

EXECUTIVE SUMMARY

Autism is a medical condition that demonstrates the social handicap behavior or incapability to respond to the stimulus as per the normal peers. The research paper highlights comparison between Random Forest Model Logistic Regression Model on the autism screening adult dataset. So, the research helps one understand the accuracy of the data with the people suffering from autism. The paper concludes with the statistical data stating that Logistic Regression Model predicting accurate result than Random Forest Model.

By Amol Shimpi

170002620

Lecturer

Dr. Fadi Fayez

Table of Contents

1. INTRODUCTION	3
2. PROBLEM STATEMENT.....	3
3. DATA PRE-PROCESSING	3
3.1 As per application in the Adult section (17 years and older),	4
3.2 As per application in the Adolescent (12-16 years),	4
3.3 As per application in the Child (4-11 years),	4
3.4 As per application in the Toddler section (12-36 months)	5
4. DATA CLEANSING	5
5. DATA UNDERSTANDING.....	6
5.1 17 years to elder attribute dataset analysis:	6
5.2 12-16 years attribute dataset analysis:	8
5.3 4-11 years attribute dataset analysis:	10
6. DISCUSSION ABOUT METHODS:	11
6.1 Logistic Regression Model	11
6.2 Random Forest Model.....	12
7. ANALYSIS.....	12
7.1 4-11 years age group for Random Forest Model:	12
7.2 12-16-year age group for Random Forest Model:	13
7.3 17 years and older age group for Random Forest Model:	14
7.4 4-11 years age group for Logistic Regression Model:	15
7.5 12-16-year age group for Logistic Regression Model:.....	15
7.6 17 years and older age group for Logistic Regression Model:	16
8. DISCUSSION ABOUT ANALYSIS	17
9. CONCLUSION	17
10. REFERENCES.....	17

1. INTRODUCTION

There are many analysis techniques available for building a predictive model. Random Forest Model and Logistic Regression Model are the most popular techniques. By comparing the accuracy of both models, we want to know which model is providing the best accuracy for ASD dataset. We are dealing with autism and we need to know the best accuracy with the least error rate model, for that we need to understand ASD (Autism spectrum Disorder).

Autism spectrum Disorder, popularly known as ASD, is used to explain Autism and other disorders that hamper the communication and one's ability to socialize in the external environment (Simon Baron-Cohen, 2001). Whereas the data can be derived in four age group and each age group will show the autism as per their perspective on the bases on some questions which have been asked. It can also be termed as a set of disabilities affecting the social interaction accompanied by handicapped behavioral symptoms. The symptoms are visible before the age of 3 years approximately (Duggar, 2012). In the application, there are options for toddlers or 4-11 years age group that will help them out to recognized if they having ASD or autism symptoms. There are so many ways to detect the autism where there is screening tool named as M-CHAT-R (Modified Checklist for Autism in Toddlers, Revised) which is for toddlers who are in-between 16-30 months of age to assess their risk for autism spectrum disorder(ASD). This tool is basically for the parents who are concerned about their children's if they having autism or not then this tool will help them out with recognized autism. It will be detected by the crying patterns of the new-born baby.

Autism is a medical condition demonstrating the abnormal behavior in the external environment and communication impairment exhibited by the repetitive actions and imagination. Autism is also a spectrum disorder as the children can be affected by a variety of symptoms, be it mild to severe. So, this symptom can be detected by this given application so-called ASD-Quiz, whereas there are 10 different types of questions in the application that might help to detect the autism as per age group. It is a complicated medical condition in which the needs of the individual are varied (Edelson, 2018). As per the research surveys conducted in the United States, in every 10000 live births, there has been detection 4-5 autistic babies (Olson, 2010).

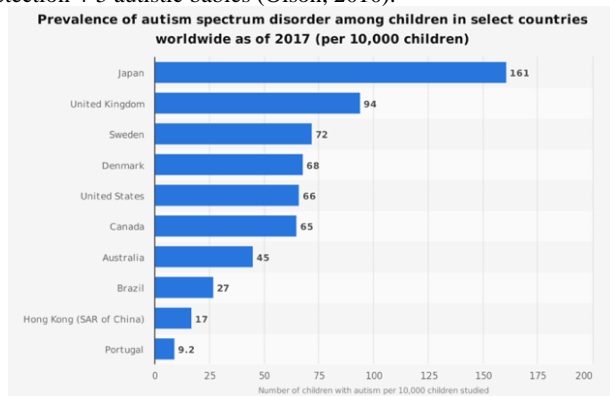


Fig 1. ASD worldwide as of 2017 (www.statista.com, 2017)

As you can see in the above image this statistic shows the occurrence of ASD among children in selected countries as of 2017

as per 10000 children. As of this time, Japan was assessed to have the highest rate of ASD among with around 161 adolescents per 10000 with this disorder. The research stated that autism affects males more as compared to the females. There are innumerable infants affected by autism by birth. As per the study done by APA (American Psychiatric association), Roots of Autism have been genetic. Major characteristics that an autistic baby would exhibit are the becoming limp, try to get away from the physical contact of the person, don't understand being picked up.

Autism Genetic Landscape

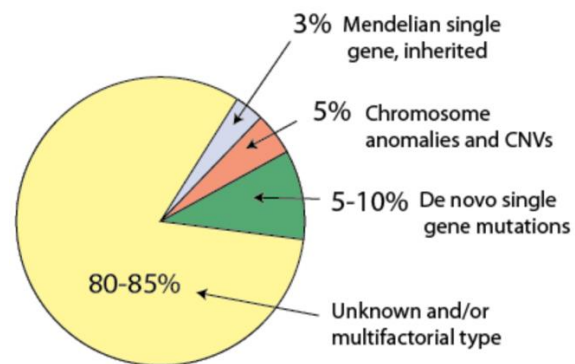


Fig 2. Autism genetic chart (The Genetics of Autism, 2012)

In the stage of autism, the child lacks in communication skills in comparison to the others of the same age. This involves repetitive behavior of hand flapping, hand biting. They also face other issues such as poor eating habits, eye contact and unable to feel the pain after being hit by rock, unable to reciprocate the feelings like a normal adult. In order to control autism, there have been various methods used by the scientists – traditional and non-traditional. Some children and adults suffering from autism are given medications too. However, until date there has been no permanent medication for autism. Ritalin is most commonly used medicine for treating Autism. However, there is no proven effectiveness of this medicine. Doctors use ABA (applied behavior analysis) and vitamin B6 combined with magnesium for the treatment (Baio, 2012). The symptoms of autism can however be treated if recognized in the early stage. It can be treated using the medical treatment and medications named as Nystatin (Edelson, 2018). There are different ways to treat their other impairments such as sensitive hearing and vision.

To conclude, Autism is a disorder faced by numerous individuals who need attention and support to overcome the challenges. After years of research, doctors and scientists have developed various methods to treat the disorder. However, it is difficult to completely recover from it, but it is possible to reduce the impact of this disorder on the individuals suffering from it.

2. PROBLEM STATEMENT

To identify the difference between a predictive model for Random Forest Model and Logistic Regression Model to get the best accurate model in the autism by given ASD dataset.

3. DATA PRE-PROCESSING

Data pre-processing is data mining technique that includes changing raw data into a clear format. Actual data is

sometimes incomplete, unpredictable in particular performance and is probably occurs many errors. The data pre-processing is a technique for solving such problems. For this we have got this raw Autism screening adult dataset from MIT (Manukau Institution of Technology) canvas of ADA (advanced data analytics) which is given by the lecturer. It contains 2497 instances and 45 various attributes; after that, the given ASD dataset file is an ODS file format and which have been converted to CSV file format for analyzing purpose because analytical tools are not accepting ODS file. As per the given dataset of ASD (Autism Screening Adult) its having 4 age group as per that age group it's asking the different questions to every age group having their own possible answers. As per the screening method type which is in integer value in form of (0,1,2,3) where,

- 0 stands for toddler (12-36 months),
- 1 is for child (4-11 years),
- 2 is for adolescent (12-16 years) and
- 3 is for adult (17 years and older).

3.1 As per application in the Adult section (17 years and older),

These are the following questions for the 17 years and elder,

1. **I often notice small sounds when others do not**
2. **I usually concentrate more on the whole picture, rather than the small details**
3. **I find it easy to do more than one thing at once**
4. **If there is an interruption, I can switch back to what I was doing very quickly**
5. **I find it easy to read between the lines when someone is talking to me**
6. **I know how to tell if someone listening to me is getting bored**
7. **I like to collect information about categories of things (e.g. types of car, types of bird, types of train, types of plant etc.)**
8. **When I'm reading a story, I find it difficult to work out the characters' intentions**
9. **I like to collect information about categories of things (e.g. types of car, types of bird, types of train, types of plant etc.)**
10. **I find it difficult to work out people's intentions.**

Whereas there are 4 possible answers which are,

- Definitely agree
- Definitely disagree
- Slightly agree
- Slightly disagree

In Q1, Q8, Q9 and Q10, if answers are Definitely agree and Slightly agree then answer code is 1 and rest of them will be 0.

In Q2, Q3, Q4, Q5, Q6 and Q7 are, then Definitely or Slightly Disagree then answer code is 1 and rest of them will be 0. If the individual score is more than 6 out of 10 consider referring them referring for a diagnostic assessment.

3.2 As per application in the Adolescent (12-16 years),

These are the following questions for the 12-16 years,

1. **S/he notices patterns in things all the time**
2. **S/he usually concentrates more on the whole picture, rather than the small details**

3. **In a social group, s/he can easily keep track of several different people's conversations**
4. **If there is an interruption, s/he can switch back to what s/he was doing very quickly**
5. **S/he is good at social chit-chat.**
6. **When s/he was younger, s/he used to enjoy playing games involving pretending with other children**
7. **S/he finds social situations easy**
8. **S/he frequently finds that s/he doesn't know how to keep a conversation going**
9. **S/he finds it difficult to imagine what it would be like to be someone else**
10. **S/he finds it hard to make new friends**

Whereas there are 4 possible answers which are,

- Definitely agree
- Definitely disagree
- Slightly agree
- Slightly disagree

In Q1, Q8, Q9 and Q10, then Definitely agree and Slightly agree then answer code is 1 and rest of them will be 0.

In Q2, Q3, Q4, Q5, Q6 and Q7 then Definitely or Slightly Disagree then answer code is 1 and rest of them will be 0. If the individual score is more than 6 out of 10 consider referring them referring for a diagnostic assessment.

3.3 As per application in the Child (4-11 years),

These are the following questions for the 4-11 years,

1. **S/he often notices small sounds when others do not**
2. **S/he usually concentrates more on the whole picture, rather than the small details**
3. **In a social group, s/he can easily keep track of several different people's conversations**
4. **S/he finds it easy to go back and forth between different activities**
5. **S/he is good at social chit-chat.** (This same question is in Adolescent (12-16 years) group which giving same answer code.)
6. **When s/he was in preschool, s/he used to enjoy playing games involving pretending with other children**
7. **S/he finds it easy to work out what someone is thinking or feeling just by looking at their face**
8. **S/he doesn't know how to keep a conversation going with his/her peers**
9. **When s/he is read a story, s/he finds it difficult to work out the character's intentions or feelings**
10. **S/he finds it hard to make new friends**

Whereas there are 4 possible answers which are,

- Definitely agree
- Definitely disagree
- Slightly agree
- Slightly disagree

In Q1, Q8, Q9 and Q10, if the Definitely agree and Slightly agree comes then answer code is 1 and rest of them will be 0.

In Q2, Q3, Q4, Q5, Q6 and Q7, then Definitely or Slightly Disagree then answer code is 1 and rest of them will be 0. If the individual score is more than 6 out of 10 consider referring them referring for a diagnostic assessment.

3.4 As per application in the Toddler section (12-36 months)

These are the following questions for the 12-36 months,

1. **Does your child look at you when you call his/her name?**
2. **If you or someone else in the family is visibly upset, does your child show signs of wanting to comfort them? (e.g. stroking hair, hugging them)**
3. **How easy is it for you to get eye contact with your child?**
4. **Does your child point to indicate that s/he wants something? (e.g. a toy that is out of reach)**
5. **Does your child point to share interest with you? (e.g. pointing at an interesting sight)**
6. **Does your child pretend? (e.g. care for dolls, talk on a toy phone) and**
7. **Does your child follow where you're looking?**
8. **Does your child use simple gestures? (e.g. wave goodbye)**
9. **Does your child stare at nothing with no apparent purpose?**
10. **Would you describe your child's first words as?**

Whereas there are 5 possible answers which are,

- Always
- Never
- Rarely
- Sometimes
- Usually

If Q1, Q2 then the answers are Never, Rarely, sometimes then answer code is 1 rest of them will be 0.

Whereas there are 5 possible answers which are,

- Impossible
- Quite difficult
- Quite easy
- Very difficult
- Very easy

In Q3 If the answers are Impossible, quite difficult and Very difficult then answer code is 1 rest of them will be 0.

Whereas there are 5 possible answers which are,

- A few time a day
- A few time a week
- Less than once a week
- Many times, a day
- Never

In Q4, Q5, Q6, Q7, Q8 and Q9 if the answers are A few times a week, less than once a week and never then answer code is 1 rest of them will be 0.

Whereas there are 5 possible answers which are,

- My child doesn't speak
- Quite typical
- Slightly unusual
- Very typical
- Very unusual

In Q10, if the answers are My child doesn't speak, slightly unusual and very unusual then answer code is 1 rest of them will be 0. If the individual score is more than 6 out of 10 consider referring them for a diagnostic assessment.

In this every screening test type having different question for all age group and also having different possible answers. So, we cannot consider in the one dataset for that we need to create four different datasets. Attribute name screening test type in that 18-36 months age group is deleted just because it's having toddler's data and all having ASD no as result. So, the toddlers having total 541 instances of data that has been removed.

4. DATA CLEANSING

Data cleansing which is also known as data cleaning where it's a process of detecting and removing corrupt or improper data from the dataset. It refers to identify the incomplete, incorrect, inaccurate parts of the data and then replacing, updating or removing uncleaned data. The data has been to be cleaned before the data can be used for our machine learning process. After cleansing there is another task to do is noise elimination. Noise is an irrelevant or meaningless data. It's mainly focused on the detection and removal of noise which is in form of low-level data errors that's the output of an improper data gathering process. Whereas this kind of data noise which is detrimental from any kind of data analysis. (Xiong)

So, at the start of the paper, we have already discussed the raw data which is completely messy or with lots of error or missing values. With the help of some methods and techniques that we have studies because some of the data were cleaned by logic and some of them filled by using RapidMiner, Excel and cleaned the given raw data into cleansed data by filling out the missing values. After observing the data, we have done some changes in the data. Firstly, case number is removed because this is only reference number or a primary key which is not an attribute. We have changed the instances of residency from Europe to England and Latin America to Argentina. Because Europe is not a country it is a continent and major country of Europe England. And Latin America is a combination of many countries we change it to Argentina.

Before analyzing or cleansing the dataset, we have to change the answer and answer code heading by giving them a proper number of attributes as in answer1 and answer code1 respectively. While analyzing the data there so many missing values in the dataset. So, it is very hard to add dataset in an analytical tool for analyzing. Only RapidMiner is able to do or fill the missing values. Because of the missing value its giving error all the time so we have to change the missing value with. The help of RapidMiner. After that we had to change the language where most of the instances were in different languages so it's quite tough to understand the data so, we had translated it to English with help of google translator so that will help to understand the data. Then we had removed the attributes name as question1 to question10 because we do not need questions for analysis because we have a numerical form of answers. But we had kept the answer code as it is. Then we had to change the ethnicity. In short, we did the aggregation on the given ethnicity its means we did cluster 5 types of similar ethnicity into one which was present in the data,

- Arab,
- Arabe,
- Arabes,
- Arabic and
- Arabs.

so, we had to change it to simple word that is Arab. Because every ethnicity having their own pronunciation so they have their own way to speak so we club these five ethnicities into one. Then we had removed some columns like email, a comment from the raw data because, we are not doing any sentimental analysis we do not need comment attribute. Also, we removed autism score to avoid overfitting error and over-learned model for predictive analysis. Then we have come to the screening test type which is mainly derived into 4 age groups but here in this raw table or dataset, there is 2 age group which is almost similar to each other's values. There are two groups of 12 to 15 years and 12 to 16 years. So, after performing aggregation on the screening test type that became one individual group that is 12 to 16 years because all question is similar to age 12 to 16 year so we do not need extra band because it is very necessary to get as much as historical data. In age attribute, we did change in some instances because there are some missing values in it represented by (?, ??) so few of them we had changed to 7 years and rest of them to 77 years. Because in the age band there is an age group of 4 to 11 years so we have taken mid value of the age band that is 7 years and so there was a two-question mark so we thought that a typing error so we considered as 77 years. So, some missing value which is manually inputted in the dataset to get easy and better results while analyzing. Also, there is an age which is 383 years that is must be typing error which is totally incorrect values that a have put in the dataset which can't be the age of anyone so it should be 38 or 83 years so we had taken 38 years.

Also, there is 1 blank value is present in the dataset that is replaced by m. As have mentioned above that there are few missing values in the dataset so after fixing it off, some of them change half of them into male and female in gender attribute. In the jaundice attribute, there are three null values that had changes to NO because there is NO in the result. After there is another attribute that is family with ASD having blank values that had changes NO because there is NO in the result. In the user app before attribute having a blank value that is replaced by NO because most of the values are NO that's why we change it to NO. In the gender attribute, we had to make changes in M to m and F to f. Because it is common or the same thing. While analyzing the data there are some attributes are less important so that column should have been removed which are not going to use in the cleansed dataset. Those are who have taken a test, use the app before. Because who has taken a test attribute is not important. In age attribute, there are one values named as 32 months so we have removed month from that instance and just kept 32 as in age. I have conformed it because this instance is coming in the toddler's data so I have changed it. Because without this change not an analytical tool was taking this data just because of this small mistake its giving error all the time for that purpose we did changes.

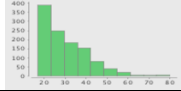


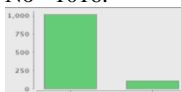
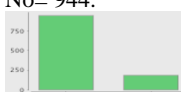
5. DATA UNDERSTANDING

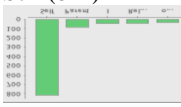



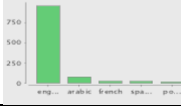
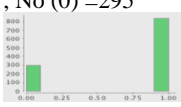
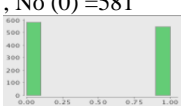
After converting file, we have faced one more problem that the dataset having different languages in different attributes and we translated it into English by using google translator. Attributes characteristics are an integer as well as there are some missing values in the dataset. Whereas the first attribute is case number which is integer value which assigns the number of instances. Then in the second attribute is Question no 1 which a string value and there is are 10 different types of questions and

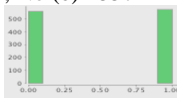
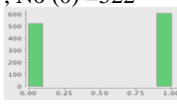
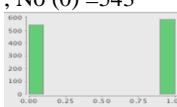

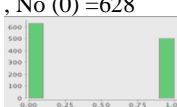
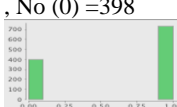
answers having their own answers code which is in binary value (1 or 0) which give the output in Yes or No. Email attribute is having lots of NULL value. After that, there is an age and gender attribute which is numerical and string variable respectively. There are 40 types of ethnicity(string) and their residency status(String). Status of those people who are born with jaundice which is Boolean value in Yes or No. Also, the status of those family members with PDD (Pervasive Development Disorder) which is a Boolean value. Screening method types are an integer which is segmented into four section that is 0,1,2,3. And also there some attributes, for e.g. language, who is taken the test, user app before, comments and class(binary). This descriptive analysis we have did with the help of Excel and RapidMiner 8.1.0.



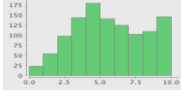
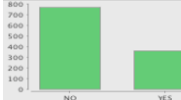
Now we made the three different excel sheet as per screening test type. One of them is 4-11years, 12-16 years and 17 years to older. So, in the 4 to 11 years having 546 instances and 22 attributes. In 12-16 years having 281 instances and 22 attributes and 17 years to older having 1129 instances and 22 attributes.

5.1 17 years to elder attribute dataset analysis:



Attribute	Type	Description	Analysis on attributes
Age	Number 17 years or older	Age in years	Min=17, Max= 80, Average= 30.15 
Gender	String Cleansed data- [Female=532, Male=596	Male or Female	Male= 596, Female= 532 
Ethnicity	String	List of common ethnicities in text format	113 types of ethnicity Least= White Irish (1) Most= White-European (331) 
Jaundice	Boolean (Yes or No) Cleansed data Yes=112, No=1016	whether the case was born with jaundice	Yes= 112 No= 1016. 
Family_with_ASD	Boolean (Yes or No) Raw data Yes=184, No=944	Whether any immediate family member has a PDD	Yes= 184, No= 944. 


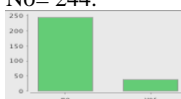

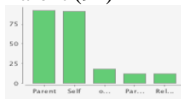

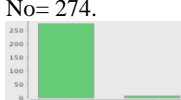


Who is taken the test	String	Parent, self, caregiver, medical staff, clinician, etc.	Least=1, Self (814) 
Residence	string	List of countries in text format	119 countries Least Zambia (1), Most US (216). 
user_app_before	Boolean (Yes or No) Cleansed data Yes=22, No=1106	Whether the user has used a screening app	Yes= 22, No= 1106. 
Screening Test Type	Integer 17 years or older.	The type of screening method chosen based on age category 3=adult	17 years and more= 1128, 
Language	Polynomial	Language spoken by people	Swahili (1), English (951) 
Answer code1	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 833 , No (0) =295 
Answer code2	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 547 , No (0) =581 


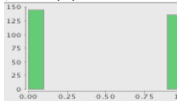
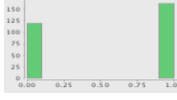
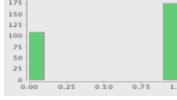
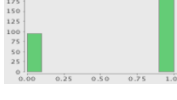

Answer code3	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 571 , No (0) =557 
Answer code4	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 606 , No (0) =522 
Answer code5	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 585 , No (0) =543 
Answer code6	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 366 , No (0) =762 
Answer code7	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 500 , No (0) =628 
Answer code8	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 730 , No (0) =398 


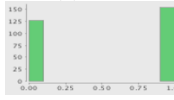
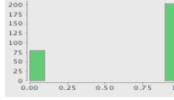
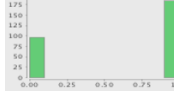
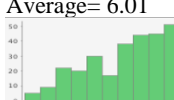
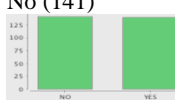
		method used	
Answer code9	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 415 , No (0) =713 
Answer code10	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 685 , No (0) =443 
autism_score	Integer That's come between 1 to 10	The final score obtained based on the scoring algorithm of the screening method used.	Min= 0, Max= 10, Average= 5.17 
Class (Result)	Polynomial That's come between yes or no	This is the results about who has autism or not.	Yes (360) No (768) 

5.2 12-16 years attribute dataset analysis:

Attribute	Type	Description	Analysis on attributes
Age	Number 12-16 years.	Age in years	Min=12, Max= 17, Average= 14.09 
Gender	String Cleansed data- [Female=159, Male=122]	Male or Female	Male= 122, Female= 159 

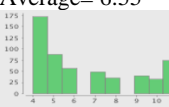

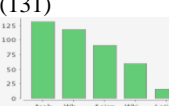
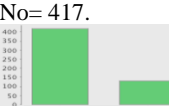
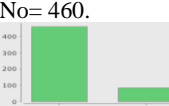


Ethnicity	String	List of common ethnicities in text format	35 types of ethnicity Least= White African (1) Most= Arab (74) 
Jaundice	Boolean (Yes or No) Cleansed data Yes=37, No=144	whether the case was born with jaundice	Yes= 37 No= 244. 
Family_with_A SD	Boolean (Yes or No) Raw data Yes=45, No=236	Whether any immediate family member has a PDD	Yes= 45, No= 236. 
Who is taken the test	String	Parent, self, caregiver, medical staff, clinician, etc.	Relative=1, Parent (92) 
Residence	string	List of countries in text format	48 countries Viet Nam (1), Most UK (76). 
user_app_before	Boolean (Yes or No) Cleansed data Yes=7, No=274	Whether the user has used a screening app	Yes= 7, No= 274. 
Screening Test Type	Integer 12-16 years	The type of screening method chosen based on age category (2=adult)	12-16 years = 281. 
Language	Polynomial	Language spoken by people	Russian (1), English (210) 

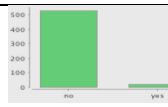

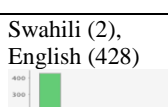
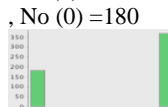
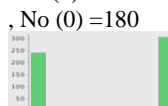
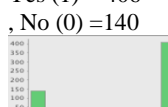
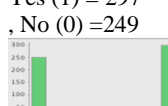
Answer code1	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 187 , No (0) =94 
Answer code2	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 137 , No (0) =145 
Answer code3	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 162 , No (0) =119 
Answer code4	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 173 , No (0) =108 
Answer code5	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 186 , No (0) =95 
Answer code6	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 177 , No (0) =104 


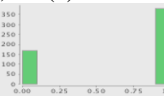
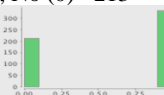

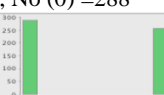

Answer code7	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 128 , No (0) =153 
Answer code8	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 154 , No (0) =27 
Answer code9	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 202 , No (0) =79 
Answer code10	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 184 , No (0) =97 
autism_score	Integer That's come between 1 to 10	The final score obtained based on the scoring algorithm of the screening method used.	Min= 0, Max= 10, Average= 6.01 
Class (Result)	Polynomial That's come between yes or no	This is the results about who has	Yes (140) No (141) 

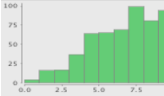
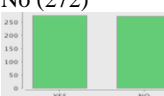
		autism or not.	
--	--	----------------	--

5.3 4-11 years attribute dataset analysis:

Attribute	Type	Description	Analysis on attributes
Age	Number 4-11 years.	Age in years	Min=4, Max= 11, Average= 6.55 
Gender	String Cleansed data- [Female=175, Male=371]	Male or Female	Male= 371, Female= 175 
Ethnicity	String	List of common ethnicities in text format	59 types of ethnicity Least= White Arish (1) Most= Arab (131) 
Jaundice	Boolean (Yes or No) Cleansed data Yes=129, No=417	whether the case was born with jaundice	Yes= 129 No= 417. 
Family_with_ASD	Boolean (Yes or No) Cleansed data Yes=86, No=460	Whether any immediate family member has a PDD	Yes= 86, No= 460. 
Who is taken the test	String	Parent, self, caregiver, medical staff, clinician, etc.	Teacher=1, Parent (92) 
Residence	string	List of countries in text format	72 countries Virgin Island (1), Most UK (104). 
user_app_before	Boolean (Yes or No)	Whether the user	Yes= 19, No= 527.

	Cleansed data Yes=19, No=527	has used a screening app	
Screening Test Type	Integer 4-11 years	The type of screening method chosen based on age category (1=adult)	4-11 years = 546. 
Language	Polynomial	Language spoken by people	Swahili (2), English (428) 
Answer code1	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 366, No (0) =180 
Answer code2	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 305, No (0) =180 
Answer code3	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 406, No (0) =140 
Answer code4	Binary (0,1)	The answer code of the question base on the screening	Yes (1) = 297, No (0) =249 

Answer code5	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 390 No (0) =156 
Answer code6	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 379 No (0) =167 
Answer code7	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 333 No (0) =213 
Answer code8	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 268 No (0) =278 
Answer code9	Binary (0,1)	The answer code of the question base on the screening method used	Yes (1) = 258 No (0) =288 
Answer code10	Binary (0,1)	The answer code of the question base on	Yes (1) = 376 No (0) =170 

		the screening method used	
autism_score	Integer That's come between 1 to 10	The final score obtained based on the scoring algorithm of the screening method used.	Min= 0, Max= 10, Average= 6.18 
Class (Result)	Polynomial That's come between yes or no	This is the results about who has autism or not.	Yes (274) No (272) 

6. DISCUSSION ABOUT METHODS:

6.1 Logistic Regression Model

Logistic Regression Model is a statistical method for examine a dataset in which there is a more than one attributes or independent variable to get the expected output. The output is measured with a characterized variable in which there should be only two possible results. In the Logistic Regression Model, the dependent variable should be binary or characterized for example it contains data as **0** (FALSE, failure) and **1** (TRUE, success).

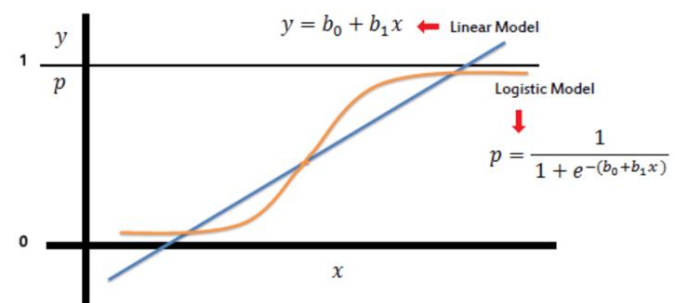


Fig 3. Logistic Regression Model

The goal of Logistic Regression Model is to find out the best fitting model to explain the connectivity between the characterized characteristic of interest for the dependent variable. Logistic Regression Model generates the coefficients of the formula to predict a logit transformation of the probability:

$$\text{logit}(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_kX_k$$

In this formula where p is the probability of presence of the characteristics of interest whereas the logit transformation is defined as the logged odds:

$$\text{odds} = \frac{p}{1-p} = \frac{\text{probability of presence of characteristic}}{\text{probability of absence of characteristic}}$$

PROS

- Provides probability for outcome

- It has low variance
- Work well with diagonal decision boundaries. (Tufts, 2015)

CONS

- Identifying independent variables
- Limited outcome variables
- Independent observation required
- Overfitting the models (Tufts, 2015)

6.2 Random Forest Model

Basically, a Random Forest Model is an algorithm which comes to supervised learning. The meaning is given in the name itself, it creates a forest and makes it random. In simple words, its create multiple decision trees and group them together to get a more accurate and stable prediction. This tree most of the time trained with the bagging method. And the basic concept of the bagging method is that a combination of learning models increases the overall results.

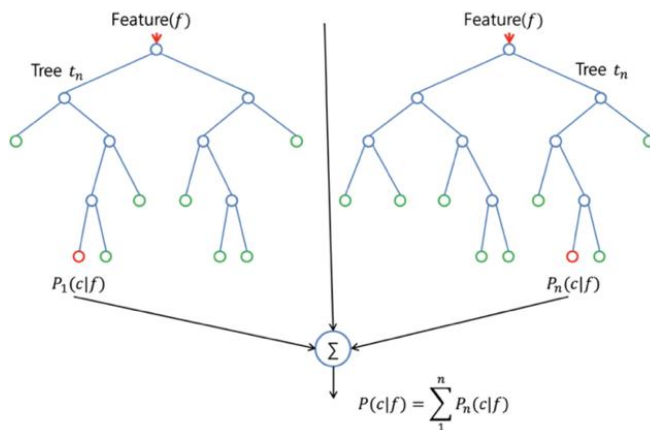


Fig 4. Random Forest Model

There are two stages in the Random Forest Model algorithm, one is to creation of Random Forest Model and the other is to do prediction from the Random Forest Model classifier which is created in the first stage. Here is the Random Forest Model creation pseudocode: (Medium, 2017)

Randomly select “K” feature from total “m” features where $k \ll m$

- Among the “K” features, calculate the node “d” with the help of best split point
- Split the node into child nodes with the help of best splits
- Repeat the 1 to 3 steps until “1” number of nodes has been reached
- Create random by repeating steps 1 to 4 for “n” number times to create “n” number of trees.

PROS

- It Decorrelates tree which is relative to bagging trees.
- It’s very important when it related to multiple features which are correlated.
- It reduces variance which is relative to regular trees. (Tufts, 2015)

CONS

- Random Forest Model is not that too easy to visually interpret. (Tufts, 2015)

7. ANALYSIS

This is the cleaned data that we have got after analyzing the raw data that we have analyzed in the WEKA 3.8.0. In this analysis we have divided the cleansed data into three different excel sheet as per age groups which are,

- 4-11 years,
- 12-16 years and
- 17 years to older.

With the help this 3 age groups we have analyzed the data and apply some algorithms for getting appropriate or expected accuracy level. So, we have did this analysis on the WEKA 3.8.0 analytical tool which we were performed 2 algorithms just for comparison to get the better results.

As well as we did cross-validation with 10 folds. This is a technique to assess predictive models by dividing the real sample into a training set to train the model. 10 folds mean its divided the real sample is arbitrarily segmented into 10 equal size subsamples. Whereas of the 10 subsamples, a single subsample is reserved as the validation data for testing the model and remaining 10-1 subsamples are used as training data and the cross-validation process is then repeated 10 times or the folds, with each of the 10 subsamples. So, we have started the analyzing the first data sheet that is 4-11 years age group.

7.1 4-11 years age group for Random Forest Model:

Firstly, we have opened the WEKA explorer and then in the processes we open the file or data sheet into it. After getting all information about the attributes then we select the select attributes option that will help to get the accurate information about the that we one is low ranked that should be removed from the attribute list. Then we found that there is a case number and autism score is the low ranked attributes. After that in the attribute list we have removed the attribute case number because it’s a reference and primary key. Also, we have removed autism score just to avoiding the over fitting error and over learned error. Then we have change the all attribute type from numeric to factor by using discretize filter. Then we moved to select attribute and choose the attribute evaluator which give the Correlation Attribute Evaluation with the help of ranker search method of cross validation which give the all information about the very single attribute who have low ranked or average ranked.

=== Attribute selection 10 fold cross-validation (stratified), seed: 1
===

average merit	average rank	attribute
0.595 +- 0.01	1 +- 0	4 Answer code4
0.475 +- 0.013	2 +- 0	6 Answer code6
0.437 +- 0.019	3.6 +- 0.92	9 Answer code9
0.416 +- 0.01	4.4 +- 1.2	5 Answer code5
0.413 +- 0.013	4.9 +- 0.7	3 Answer code3
0.414 +- 0.014	5.2 +- 1.08	8 Answer code8
0.377 +- 0.013	7.2 +- 0.4	1 Answer code1
0.366 +- 0.011	7.7 +- 0.64	10 Answer code10
0.33 +- 0.014	9 +- 0	7 Answer code7
0.228 +- 0.015	10 +- 0	2 Answer code2
0.088 +- 0.009	11.1 +- 0.3	20 Who is taken the test
0.073 +- 0.006	12 +- 0.45	13 ethnicity
0.06 +- 0.004	13.4 +- 0.49	16 Residence
0.06 +- 0.009	13.5 +- 0.67	19 Language
0.038 +- 0.003	15.1 +- 0.3	11 Age

0.019 +- 0.01 17 +- 1.34 14 Jaundice
0.017 +- 0.008 17.4 +- 1.02 12 gender
0.016 +- 0.006 17.7 +- 0.9 15 Family_with_AS
0.013 +- 0.009 17.8 +- 1.17 17 user_app_before
0 +- 0 20 +- 0 18 Screening Test Type

Then we found the screening test type having very low rank which is 0 so we have removed that attribute. Then after we have tried one more wrapping method that is wrapper subset evaluator with the help of Greedy stepwise method which is ranking all the attributes as 0 which is incorrect so we have drop this method and we have considered as not perfect method for this dataset. We got this result from the Greedy stepwise method, === Attribute selection 10-fold cross-validation (stratified), seed:1===

number of folds (%) attribute

0(0 %) 1 Answer code1
0(0 %) 2 Answer code2
0(0 %) 3 Answer code3
0(0 %) 4 Answer code4
0(0 %) 5 Answer code5
0(0 %) 6 Answer code6
0(0 %) 7 Answer code7
0(0 %) 8 Answer code8
0(0 %) 9 Answer code9
0(0 %) 10 Answer code10
0(0 %) 11 Age
0(0 %) 12 gender
0(0 %) 13 ethnicity
0(0 %) 14 Jaundice
0(0 %) 15 Family_with_AS
0(0 %) 16 Residence
0(0 %) 17 user_app_before
0(0 %) 18 Screening Test Type
0(0 %) 19 Language
0(0 %) 20 Who is taken the test

Because of this result we have not used this method for further dataset so only the correct method is 10-fold cross validation that we have used for the analysis. And we have used 19 best attributes as suggested by the ranking method for this analysis instead of 22 attributes. We applied classifier on the dataset that is Random Forest Model and test option is cross validations with 10 folds. After performing Random Forest Model, we have got 87.17% of correctly classified instances and 12.82% of incorrectly classified instances.

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	476	87.1795 %
Incorrectly Classified Instances	70	12.8205 %
Kappa statistic	0.7436	
Mean absolute error	0.2571	
Root mean squared error	0.3186	
Relative absolute error	51.4181 %	
Root relative squared error	63.7234 %	
Total Number of Instances	546	

=== Detailed Accuracy by Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC
ROC Area	PRC Area	Class			

	0.860	0.117	0.880	0.860	0.870	0.744	0.954
0.955 NO							
	0.883	0.140	0.864	0.883	0.874	0.744	0.954
0.958 YES							
Weighted Avg.	0.872	0.128	0.872	0.872	0.872	0.744	0.954
0.954 0.956							

=== Confusion Matrix ===

N=546 Actual value	Predicted values	
	FALSE	TRUE
0	234(TN)	38(FP)
1	32(FN)	242(TP)

True positive (TP): these are the cases in which we predicted yes (they have autism) or they do have autism.

True negative (TN): we predicted no, and they don't have autism

False positive (FP): we predicted yes, but they don't have autism

False negative (FN): we predicted no, but they do have autism.

Accuracy rate: To calculate how often the classifier is correct we use $((TP+TN)/Total)$

By confusion matrix this is the Accuracy rate $(234+242)/(234+242+38+32) = 0.87=87\%$

Error rate: To calculate how often it is wrong we use $(FP+FN)/Total$

By confusion matrix this is the Error rate: $(38+32)/(234+242+38+32) = 0.128=12.82\%$

7.2 12-16-year age group for Random Forest Model:

In this group we did the same for case number and autism score. That we removed this attribute where case number because it's a reference and a primary key in the dataset. As well as the autism score is removed because autism score just to avoiding the over fitting error and over learned error. Then we have change the all attribute type from numeric to factor by using discretize filter. Then we moved to select attribute and choose the attribute evaluator which give the Correlation Attribute Evaluation with the help of ranker search method of cross validation which give the all information about the very single attribute who have low ranked or average ranked.

=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

average merit	average rank	attribute
0.616 +- 0.016	1.1 +- 0.3	6 Answer code6
0.595 +- 0.013	1.9 +- 0.3	3 Answer code3
0.553 +- 0.012	3 +- 0	4 Answer code4
0.517 +- 0.011	4 +- 0	5 Answer code5
0.389 +- 0.017	5.6 +- 1.02	7 Answer code7
0.374 +- 0.025	6.8 +- 1.47	2 Answer code2
0.37 +- 0.013	6.8 +- 1.17	9 Answer code9
0.359 +- 0.02	7.7 +- 0.9	1 Answer code1
0.334 +- 0.016	9 +- 0.89	10 Answer code10
0.333 +- 0.024	9.1 +- 1.37	8 Answer code8
0.281 +- 0.014	11 +- 0	19 Language
0.15 +- 0.011	12 +- 0	13 ethnicity
0.125 +- 0.007	13.4 +- 0.49	20 Who is taken the test
0.127 +- 0.006	13.6 +- 0.49	16 Residence
0.068 +- 0.016	15.7 +- 0.78	17 user_app_before
0.064 +- 0.007	15.8 +- 0.75	11 Age
0.054 +- 0.012	16.8 +- 0.87	14 Jaundice
0.021 +- 0.017	18.3 +- 0.64	12 gender
0.016 +- 0.014	18.4 +- 0.92	15 Family_with_AS

0 +- 0 20 +- 0 18 Screening Test Type

Then we found the screening test type having very low rank which is 0 so we have removed that attribute. We applied classifier on the dataset that is Random Forest Model and test option is cross validations with 10 folds. After performing Random Forest Model, we have got 84.69% of correctly classified instances and 15.30% of incorrectly classified instances.

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	238	84.6975 %
Incorrectly Classified Instances	43	15.3025 %
Kappa statistic	0.6941	
Mean absolute error	0.2589	
Root mean squared error	0.3252	
Relative absolute error	51.7895 %	
Root relative squared error	65.0451 %	
Total Number of Instances	281	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC
ROC Area PRC Area Class						
0.780 0.086 0.902	0.780	0.837	0.701	0.949		
0.952 NO						
0.914 0.220 0.805	0.914	0.856	0.701	0.949		
0.952 YES						
Weighted Avg.	0.847	0.153	0.854	0.847	0.846	0.701
0.949 0.952						

=== Confusion Matrix ===

N=281	Predicted values	
Actual value	FALSE	TRUE
0	110(TN)	31(FP)
1	12(FN)	128(TP)

True positive (TP): these are the cases in which we predicted yes (they have autism) or they do have autism.

True negative (TN): we predicted no, and they don't have autism

False positive (FP): we predicted yes, but they don't have autism

False negative (FN): we predicted no, but they do have autism.

Accuracy rate: To calculate how often the classifier is correct we use $((TP+TN)/Total)$

By confusion matrix this is the Accuracy rate $((110+128)/(110+128+31+12) = 0.84=84.7\%$

Error rate: To calculate how often it is wrong we use $((FP+FN)/Total)$

By confusion matrix this is the Error rate: $((31+12)/(110+128+31+12)=0.153=15.30\%$

7.3 17 years and older age group for Random Forest Model:

In this group we did the same for case number and autism score. That we removed this attribute where case number because it's a reference and a primary key in the dataset. As well as the autism score is removed because autism score just to avoiding the over fitting error and over learned error. Then we have change the all attribute type from numeric to factor by using discretize filter. Then we moved to select attribute and choose the attribute evaluator which give the Correlation Attribute Evaluation with the help of ranker search method of cross validation which give the all information about the very single attribute who have low ranked or average ranked.

=== Attribute selection 10-fold cross-validation (stratified), seed: 1 ===

average merit average rank attribute

0.618 +- 0.008	1.1 +- 0.3	6 Answer code6
0.598 +- 0.008	1.9 +- 0.3	9 Answer code9
0.576 +- 0.005	3 +- 0	5 Answer code5
0.471 +- 0.006	4.1 +- 0.3	4 Answer code4
0.444 +- 0.013	4.9 +- 0.3	3 Answer code3
0.403 +- 0.008	6 +- 0	10 Answer code10
0.381 +- 0.009	7 +- 0	7 Answer code7
0.31 +- 0.009	8.1 +- 0.3	2 Answer code2
0.295 +- 0.008	8.9 +- 0.3	1 Answer code1
0.259 +- 0.009	10 +- 0	8 Answer code8
0.156 +- 0.008	11 +- 0	15 Family_with_ASD
0.134 +- 0.002	12 +- 0	13 ethnicity
0.103 +- 0.004	13.1 +- 0.3	16 Residence
0.084 +- 0.015	14.8 +- 0.75	14 Jaundice
0.082 +- 0.01	15 +- 1	17 user_app_before
0.068 +- 0.008	16.3 +- 0.64	20 Who is taken the test
0.066 +- 0.018	16.7 +- 2	12 gender
0.061 +- 0.008	17.4 +- 1.02	19 Language
0.05 +- 0.004	18.7 +- 0.46	11 Age

0 +- 0 20 +- 0 18 Screening Test Type

Then we found the screening test type having very low rank which is 0 so we have removed that attribute. We applied classifier on the dataset that is Random Forest Model and test option is cross validations with 10 folds. After performing Random Forest Model, we have got 92.37% of correctly classified instances and 7.62% of incorrectly classified instances.

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	1042	92.3759 %
Incorrectly Classified Instances	86	7.6241 %
Kappa statistic	0.8243	
Mean absolute error	0.1641	
Root mean squared error	0.2475	
Relative absolute error	37.7577 %	
Root relative squared error	53.0856 %	
Total Number of Instances	1128	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC
ROC Area PRC Area Class						
0.945 0.122 0.943	0.945	0.944	0.824	0.980		
0.991 NO						
0.878 0.055 0.883	0.878	0.880	0.824	0.980		
0.958 YES						
Weighted Avg.	0.924	0.101	0.924	0.924	0.924	0.824
0.980 0.980						

=== Confusion Matrix ===

N=1128	Predicted values	
Actual value	FALSE	TRUE
0	726(TN)	42(FP)
1	44(FN)	316(TP)

True positive (TP): these are the cases in which we predicted yes (they have autism) or they do have autism.

True negative (TN): we predicted no, and they don't have autism

False positive (FP): we predicted yes, but they don't have autism

False negative (FN): we predicted no, but they do have autism.

Accuracy rate: To calculate how often the classifier is correct we use $((TP+TN)/Total)$

By confusion matrix this is the Accuracy rate (726+316)/(726+316+44+42) = 0.92=92.38%

Error rate: To calculate how often it is wrong we use (FP+FT)/Total

By confusion matrix this is the Error rate:(42+44)/(726+316+42+44) = 0.076=7.62%

7.4 4-11 years age group for Logistic Regression Model:

In this group we did the same for case number and autism score. That we removed this attribute where case number because it's a reference and a primary key in the dataset. As well as the autism score is removed because autism score just to avoiding the over fitting error and over learned error. Then we have change the all attribute type from numeric to factor by using discretize filter. Then we moved to select attribute and choose the attribute evaluator which give the Correlation Attribute Evaluation with the help of ranker search method of cross validation which give the all information about the very single attribute who have low ranked or average ranked.

=== Attribute selection 10 fold cross-validation (stratified), seed: ===

average merit	average rank	attribute
0.595 +- 0.01	1 +- 0	4 Answer code4
0.475 +- 0.013	2 +- 0	6 Answer code6
0.437 +- 0.019	3.6 +- 0.92	9 Answer code9
0.416 +- 0.01	4.4 +- 1.2	5 Answer code5
0.413 +- 0.013	4.9 +- 0.7	3 Answer code3
0.414 +- 0.014	5.2 +- 1.08	8 Answer code8
0.377 +- 0.013	7.2 +- 0.4	1 Answer code1
0.366 +- 0.011	7.7 +- 0.64	10 Answer code10
0.33 +- 0.014	9 +- 0	7 Answer code7
0.228 +- 0.015	10 +- 0	2 Answer code2
0.088 +- 0.009	11.1 +- 0.3	20 Who is taken the test
0.073 +- 0.006	12 +- 0.45	13 ethnicity
0.06 +- 0.004	13.4 +- 0.49	16 Residence
0.06 +- 0.009	13.5 +- 0.67	19 Language
0.038 +- 0.003	15.1 +- 0.3	11 Age
0.019 +- 0.01	17 +- 1.34	14 Jaundice
0.017 +- 0.008	17.4 +- 1.02	12 gender
0.016 +- 0.006	17.7 +- 0.9	15 Family_with_ASD
0.013 +- 0.009	17.8 +- 1.17	17 user_app_before
0 +- 0	20 +- 0	18 Screening Test Type

Then we found the screening test type having very low rank which is 0 so we have removed that attribute. We applied classifier on the dataset that is Logistic Regression Model and test option is cross validations with 10 folds. After performing Logistic Regression Model, we have got 96.33% of correctly classified instances and 3.66% of incorrectly classified instances.

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	526	96.337 %
Incorrectly Classified Instances	20	3.663 %
Kappa statistic	0.9267	
Mean absolute error	0.0351	
Root mean squared error	0.1785	
Relative absolute error	7.0206 %	
Root relative squared error	35.6958 %	
Total Number of Instances	546	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC
ROC Area	0.949	0.022	0.977	0.949	0.963	0.927
0.996 NO	0.978	0.051	0.950	0.978	0.964	0.927
0.992 YES	0.963	0.037	0.964	0.963	0.963	0.927
Weighted Avg.	0.995	0.994				

=== Confusion Matrix ===

N=546	Predicted values	
Actual value	FALSE	TRUE
0	258(TN)	14(FP)
1	6(FN)	268(TP)

True positive (TP): these are the cases in which we predicted yes (they have autism) or they do have autism.

True negative (TN): we predicted no, and they don't have autism

False positive (FP): we predicted yes, but they don't have autism

False negative (FN): we predicted no, but they do have autism.

Accuracy rate: To calculate how often the classifier is correct we use ((TP+TN)/Total

By confusion matrix this is the Accuracy rate (258+268)/(258+268+6+14) = 0.96=96.33%

Error rate: To calculate how often it is wrong we use (FP+FT)/Total

By confusion matrix this is the Error rate:(6+14)/(258+268)/(258+268+6+14) = 0.03=3.66%.

7.5 12-16-year age group for Logistic Regression Model:

In this group we did the same for case number and autism score. That we removed this attribute where case number because it's a reference and a primary key in the dataset. As well as the autism score is removed because autism score just to avoiding the over fitting error and over learned error. Then we have change the all attribute type from numeric to factor by using discretize filter. Then we moved to select attribute and choose the attribute evaluator which give the Correlation Attribute Evaluation with the help of ranker search method of cross validation which give the all information about the very single attribute who have low ranked or average ranked.

=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

average merit	average rank	attribute
0.616 +- 0.016	1.1 +- 0.3	6 Answer code6
0.595 +- 0.013	1.9 +- 0.3	3 Answer code3
0.553 +- 0.012	3 +- 0	4 Answer code4
0.517 +- 0.011	4 +- 0	5 Answer code5
0.389 +- 0.017	5.6 +- 1.02	7 Answer code7
0.374 +- 0.025	6.8 +- 1.47	2 Answer code2
0.37 +- 0.013	6.8 +- 1.17	9 Answer code9
0.359 +- 0.02	7.7 +- 0.9	1 Answer code1
0.334 +- 0.016	9 +- 0.89	10 Answer code10
0.333 +- 0.024	9.1 +- 1.37	8 Answer code8
0.281 +- 0.014	11 +- 0	19 Language
0.15 +- 0.011	12 +- 0	13 ethnicity
0.125 +- 0.007	13.4 +- 0.49	20 Who is taken the test
0.127 +- 0.006	13.6 +- 0.49	16 Residence
0.068 +- 0.016	15.7 +- 0.78	17 user_app_before
0.064 +- 0.007	15.8 +- 0.75	11 Age
0.054 +- 0.012	16.8 +- 0.87	14 Jaundice

0.021 +- 0.017 18.3 +- 0.64 12 gender
0.016 +- 0.014 18.4 +- 0.92 15 Family_with_ASD
0 +- 0 20 +- 0 18 Screening Test Type

Then we found the screening test type having very low rank which is 0 so we have removed that attribute. We applied classifier on the dataset that is Logistic Regression Model and test option is cross validations with 10 folds. After performing Logistic Regression Model, we have got 87.54% of correctly classified instances and 12.45% of incorrectly classified instances.

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	246	87.5445 %
Incorrectly Classified Instances	35	12.4555 %
Kappa statistic	0.7509	
Mean absolute error	0.1263	
Root mean squared error	0.3449	
Relative absolute error	25.2517 %	
Root relative squared error	68.9761 %	
Total Number of Instances	281	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC
ROC Area	0.865	0.114	0.884	0.865	0.875	0.751
PRC Area	0.949	NO				
Class	0.886	0.135	0.867	0.886	0.876	0.751
YES	0.938	YES				
Weighted Avg.	0.875	0.124	0.876	0.875	0.875	0.751
0.947	0.94					

3=== Confusion Matrix ===

N=281	Predicted values	
Actual value	FALSE	TRUE
0	122(TN)	19(FP)
1	16(FN)	124(TP)

True positive (TP): these are the cases in which we predicted yes (they have autism) or they do have autism.

True negative (TN): we predicted no, and they don't have autism

False positive (FP): we predicted yes, but they don't have autism

False negative (FN): we predicted no, but they do have autism.

Accuracy rate: To calculate how often the classifier is correct we use $((TP+TN)/Total)$

By confusion matrix this is the Accuracy rate $(122+124)/(122+124+16+19) = 0.87=87.54\%$

Error rate: To calculate how often it is wrong we use $(FP+FN)/Total$

By confusion matrix this is the Error rate: $(20+24)/(122+124+16+19) = 0.12=12.45\%$.

7.6 17 years and older age group for Logistic Regression Model:

In this group we did the same for case number and autism score. That we removed this attribute where case number because it's a reference and a primary key in the dataset. As well as the autism score is removed because autism score just to avoiding the over fitting error and over learned error. Then we have change the all attribute type from numeric to factor by using discretize filter. Then we moved to select attribute and choose the attribute evaluator which give the Correlation Attribute Evaluation with the help of ranker search method of cross validation which give the all

information about the very single attribute who have low ranked or average ranked.

=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

average merit	average rank	attribute
0.618 +- 0.008	1.1 +- 0.3	6 Answer code6
0.598 +- 0.008	1.9 +- 0.3	9 Answer code9
0.576 +- 0.005	3 +- 0	5 Answer code5
0.471 +- 0.006	4.1 +- 0.3	4 Answer code4
0.444 +- 0.013	4.9 +- 0.3	3 Answer code3
0.403 +- 0.008	6 +- 0	10 Answer code10
0.381 +- 0.009	7 +- 0	7 Answer code7
0.31 +- 0.009	8.1 +- 0.3	2 Answer code2
0.295 +- 0.008	8.9 +- 0.3	1 Answer code1
0.259 +- 0.009	10 +- 0	8 Answer code8
0.156 +- 0.008	11 +- 0	15 Family_with_ASD
0.134 +- 0.002	12 +- 0	13 ethnicity
0.103 +- 0.004	13.1 +- 0.3	16 Residence
0.084 +- 0.015	14.8 +- 0.75	14 Jaundice
0.082 +- 0.01	15 +- 1	17 user_app_before
0.068 +- 0.008	16.3 +- 0.64	20 Who is taken the test
0.066 +- 0.018	16.7 +- 2	12 gender
0.061 +- 0.008	17.4 +- 1.02	19 Language
0.05 +- 0.004	18.7 +- 0.46	11 Age
0 +- 0	20 +- 0	18 Screening Test Type

Then we found the screening test type having very low rank which is 0 so we have removed that attribute. We applied classifier on the dataset that is Logistic Regression Model and test option is cross validations with 10 folds. After performing Logistic Regression Model, we have got 95.03% of correctly classified instances and 4.96% of incorrectly classified instances.

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	1072	95.0355 %
Incorrectly Classified Instances	56	4.9645 %
Kappa statistic	0.8873	
Mean absolute error	0.0491	
Root mean squared error	0.2198	
Relative absolute error	11.2929 %	
Root relative squared error	47.1455 %	
Total Number of Instances	1128	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC
ROC Area	0.952	0.053	0.975	0.952	0.963	0.888
PRC Area	0.986	NO				
Class	0.947	0.048	0.902	0.947	0.924	0.888
YES	0.932	YES				
Weighted Avg.	0.950	0.051	0.952	0.950	0.951	0.888
0.975	0.969					

=== Confusion Matrix ===

N=1128	Predicted values	
Actual value	FALSE	TRUE
0	731(TN)	37(FP)
1	19(FN)	341(TP)

True positive (TP): these are the cases in which we predicted yes (they have autism) or they do have autism.

True negative (TN): we predicted no, and they don't have autism

False positive (FP): we predicted yes, but they don't have autism
False negative (FN): we predicted no, but they do have autism.

Accuracy rate: To calculate how often the classifier is correct we use $((TP+TN)/Total)$

By confusion matrix this is the Accuracy rate $(731+341)/(731+341+19+37) = 0.95=95.03\%$

Error rate: To calculate how often it is wrong we use $(FP+FT)/Total$

By confusion matrix this is the Error rate: $(19+37)/(731+341+19+37) = 0.04=4.96\%$.

8. DISCUSSION ABOUT ANALYSIS

This is another age group vice accuracy for analyzed algorithms that we have used on ASD dataset. As per the result that we have come to this conclusion at some point of stage RapidMiner is working incorrect and giving unexpected results rather than Logistic Regression Model is working absolutely fine on every screening test type and generating accurate results or output. Because at every age group is giving higher results for Logistic Regression Model only in the 12-16 age group giving approx. same results. If we take close look at all three age groups by comparing both methods for the same age group.

Difference between all two-method in-between all screening test type:

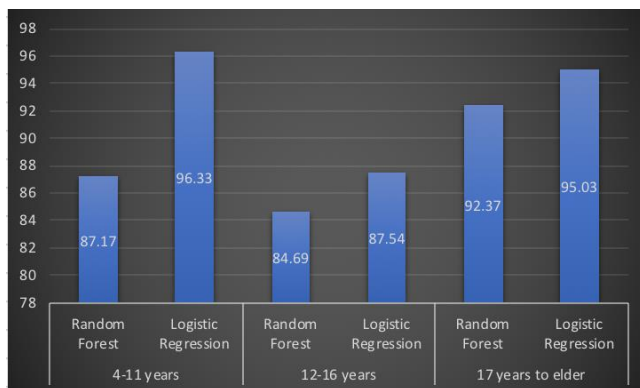


Fig 5. Comparison between Random Forest Model and Logistic Regression Model for all screening test type

In this 4-11 years is giving the major difference between these two methods so here the Random Forest Model is giving 87.17 % of accuracy while testing the dataset but Logistic Regression Model is giving higher than Random Forest Model is 96.33% accuracy. After this in 12-16 years is giving the major difference between these two methods so here the Random Forest Model is giving 84.69% of accuracy and Logistic Regression Model is giving slightly lower accuracy which is almost similar in both methods which are 87.54% accuracy. And 17 years to elder is giving the major difference between these two methods so here the Random Forest Model is giving 92.37 % of accuracy while testing the dataset but Logistic Regression Model is giving higher than Random Forest Model is 95.03% accuracy.

In short after analyzing the ASD dataset we came to this conclusion that Logistic Regression Model method is a better than Random Forest Model. Where Logistic Regression Model is giving accurate prediction or results rather than Random Forest Model.

Because Logistic Regression Model is very sensitive and it has efficient way of implementation.

9. CONCLUSION

Finally, we came to this conclusion that Logistic Regression Model is a good algorithm just because as per the predictive analysis in the Random Forest Model is predicting lower accuracy in all screening test type in the 4-11 years screening test type if giving 9.16% of difference, in 12-16 years age group is giving 2.85% and 17 years to elder age group is giving 2.66% which is lower than Logistic Regression Model. So, for the AUTISM adult dataset Logistic Regression Model is the most fitted algorithm rather than the Random Forest Model.

10. REFERENCES

- 1) McConachie, H. (2012). Effective Therapy for Anxiety in Young People with Autism Spectrum Disorder. *Http://isrctn.org/*. doi:10.1186/isrctn11219568
- 2) B. (n.d.). Using Machine Learning for Detection of Autism Spectrum Disorder. Retrieved from <http://referaat.cs.utwente.nl/conference/26/paper/7616/using-machine-learning-for-detection-of-autism-spectrum-disorder.pdf>
- 3) H., G., M., & V. (n.d.). Enhancing Data Analysis with Noise Removal. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*. Retrieved from <http://datamining.rutgers.edu/publication/tkdehcleaner.pdf>
- 4) Logistic Regression Model vs Decision Trees vs SVM: Part II. (n.d.). Retrieved April 07, 2018, from <https://www.edvancer.in/logistic-regression-vs-decision-trees-vs-svm-part2/>
- 5) Vanschoren, J. (n.d.). OpenML. Retrieved April 07, 2018, from <https://www.openml.org/a/estimation-procedures/1>
- 6) Al-jawahiri, R., & Milne, E. (2017). Retrieved April 07, 2018, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5237363/>
- 7) M. (n.d.). Impact Autism Spectrum Disorders Has On Parents. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.389.9135&rep=rep1&type=pdf>
- 8) T., C., M., M., & T. (2014). Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2010. *MMWR*, 62, 2nd ser. Retrieved from <http://tacanow.org/wp-content/uploads/2012/04/ADDMMWR-03-38-2014.pdf>
- 9) Autism Spectrum Disorder (ASD). (2016, July 11). Retrieved April 07, 2018, from <https://www.cdc.gov/ncbddd/autism/data.html>
- 10) C. (2010). The Effects of Early Intervention on Children with Autism Spectrum Disorders. *Southern Illinois University Carbondale OpenSIUC*. Retrieved from

http://opensiuc.lib.siu.edu/cgi/viewcontent.cgi?article=1259&context=gs_rp

- 11) S., S., R., J., & E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Males and Females, Scientists and Mathematicians. *Journal of Autism and Developmental Disorders*, 31, 1st ser. Retrieved from http://docs.autismresearchcentre.com/papers/2001_BCetal_AQ.pdf
- 12) L., D., B., & J. (2007). Maternal and Paternal Age and Risk of Autism Spectrum Disorders. Retrieved from

http://www.princeton.edu/~sswang/croen_paternal-maternal-age-ASD.pdf

- 13) (n.d.). Retrieved April 11, 2018, from https://www.autism.com/pro_research
- 14) Autism Rates across the Developed World. (2017, October 06). Retrieved April 11, 2018, from <https://www.focusforhealth.org/autism-rates-across-the-developed-world/>