****

**CLOUD COMPUTING**

**CSEE 5525-2020 SPRING**

**PROJECT REPORT**

**AUTISM DATA ANALYSIS**

**TEAM 7**

**Pavan Kumar Chongala**

**Naga Tanmaiyee Nagalla**

**Venkata Sreesudha Madabhushi**

**CONTENTS**

**MOTIVATION**

**DATA RESOURCES**

**OBJECTIVE**

**QUESTIONS**

**TOOLS**

**SYSTEM ARCHITECTURE**

**IMPLEMENTATION**

* **DATA VISUALIZATION**
* **STATISTICAL ANALYSIS**
* **WEBSITE DEPLOYMENT**

**CHALLENGES FACED**

**LEARNING OUTCOMES**

**CONCLUSION**

**LINKS**

* **WEBSITE**
* **GITHUB**
* **TRELLO**

**MOTIVATION**

ASD – Autism Spectrum Disorder is a neurological disorder with a wide range of conditions such as difficulty with social skills, communication skills, repetitive behaviors resulting in developmental disability. According to CDC 1 among every 59 are diagnosed with ASD.

There is no known single cause for ASD as well as a complete cure for the condition. The signs usually appear by age 2 or 3 and according to studies, ASD is more in men than female. Even though it cannot be cured completely, early diagnosis would lead to a better treatment to improve the quality of life for patients.

The ASD diagnosis procedures are expensive and time consuming. But diagnosis is the only key in order to provide any kind of treatment for the patients. In order to make the diagnosis process time- efficient and affordable by everyone, we need better screening procedures. This can be done extensive analyzing of the past data the tells about the trends, symptoms and many more factors and models can be built based on this analysis which ca be used for an efficient diagnosis.

**DATA RESOURCES**

* <https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult>.
* <https://www.manukau.ac.nz/about/faculties-schools/business-and-information-technology/more-information-for-students/lecturer-profiles/dr-fadi-fayez-thabtah>.

**OBJECTIVE**

* ASD Analysis
* To do analysis on the Autism Spectrum Disorder (ASD) data of both adults and toddlers’ datasets by visualizing the data with the help of most influential factors.
* Evaluation of Trends
* Diagnosis based on the behavioral trends.
* Determining Factors :
* Age
* Gender
* Ethnicity
* Jaundice
* Family history
* To deploy a website by pushing the obtained the information into a cloud.

**QUESTIONS**

The following are the queries that we considered for the data analysis and visualization of the ASD datasets

1. How does Jaundice from birth influences the ASD in adults and toddlers?
2. At what age does ASD usually appear in adults?
3. At what age does ASD usually appear in toddlers?
4. In what regions ASD cases are more?
5. How is ASD distributed based on gender in different ethinicities?
6. How is ASD distributed in family?

**TOOLS, LANGUAGES AND PLATFORM**

* Python
* Bootstrap
* Javascript
* HTML
* CSS
* GoogleColab
* Flask
* AWS S3 Bucket

**SYSTEM ARCHITECTURE**

Data collection

Python

Data Analysis and Visualization

Google Colab

Website Development

By HTML and Flask

AWS S3 Database

AWS EC2 Deployment

**IMPLEMENTATION AND VISUALIZATION**

**RETRIEVING THE DATA**

A screenshot of a cell phone

Description automatically generated

**READING AND DISPLAYING THE DATA**

A screenshot of a cell phone

Description automatically generated

**SUMMARIZING THE DATA**

A screenshot of a social media post

Description automatically generated

**DATA VISUALIZATION**

**ASD POSITIVE PATIENTS WITH JAUNDICE FROM BIRTH---ADULTS AND TODDLERS**

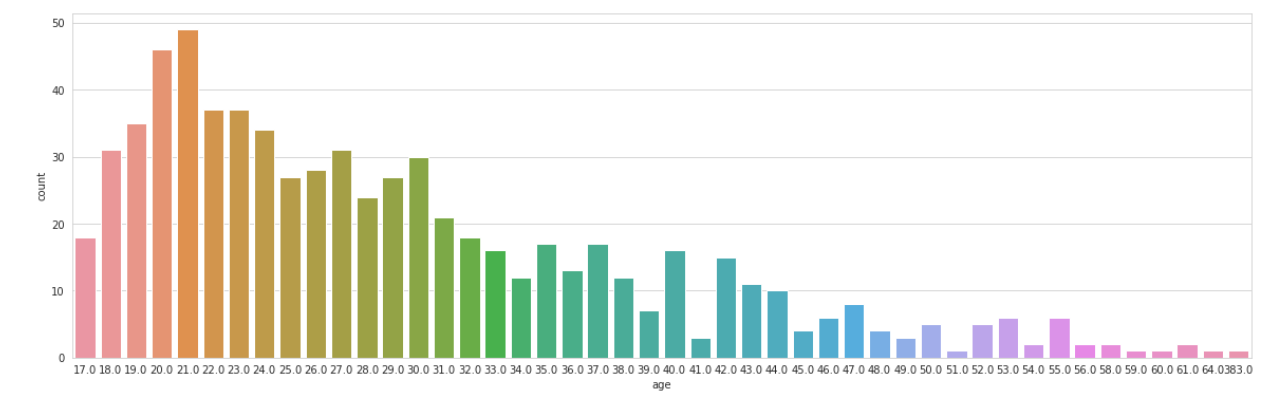
A screenshot of a cell phone

Description automatically generated

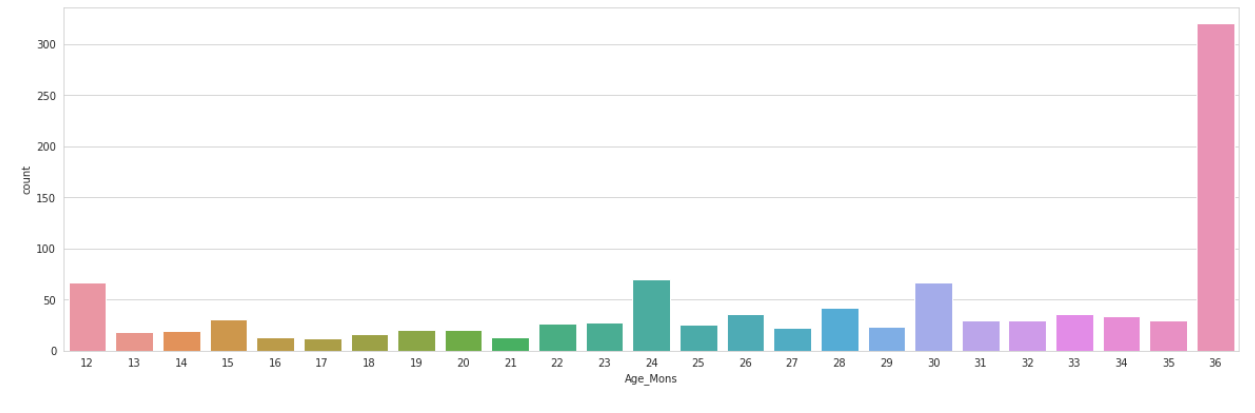
Adults: ASD is higher in females born with Jaundice.

Toddlers: ASD is higher in males (boys) born with jaundice.

**ASD POSITIVE BASED ON AGE - ADULTS**



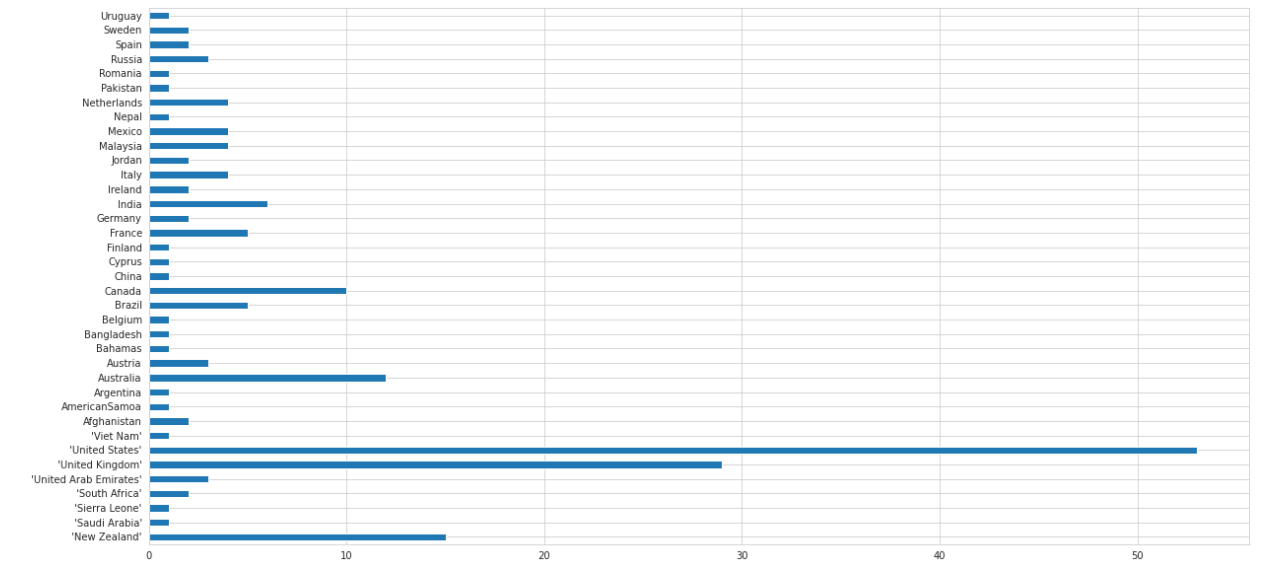
**ASD POSITIVE BASED ON AGE - TODDLERS**



Adults: ASD patients are more in number at the age ranging from 18 to 30, with peak at age of 20-21.

Toddlers: ASD is higher in kids of age 3, since the earliest signs can be seen at ages 2 or 3.

**VISUALIZATION OF ADULTS LIVING IN DIFFERENT CITIES**



Developed countries like USA, UK, Canada, Australia show higher cases of ASD with USA at top.

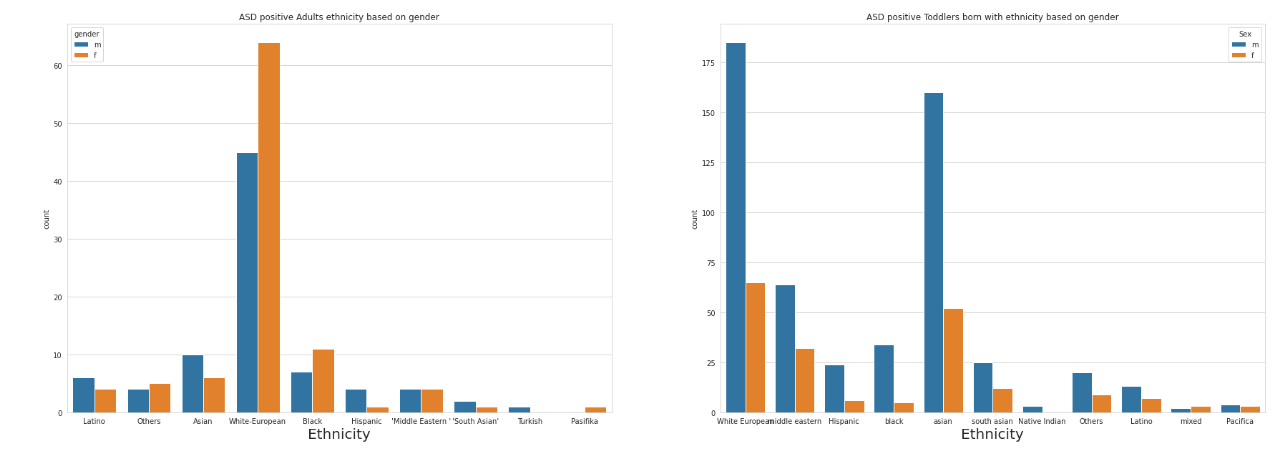
**ADULTS: ASD POSITIVE – COUNTRY WISE DISTRIBUTION**

A screenshot of a cell phone

Description automatically generated

With USA at top and UK at second, ASD cases are higher among males than females but in other countries females are more affected than males.

**ETHNICITY OF ADULTS AND TODDLERS BASED ON GENDER IN ASD POSITIVE**



**ADULTS: ASD POSITIVE PATIENTS OF WHITE AND EUROPEAN ETHNICITIES**

A screenshot of a cell phone

Description automatically generated

In general, ASD cases are more in USA compared to other countries.

**DISTRIBUTION OF AUTISM IN FAMILY WITHIN DIFFERENT ETHNICITY**

A screenshot of a cell phone

Description automatically generated

In both adults and toddlers, White-European ethnicities are at a higher probability of getting ASD if the there is a family history of ASD.

**STATISTICAL ANALYSIS**

**LOGISTIC REGRESSION**

A screenshot of a social media post

Description automatically generated

A screenshot of a social media post

Description automatically generated

**PREDICTION USING CONFUSION MATRIX**

A screenshot of a social media post

Description automatically generated

**RANDOM FOREST CLASSIFIER**

A screenshot of a social media post

Description automatically generated

**SCALING THE VARIABLES**

A screenshot of a cell phone

Description automatically generated

**KNN CLASSIFIER**

A screenshot of a social media post

Description automatically generated

**PREDICTION USING CONFUSION MATRIX**

A screenshot of a social media post

Description automatically generated

Three prediction models have been considered and compared in order to provide an efficient model which provides best accuracy.

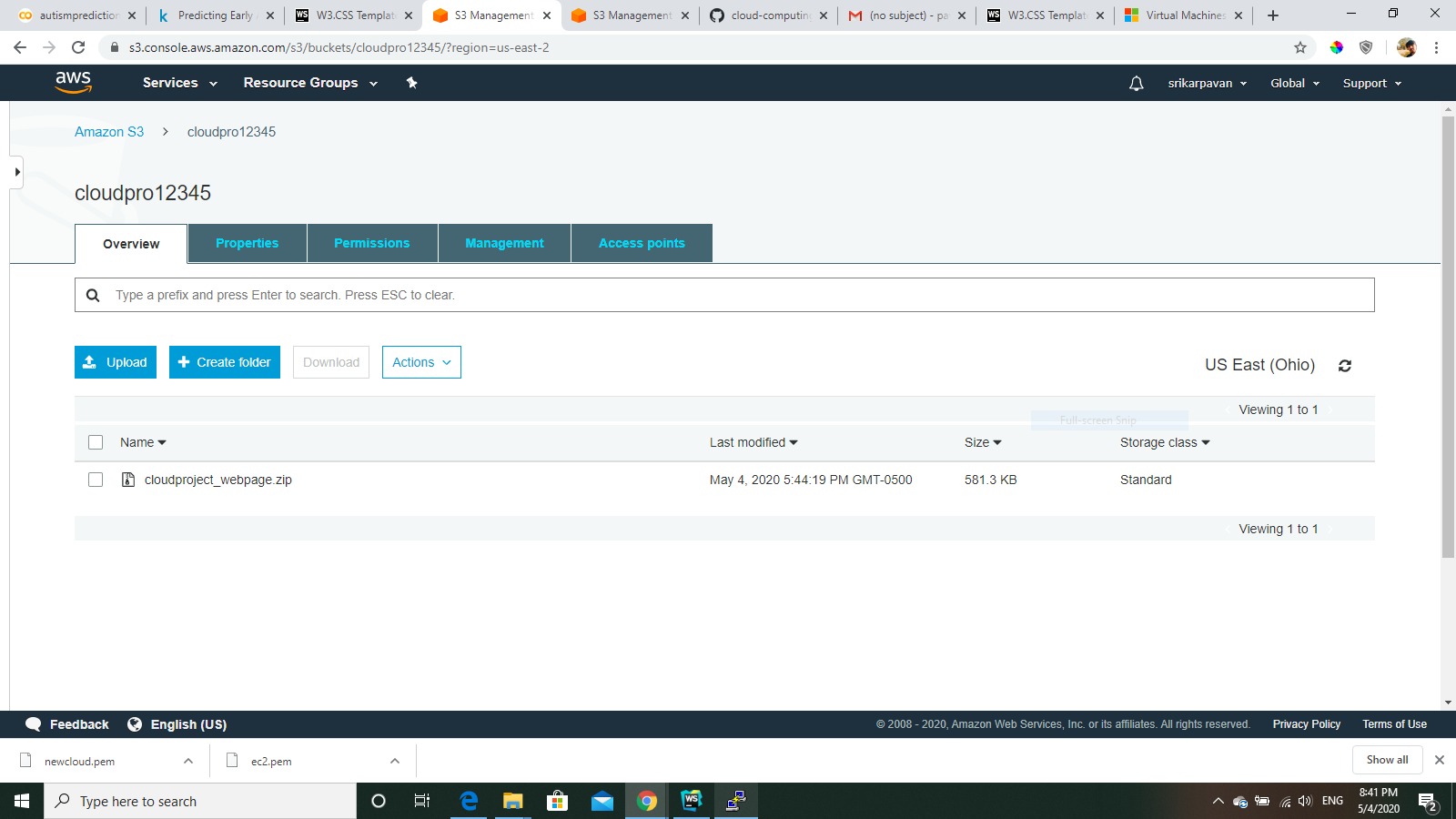
* **Logistic Regression:** It uses a logistic function for modelling of a binary dependent variable.
* **Model accuracy: 63%.**
* **Random Forest:** Decision trees are used by random forest model and gives the output, mode of the classes.
* **Model accuracy: 70%.**
* **KNN Classifier:** It uses ‘K’ nearest known neighbors and assigns the output to that class.
* **Model accuracy: 71%.**
* Among these 3 models, **KNN classifier is a better model** since the accuracy rate is higher than others.

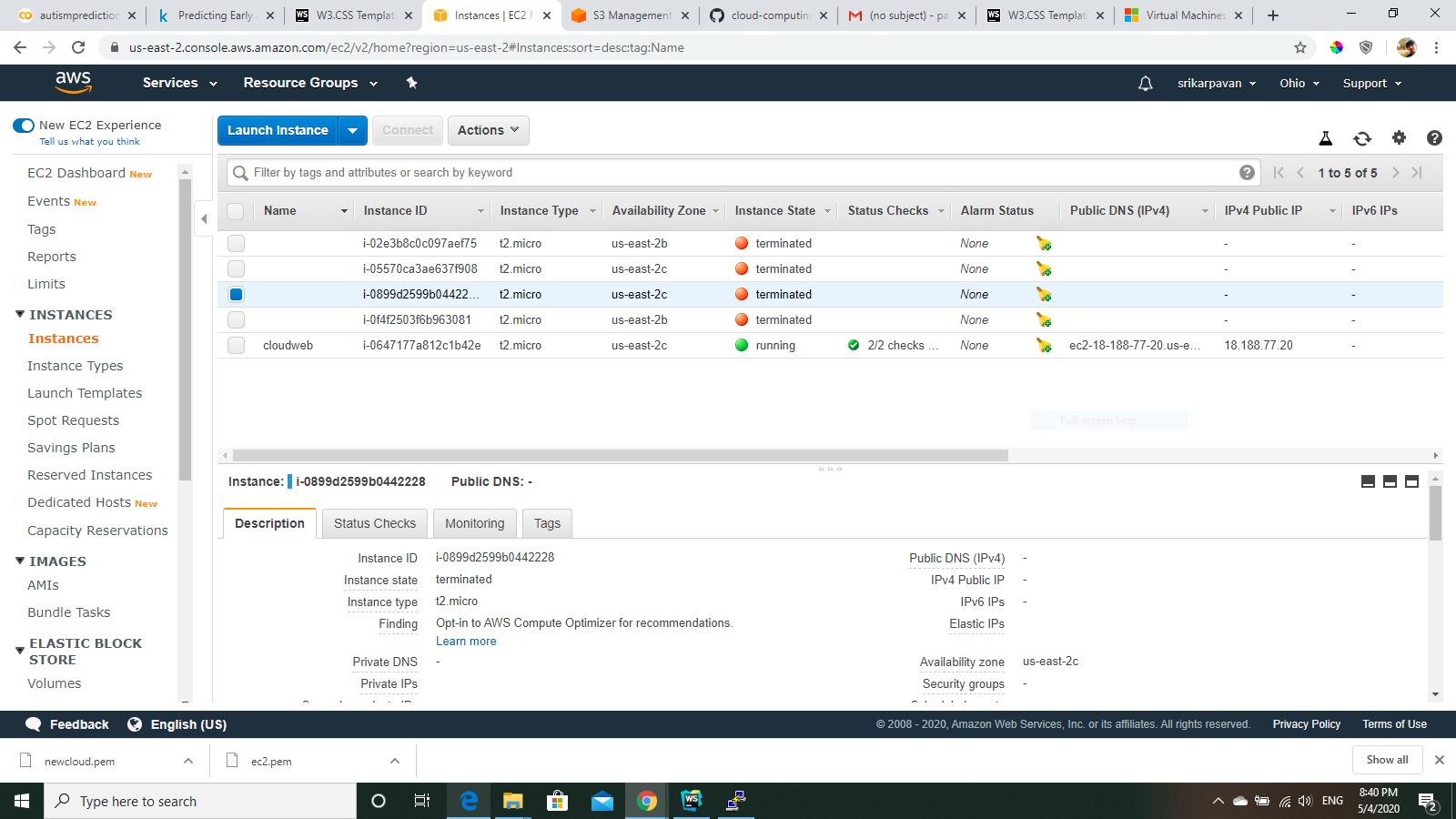
**WEBSITE DEPLOYMENT**

**(CLOUD)**

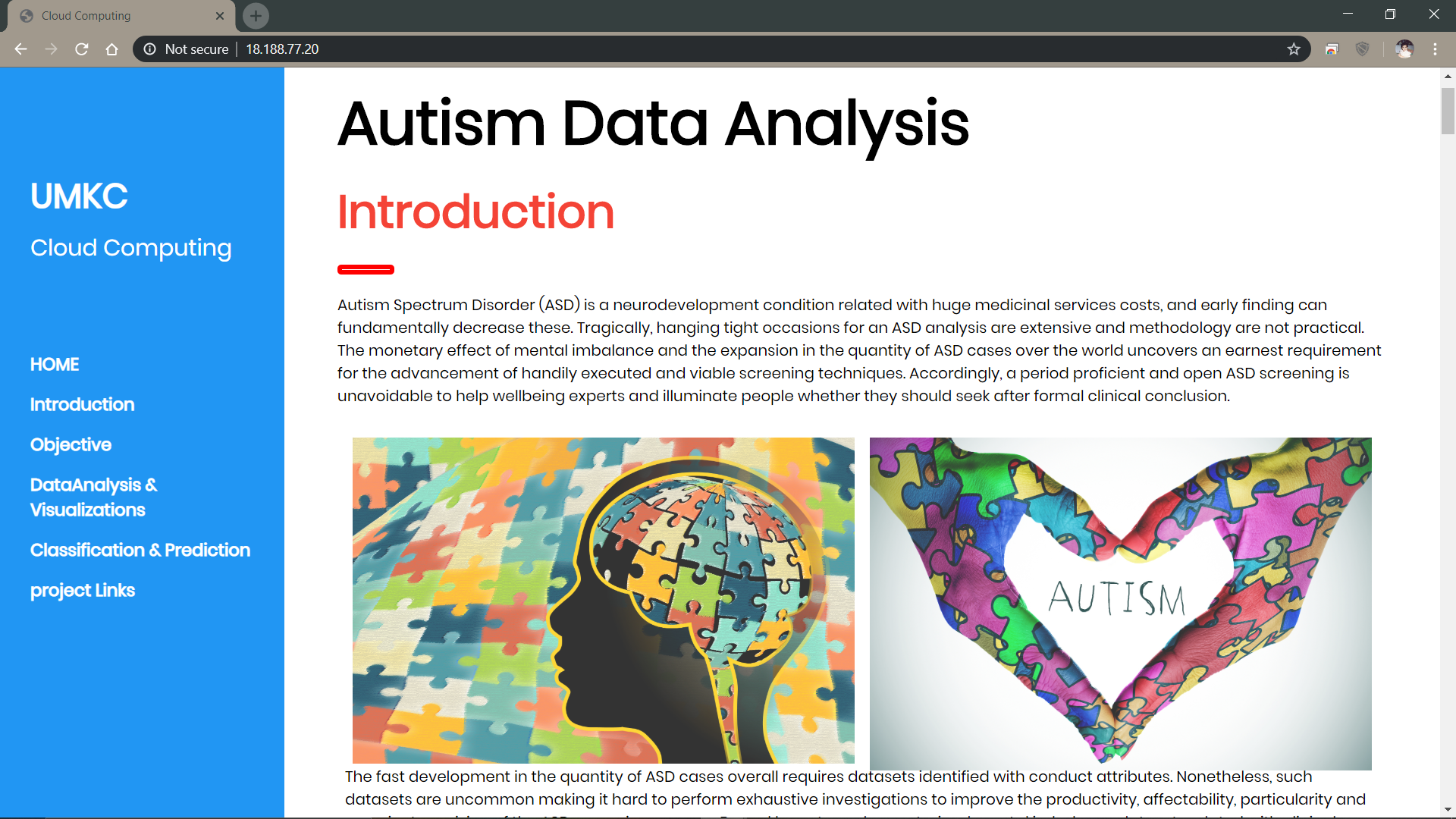
**AMAZON AWS S3 BUCKET**

* An **Amazon S3 bucket** is a public cloud storage resource available in **Amazon Web Services**' (**AWS**) Simple Storage Service (**S3**), an object storage offering.
* **Amazon S3 buckets**, which are similar to file folders, store objects, which consist of data and its descriptive metadata.
* It is highly scalable, fast, inexpensive storage, reliable and highly trusted database.





**WEBSITE**



**CHALLENGES FACED**

* During collecting the data.
* Prediction analysis.
* Deploying the website.
* Server problems during hosting the site.

**LEARNING OUTCOMES**

* Data analysis and visualization.
* Deploying the website using AWS3 Bucket.

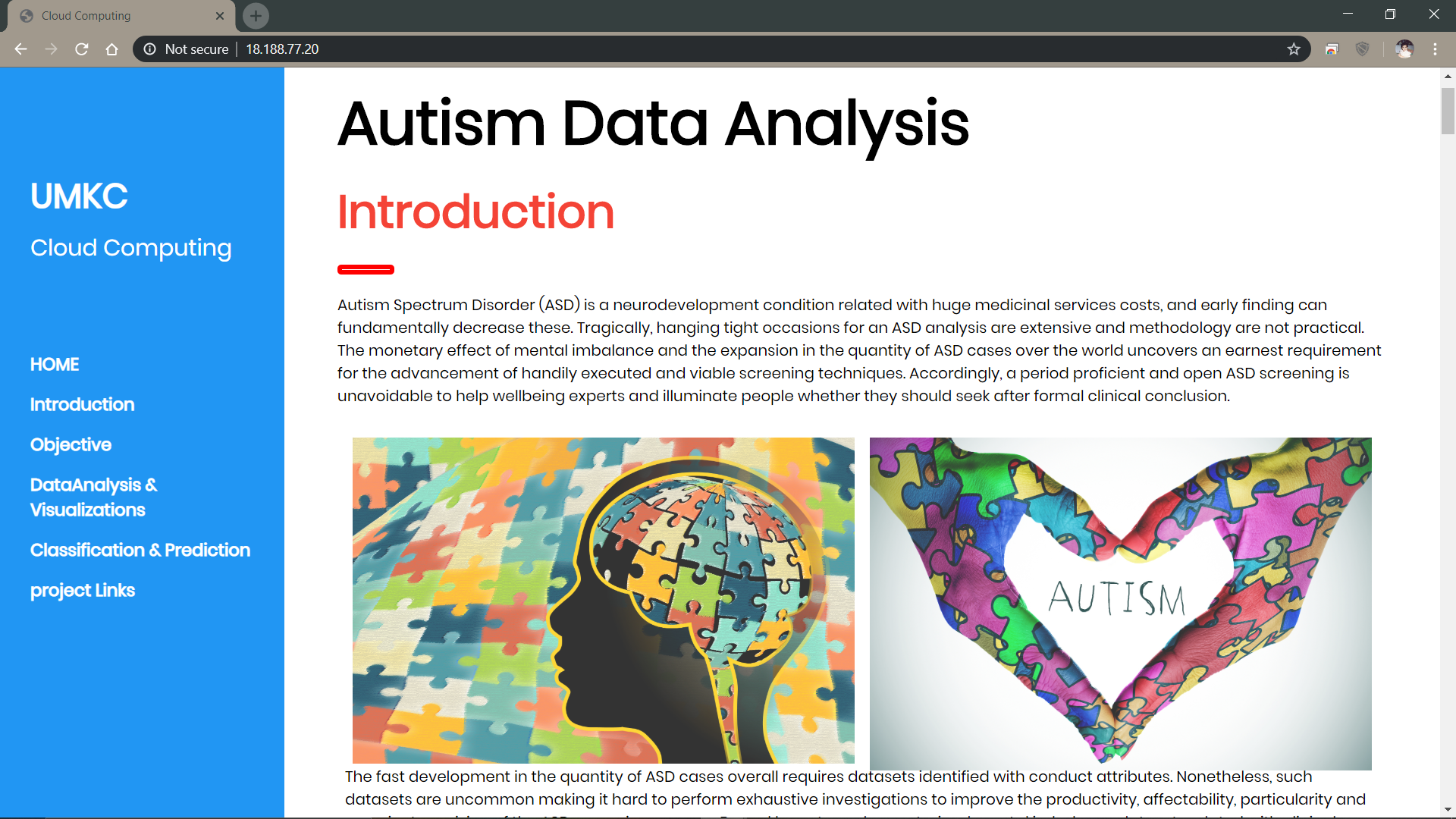
**CONCLUSION**

* We have done analysis on the Autism Spectrum Disorder (ASD) data of both adults and toddlers data set.
* We have done data visualizations based on the factors age, country of resident, family history and jaundice.
* We created a web site and pushed our data into the Amazon Cloud.

**LINKS**

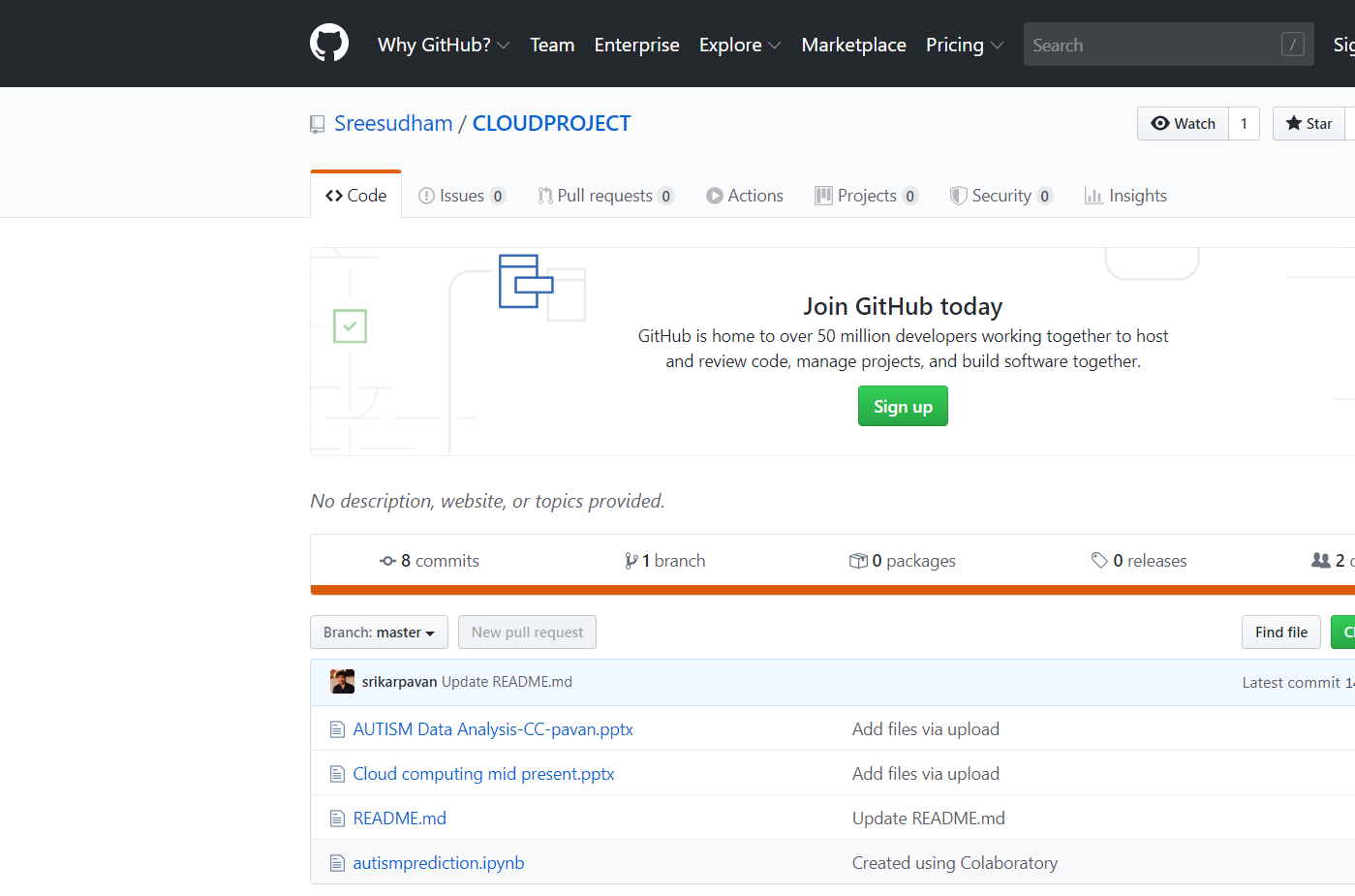
**WEBSITE**

WEBSITE LINK: <http://18.188.77.20/>



**GITHUB**

GITHUB LINK: <https://github.com/Sreesudham/CLOUDPROJECT>



**TRELLO**

TRELLO LINK: <https://trello.com/b/g1hUCEKs/autism-data-analysis>

