**Framework for Automatic Diagnosis of Psychological Disorders**

Girish P Mallya

School of CSE

KLE Technological University

Hubli, Karnataka, India

girishm60@gmail.com

Amaresh Hiremani

School of CSE

KLE Technological University

Hubli, Karnataka, India

mrshhiremani@gmail.com

Ganesh Gajakosh

School of CSE

KLE Technological University

Hubli, Karnataka, India

ganeshgajakosh01@gmail.com

Darshan Badiger

School of CSE

KLE Technological University

Hubli, Karnataka, India

darhsanbadiger@gmail.com

Ms. P.G. Sunitha Hiremath

School of CSE

KLE Technological University

Hubli, Karnataka, India

pgshiremath@bvb.edu

***Abstract*— Diagnosis of mental health disorders can be challenging even for experienced professionals. An interview is conducted by medical professionals and based on the interview the disorder is identified. World Health Organization has predicted that nearly one in five people suffers from mental illness. If mental illness is diagnosed at an early stage it can be easier to treat, therefore, the focus is on the age group of 5-24. Data regarding symptomatology and diagnoses are collected with the collaboration of psychiatrist Dr. Anand Pandurangi. A web application developed with MEAN stack is provided to the receptionists at the clinic to help collect the preliminary details of the patient, and the same is provided to the doctor to write the diagnoses. Based on this diagnosis the symptoms are generated and stored in a database. This data is used to build a predictor model. The predicted output is given to the doctor as a suggestion. LogisticRegressionCV and KNeighbor classifier are used to classify the disorder based on the data of the patient. Based on the outcome of this result, we can avoid inaccurate diagnoses and subsequently prevent the irreversible damage to one’s mental health. Data analysis is performed to identify the causes and relationships between the symptoms and diseases.**

# ***Keywords: Cross validation, Logistics Regression, KNeighbor classifier, MEAN stack, WHO***

# Introduction

Mental health can be defined as the way in which people react to the situations they face in life. These situations could be social, emotional or psychological. It adversely affects one’s productivity in professional life, can ruin relationships in personal life and may end up being a problem to society rather than contributing towards its well-being. As such, it becomes important to understand the various facets of mental health issues, their symptoms, and their causes. Diagnosis of mental health illnesses is hard as there is no objective metric to determine one’s mental health status. It is dependent on the subjective knowledge of the psychiatrist. If we achieve a better understanding of these symptoms, their causes, and how they are correlated to one another, this can provide insights on prevention or treatment of mental illnesses.

An effective strategy to tackle mental health illness is to diagnose and treat it at early ages. Therefore, the focus of the study is on the age group 5-24 years. This age group is highly prone to mental illness and do not have the requisite life experience to handle the stress and anxiety. This age group also plays a crucial role in the individual’s personality development, which in turn affects the productivity of the individual and by extension, his contributions to the nation. This younger generation is responsible for carrying the nation to glory in the coming years. NMHS has revealed that one among six Indian needs mental health care [1]. To understand these mental health illnesses and their effects, we contacted a local psychiatrist, Dr. Anand Pandurangi of Shree Psychiatric Clinic, Dharwad.

# Literature Review

There have been many studies conducted by various organizations around the world to improve the mental healthcare. Ms. Sumathi M.R. & Dr. B. Poorna in [2] performed a prediction of Mental Health disorders among children using 8 machine learning techniques, This study focused mainly on the 5 most common mental health illnesses amongst children - Attention problem, Academic Problem, Anxiety Problem, Attention Deficit Hyperactivity Disorder (ADHD) and Pervasive Developmental Disorder (PDD), the factors, symptoms and various test results that are observed by the professionals are given as input to the techniques and the psychological problem diagnosed is retrieved as their output. Successful and accurate predictions were generated by 3 machine learning techniques viz.  Multilayer Perceptron, Multiclass Classifier, and LAD Tree. The research was performed on a small dataset (n=60).

Various researchers in the USA and China studied data about misdiagnosis among patients and its effects in [3] and attempted to provide a more intelligent solution to the same by introducing a semi-autonomous way that aids in the preliminary diagnosis of the patients.

They achieved the same by integrating the technologies of the genetic algorithm, classification data mining, and machine learning. The genetic algorithm was used here to extract the information regarding the mental illness from the symptoms entered. This study focuses on the mental health issues illustrated in ‘Diagnostic Statistical Manual IV, Fourth Edition Text Revision (DSM-IV-TR). Keywords from this textbook about each disease have been loaded into a database. A genetic algorithm has been developed to extract keywords from a patient’s symptoms (using BPEL), these keywords are matched with text in the database to extract a classification.

The algorithm is tested on a set of sample data using the same random sample set of 15 symptoms with 600 words each. The algorithms produced a better fitness on average compared to the k-means algorithm (average fitness of 1.569 compared to average fitness of 1.678 of K-Means). The limitation of this research is that the study has concentrated on a sample of 15 symptoms.

In [4] multi-criteria decision support method called MACBETH (Measuring Attractiveness by a Categorical Based Evaluation Technique) is used. This study emphasizes on the difficulty in diagnosis of the mental disorder and its severity, and tries to improve the quality of the diagnosis, focusing on the Obsessive Compulsive Disorder was used to generate set of rules.

In [5], to predict the mental health disorder the paper describes the use of a machine learning technique called Bayesian network and Fuzzy clustering. The dataset for this research is collected from a psychologist, consisting of 77 of children profile after pre-processing. Applying a feature selection algorithm to select the relevant attributes a Bayesian Network was constructed using the GeNie tool and the parameters were learned from the dataset. The model was validated with 2-fold cross-validation. This network is called the Expert Knowledge-based Bayesian Network (EKBN). Using the GeNie tool, a BN was generated automatically from data (DBN). The predictive accuracies of EKBN and DBN were 48% and 58% respectively.

The research conducted [6], by Changye Zhu,Baobin Li,Ang Li and Tingshao Zhu aimed to Predict Depression from the internet behaviors by Time-Frequency Features. For the purpose of study 728 postgraduate students were recruited. In this study scores on a depression questionnaire and digital records of Internet behaviors are obtained. This research used clustering and time-frequency analysis (such as DFT) to extract features and build classification models to recognize Depression. Usage of Time-frequency features had improved the performance of models. The precision and recall of Naïve Bayes algorithm compared with Back Propagation neural network and Decision tree was better with a score of 75.6 and 62.3 respectively considering all the features. They are conducting more studies to improve accuracy by using techniques like wavelet transforms in their further approach.

# data description

The symptomatology and diagnoses were collected with the help of the psychologist and psychiatrists at the clinic. Ethical clearance for usage of the data was also obtained.

The socio-demographic details and information about the identity of the patients were not revealed to us.

The symptoms are represented in separate columns in Binary wherein 1 represents the patient having the symptom and 0 represents the patient not having the symptom. The symptoms include- Dull, Self-talking, Aggression, Irritated, Suicide Attempts and some physique related symptoms like sleeplessness, weakness, restlessness etc. There were 40 symptoms considered for the study.

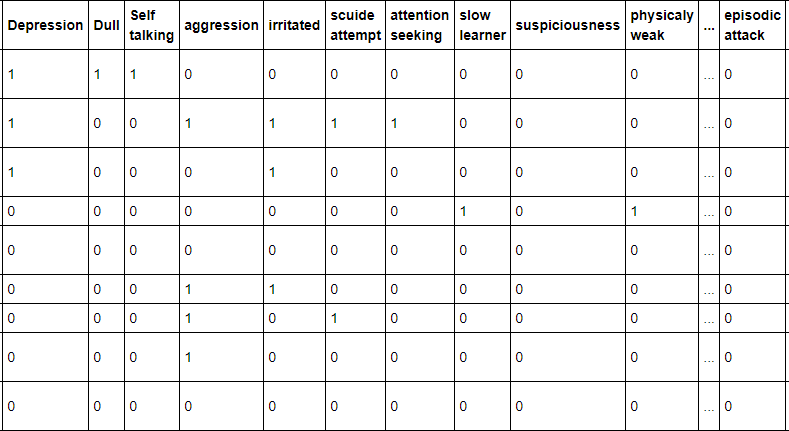


Figure 1: Data Sample.

Figure 1 shows the structure of the dataset being used in the study.

# data preprocessing

For a relatively smaller size of data (N=250), the highly correlated features can produce harmful bias in the data. This can subsequently affect the performance of the predictive model. Pearson’s correlation heat map revealed some highly correlated features like ‘suspiciousness’ and ‘delusion’ with a Pearson’s correlation of 0.75. A statistical technique called multicollinearity analysis [7] was adopted to tackle this issue. One feature is predicted from the other features with a substantial degree of accuracy in this technique. Using this, Variance Inflation Factor (VIF) for each feature in the predictive model was calculated. It is the ratio of variance in a model with multiple features, divided by the variance of a model with one feature alone. For severely correlated features a very high VIF score of 5 or higher was obtained. All the features which produced VIF of higher than 5 were eliminated in the model [8].

Suppose there is a dependent variable ‘Y’, with independent variables,and. So, the regression equation:

(1)

Where is the change in Y for a 1-unit change in, while is held constant, and is the change in Y for a 1-unit change in, while and is held constant.It is required to find VIF

VIF= (2)

Here is the coefficient of determination obtained by regressing over all the other independent variables.

For instance, is obtained by regressing over all other independent variables.

(3)

1. vif factor of few features

|  |  |  |
| --- | --- | --- |
|  | **VIF Factor** | **Features** |
| 1 | 3.59 | Depression |
| 2 | 8.545 | Dull |
| 3 | 4.84 | Self-talking |
| 4 | 5.86 | Irritated |
| 5 | 3.75 | Attention seeking |
| 6 | 4.64 | Slow learner |
| 7 | 8.84 | Suspiciousness |
| 8 | 5.05 | Physical weak |
| 9 | 6.53 | Anxiety |
| 10 | 4.26 | Impulsive |
| 11 | 6.34 | Wandering tendency |
| 12 | 7.95 | Delusional |
| 13 | 9.784 | Emotionally disturbed |

Table1 shows the VIF factors obtained for the first few features. All the features with VIF factor of above 5 were eliminated. The resulting data had 21 features left.

# Design

# *System model*

## 

Figure 2: System model with detail components.

The diagrammatic flow of the system is shown in figure 2.A web application is created where a new patient can register by providing their details. The registration process will be done by the receptionist on the web application in the hospital and the registered patient can take an appointment with his preferred date and time. Then the list of appointment will be displayed in a date-wise manner to the doctor who will access the app on an iPad. Here he can select the appointment and then write the case card which includes the symptoms, prescription, and disorder the patient is suffering. From the symptoms, the features required to predict the disorder will be extracted by the custom script written in JavaScript. Considering the key symptoms, a request is made with the features via post method and once request received by the server it will send back the predicted label with accuracy which gets listed to doctor as a suggestion. The website helps to organize the patient details. This lessens the burden of organizing the hard copy versions of the case cards which is in use currently. Since the psychological details of patients and personal information is considered privileged, there is a need for high security in the database. Therefore, every patient’s details are encrypted and stored in a MongoDB server. The data extracted from the database is pre-processedand used to perform analysis and predictive modelling. The prediction is shown to the doctor at the time of diagnosing the patient to help assist in the identification of the disorder.

# results

# *descriptive analysis*

While observing the data that was available for the study, some analysis was done and some trends were discovered in the data. Within the data, it was found out that some of the disorders are more common. Disorders like Depression, Schizophrenia, Mental Retardation and Seizure were found to have the most numbers of cases. Figure 3 shows the distribution of disorders observed in a sample of 250 patients. Furthermore, interesting trends about these top Disorders are discovered, which are backed by the already performed researches in this domain

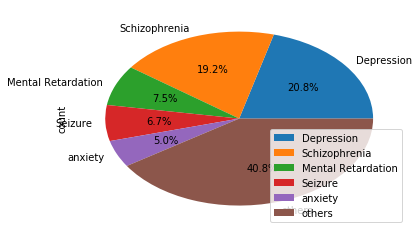


Figure 3: Pie chart showing the disorder spread.

Depression is the most commonly observed psychological Disorder. [9] Has also indicated that Depression is the leading cause of disability worldwide.

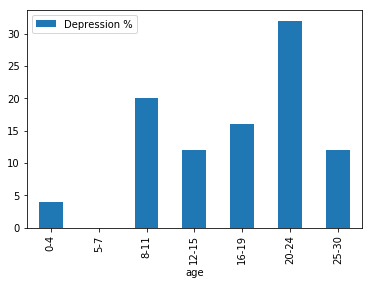
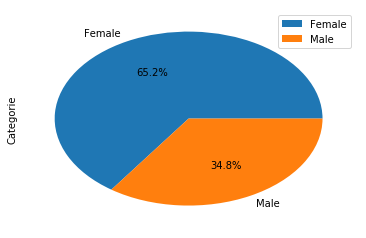


Figure 4: Depression vs. Age.

Figure4 indicates that depression is mostly observed in the 20-24 age group (Young adults) [10] [11].

Figure 5: Pie chart of male to female case of Depression.



Depression is more prevalent in females than in males. A study [12] has indicated that this could be possibly due to hormonal changes between the two genders.

Schizophrenia which is the second most common disorder in the dataset had a similar visualization to the Depression.

Figure 6: Pie chart of male to female case of Schizophrenia.

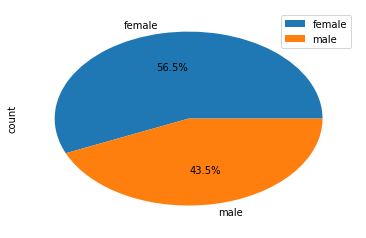
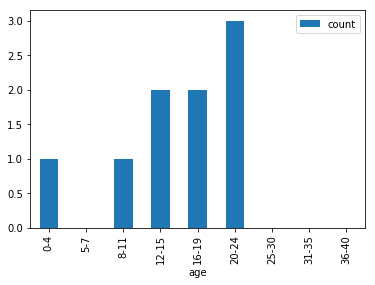


Figure 6 indicates that the prevalence of Schizophrenia is slightly more in females than males, although it is a very small difference as compared to Depression or Mental Retardation.

Mental Retardation begins in childhood or adolescence before the age of 18 and most commonly seen to persist throughout adult life as shown in figure7.

Figure 7: Mental Retardation in various age groups.



Mental Retardation is found to be observed more in male than in female’s fig 8.

Figure 8: Pie chart of male to female case of Mental Retardation**.**

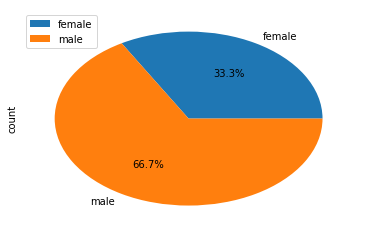


Figure 8, reveals that Mental Retardation is seen more commonly in males than females as suggested by a study in [13], it is assumed to be due to mutations on the X chromosome [14].

# *association rule mining*

Association is a data mining function that is used to determine the co-occurrence of two or more items in a dataset. These are expressed by Association rules and the co-occurrence of these items is objectively determined using certain metrics like Support and Confidence [15].

This is used in our study to determine the co-occurrence between various symptoms for a specific disorder or the co-occurrence of various symptoms that lead to a specific disorder. Kulczynski interestingness measure is used to determine the usefulness of the rules [16]. If Kulczynski is near 0 or 1, this can conclude that the association rule is interesting, it can be said that it is negatively or positively associated respectively. If Kulczynski score is close to 0.5, the respective rule is considered to be skewed and uninteresting with respect to the study. For a given rule, A->B,

Kulczynski =(P (A|B) +P (B|A)) (4)

Association can exist between two objects or more than two objects. A two object association rule is represented as X->Y, which is read as X implies Y or X leads to Y. Similarly, a three object association rule can be expressed as X->Y, Z or X, Y -> Z. Rules are mined with minimum support of 20% and minimum confidence of 60%.

1. Two symptom association rules for Schizophrenia

|  |  |  |  |
| --- | --- | --- | --- |
| **SL NO** | **Rule** | **Confidence** | **Kulczyn-ski score** |
| 1 | self-talking🡪 aggression | 0.87 | 0.78 |
| 2 | impulsive🡪 aggression | 1.00 | 0.65 |

1. Three symptom association rules for Schizophrenia

|  |  |  |  |
| --- | --- | --- | --- |
| **SL NO** | **Rule** | **Confidence** | **Kulczyn-ski score** |
| 1 | irrelevant behaviour🡪self-talking, suspiciousness | 0.7 | 0.75 |
| 2 | lazy, aggression 🡪 abusive language | 1.00 | 0.75 |

1. Two symptom association rules for depression

|  |  |  |  |
| --- | --- | --- | --- |
| **SL NO** | **Rule** | **Confidence** | **Kulczyn-ski score** |
| 1 | low appetite 🡪 panic attack | 1.00 | 0.90 |
| 2 | anxiety 🡪 panic attack | 1.00 | 0.65 |
| 3 | fear 🡪aggression | 1.00 | 0.66 |

1. Three symptom association rules for depression

|  |  |  |  |
| --- | --- | --- | --- |
| **SL NO** | **Rule** | **Confidence** | **Kulczyn-ski score** |
| 1 | low appetite🡪panic attack, sleeplessness | 1.00 | 1.00 |
| 2 | aggression, sleeplessness🡪fear | 1.00 | 1.00 |

# *predictive model*

The predictive model for our system to predict the mental disorder is trained based on the symptoms extracted from the manually collected patient details described in data description phase. Once the application goes to implementation phase we can train the model on the data being collected in the database. The reduced features were obtained after the elimination of highly correlated features as mentioned in the pre-processing section. Models are trained with the help of scikit-learn package [18] with parameters tuned using GridSearchCV and evaluated with the help of 5-fold cross-validation. The model is evaluated using precision, which represents the ratio of predictions that turn out to be positive in the group of records that were predicted to be positive by the model; precision is sometimes also referred to as the positive predictive value (PPV). The second metric used is recall. The recall is a measure of actual true positive rate, which is the ratio of correct instances among all instances that actually belong to the relevant class. The third metric is F1-score also called as f-measure. F-measure is the harmonic mean of precision and recall. An ideal model would yield a high F-measure of 1.

The first model used in our study is the KNeighborClassifier. It is a non-generalized learning model, sometimes also called an instance-based learning model. It does not try to develop an internal statistical mode like how it is done in the case of Linear Regression with a generalized cost function, but simply stores instances of the training data [19]. These instances of the data are assigned to a particular class using one of three algorithms: Brute Force, K-D tree, Ball-tree. The algorithms used in the study are the Brute force algorithm which utilizes Euclidean distance to calculate the distance between the point of interest and to all training set for n samples in d dimensions. Classification is computed from a simple majority vote of the nearest neighbours of each point: a query point is assigned the data class which has the most representatives within the nearest neighbours of the point.

https://hlab.stanford.edu/brian/euclid2.gif (5)

Euclidean distance between two points i and j in an n-dimensional Euclidean plane is given by the equation above. This model is an inherently multiclass classifier which suits the type of classification for our study. This model yielded a precision of 0.58, recall of 0.77 and F-1 of 0. 657. The second model considered in the study is the LogisticRegressionCV [20]. This model is an implementation of logistic regression using bilinear, newton-cg, and sag of lbfgs optimizer. Also, this model is cross-validated by default which can help tackle the problem of overfitting or underfitting. Underfitting is a particular problem for smaller datasets such as the one considered in this study. Also, Logistic Regression based models consider the entire feature vector during training. The model produced a recall of 0.90, precision of 0.73 and F-1 Score of 0.80. The third model used for the study is the GradientBoostingMode [17]. GB is a forward stage-wise learning model; one of GB’s greatest advantages is that this can use the arbitrary loss functions for each stage in the GB algorithm. The following equation depicts the typical loss function called squared error loss function:

H(x) = (6)

Here y is expected value and (x) is the predicted value h(x) is the loss which gradient descent will minimize. Here every stage will correct its predecessor hence the final h(x) loss will be reduced. In each stage, a select number of regression trees are fit on the negative gradient of the binomial or multinomial deviance loss function. The model produced a precision of 0.72, recall of 0.8 and F-1 Score of 0.757. LogisticRegressionCV seemed to be the best model obtained with the current amount of data available, and the same is used in the implementation of the system on the PythonAnywhere cloud which sends real-time prediction to the doctor.

# acknowledgment

This research was partially supported by Dr. Anand Pandurangi, Shanti Dhama, Dharwad. We thank the Medical Staff and Doctors from Shanti Dhama who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations or conclusions of this paper.

# references

1. Nimhans Bengaluru 2015-2016, “prevalence, patterns and outcomes”.
2. Ms. Sumathi M.R, Dr. B. Poorna, “prediction of mental health among children using machine learning” International Journal of Advanced Computer Science and applications, 2016.
3. Ghassan Azar, Clay Gloster,” Intelligent Data Mining and Machine Learning for Mental Health Diagnosis using Genetic Algorithm”,2014.
4. Luciano Comin Nunes, Placido Rogério Pinheiro” An expert system applied to the diagnosis of psychological disorders”, IEEE, 2009.
5. Sumathi M.R and Poorna B “Prediction of Mental Health Problems among Children using Fuzzy Clustered Bayesian Model”, International Journal of Control Theory and Applications.
6. C. Zhu, B. Li, A. Li, and T. Zhu, “Predicting Depression from Internet Behaviors by Time-Frequency Features,” 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), 2016.
7. Staff, "Multicollinearity", Investopedia, 2018. [Online]. Available: https://www.investopedia.com/terms/m/multicollinearity.asp. [Accessed: 15- Apr- 2018].Depression a Growing, Deadly Threat for Teens, Young Adults, March 2018.Nimhans Bengaluru 2015-2016, “prevalence, patterns and outcomes”.
8. R. O’brien, "A Caution Regarding Rules of Thumb for Variance Inflation Factors", Quality & Quantity, vol. 41, no. 5, pp. 673-690, 2007.Staff, "Multicollinearity", Investopedia, 2018. [Online]. Available: https://www.investopedia.com/terms/m/multicollinearity.asp. [Accessed: 15- Apr- 2018].Staff, "Multicollinearity", Investopedia, 2018. [Online]. Available: https://www.investopedia.com/terms/m/multicollinearity.asp. [Accessed: 15- Apr- 2018].
9. Adaa.org. (2018). Facts & Statistics | Anxiety and Depression Association of America, ADAA. [online] Available at: https://adaa.org/about-adaa/press-room/facts-statistics# [Accessed 14 Apr. 2018].
10. Child Trends. (2018). Young Adult Depression - Child Trends. [online] Available at: https://www.childtrends.org/indicators/young-adult-depression/, [Accessed 14 Apr. 2018].
11. Depression a Growing, Deadly Threat for Teens, Young Adults, March 2018.Nimhans Bengaluru 2015-2016, “prevalence, patterns and outcomes”.
12. Paul R. Albert “why is depression more prevalent in women”, J Psychiatry Neurosci. 2015 Jul.
13. X linked mental retardation: a clinical guide , MarchUniversity Science, 1989.
14. J. Pediatr. (Rio J.) - vol.80 issue2 suppl.", Scielo.br, 2018. [Online]. Available: http://www.scielo.br/scielo.php?pid=0021-755720040003&script=sci\_issuetoc. [Accessed: 19- Apr- 2018].
15. M. Hahsler, "A Probabilistic Comparison of Commonly Used Interest Measures for Association Rules", Michael.hahsler.net. [Online]. Available: http://michael.hahsler.net/research/association\_rules/measures.html. [Accessed: 17- Apr- 2018].
16. J. Han, M. Kamber and J. Pei, Data mining, 3rd ed. Amsterdam [etc.]: Morgan Kaufmann, an imprint of Elsevier, 2012, p. Chapter 6.
17. P. Miller, D. McArtor and G. Lubke, "A Gradient-Boosting Machine for Hierarchically Clustered Data", Multivariate Behavioral Research, vol. 52, no. 1, pp. 117-117, 2017.
18. G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa and A. Mueller, "Scikit-learn", GetMobile: Mobile Computing and Communications, vol. 19, no. 1, pp. 29-33, 2015.
19. "K-nearest neighbors algorithm", En.wikipedia.org, 2018. [Online]. Available: https://en.wikipedia.org/wiki/K-nearest\_neighbors\_algorithm. [Accessed: 17- Apr- 2018].Staff, "Multicollinearity", Investopedia, 2018. [Online]. Available: https://www.investopedia.com/terms/m/multicollinearity.asp. [Accessed: 15- Apr- 2018].
20. "3.2.4.1.5. sklearn.linear\_model.LogisticRegressionCV — scikit-learn 0.19.1 documentation", Scikit-learn.org, 2018. [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LogisticRegressionCV.html. [Accessed: 16- Apr- 2018].