CISC 520-50- B-2023/Summer - Data Engineering and Mining

Deliverable 3: Final Project Report

Saurabh Shirish Prabhu

SPrabhu1@my.harrisburgu.edu

Introduction and Background

Dataset used:

1. National Longitudinal Study of Adolescent to Adult Health (Add Health) Wave I, 1994-1995 and
https://dataverse.unc.edu/dataset.xhtml?persistentId=doi:10.15139/S3/11900

2. National Longitudinal Study of Adolescent to Adult Health (Add Health) Wave IV, 2008
https://dataverse.unc.edu/dataset.xhtml?persistentId=doi:10.15139/S3/11920

The National Longitudinal Study of Adolescent to Adult Health (Add Health) is a longitudinal study of a nationally representative sample of adolescents in grades 7-12 in the United States during the 1994-95 school year. Add Health is a school-based longitudinal study of a nationally-representative sample of adolescents in grades 7-12 in the United States in **1994-95**. Data have been collected from adolescents, their fellow students, school administrators, parents, siblings, friends, and romantic partners through multiple data collection components, including four respondent in-home interviews. In addition, existing data bases with information about respondents' neighborhoods and communities have been merged with Add Health data, including variables on income and poverty, unemployment, availability and utilization of health services, crime, church membership, and social programs and policies.

The Add Health cohort has been followed into young adulthood with four in-home interviews, **the most recent in 2008**, when the sample was aged 24-32*. Add Health combines longitudinal survey data on respondents' social, economic, psychological and physical well-being with contextual data on the family, neighborhood, community, school, friendships, peer groups, and romantic relationships, providing unique opportunities to study how social environments and behaviours in adolescence are linked to health and achievement outcomes in young adulthood. The fourth wave of interviews expanded the collection of biological data in Add Health to understand the social, behavioural, and biological linkages in health trajectories as the

This study spans over 14 years from 1994 until most recent year 2008

Wave I The public use dataset for Wave I contains information collected in 1994-95 from Add Health's nationally representative sample of adolescents. This dataset includes Wave I respondents and consists of one-half of the core sample, chosen at random, and one-half of the oversample of African-American adolescents with a parent who has a college degree. The total number of Wave I respondents in this dataset is approximately 6,500.

Wave IV was designed to study the developmental and health trajectories across the life course of adolescence into young adulthood. Taking place in 2008, approximately 92.5% of the original Wave I respondents were located and 80.3% of eligible cases were interviewed. The Wave IV public use file contains data on 5,114 respondents, aged 24 to 32*. In Wave IV,

Description of data quality:

The dataset combines longitudinal survey data on respondents' social, economic, psychological and physical well-being with contextual data on the family, neighborhood, community, school, friendships, peer groups, and romantic relationships. The dataset has been followed into young adulthood with four in-home interviews, the most recent in 2008.

The quality of the data is high, and it is considered to be one of the most comprehensive datasets on adolescent health and development. The dataset is also publicly available for researchers to use. However, there are some known sources of errors or biases. For example, the sample is not representative of all adolescents in the United States because it excludes those who dropped out of school before grade 7 or who were not enrolled in school during the 1994-95 school year. Additionally, there may be some measurement error due to self-reported data. Add Health oversampled schools with larger proportions of black and Hispanic students.

The dataset is maintained at Odum Institute Data Archive
The Odum Institute Data Archive is a research data stewardship organization, preserving and broadening research data assets for scientific inquiry and reproducibility. I

Hypothesis

How are various features related to educational resources, health resources and various opportunities influencing participants highest education level?

Mining Methods and Analysis Proposal

In this study following methods were proposed

Outlier Detection: This technique is used to identify unusual patterns or observations in a dataset.
Genetic Algorithm: This technique is used to optimize complex problems by simulating the process of natural selection.
Regression analysis is proposed to understand various patterns in adolescent to adult health patterns. Logistic regression is proposed to understand relationships between binary target variable and various features.
Classification is proposed to understand Y/N type participants education level patterns
Prediction: is proposed to check if the classification and regression models can predict participants education level patterns.
Feature Selection: Feature selection techniques help in choosing the most relevant and informative features for building models, reducing complexity and improving model performance.
Ensemble Methods: Ensemble methods combine multiple models to improve prediction accuracy and reduce overfitting. Bagging (e.g., Random Forest) and Boosting (e.g., Gradient Boosting Machines) are common ensemble techniques.

Exploratory Data Analysis

Wave I 1994 dataset contains 6504 rows × 2794 columns and Wave IV 2008 dataset contains 5114 rows × 920 columns. Two datasets were merged together to establish relation between past and future of participating populations. Merged dataset contains 6504 rows x 3713 columns.

One basic question was investigated to understand influence of TV watching hours on grades of children.

Does watching TV affect participants grades in mathematics subject?

Grades variable was categorical with 9 levels.

Distribution of Grades variable

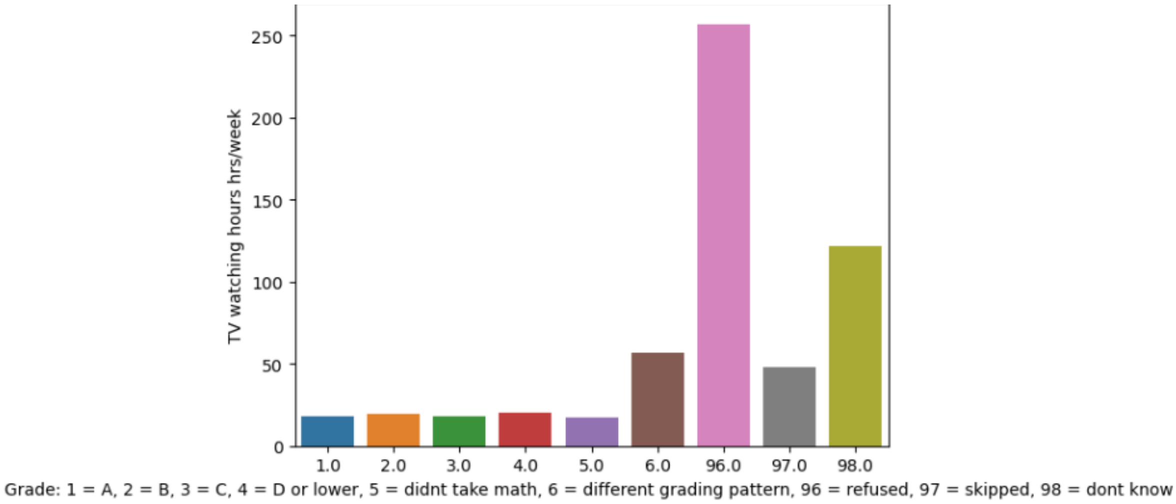


Fig 1: H1ED12 – grades

Grades variable H1ED12 was converted to categorical with 2 levels (A and B or less) and TV watching hours variable H1DA8 was plotted as below:

Grades vs TV watching hours

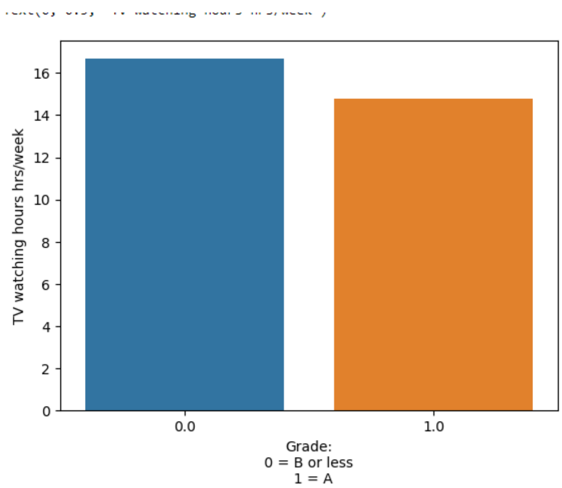


Fig 2: H1DA8 – number of hrs spent on watching tv vs H1ED12 – grades

Plotly library was used to test effect of males and females on grades and TV watching hours.

Grades vs TV watching hours vs gender

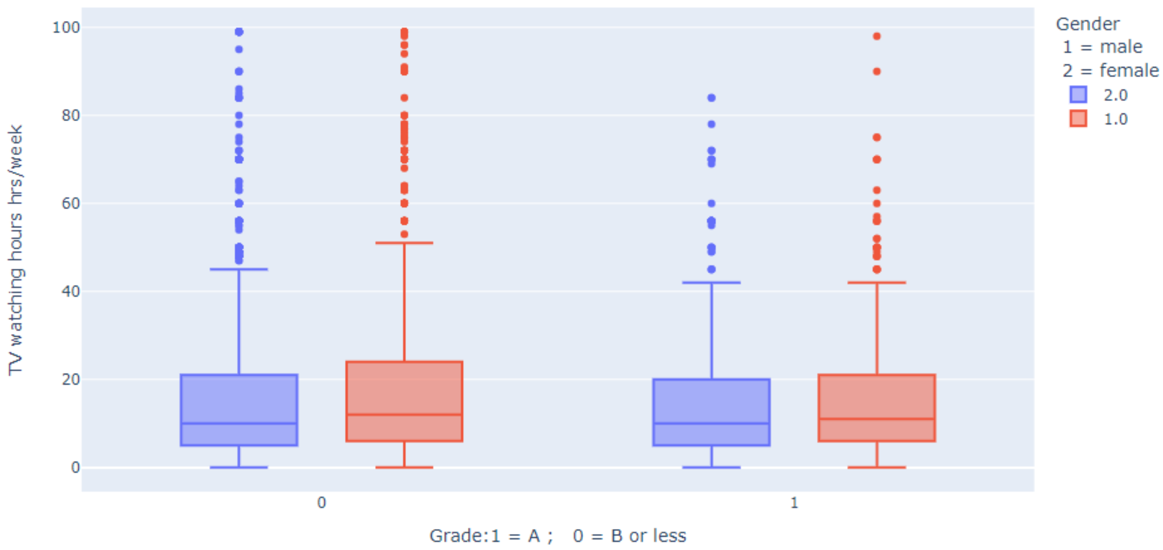


Fig 3: H1DA8 – number of hrs spent on watching tv vs H1ED12 – grades vs BIO_SEX Gender

New feature BMI:

A feature named BMI was created using height and weight of participants.

	AID	H1GH60	HEIGHT94	BMI94
0	b'57100270'	99.790321	157.48	40.234131
1	b'57101310'	68.946040	182.88	20.612654
3	b'57103869'	102.058283	205.74	24.108368
4	b'57104553'	90.718474	170.18	31.321007
5	b'57104649'	54.431084	162.56	20.595703
...	...	...	...	...
6499	b'99719930'	50.348753	165.10	18.469349
6500	b'99719939'	65.770894	175.26	21.410418
6501	b'99719970'	63.502932	165.10	23.294675
6502	b'99719976'	73.481964	165.10	26.955266
6503	b'99719978'	61.234970	165.10	22.462722

6291 rows × 4 columns

Fig 4: Table showing H1GH60 Weight and HEIGHT94 Height used to calculate BMI94- BMI in 1994

	AID	H4WGT	H4HGT	BMI08
1	b'57101310'	113.9	180.0	35.154321
3	b'57103869'	107.8	202.0	26.418979
7	b'57109625'	68.0	161.0	26.233556
9	b'57111071'	89.4	177.0	28.535861
11	b'57113943'	150.6	185.5	43.766029
...	...	...	...	...
6499	b'99719930'	58.6	167.5	20.886612
6500	b'99719939'	97.0	174.5	31.855239
6501	b'99719970'	80.6	178.0	25.438707
6502	b'99719976'	75.3	165.0	27.658402
6503	b'99719978'	78.2	180.0	24.135802

5042 rows × 4 columns

Fig 5: Table showing H4WGT Weight and H4HGT Height used to calculate BMI08- BMI in 2008

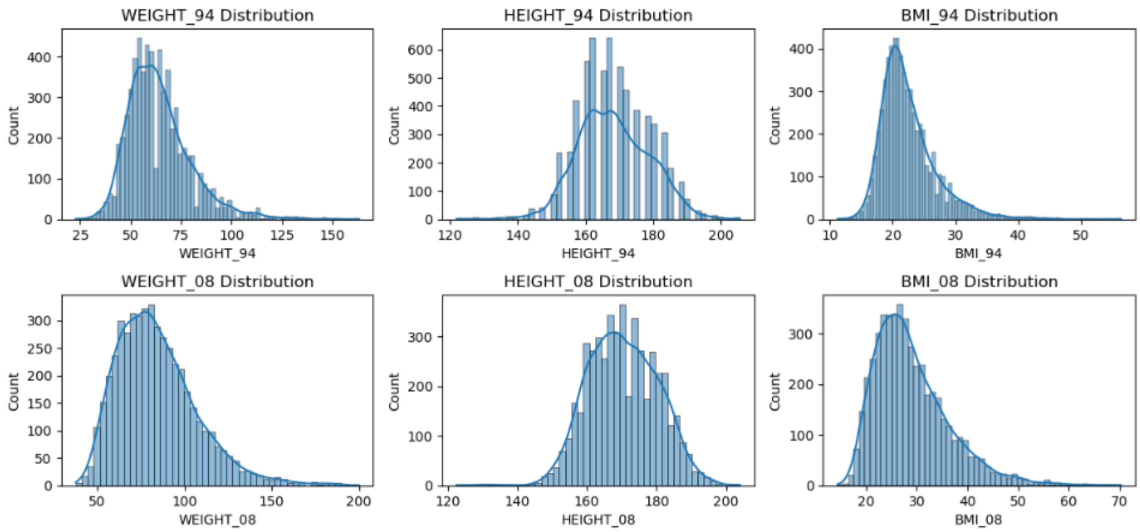


Fig 6: Histograms showing Weight, Height and BMI in 2008 and 1994

Gender was used to check how the distribution differs for males and females.

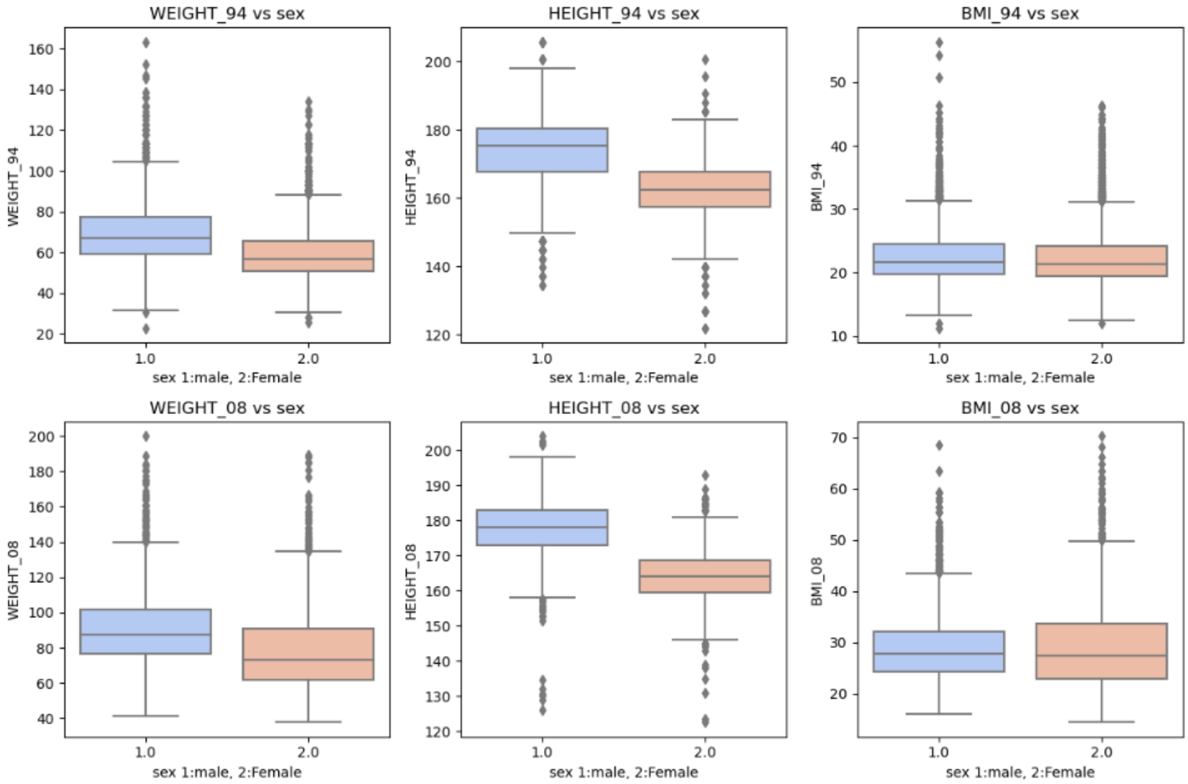


Fig 7: Boxplot showing Weight, Height and BMI in 2008 and 1994 vs Gender

Difference between BMI 2008 and BMI 1994 was calculated to check how the participants health is trending.

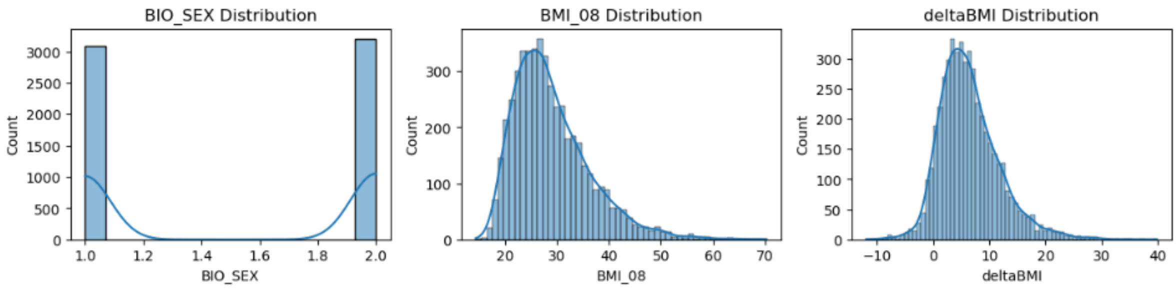


Fig 7: histogram showing BMI in 2008 , difference between BMI_08 and BMI_94

Mining method and evaluation

Feature selection and Feature Engineering

Feature	Description	Type
AID	Participant unique identifier	Object

Feature	Description	Type
BIO_SEX	**Gender** 1 R is male, 2 R is female, 6 refused	Categorical
H4OD1Y	**Respondent's date of birth** – year 1974 - 1983	Numerical
H1DA6	**During the past week, how many times did you do exercise:** 0 : not at all, 1 : 1 or 2 times, 2 : 3 or 4 times, 3 : 5 or more times, 6 : refused,8 : don't know	Categorical
H4DA1	**how many hours did you watch television?** 1-150 hours, 996: refused, 998: dont know	Numerical
H4SP1H	**what time do you usually wake up?** 1-12 hours, 96: refused, 98: dont know	Numerical
H4PE7	**I'm always optimistic about my future-** 1: strongly agree ,2: agree ,3: neither agree nor disagree ,4: disagree ,5: strongly disagree ,6: refused ,8: don't know , .: missing	Categorical
H4GH1	**how is your health?**-1: excellent ,2: very good ,3: good ,4: fair ,5: poor	Categorical
H1ED14	**Grade in science?** - 1: A ,2: B ,3: C ,4: D or lower ,5: didn't take this subject ,6: took the subject, but it wasn't graded this way ,96: refused ,97: legitimate skip ,98: don't know	Categorical
H1ED13	**Grade in history or social studies?**:A ,2: B ,3: C ,4: D or lower ,5: didn't take this subject ,6: took the subject, but it wasn't graded this way ,96: refused ,97: legitimate skip ,98: don't know	Categorical
H1ED12	**Grade in mathematics?** A ,2: B ,3: C ,4: D or lower ,5: didn't take this subject ,6: took the subject, but it wasn't graded this way ,96: refused ,97: legitimate skip ,98: don't know	Categorical
H1GH51	**How many hours of sleep do you usually get?** 1-20 hours ,96: refused ,98: don't know	Numerical
H1DA8	**How many hours a week do you watch television?** 0 hrs, 1-99 hrs, 996 refused, 998 don't know	Numerical
H1GH1	**how is your health?** 1: excellent ,2: very good ,3: good ,4: fair ,5: poor, 6:refused, 8: don't know	Categorical
H1DA10	**How many hours a week do you play video or computer games?** 0: don't play, 1 - 99 hrs, 996: refused, 998 don't know	Numerical
BMI_08	**Bmi in 2008 calculated in above steps**	Numerical
deltaBMI	**Change Bmi from 1994 to 2008 as calculated in above steps**	Numerical
H4ED2	(***TARGET***) **highest level of education:** 1: 8th grade or less ,2: some high school ,3: high school graduate ,4: some vocational/technical training (after high school) ,5: completed vocational/technical training (after high school) ,6: some college ,7: completed college (bachelor's degree) ,8: some graduate school ,9: completed a	Categorical

Feature	Description	Type
	master's degree ,10: some graduate training beyond a master's degree ,11: completed a doctoral degree ,12: some post baccalaureate professional education (e.g., law school) ,13: completed post baccalaureate professional education (e.g., law school, med school, nurse) ,98: don't know	

Fig 8: Table showing features under investigation

Overall distribution of above features was plotted below

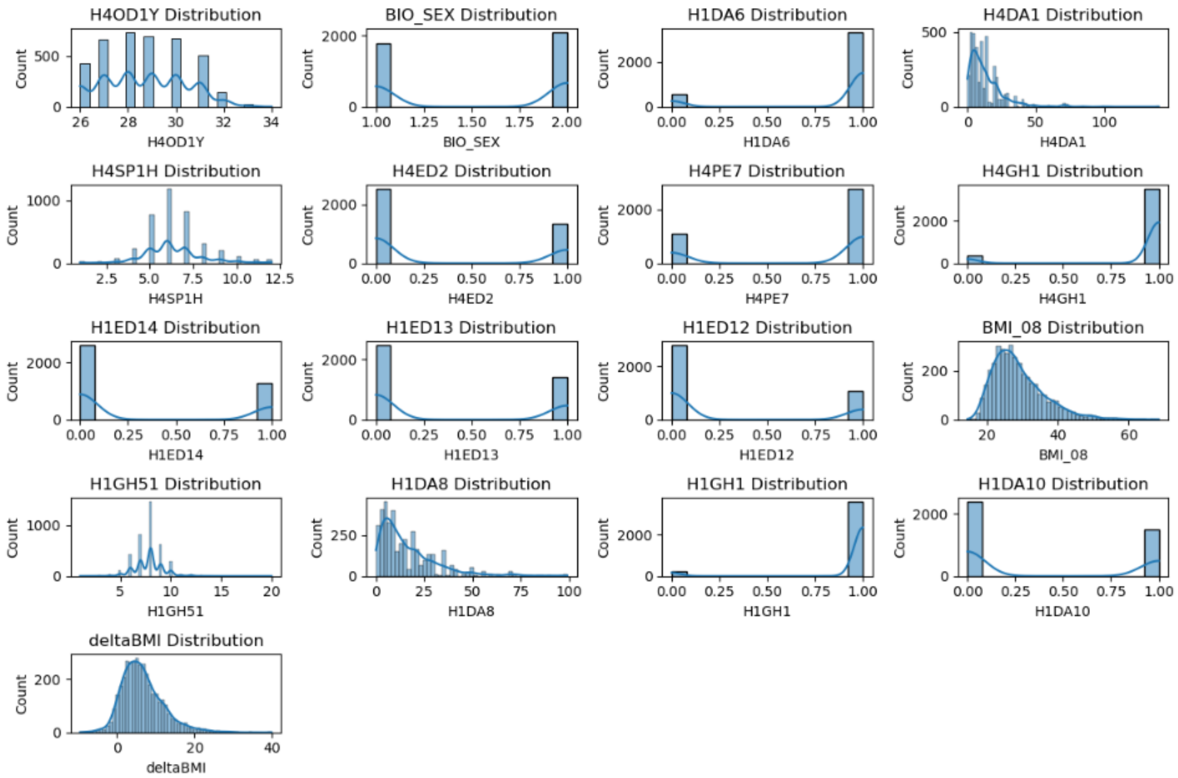


Fig 8: distribution of features under investigation

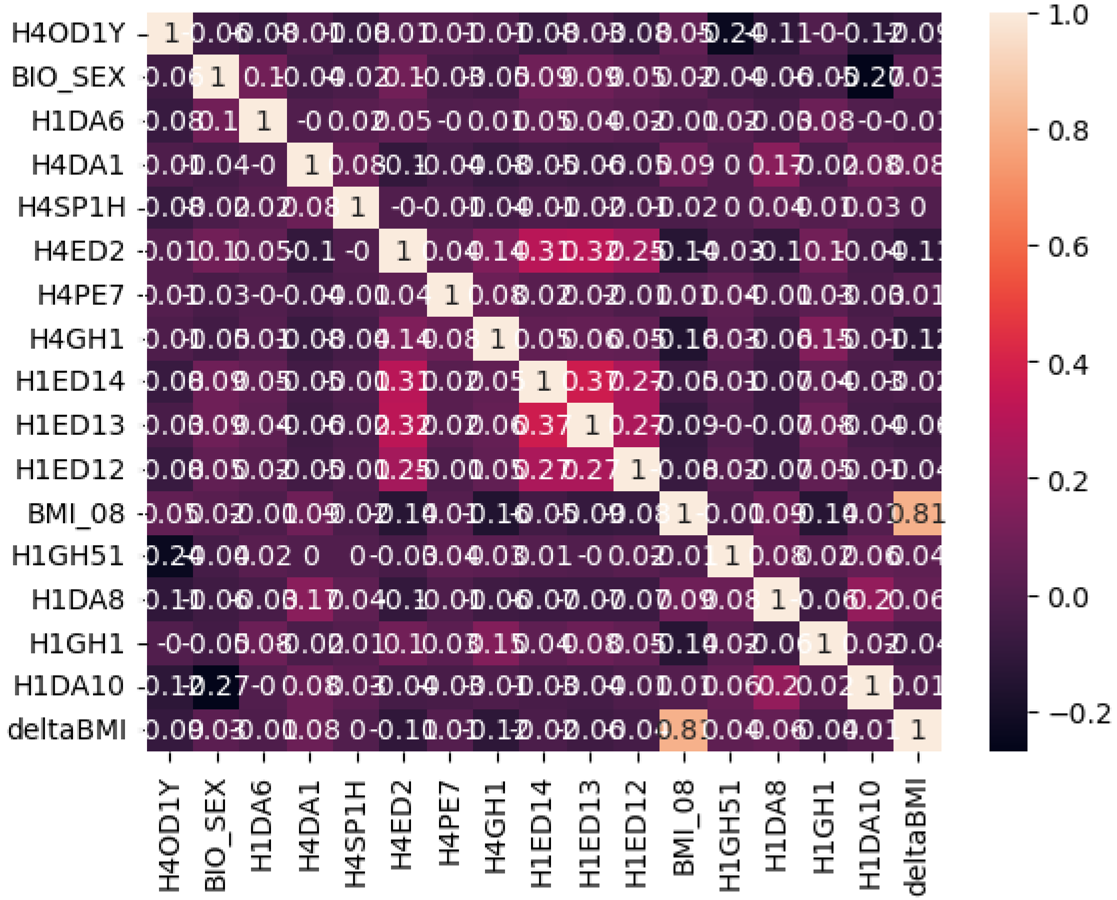


Fig 9: correlation plot of features under investigation (unnormalized)

Dealing with outliers

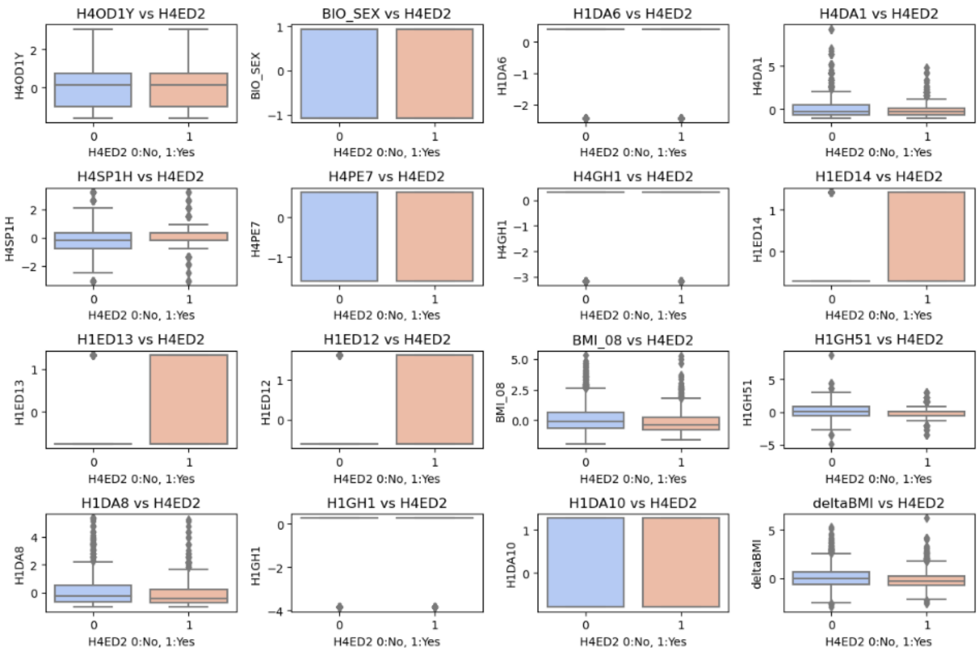


Fig 10: plots of features under investigation showing outliers (unnormalized) vs hypothesis question target

Outliers were removed using interquartile range method

In [45]:
```
## remove outliers from continuous features

Q1 = data[['BMI_08', 'H1GH51', 'H1DA8', 'deltaBMI']].quantile(0.25)
Q3 = data[['BMI_08', 'H1GH51', 'H1DA8', 'deltaBMI']].quantile(0.75)
IQR = Q3 - Q1

# Create masks for each column separately
mask_bmi = (data['BMI_08'] < (Q1['BMI_08'] - 1.5 * IQR['BMI_08'])) | (data['BMI_08'] > (Q3['BMI_08'] + 1.5 * IQR['BMI_08']))
mask_h1gh51 = (data['H1GH51'] < (Q1['H1GH51'] - 1.5 * IQR['H1GH51'])) | (data['H1GH51'] > (Q3['H1GH51'] + 1.5 * IQR['H1GH51']
mask_h1da8 = (data['H1DA8'] < (Q1['H1DA8'] - 1.5 * IQR['H1DA8'])) | (data['H1DA8'] > (Q3['H1DA8'] + 1.5 * IQR['H1DA8']))
mask_deltabmi = (data['deltaBMI'] < (Q1['deltaBMI'] - 1.5 * IQR['deltaBMI'])) | (data['deltaBMI'] > (Q3['deltaBMI'] + 1.5 * I

# Combine masks using logical OR (|) operator
mask_combined = mask_bmi | mask_h1gh51 | mask_h1da8 | mask_deltabmi

# # Replace outliers with NaN
# data[['BMI_08', 'H1GH51', 'H1DA8', 'deltaBMI']][mask_combined] = np.nan

# Replace outliers with NaN using .loc
data.loc[mask_combined, ['BMI_08', 'H1GH51', 'H1DA8', 'deltaBMI']] = np.nan
```

```
In [47]:  data.isna().sum()

Out[47]: H4OD1Y        0
         BIO_SEX       0
         H1DA6         0
         H4DA1         0
         H4SP1H        0
         H4PE7         0
         H4GH1         0
         H1ED14        0
         H1ED13        0
         H1ED12        0
         BMI_08      365
         H1GH51      365
         H1DA8       365
         H1GH1         0
         H1DA10        0
         deltaBMI    365
         H4ED2         0
         dtype: int64
```

```
-------Total Outlier Count - Pre Masking-------

 365


-------Total Outlier Count - Post Masking-------

 H4OD1Y      0
BIO_SEX     0
H1DA6       0
H4DA1       0
H4SP1H      0
H4PE7       0
H4GH1       0
H1ED14      0
H1ED13      0
H1ED12      0
BMI_08      0
H1GH51      0
H1DA8       0
H1GH1       0
H1DA10      0
deltaBMI    0
H4ED2       0
dtype: int64
```

Fig 11: plots of features under investigation showing outliers and statistics.

Normalization:

Normalization is scaling data to a common range (usually 0 to 1) to maintain relative proportions between features.

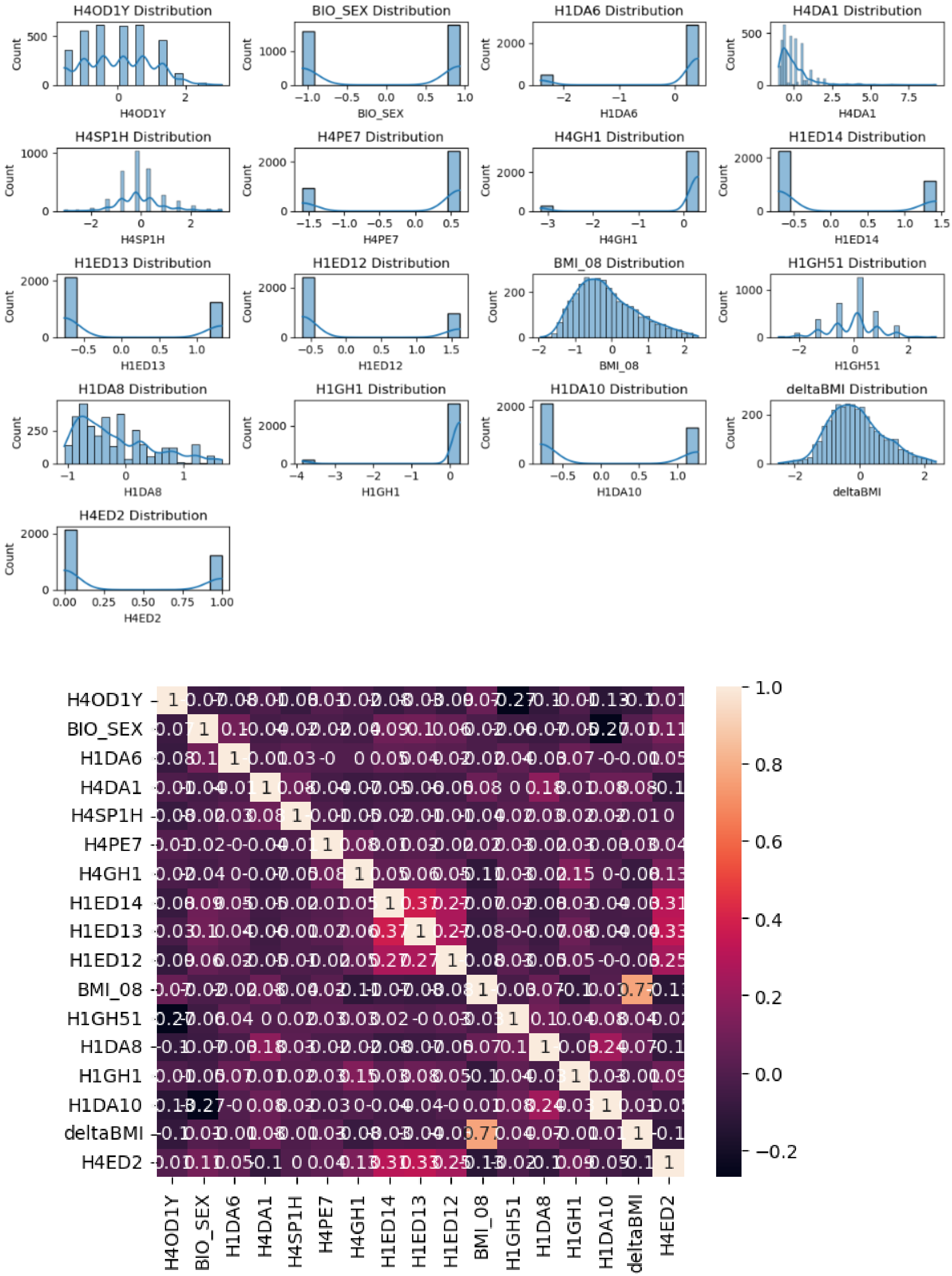


Fig 12: plots of features under investigation (normalized).

Top features were as follows

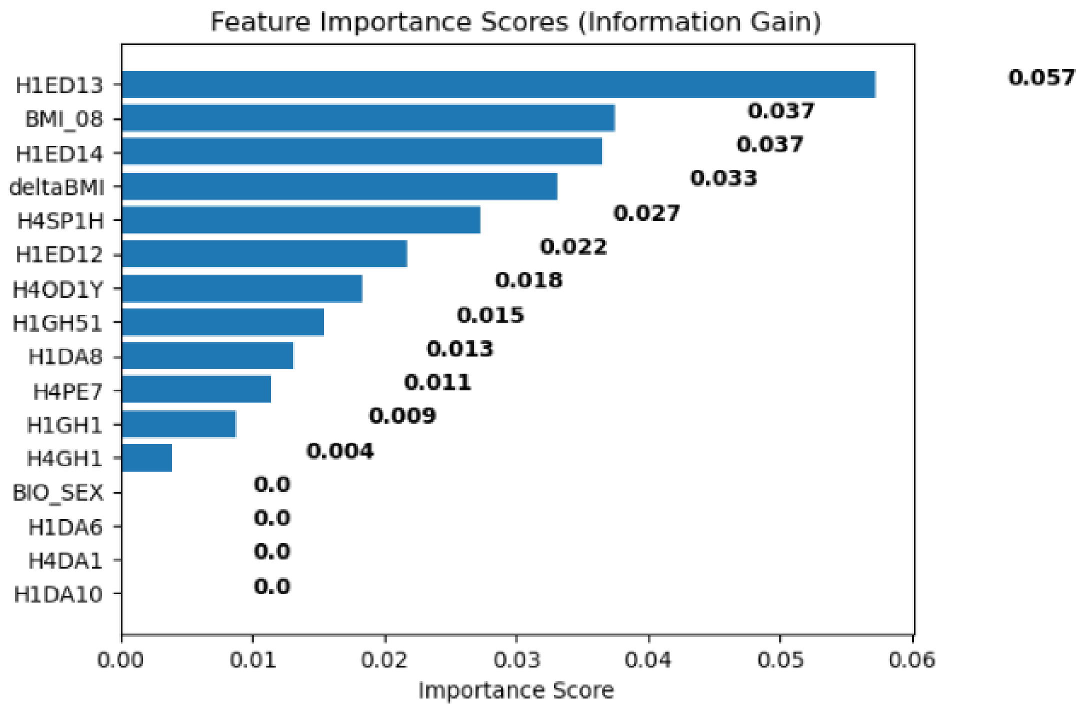


Fig 13: top features using classification based algorithm **mutual_info_classif**

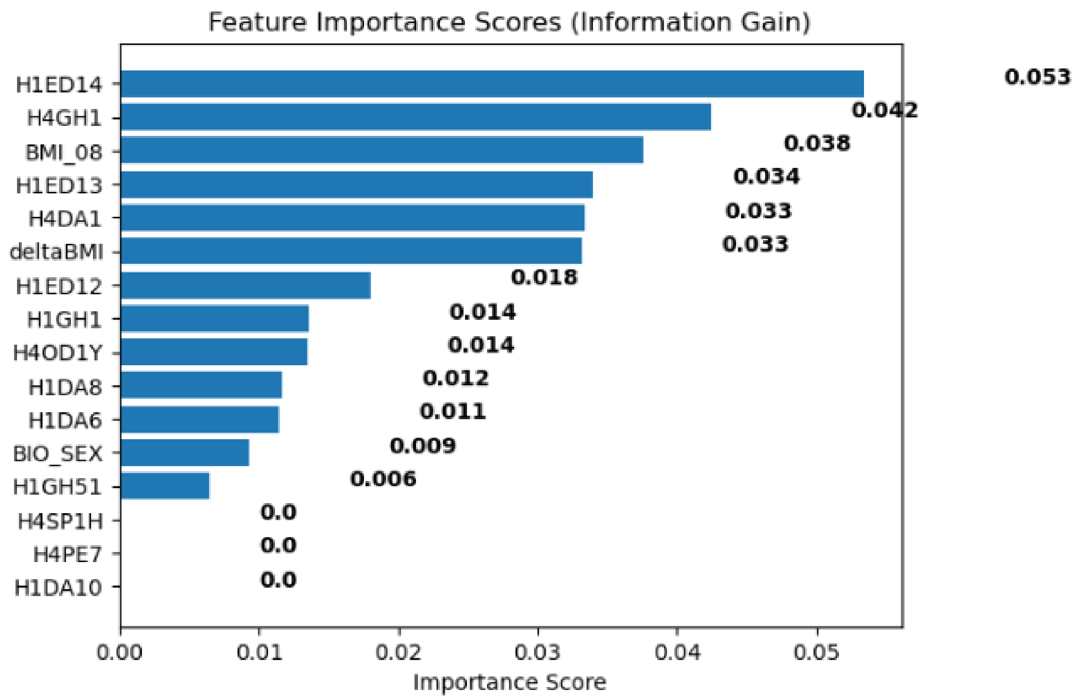


Fig 14: top features using regression based algorithm **mutual_info_regression**

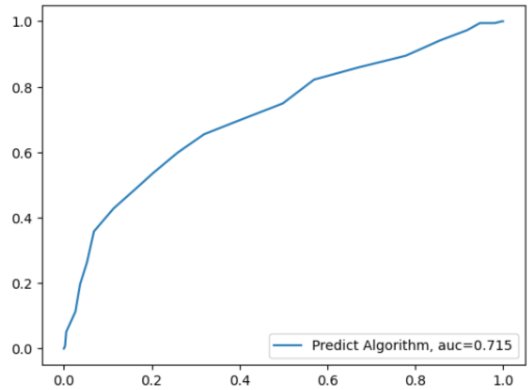
Results Evaluation

Ensemble modelling

An ensemble model was trained using following 4 classification based models

1. KNeighbors
2. RandomForest
3. LogisticRegression
4. Gaussian Naive Bayes

Model 1 : KNeighbors

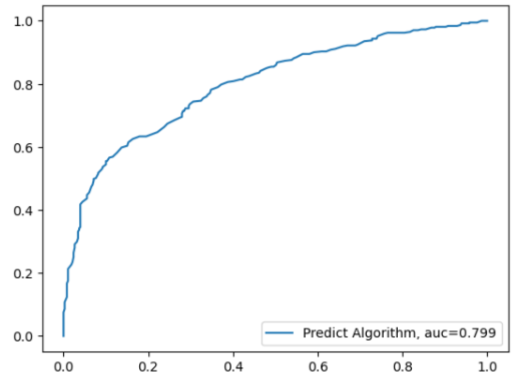


Confusion matrix

```
array([[258, 121],
       [128, 243]], dtype=int64)
```

```
Sensitivity 0.6549865229110512
Specificity: 0.6675824175824175
```

Model 2 : RandomForest

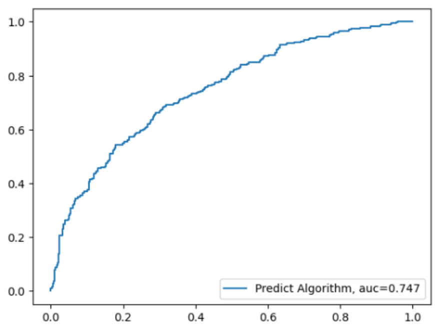


Confusion matrix

```
array([[267, 112],
       [103, 268]], dtype=int64)
```

```
Sensitivity 0.7223719676549866
Specificity: 0.7052631578947368
```

Model 3: Logistic regression

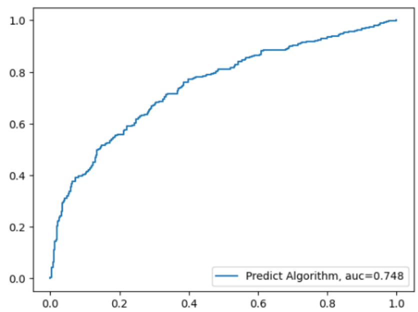


Confusion matrix

```
array([[272, 107],
       [134, 237]], dtype=int64)
```

```
Sensitivity 0.6388140161725068
Specificity: 0.688953488372093
```

Model 4: Gaussian Naive Bayes



Confusion matrix

```
array([[220, 159],
       [ 81, 290]], dtype=int64)
```

```
Sensitivity 0.7816711590296496
Specificity: 0.6458797327394209
```

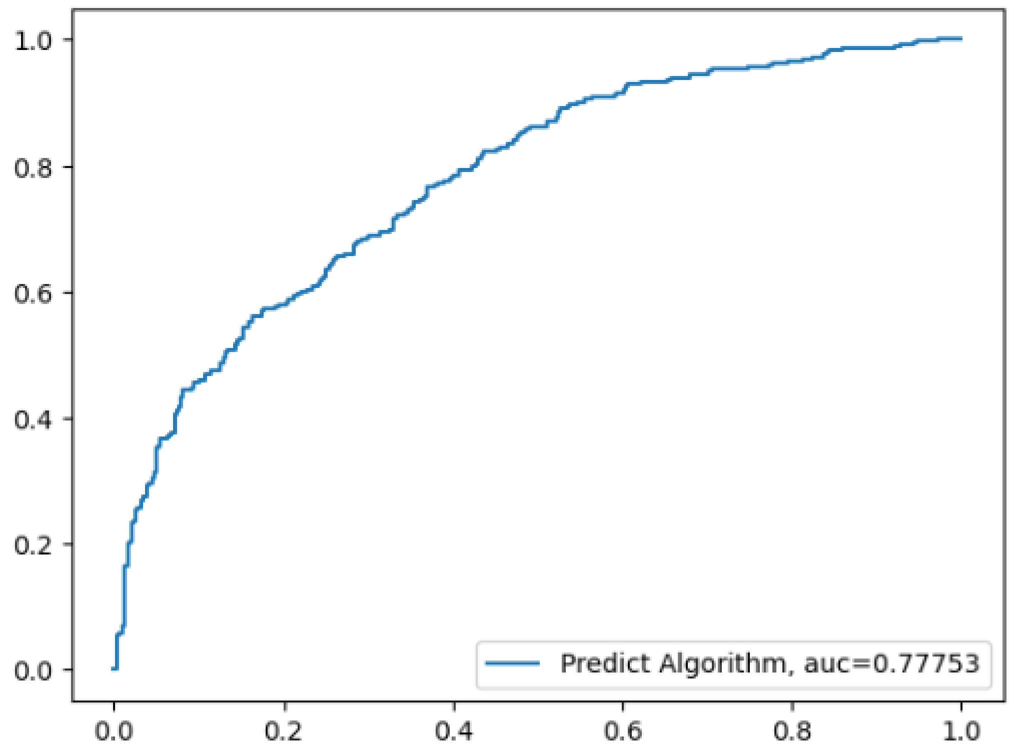
checking scores

```
print('knn: {}'.format(knn_best.score(X_test, y_test)))
print('rf: {}'.format(rf_best.score(X_test, y_test)))
print('log_reg: {}'.format(logreg.score(X_test, y_test)))
print('gnb: {}'.format(model_gb.score(X_test, y_test)))
```

```
knn: 0.668
rf: 0.7133333333333334
log_reg: 0.6786666666666666
gnb: 0.68
```

Fig 15: 4 classification models and their scores

Ensemble model:



Confusion matrix

```
array([[253, 126],
       [105, 266]], dtype=int64)
```

```
Sensitivity 0.7169811320754716
Specificity: 0.6785714285714286
```

Fig 16: Ensemble model scores

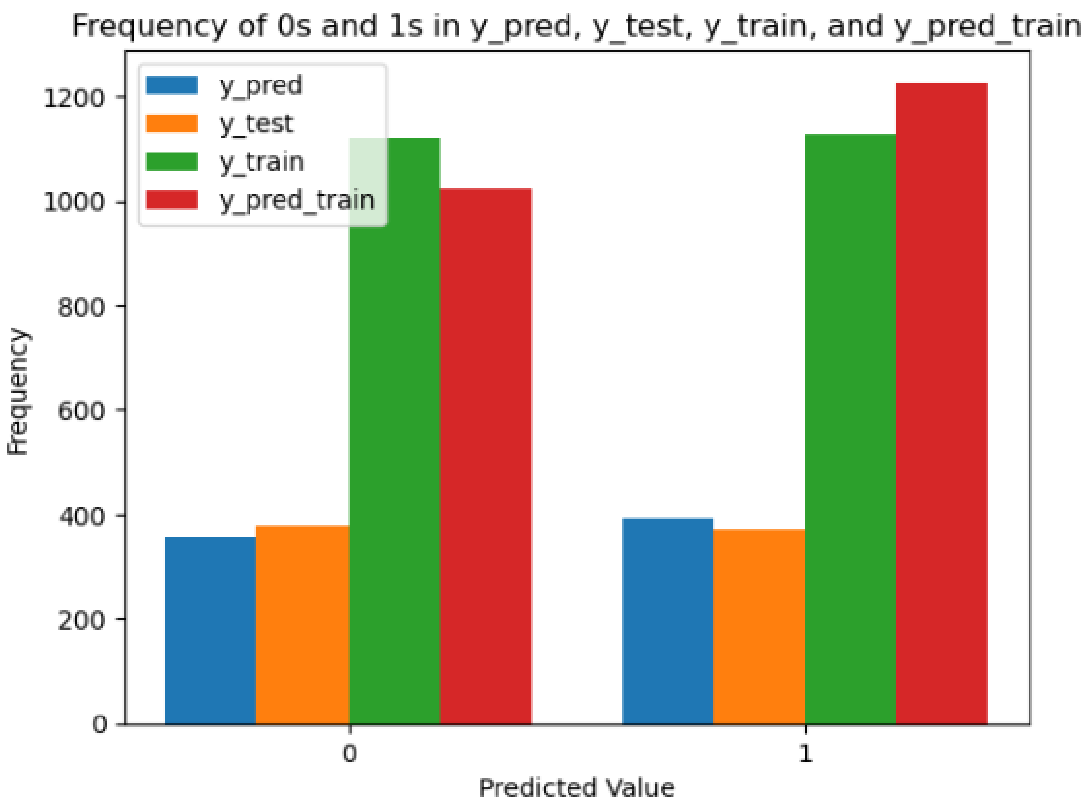


Fig 17: train, test and predict frequency.

Final Discussion and Conclusion

17 features were shortlisted from the dataset for investigation of whether the participants will end up studying till "graduate level or more" or "stop at college level".

Features BMI_08 and BMI_94 were calculated based on participants height and weight information. deltaBMI is difference between BMI_08 and BMI_94 indicating participants overall BMI / health trent.

Interquantile range method was used to remove outliers from numerical features. Q1 and Q3 quantiles were calculated to estimate IQR Interquartile range. Features had non-uniform units - kg, lbs, hrs, yes/no etc. The dataset was scaled to remove unnecessary influence of feature units.

According to feature scoring algorithm the top three features with mutual_info_classifier were H1ED13 Score: 0.057, BMI_08 Score: 0.037, H1ED14 Score: 0.036 mutual_info_regression were H1ED14 Score: 0.053, H4GH1 Score: 0.042,BMI_08 Score: 0.037 In addition, Recursive Feature Elimination, Cross-Validated (RFECV) feature selection method was given a try in the feature engineering process. Based on feature scores, two features were removed and remaining 15 features were used to train classification ML models.

An **ensemle model** (*Sensitivity 0.71, Specificity: 0.68 AUC: 0.77*) was generated using using following 4 models - KNeighbors (score: 0.668) , RandomForest (score: 0.7133) , LogisticRegression (score: 0.678), Gaussian Naive Bayes(score: 0.68).

Future Scope

Hyperparameter tuning is an essential step in optimizing the performance of a classification model. This is a time intensive process and due to time constraints, this step was not implemented in this report. In future, Hyperparameter tuning could be implemented to improve performance of models.

References:

1. Harris, K. M., & Udry, R. J. (2015). National Longitudinal Study of Adolescent to Adult Health (Add Health) Wave I, 1994-1995 [Data set]. UNC Dataverse. https://doi.org/10.15139/S3/11900
2. Harris, K. M., & Udry, R. J. (2015). National Longitudinal Study of Adolescent to Adult Health (Add Health) Wave IV, 2008 [Data set]. UNC Dataverse. https://doi.org/10.15139/S3/11920
3. Lecy, N., & Osteen, P. (2022). The Effects of Childhood Trauma on College Completion. Research in Higher Education, 63, 1058-1072. https://doi.org/10.1007/s11162-022-09677-9
4. The National Longitudinal Study of Adolescent to Adult Health (Add Health). https://addhealth.cpc.unc.edu/
5. Odum Institute Data Archive. https://odum.unc.edu/archive/
6. Heatmaps in Python. https://plotly.com/python/heatmaps/
7. Setting the Font, Title, Legend Entries, and Axis Titles in Python. https://plotly.com/python/figure-labels/
8. EDA. https://github.com/SaurabhPrabhu94/ANLY-530-Group-Project-Heart/tree/main
9. Binary Classification. https://www.learndatasci.com/glossary/binary-classification/#:~:text=Now%2C%20for%20the%20targets%3A%20dataset%20%5B%27target%27%5D.head%20%28%29%200,1%20357%200%20212%20Name%3A%20target%2C%20dtype%3A%20int64
10. Brown, D. W., Anda, R. F., Tiemeier, H., Felitti, V. J., Edwards, V. J., Croft, J. B., & Giles, W. H. (2009). Adverse childhood experiences and the risk of premature mortality. American Journal of Preventive Medicine, 37, 389-396. https://doi.org/10.1016/j.amepre.2009.06.021
11. Lecture notes
12. GeeksforGeeks. (n.d.). Data Mining Techniques. https://www.geeksforgeeks.org/data-mining-techniques/
13. Investopedia. (n.d.). Data Mining. https://www.investopedia.com/terms/d/datamining.asp
14. IBM. (n.d.). Data Mining. https://www.ibm.com/topics/data-mining
15. JavaTpoint. (n.d.). Data Processing in Data Mining. https://www.javatpoint.com/data-processing-in-data-mining
16. Springboard. (n.d.). Data Mining. https://www.springboard.com/blog/data-science/data-mining/
17. Barnett, R. (1990). The Idea of Higher Education. ISBN-0-335-09420-1.