IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 05-10

Exploratory Data Analysis of Autism Data

Dr. R. Uma Rani M.C.A., M.Phil., Ph.D, R. Suguna M.C.A., M.Phil
Associate Professor Of Computer Science, Sri Sarada College For Women (Autonomous),Salem-16.
Ph.D Research Scholar, Department Of Computer Science, Sri Sarada College For Women (Autonomous), Salem-16.

Abstract: *Data is growing every day. Now we are having lot of data from various e-components and social media, medical field and finance domains. Analysing the data gives us most useful information to face the future in various ways. The term data science is a broad place to provide a deep insight on the huge amount of data. Data science is gathering knowledge from data with the computational algorithms and statistical methods or mathematical models with effective visualization of data. This paper explores the comparison of various classification methods and various statistical models for the autism data.*

Keywords: *Data Analytics, Autism Spectrum Disorder, Classification, Statistical Models.*

I. Introduction

Data analytics is the process of analyzing the dataset to find full of meaning just round the corner of data using computational algorithms and statistical methods. Data analytics is used to explore and analyze data using statistical methods and models. Data science is a control which groups techniques and methods from various domains to study about data and data analytics is a part of data science. The data analysis is the process of breaking down a complex object into its simple forms and data analytics is the science of analysis whereby statistics, data mining, computer technology and mathematics can be used in doing analysis in particular dataset. In the earlier days the data analytics was carried out in the domains of finance, marketing, psychology, economics, social science for analyzing trends and future predictions. Statistical techniques along with mathematical models were used to understand and explore data effectively. The organizations understood the importance of data and treat the data as an asset like other properties. Data analytics is gaining deep momentum from data and different stages of analytics are as descriptive, diagnostics, predictive, prescriptive analytics. They provide more flexibility for users regarding data. The decisions are taken from prediction data. The overall performance of the data analytics is not about more data, it is about deeper look of data.

1.1 Data analytics vs. big data analytics

Data analytics is used to explore and analyze datasets using statistical methods or mathematical models. Big data analytics is used to analyze with the characteristics of volume, velocity and variety by integrating statistics, mathematics and computational algorithms in big data platform. The data under big data is classified as structured data which is in the form of rows and columns and data are stored in database ,then the unstructured data is which doesn’t contain any specific structure. Semi- structured data means data is having specific structure but varies with respect to particular domain such as emails and chat messages [1]. There are various types of data analytics methods of data analytics for the autism dataset as

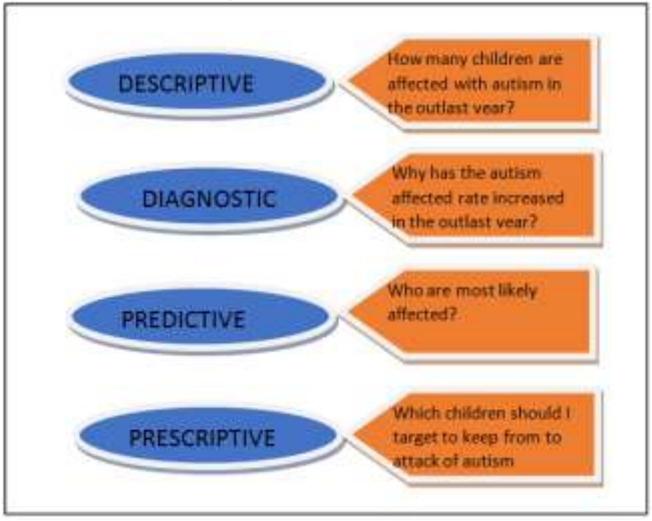
DESCRIPTIVE
How many children are affected with autism in the outlast year?
DIAGNOSTIC
Why has the autism affected rate increased in the outlast year?
PREDICTIVE
Who are most likely affected?
PRESCRIPTIVE
Which children should I target to keep from to attack of autism

Figure1. Types of Data Analytics

1.2 Data analytics in healthcare

Data analytics in healthcare provides evidential insights with a lower cost for providing right patient care management. Analytics used in various health care datasets with machine learning algorithms helps in improving patient health and prevent factors causing epidemics [1].

1.3 Autism Spectrum disorder

Autism disorder is one of the most sensitive problems in the society. The ASD is identified when the child is three years old. There are multiple symptoms such as difficulties to communicate with others, to understanding of other talk and low ability of studying, playing, smiling, loneliness also. Social interaction is hard for these children who are affected with the ASD. There are various types of Autism Disorder include

1. Asperger's Syndrome
2. Pervasive development disorder, not otherwise specified(PDD-NOS)
3. Autistic Disorder
4. Rett Syndrome
5. Childhood Disintegrative disorder

(i) Asperger's Syndrome

This is diagnosed in the age between five and nine and may found later than that age also. The symptoms of it are poor social interactions, lack of understanding of the others body language.

(ii) Pervasive development disorder, not otherwise specified(PDD-NOS)

This type of problem is sequence of disorder that takes in delays in development such as social interaction, communication, walking, talking and lack of ability to use their imagination.

(iii) Autistic Disorder

This type is also having the difficulty with communication and these children show actions to express their thoughts, then it will continue when they grow up. Their intelligence level is either below-average or above-average which is called as “high functioning”. The symptoms are as lack of eye contact, lack of modulation in their speech.

(iv) Rett Syndrome

This type of autism specially occurred in girls. It explicitly shows from the age of six months to end of their life. The effect of the rett syndrome varies from one child to another. The symptoms such as breathing problems, sleep problems, teeth grinding, slowness of growth, different way of walking and slowly they loss their abilities in each level of their age. Finally at age of 10 and above the physical decline may be very severe.

(v) Childhood Disintegrative disorder

Most of the children have normal development till 2 years. Then they slowly lose their abilities but it happens only few months. The behaviours such as anxiety, bladder control also occurred. The symptoms of this disorder are the children doing some activities repeatedly and move from one activity to another is so hard. Their self feeding also difficult for them and loss of the activities cannot recover [7].

II. Literature Review

Healthcare industries data are more complex and growing every day because of new diseases are discovering and new technologies are discovered. So data analytics tools provide support efficiency in healthcare data. There are various tools available for data analytics Advanced data visualization, presto, hive, vertica, key performance indicators (KPI), jaql, avro are which will improve the health care data [8].

The evaluation of behavioural factor associations and classified the behaviours using classification based association (CBA).The experiments used actual patient profiles from two hospitals in Thailand [2].By using multimedia intervention tool the social skills of the ASD children such as simulating the situation and understanding level were analyzed. There are 84 children belongs to NGOs and clinics selected for the study. The algorithms of classification such as J48, SVM, JRip, voted perceptron and multilayer preceptron were applied in the training and test data with the use of multimedia intervention tool [3].

Agriculture is the backbone of India .Soil is the boon for agriculture. The chemical fertilizer is the curse for soil. By using the food which is produced by the soil collaborates with chemical fertilizers will affect the human brain nerves easily, especially the children. This is one of the reasons for the autism. The study examined the agriculture has any impact on autism with the help of various classification algorithms [4]. Twitter is a highly popular media for information sharing. By using twitter as a data mining source and the tweets collected regarding autism spectrum disorder from the interested individuals. Content analysis and hash tag analysis, speech analysis was done with the various data mining techniques applied on the pre-processed tweets [5].

Collection of large corpus of medical literature related to autism spectrum disorder was analyzed with the various topic models. Topic model provides a statistical framework along with the digital archives of scholarly literature. Topic model enable us to learn the hidden distribution from a large archives of documents .The Bayesian non-parametric model was used after pre-processing [6]. The data of Autism was classified decision trees with classes as mild ,moderate, sivere.The machine learning algorithms with artificial neural networks, SVM and fuzzy logic was applied[7].

III. Method

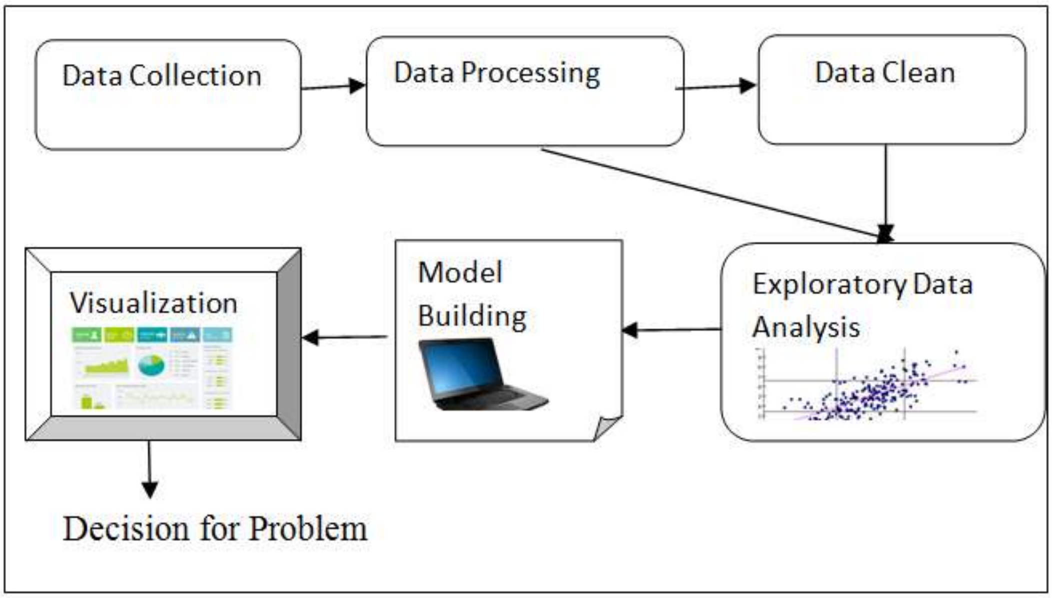

Figure 2: Data Analytics Process Model

3.1 Data Collection

The dataset is available in various websites which are collected and processed earlier. The autism data set collected from UCI repository.

3.2 Data Processing

Pre-processing is the process of handling the dataset with the missing values and transformation of data and dimensionality reduction in the dataset is called as cleaning the data. The dimensionality reduction for the autism provides the high correlation among the variables using PCA [8]. There are various tools for Processing and preparing the data .

- Data Wrangler- is an effective tool for data cleaning and transformation and it was developed by stanford university.
- D3.js (Data-Driven Documents) - was developed by Mike Bostock. D3 is a JavaScript library for visualizing data and manipulating the document object model that runs in a browser without a plugin.
- OpenRefine – Open source tool for working with messy data,It is GUI tool and popular for transformations,

3.3 EDA: Exploratory Data analysis

The main work of EDA is to understand the relationships among the variables to notify selection of the variables and methods. After cleaning the data, it needs to be modelling [1].A particular mathematical or statistical technique can be selected then applied on data. The various statistical models and mathematical models are available for apply on the data to get meaningful knowledge for the problem [11]. Data analytics is carried out using statistical methods and computational algorithms. Some data exploration methods take place in preparation of data only. Here it concentrates on the high quality of data [1]. The general data analytics methods are,

Statistical Methods	Computational Algorithms
Hypothesis Testing	Probabilistic Models
Covariance	Rule Based Learners
Correlation	Classifiers
T-Test,F-Test,pValue	Decision Trees
False Discovery Rate	Neural Networks
Bayesian Model	Association Rule mining
	Clustering
	Regression Analysis
	NLP

Table1. Data Analytics methods

3.4 Model Building

The analytical model should be developed based on dataset. The data should be dividing into test data and training data .The test data can used to build the model. The training dataset is used for conducting the initial experiments. The test sets is used for validation are arrived at once the initial experiments and models have been run. Some tools for model building are R studio, Octave, WEKA and Python as the open source tools. In this paper the comparison made with various classification algorithms like naïve bayes, regression, logistic regression, C4.5, random forest and j48, KNN [11].

3.5 Statistical Methods

Hypothesis testing is the statistical methods which also called as confirmatory data analysis to do the statistical decisions according to the data. From the autism data the variables age and result of autism test for the children were used to do the hypothesis testing .Based on the data, the null hypothesis and alternate hypothesis are applied. So, Paired T-test applied in the data. False discovery rate will be calculated when multiple comparisons is applied on data.

IV. Result And Discussion

The Classification algorithms are compared for the autism dataset. The table below shows the algorithms accuracy on autism dataset. The KNN algorithm gives better result compare with other algorithms such as Naïve Bayes, Logistic Regression and J48.

Algorithm	Accuracy level
Naïve Bayes	76.98%
Logistic Regression	87.07%
Random Forest	90%
J48	87%
K-NN	95%
C4.5	93%

Table 2. Comparison of classification algorithms

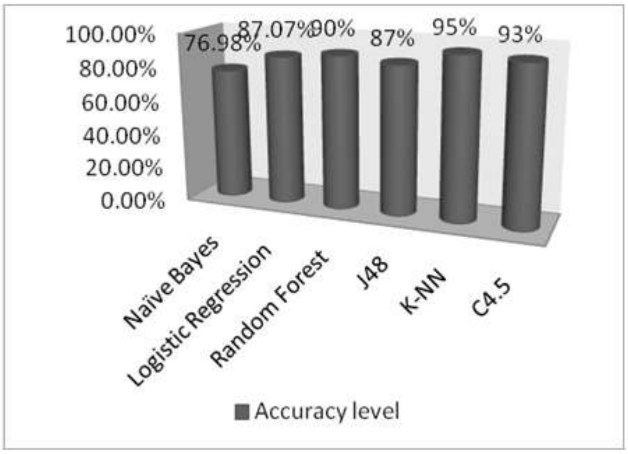

Figure 3: performance of classification algorithms

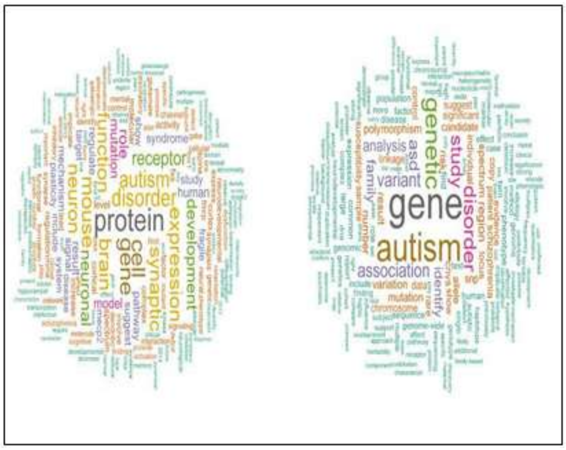

Figure 4: word wrap for autism dataset

The above Figure-3 shows that the result of comparison of classification algorithms.KNN gives better classification among the other algorithms. Figure-4 shows that the result of frequently used words about autism which was collected from the journals of autism in the year 2016 and 2017.

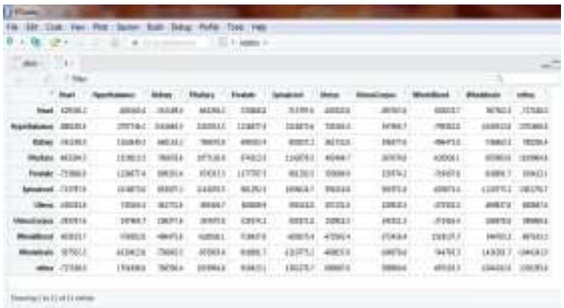

Figure 6: Correlation matrix of autism data

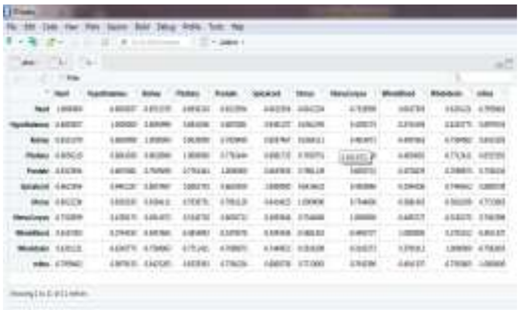

Figure 6: Covariance matrix of autism data

The above figure shows that the correlation among the variables of autism data. The covariance matrix shows that combinations of similar variables. The figure-7 shows the result of T-test. The hypothesis testing among the screening test result is depending on the age of autism children. Alternate hypothesis is used in the data. Then the Paired T-test and wilcoxon rank test calculated among the variables.

Figure 7: Hypothesis testing on age and result

V. Conclusion

Autism is a big social relevant problem which can affect anyone. It is not brain disease it is the effect of damaged neuron cell of brain. In this paper various types of classification algorithms were compared and statistical methods are used which are useful in data analytics. The diagnostics analytics were used on the autism data. Statistical methods such as hypothesis testing and correlation were applied on the data.

VI. Future Scope

In future we can explore the each analytics methods with autism data for fruitful decision making regarding autism problem. By this study the age of the children not a factor for autism is assumed based on the statistical methods. For the future scope the other analytical methods such as prescriptive a, descriptive, predictive methods can be applied on the data.

References

[1]. V.Bhuvaneswari, T.Devi,"Big data analytics-A Practioner's Approach",2016.

[2]. Dr.M.Manimekalai, A.E.Arthipriya,"Evaluating the behavioural and developmental interventions for autism spectrum disorder", International journal of information science and application, 2014.

[3]. Richa misra, Divya bhatnagar,"Analysing the social awareness in autistic children trained through multimedia intervention tool using data mining",International journal of Advanced computer science and applications,2017.

[4]. Dr.Yamini, M.Premasundari,"A Review on classification technique with autism spectrum disorder and agriculture", International journal of advanced research in computer science",2017.

[5]. Adham Beyikhoshk," Data mining and Autism Spectrum Disorder: A pilot study",International Conference on Advances in Social Networks Analysis and Mining, 2014.

[6]. Adham Beykikhoshk ; Dinh Phung ; , Analysing the History of Autism Spectrum Disorder using Topic Models, IEEE International Conference on Data Science and Advanced Analytics ,2016.

[7]. M.S.Mythili,A.R.Mohammed shanavas,"A Study on Autism spectrum disorders using classification techniques",2014.

[8]. Dr. R.Umarani, R.Suguna"A study of autism spectrum disorder using principal component analysis and fuzzy cmeans clustering",

[9]. G.Leory,A.Irmscher and M.H.Charlop-christy,"Datamining Techniques to study therapy success with autistic children",2006 International conference on datamining,26-29,june 2006,monte carlo resort, as vegas,USA.

[10]. Centers for Disease Control and Preventionhttp://www.cdc.gov/ncbddd/autism Eunice Kennedy Shriver National Institute**.**

[11]. EMC Educational Services, "Data Science and Big Data Analytics" ,Wiley ,India.

[12]. Autism Speaks 100 days toolkit,http://www.cdc.gov/ncbddd/autism.

[13]. Daniel Bone, Matthew S. Goodwin,"Applying Machine Learning to Facilitate Autism DiagnosticsPitfalls and promises"," Journal of Autism and Developmental Disorders", 2014 .

[14]. A Survey of Parents with Children on the Autism Spectrum: Experience with Services and Treatments, Tracy A Becerra, PhD, OTR/L; Maria L Massolo, PhD; Vincent M Yau, PhD; Ashli A Owen-Smith, PhD; Frances L Lynch, PhD; Phillip M Crawford, MS; Kathryn A Pearson; Magdalena E Pomichowski, MPH;