Sachin Ramesh SU ID-384471510

Subgroup Analysis

The purpose of this project is to analyze the subgroup in the dataset consisting of the subjects studied for ASD, ASD+ADHD, VCFS, VCFS+ASD tests.

Subgroup analysis is an important aspect of design and analysis of clinical trials and one that can lead to misinterpretation of data. Subgroup analyses are often performed when no overall effect is found for a trial. They can also be used to search for high-risk or unusual groups with a marked effect. Subgrouping can help to find the underlying categories, which can be further analyzed in understanding the data.

Problem Definition:

In this project I am trying to answer/solve below 3 problems:

- Building a model to predict the categories, given features.
- Finding new groups, pattern or labels from the current dataset.
- If there is a new label can it be used to predict categories

About the data, there are 4 major categories and 90 Features associated with it

Models/Methods/Algorithms:

Missing Values:

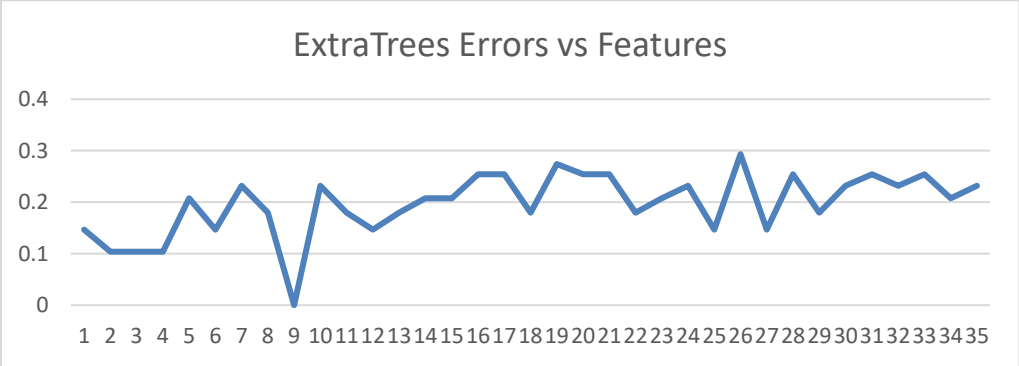
- There were many missing values in Age, I tried to handle these values by taking relative average of the Diagnosis and sex. for eg: for VCFS+ASD I filtered the results by sex say male, and took the average for the all males of VSFS+ASD and filled in that average value in the missing age values of male.
- For the features ADI AGE, BASC AGE, VINE AGE, the missing values were handled with ridge regression progressively trained on sex and age features. For eg: for the values present in ADI AGE, made that as target and Sex and Age being training set features, I predicted for the missing value of ADI AGE, and this predicted value was added again to training set to predict the next missing value and so on.
- The missing values of the features VCI ,PRI ,FSIQ, ADI_TOTA, ADI_TOTBV, ADI_TOTBNV, ADI_TOTC, ADI_TOTD, ADIDIRGZ, ADISOCSM, ADIRNGFE ,ADITOTA1, ADIIPw_P, ADINTICH, ADIRSPCH, ADIGPw_PF, ADITOTA2, ADISDATN, ADISHARE, ADISSEO, ADITOTA3, ADIUOBC, ADIOFRCM, ADISOCOVT, ADIINFAC, ADIASR, ADITOTA4, ADIPEI, ADINODNG, ADIHEDSH, ADIC_IG, ADITOTB1, ADISIA, ADIIMGPL, ADISOCPL, ADITOTB4, ADISV_C, ADIRC, ADITOTB2V, ADISUDE, ADIINQUS, ADIPR, ADIN_IL, TOT_B3V, ADIUP, ADICI, ADITOTC1, ADIVR, ADIC_R, ADITOTC2, ADIHFM_SBM, ADITOTC3 ADIPREOCC, ADITOTC4, ADIAPFN, ADIAFSW, ADIAFP, ADIAWAFE, ADIIJAAF, BPEXT, BPINT, BPBEHSYM, BPADPCOM, BPHYPER, BPAGGR, BPCONDCT, BPANXTY, BPDEPRS, BPSOMATZ, BPATYPCL, BPWITHDR, BPATTEN, BPSOCSK, BPLEADER were calculated with the average values as there were only 20-30 values missing among 369 rows

- The features ADIQIRSI, ADICOMM, ADIRBSP, ADIB36M, ADIAUTSM has binary values Y and N. I used logistic regression to predict the missing values of these features using all other features.
- BP_ADL, BP_FUNCCOM, BPADAPT, VINECOMM, VINEDLIV, VINESOC, VINECOMP each has at most 200 missing values, taking average would not provide meaningful values, so I decided to predict these values with the remaining features using ridge regression progressively (adding predicted values to the training set to predict next value)
- To validate these values, I tried to divide the Diagnosis (ASD, ADHD+ASD, VCFS, VCFS+ASD) into training and test data categorically and predicted the diagnosis with all the other features (using ridge regression). The result was as expected proving that missing values that were filled in were meaningful.

Feature Selection:

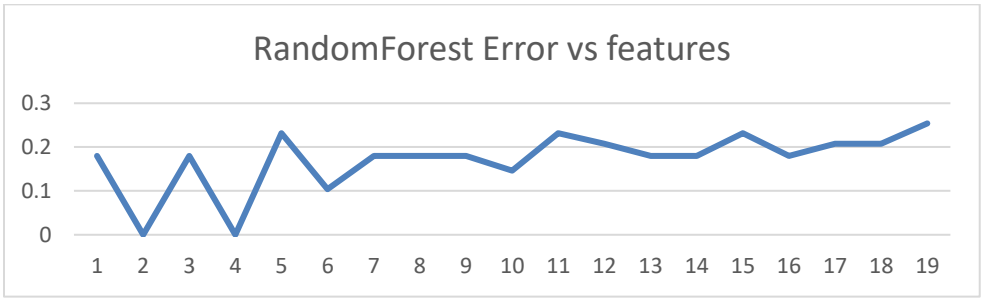
- The last three columns ADI, BASC, VINE indicates the categories of the tests performed on each subject. As per discussion with Dr. Russo, these columns were inconsistent and can be ignored. So, I removed these columns.
- Also, the first two columns Subject ID, Lab, years in months were also removed as they do not provide any meaningful intuition to the data and its already mentioned which lab conducted which of the diagnosis
- The Total features were removed as they were mathematical sum of other existing features having which was indicating the summed-up duplicates.
- Using the dataset with above changes I used Machine learning methods to select relevant features. The feature importance was measured using the co-efficients . Data was experimented with four methods to extract relevant features
    - ExtraTreeClassifier: Even though the performance is comparable to other decision tree algorithms, this algorithm produces piece-wise multilinear approximations.
    - rmse error corresponding to number of features with high value of importance, in predicting the diagnosis category was calculated.

      This method yielded 36 top features with below trend:



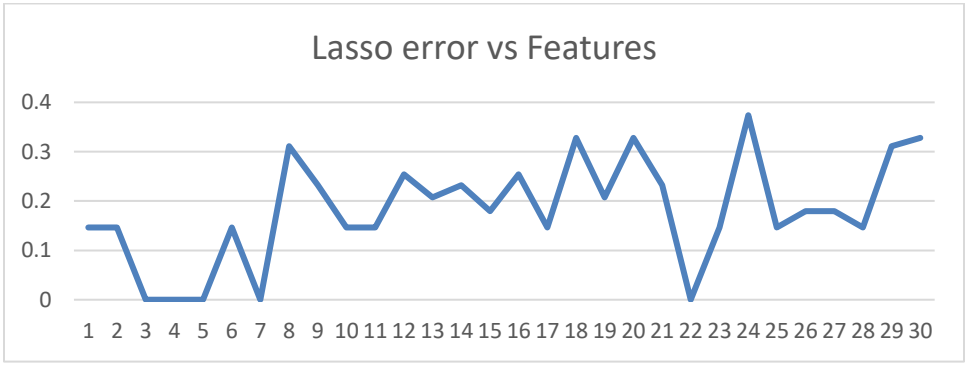
Based on above observation the top 8-10 features would give better error rate and will be more useful.

- Random Forests: Primitive version of extra tree classifier, yielded comparably satisfactory results.



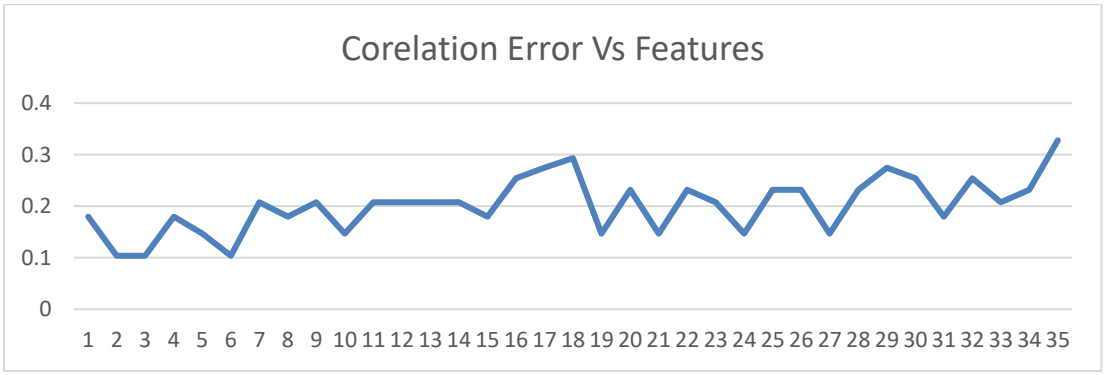
Even though top 4 features give lowest rmse choosing top 6 features would help to determine underlying meaning of the data.

- Lasso cross validation: Selects Features by shrinking non-contributing features to 0



Choosing top 8 or 23 key features would yield better result from above observation

- DesisionTreeRegression: I used this method to observe key features by co-relating each of them to be predicted by all other features in the model. By this method, if the score of the current target feature is less, then it is important in predicting other features so has more value. This indicates the dependencies of the tests for predicting category as a whole. The trend was:



Error rate is relatively uniform choosing top 6-10 features would help the cause

Based on the error trends the number of features corresponding to low error were selected from each model. The final feature set for further analysis was calculated based on AUC explained in next section

AUC:

Based on the features from above 4 models, each of the feature set were experimented with logistic regression (norm 2) and svm. Area under curve was calculated with different versions of data to understand the features. The area under (a ROC) curve is a summary measure of the accuracy of a quantitative diagnostic test.

Dimensionality Reduction:

I used PCA method to reduce the dimension of the dataset. The purpose of dimension reduction was to check if there were any subgroups in the data, clustering reduced dimensions. The number of dimensions helps to discover which dimensions about the data best maximize the variance of features involved. PCA will also report the *explained variance ratio* of each dimension — how much variance within the data is explained by that dimension alone.

The number of dimensions used was 2. This is because first two components represented 60% of the data.

Clustering:

Observations were made on two clustering algorithms.

- Kmeans Algorithm is robust, reliable and fast. It creates specified number of tightly bound clusters guranteed to converge after some iterations. As this is a supervised learning method it works at its best when the data is well defined and has features and labels have good co-relation and are clearly distinguished. Kmeans uses hard classification. It updates the mean point of the cluster using the popints in the same class.

- Gaussian Mixture model is highly flexible in terms of cluster Co-Variance. The membership of the data points can belong to more than one cluster which makes it easier ti identify related and relevant clusters. Uses soft classification, assumes means and deviations, each point records possibilities for all the classes.

Following steps were taken for creating clusters:

- Fitting a clustering algorithm to the reduced_data
- Predict the cluster for each data point in reduced_data using clusterer.predict
- Find the cluster centers using the algorithm's respective attribute.
- Predict the cluster for each sample data point in pca_samples

- calculate the silhouette score of reduced_data against preds to determine the number of clusters
- Visualizing the cluster

The maximum silhouette score for GMM method was 3 and maximum silhouette score for kmeans was 2. Both the clusters are created and analyzed. The true centers were recovered to map the cluster to each rows for analysis. Sample of 4 rows was chosen each from 4 distinct categories. The mapping of samples to the clusters was analyzed.

Libraries used:
Pandas
Numpy
sklearn.linear_model .LassoCV
sklearn.ensemble.ExtraTreesClassifier
sklearn.ensemble.RandomForestClassifier
sklearn.linear_model.LogisticRegression
sklearn.metrics.roc_curve, auc
sklearn.metrics.pairwise.cosine_similarity

Results and Plots:

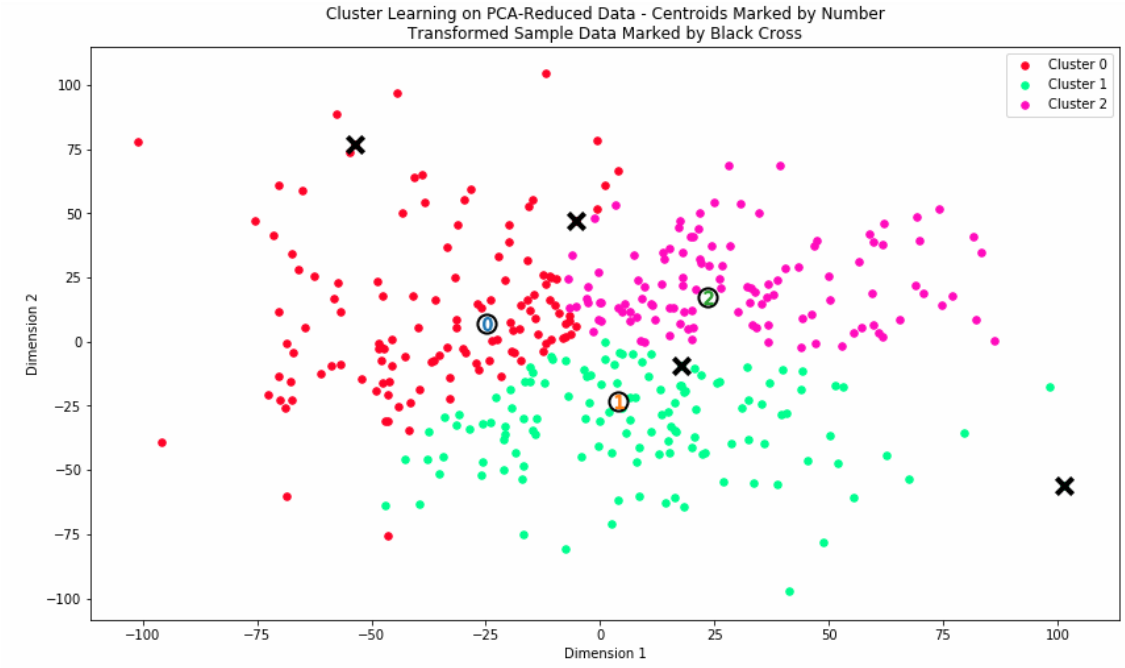
- The first part of the problem was to build a predictive model to predict any future tests into diagnosis. This was achieved by choosing Lasso features, from below observations:

	Logistic Regression		SVM	
Data Version	AUC	rmse	AUC	rmse
Without Eliminating any features	0.98	0.26	0.99	0.31
Eliminating only total values	0.99	0.29	0.98	0.28
Top Feature selection from ExtraTreeClassifier	0.94	0.33	0.92	0.34
Top Feature selection from LassoCV	0.96	0.37	0.98	0.37
Top feature selection from RandomForests	0.96	0.32	0.97	0.34
Top Feature selection from decisiontree regressor	0.93	0.43	0.9	0.4

- Clustering the reduced dimensions yielded two types of clusters from 2 different clustering algorithms
- Kmeans 2 cluster algorithm plot is as shown below:

- The samples were chosen from each one of the diagnosis category in order I,e first sample →ADHD+ASD, second sample →ASD, third sample →VCFS, fourth →VCFS+ASD

  Sample point 0 predicted to be in Cluster 0
  Sample point 1 predicted to be in Cluster 0
  Sample point 2 predicted to be in Cluster 1
  Sample point 3 predicted to be in Cluster 1
- We can draw a conclusion that the whole data is divided into two main clusters into dominating diagnosis categories ASD and VCFS. ADHD+ASD has merged into ASD and VCFS+ASD has merged into VCFS.
- The 3 clusters from GMM method is visualized as below:



Sample point 0 predicted to be in Cluster 1
Sample point 1 predicted to be in Cluster 1
Sample point 2 predicted to be in Cluster 0
Sample point 3 predicted to be in Cluster 0

- One of the possible conclusions here is that cluster 1 can be the ASD, cluster 0 can be VCFS and t he third cluster can contain the data grouped as ADHD.
- Another possibility is ADHD and ASD can be grouped to cluster 1, VCFS could be grouped to clust er 0 and the cluster 2 could be showing another group with different diagnosis which overlaps w ith the same tests as our four original categories.
- The above two plots answer our second problem of subgrouping
- As for the third problem, with the reasonable domain knowledge to segregate the tests into sep arate diagnosis, it can be predicted to which of the three clusters that a given test results can fall into.

Other Observations:

- FSIQ even though chosen as one of the key features, alone cannot predict the diagnosis category . I tried this with regression and the error rate was 0.72
- Both in Genetic(VCFS) and Behavioral (ASD, ADHD) diagnosis male population seems to dominat e among the subjects. But, the sex feature found to be not so important in determining the diag nosis.
- I tried finding mathematical similarity(cosine) among each diagnosis with the same sample data used for clustering. The result was as expected except for the third sample. Below are the result s:

Sample	0 is	98.1501001813% similar to ADHD+ASD
Sample	0 is	98.1895530703% similar to ASD
Sample	0 is	96.8869244561% similar to VCFS+ASD
Sample	0 is	92.9399297778% similar to VCFS
Sample	1 is	94.7307551712% similar to ADHD+ASD
Sample	1 is	95.0155504621% similar to ASD
Sample	1 is	92.9793469021% similar to VCFS+ASD
Sample	1 is	85.756118627% similar to VCFS
Sample	2 is	97.1200298435% similar to ADHD+ASD
Sample	2 is	96.7999136225% similar to ASD
Sample	2 is	96.3585031328% similar to VCFS+ASD
Sample	2 is	94.9839162258% similar to VCFS
Sample	3 is	95.3522592678% similar to ADHD+ASD
Sample	3 is	94.6302736827% similar to ASD
Sample	3 is	95.2320906309% similar to VCFS+ASD
Sample	3 is	98.0746301437% similar to VCFS

The third same is expected to be more similar to VCFS.

Limitations:

- The features represent the tests conducted on subjects, some of the tests may be missed purpo sefully, but we assume it as a missing data and estimate the values.

- Limited domain knowledge, due to which all the attributes are just observed with mathematical importance, with the knowledge of underlying meaning more specific and meaningful models ca n be computed.
- The clustering conclusions are subjective based on observation. With the known probability of s ubgroups existence, more meaningful conclusions can be drawn.

REFERENCES:

1) http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001544
2) http://centaur.reading.ac.uk/32453/1/journal.pbio.1001544.pdf
3) http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0074873&type=printable