Business Intelligence Project

**Detecting Autism**

Submitted by

**TEAM 11**

Aakanksha kumari - 03204092018

Himanshi gupta- 04104092018

Priyanka- 02104092018

Kanika Kashyap- 03404092017

Nidhi Gangwar- 04504092017

Sonia-01004092018

Under the Supervision of

**Mr. Rishabh Kaushal**

Assistant Professor

Department of Information Technology



Department of Information Technology

Indira Gandhi Delhi Technical University for Women

Kashmere Gate, Delhi - 110006

March 2020

Contents

**1 Introduction**

* 1. Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .1

1.1 Objective . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .1

**2 Literature Survey**

2.1 Research questions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .2

**3 Proposed Methodology**

3.1 Dataset . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

3.2 Attribute Description . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

3.3 Details about the Organization . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

3.4 Data Pre-processing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .5

3.4.1 Data Files Combining . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

3.4.2 Data Cleaning . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

3.5 Feature Computation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

**4 Data Exploration**

4.1 Visualizations . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

**5 Algorithms**

5.1 Random Forest Cart . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .12

5.1.1 Description . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

5.1.2 Graphical Representation of the Results . . . . . . . . . . . . . . . . 13

5.1.3 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14

5.2 Decision Tree Cart . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .15

5.2.1 Description . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

5.2.2 Graphical Representation of the Results . . . . . . . . . . . . . . . . 15

5.2.3 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17

5.3 Merge Random Forest Cart with ID3 algorithm . . . . . . . . . . . . . . . . .18

5.3.1 Description . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18

5.3.2 Graphical Representation of The Results . . . . . . . . . . . . . . . . . .18

5.3.3 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20

5.5 Comparison of Implemented Algorithms . . . . . . . . . . . . . . . . . . . . 25

**6 Future Work**

**Chapter 1**

**Introduction**

Autism spectrum disorder is a neurodevelopmental disorder that affects a person’s interaction, communication and learning skills. Although diagnosis of autism can be done at any age, its symptoms generally appear in the first two years of life and develops through time [1]. Autism patients face different types of challenges such as difficulties with concentration, learning disabilities, mental health problems such as anxiety, depression etc, motor difficulties, sensory problems and many others. Current explosion rate of autism around the world is numerous and it is increasing at a very high rate. According to WHO [2], about 1 out of every 160 children has ASD. Some people with this disorder can live independently, while others require life-long care and support. Diagnosis of autism requires significant amount of time and cost. Earlier detection of autism can come to a great help by prescribing patients with proper medication at an early stage. It can prevent the patient’s condition from deteriorating further and would help to reduce long term costs associated with delayed diagnosis. Thus a time efficient, accurate and easy screening test tool is very much required which would predict autism traits in an individual and identify whether or not they require comprehensive autism assessment.

**1.1 Problem Statement**

-> Autism spectrum disorder impacts the nervous system and affects the overall cognitive, emotional, social and physical health of the affected individual.

->Number of questions are asked from a person and on the basis of score he gained we predict if the person has autism.

**1.1.1 Objective**

The objective of this work is to propose an autism prediction model using ML techniques and to develop a mobile application that could effectively predict autism traits of an individual of any age. In other words, this work focuses on developing an autism screening application for predicting the ASD traits among people of age 18 and more. The rest of the paper is organized as follows. Section II discusses the related researches previously done in this area. Section III presents the research methodology. Detailed implementation of the proposed system is discussed in Section IV and the implemented system is evaluated in Section V. Section VI brief discusses how the proposed algorithm was merged into a mobile application. Finally, Section VII concludes the paper by highlighting the research contributions, research limitations and future plans to extend this work.

**Chapter 2**

**Literature Survey**

**2.1 Research questions**

1. Is it hard for you to communicate?

2. Does this mean you don’t get irony or sarcasm?

3. Is it hard for you to make eye-contact?Why?

4. Can you read facial expression?

5.So do you ever offend anyone?

6. what about relationships? Are they hard?

7.Do you get sensory overload? How does it feel like?

8. Do you need to stick to a routine?

9.Are you often forced to overthink things?

10. Do you like having autism?

**Chapter 3**

**Proposed Methodology**

**3.1 Dataset**

* Number of instances:704
* Number of attributes:21
* Whether labeled or unlabeled: labelled
* Type of label information (if present): categorical, binary, numerical

|  |  |
| --- | --- |
| ATTRIBUTE NAME | ATTRIBUTE TYPE |
| A1\_Score | Binary |
| A2\_Score | Binary |
| A3\_Score | Binary |
| A4\_Score | Binary |
| A5\_Score | Binary |
| A6\_Score | Binary |
| A7\_Score | Binary |
| A8\_Score | Binary |
| A9\_Score | Binary |
| A10\_Score | Binary |
| Age | Numeric |
| Gender | Categorical |
| Ethnicity | Categorical |
| Jundice | Categorical |
| Austim | Categorical |
| contry\_of\_res | Categorical |
| used\_app\_before | Categorical |
| result | Categorical |
| age\_desc | Categorical |
| relation | Categorical |
| Class/ASD | Categorical |

**3.2 Attribute Description**

* **A1\_Score -** Question 1 Answer Binary (0, 1) The answer code of the question based on the screening method used
* **A2\_Score -** Question 2 Answer Binary (0, 1) The answer code of the question based on the screening method used
* **A3\_Score -** Question 3 Answer Binary (0, 1) The answer code of the question based on the screening method used
* **A4\_Score -** Question 4 Answer Binary (0, 1) The answer code of the question based on the screening method used
* **A5\_Score -** Question 5 Answer Binary (0, 1) The answer code of the question based on the screening method used
* **A6\_Score -** Question 6 Answer Binary (0, 1) The answer code of the question based on the screening method used
* **A7\_Score -** Question 7 Answer Binary (0, 1) The answer code of the question based on the screening method used
* **A8\_Score -** Question 8 Answer Binary (0, 1) The answer code of the question based on the screening method used
* **A9\_Score -** Question 9 Answer Binary (0, 1) The answer code of the question based on the screening method used
* **A10\_Score -** Question 10  Answer Binary (0, 1) The answer code of the question based on the screening method used
* **Age -** Age of the person
* **Gender -** Gender  of the Person
* **Ethnicity -** String List of common ethnicities
* **Jundice -** (yes or no) Whether the case was born with jaundice
* **Austim -** Whether autism patient or not
* **country\_of\_res -** Name of the country
* **used\_app\_before -** Whether application was used before or not
* **result -** The final score obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner.
* **age\_desc -** Belongs to which group of age
* **relation -** Who is completing the test String Parent, self, caregiver, medical staff, clinician ,etc.
* **Class/ASD -** Belongs to which class YES or NO

**3.3 Data Pre-processing**

**3.3.1 Data Collection**

To develop an effective predictive model, AQ-10 dataset was used which consists of three different datasets based on AQ-10 screening tool questions. This dataset contain data of ages of 18 or more (adult). AQ-10 or Autism Spectrum Quotient tool is used to identify whether an individual should be referred for a comprehensive autism assessment. AQ-10 screening questions focus on different domains such as- attention to detail, attention switching, communication, imagination and social interaction. Scoring method of the questions is that only 1 point can be scored for each of the 10 questions. User may score 0 or 1 point on each question based on their answer. Dataset of adult contain 704 instances. Dataset contains twenty-one attributes which are a mix of numerical and categorical data that includes: Age, Gender, Ethnicity, If born with Jaundice, Family member with PDD, Country of Residence, Used the screening app before, Screening method type, Question 1-10, Result and Class.

**3.3.2 Data Cleaning**

The collected data were cleaning to remove irrelevant features. For example, the ID column was irreverent to develop a prediction model, thus it was removed. To handle null values, listwise deletion technique was applied where a particular observation was deleted if it had one or more missing values. Then to extract unnecessary features from the dataset, decision tree algorithm was used. Results showed dropping ‘relation’, ‘age desc’, ‘used app before’ and ‘age’ columns would result in more accurate classification and so those columns were dropped.

Chapter 4

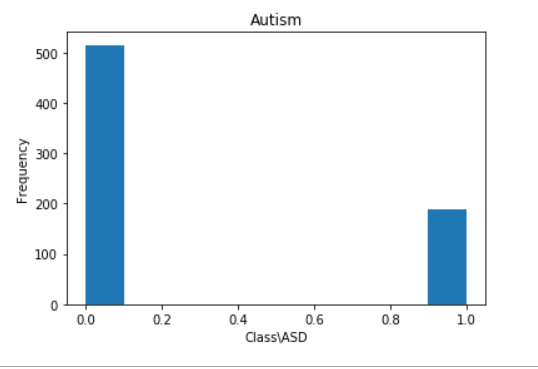
Data Exploration

**4.1 Metrics**

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | Mean | Median | Mode |
| A1\_Score | 0.721591 | 1.0 | 0 |
| A2\_Score | 0.453125 | 0.0 | 1 |
| A3\_Score | 0.457386 | 0.0 | 0 |
| A4\_Score | 0.495739 | 0.0 | 0 |
| A5\_Score | 0.498580 | 0.0 | 0 |
| A6\_Score | 0.284091 | 0.0 | 0 |
| A7\_Score | 0.417614 | 0.0 | 0 |
| A8\_Score | 0.649148 | 1.0 | 1 |
| A9\_Score | 0.323864 | 0.0 | 0 |
| A10\_Score | 0.573864 | 1.0 | 1 |
| result | 4.875000 | 4.0 | 4 |

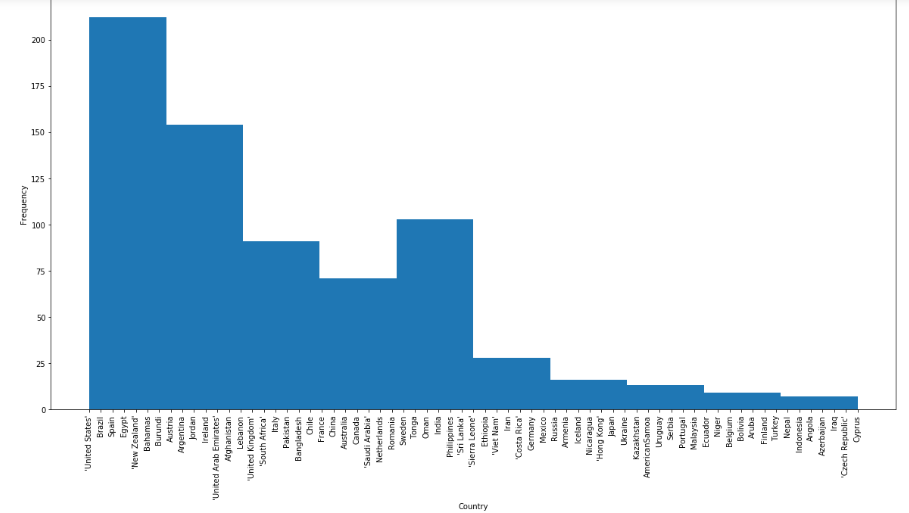
**4.2 Visualizations**

**4.2.1 Number of patients**



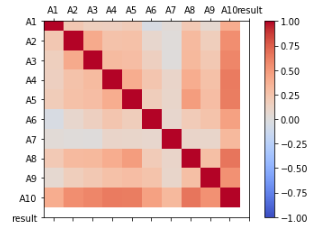
Graph shows the line plot between the Class/ASD on X-Axis and Frequency on Y-axis. On x-axis 0 represent who does not have autism and 1 represent who have autism.

**4.2.2 Number of country**



Graph shows the line plot between the Country on X-Axis and Frequency on Y-axis. On x-axis depict the various country answered the questions.

**4.2.3 Correlation matrix**



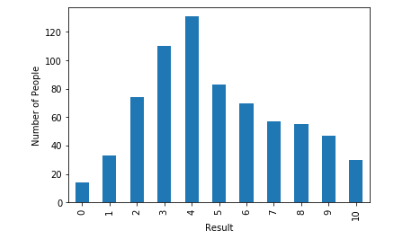
Graph shows correlation between questions and result.

**4.2.4 Jaundice v/s Autism**



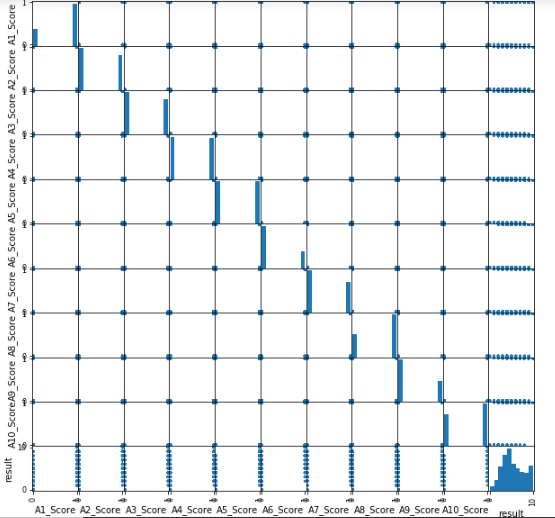
Graph shows the scatter plot between jaundice patients on Y-axis and Class/ASD on X-axis .

**4.2.5 Score of patient to detect autism**



Graph shows bar graph between number of people on Y-axis and result i.e. score on the basis of questions on X -axis.

**4.2.6 Scatter matrix**



Graph shows scatter matrix between colmuns of dataset.Diagonals represents yes and no of each attribute from A1\_score to A2 score and last diagonal which is result shows the overall score of A1 to A10.

Chapter 5

Algorithms

**5.1 Random forest Cart**

**5.1.1 Description**

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

**5.2 Decision Tree Cart**

**5.2.1 Description**

A **decision tree** is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

A decision tree is a flowchart-like structure in which each internal node represents a “test” on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

Tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods. Tree based methods empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map non-linear relationships quite well. They are adaptable at solving any kind of problem at hand (classification or regression). Decision Tree algorithms are referred to as **CART** **(Classification and Regression Trees)**.

**5.3 Merge Random forest cart with ID3 Algorithm**

**5.3.1 Description of ID3 Algorithm**

In order to improve the performance, a prediction model is proposed that merged the concept of random forest- CART with the concept of random forest - ID3 [Algorithm 3]. The algorithm for the proposed prediction model can be split into two phases like before: generating the merged random forest and classifying test data. Difference of it from Algorithm 2 is that here randomness is increased more by generating and adding ID3 decision trees to the random forest. Algorithm 3 tends to work better than Algorithm 2 because addition of ID3 decision trees limits overfitting and thus further reduces error compared to Algorithm 2.

**Chapter 6**

Future Work

Our future work will focus to collect more data from

various sources and to improve the proposed machine learning

classiﬁer to enhance its accuracy. A user study will also be

conducted to evaluate the usability and user experience (UX)

Our future work will focus to collect more data from various sources and to improve the proposed machine learning classiﬁer to enhance its accuracy. A user study will also be conducted to evaluate the usability and user experience (UX) of the mobile application.

The proposed Merged Random Forest algorithm was in-

tegrated in a screening android application with the help of

The proposed Merged Random Forest algorithm will be integrated in a screening android application with the help of Amazon Web Service (AWS). Using AWS, an API will be created to call from the android app. Home screen and a transition from home screen of the application is showed in Figure below. The application will be divided for 3 different age groups. Different questions will be used for different age groups based on the three AQ-10 screening tool versions. Based on the answer of all the questions the application showed whether or not the user has autism traits.

