Machine Learning

Project Report

Project Title

Application of machine learning in early detection of ADHD

Submitted By

Name: Himel Mazumder

ID: 011 161 045

Section: B

Contents ----------

Abstract

Attention deficit hyperactivity disorder (ADHD) is one of the most prevalent neurobehavioral illnesses in children. This project aims to find out a machine learning based approach to detect whether a child having ADHD or not. The dataset for this project was collected from the 2019 National Survey of Children's Health (2019 NSCH Data Release). For the feature engineering/evaluation, two statistical hypothesis testing methods were used. Missing values in most of the feature variables of the dataset were handled using KNN imputation method. At this stage, around 2557 (9.95%) of children were having ADHD, and the rest of the children were healthy. Due to such high imbalance in dataset class label, I adopted oversampling approach to make class label balanced. Different encoding techniques were applied to target variable and different types of feature variables. Standardization was used to scale the values of specific feature variables. Four ML-based classifiers then were adopted for the prediction of children with ADHD. These are Naïve Bayes (NB), Random Forest (RF), K-Nearest Neighbor (KNN), and Multi-Layer perceptron (MLP). These classifiers were trained with K-Fold Cross Validation technique. Findings showed that RF based classifier provided the highest classification accuracy of 92.8%, sensitivity of 97.4%, specificity of 88.3% with a value of k equals 10. In the end, I performed t test on the two most performing classifiers namely as RF and KNN to see if there exists a significant difference between the two models. In the t test, significant difference between the two models was established and random forest again emerged victorious. This system will be helpful for early detection and diagnosis of ADHD.

1 Introduction

Attention deficit hyperactivity disorder (ADHD) is one of the most frequent neurodevelopmental behavioral disorders among children. Children with ADHD shows symptoms such as hyperactivity, inattention, and impulsivity. According to the Centers for Disease Control (CDC) and prevention, the number of children in the USA who have been diagnosed with ADHD has fluctuated over time as follows: about 4.4 million children between the ages of 2 and 17 years were diagnosed with ADHD in 2003, 5.4 million children in 2007, 6.4 million children in 2011, and 6.1 million children in 2016. About 12.9% of male children and 5.6% of females were diagnosed with ADHD. It can be understood that the number of children with ADHD has been increasing day by day. Researchers are trying to determine the risk factors to reduce the number of children with ADHD. A study showed that genetic factors played a significant role and were linked with ADHD. Genetic factors are responsible for almost 75% of the risk of ADHD in younger children. Besides the genetic factors, there are several risk factors for ADHD such as brain injury, alcohol/tobacco use during pregnancy, and premature delivery. Studies also showed that age, sex, asthma, race, anxiety, depression, obesity, cigarette smoking, and

socio-economic status were also associated with children with ADHD. As a result, it has become important to propose a prediction model. In this regard, in comparison with conventional approaches, machine learning based models may be used for prediction.

2 Dataset

The dataset used for this project was extracted from the 2019 NSCH Data Release. This is a United States government official survey based on child health and well-being. Participants were 29433 youths aged 0 to 17 years from the NSCH, 2019. The dataset contained a number of instances with missing values. I excluded a porting of such instances. For the rest, I imputed the missing values. After that, about 25680 instances were considered for the final analysis. Among them, 2557 children were with ADHD and the rest of the children were healthy.

2.1 Outcome/Target Variables

I considered the column named "K2Q31A" as the outcome variables. In the dataset variable list manual, the K2Q31A column's description is as follows, "Has a doctor or other health care provider EVER told you that this child has Attention Deficit Disorder or Attention-Deficit/Hyperactivity Disorder, that is, ADD or ADHD?". This column has two unique values. These are

1. yes
2. no

As the Outcome variable name were written in codes in the original dataset, I changed its name to "ADD/ADHD" for better readability.

2.2 Predictor/Feature Variables

There was total 443 columns in the original dataset. Not all of them were related to ADHD. Based on literature review, the predictor/feature variables I found related to ADHD are the following.

Variable Names	Question Types	Categories
Child's age	Child age in years	Continuous
Sex	Sex of the child	Male and Female
Mother's age	Mother's age in years	Continuous
Allergies	Has a doctor ever told you that the selected child (S.C.) has allergies?	Yes and No
Arthritis	Has a doctor ever told you that S.C. has arthritis?	Yes and No
Asthma	Has a doctor ever told you that S.C. has asthma?	Yes and No
Brain injury	Has a doctor ever told you that S.C. has a brain injury	Yes and No
Headaches	Has a doctor ever told you that S.C. has frequent or severe headaches or migraine?	Yes and No
Anxiety	Has a doctor ever told you that S.C. had anxiety problems?	Yes and No
Depression	Has a doctor ever told you that S.C. had depression problems?	Yes and No
Insurance	Is S.C. currently covered by any kind of health insurance plan?	Yes and No
Alcohol	To the best of your knowledge, has S.C. ever experienced lived with anyone who had a problem with alcohol or drugs	Yes and No
Race	What is this child's race?	White, Black, and Other
Family structure	Family structure	Two-parent-biological/step/adopted and Other-single mother/ father/other
Mother's education	Highest level of education	<High school, High school, and > High school
Very LBW	Is child-birth weight <1.5 kg?	Yes and No
LBW	Is child-birth weight <2.5 kg?	Yes and No
Premature	Premature birth (>3 weeks before due date)	Yes and No
Poverty	Income-based on federal poverty level status	<200% and >=200%

So, I only considered the aforementioned columns as predictor/feature variables for the project. As some of the feature variable names were written in codes in the original dataset, I, for the purpose of better readability, changed those to names that are relevant and easily understandable with the help of dataset variable list manual.

2.3 Derived Feature Variables

There are two risk factors namely as Mother's Education and Poverty/Poverty Level weren't given directly in the original dataset. These two feature variables were derived from other different columns in the dataset.

2.3.1 Mother's Education

Mother's Education was derived from the following columns.

'A1_SEX' -> Gender of Adult 1

'A2_SEX' -> Gender of Adult 2

'A1_RELATION' -> How Adult 1 is related to this child.

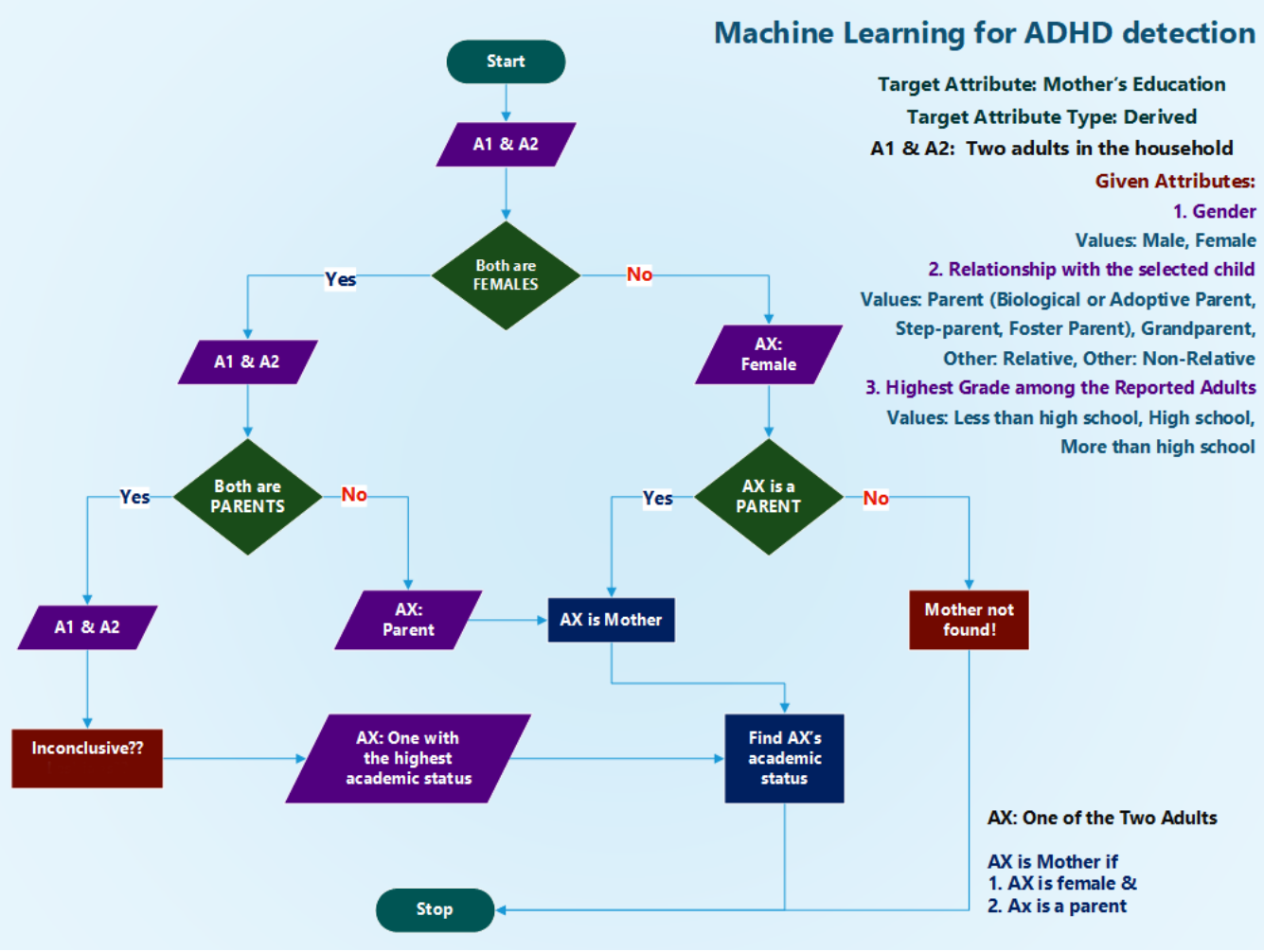
'A2_RELATION' -> How Adult 2 is related to this child.

'A1_GRADE' -> Highest grade or level of school Adult 1 has completed.

'A2_GRADE' -> Highest grade or level of school Adult 2 has completed.

'HIGRADE' -> Highest Level of Education among Reported Adults

The flowchart for the method of derivation is given below.



2.3.2 Poverty Level

In the original dataset, multiple imputation method was used for handling missing values in the poverty level column. That resulted in six different implicates of the poverty level column along with an extra imputation flag column which indicates whether the value of

an instance was imputed or not. If the value of imputation flag column is 1 for an instance then it's value was imputed. Poverty Level was derived from the following columns using multiple imputation estimate technique.

FPL_I1 - Family Poverty Ratio, First Implicate

FPL_I2 - Family Poverty Ratio, Second Implicate

FPL_I3 - Family Poverty Ratio, Third Implicate

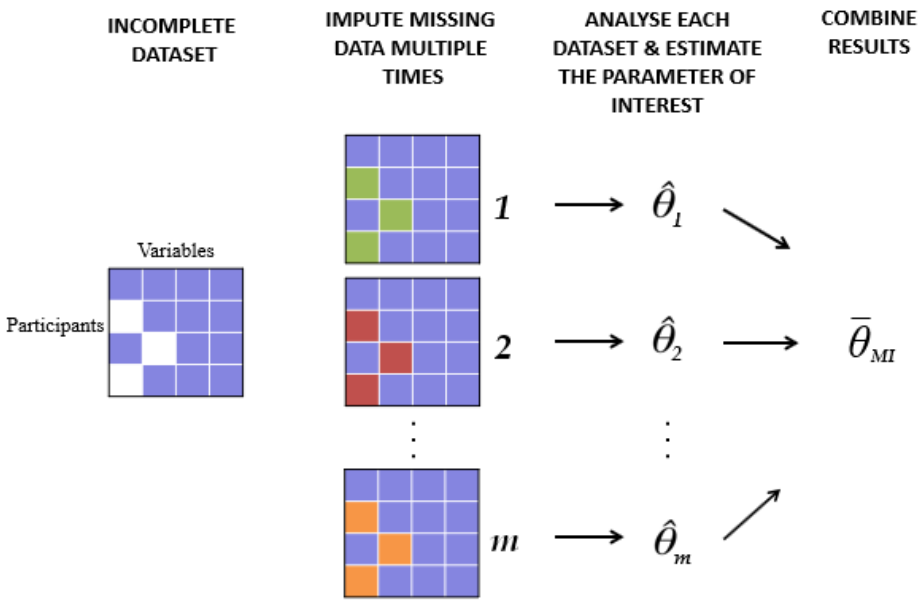
FPL_I4 - Family Poverty Ratio, Fourth Implicate

FPL_I5 - Family Poverty Ratio, Fifth Implicate

FPL_I6 - Family Poverty Ratio, Sixth Implicate

FPL_IF - Imputation Flag for FPL

Methodology for multiple imputation estimation:



The resulted values in the Poverty Level column were between 50 and 400. I binarized the values such as 1. < 200% & 2. >= 200%

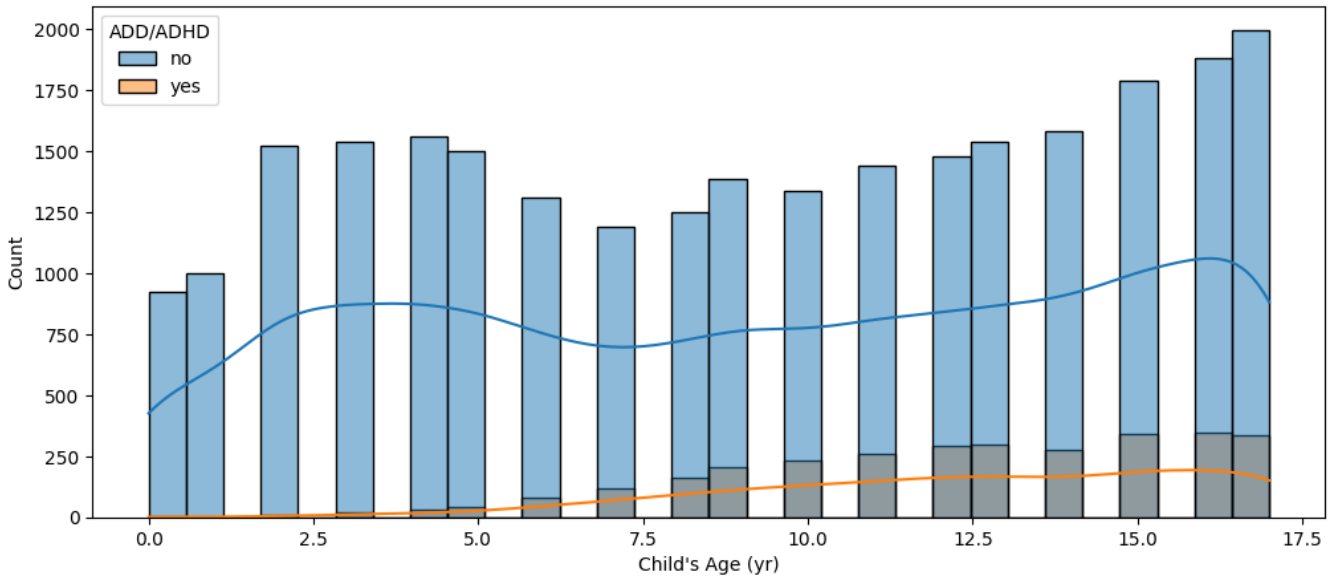
2.4 Descriptive Analysis

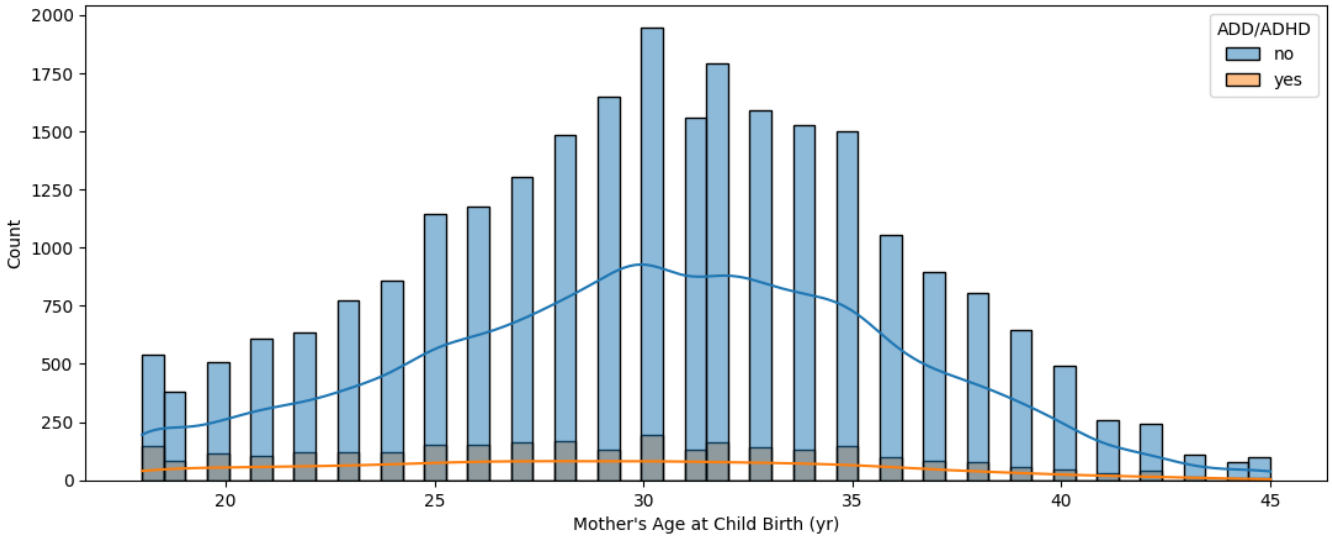
There were 2 types of feature variables in the dataset. They are 1. Numerical (discrete) 2. Categorical

2.4.1 Descriptive Analysis for numerical features

From this descriptive analysis, we can see that the age of selected children ranges from 0 to 17 years. The mothers' age at child birth ranges from 18 to 45 years. Most of the mothers were between 25 to 35 years at the time of their child's birth.

	Child's Age (yr)	Mother's Age at Child Birth (yr)
count	29433.000000	28770.000000
mean	9.524038	30.142127
std	5.163991	5.813453
min	0.000000	18.000000
25%	5.000000	26.000000
50%	10.000000	30.000000
75%	14.000000	34.000000
max	17.000000	45.000000



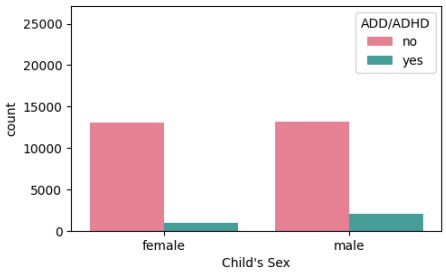


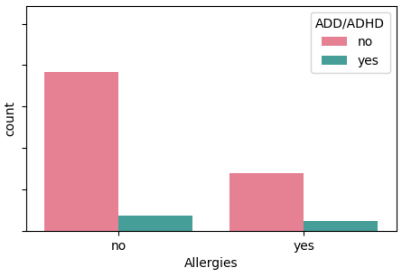
2.4.2 Descriptive Analysis for categorical features

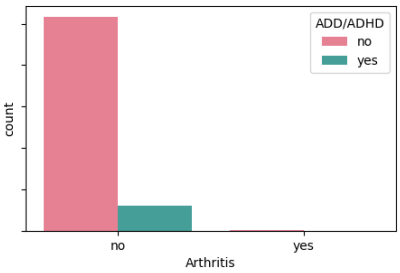
From the analysis we can see that, most of the selected children were male having no allergies, arthritis, asthma, brain injury, headaches, anxiety, and depression. Most of the children's families were covered by health insurance. Most of the selected children have no experience of living with anyone with alcohol/drug problems and most of the children are white by race. Most of the selected children had two biological/ adoptive parents and they were currently married. Education of the mothers of most of the children is more than high school. Most of the children's families' poverty level is above 200% in the scale of federal poverty level.

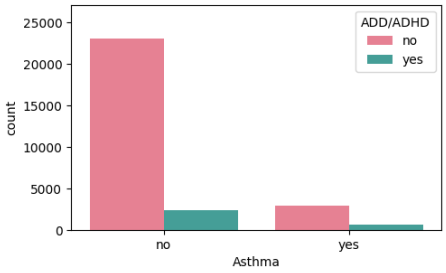
	Child's Sex	Allergies	Arthritis	Asthma	Brain Injury	Headaches	Anxiety	Depression	Health Insurance Coverage
count	29433	29357	29144	29174	29321	29358	29347	29354	29302
unique	2	2	2	2	2	2	2	2	2
top	male	no	no	no	no	no	no	no	yes
freq	15323	21140	29039	25609	27971	27987	25894	27751	28102

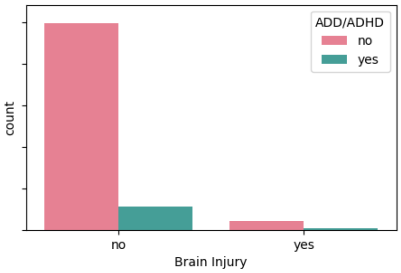
Lived with Anyone with Alcohol/Drug Problem	Child's Race	Family Structure	Very Low Birth Weight	Low Birth Weight	Premature Birth	Mother's Education	Poverty Level
28415	29433	28806	28167	28167	28839	25820	29433
2	3	8	2	2	2	3	2
no	white	Two biogical/adoptive parents, currently married	no	no	no	more than high school	>=200%
25666	23153	19115	27820	25795	25716	21250	23567

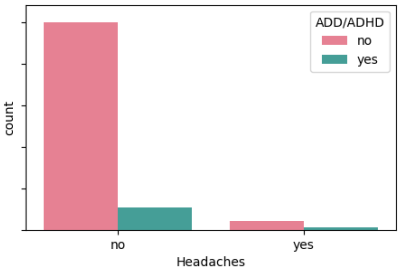

ADD/ADHD
no
yes
count
25000
20000
15000
10000
5000
0
female
male
Child's Sex

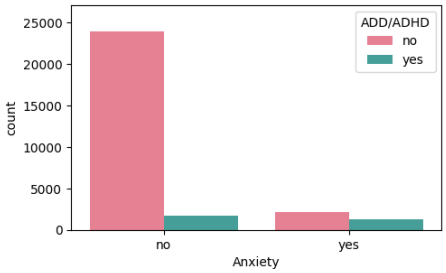

ADD/ADHD
no
yes
count
no
yes
Allergies

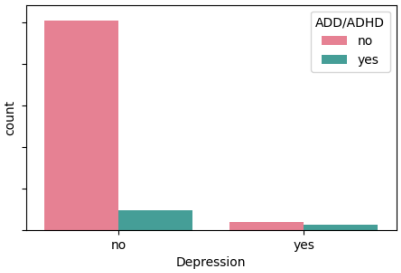

ADD/ADHD
no
yes
count
no
yes
Arthritis

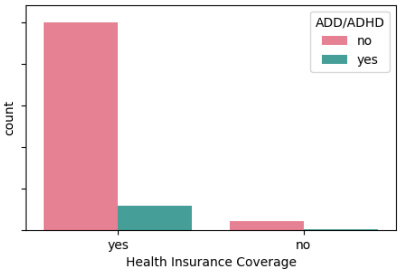

ADD/ADHD
no
yes
count
25000
20000
15000
10000
5000
0
no
yes
Asthma

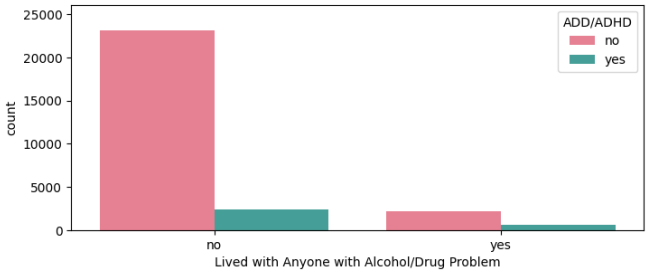

ADD/ADHD
no
yes
count
no
yes
Brain Injury

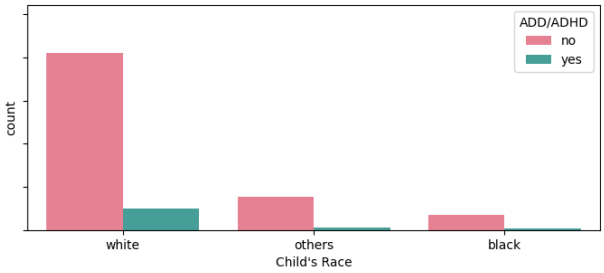

ADD/ADHD
no
yes
count
no
yes
Headaches

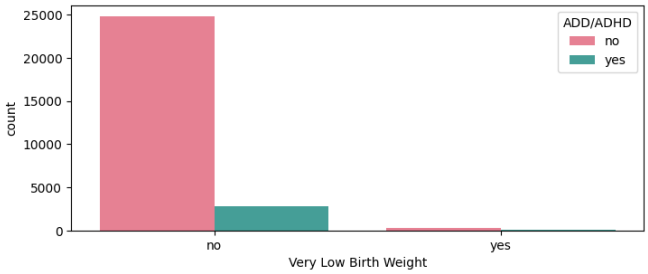

ADD/ADHD
no
yes
count
25000
20000
15000
10000
5000
0
no
yes
Anxiety

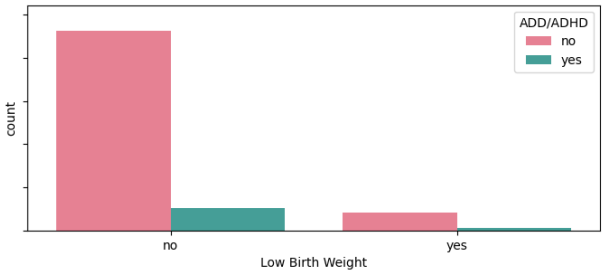

ADD/ADHD
no
yes
count
no
yes
Depression

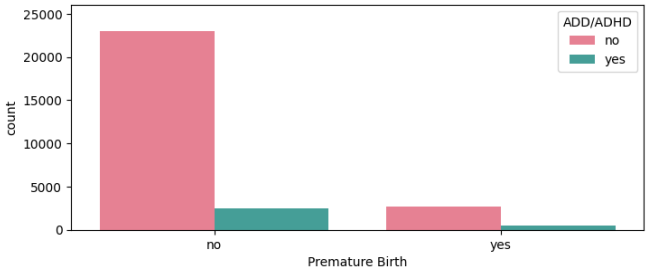

ADD/ADHD
no
yes
count
yes
no
Health Insurance Coverage

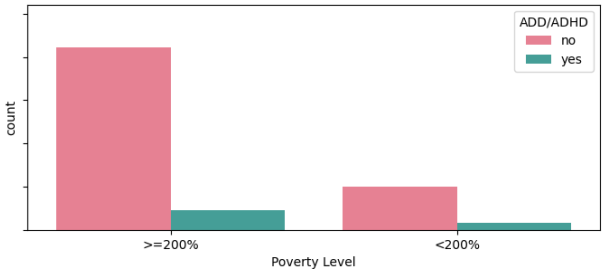

ADD/ADHD
no
yes
count
25000
20000
15000
10000
5000
0
no
yes
Lived with Anyone with Alcohol/Drug Problem

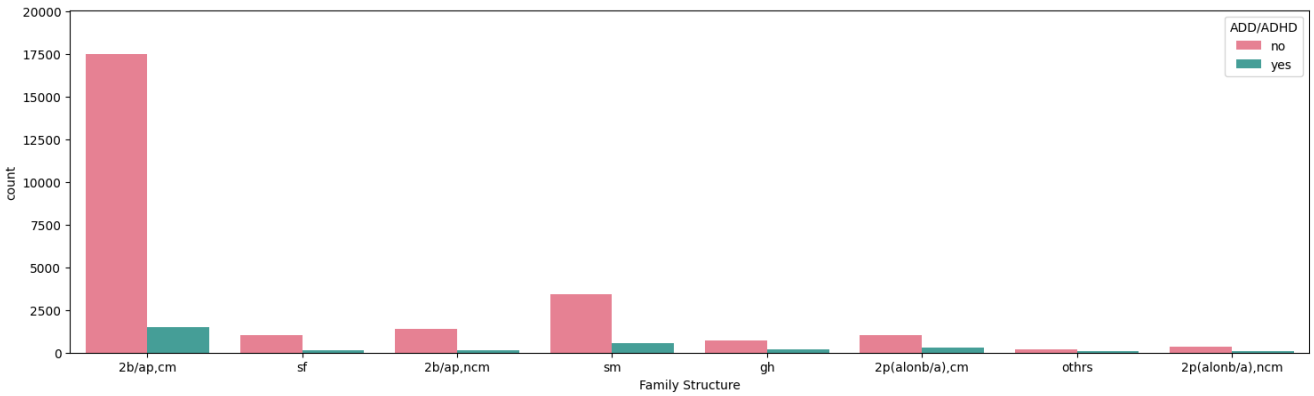

ADD/ADHD
no
yes
count
white
others
black
Child's Race

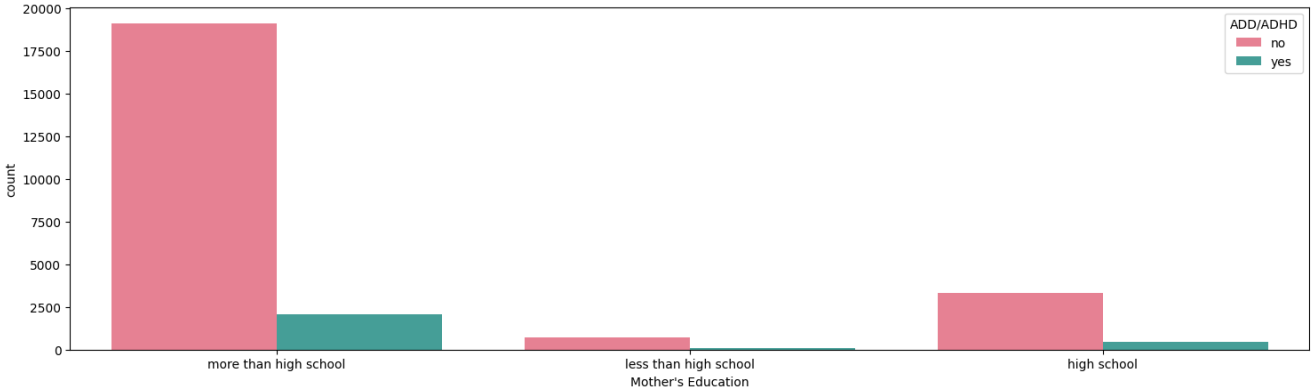

ADD/ADHD
no
yes
count
25000
20000
15000
10000
5000
0
no
yes
Very Low Birth Weight


ADD/ADHD
no
yes
count
no
yes
Low Birth Weight


ADD/ADHD
no
yes
count
25000
20000
15000
10000
5000
0
no
yes
Premature Birth


ADD/ADHD
no
yes
count
>=200%
<200%
Poverty Level





3 Methods

3.1 Feature Engineering/Evaluation

Two statistical hypothesis testing were adopted to evaluate the feature variables to see if there exists a relation between the target variable and these feature variables. For categorical features I used Independent two tailed Chi Square test and for numerical features independent two tailed t test was involved. I considered a value of 0.05 as significance value.

3.1.1 Chi Square test for Categorical Feature Variables

Formula:

The Formula for Chi Square Is

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

c = degrees of freedom

O = observed value(s)

E = expected value(s)

Results:

Only Health Insurance Coverage found to have no significant relation with target variable by Chi Square test. So, this column was dropped.

```
ADD/ADHD and  Child's Sex : significant relation exists
ADD/ADHD and  Allergies : significant relation exists
ADD/ADHD and  Arthritis : significant relation exists
ADD/ADHD and  Asthma : significant relation exists
ADD/ADHD and  Brain Injury : significant relation exists
ADD/ADHD and  Headaches : significant relation exists
ADD/ADHD and  Anxiety : significant relation exists
ADD/ADHD and  Depression : significant relation exists
***ADD/ADHD and  Health Insurance Coverage : no significant relation
ADD/ADHD and  Lived with Anyone with Alcohol/Drug Problem : significant relation exists
ADD/ADHD and  Child's Race : significant relation exists
ADD/ADHD and  Family Structure : significant relation exists
ADD/ADHD and  Mother's Education : significant relation exists
ADD/ADHD and  Very Low Birth Weight : significant relation exists
ADD/ADHD and  Low Birth Weight : significant relation exists
ADD/ADHD and  Premature Birth : significant relation exists
ADD/ADHD and  Poverty Level : significant relation exists
```

3.1.2 t-test for Numerical Feature Variables

For independent two tailed t test, two distinct columns were made from each feature variable and target variable. There are two unique values in the target variable column. These are 1. yes & 2. no

One column includes all the instances of feature variable that belongs to class “yes” and another column includes all the instances of feature variable that belongs to class “no”. the variances of these two columns were calculated. If the ratio of the larger variance and smaller variance is less than or equal to 4 then an equal variance is considered. For an equal variance situation, I used student’s t test. Otherwise, welche’s t-test was used.

Student’s t-test:

Test statistic: $(\bar{x}_1 - \bar{x}_2) \ / \ s_p(\sqrt{1/n_1 + 1/n_2})$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample means, n_1 and n_2 are the sample sizes for sample 1 and sample 2, respectively, and where s_p is calculated as:

$s_p = \sqrt{(n_1\text{-}1)s_1^2 + (n_2\text{-}1)s_2^2 \ / \ (n_1+n_2\text{-}2)}$

where s_1^2 and s_2^2 are the sample variances.

Degrees of freedom: $n_1 + n_2 - 2$

Welch’s t-test

Test statistic: $(\bar{x}_1 - \bar{x}_2) \ / \ (\sqrt{s_1^2/n_1 + s_2^2/n_2})$

Degrees of freedom: $(s_1^2/n_1 + s_2^2/n_2)^2 \ / \ \{ [(s_1^2 / n_1)^2 / (n_1 - 1)] + [(s_2^2 / n_2)^2 / (n_2 - 1)] \}$

Result:

All the feature variables found related to the target variable.

```
student t test
Ttest_indResult(statistic=32.41319214590614, pvalue=1.8313011565249034e-226)
reject null hypothesis:
significant difference between Child's Age (yr) with respect to child having ADHD or not

student t test
Ttest_indResult(statistic=-12.672223634891715, pvalue=1.0580924883357795e-36)
reject null hypothesis:
significant difference between Mother's Age at Child Birth (yr) with respect to child having ADHD or not
```

3.2 Handling Missing Values

The dataset contained a number of missing values. A list of number of missing values in each variable is given below.

```
[ ]  dataframe_v4_numeric.isnull().sum()
```

```
Child's Age (yr)                              0
Mother's Age at Child Birth (yr)            663
Child's Sex                                   0
Allergies                                    76
Arthritis                                   289
Asthma                                      259
Brain Injury                                112
Headaches                                    75
Anxiety                                      86
Depression                                   79
Lived with Anyone with Alcohol/Drug Problem  1018
Child's Race                                  0
Family Structure                            627
Very Low Birth Weight                       1266
Low Birth Weight                            1266
Premature Birth                              594
Mother's Education                          3613
Poverty Level                                 0
ADD/ADHD                                    187
dtype: int64
```

```
dataframe_v4_numeric.shape
```

```
(29433, 19)
```

```
dataframe_v4_numeric["ADD/ADHD"].value_counts()
```

```
2.0    26197
1.0     3049
Name: ADD/ADHD, dtype: int64
```

First, I dropped the instances in which all values were missing. As I didn't want to impute the missing values in the target variable, I also dropped each instance in which target column value was missing. Mother's Education is a derived feature. If a child did not live with his/her mother, then imputation in mother's education might go in wrong direction. So, I did not impute the missing values in Mother's Education column. I dropped all the instances having missing value in Mother's Education column. After dropping instances, the total number of instances came down to 25680. The total missing value count for all the variables are given below.

```
dataframe_v4_numeric.isnull().sum()
```

```
Child's Age (yr)                              0
Mother's Age at Child Birth (yr)            469
Child's Sex                                   0
Allergies                                    58
Arthritis                                   228
Asthma                                      213
Brain Injury                                 84
Headaches                                    52
Anxiety                                      65
Depression                                   59
Lived with Anyone with Alcohol/Drug Problem  365
Child's Race                                  0
Family Structure                              5
Very Low Birth Weight                       920
Low Birth Weight                            920
Premature Birth                             462
Mother's Education                            0
Poverty Level                                 0
ADD/ADHD                                      0
dtype: int64
```

```
dataframe_v4_numeric.shape
```

```
(25680, 19)
```

```
dataframe_v4_numeric["ADD/ADHD"].value_counts()
```

```
2.0    23123
1.0     2557
Name: ADD/ADHD, dtype: int64
```

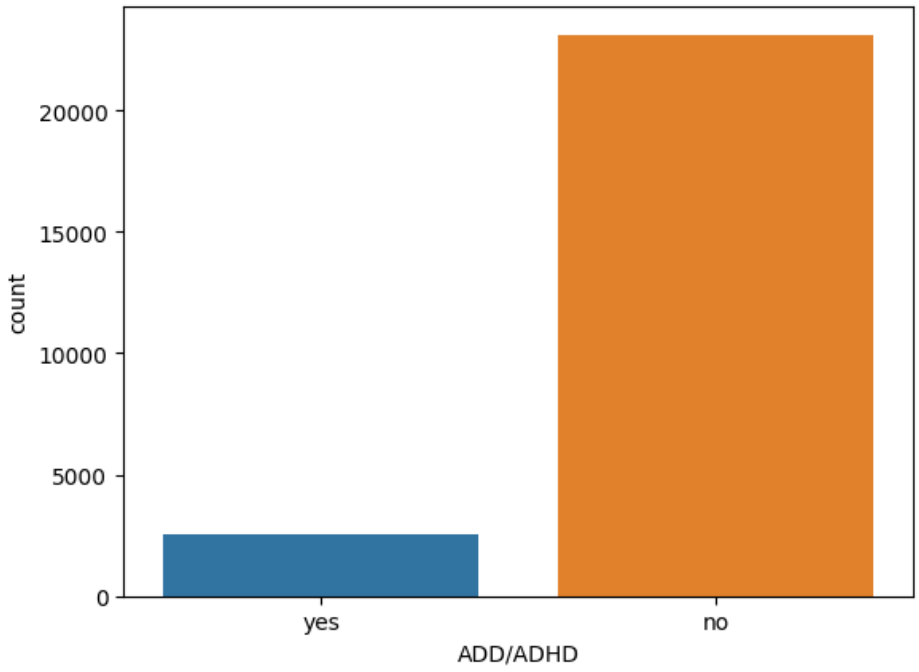
For the rest of the missing values, I adopted KNN imputation method with value of k equals 1 as it is convenient for categorical features where we don't want any fraction values.

```
dataframe_v4_imputed.isnull().sum()
```

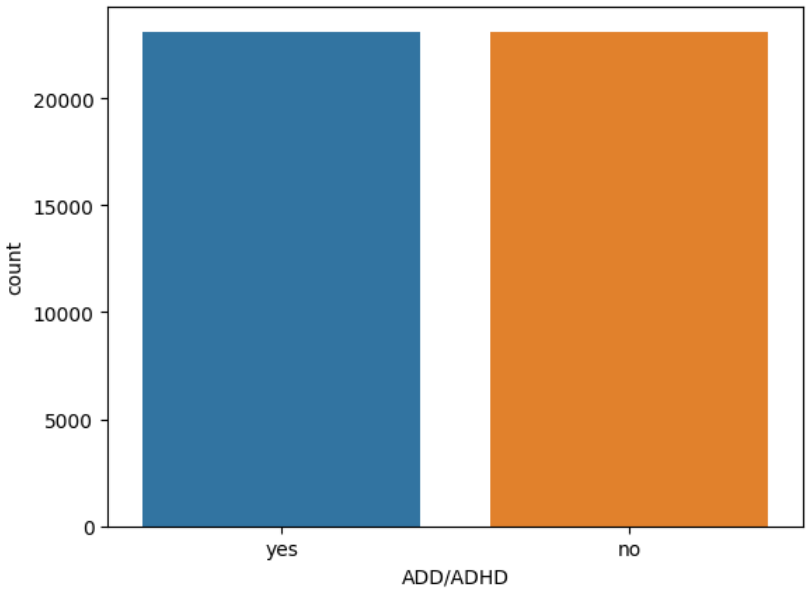
```
Child's Age (yr)                              0
Mother's Age at Child Birth (yr)              0
Child's Sex                                   0
Allergies                                     0
Arthritis                                     0
Asthma                                        0
Brain Injury                                  0
Headaches                                     0
Anxiety                                       0
Depression                                    0
Lived with Anyone with Alcohol/Drug Problem   0
Child's Race                                  0
Family Structure                              0
Very Low Birth Weight                         0
Low Birth Weight                              0
Premature Birth                               0
Mother's Education                            0
Poverty Level                                 0
ADD/ADHD                                      0
dtype: int64
```

3.3 Oversampling

Around 2557 (9.95%) of children were having ADHD, and the rest of the children were healthy.



Due to such high imbalance in dataset class label, I adopted resampling using random oversampling approach to make class label balanced.



3.4 Applying Encoder

I used different encoders on different variables. For target variable, I used label encoding. For Categorical (Ordinal) valued feature variables, I used ordinal encoding.

	Mother's Education	Poverty Level		ADD/ADHD
0	2.0	1.0	30	0
1	2.0	1.0	31	0
2	2.0	1.0	32	0
3	2.0	1.0	33	1
4	1.0	1.0	34	0
5	2.0	1.0	35	0
6	2.0	1.0	36	1

I used One Hot Encoder (OHE) for Categorical (Nominal) valued feature variables.

BI_yes	HEA_yes	ANX_yes	DEP_yes	...	FS_Two biogical/adoptive parents, currently married	FS_Two biogical/adoptive parents, not currently married	FS_Two parents (at least one not biological/adoptive), currently married	FS_Two parents (at least one not biological/adoptive), not currently married
0.0	0.0	1.0	0.0	...	1.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0

3.5 Scaling using Standardization

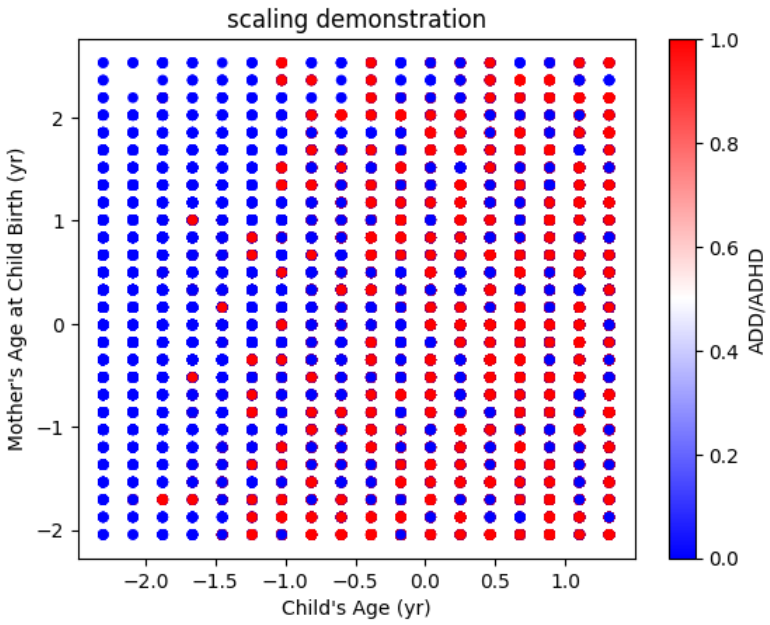
Standardization was used to scale the values of numerical valued feature variables.

Formula:

$$z = \frac{x_i - \mu}{\sigma}$$

Status of Child's Age and Mother's Age at Child Birth after scaling.

	Child's Age (yr)	Mother's Age at Child Birth (yr)
count	7.0	7.0
mean	0.0	0.0
std	1.0	1.0
min	-1.0	-1.0
25%	0.0	-1.0
50%	0.0	-0.0
75%	1.0	0.0
max	1.0	1.0



3.6 Machine Learning Techniques

Four machine learning classifiers were used. Each classifier was trained using K-Fold Cross Validation technique. The best value of k is chosen based on the performance scores after testing the model with few predefined values of k.

Predefined values of k = [2, 3, 4, 5, 6, 7, 8, 9, 10]

Four machine learning based classifiers are 1. Naïve Bayes (NB), 2. Random Forest (RF), 3. KNN, and 4. MLP. For each classifier, sklearn library was used.

3.6.1 Naïve Bayes (NB)

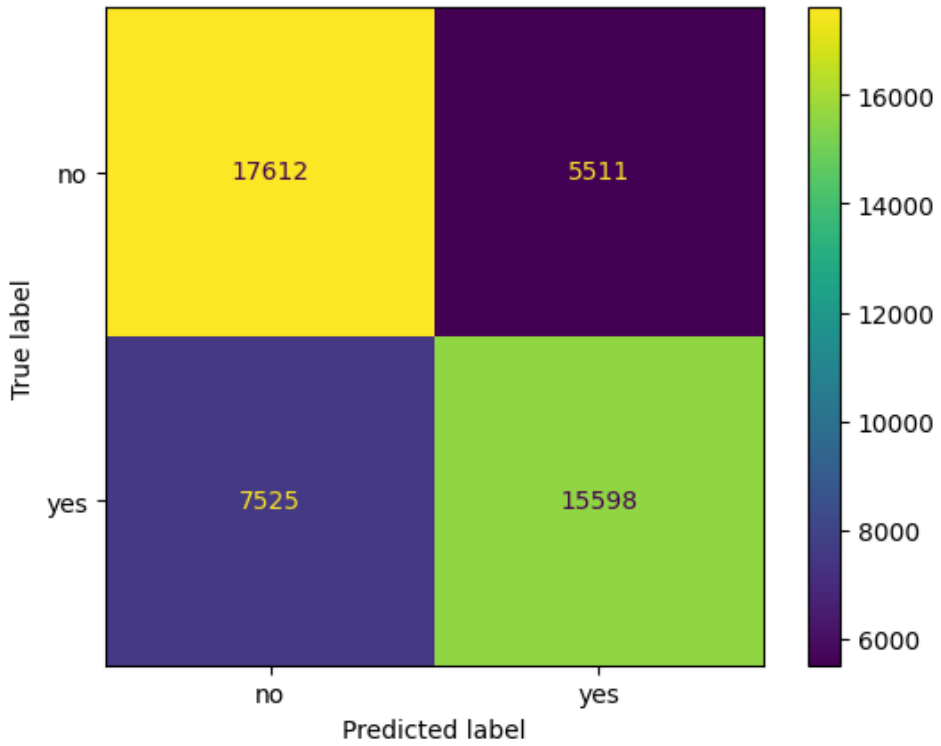
Only encoded dataset has been used to train NB classifier. Scaled dataset wasn't used for NB as negative values created by scaling causes problem for Naive Bayes. Two numerical valued feature variables namely as Child's Age and Mother's Age at Child Birth were converted to categorical valued variable using binning technique. This was done so that I could use CategoricalNB() from sklearn library.

```
# converting numerical columns to categorical columns using binning
# child's age & mother's age

dataframe_v6_encoded_X_NB = dataframe_v6_encoded_X.copy()
# 0-2 toddler 1
# 3-5 early childhood 2
# 6-11 middle childhood 3
# 12-18 early adolescence 4
dataframe_v6_encoded_X_NB["Child's Age (yr)"] = pd.cut(x=dataframe_v6_encoded_X_NB["Child's Age (yr)"], bins=[-1, 2, 5, 11, 18],
                                                       labels=[1, 2, 3, 4])


# 13-19 teen 1
# 20-30 adult 2
# 31-50 middle aged 3
dataframe_v6_encoded_X_NB["Mother's Age at Child Birth (yr)"] = pd.cut(x=dataframe_v6_encoded_X_NB["Mother's Age at Child Birth (yr)"],
                                                                       bins=[12, 19, 30, 50],
                                                                       labels=[1, 2, 3])
```

Performance:

For k = [2, 3, 4, 5, 6, 7, 8, 9, 10] we have accuracies [0.718, 0.7178, 0.7181, 0.718, 0.7178, 0.7179, 0.7179, 0.7177, 0.7179]. Best value of K is 4, for k = 4 we get

```
accuracy:  0.718116
error rate:  0.281884
sensitivity:  0.674566
specificity:  0.761666
```



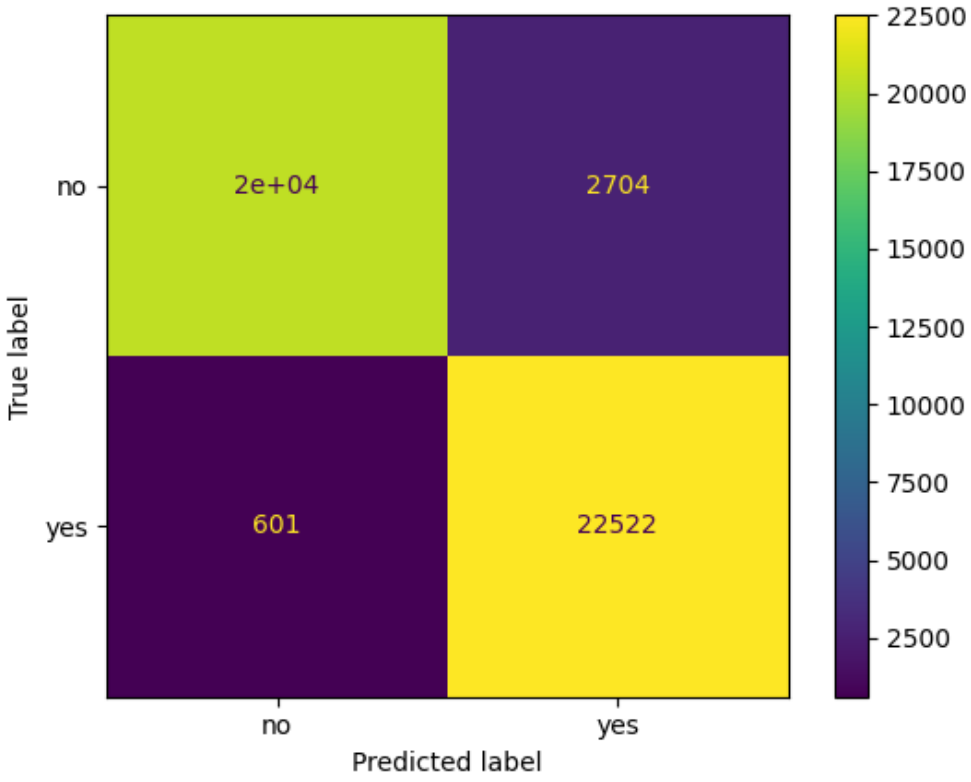
3.6.2 Random Forest (RF)

Only encoded dataset was used. Random Forest doesn't require scaling. Number of trees in the forest is 100.

Performance:

For k = [2, 3, 4, 5, 6, 7, 8, 9, 10] we have accuracies [0.9036, 0.9178, 0.9217, 0.9245, 0.9257, 0.9264, 0.9274, 0.927, 0.9286]. Best value of K is 10, for k = 10 we get

```
accuracy:  0.928534
error rate:  0.071466
sensitivity:  0.974009
specificity:  0.88306
```



3.6.3 KNN

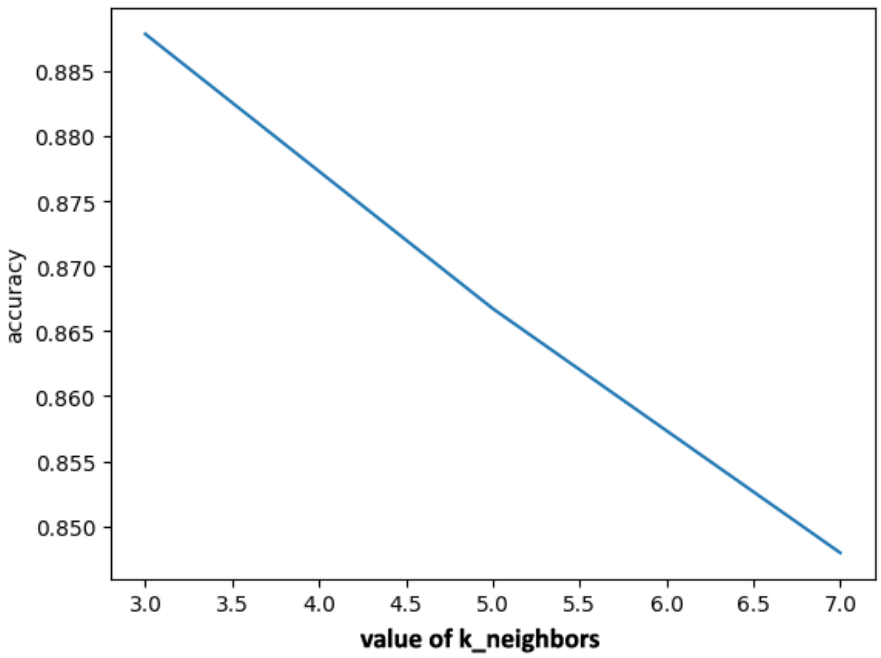
KNN performs better on scaled dataset. Best value for number of nearest neighbors is chosen based on the performance scores after testing the model with few predefined values of number of nearest neighbors.

predefined values of number of nearest neighbors, k_neighbors = [3, 5, 7]

Performance:

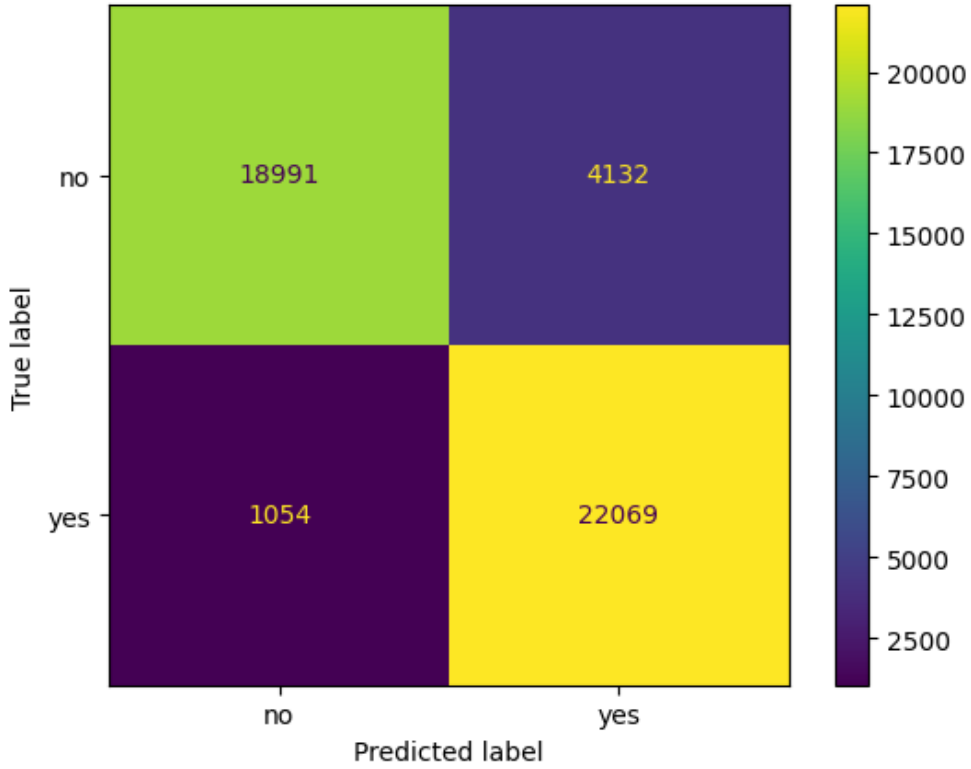
For k = [2, 3, 4, 5, 6, 7, 8, 9, 10], the best value of K is 10 for all the values of k_neighbors. For k_neighbors = [3, 5, 7] we have accuracies [0.887861, 0.866734, 0.847944]

It's observed that with the increase in k_neighbors values, the accuracies decrease.



So, the best value for k_neighbors = 3. For this value of k_neighbors we get,

```
accuracy:   0.887861
error rate:   0.112139
sensitivity:   0.954418
specificity:   0.821303
```



3.6.4 Multi-Layer Perceptron (MLP)

MLP classifier requires scaled dataset. For the MLP classifier, I have assigned the following values to the sklearn MLP classifier parameters.

hidden_layer_sizes = (100, i) (default) (ith element represents the number of neurons in the ith hidden layer)

random state = 1 (determines random number generation for weights and bias initialization)

max_iter = 500 (maximum number of iterations. solver iterates until convergence)

solver = adam (adam refers to a stochastic gradient-based optimizer proposed by Kingma, Diederik, and Jimmy Ba)

activation = relu

learning rate = 0.01

early stopping = True (terminate training when validation score is not improving. 10% of training data set as validation)

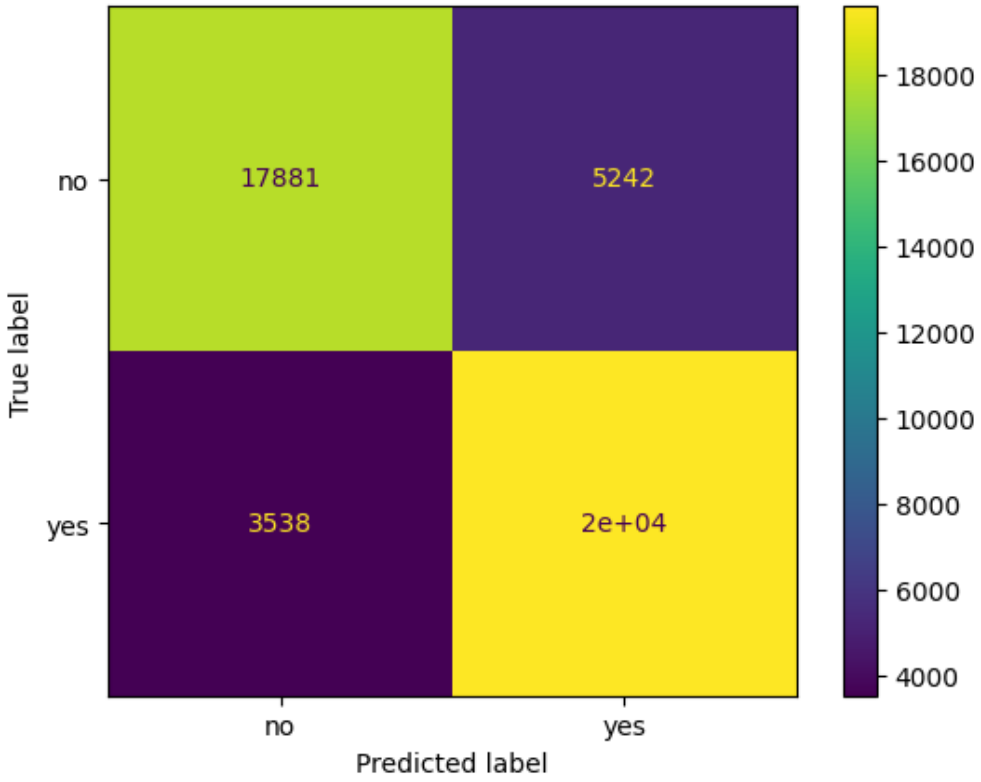
Performance:

For k = [2, 3, 4, 5, 6, 7, 8, 9, 10] we have accuracies

[0.7907, 0.805, 0.8085, 0.8096, 0.8051, 0.8101, 0.8085, 0.8092, 0.8074]

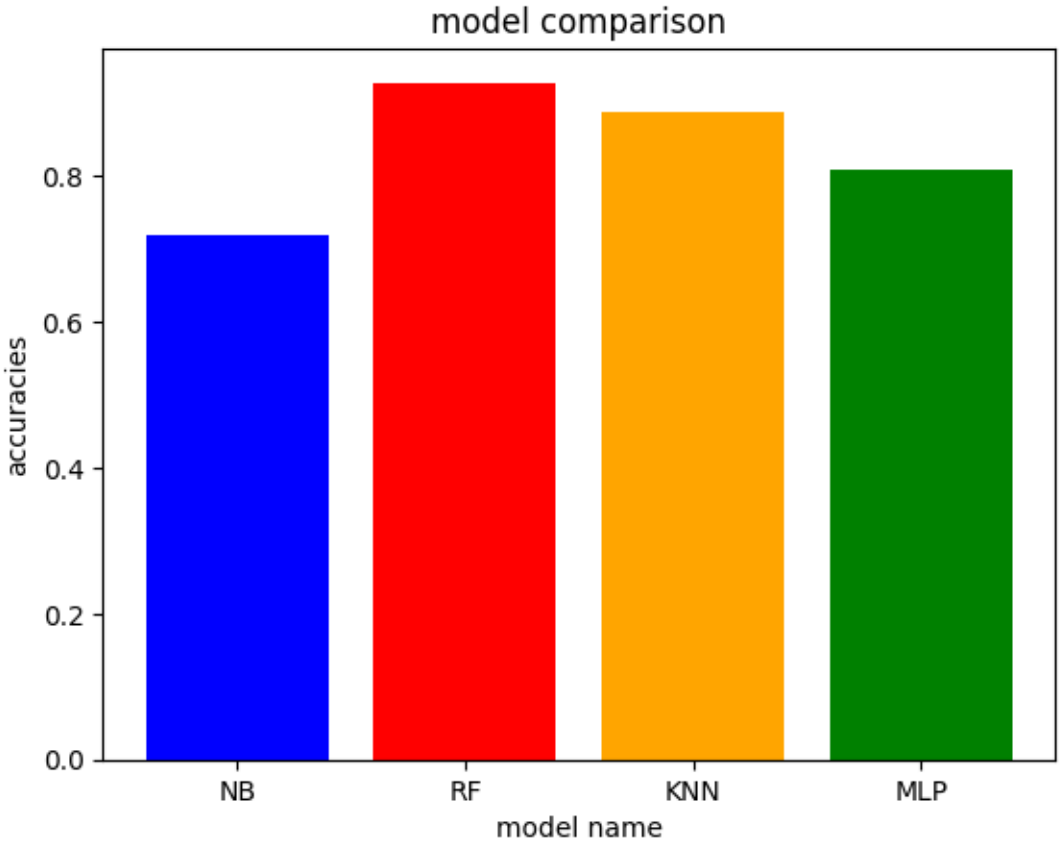
Best value of K is 4, for k = 7 we get

```
accuracy:   0.810146
error rate:   0.189854
sensitivity:   0.846992
specificity:   0.773299
```



4 Results

The comparison of accuracy scores of four aforementioned classifiers are given below.



From this we can conclude that the best machine learning classifier for early prediction of ADHD is Random Forest. At the second place we have KNN classifiers. I have performed a t-test between the two models to see if there exists any statistically significant difference between Random Forest and KNN. T-test was performed with a significance level of 0.05.

The test result showed that there is a significant difference between Random Forest and KNN. It appeared that Random Forest is better than KNN.

```
Null Hypothesis is False: Random Forest model is better
```

5 Conclusion

This project illustrated that RF-based classifier could provide excellent accuracy in correctly classifying and predicting children with ADHD. This study will assist physicians in detecting and treating children with ADHD at an early stage.