Statistical Computational Lab
Project Report
Autistic Spectrum Disorder Detection

Siddharth Das
Dept. of Elec. and Telecomm.
Sardar Patel Institute of Technology
Mumbai, India
siddharth.das@spit.ac.in

Anushka Chintrate
Dept. of Elec. and Telecomm.
Sardar Patel Institute of Technology
Mumbai, India
anushka.chintrate@spit.ac.in

Sonal Kamble
Dept. of Elec. and Telecomm.
Sardar Patel Institute of Techology
Mumbai, India
sonal.kamble@spit.ac.in

***Abstract*—The importance of healthcare is embedded in our lives so we have opted a dataset consisting of Autistic Spectrum Disorder (ASD). ASD is a broad term used to describe a group of neuro-developmental disorders. The awareness of ASD and its number of cases are increasing , there is an urgent need to develop effective screening methods. In our project we are using ASD dataset and then applying Random Forest and Logistic Regression algorithms to test and to observe which algorithm performs better.**

I. Introduction

Autistic Spectrum Disorder (ASD) is a neuro-development condition associated with significant healthcare costs, and early diagnosis can significantly reduce these. Unfortunately, waiting times for an ASD diagnosis are lengthy and procedures are not cost effective. The economic impact of autism and the increase in the number of ASD cases across the world reveals an urgent need for the development of easily implemented and effective screening methods. Therefore, a time-efficient and accessible ASD screening is imminent to help health professionals and inform individuals whether they should pursue formal clinical diagnosis. The rapid growth in the number of ASD cases worldwide necessitates datasets related to behaviour traits. However, such datasets are rare making it difficult to perform thorough analyses to improve the efficiency, sensitivity, specificity and predictive accuracy of the ASD screening process. Presently, very limited autism datasets associated with clinical or screening are available and most of them are genetic in nature. Hence, we propose a new dataset related to autism screening of adults that contained 20 features to be utilised for further analysis especially in determining influential autistic traits and improving the classification of ASD cases.

II. METHODOLOGY

A. Data Collection

The outcome of this step is generally a representation of data which we will use for training.

Under the support and guidance of Prof. Dayanand Ambawade


```
# Encode labels in column 'species'.
print(df['Sex'].unique())
df['Sex']= label_encoder.fit_transform(df['Sex'])
print(df['Sex'].unique(), '\n')

print(df['Ethnicity'].unique())
df['Ethnicity']= label_encoder.fit_transform(df['Ethnicity'])
print(df['Ethnicity'].unique(), '\n')

print(df['Jaundice'].unique())
df['Jaundice']= label_encoder.fit_transform(df['Jaundice'])
print(df['Jaundice'].unique(), '\n')

print(df['Family_mem_with_ASD'].unique())
df['Family_mem_with_ASD']= label_encoder.fit_transform(df['Family_mem_with_ASD'])
print(df['Family_mem_with_ASD'].unique(), '\n')

print(df['Who completed the test'].unique())
df['Who completed the test']= label_encoder.fit_transform(df['Who completed the test'])
print(df['Who completed the test'].unique(), '\n')
```

```
['f' 'm']
[0 1]

['middle eastern' 'White European' 'Hispanic' 'black' 'asian'
 'south asian' 'Native Indian' 'Others' 'Latino' 'mixed' 'Pacifica']
[ 8  5  0  7  6 10  2  3  1  9  4]

['yes' 'no']
[1 0]

['no' 'yes']
[0 1]

['family member' 'Health Care Professional' 'Health care professional'
 'Self' 'Others']
[4 0 1 3 2]
```

Fig. 1. Label Encoding

B. Data Preparation

Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.)

Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data.

Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis.

Data was appropriately encoded using label encoding; used to handle categorical features.

C. Choose a Model

Choice of algorithm is on the basis of application.

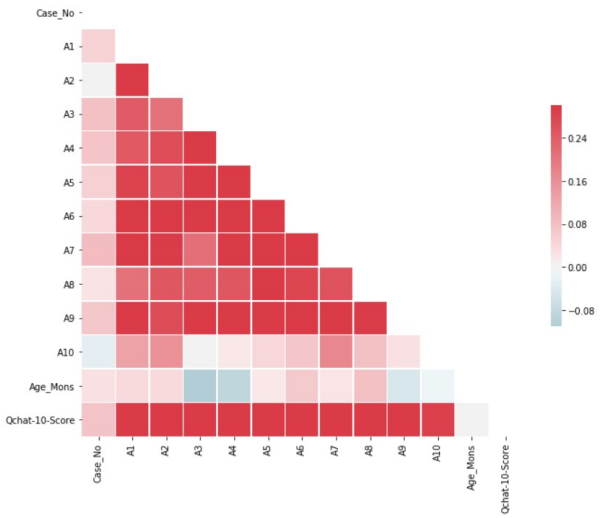

Fig. 2. Heat-map showing feature correlation

D. Train the Model

The goal of training is to answer a question or make a prediction correctly as often as possible.

Each iteration of process is a training step.

E. Evaluate the Model

Uses some metric or combination of metrics to ”measure” objective performance of model.

Test the model against previously unseen data.

F. Parameter Tuning

Tune model parameters for improved performance.

Simple model hyper parameters may include: number of training steps, learning rate, initialization values and distribution, etc.

G. Make Predictions

Using further (test set) data which have, until this point, been withheld from the model (and for which class labels are known), are used to test the model; a better approximation of how the model will perform in the real world.

III. RESULT ANALYSIS

Feature extraction gives the following results:-

A. Heatmap

B. Confusion Matrix

IV. CONCLUSION

This simulation shows us the results of both the ML algorithms: ’Logistic Regression’ and ’Random Forest’.

Linear models are composed of one or multiple independent variables that describes a relationship to a dependent response variable. Mapping qualitative or quantitative input features to a target variable that is attempted to being predicted such as financial, biological, or sociological data is known as

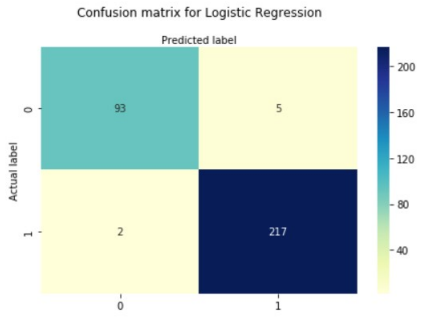

Fig. 3. Results for Logistic Regression

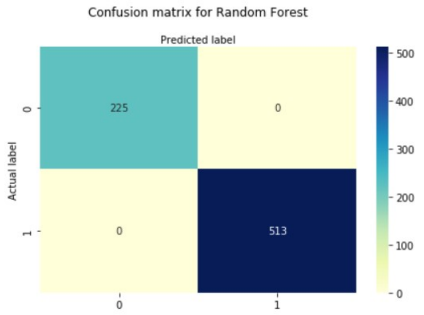

Fig. 4. Result for Random Forest

supervised learning in machine learning terminology if the labels are known. One of the most common utilized linear statistical models for discriminant analysis is logistic regression. This algorithm has poorer boundary mapping capability than the two. Logistic Regression gives up accuracy for model interpretability.

Random forest is an ensemble-based learning algorithm which is comprised of n collections of de-correlated decision trees. It is built off the idea of bootstrap aggregation, which is a method for re-sampling with replacement in order to reduce variance. Random Forest uses multiple trees to average (regression) or compute majority votes (classification) in the terminal leaf nodes when making a prediction. Built off the idea of decision trees, random forest models have resulted in significant improvements in prediction accuracy as compared to a single tree by growing ’n’ number of trees; each tree in the training set is sampled randomly without replacement. Decision trees consist simply of a tree-like structure where the top node is considered the root of the tree that is recursively split at a series of decision nodes from the root until the terminal node or decision node is reached. This algorithm has a very good boundary shaping capability. Random Forest sacrifices interpretability for accuracy of model.

So, applications where interpretation of the model is of importance, use Logistic Regeression. While Random Forest is used for better model fitting on the data.

ACKNOWLEDGMENT

We wish to acknowledge the help provided by Prof. Dayanand Ambawade for their patient guidance, constant support and encouragement and useful critiques to complete this project.

REFERENCES

[1] https://github.com/

[2] https://www.healthline.com/health/autism

[3] https://www.nhsinform.scot/illnesses-and-conditions/brain-nerves-and-spinal-cord/autistic-spectrum-disorder-asd

[4] https://towardsdatascience.com/the-7-steps-of-machine-learning

[5] https://www.datacamp.com/