**Predicting Autism Spectrum Disorder using Supervised Learning Algorithms**

**TABLE OF CONTENTS**

**S.No Contents**

| 1. | Abstract |  |
| --- | --- | --- |
| 2. | Introduction |  |
| 3. | Literature Review |  |
| 4. | Problem Identification & Objectives |  |
| 5. | System Methodology |  |
| 6. | Overview of Technologies |  |
| 7. | Implementation |  |
|  | 7.1.Coding |  |
|  | 7.2.Testing |  |
|  | 7.3 Deploying Flask Application |  |
| 8. | Results & Discussion |  |
| 9. | Comparative Study |  |
| 10. | Conclusion & Future Scope |  |
| 11. | References |  |

1. **ABSTRACT**

An autistic spectrum disorder (ASD) is a neurological disorder, which can impair a person's language skills, speech, cognitive abilities, and social skills. The symptoms typically occur within the first two years after birth, that is, during the developmental stage, and it affects approximately 1% of the worldwide population. Although Autism Spectrum Disorder (ASD) is primarily due to genetics and environmental factors, it can be treated at an early age if detected and treated appropriately.

Most children with autism do not obtain an accurate diagnosis, and usually pass the point of not returning. There are times when guardians are reluctant to accept that their child's psychological development differs from his/her physical development. The delay in diagnosis prevents a child from receiving the assistance they need to develop. The earliest examination, screening, and assessment of a child's development is crucial for ensuring that an autistic child will receive adequate care and support to assist them in reaching their fullest potential.

To diagnose autism spectrum disorder (ASD) in the present day, clinical standardised tests are the only methods available. In addition to requiring a longer diagnostic time, it also involves a significant increase in medical costs.

Traditional procedures are being improved using machine learning approaches to increase efficiency and precision. ML algorithms like support vector machine (SVM) and Random forest classifier (RFC) were applied to our dataset to build predictive models. It is the primary goal of our study to determine whether a child is susceptible to ASD in its early stages, which will facilitate diagnosis.

We have utilised a step by step approach to break down the information of the past ten years. The anticipated data of patients with chemical imbalance and non chemical imbalance will be seen as new information and will be utilised for noticing results for approaching patients. Results were evaluated using the average value. The use of SVM and RF for ASD classification is shown to be effective. A 100% accuracy rate was achieved by the RF method especially, for all datasets.

It is crucial to diagnose ASD early. If the amount of information tests is sufficiently large, we can achieve a great level of precision for AI based ASD finding. RF attains the best exhibition for the ASD information order among the three arrangement strategies.

**2. INTRODUCTION**

Autism Spectrum Disorder is a neurobehavioral condition characterised by deficiencies when involved in any kind of social interaction or communication along with restricted or repetitive movements. Autistic individuals frequently exhibit alteration in their motor activities such as poor hand-eye coordination, loss in balance of self's body and gait patterns.

In recent years, the autism research community has become increasingly interested in motor impairments. Untreated motor issues in autism can lead to lifelong ill-effects be it physical, psychological, and behavioural. There is also a desire to understand whether movement characteristics can be used as early diagnosis, which is relevant to this particular work.

In current times with the steady progress and innovation technology a colossal amount of data can be stored for future study. Mining the relevant information is an essential errand to settle on choices in light of the gathered information. Lately, AI has procured numerous accomplishments and it is acquiring significance in applied sciences, for example, medication and biomedicine. In clinical direction and diagnostics, AI techniques are utilised or prescribed to help information understanding. In this way, techniques for screening sicknesses with the guide of AI are extensively contemplated.

Diagnosis of Autism is possible at any age, but such delay can have lasting ill-effects. The initial symptoms of the ASD conditions begin appearing early on in an individual's childhood and continue into their adolescence and later in their adulthood. There has been an increased adoption of machine learning techniques in the medical research domain in recent times. This project attempts to investigate the possibility of using Random forest, AdaBoost, and Support Vector Machine for detecting and analysing ASD symptoms in children, adolescents, and adults. The proposed methods are tested using a publicly available non-clinical ASD dataset.

In this study, we investigated whether development attributes estimated during impersonation can be utilised to separate individuals with ASC from neurotypical controls utilising AI (ML) techniques. Rather than utilising emotional scores acquired from conduct tests and surveys, we were keen on the attainability of utilising painless quantitative measures to equitably evaluate contrasts and arrange chemical imbalance from neurotypical controls. We present an information driven approach and AI based techniques to resolve essential issues with respect to distinguishing expressive test conditions and elements from high-layered information and regularly when just a small size of tests are accessible.

**3. LITERATURE REVIEW**

The course of early screening isn't extremely reliable in light of the fact that it depends on a questionnaire that guardians reply about their ward's way of behaving (usually at the child’s 18 month check-up). Furthermore, these surveys regularly produce misleading false positives. Truth be told, of the kids whose guardians report early indications of ASD on the poll, just 50% have that diagnosis affirmed by an authorised ASD clinician. Also, in light of the fact that there are scarcely any authorised ASD clinicians qualified to circle back to the many guardians who report through ASD through the poll, the wait time that child will get a diagnosis could be well after the child's third birthday postponing treatment past the critical window of time to further develop results for child's with ASD possibly.

1. M.S. Mythili, A.R. Mohammed Shanavas made use of classification techniques in their study on ASD. The objective of their research was to study the autism disorder and the various levels it exists in. In this study the subject’s behavioural and social interactions were analysed with the use of SVM, Neural network and Fuzzy techniques with WEKA tools.
2. Wall made use of Alternating Decision Trees (ADTree) to reduce the time takes to screen and detect ASD traits in a subject in an effort to make the diagnosis procedure significantly faster. They observed higher levels of accuracy with a data of 891 individuals by using a revised ADI-R method. Although the test failed to accurately predict ASD in varying age groups since the study was limited within the age bracket of 5 years to 17 years.
3. A searching method was proposed by A. Kosimchi, V.Suchat, M.Duda and D.P. Wall over the set of traits in Autism prediction. The authors used a ML approach in this study for the evaluation of the clinical diagnosis of ASD. This study employs 8 different ML algorithms to perform step-by-step backward recognition on 4540 sheets.
4. Bone used a Support Vector Machine in their study, which included 1264 people with ASD traits and 462 people without ASD traits. In this study they obtained 89.2% percent sensitivity and 59% specificity.
5. From the preceding section, it is evident that deep learning-based models have a great potential to detect ASD in human populations. The majority of works cited so far have relied on traditional machine learning algorithms, which are limited in terms of performance. In this study, multiple machine learning models were compared to see which was the most efficient.

**4. PROBLEM IDENTIFICATION AND OBJECTIVES**

Autism Spectrum Disorder (ASD) is gaining traction quicker than ever before. Autism spectrum disorder is a neurodevelopmental disorder that impacts a person's ability to interact, communicate, and learn. Autism features can be detected by screening tests, however they are very costly and time taking process. Autism is now usually identified at an initial stage with advances in Artificial Intelligence and machine learning (ML).

The issue, as previously outlined, occurrence of ASD, is increasing day by day ,despite little recognition and effort being put into addressing it. With an ever-increasing number of cases of ASD in children, we need to devise a method for easing ASD aggravation. We intend to use this review to break down the barriers to ASD identification information and to assist future studies by sharing our findings.

**4.1 Background**

* According to the CDC’s (ADDM) Network, around 1 in 44 individuals has been diagnosed with autism spectrum disorder (ASD).
* ASD has been observed in people of various races, ethnicities, and economic groups.
* ASD affects more than four times as many boys as it does girls.
* During the study period of 2009-2017, about 1 in 6 (17%) children aged 3–17 years were recognized with a learning disorder, according to parents' reports. The observations  are behavioral problems.

Source: <https://www.cdc.gov/ncbddd/autism/data.html>

**4.2 Treatments and early intervention**

* There are numerous therapy options available. Although some therapies require more than one strategy, these treatments can typically be divided into the following categories:
* Behavioural, Developmental, Educational, Social-Relational, Pharmacological, Psychological, Complementary and Alternative Medicine.

**4.3 Aims and goals of the project**

In this work, we wish to present a faster, more simple machine learning-based method that may be used to identify ASDs in the future and detect the associated traits that cause the ASD. Machine learning has the potential to improve symptomatic and treatment research in the social sciences, and it could be especially useful in studying the widely prevalent and heterogeneous condition of ASD. As more advanced computing and designing methods have been used to solve the difficulties of cross-subject applications, machine learning has shown promising results in identifying a variety of clinical side effects, dramatically boosting the likelihood of a cure for a large number of patients.

Applying the right model is critical because an increasing number of young people are experiencing symptoms, and there will be a significant amount of middle-age psychological disorder in the future, and we need to identify the adverse effects ahead of time so that doctors can treat them.

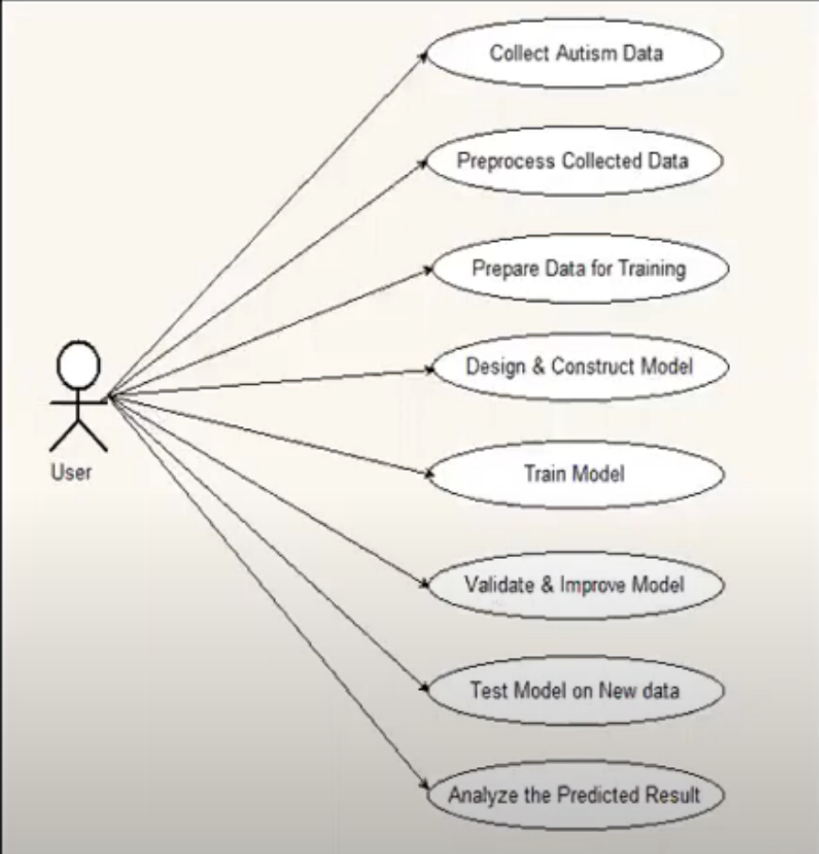
**4.4 Ethics of the project**

Ethics is a wide term that refers to a collection of standards that regulate our expectations as well as our personal as well as others' actions. For starters, it promotes the ideals of  respect and equality that are necessary for collaborative work. Secondly, it requires that professionals can face consequences for their activities. Finally, strong research ethics promote crucial social and moral ideals, such as not harming others.

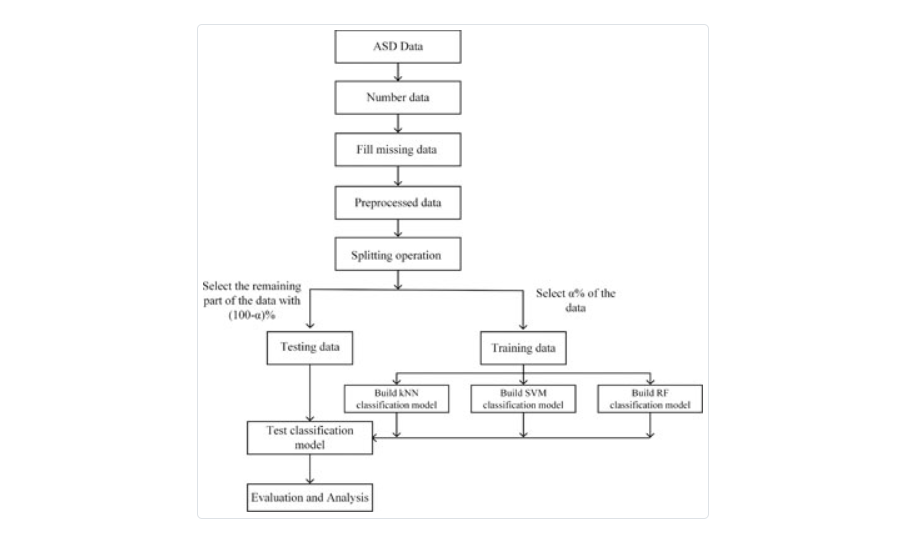
The dataset that we are using in this project is easily acquired from the UCI machine learning repository ensuring the safety of the exploration members. The goal of this project is to expedite the investigation of chemical imbalances and to make the ASD community more inclusive and welcoming. If any moral concerns arise in our venture, we will first speak with those who support our initiative while disregarding any morality, and then handle the problem.

**5. SYSTEM METHODOLOGY**

**UML Diagram:**

****

**Figure 5.1 UML Diagram, depicts the activities that take place in the course of implementing the project.**

**Architecture:**

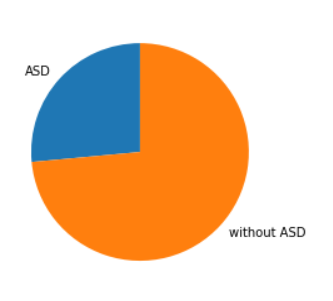
Build AdaBoost

Classification model

**Figure 5.2 Architecture of the project**

In this dissertation, five phases are involved: data collection, data pre-processing, developing the prediction model, assessing the prediction model, and developing an application in Flask. The following subsections briefly describe each phase:

**5.1 DATA COLLECTION**

 The dataset for this study was gathered from the UCI Repository, which is open to the public. Dataset Name-ASD Screening Data for Adult, Attribute Type-continuous and binary, Number of Instances-704. There are sixteen characteristics in the dataset, which are a combination of categorical and numerical data of which include: Question 1-10, Age, Gender, if the person born with jaundice, Any Family member with ASD, who is completing the test, and Class. The AQ-10 screening questions cover a variety of domains, including attention switching, imagination, communication, and social interaction. The questions are graded using a one-point scoring system for each of the ten questions. On each question, the user might earn 0 or 1 point depending **on their response.**

**Figure 5.3 Classification of Dataset**

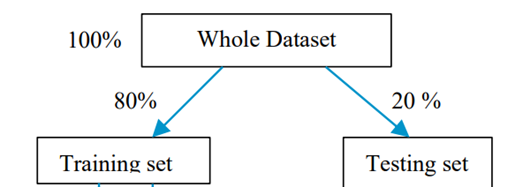
**5.2 DATA PRE-PROCESSING**

Data Pre-processing is for removing outliers and noise in the raw data and makes it available for training the model. In simple way, Data Pre-processing is the major step in the project evaluation to obtain the best accuracy. Thus, raw data is converted into something usable and understandable. Real-world data is frequently partial and inaccurate because it contains many errors and outliers. Several methods are to handle data, such as handling incomplete data, outlier analysis, data reduction, and discretization. The missing values in these datasets were solved using the imputation method.

**5.3 DEVELOPING PREDICTING MODEL**

Considering an 80:20 ratio, the complete dataset is categorized into training set and testing set respectively as per requirements. In the wake of implementing several types of supervised learning systems such as random forests (RF), SVM, and AdaBoost, a system to predict ASD could be put into place.

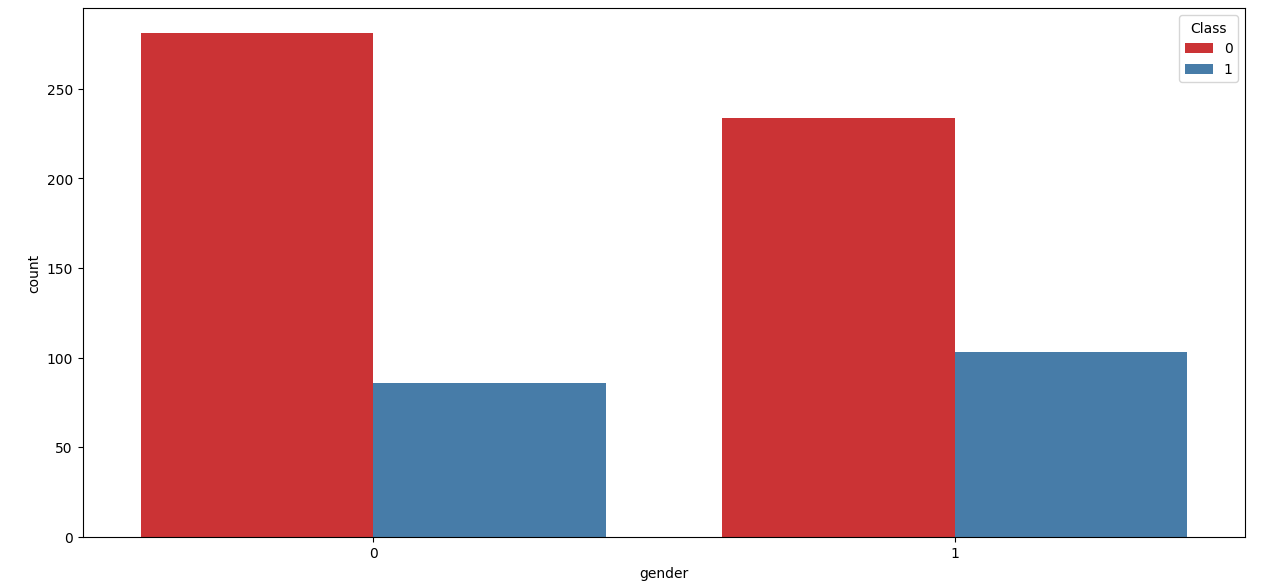
Algorithms were created and their accuracy was tested in order to generate autism trait predictions. Following the completion of types of supervised learning like random forests (RF), SVM, AdaBoost. Random forest is used for implementing the ASD predictive system.



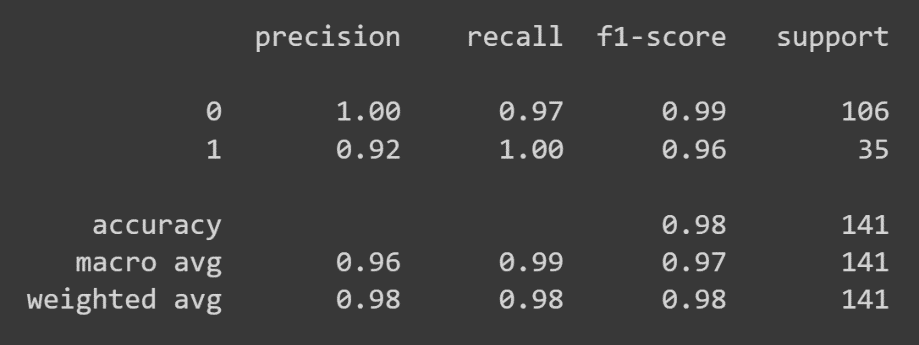
**Figure 5.4 Division of dataset for training and testing**

**5.4 EVALUATING THE PREDICTIVE MODEL**

In addition to assessing accuracy, sensitivity, specificity and precision, the proposed model was also tested using the leave-one-out strategy on the AQ-10 dataset. As part of the validation process, field observations were conducted at various places using forms to collect over 189 ASD cases and 515 cases without ASD from a special education institute for people with special needs.



|  |
| --- |
| **Figure 5.5: The person is autistic or not with respect to the gender** |



**Figure 5.6 The results of the Random Forest Algorithm with various classification metrics**.

|  |
| --- |
| **Figure 5.7 The results of the SVM Algorithm** |
| **Figure 5.8 The results of the AdaBoost Algorithm** |

The precision, recall, f1-score and support of the random forests algorithm are shown in the above figure.

**6. OVERVIEW OF TECHNOLOGIES**

**6.1. IMPLEMENTATION TECHNOLOGIES**

**6.1.1 PYTHON**

* Python is an interpreted and simple programming language.   This is identical to the programming languages PERL and PHP. Python is platform independent, which means that programs are written on one system can be used on the other without (or with little) changes.
* Python is an excellent language for beginners.
* Python has an extensive selection of libraries and frameworks.

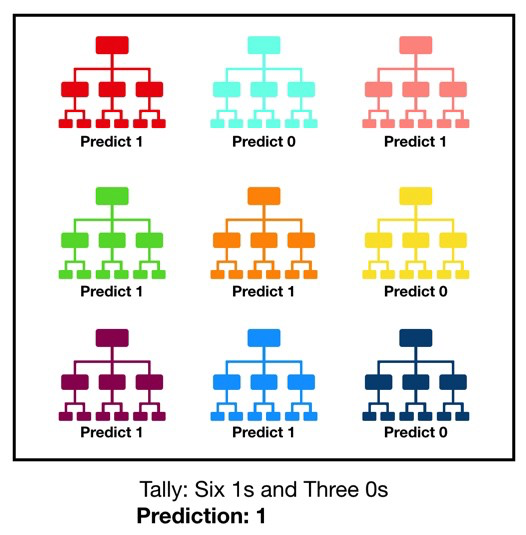
**6.1.2 PYTHON LIBRARIES**

* + NumPy or Numerical Python is an array of multi-dimensional array objects.
  + Panda: Gives us a set of tools to perform data analysis.
  + Matplotlib: It is a simulation framework in python used for two-dimensional plotting.
  + Scikit-learn offers several supervised and unsupervised Machine Learning.

Several machine learning algorithms have been implemented to achieve the final results.

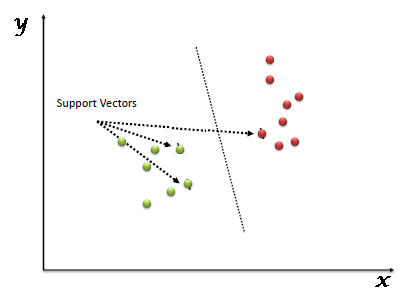
**6.1.3 RANDOM FOREST CLASSIFICATION**

The best predictors from a subset of predictors chosen at random are used to split each node in a random forest. This seemingly paradoxical method outperforms other classifiers such as SVM, and neural networks. Random Forest classifier was used to improve the model's accuracy. The algorithm can be divided into two steps in this case as well: producing random forest and classifying test data.



**Figure 6.1 Visualisation of a random forest model making prediction**

**6.1.4 SUPPORT VECTOR MECHANISM**

A SVM model (Support vector) is machine learning method used for regression and classification that study and examines data. SVM is a supervised learning method which divides the data into two groups based on the hyperplane N dimension. N represents the total of different features used to classify the data points.

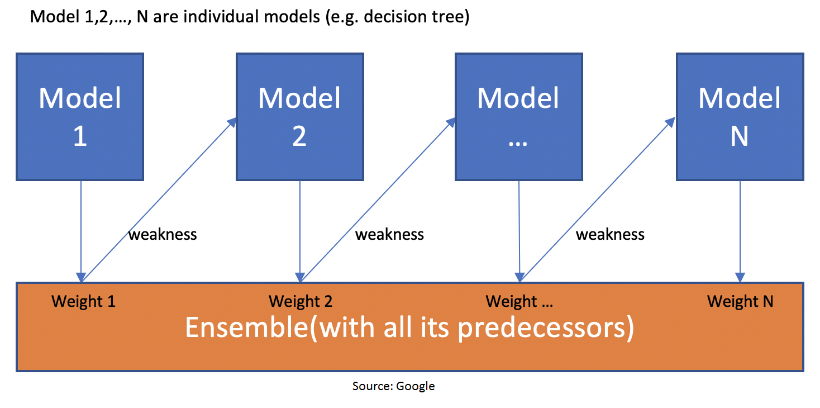
**Figure 6.2 SVM plotting**

**6.1.5 Adaboost algorithm**

One of several boosting algorithms AdaBoost is first boosting algorithm in solving practices. It is used to merge several "weak classifiers" into one "strong classifier."

Weak classifiers are introduced one by one and trained on a weighted dataset. After creating a predetermined number of weak learners (a user parameter), the process is repeated until the training dataset cannot be improved anymore. AdaBoost is frequently referred to as the best out-of-the-box classifier (with decision trees as the weak learners).

 AdaBoost algorithm offers the advantage of predicting more accurate results and has better performance than the current framework.



**Figure 6.3 Working of AdaBoost**

**6.2 REQUIREMENT ANALYSIS**

**6.2.1 Non-Functional Requirements**

* Reliability: During runtime, the system will always be up to perform the requested task 99.9% of the time except for some abnormal disruptions.
* Performance: Numerous users should be able to use the System at all time
* Maintainability: Keeping an eye on and administering the programme is critical, and it should be methodical in its approach.
* Portability: The system should be able to seamlessly transition to another application.
* Scalability: The system will be versatile enough to add new features in the future.
* Flexibility: A system's ability to respond to changes in its surroundings and adapt to changes in market tactics and rules is referred to as flexibility.

**6.2.2 Functional requirements**

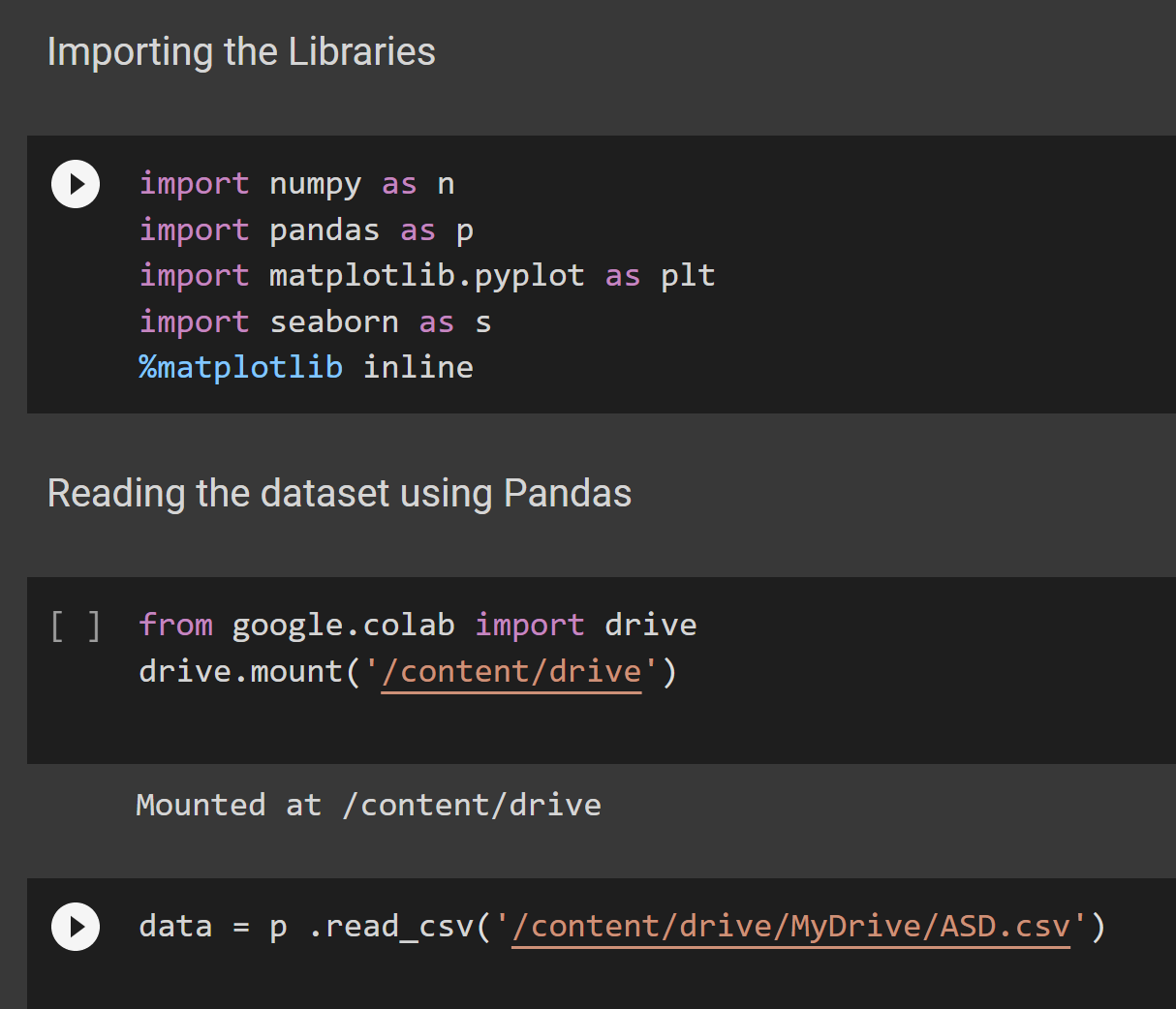
One of several boosting algorithms AdaBoost is first boosting algorithm in solving practices. It is used to merge several "weak classifiers" into one "strong classifier."

Weak classifiers are added one by one, and they are trained with weighted dataset. After creating a predetermined number of weak learners (a user parameter), the process is repeated until the training dataset cannot be improved anymore. AdaBoost is frequently referred to as the best out-of-the-box classifier (with decision trees as the weak learners).

AdaBoost algorithm offers the advantage of predicting more accurate results and has better performance than the current framework.

**7. IMPLEMENTATION**

#### **7.1 Coding**

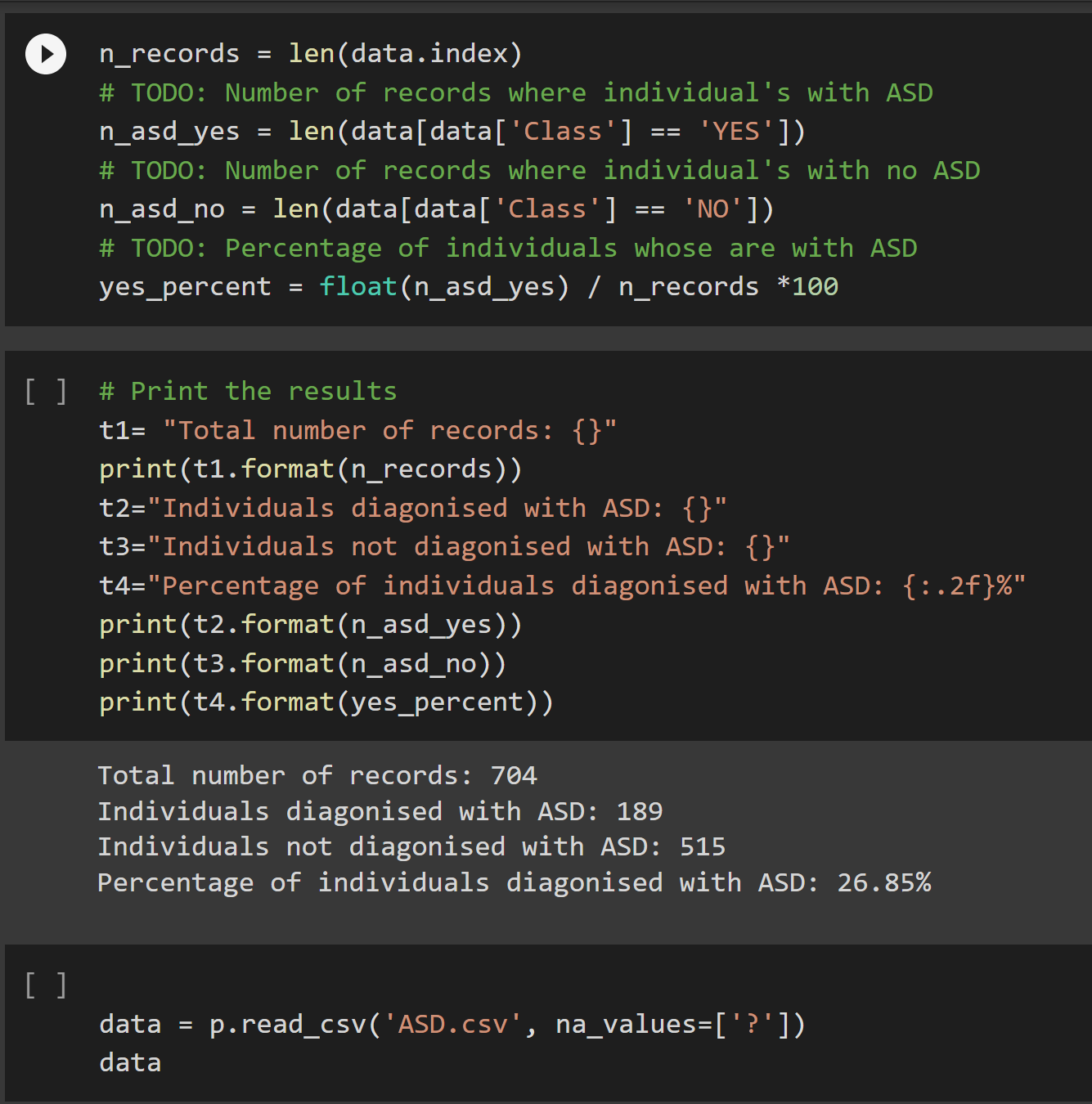
****

Imported the required libraries such as :

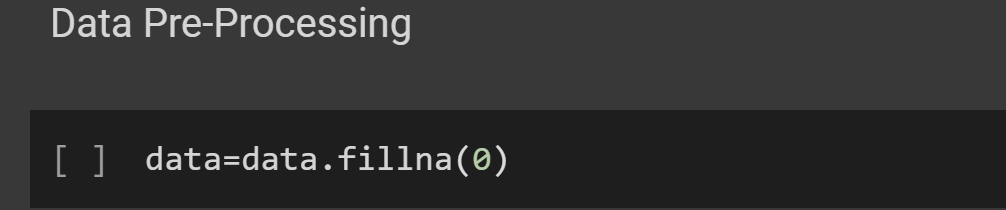
* Numpy: it is used to work on high-level mathematical functions.
* Pandas: Helps us to manipulate data and also analyse data.
* Matplotlib: used to represent the data in the form of graphs, pie charts etc.,
* Seaborn: it is an data visualization tool based on matplotlib.

We have executed our project in Google collab in order to complete and divide work over the internet easily. So Dataset is uploaded in the google drive. Thus, mounting the drive is needed to retrieve the dataset.

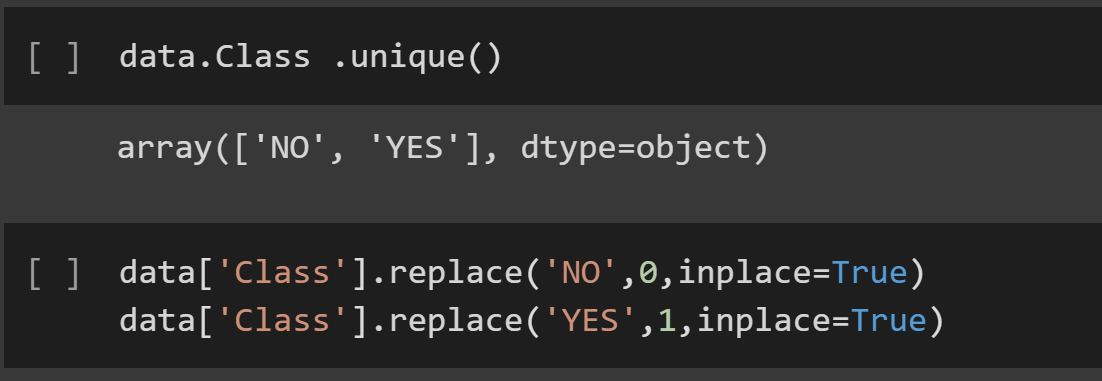
Then using pandas, read\_csv method is used to read the data and followed by the location/ path of the file.

****

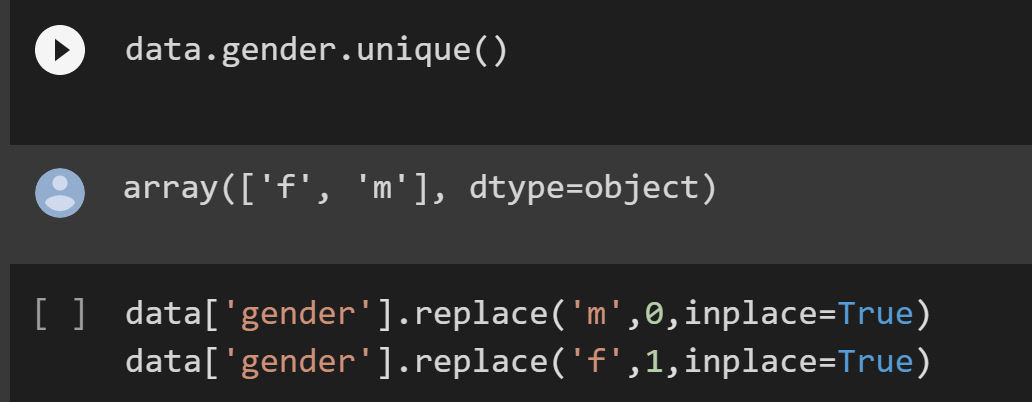
Printing the no of records and patients with & without ASD in the used Dataset.



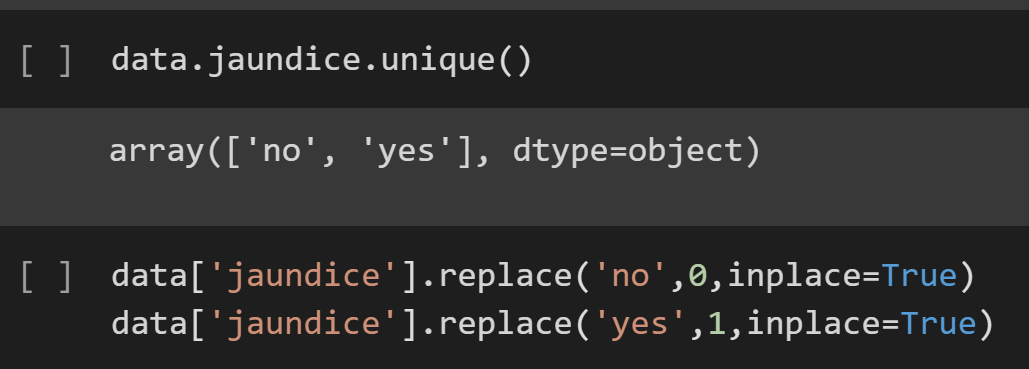
“Nan” values are replaced using zeros using fillna() method.



Unique elements in the column “Class” is found using unique() method and replaced into binary values.



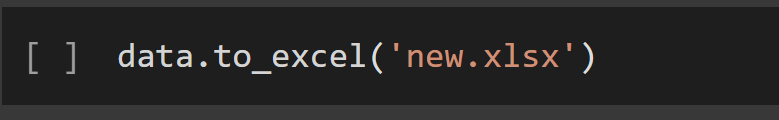
Unique elements in the column “gender” is found using unique() method and replaced into binary values.

****

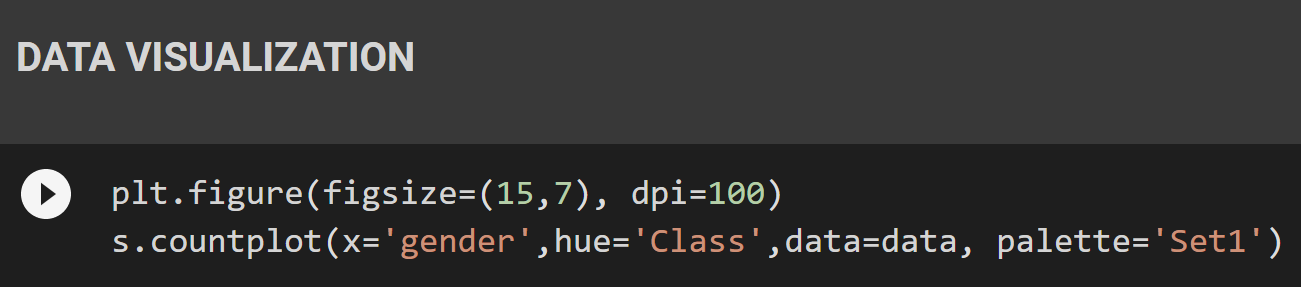
Unique elements in the column “jaundice” is found using unique() method and replaced into binary values.

****

In the same way using unique method, all the columns are converted into numerical values in order to make execution simple.

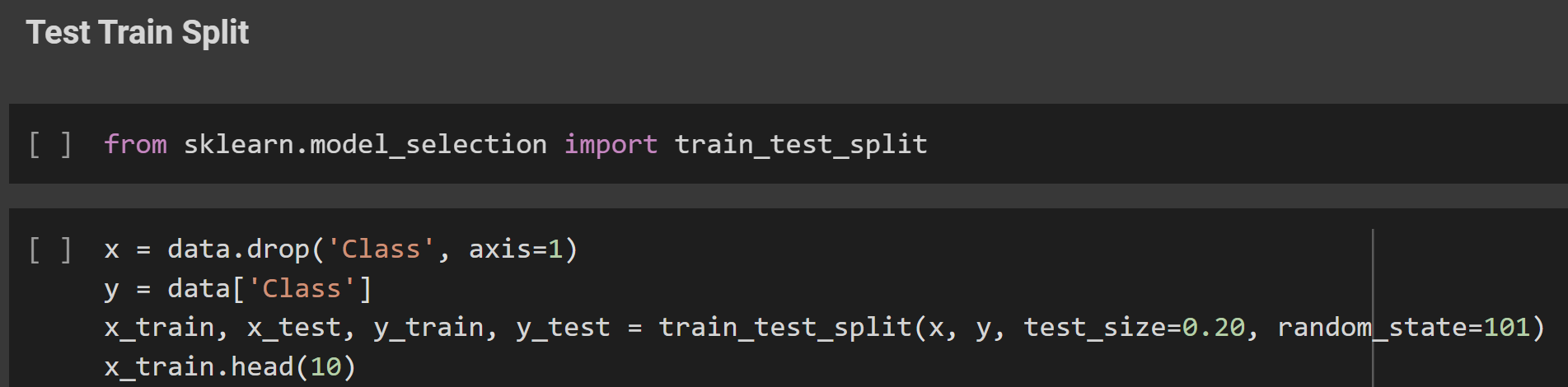
****

Creates a new excel sheet of new numerical dataset.

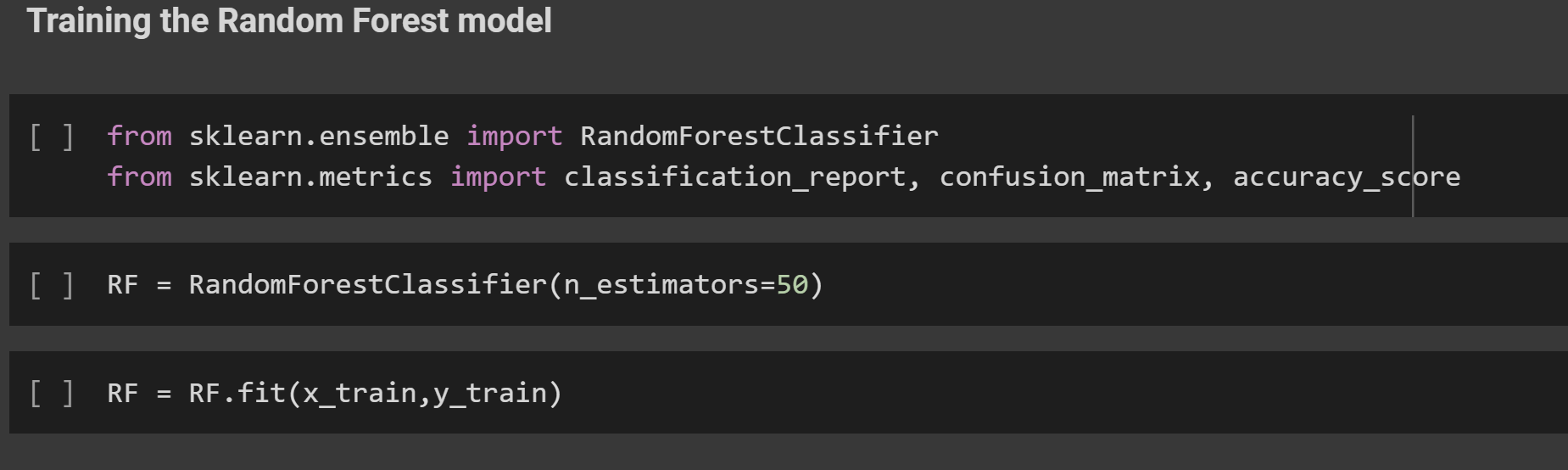


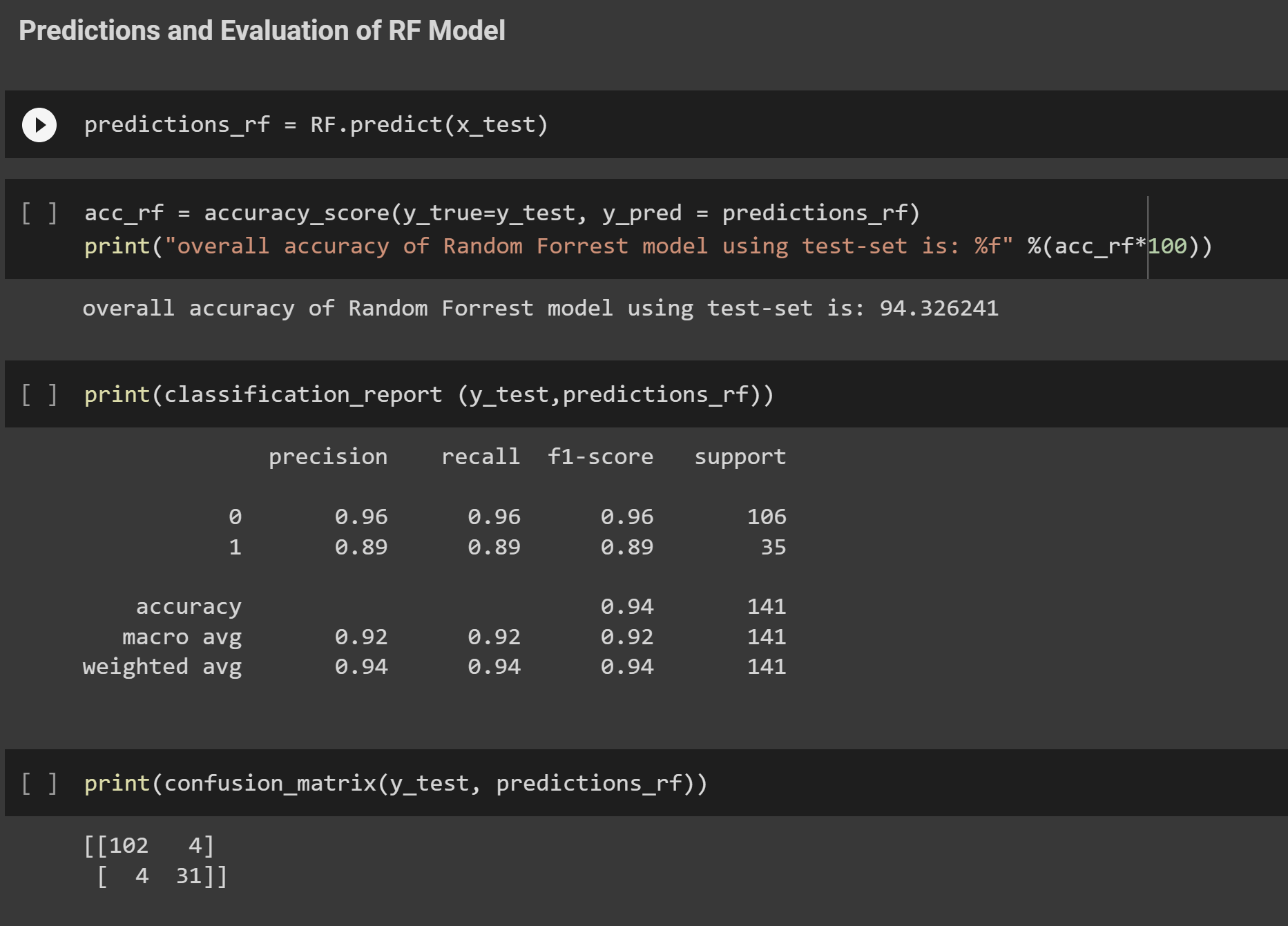
Data visualization done using seaborn library.

**Spliting the Dataset using train\_test\_split method in sklearn:**

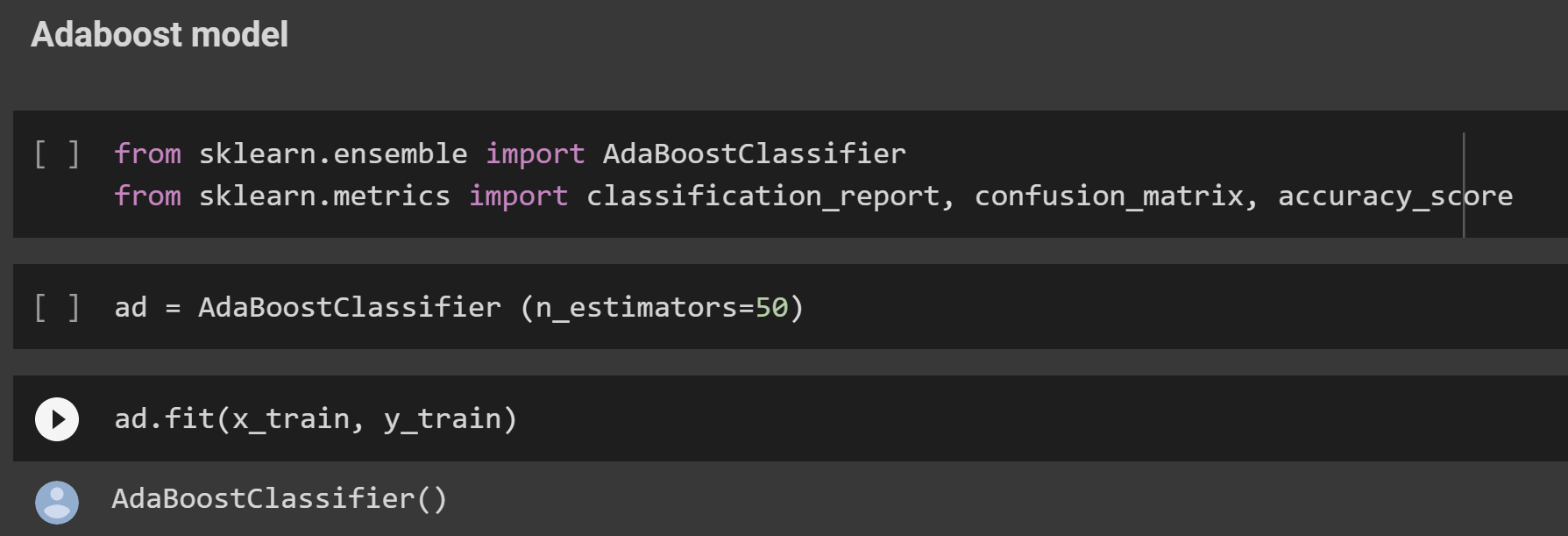
****

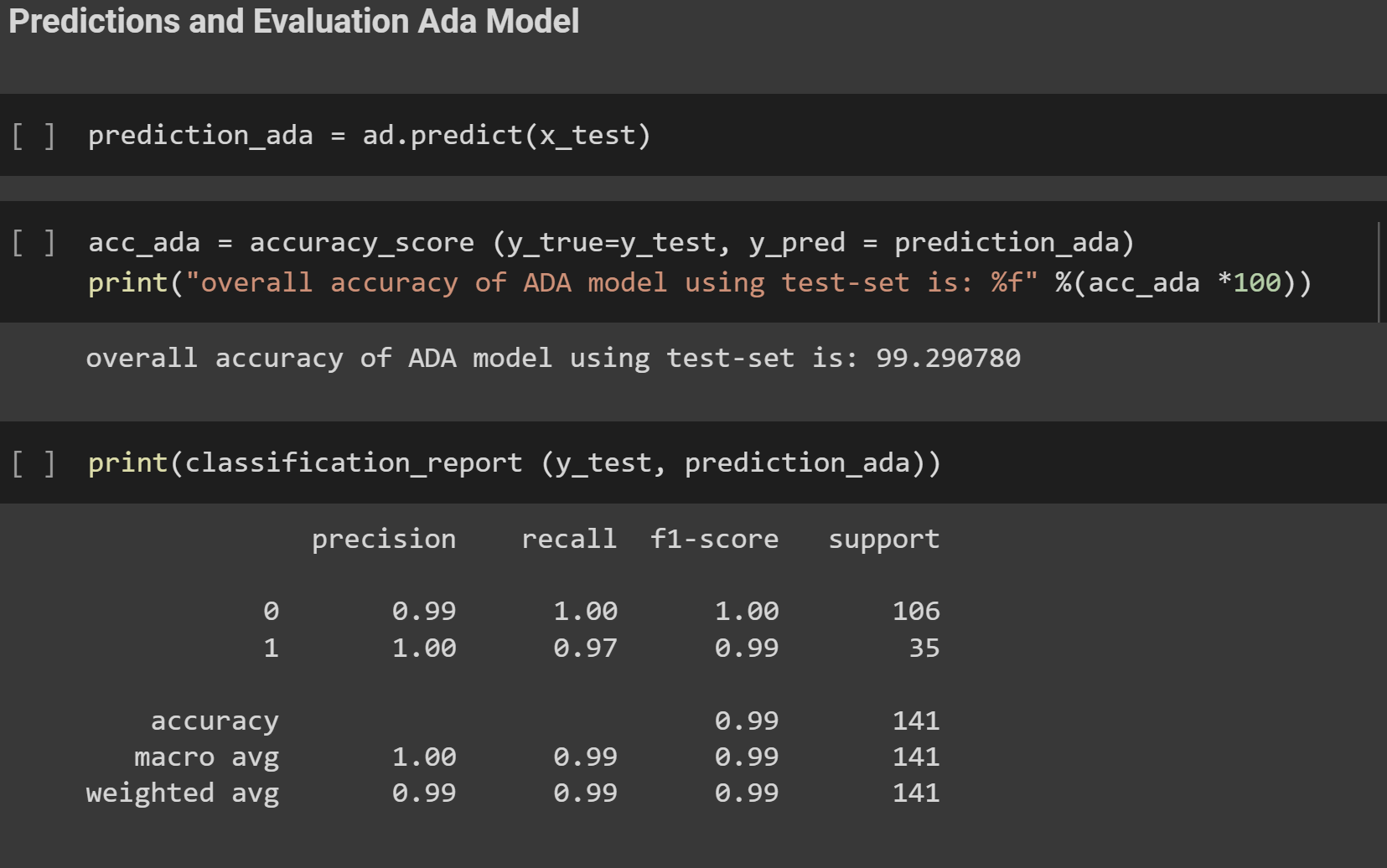
**Random Forest:**

****

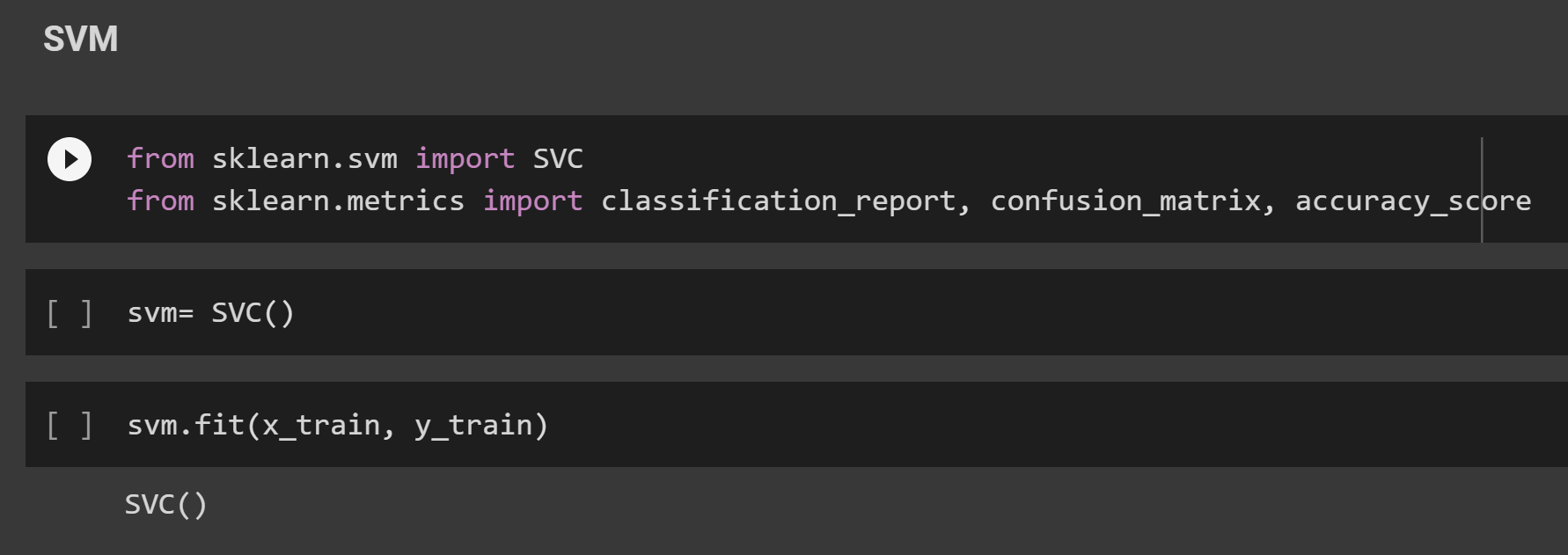
****

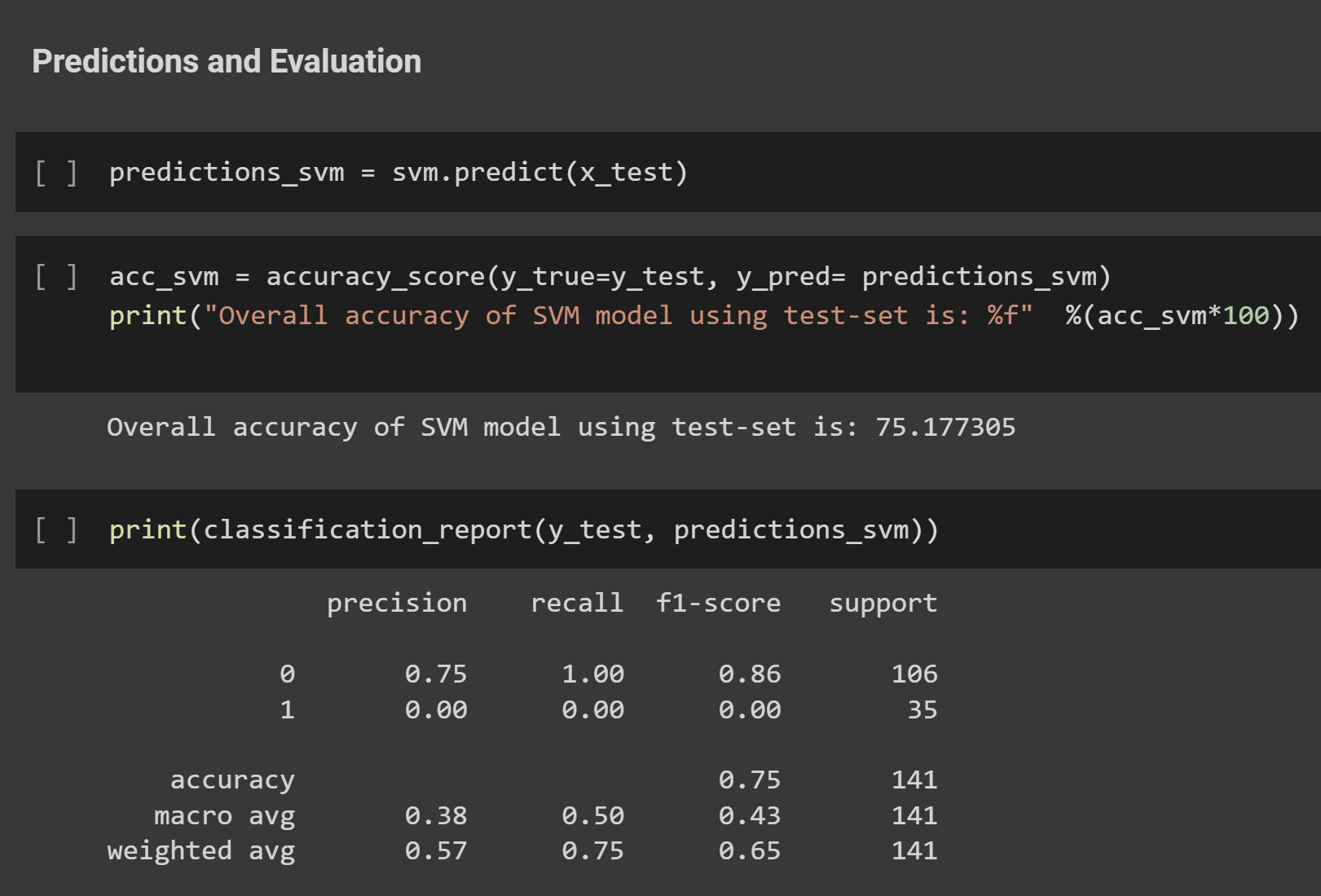
**AdaBoost:**

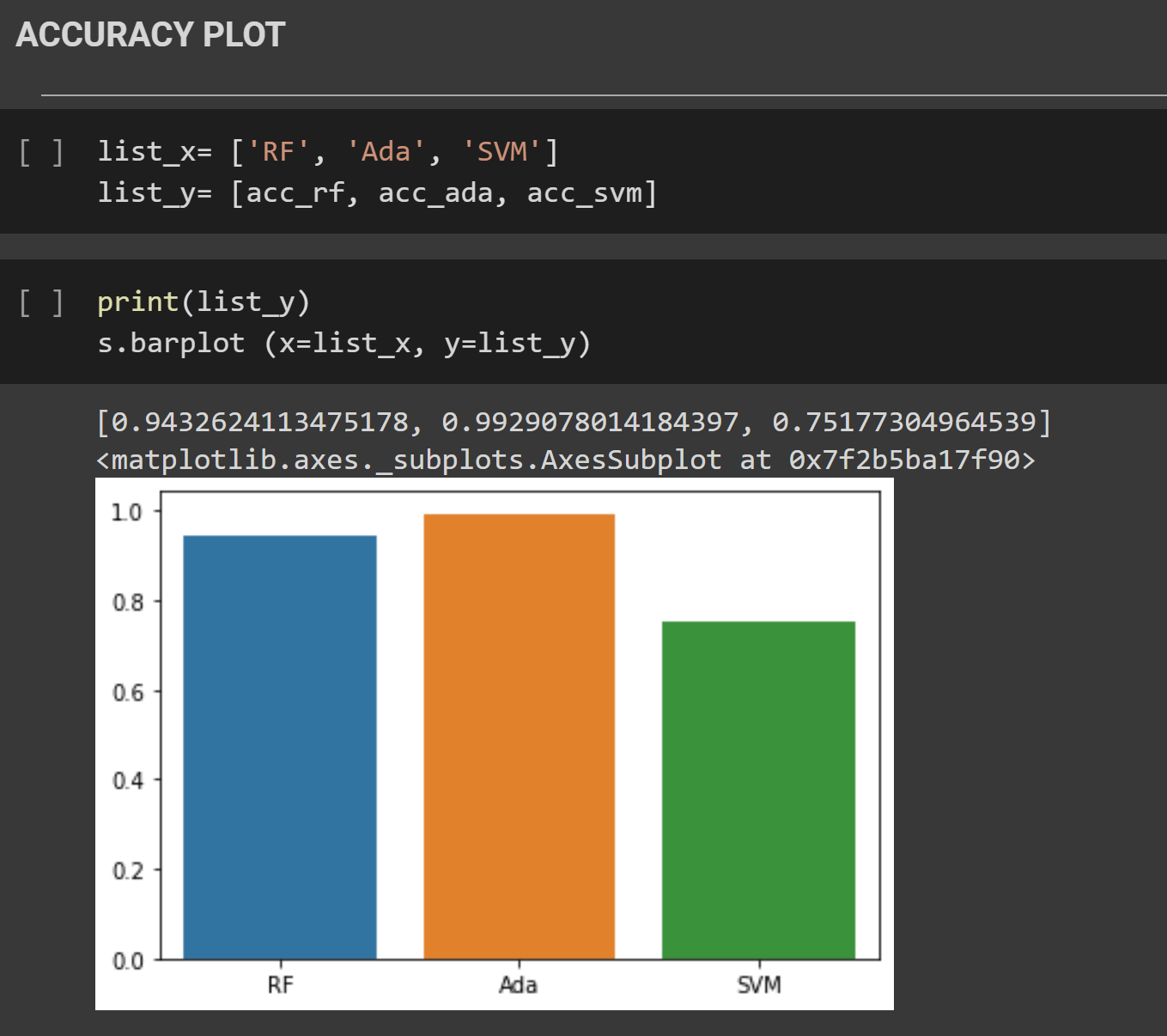
****

****

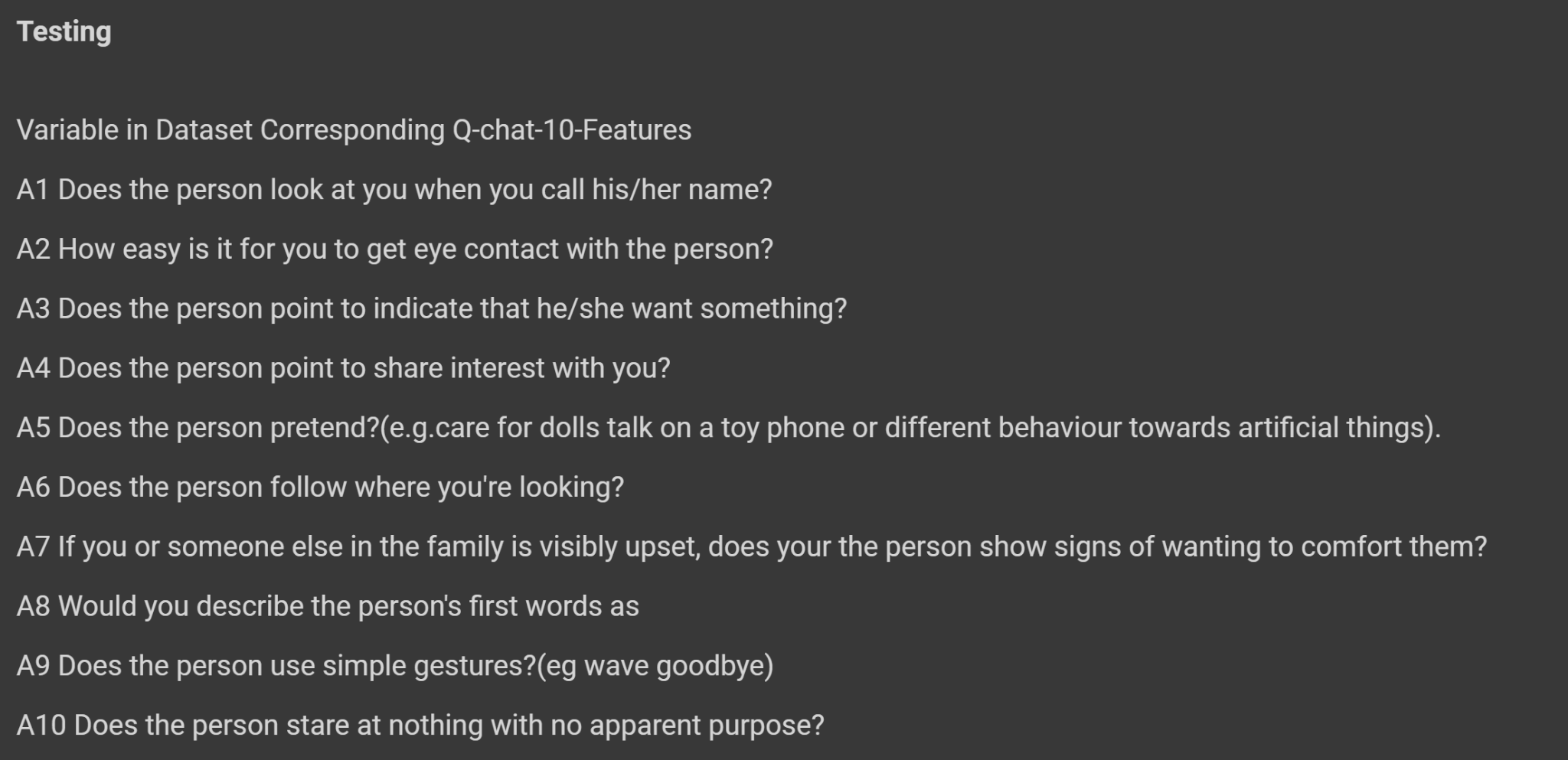
**SVM:**

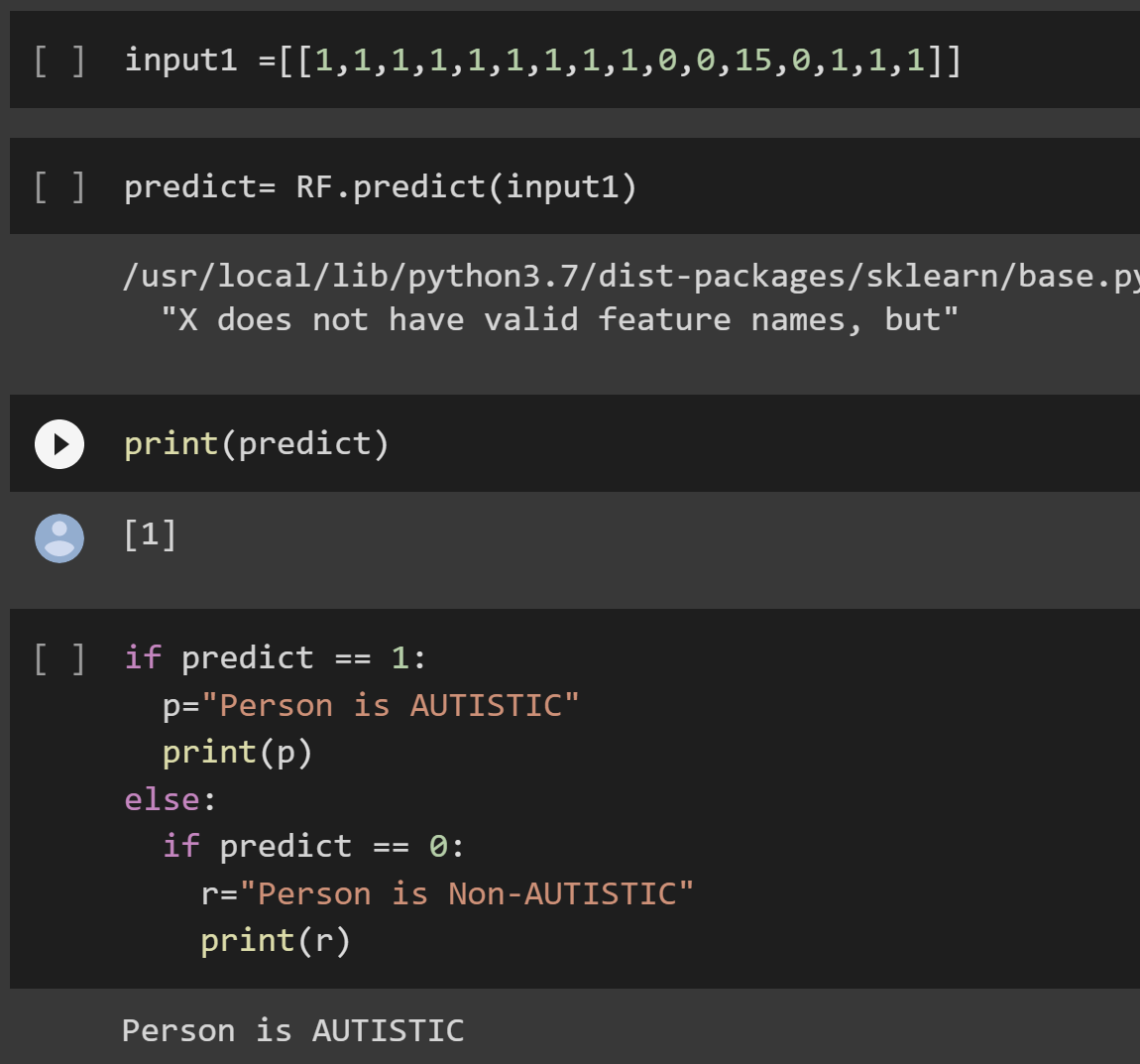
****

****

****

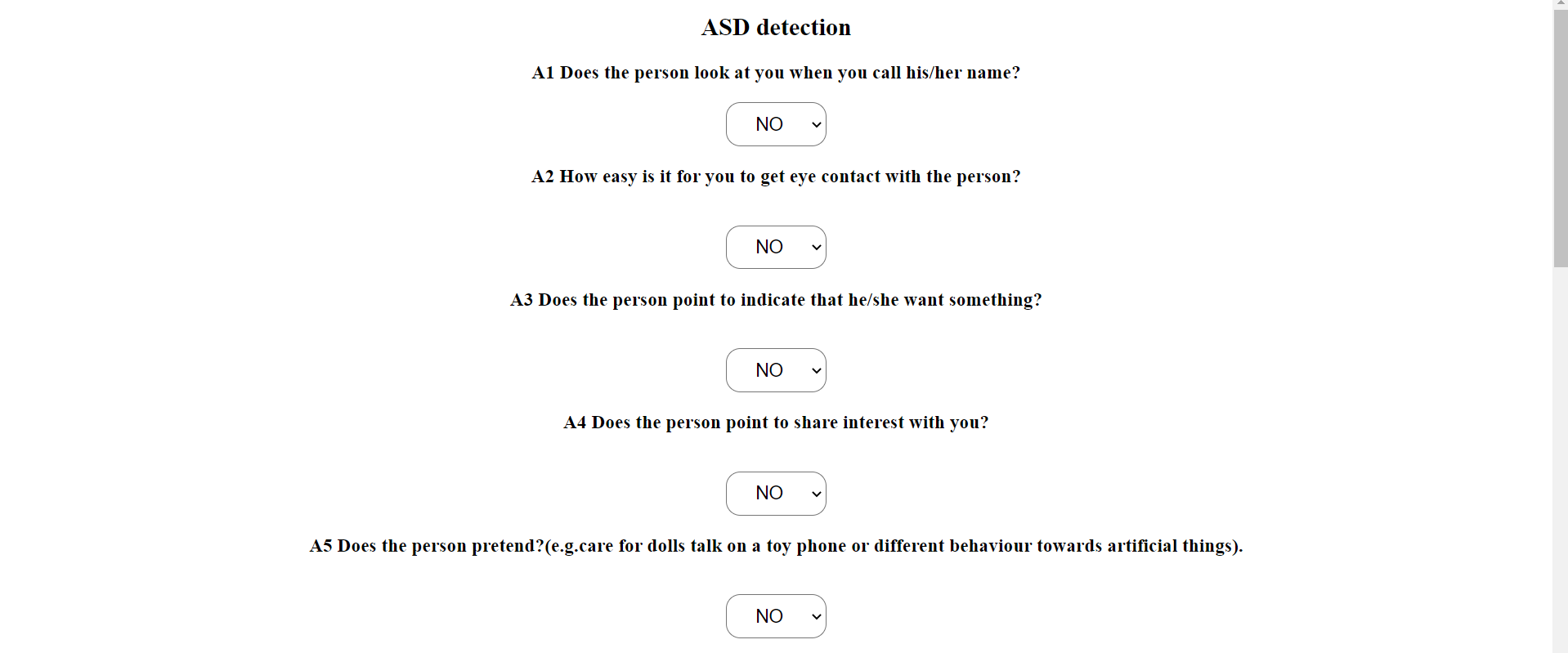
#### **7.2 Testing**

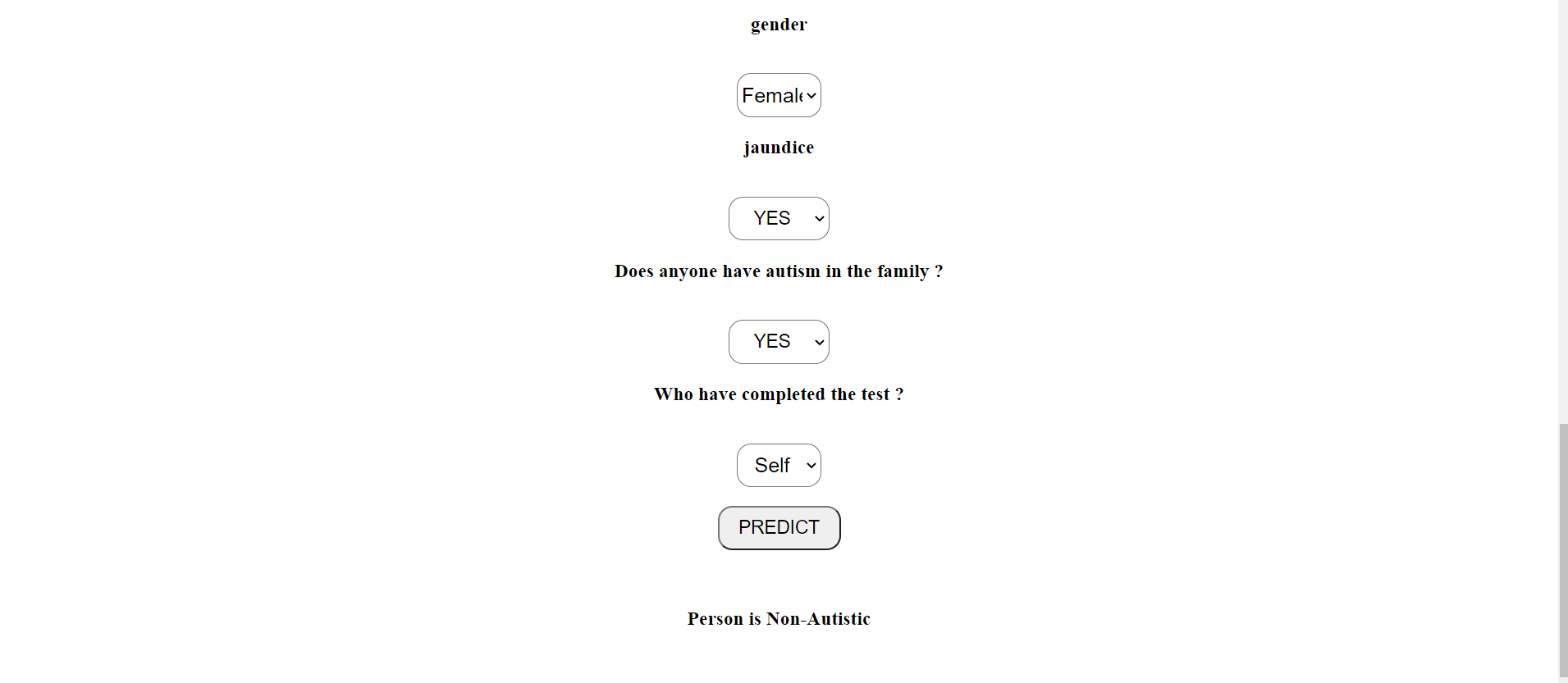
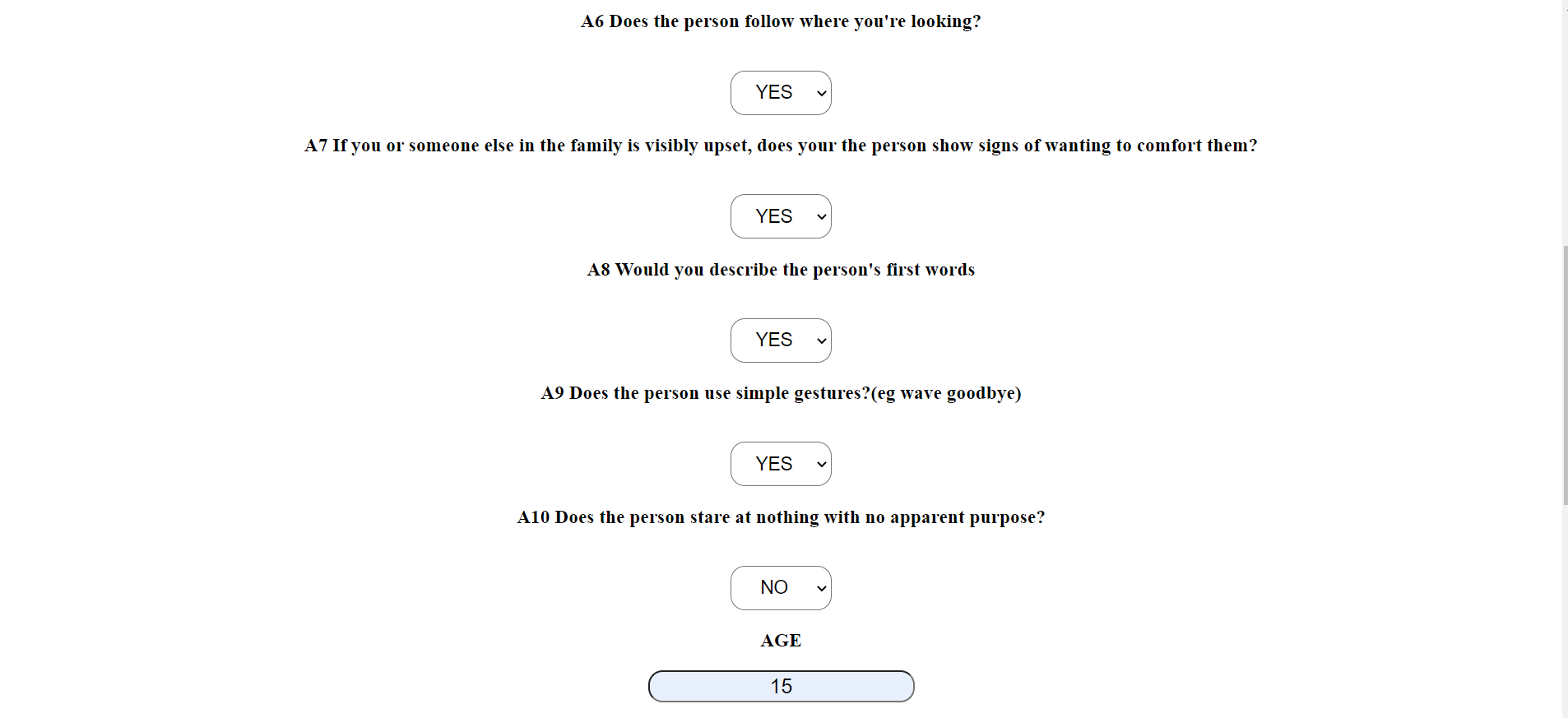
****

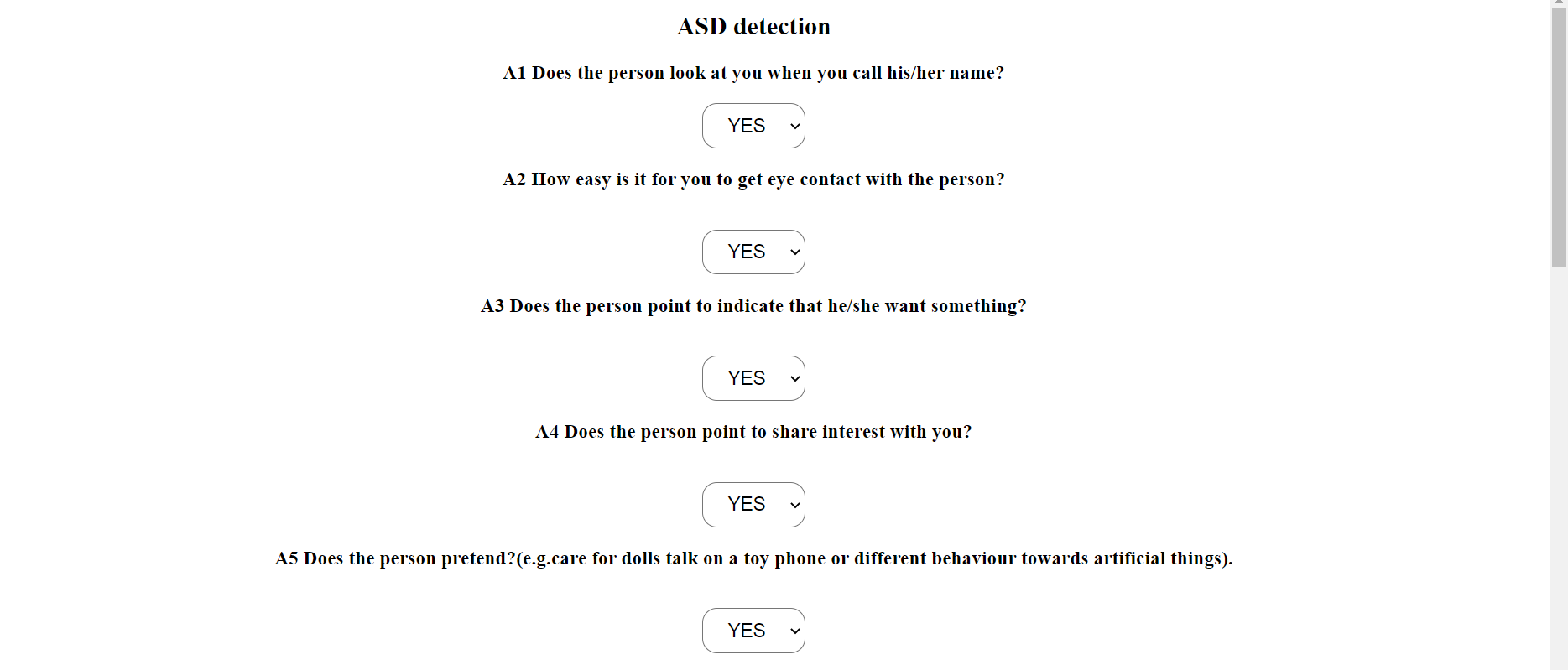
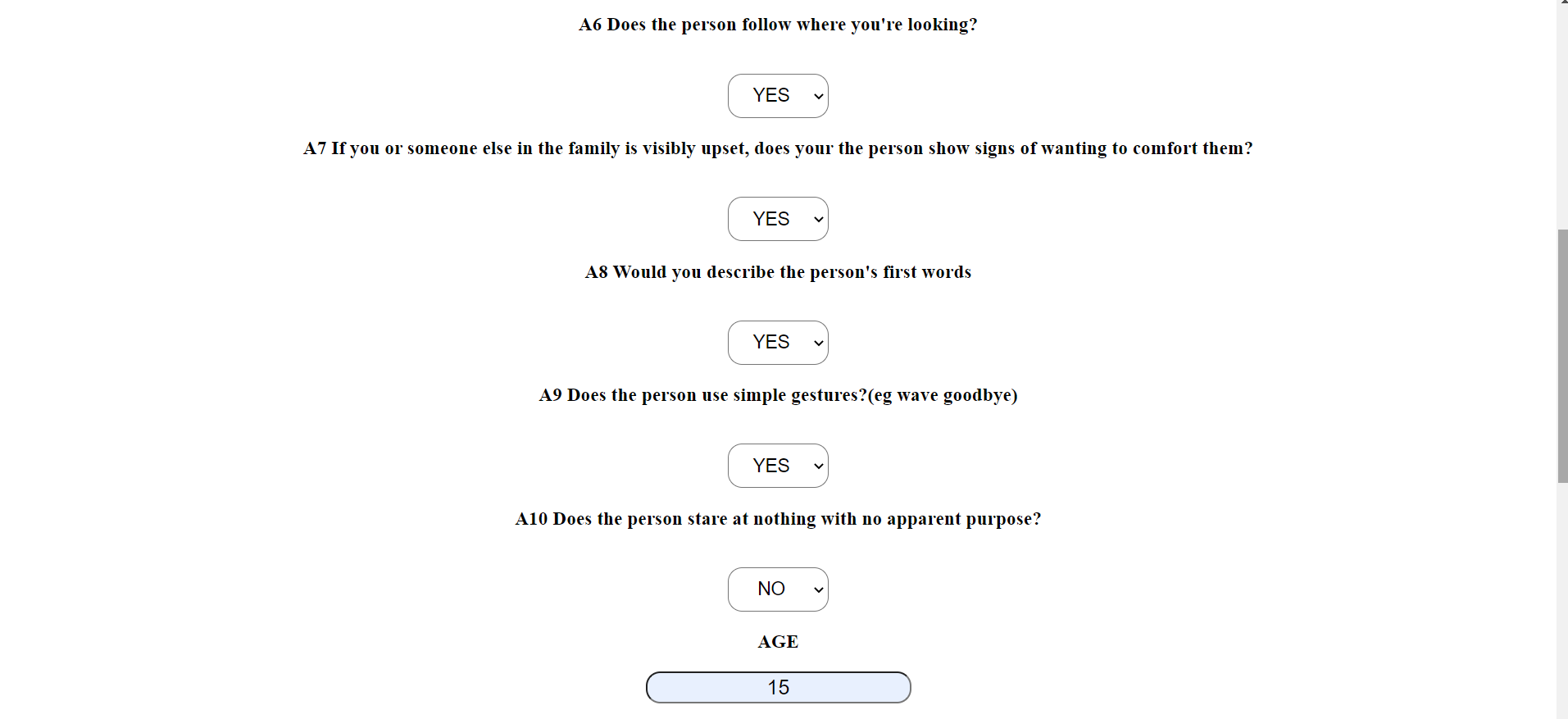
****

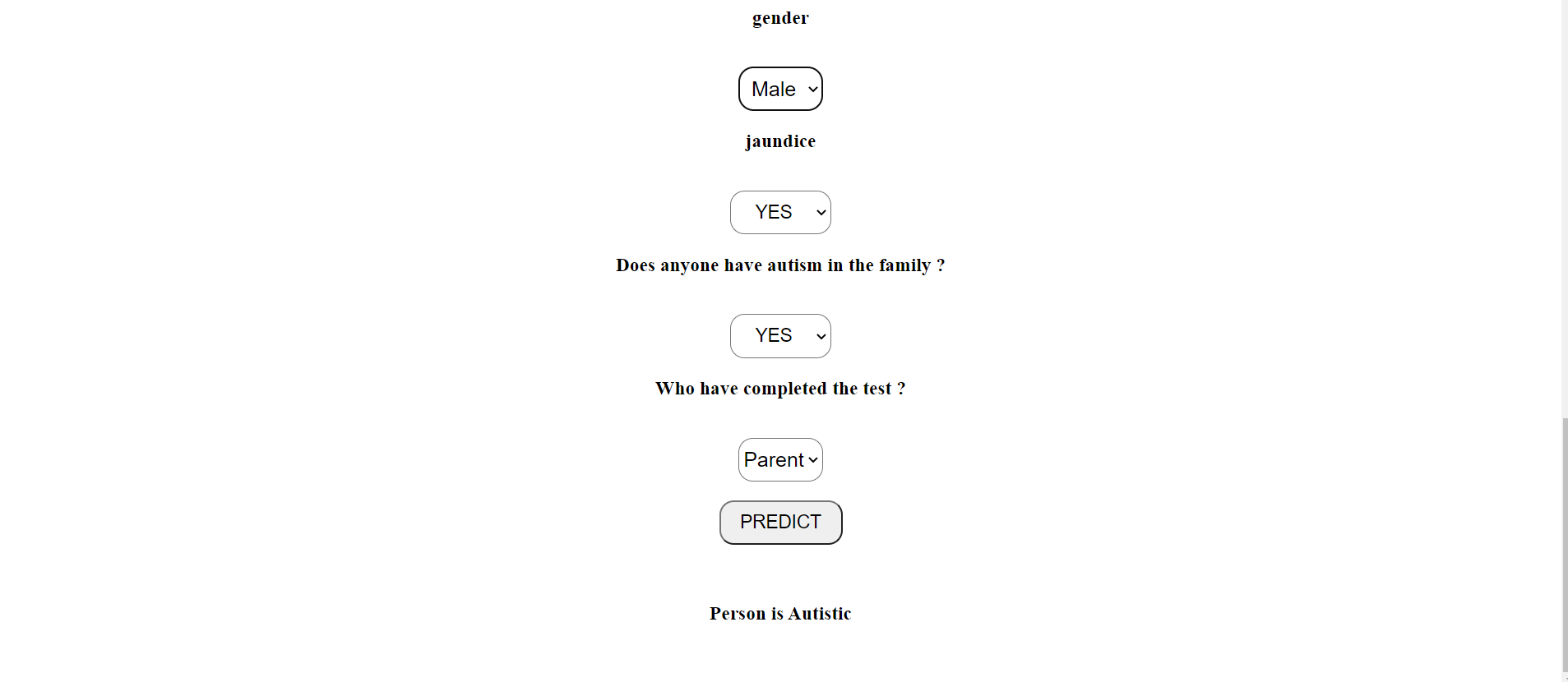
**7.3 Model deployment using flask:**

* In order to make the application user-friendly, the Random Forest algorithm was integrated with flask into a screening web application.
* Based on the AQ-10 characteristics and some other inputs, various questions were asked for different age groups. According to the inputted answers by the user, the application indicated whether the person is autistic or not.

 **7.3.1 Test case 1: Obtaining person is Non-Autistic:**



**7.3.1 Test case 2: Obtaining person is Autistic:**

****

**8. RESULT & DISCUSSION**

The dataset for this study was gathered from the UCI Repository, which is open to the public. Dataset Name-ASD Screening Data for Adult, Attribute Type-continuous and binary, Number of Instances-704. The dataset includes 16 parameters, the first 10 questions contain the behavioural characteristics of the person based on communication and social interaction. Next parameters are gender, age, person born with jaundice, any family member containing autism and who completed the dataset. The questions are graded using a one-point scoring system for each of the ten questions. On each question, the user might earn 0 or 1 point depending on their response. The entire dataset has been divided into two sections, one for training and the other for testing, using an 80:20 ratio. Algorithms were created and their accuracy was tested in order to generate autism trait predictions. With the adoption of many machine learning algorithms such as random forests, Support vector machine and AdaBoost. The implementation of the ASD predictive system was proposed using random forest. Using the AQ10 dataset and real-world data, the proposed predictive model was assessed for efficiency, specificity, precision as well as false positive rate. On the AQ-10 dataset, the leave one out technique is used to test the performance of the suggested model. Using the AQ10 dataset and real-world data, the proposed predictive model was assessed for accuracy, correctness, validity, efficiency as well as false positive rate. A total of 704 records were examined, with 189 being identified as autistic, accounting for around 26.85% of the total and 515 records are identified as non-autistic accounting for 73.15%.

**9. COMPARATIVE STUDY**

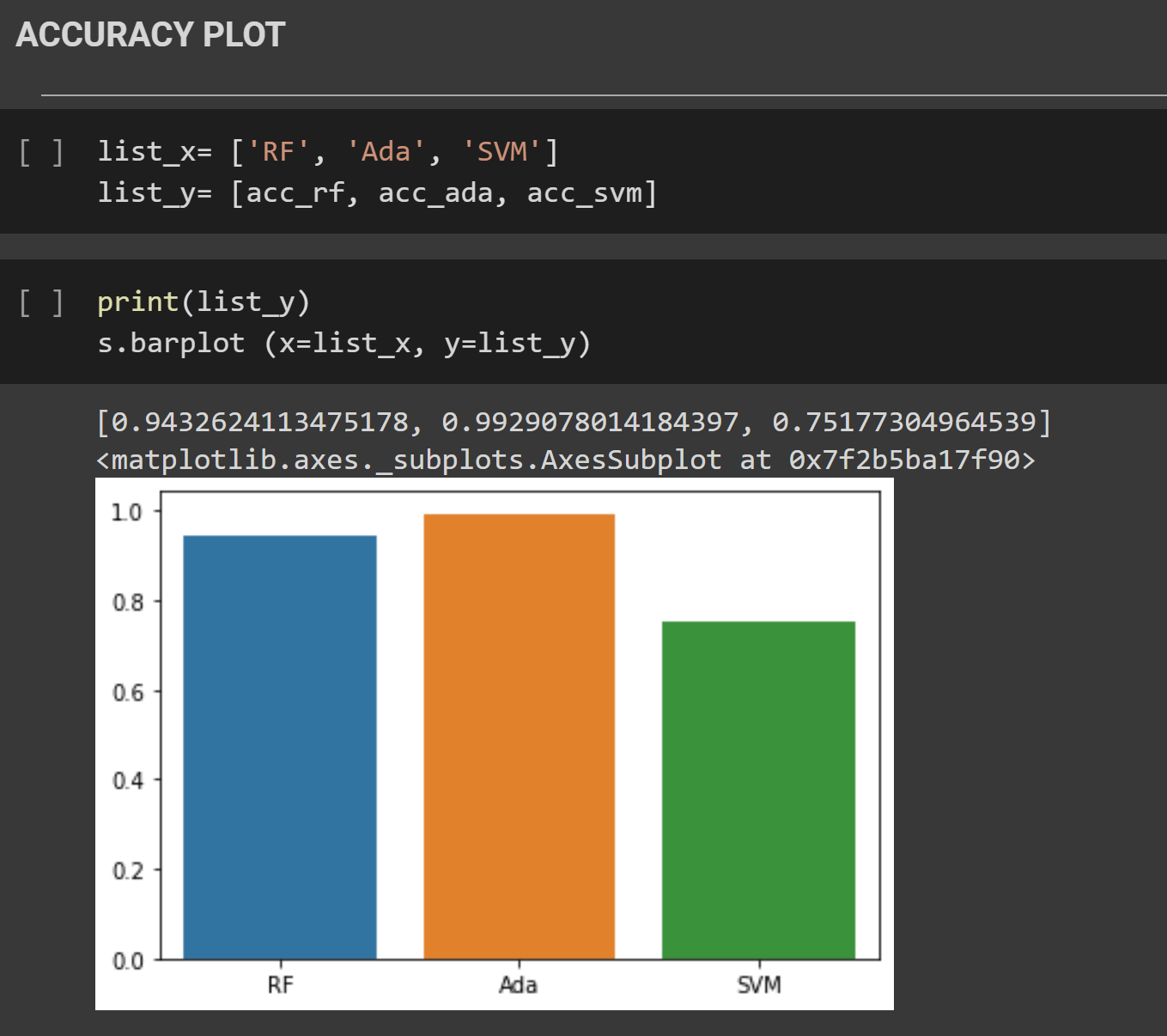
****

Figure 9.1 Comparison of the different algorithms used

Among trained algorithms AdaBoost has the higher accuracy when compared to other algorithms.

**10. CONCLUSION AND FUTURE SCOPE**

## The study is about to identify autism by investigating patients’ behavioural characteristics. After exploring the ASD dataset with different kinds of learning algorithms, We have arrived at the conclusion that all of the models work well with the data. We have used different metrics to measure the performance of our models, and it seems like all of the metrics indicated an almost perfect classification of the ASD cases.

## We plan to build a more accurate model, we need to have access to larger datasets. Here the number of instances after cleaning the data were not so sufficient enough so that we can claim that this model is optimum. As this dataset is only publicly available from December 2017, we believe it may not work that deal with this dataset are available online.

## In that consideration, These models can serve as benchmark models for any machine learning researcher/practitioner who will be interested to explore this dataset further. With this fact in mind, It is a very well developed model that can detect ASD in individuals with certain given attributes.

## When compared to the results of another recent study, we found that the AdaBoost classifiers performed better than the SVM classifiers when all of the features attributes were included after missing values were handled. After handling missing values, AdaBoost based models show the same accuracy of prediction for ASD dataset. Finally, we created a front-end application to our project as a user friendly application. In future we would like to add another feature which is obtaining output in the form of audio.

**11. REFERENCES**

[1] Applying machine learning to identify autistic adults using imitation: An exploratory study

Baihua Li ,Arjun Sharma, James Meng, Senthil Purushwalkam, Emma Gowen.

[2] Thabtah, Fadi. "Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. (2018) " Informatics for Health and Social Care : 1-20.

[3] Thabtah, Fadi, Firuz Kamalov, and Khairan Rajab. (2018) "A new computational intelligence approach to detect autistic features for autism screening." International journal of medical informatics 117: 112-124.

[4] Vaishali, R., and R. Sasikala. "A machine learning based approach to classify Autism with optimum behaviour sets. (2018) " International Journal of Engineering & Technology 7(4): 18.

[5] Testing the accuracy of an observation-based classifier for rapid detection of autism risk

M Duda, J A Kosmicki & D P Wall.

[6] Daniel Bone, Matthew S. Goodwin, Matthew P. Black, Chi-Chun Lee, Kartik Audhkhasi, and Shrikanth Narayanan. (2015) "Applying machine learning to facilitate autism diagnostics: pitfalls and promises." Journal of autism and developmental disorders 45(5): 1121-1136.

[7] Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework

Wenbo Liu,Ming Li,Li Yi.

[8] Dennis P. Wall, Rebecca Dally, Rhiannon Luyster, Jae-Yoon Jung, and Todd F. DeLuca. (2012) “Use of artificial intelligence to shorten the behavioral diagnosis of autism.” PloS one, 7(8): e43855.

[9] Detecting Abnormal Word Utterances in Children With Autism Spectrum Disorders: Machine-Learning-Based Voice Analysis Versus Speech Therapists

Yasushi Nakai, Tetsuya Takiguchi, Gakuyo Matsui, Noriko Yamaoka, Satoshi Takada Satoshi Tak

[10] Fadi Thabtah. (2017) “ASD Tests. A mobile app for ASD screening.” www.asdtests.com [accessed December 20th, 2017].

[11] Baihua Li, Arjun Sharma, James Meng, Senthil Purushwalkam, and Emma Gowen. (2017) “Applying machine learning to identify autistic adults using imitation: An exploratory study.” PloS one, 12(8): e0182652. [12] Fadi Fayez Thabtah (2017), “Autistic Spectrum Disorder Screening Data for Adult”.,

[13] M. S. Mythili, and AR Mohamed Shanavas. (2014) “A study on Autism spectrum disorders using classification techniques.” International Journal of Soft Computing and Engineering (IJSCE), 4: 88-91.

[14] J. A. Kosmicki, V. Sochat, M. Duda, and D. P. Wall. (2015) “Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning.“ Translational psychiatry, 5(2): e514.