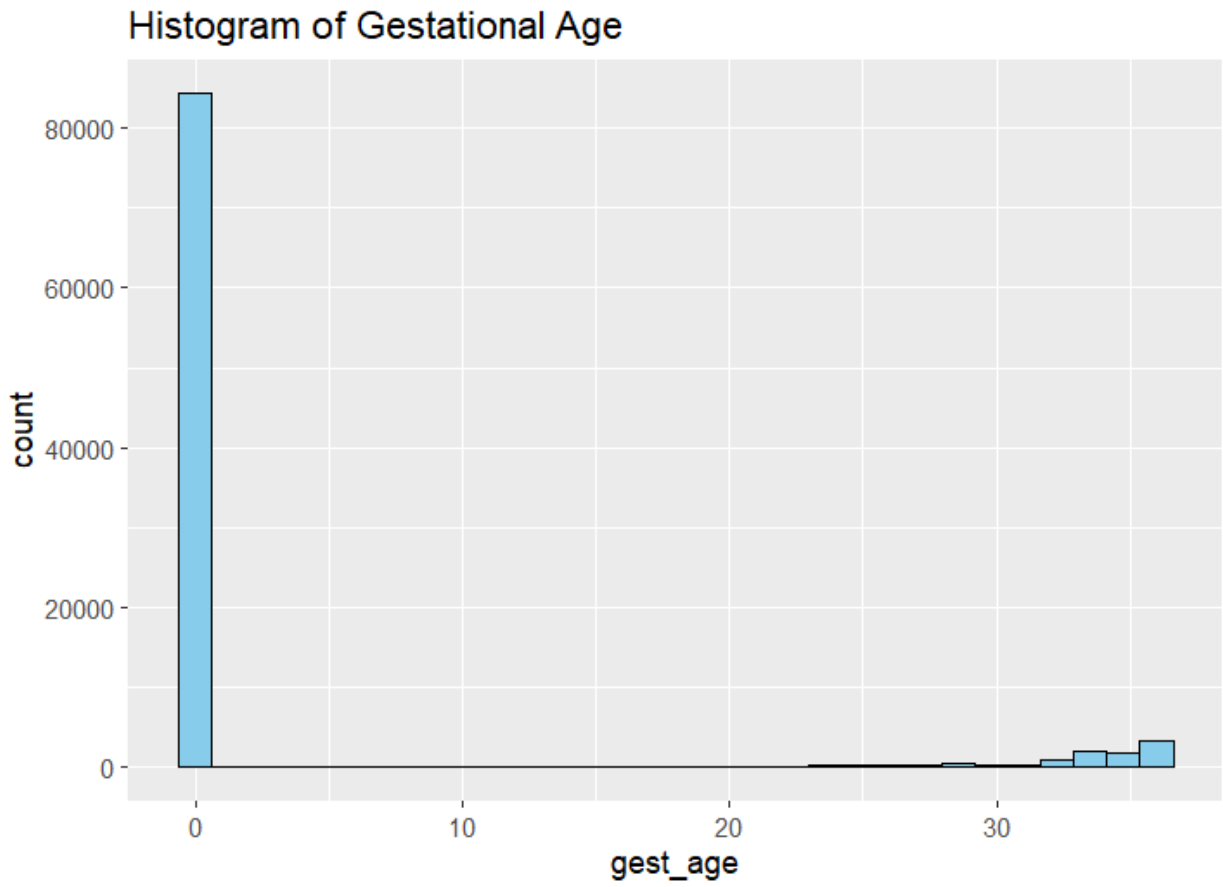
Numerical columns : ['gest_age']

Results

Df size after removing unaffected people = 94057

Number of X features = 62

Results for gestational age:



Results for categorical columns

Column_Name	Value_1_Count	Value_0_Count
birth_def_bone_club	200	93857
birth_def_bone_miss	356	93701
birth_def_bone_polydact	73	93984
birth_def_bone_spine	154	93903

birth_def_cleft_lip	106	93951
birth_def_cleft_palate	147	93910
birth_def_cns_brain	280	93777
birth_def_cns_myelo	63	93994
birth_def_gi_esoph_atres	32	94025
birth_def_gi_hirschprung	15	94042
birth_def_gi_intest_malrot	45	94012
birth_def_gi_pylor_sten	38	94019
birth_def_thorac_cdh	23	94034
birth_def_thorac_heart	664	93393
birth_def_thorac_lung	58	93999
birth_def_urogen_hypospad	257	93800
birth_def_urogen_renal	83	93974

birth_def_urogen_renal_agen	20	94037
birth_def_urogen_uter_agen	5	94052
birth_etoh_subst	1092	92965
birth_ivh	637	93420
birth_oxygen	5461	88596
birth_pg_inf	228	93829
birth_prem	10380	83677
cog_med	4348	89709
dev_id	16516	77541
dev_lang_dis	41394	52663
dev_ld	19937	74120
dev_motor	13796	80261
dev_mutism	1013	93044

dev_soc_prag	13506	80551
dev_speech	21610	72447
eating_probs	29433	64624
eating_disorder	1840	92217
encopres	5850	88207
enures	6368	87689
feeding_dx	11593	82464
growth_low_wt	2974	91083
growth_macroceph	2473	91584
growth_microceph	796	93261
growth_obes	2302	91755
growth_short	2015	92042
mood_anx	25518	68539

mood_bipol	4814	89243
mood_dep	16045	78012
mood_dmd	2612	91445
mood_hoard	1770	92287

Filename: eda_model.r

Interpretation

Many columns have value <1000 for Value=1 and can be ignored. These are:

'birth_def_bone_club', 'birth_def_bone_miss', 'birth_def_bone_polydact', 'birth_def_bone_spine', 'birth_def_cleft_lip', 'birth_def_cleft_palate', 'birth_def_cns_brain', 'birth_def_cns_myelo', 'birth_def_gi_esoph_atres', 'birth_def_gi_hirschprung', 'birth_def_gi_intest_malrot', 'birth_def_gi_pylor_sten', 'birth_def_thorac_cdh', 'birth_def_thorac_heart', 'birth_def_thorac_lung', 'birth_def_urogen_hypospad', 'birth_def_urogen_renal', 'birth_def_urogen_renal_agen', 'birth_def_urogen_uter_agen', 'birth_ivh', 'birth_pg_inf', 'growth_microceph', 'neuro_inf', 'neuro_lead', 'neuro_tbi', 'visaud_blind', 'visaud_catar'

EXTRA - FEATURES TO BE USED

sex

birth_def_bone_club

birth_def_bone_miss

birth_def_bone_polydact

birth_def_bone_spine

birth_def_cleft_lip

birth_def_cleft_palate

birth_def_cns_brain

birth_def_cns_myelo

birth_def_fac

birth_def_gi_esoph_atres

birth_def_gi_hirschprung

birth_def_gi_intest_malrot

birth_def_gi_pylor_sten

birth_def_thorac_cdh

birth_def_thorac_heart

birth_def_thorac_lung

birth_def_urogen_hypospad

birth_def_urogen_renal

birth_def_urogen_renal_agen

birth_def_urogen_uter_agen

birth_etoh_subst

birth_ivh

birth_oth_calc

birth_oxygen

birth_pg_inf

cog_med

dev_id

dev_lang

dev_lang_dis

dev_ld

dev_motor

dev_mutism

dev_soc_prag

dev_speech

eating_probs

eating_disorder

feeding_dx

gest_age

growth_low_wt

growth_macroceph

growth_microceph

growth_obes

growth_short

med_cond_birth

med_cond_birth_def

med_cond_growth

med_cond_neuro

med_cond_visaud

neuro_inf

neuro_lead

neuro_sz

neuro_tbi

```
features = [
    "sex", "gest_age", "eating_probs", "feeding_dx",
    "med_cond_birth", "birth_oth_calc",
    "med_cond_birth_def", #"birth_def_oth_calc" is not used because all are nans
    "med_cond_growth", "growth_oth_calc",
    "med_cond_neuro", "med_cond_visaud"
]

# use this set of features for SPARK experiments
features = [
    "sex", 'mother_highest_education', 'father_highest_education',
'annual_household_income',
    "smiled_age_mos", "sat_wo_support_age_mos", "crawled_age_mos", "walked_age_mos",
    "fed_self_spoon_age_mos", "used_words_age_mos", "combined_words_age_mos",
    "combined_phrases_age_mos", "bladder_trained_age_mos", "bowel_trained_age_mos",
"hand",
    "twin_mult_birth", 'num_asd_parents', 'num_asd_siblings'
]
```

EXPERIMENT-12

DATE: 28/02/24

Description and Input

We modified the above XGBoost Classifier on the basic medical screening dataset to make a 3-way classification model that classifies patients into individuals with ASD only, ADHD only or both by replacing nan values in the gest_age column with 40. Some features are removed as well.

COLUMNS CHOSEN FOR FEATURES ARE:

Categorical columns : ['birth_etoh_subst', 'birth_oxygen', 'birth_prem', 'cog_med', 'dev_id', 'dev_lang_dis', 'dev_ld', 'dev_motor', 'dev_mutism', 'dev_soc_prag', 'dev_speech', 'eating_probs', 'eating_disorder', 'encopres', 'enures', 'feeding_dx', 'growth_low_wt', 'growth_macroceph', 'growth_obes', 'growth_short', 'mood_anx', 'mood_bipol', 'mood_dep', 'mood_dmd', 'mood_hoard', 'mood_or_anx', 'mood_sep_anx', 'mood_soc_anx', 'neuro_sz', 'sleep_dx', 'sleep_eat_toilet', 'sleep_probs', 'visaud_deaf', 'visaud_strab']

Numerical columns : ['gest_age']

Results

Number of X features = 35

Df size after removing unaffected people = 94057

MEAN AND STANDARD DEVIATION OF VARIOUS METRICS

Metric	Mean	Standard Deviation
Accuracy	0.649957	0.00502995
Sensitivity	0.53168	0.0119204

| Specificity | 0.928131 | 0.00273799 |

+----------+----------+----------+

| F1 score | 0.639835 | 0.0055437 |

+----------+----------+----------+

| Precision | 0.642732 | 0.00539781 |

+----------+----------+----------+

Filename: model_3way_basicmedical_3.py

EXPERIMENT-13

DATE: 29/02/24

Description and Input

We modified the above XGBoost Classifier on the basic medical screening dataset to make a 3-way classification model that classifies patients into individuals with ASD only, ADHD only or both by using 11 features.

COLUMNS CHOSEN FOR FEATURES ARE:

Categorical columns : ['birth_etoh_subst', 'birth_oxygen', 'birth_prem', 'cog_med', 'dev_id', 'dev_lang_dis', 'dev_ld', 'dev_motor', 'dev_mutism', 'dev_soc_prag', 'dev_speech', 'eating_probs', 'eating_disorder', 'encopres', 'enures', 'feeding_dx', 'growth_low_wt', 'growth_macroceph', 'growth_obes', 'growth_short', 'mood_anx', 'mood_bipol', 'mood_dep', 'mood_dmd', 'mood_hoard', 'mood_or_anx', 'mood_sep_anx', 'mood_soc_anx', 'neuro_sz', 'sleep_dx', 'sleep_eat_toilet', 'sleep_probs', 'visaud_deaf', 'visaud_strab']

Numerical columns : ['gest_age']

Results

Number of X features = 11

Df size after removing unaffected people = 94057

MEAN AND STANDARD DEVIATION OF VARIOUS METRICS

Metric	Mean	Standard Deviation
Accuracy	0.563212	0.00516725
Sensitivity	0.559269	0.0169223
Specificity	0.906569	0.00333269

F1 score	0.494427	0.00511255
Precision	0.537941	0.00903289

Filename: model_3way_basicmedical_4.py

EXPERIMENT-14

DATE: 04/03/24

Description and Input

We modified the above XGBoost Classifier on the basic medical screening dataset to make a 2-way classification model that classifies patients into individuals with ADHD or not using 11 features. We also tried out the same code using other classifiers used earlier to compare the values of the different metrics. We tried this with and without class balance.

COLUMNS CHOSEN FOR FEATURES ARE:

Categorical columns : 'sex', 'eating_probs', 'feeding_dx',

'med_cond_birth', 'birth_oth_calc',

'med_cond_birth_def',

'med_cond_growth', 'growth_oth_calc',

'med_cond_neuro', 'med_cond_visaud'

Numerical columns : ['gest_age']

Results

Number of X features = 11

Number of samples = 192557

Number of individuals affected by ADHD: 44253

Number of unaffected individuals: 148304

CASE 1: WITHOUT CLASS BALANCE

METRIC	XGBOOST		LOGISTIC REGRESSION		DECISION TREES		RANDOM FOREST	
	MEAN	SD	MEAN	SD	MEAN	SD	MEAN	SD

Accuracy	0.76983459	0.00074839 3	0.769367	0.00065391 3	0.768785	0.00089607 9	0.768811	0.00080017 2
Sensitivity	0.983621	0.00122201	0.982819	0.00105313	0.98176	0.00149272	0.980095	0.00096415 4
Specificity	0.053375	0.00434085	0.0540304	0.00216125	0.0550472	0.002859	0.0607419	0.00263818
F1 score	0.690739	0.00175082	0.690699	0.00086351 9	0.690706	0.00107925	0.692663	0.00120247
Precision	0.711627	0.00381884	0.709654	0.00296767	0.707354	0.00381578	0.708462	0.00335527
ROC-AUC Score	0.650792	0.00535774	0.64802	0.00488875	0.642575	0.00624813	0.644886	0.00597992

Filename: model_2way_basicmedical_adhd_1.py

<u>CASE 2: WITH CLASS BALANCE</u>

METRIC	**XGBOOST**		**LOGISTIC REGRESSION**		**DECISION TREES**		**RANDOM FOREST**	
	MEAN	**SD**	**MEAN**	**SD**	**MEAN**	**SD**	**MEAN**	**SD**
Accuracy	0.618557	0.0052565	0.614908	0.00680877	0.614184	0.00471841	0.615868	0.00474355
Sensitivity	0.72316	0.00538263	0.762389	0.00675764	0.730165	0.0059262	0.727521	0.00619465
Specificity	0.513954	0.00987442	0.467427	0.0162587	0.498204	0.00897333	0.504215	0.00968887
F1 score	0.6143	0.005	0.606	0.007	0.6089	0.005	0.6109	0.005

	19	59345	282	91234	06	02617	98	08422
Precision	0.62399	0.00517802	0.625864	0.00626792	0.620689	0.00474982	0.621963	0.00471774
ROC-AUC Score	0.648531	0.00503725	0.646848	0.00528093	0.643474	0.00505755	0.644431	0.00493622

Filename: model_2way_basicmedical_adhd_2.py

EXPERIMENT-15

DATE: 05/03/24

Description and Input

We modified the above XGBoost Classifier on the basic medical screening dataset to make a 2-way classification model that classifies patients into individuals with ASD or not using 11 features. We also tried out the same code using other classifiers used earlier to compare the values of the different metrics. We tried this with and without class balance.

COLUMNS CHOSEN FOR FEATURES ARE:

Categorical columns : 'sex', 'eating_probs', 'feeding_dx',

'med_cond_birth', 'birth_oth_calc',

'med_cond_birth_def',

'med_cond_growth', 'growth_oth_calc',

'med_cond_neuro', 'med_cond_visaud'

Numerical columns : ['gest_age']

Results

Number of X features = 11

Number of samples = 192557

Number of individuals affected by ASD 79631

Number of unaffected individuals: 112926

CASE 1: WITHOUT CLASS BALANCE

METRIC	XGBOOST		LOGISTIC REGRESSION		DECISION TREES		RANDOM FOREST	
	MEAN	SD	MEAN	SD	MEAN	SD	MEAN	SD

Accuracy	0.817233	0.00186797	0.812035	0.00214957	0.816028	0.00190885	0.816605	0.00171887
Sensitivity	0.916565	0.00184422	0.934107	0.00160649	0.9173	0.00192565	0.916901	0.00201713
Specificity	0.67637	0.00460801	0.638922	0.00608115	0.672414	0.00457362	0.674373	0.00439409
F1 score	0.812945	0.00200906	0.805614	0.00244153	0.811567	0.00204757	0.812231	0.00184546
Precision	0.821525	0.00179507	0.821631	0.0017688	0.820616	0.00185936	0.821041	0.00168447
ROC-AUC Score	0.86552	0.00182675	0.861471	0.00193155	0.863416	0.00208145	0.8645	0.00187666

Filename: model_2way_basicmedical_asd_1.py

CASE 2: WITH CLASS BALANCE

METRIC	**XGBOOST**		**LOGISTIC REGRESSION**		**DECISION TREES**		**RANDOM FOREST**	
	MEAN	**SD**	**MEAN**	**SD**	**MEAN**	**SD**	**MEAN**	**SD**
Accuracy	0.797453	0.00290681	0.79017	0.00276503	0.796034	0.00285305	0.7967	0.00299884
Sensitivity	0.910412	0.00206959	0.92852	0.00251392	0.911052	0.00217631	0.911354	0.00212653
Specificity	0.684495	0.00455889	0.651819	0.00426426	0.681016	0.00444535	0.682046	0.00470103

F1 score	0.794834	0.00301352	0.786073	0.00288724	0.793298	0.00295796	0.79399	0.0031119
Precision	0.813454	0.00267131	0.814233	0.00268003	0.812577	0.00265079	0.813169	0.00275482
ROC-AUC Score	0.865677	0.00196406	0.861935	0.0021349	0.863616	0.00202295	0.864549	0.00207078

Filename: model_2way_basicmedical_asd_2.py

EXPERIMENT-16

DATE: 06/03/24

Description and Input

Analysis of the features being used as model input from the different datasets and finding out which feature belongs to which dataset and which are common to multiple datasets

DATASETS CHOSEN FOR MODEL:

1) df1 = basic_medical_screening-2023-07-21.csv
2) df2 = background_history_adult-2023-07-21.csv
3) df3 = background_history_child-2023-07-21.csv
4) df4 = background_history_sibling-2023-07-21.csv
5) df5 = roles_2023-07-17.csv
6) df6 = individuals_registration-2023-07-21.csv

Results

SL.NO	FEATURE	DATASETS TO WHICH IT BELONGS
1)	sex	All 6
2)	mother_highest_education	df2, df3, df4
3)	father_highest_education	df2, df3, df4
4)	annual_household_income	df2, df3, df4
5)	smiled_age_mos	df3, df4
6)	sat_wo_support_age_mos	df3, df4

7)	crawled_age_mos	df3, df4
8)	fed_self_spoon_age_mos	df3, df4
9)	used_words_age_mos	df3, df4
10)	combined_words_age_mos	df3, df4
11)	combined_phrases_age_mos	df3, df4
12)	bladder_trained_age_mos	df3, df4
13)	bowel_trained_age_mos	df3, df4
14)	twin_mult_birth	df3, df4
15)	num_asd_parents	df6
16)	num_asd_siblings	df6

EXPERIMENT-17

DATE: 13/03/24

Description and Input

Successfully implemented the 2-way classification model for ASD using 4 datasets and XGBoost classifier - with and without Class balancing.

DATASETS CHOSEN FOR MODEL:

1) df1 = basic_medical_screening-2023-07-21
2) df2 = background_history_child-2023-07-21
3) df3 = background_history_sibling-2023-07-21
4) df4 = individuals_registration-2023-07-21

Results

Number of X features = 28, Number of samples = 65897 (without class balance),38360(with class balance)

ASD- affected = 46717, Unaffected = 19180

METRIC	XGBOOST		XGBOOST BALANCED	
	MEAN	SD	MEAN	SD
Accuracy	0.836396	0.00511932	0.814964	0.00758401
Sensitivity	0.682586	0.0126251	0.830918	0.00887275
Specificity	0.899544	0.0035826	0.799009	0.00771655
F1 score	0.834504	0.00531861	0.814914	0.00758268
Precision	0.833499	0.00540914	0.8153	0.00761652
ROC-AUC Score	0.895154	0.00558414	0.891596	0.00732261

Filename: model_2way_asd_multidata.py & model_2way_asd_multidata_balanced.py

EXPERIMENT-18

DATE: 13/03/24

Description and Input

Successfully implemented the 2-way classification model for ADHD using 4 datasets and XGBoost classifier - with and without Class balancing.

DATASETS CHOSEN FOR MODEL:

1. df1 = basic_medical_screening-2023-07-21
2. df2 = background_history_child-2023-07-21
3. df3 = background_history_sibling-2023-07-21
4. df4 = individuals_registration-2023-07-21

Results

Number of X features = 28, Number of samples = 65897 (without class balance), 40188 (with class balance)

ADHD- affected = 20094, Unaffected = 45803

METRIC	XGBOOST		XGBOOST BALANCED	
	MEAN	SD	MEAN	SD
Accuracy	0.710138	0.0051083	0.596397	0.00620875
Sensitivity	0.891732	0.00510639	0.585798	0.0104099
Specificity	0.296207	0.0131839	0.606998	0.0110402
F1 score	0.680386	0.00609839	0.596321	0.00620687
Precision	0.682649	0.00694207	0.596471	0.00621796
ROC-AUC Score	0.722084	0.00316626	0.635234	0.00931322

Filename: model_2way_adhd_multidata.py & model_2way_adhd_multidata_balanced.py

EXPERIMENT-19

DATE: 13/03/24

Description and Input

Successfully implemented the 3-way classification model for classification of individuals into patients affected with ASD-only, ADHD-only or both using 4 datasets and XGBoost classifier - with and without Class balancing.

DATASETS CHOSEN FOR MODEL:

1. df1 = basic_medical_screening-2023-07-21
2. df2 = background_history_child-2023-07-21
3. df3 = background_history_sibling-2023-07-21
4. df4 = individuals_registration-2023-07-21

Results

Number of X features = 28, Number of samples = 65897

ADHD- only= 2793, ASD- only = 29416, Both = 17301

Prediction	Value
0	ADHD only
1	ASD only
2	Both

METRIC	XGBOOST		XGBOOST BALANCED	
	MEAN	SD	MEAN	SD
Accuracy	0.637407	0.00591548	0.585153	0.0147246
Sensitivity	0.191192	0.023343	0.724324	0.0173023

Specificity	0.985872	0.00129488	0.797537	0.0198403
F1 score	0.624134	0.00557477	0.582287	0.0155735
Precision	0.62149	0.00572235	0.582548	0.0157496

Filename: model_3way_multidata.py & model_3way_multidata_balanced.py

EXPERIMENT-20

DATE: 20/03/24

Description and Input

Obtained 2 sets of genomic data (.TSV file) on variant counts - ASD prediction in one and ADHD prediction in the other as the last column. This data contains numerical information about the number of each type of variant. I merged each of these datasets with the previous 4 datasets and counted the number of NANs for features(columns) in this merged df - for each TSV file.

DATASETS CHOSEN FOR MODEL:

1) Server_final_updated_data_with_target_all_variants + previous 4 datasets
2) Server_final_updated_data_with_target_all_variants_with_adhd + previous 4 datasets

Results

The results were stored in 2 excel sheets :

1) nan_counts_asd
2) nan_counts_adhd

Filename: dataset_genomic_features_adhd.py and dataset_genomic_features_asd.py

EXPERIMENT-21

DATE: 20/03/24

Description and Input

Successfully implemented the 2-way classification model for ASD and ADHD using 5 datasets and XGBoost classifier - with Class balancing.

DATASETS CHOSEN FOR MODEL:

1) df1 = basic_medical_screening-2023-07-21
2) df2 = background_history_child-2023-07-21
3) df3 = background_history_sibling-2023-07-21
4) df4 = individuals_registration-2023-07-21
5) df5 = server_final_updated_data_with_target_all_variants.tsv

Results

Number of X features = 43

MEAN AND STANDARD DEVIATION OF VARIOUS METRICS - ASD

Metric	Mean	Standard Deviation
Accuracy	0.842813	0.00702391
Sensitivity	0.848936	0.00850976
Specificity	0.836691	0.0117813
F1 score	0.842799	0.0070277
Precision	0.842942	0.00700073

ROC-AUC Score:	0.842814	0.00702212

MEAN AND STANDARD DEVIATION OF VARIOUS METRICS - ADHD

Metric	Mean	Standard Deviation
Accuracy	0.649302	0.00659275
Sensitivity	0.611879	0.0101125
Specificity	0.686727	0.0158678
F1 score	0.648765	0.0064883
Precision	0.650235	0.00685656
ROC-AUC Score	0.649303	0.00659524

Filename: model_2way_asd_genomic.py & model_2way_adhd_genomic.py

EXTRA - NATURE PAPER FEATURES PRESENT IN BASIC MEDICAL SCREENING AND TRIED

Features

Preterm birth

Hypoxia at birth

Fetal alcohol syndrome - did not help alone

Bleeding into brain

Traumatic brain injury

Brain infection

Infection in pregnancy

Lead poisoning

EXPERIMENT-22

DATE: 28/03/24, 05/04/24

Description and Input

Successfully implemented the 2-way classification model using multiple datasets and XGBoost classifier - with Class balancing for classifying individuals into:

1. affected by ASD only v/s both ASD and ADHD
2. Affected by ADHD only v/s both ASD and ADHD

DATASETS CHOSEN FOR MODEL:

1) df1 = basic_medical_screening-2023-07-21
2) df2 = background_history_child-2023-07-21
3) df3 = background_history_sibling-2023-07-21
4) df4 = individuals_registration-2023-07-21

Results

CASE 1: ADHD ONLY V/S BOTH

The number of individuals affected is -> ADHD only: 2793, Both: 17301

MEAN AND STANDARD DEVIATION OF VARIOUS METRICS

Metric	Mean	Standard Deviation
Accuracy	0.725921	0.0180678
Sensitivity	0.740764	0.0294121
Specificity	0.711065	0.0225297
F1 score	0.725764	0.0180266

Precision	0.726458	0.0183136
ROC-AUC Score:	0.797711	0.0185465

Filename: model_adhdonly_both.py

CASE 2: ASD ONLY V/S BOTH

The number of individuals affected is -> ASD only: 29416, Both: 17301

MEAN AND STANDARD DEVIATION OF VARIOUS METRICS

Metric	Mean	Standard Deviation
Accuracy	0.660858	0.00533556
Sensitivity	0.59141	0.00780647
Specificity	0.730305	0.00795559
F1 score	0.659202	0.00535843
Precision	0.664047	0.00550446
ROC-AUC Score:	0.715058	0.00476497

Filename: model_asdonly_both.py

EXPERIMENT-23

DATE: 05/04/24

Description and Input

Successfully implemented the 2-way classification model removing dataset 3 and XGBoost classifier - with Class balancing for classifying individuals into:

1. Affected by ASD only v/s both ASD and ADHD
2. Affected by ADHD only v/s both ASD and ADHD -> This crashed as no of individuals with ADHD only became 0

DATASETS CHOSEN FOR MODEL:

1) df1 = basic_medical_screening-2023-07-21
2) df2 = background_history_child-2023-07-21
3) df3 = background_history_sibling-2023-07-21 - REMOVED DATASET
4) df4 = individuals_registration-2023-07-21

Results

ASD ONLY V/S BOTH

The number of individuals affected are:

ASD only: 29395

Both: 17301

MEAN AND STANDARD DEVIATION OF VARIOUS METRICS

Metric	Mean	Standard Deviation
Accuracy	0.657419	0.00570264
Sensitivity	0.578291	0.00892872

| Specificity | 0.736547 | 0.00660572 |

+----------+----------+----------+

| F1 score | 0.655249 | 0.00582323 |

+----------+----------+----------+

| Precision | 0.661479 | 0.00570259 |

+----------+----------+----------+

| ROC-AUC Score: | 0.707441 | 0.00707333 |

+----------+----------+----------+

Filename: model_adhdonly_both_nodf3.py

EXPERIMENT-24

DATE: 05/04/24

Description and Input

Successfully implemented the 2-way classification model using the basic medical screening dataset and XGBoost classifier - with Class balancing for classifying individuals into:

1. Affected by ASD only v/s both ASD and ADHD
2. Affected by ADHD only v/s both ASD and ADHD

DATASET CHOSEN FOR MODEL:

basic_medical_screening-2023-07-21

Results

CASE1: ASD ONLY V/S BOTH

The number of individuals affected are:

ASD only: 49804

Both: 29827

Number of features = 11+5(nature paper) = 16, No of samples = 59654

MEAN AND STANDARD DEVIATION OF VARIOUS METRICS

Metric	Mean	Standard Deviation
Accuracy	0.546015	0.00541671
Sensitivity	0.672009	0.0152007
Specificity	0.420022	0.0132683

| F1 score | 0.538609 | 0.00540818 |

+----------+----------+----------+

| Precision | 0.549202 | 0.00597462 |

+----------+----------+----------+

| ROC-AUC Score: | 0.563649 | 0.00699528 |

+----------+----------+----------+

Filename: model_asdonly_both_basic_medical.py

CASE2: ADHD ONLY V/S BOTH

Sample size = 28852, Features = 11

The number of individuals affected are:

ADHD only: 14426, Both: 29827

MEAN AND STANDARD DEVIATION OF VARIOUS METRICS

+----------+----------+----------+

| Metric | Mean | Standard Deviation |

+==========+==========+==========+

| Accuracy | 0.774332 | 0.00629648 |

+----------+----------+----------+

| Sensitivity | 0.84944 | 0.00788158 |

+----------+----------+----------+

| Specificity | 0.699226 | 0.0111601 |

+----------+----------+----------+

| F1 score | 0.773037 | 0.00642595 |

+----------+----------+----------+

| Precision | 0.780723 | 0.00613561 |

+----------+----------+----------+

| ROC-AUC Score: | 0.837856 | 0.00623604 |

+----------+----------+----------+

Filename: model_adhdonly_both_basic_medical.py

EXPERIMENT-25

DATE: 18/04/24

Description and Input

A repetition of exp 24 - classifying individuals into those Affected by ADHD only v/s both ASD and ADHD using 4 different classifiers and comparing results -> **THIS IS THE FINAL RESULT FOR POSTER**

DATASET CHOSEN FOR MODEL:

basic_medical_screening-2023-07-21

Results

Sample size = 28852, Features = 11

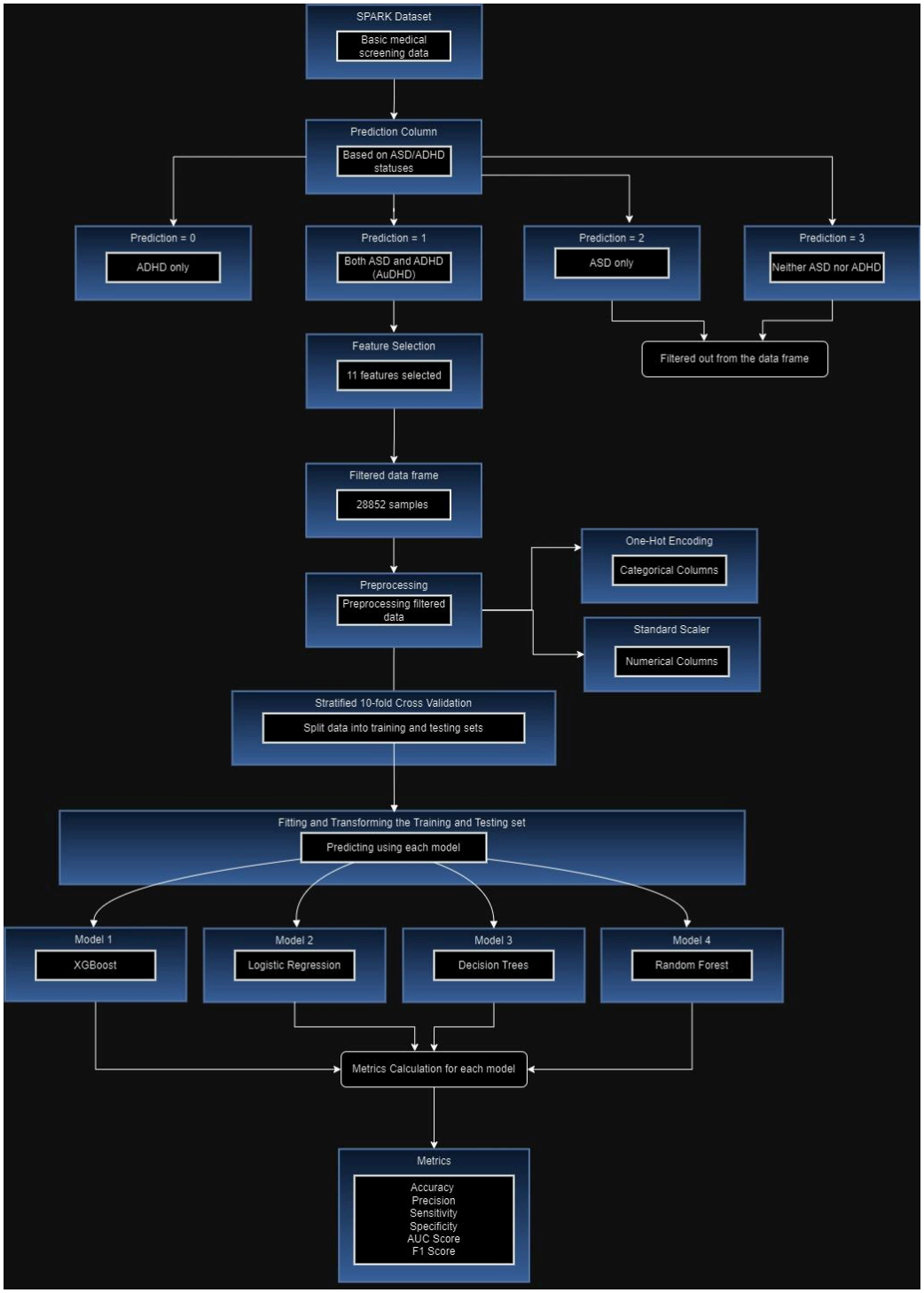
The number of individuals affected are:

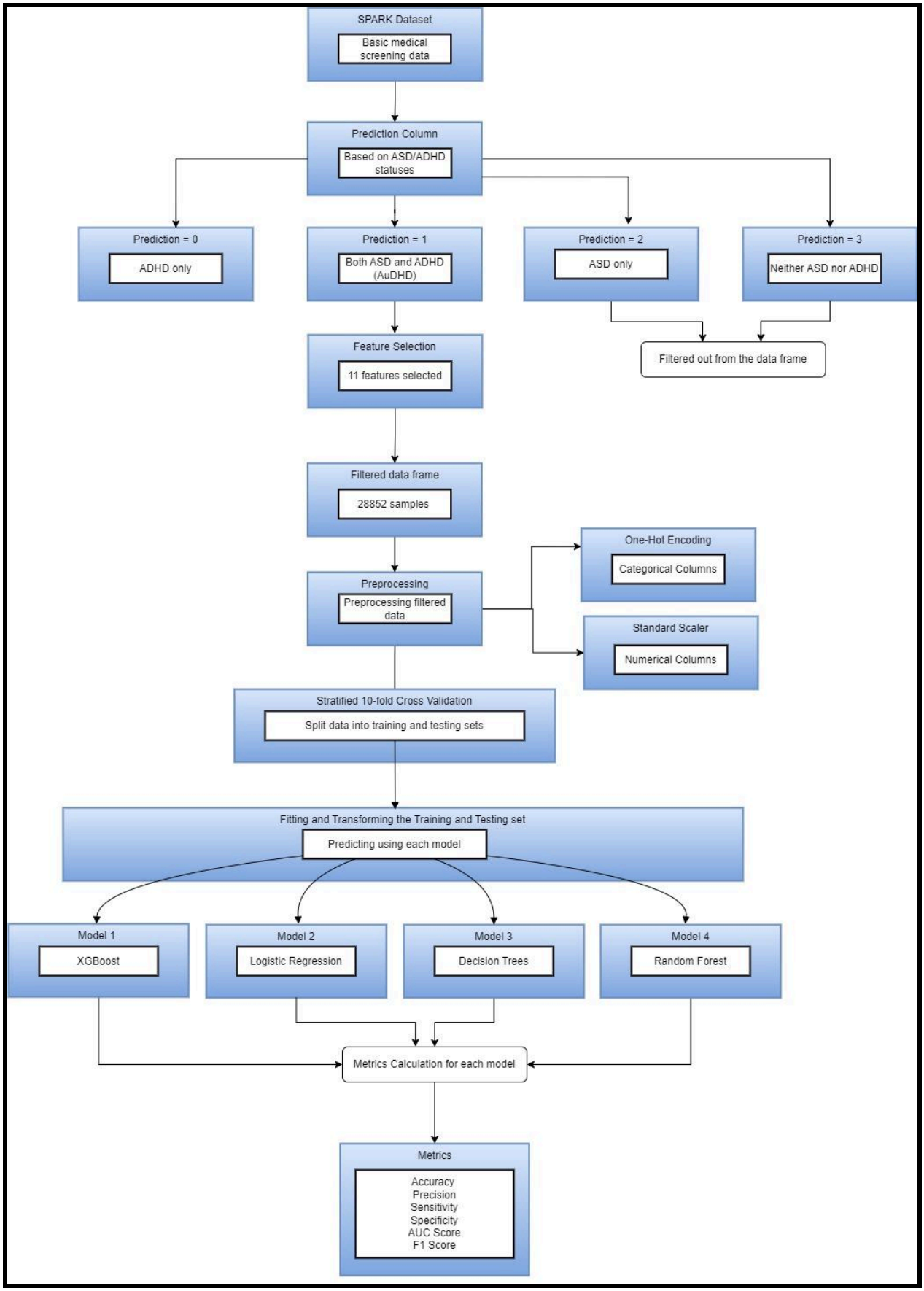
ADHD only: 14426, Both: 29827

METRIC	XGBOOST		LOGISTIC REGRESSION		DECISION TREES		RANDOM FOREST	
	MEAN	SD	MEAN	SD	MEAN	SD	MEAN	SD
Accuracy	0.774332	0.00629648	0.763449	0.00561958	0.769098	0.0066501	0.771871	0.00624375
Sensitivity	0.84944	0.00788158	0.872315	0.00733369	0.852351	0.00829936	0.850618	0.00836235
Specificity	0.699226	0.0111601	0.654584	0.0094651	0.685847	0.0110038	0.693125	0.0103141
F1 score	0.773037	0.00613561	0.760599	0.00577574	0.767472	0.00678412	0.770434	0.006342

Precision	0.780723	0.00613561	0.776608	0.0057287	0.776827	0.00661005	0.778843	0.00626712
ROC-AUC Score	0.837856	0.00623604	0.776608	0.00680111	0.829393	0.00621205	0.83409	0.00598536

FLOWCHART FOR WORKFLOW




SPARK Dataset
Basic medical screening data
Prediction Column
Based on ASD/ADHD statuses
Prediction = 0
ADHD only
Prediction = 1
Both ASD and ADHD (AuDHD)
Prediction = 2
ASD only
Prediction = 3
Neither ASD nor ADHD
Filtered out from the data frame
Feature Selection
11 features selected
Filtered data frame
28852 samples
Preprocessing
Preprocessing filtered data
One-Hot Encoding
Categorical Columns
Standard Scaler
Numerical Columns
Stratified 10-fold Cross Validation
Split data into training and testing sets
Fitting and Transforming the Training and Testing set
Predicting using each model
Model 1
XGBoost
Model 2
Logistic Regression
Model 3
Decision Trees
Model 4
Random Forest
Metrics Calculation for each model
Metrics
Accuracy
Precision
Sensitivity
Specificity
AUC Score
F1 Score

EXPERIMENT-26

DATE: 01/05/24

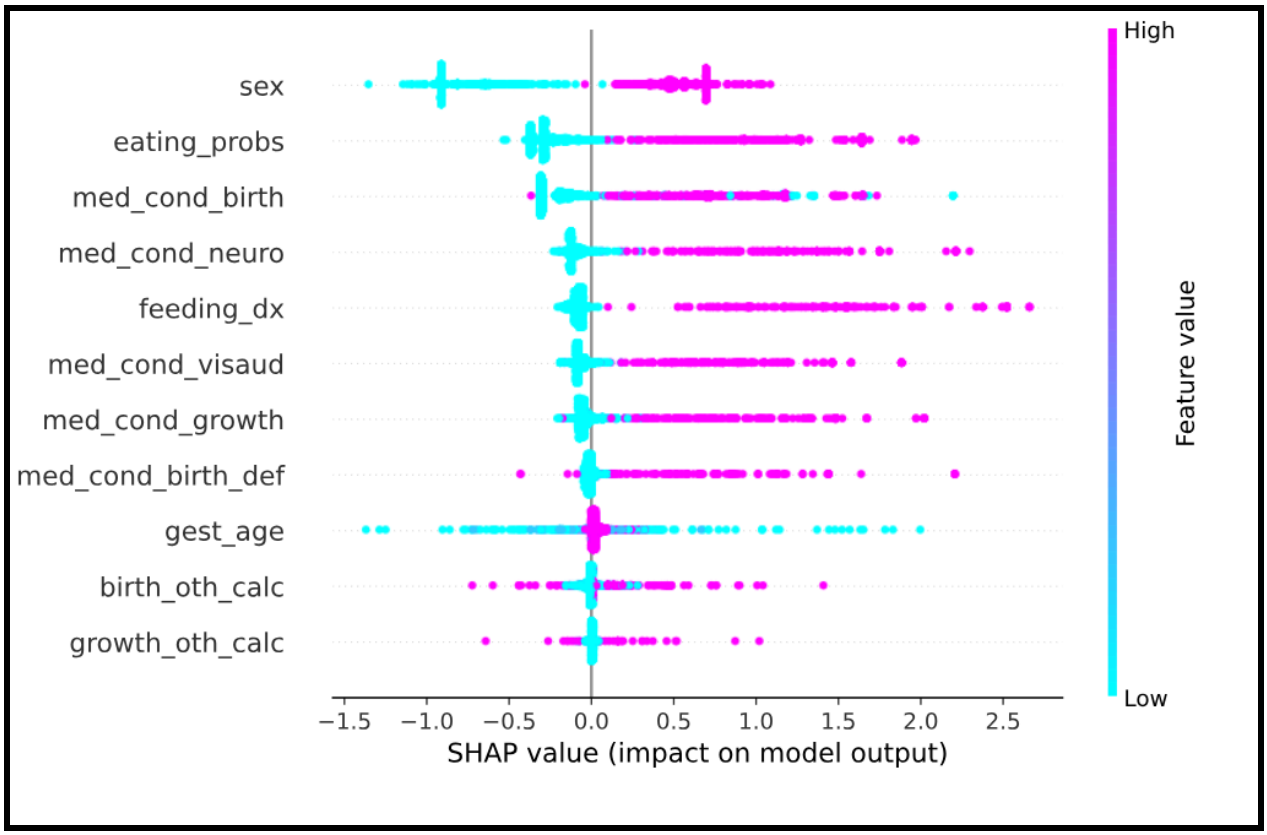
Description and Input

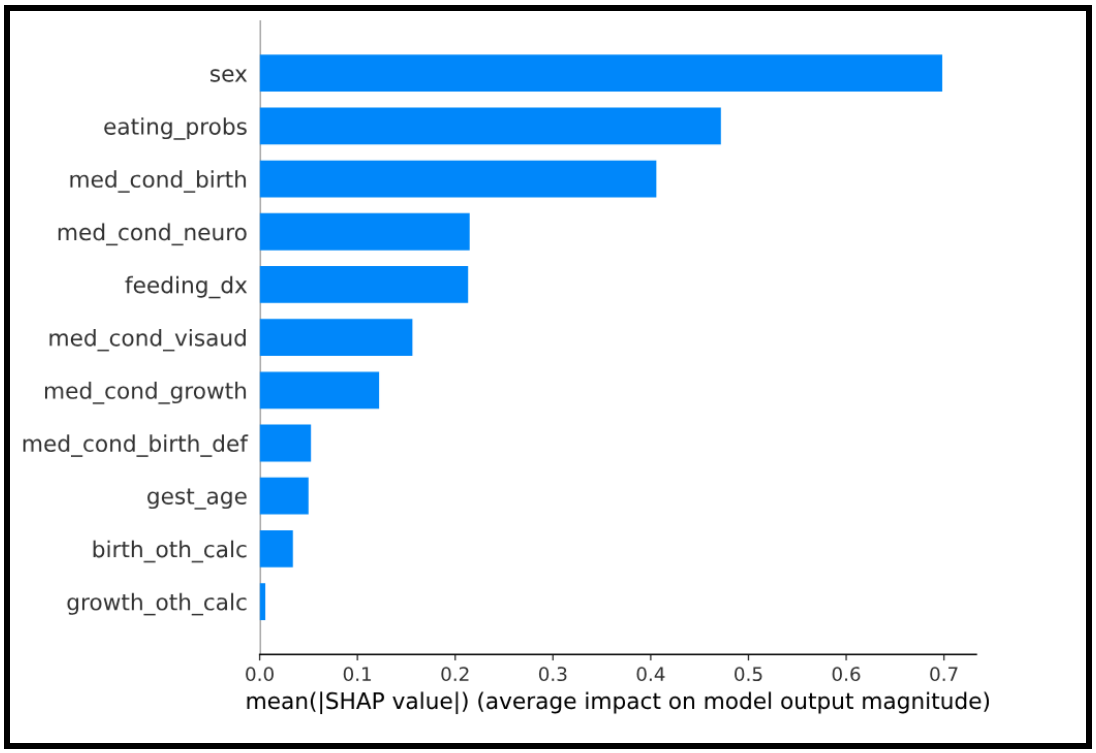
The model used in experiment 24 (XGBoost classifier) was subjected to SHAP analysis after preprocessing the encoded categorical columns. Different plots were obtained

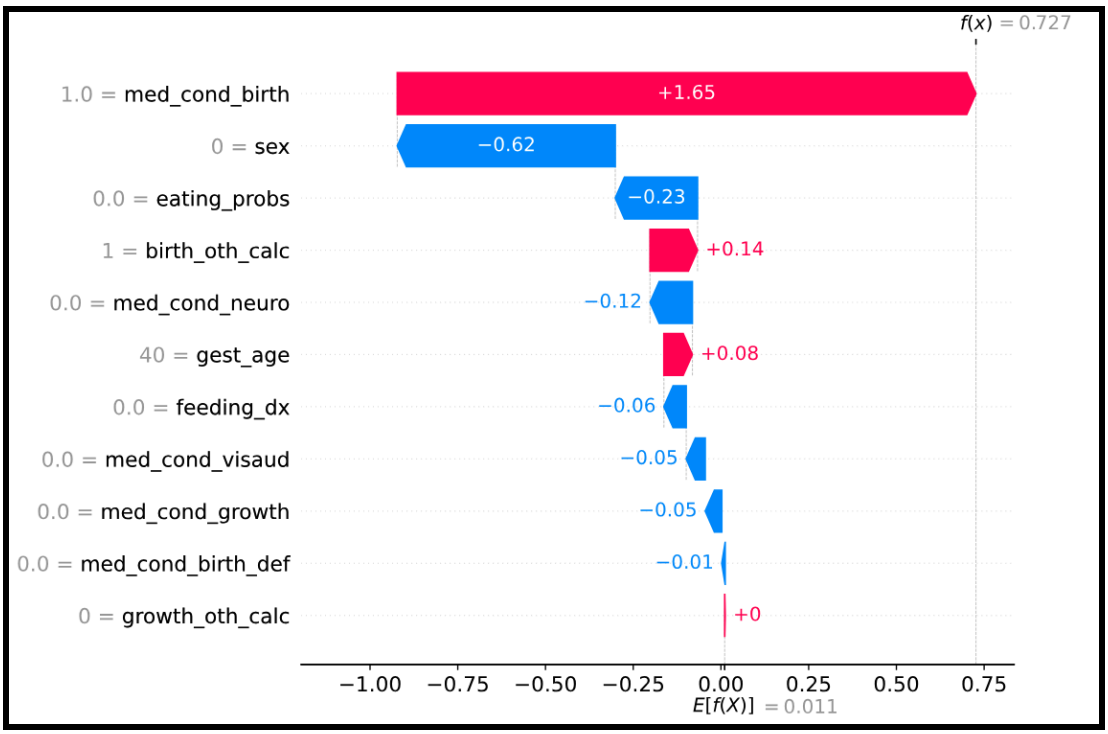
DATASET CHOSEN FOR MODEL:

basic_medical_screening-2023-07-21

Results







Filename: model_adhdonly_both_basic_medical_new.py

EXPERIMENT-27

DATE: 03/05/24

Description and Input

The model used in experiment 24 (ADHD only v/s AuDHD) was implemented without using sex as one of the features

DATASET CHOSEN FOR MODEL:

basic_medical_screening-2023-07-21

Results

Sample size = 28852, Features = 10

The number of individuals affected are:

ADHD only: 14426, Both: 29827

METRIC	XGBOOST		LOGISTIC REGRESSION		DECISION TREES		RANDOM FOREST	
	MEAN	SD	MEAN	SD	MEAN	SD	MEAN	SD
Accuracy	0.775233	0.00546441	0.76993	0.00648469	0.771732	0.00502226	0.773361	0.00563606
Sensitivity	0.847914	0.00814457	0.853044	0.00851792	0.850064	0.00809217	0.848954	0.00824876
Specificity	0.702553	0.00963431	0.686818	0.0144099	0.693402	0.00827227	0.69777	0.0097316
F1 score	0.774027	0.00553714	0.768301	0.00674435	0.770313	0.00506351	0.772046	0.00571314

Precision	0.781233	0.00551508	0.777702	0.0059043	0.778622	0.00525703	0.779815	0.00569779
ROC-AUC Score	0.78951	0.0066396	0.789977	0.00632248	0.784112	0.00576395	0.786397	0.00687734

Filename: model_adhdonly_both_basic_medical_2.py

EXPERIMENT-28

DATE: 04/05/24

Description and Input

The plots for model used in experiment 24 (ADHD only v/s AuDHD) was implemented

DATASET CHOSEN FOR MODEL:

basic_medical_screening-2023-07-21

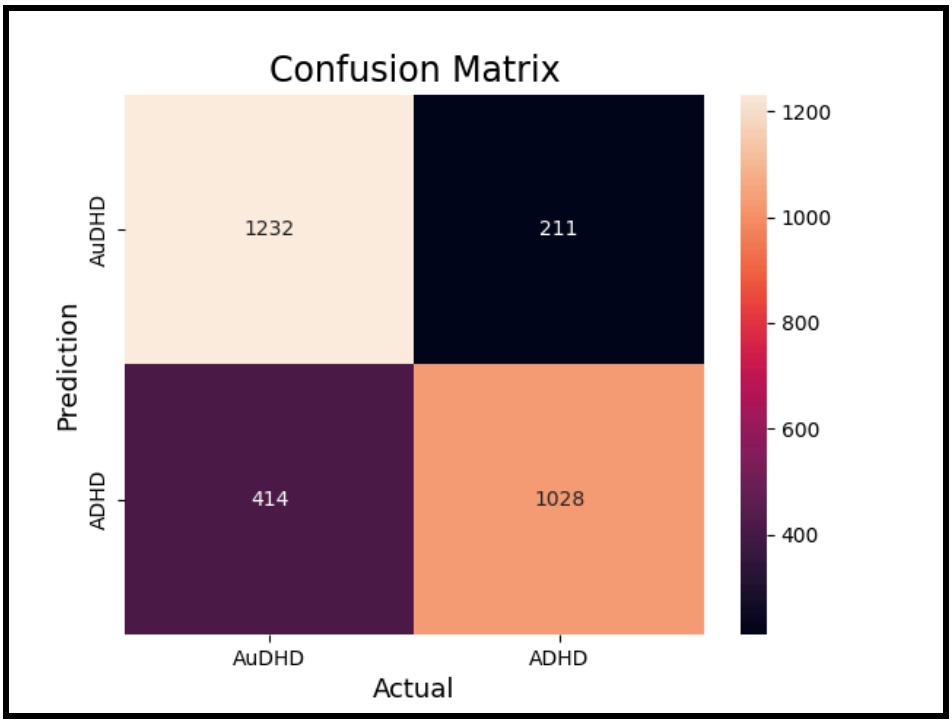
Results

Sample size = 28852, Features = 11

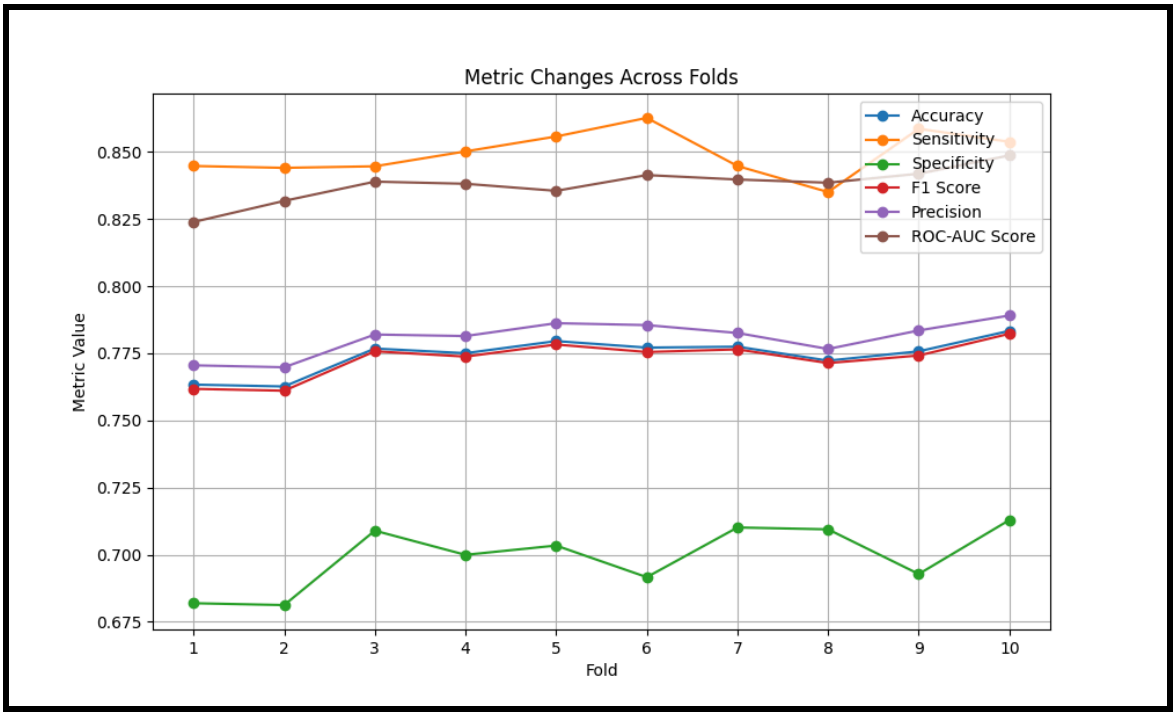
The number of individuals affected are:

ADHD only: 14426, Both: 29827

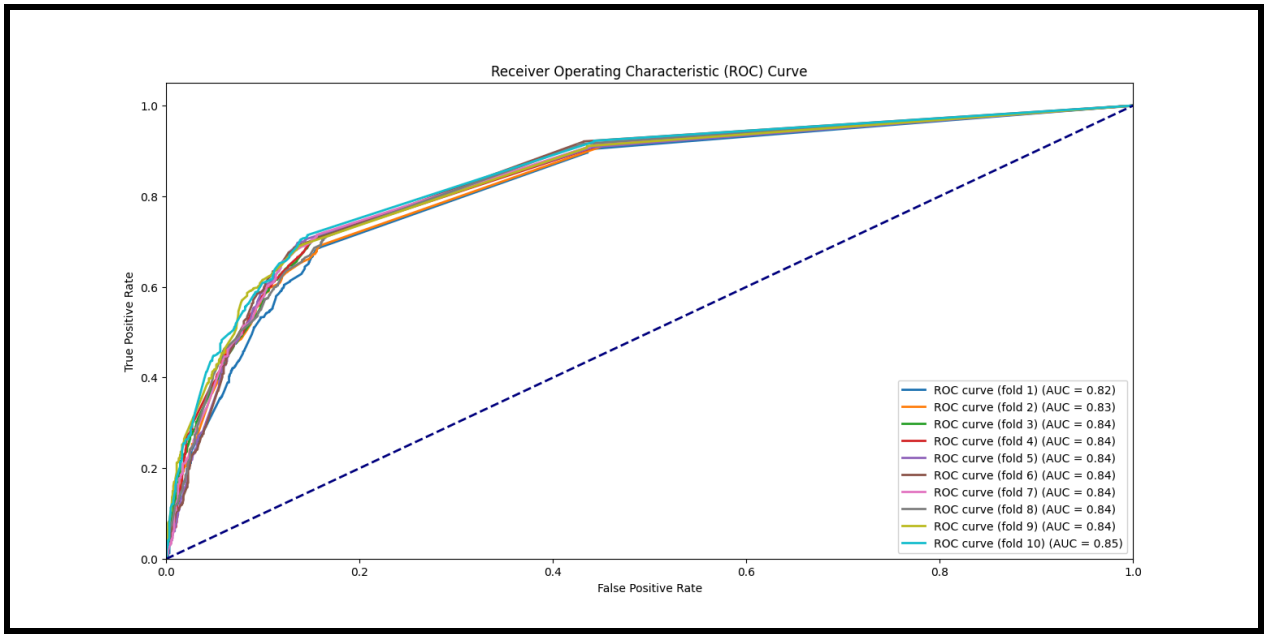
Confusion matrix plot



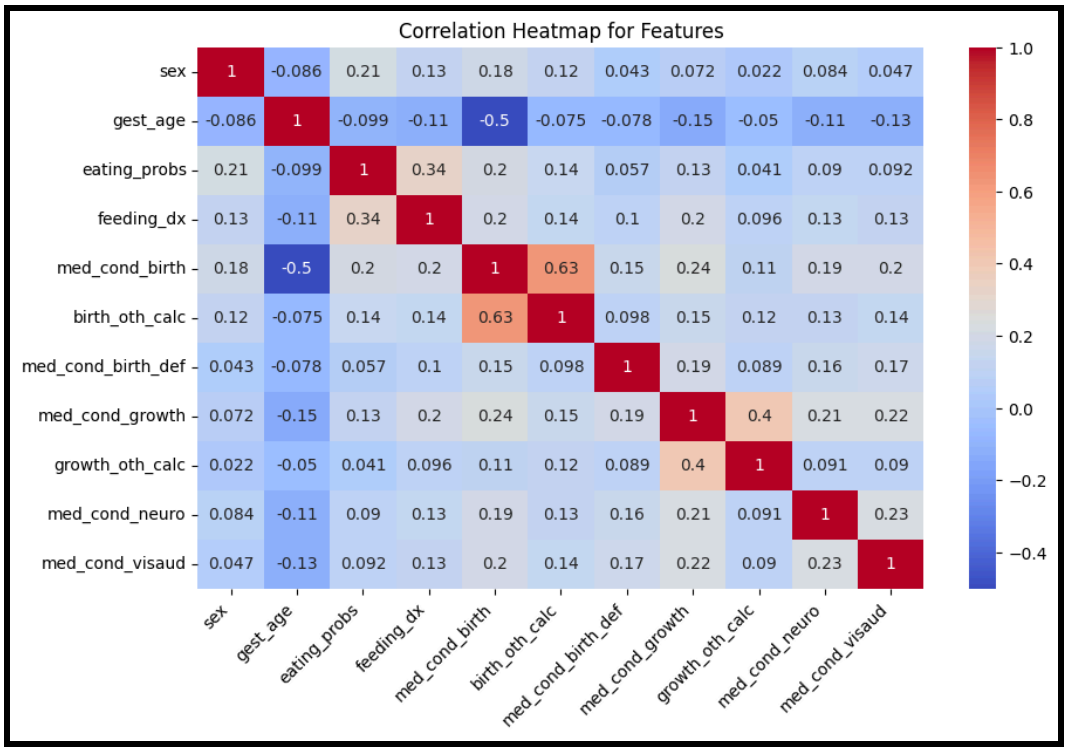
Metrics fold changes



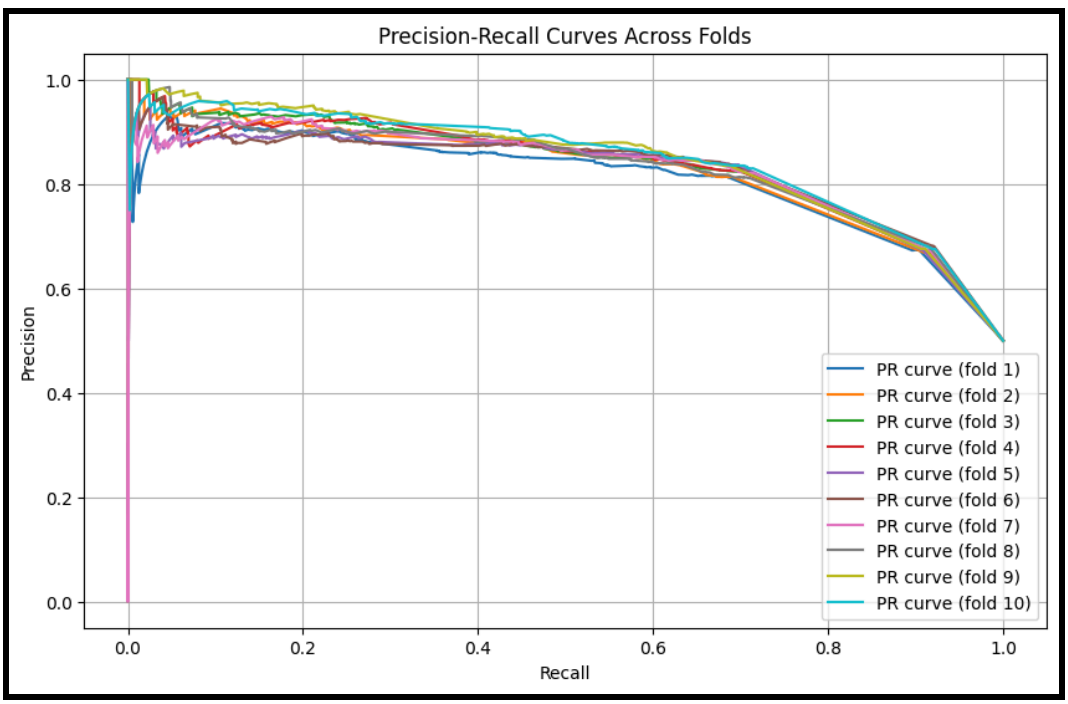
ROC-Curve(All 10 folds)



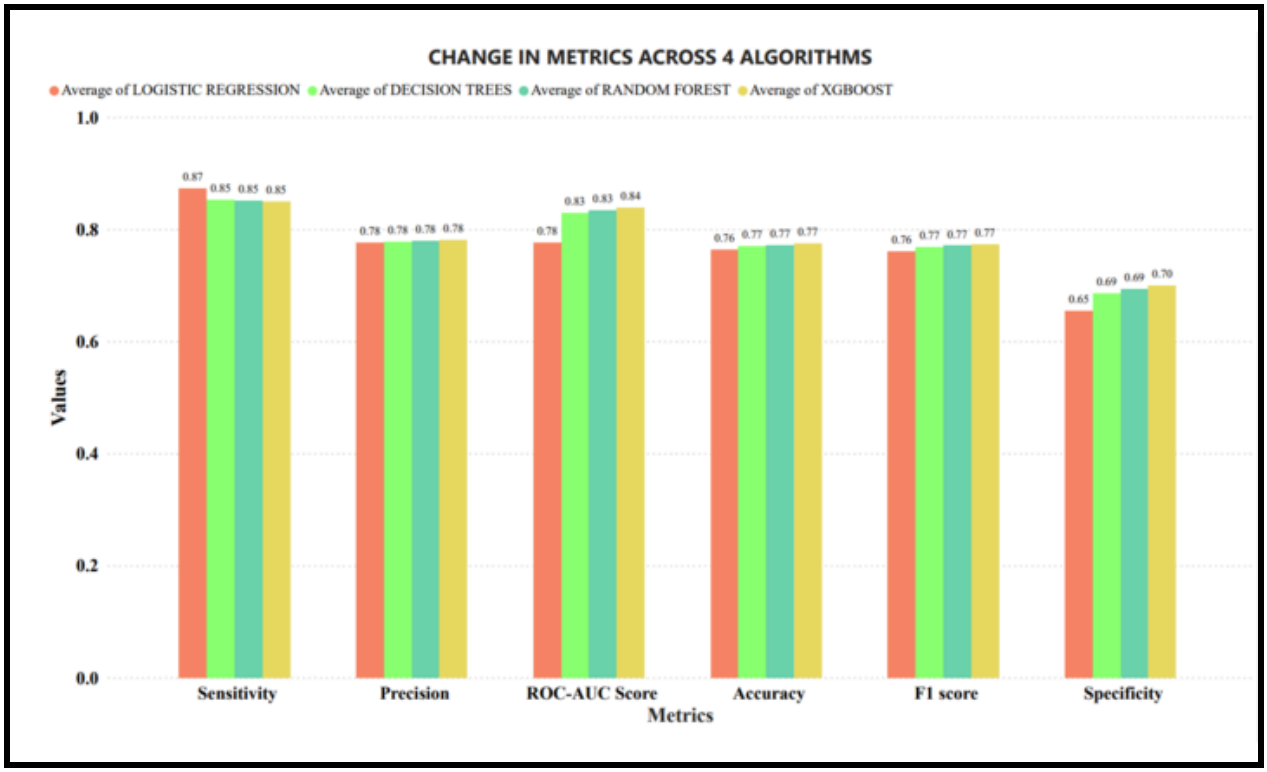
Feature Correlation Heatmap



Precision-Recall curve



Metrics Change Across the 4 Classifiers



Filename: model_adhdonly_both_basic_medical_new.py

EXPERIMENT-29

DATE: 05/05/24

Description and Input

The model used in experiment 24 (ADHD only v/s AuDHD) was implemented by class balancing based on sex instead of AuDHD/ADHD

DATASET CHOSEN FOR MODEL:

basic_medical_screening-2023-07-21

Results

The number of males in the dataset are: 27382

The number of females in the dataset are: 16871

The number of individuals affected are:

ADHD only: 14426, Both: 29827

Sample size = 33742, Features = 11

METRIC	XGBOOST		LOGISTIC REGRESSION		DECISION TREES		RANDOM FOREST	
	MEAN	SD	MEAN	SD	MEAN	SD	MEAN	SD
Accuracy	0.794736	0.00697491	0.794558	0.00638615	0.790617	0.00742432	0.792869	0.00725836
Sensitivity	0.652253	0.0112078	0.652577	0.0114871	0.654597	0.0117495	0.653951	0.0113246
Specificity	0.877246	0.00634492	0.876778	0.0054623	0.869384	0.0066536	0.873315	0.00647984
F1 score	0.791138	0.00711084	0.790991	0.00658026	0.787449	0.0075520	0.789507	0.00738632

						6		
Precision	0.791842	0.0072634	0.791651	0.00665449	0.787596	0.00768917	0.7899	0.00753867
ROC-AUC Score	0.839879	0.00626137	0.835232	0.00668988	0.834819	0.00676621	0.837127	0.00693749

Filename: model_adhdonly_both_basic_medical_sex_classbalance.py

EXPERIMENT-30

DATE: 06/05/24

Description and Input

The model used in experiment 24 (ADHD only v/s AuDHD) was implemented by age stratification. 3 sets of experiments were done (i.e. 3 age groups):

1. 0-2 years (includes age 0 and 1 years)
2. 2-6 years (includes age 2 to 5 years)
3. 6-10 years (inclues age 6 to 9 years)

DATASET CHOSEN FOR MODEL:

basic_medical_screening-2023-07-21

Results

CASE 1: 0-2 YEARS —--> NOT RELIABLE

The number of individuals affected are:

ADHD only: 19

Both: 11

Sample size = 22, Features = 11

METRIC	XGBOOST		LOGISTIC REGRESSION		DECISION TREES		RANDOM FOREST	
	MEAN	SD	MEAN	SD	MEAN	SD	MEAN	SD
Accuracy	0.516667	0.283333	0.733333	0.351188				
Sensitivity	0.65	0.45	0.8	0.4				
Specificity	0.35	0.45	0.7	0.458258				

F1 score	0.433333	0.3	0.683333	0.397562				
Precision	0.391667	0.316338	0.661111	0.420317				
ROC-AUC Score	0.6	0.43589	0.7	0.458258				

CASE 2: 2-6 YEARS

The number of individuals affected are:

ADHD only: 254

Both: 2712

Sample size = 508, Features = 11

METRIC	XGBOOST		LOGISTIC REGRESSION		DECISION TREES		RANDOM FOREST	
	MEAN	SD	MEAN	SD	MEAN	SD	MEAN	SD
Accuracy	0.627804	0.0592887	0.627961	0.0597866	0.610431	0.0645274	0.586627	0.0638246
Sensitivity	0.692615	0.0475223	0.672615	0.100531	0.685077	0.0605929	0.649077	0.118875
Specificity	0.563385	0.111669	0.582769	0.0992082	0.535692	0.115115	0.523846	0.0837275
F1 score	0.624627	0.061317	0.624582	0.0610932	0.606301	0.06641	0.582442	0.0631818
Precision	0.631581	0.0605925	0.631822	0.0592133	0.614226	0.0648531	0.591423	0.0677295
ROC-AUC Score	0.620969	0.0660103	0.657526	0.0704169	0.583714	0.0849531	0.596671	0.0773416

CASE 3: 6-10 YEARS

The number of individuals affected are:

ADHD only: 1295

Both: 8433

Sample size = 2590, Features = 11

METRIC	XGBOOST		LOGISTIC REGRESSION		DECISION TREES		RANDOM FOREST	
	MEAN	SD	MEAN	SD	MEAN	SD	MEAN	SD
Accuracy	0.643629	0.0351775	0.647104	0.0354372	0.620077	0.0245531	0.644402	0.0244466
Sensitivity	0.752153	0.0365042	0.742934	0.040305	0.729004	0.0278377	0.716577	0.0285023
Specificity	0.535075	0.052298	0.551366	0.0633908	0.511187	0.0310315	0.572194	0.053503
F1 score	0.638991	0.0365761	0.643172	0.0366321	0.615403	0.0251177	0.642005	0.0257016
Precision	0.650905	0.0354967	0.653502	0.0349924	0.626202	0.0258833	0.647987	0.0231872
ROC-AUC Score	0.683581	0.0333613	0.702239	0.0353258	0.660552	0.0208956	0.684112	0.0246874

Filename: model_adhdonly_both_basic_medical_age_strat.py

EXPERIMENT-31

DATE: 08/05/24

Description and Input

The model used in experiment 24 (ADHD only v/s AuDHD) was subjected to various types of statistical analysis

DATASET CHOSEN FOR MODEL:

basic_medical_screening-2023-07-21

Results

1. **Confusion Matrix Analysis:**

   Chi-Squared Test:

   Chi-Squared Statistic: 942.9356814557027

   P-value: 4.553633162985258e-207

   Fisher's Exact Test:

   Odds Ratio: 14.498431668841725

   P-value: 1.0934394995396334e-221

   Interpretation:

   Based on the results of the statistical tests:

   1.Chi-Squared Test:

   - Chi-Squared Statistic: 942.94
   - P-value: 4.55e-207

The chi-squared statistic is a measure of the discrepancy between the observed and expected frequencies in the contingency table. A larger chi-squared statistic indicates a greater discrepancy. The extremely small p-value (close to zero) suggests strong evidence against the null hypothesis, indicating that there is a significant association between the predicted and actual class labels.

2. Fisher's Exact Test:

- Odds Ratio: 14.50
- P-value: 1.09e-221

Fisher's exact test is another method for determining the association between categorical variables, particularly when the sample size is small. The odds ratio measures the strength of association between the two variables. In this case, the odds ratio of 14.50 indicates a strong association. The very small p-value (close to zero) suggests strong evidence against the null hypothesis, indicating that there is a significant association between the predicted and actual class labels.

In summary, both tests provide strong evidence to reject the null hypothesis, indicating a significant association between the predicted and actual class labels. This suggests that the model's predictions are not random and are indeed associated with the true class labels.