Classification of Eating Disorder using Logistic Regression

Jay Khandelwal
Student, Department of computer applications
Sardar Patel Institute of Technology
Mumbai, India

Jay Visave
Student, Department of computer applications
Sardar Patel Institute of Technology
Mumbai, India

Ninad Patwardhan
Student, Department of computer applications
Sardar Patel Institute of Technology
Mumbai, India

ABSTRACT
Abnormal eating behavior which has a negative impact on a human body is called an eating disorder. Eating disorders generally occur in teenage years and cause disturbances in diet and health. There are various types of eating disorders where the two most prominent types are Anorexia Nervosa (AN) and Bulimia Nervosa (BN). Anorexia refers to eating in very small quantities which is not sufficient to a human body and Bulimia refers to heavy or excess eating which is harmful as well. In this paper, the focus was on Anorexia Nervosa. People in anorexia tend to over consciously control their weight and shape. In order to prevent weight gain, people with anorexia strictly limit the amount of food they consume, they even vomit purposely after eating to reduce the calorie intake or even take misleading medicines to control their weight. The disease can take over one's life and can be very tough to overcome. But proper treatment can help an individual to get to know themselves and move towards a healthier lifestyle. For people to be able to recognize if they have this disease or not, have developed a machine learning model using logistic regression to predict the possibility of Anorexia. The used dataset is a valuation in adults with a history of anorexia nervosa from Department of Neuroscience, University of Pennsylvania published by Duke Digital Repository. The dataset consists of records of people who are suffering from this disease. It consists of various factors like current BMI, age etc.

Keywords
Eating Disorder, Logistic Regression

1. INTRODUCTION
Eating disorders (EDs) are disorders that appear during the teenage years that cause major disturbances to an individual's usual diet, like eating extremely small quantities of food substances or highly overeating. Females have been demonstrated as a potent risk factor for eating disorders, but how much this association can be attributed to biological rather than social factors are uncertain. The two prominent clinical constitutions of EDs are anorexia nervosa (AN) and bulimia nervosa (BN).

2. TYPES OF EATING DISORDERS

2.1 Anorexia Nervosa (AN)
AN is a serious mental disorder that is found to be fatal in approximately 10% of cases. According to the new DSM-5 criteria, to be diagnosed as having AN, an individual must display (a) persistent restriction of energy intake relative to requirements resulting in a significantly low body weight; (b) intense fear of gaining weight or becoming fat, although they're underweight; and (c) disturbance within the way within which one's weight or shape is experienced, undue influence of weight or shape on self-evaluation, or denial of the seriousness of the present low weight.[2]

2.2 Bulimia Nervosa (BN)
Bulimia Nervosa is characterized by unmanageable episodic overeating, called binging, followed by purging with methods such as self-induced emesis or inappropriate use of laxatives. Binging is eating larger quantities of food than a person would usually eat in a short span, usually in less than two hours.[2]

2.3 Pica
Pica disorder that involves eating substances that are not considered food is another eating disorder. People with pica disorder crave non-food substances, such as ice, dirt, soil, chalk, soap, paper, hair, cloth, wool, pebbles, laundry detergent, or cornstarch. Pica can be seen in adults, children and adolescents. This type of disorder is most adequately found in children, pregnant women, and individuals with mental disabilities. Individuals with pica disorder may be at a high risk of poisoning, infections, gut injuries, and nutritional deficiencies depending on the substances ingested. Pica may be termed as fatal when involved in ingesting harmful substances.[2]

2.4 ARFID
Avoidant/restrictive food intake disorder (ARFID) is a new name for an old disorder which refers to what was known as a "feeding disorder of infancy and early childhood," a diagnosis previously reserved for children under seven years of age. ARFID generally develops during early childhood, it can persist into adulthood, found to be equally common among men and women. People with this disorder experience disturbed eating either due to a lack of eating interest or distaste for certain smells, tastes, colors, textures, or temperatures.[5]

2.5 Other Disorders
Other types of eating disorders fall under one of three categories: (a) Purging disorder in which individuals often use purging behaviors, such as vomiting, laxatives, diuretics, or excessive exercising, to control their weight or shape. However, they do not binge. (b) Night eating syndrome in which individuals frequently eat excessively, often after awakening from sleep. (c) Other specified feeding or eating disorders (OSFED) while not found in the DSM-5 includes any other conditions that have symptoms similar to those of an eating disorder but don't fit into any of the categories above.[2]

3. ANOREXIA NERVOSA

Until now, diagnosis of EDs are based only on clinical trials complemented by exams and tests which includes: (a) A physical diagnosis where a doctor will examine a person to rule out other medical conditions for eating problems the person experiences. The individual may also order a lab examination. (b) Psychological tests are conducted where a psychologist will probably ask about thoughts, feelings and eating habits for diagnosis. A complete psychological self-assessment questionnaire may also be a part of such examination. (c) Other studies involving different methodologies. Additional assessments may be done to diagnose any complications related to eating disorders in a person. However, EDs diagnosis is unstable, with clinical characteristics varying over time and often shifting from anorexia to bulimia. For this reason, there is an emergent need to identify biologic symptoms which may be used for assisting and improving early diagnosis, treatment planning, and monitoring of disease progression. In the last 10 years, considerable efforts have been expended in developing advanced neuroimaging techniques.[3]

As a result, an abundance of functional and structural neuroimaging studies has been conducted to unravel the physiology of abnormal states, specifically mechanisms of EDs. The vast majority of these studies reported in AN patient's global reductions of total gray and white matter, as well as cortical thickness, a number of recent studies have emphasized regional group differences. Remarkable results have been attained, the disadvantage of these studies is that they reported neurobiological abnormalities comparing patients and controls at a group level, while limiting clinical trials at the individual level. Due to this, alternative kinds of analyses of neuroimaging data are being focused for development. A growing interest within the neuroimaging community in classification methods, including machine learning methods has been seen in the last couple of years. These methods are based on algorithms that are able to automatically extract multiple parts of information from image data sets without requiring a-priori hypotheses of where they may be found on images. The aim of these techniques is to maximize the distance between image groups in order to classify individual structural or functional brain images. Studies have tested the clinical relevance of such methods showing very favorable findings mainly in the neurological studies. For instance, machine learning techniques are able to classify very reliable imaging biological symptoms allowing individual diagnoses of Alzheimer's disease, Mild Cognitive Impairment and Parkinson's disease with an accuracy above 90%. In psychiatric studies, this kind of advanced neuroimaging method is in its relative infancy. Although some interesting applications have been made in patients with posttraumatic stress disorders, depression disorders, and first-episode psychosis, there are no studies investigating the potential role of these methods in EDs.

For these conditions, the study was aimed at employing a validated supervised machine learning method to classify features useful to distinguish individuals with diagnosis of ED patients from healthy individuals by means of symptoms examined via a questionnaire. This method makes use of the dataset that is a valuation in adults with a history of anorexia nervosa from the Department of Neuroscience, University of Pennsylvania published by Duke Digital Repository [4] in order to extract the most informative features from the existing dataset, while the Logistic Regression approach was used to perform classification.[6]

4. CLASSIFICATION MODELS

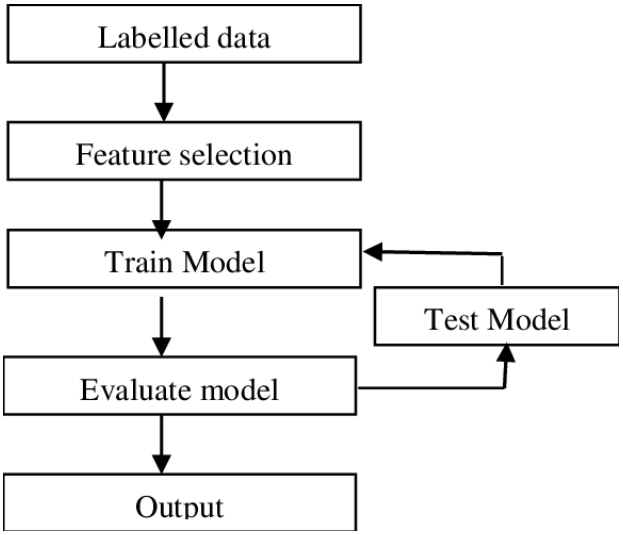


Fig 1: Steps taken to develop the model

4.1 Logistic Regression

Logistic Regression utilizes regression methodology to do classification. The reason for the model's success is its power of explain ability i.e., calling-out the contribution of individual predictors quantitatively.[8] The model uses Maximum Likelihood for fitting a sigmoid-curve on the target variable distribution. The model's susceptibility to multicollinearity, applying it stepwise, turns out to be a better way in finalizing the predictors of the model. The algorithm is a popular choice in many natural language processing tasks e.g. predict the likelihood of disease or illness for a given population, toxic speech detection, topic classification, etc.

4.2 Artificial Neural Networks

Artificial Neural Networks (ANN) are suitable for large and complex datasets. Their structure comprises layers of intermediate nodes which are mapped together to the multiple inputs and the target output. It is a self-learning algorithm, in that it starts out with a random mapping and thereafter, iteratively adjusts the related weights by itself to fine-tune to the desired output for all records.[8] The multi-layer network provides a deep learning capability for extracting higher-level features from the raw data, providing high prediction accuracy but needs to be scaled numeric features. It has a huge range of applications in upcoming fields including Computer Vision, NLP, Speech Recognition, etc.

4.3 Naïve Bayes

Naive Bayes makes an assumption that each feature makes an independent and equal contribution to the outcome. This is the algorithm that's most commonly used to filter through spam emails. It applies a posterior probability using Bayes Theorem to perform categorization on the unstructured data.[8] While doing this, it makes a naïve assumption that the predictors are independent, which may not be true. The model works well with a small training dataset, requiring all the classes of the categorical predictor are present.

4.4 Random Forest

A Random Forest is a reliable ensemble of multiple Decision Trees though more popular for classification, than regression applications. The individual trees are built using aggregation of bootstraps which are nothing but multiple train datasets created sampling of records with replacement and split using

fewer features. The resulting diverse forest of uncorrelated trees exhibits reduced variance and therefore is more robust towards change in data and carries its prediction accuracy to new data.[8] The algorithm does not work well for datasets having a lot of outliers, something which needs addressing prior to the model building. It has wide applications across Financial, Retail, Aeronautics, and many other domains.

4.5 KNN

K-Nearest Neighbor (KNN) algorithm predicts based on the specified number (k) of the nearest neighboring data points. Here, the pre-processing of the data is important as it impacts the distance measurements significantly. The model does not have a mathematical formula, nor any descriptive ability. Here, the parameter 'k' needs to be chosen wisely, a lower value than optimal leads to bias, whereas a higher value impacts prediction accuracy. It is a simple, fairly accurate model preferable mostly for smaller datasets, owing to huge computations involved on the continuous predictors. At a simple level, KNN may be used in a bivariate predictor setting e.g., height and weight, to determine the gender given a sample.

5. CODE SNIPPET

```
# Normalization of the dataset
X = preprocessing.StandardScaler().fit(X).transform(X)

# Train-and-Test -Split
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(
        X, y, test_size = 0.2, random_state = 4)
print ('Train set:', X_train.shape,  y_train.shape)
print ('Test set:', X_test.shape,  y_test.shape)

# model_train
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()
logreg.fit(X_train, y_train)
y_pred = logreg.predict(X_test)
```

6. FINDING AND ANALSYIS

Logistic regression is used to predict the eating disorder of individuals based on one or multiple predictor variables. It is used to model a binary outcome, that is a variable, which can have only two possible values: no disorder or disorder.

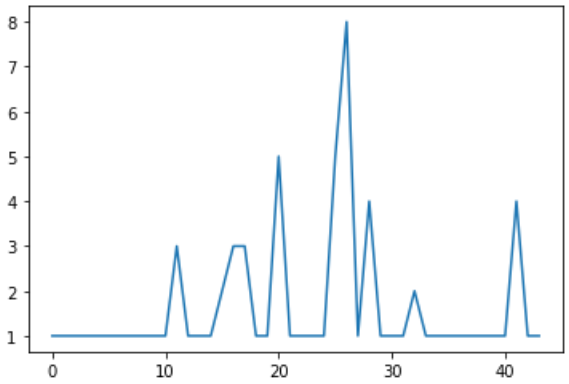


Fig. 2 Participant reported time since engagement in the disordered behaviour of binge eating (XAxis – Age rangein Years)

Table 1. Description of Binge Eating Plot (Y Axis)

Value	Description
1	Never Did This
2	Over 5 years
3	Over 2 years
4	Over 1 year
5	Less than 1 year
6	Less than 6 months
7	Less than 1 month
8	Still struggling

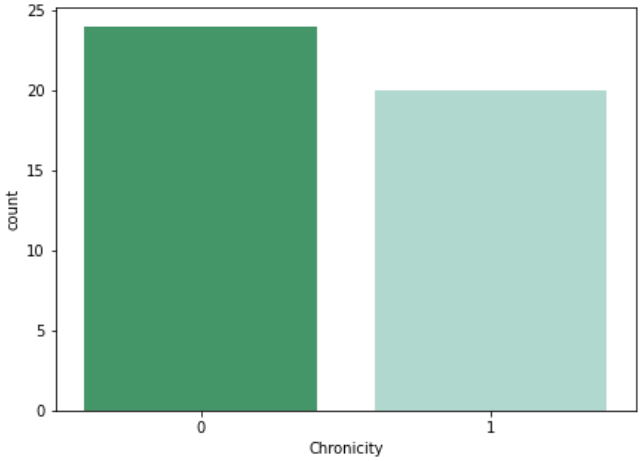
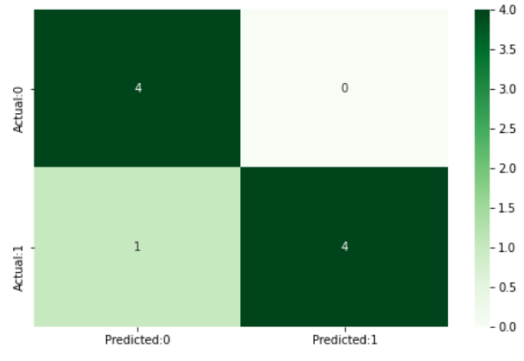


Fig. 3 Number of patients affected by eating disorder. (X Axis - Chronicity – Presence of Disorder,Y Axis – Number of Patients Examined)

Table 2. Description of Patients affected

Value	Description
0	No Disorder
1	Disorder



The details for confusion matrix is =

	precision	recall	f1-score	support
0	0.80	1.00	0.89	4
1	1.00	0.80	0.89	5
accuracy			0.89	9
macro avg	0.90	0.90	0.89	9
weighted avg	0.91	0.89	0.89	9

Fig. 4 Confusion Matrix

As per the results the ML model using Logistic Regression was found to have an accuracy of 89% obtained using Confusion Matrix where the model was trained over 80-20(training-testing) data.

7. CONCLUSION

To date no relevant studies, exist that have addressed a statistical detection of an ED by means of ML approaches using Logistic Regression. EDs in general are often undetected and untreated, which leads to a worsening of the individuals' symptoms and higher costs for medicating individuals in advanced disorder stages. ED is the most common ED and therefore an important health problem worldwide resulting in obesity.

For this reason, it is essential to make use of ML approaches in the context of a BED in order to facilitate identification of the disorder. Hence, the approach uses a Logistic Regress, and both classify with an accuracy of 89% and identify highly important variables for an accurate classification out of a huge number of input variables. For distinguishing between participants affected by an ED and not affected by an ED, the model highlighted these parameters as the most important ones.

8. FUTURE SCOPE

BED is a disorder increasing in adolescents.[1] If the disorder is not identified and treated at an early stage, the impact could be worse as the age increases. Apart from the mentioned disorder, there are more types of eating disorders such as bulimia nervosa, rumination disorder, binge eating disorder, Pica, ARFID (Avoidant/Restrictive food intake disorder), etc. The aim would be to use these insights and the algorithm to detect the above-mentioned disorders as well. Furthermore, we would like to extend the current scope by using a variety of machine learning algorithms.

9. REFERENCES

[1] Raab, Dominik & Baumgartl, Hermann & Buettner, Ricardo. (2020). Machine Learning Based Diagnosis of Binge Eating Disorder Using EEG Recordings.

[2] D. R. Dwiki Putri, L. Sipahutar, M. Reza Fahlevi, R. Utami, F. Pranita Nasution and E. Syahrin, "Identification of Anorexia Nervosa and Bulimia Nervosa Eating Disorders Using the Dempster Shafer Method," 2020 8th International Conference on Cyber and IT Service Management (CITSM), 2020, pp. 1-5, doi: 10.1109/CITSM50537.2020.9268817.

[3] C. S. D. Leon, D. A. V. Casas and J. P. J. M. Lopez, "IoT technological model to improve the control and monitoring of patients with eating disorders (ED): Anorexia and Bulimia in a mental health hospital," 2021 IEEE Sciences and Humanities International Research Conference (SHIRCON), 2021, pp. 1-4, doi: 10.1109/SHIRCON53068.2021.9652296.

[4] Sweitzer, M. M., Watson, K. K., Erwin, S. R., Winecoff, A. A., Datta, N., Huettel, S., ... Zucker, N. L. (2018). Data from: Neurobiology of social reward valuation in adults with a history of anorexia nervosa. Duke Digital Repository. https://doi.org/10.7924/r45t3km1m

[5] Janet Polivy and C. Peter Herman. (2002). Causes of Eating Disorders.

[6] P. Reynolds, "The Biology Behind Eating Disorders," in IEEE Pulse, vol. 11, no. 6, pp. 17-20, Nov.-Dec. 2020, doi: 10.1109/MPULS.2020.3037985.

[7] Lauren N. Forrest, Valentina Ivezaj, Carlos M. Grilo. (2021) Machine learning v. traditional regression models predicting treatment outcomes for binge-eating disorder from a randomized controlled trial.

[8] Definition data about various eating disorders. https://www.analyticsvidhya.com/blog/2020/11/popular-classification-models-for-machine-learning/