AUTISM SPECTRUM DISORDER DETECTION IN EARLY AGES USING AUTOMATED MACHINE LEARNING

KHAFSA EHSAN

01-241211-003

A thesis submitted in fulfillment of the
requirements for the award of the degree of
Master of Science (Software Engineering)

Department of Software Engineering

BAHRIA UNIVERSITY ISLAMABAD

APRIL 2023

APPROVAL FOR EXAMINATION

Scholar's Name: _Khafsa Ehsan_ Registration No. __01-241211-003__

Program of Study: __MS (Software Engineering)__

Thesis Title: ___Autism Spectrum Disorder Detection in Early Ages Using Automated Machine Learning__

It is to certify that the above scholar's thesis has been completed to my satisfaction and, to my belief, its standard is appropriate for submission for examination. I have also conducted plagiarism test of this thesis using HEC prescribed software and found similarity index 15% that is within the permissible limit set by the HEC for the MS degree thesis. I have also found the thesis in a format recognized by the BU for the MS thesis.

Principal Supervisor's

Signature:_____

Date: _____

Name:_____

AUTHOR'S DECLARATION

I, Khafsa Ehsan hereby state that my MS thesis titled "Autism Spectrum Disorder Detection in Early Ages Using Automated Machine Learning" is my own work and has not been submitted previously by me for taking any degree from this university Bahria University Islamabad or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw/cancel my MS degree.

Name of scholar: Khafsa Ehsan (01-241211-003)

Date: _____

PLAGIARISM UNDERTAKING

I, Khafsa Ehsan, solemnly declare that research work presented in the thesis titled "Autism Spectrum Disorder Detection in Early Ages Using Automated Machine Learning" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero-tolerance policy of the HEC and Bahria University towards plagiarism. Therefore I as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS degree, the university reserves the right to withdraw/revoke my MS degree and that HEC and the University has the right to publish my name on the HEC/University website on which names of scholars are placed who submitted plagiarized thesis.

Scholar / Author's Sign: _____

Name of the Scholar: Khafsa Ehsan (01-241211-003)

DEDICATION

To my beloved mother and father

ACKNOWLEDGEMENT

I would start by thanking ALLAH Almighty, with gratitude for giving me strength in every aspect of life and helping me in this thesis as well.

I wish to express my sincere appreciation to my thesis supervisor, Dr. Kashif Sultan, for his support, guidance, and valuable feedback throughout this thesis. His expertise and encouragement have been instrumental in shaping my research and helping me to overcome the challenges that I faced. I am also very thankful to Dr. Raja Suleman for guidance, advice, and motivation.

I want to acknowledge my parents, who supported me and. This thesis work is dedicated to my parents, who have been a constant source of support during the challenges of life and prayed for my success. I am truly thankful for the support in every aspect whether that is financial, emotional, or mental. I would like to thank my sisters, Aqsa Ehsan and Taiba Ehsan for their constant support that kept me motivated.

Lastly, I would like to acknowledge the contributions of all the participants who took part in my study, without whom this research would not have been possible.

ABSTRACT

Autism spectrum disorder (ASD) is a neurodevelopmental condition characterized by a diverse range of symptoms and levels of ability. Detection of ASD at early ages is desirable as it allows for early intervention, which can improve the child's condition. However, the conventional diagnostic process requires hours of clinical examination, which can be time-consuming and expensive. The thesis proposes the use of auto-ML as a tool to simplify the diagnosis process and improve precision. The study collected data from multiple rehab centers in Pakistan and applied the auto-ML framework TPOT to the dataset for ASD detection. The results showed that TPOT gave the best pipeline for the dataset, with the highest accuracy of 79%, and it was verified. The study contributes to the field of ASD diagnosis by utilizing auto-ML to identify the likelihood of ASD in children during the early stages of development. The study also provides an evaluation of precision, recall, and F1-Score metrics to verify the accuracy of the diagnosis. Overall, this thesis presents a promising approach to improve the detection of ASD in children, which can ultimately lead to better outcomes for affected individuals and their families.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ASD	-	Autism Spectrum Disorder
AutoML	-	Automated Machine Learning
ML	-	Machine Learning
WHO	-	World Health Organization
PDA	-	Pervasive Development Ailment
ADOS	-	Autism Diagnostic Observation Schedule
ADI-R	-	Autism Diagnostic Interview-Revised
DSM-5	-	Diagnostic and Statistical Manual of Mental Disorders 5
AQ	-	Autism Spectrum Quotient
SCQ	-	Social Communication Questionnaire
M-CHAT	-	Modified Checklist for Autism in Toddler
CARS-2	-	Childhood Autism Rating Scale
STAT	-	Screening Tool for Autism in Toddlers and Young Children
SVM	-	Support Vector Machines
RFC	-	Random Forest Classifier

NB	-	Naïve Bayes
LR	-	Logistic Regression
kNN	-	k-Nearest Neighbors
Q-CHAT-10	-	Quantitative Checklist for Autism in Toddlers-10
DD	-	Developmental Delay
TD	-	Typically Developing
JADBio	-	Just Add Data Bio
PDQ	-	Performance Developmental Quotients
TP	-	True Positive
FP	-	False Positive
TN	-	True Negative
FN	-	False Negative
AUC-ROC	-	Area Under Curve-Receiver Operating Curve
GB Classifier	-	Gradient Boosting Classifier

CHAPTER 1

INTRODUCTION

Autism is a condition generally referred to as Autism Spectrum Disorder (ASD), with the term "spectrum", denotes the range of symptoms, abilities, and extent of impairment or disability that individuals with the disorder may exhibit [1]. Autism is not a sickness or disease. It is neurological developmental disorder which has lifelong impacts on the communication, social skills, and behavior of individuals [2] [3] [4]. ASD goes unnoticed in toddlers until it has reached a certain age. Its symptoms mostly appear in the first 5 years of life [5], but it can be diagnosed as early as first 18 months to 2 years [6] [7] [8]. Root causes of ASD have not found by the scientists yet [9]. But its causes can be genetic that either of parents have this disorder or any other family member, any kind of complications during pregnancy, the child might haven't received proper vaccination and so on [2] [3] [10].

Based on the statistical information provided by the World Health Organization (WHO), in initial 2 years of life, it affects around 1% of the population around the globe [3]. ASD detection is a challenging task while there are many other mental disorders with overlapping conditions [11]. As its symptoms appear at an early age of 3 years and continue for the rest of the life [12] [13]. Although, there is no treatment for this condition, but early detection can reduce the effects by means of therapy [10] [14] [15]. The only proven therapy is behavioral therapy if it's started at an early age [15] [16] [17]. Autism is categorized into 3 categories on the basis of severity levels (low, medium, high). They are: Autistic ailment, Asperger ailment and PDA [2]. The following are some of the hurdles experienced by children with ASD [12] [18] [19]:

- Lack of concentration
- Unable to maintain eye contact.
- Repetition of words/phrases
- No social interaction with others

- Unable to understand and express sounds, gestures, and facial expressions.
- Sensitivity to sounds, touch, smell.

ASD symptoms vary from one child to another, they generally fall in three categories [4] [18] [20]:

- **Social impairment** – such as sharing emotions, holding a conversation.
- **Communication difficulties** – can be either verbal (expressed through speech) or non-verbal (like facial expressions, eye contact, and gestures).
- **Repetitive and stereotyped behavior** – repeating words or actions.

Formal diagnosis of ASD is an extensive process in which the average wait period for ASD detection in United Kingdom is over 3 years [21] [22]. ASD diagnosis can be made at any age, but early diagnosis is also beneficial for both patient and family as it will improve the condition and reduced the cost linked with delayed diagnosis [4] [5] [23]. Researchers highlighted the significance of early identification [24]. Timely intervention for enhancing language and communication abilities, as well as the overall well-being of children with autism [18] [25] [26]. Due to the lack of local assessment diagnostic tools, limited speech-therapy facilities especially in rural areas, autistic children do not get to rehabilitation facility till the age of seven when they get into compulsory schooling. Therefore, for ASD an early intervention is needed through a speedy diagnosis process. Traditionally, ADOS is used to determine the difference between the behavior of a child with ASD and without ASD. The ADI-R and the DSM-5 are recognized as reliable diagnostic tools [2] [18].

ASD diagnosis can be made at an early age of 18 – 24 months [25]. The symptoms which appear at this age can be distinguished from the typical developmental characteristics of a child [14]. The formal diagnosis of ASD is frequently delayed until the age of 4 years due to variation in symptoms [3]. There are various screening methods for ASD such as AQ, SCQ, and M-CHAT, CARS-2, and STAT [27]. The aim of screening and monitoring is to detect children who are at risk or have been exposed

to ASD during the early stages. Individuals such as parents, caregivers, teachers, and those without specialized education or training can perform these screenings. Information from these processes can be helpful for professionals to understand the behavior and condition of children [1] [18] [19].

Early diagnosis leads to early treatment which in return improves the quality of life for people with ASD [21] [28] [29]. Data from various screening methods have been used to detect autism with the help of machine learning (ML) technique. In various studies, Q-CHAT-10 [7] is used for ASD screening in children [29] [30] [31]. The objective of this study is to utilize auto ML tools to identify whether a child is prone to ASD or not at the very beginning stages.

1.1. Motivation

Early diagnosis of ASD is a challenging task in clinical practices. As, autism is a life-long developmental disorder. Detection does not provide the perfect solution. It is much easier to prevent the effects of ASD at early stages. Early diagnosis is very important as it leads to early treatment which improves the condition of a child. With the advancement of applications used in day-to-day life, machine learning is becoming an emerging field, and it is increasing interaction with technology at a rapid pace.

Early autism diagnosis helps to reduce time, effort and cost used for autism diagnosis. It also increases the rate of betterment in the condition of autistic children.

1.2. Research Gap

ASD is also known as “behavioral disease”, as it affects social communication and interaction. Due to which children with autism are often socially isolated as mingling

with peers is a huge challenge faced by them. Identification of ASD is difficult in early childhood due to a variety of symptoms. ASD diagnosis is an expensive and lengthy process carried out by professionals in a clinical environment. Early diagnosis leads to early treatment which helps to improve the language, communication, and overall well-being of children. For this purpose, ML techniques can provide faster detection of ASD in children.

1.3. Problem Statement

The autism spectrum disorder (ASD) detection is difficult in children due to the wide range of symptoms that present themselves, which requires expensive and lengthy processes by professionals in clinical environments. Early detection of ASD is critical for improving outcomes in children with ASD yet remains a challenge. Moreover, the scarcity of large and diverse datasets for training and testing machine learning models presents a further obstacle in ASD research. Therefore, the research aims to explore strategies to overcome these challenges and enhance the accuracy of early ASD detection through the development of automated machine learning model.

1.4. Research Questions

RQ 1: How proposed auto-machine learning (auto-ML) framework will be helpful in predicting ASD?

RQ 2: How proposed framework will be compared with traditional methods for ASD detection?

1.5. Research Objectives

The objective of this study is to reduce the human involvement in autism prediction by utilizing machine learning techniques. The study aims to determine if a child is susceptible to ASD through ML techniques, which can aid in the early diagnosis of ASD. This can result in better treatment for children with ASD at an early age of the condition.

1.6. Contribution of the study

This section of the thesis discusses the significant contributions that have arisen as a result of conducting this study:

- Dataset collected through survey using Q-CHAT-10 questionnaire.
- This study implements the auto ML on dataset collected using TPOT library. We evaluated our models using various metrics, such as precision, recall, F1-score, PR, and AUC-ROC curves.
- We verified the pipelines generated by TPOT on our dataset by manually creating the model with same parameters as TPOT generated. Evaluation metrics of both auto ML and manual models were compared to evaluate the effectiveness of outcomes.

1.7. Outline of this thesis

The organization of this paper is as follows: **Chapter 1** “Introduction” section includes the introduction of the study. **Chapter 2** “Literature Review” section summarizes the previous works on ML that are related to ASD. **Chapter 3** “Research

Methodology" section explains the working and methodology of the auto ML system that we have proposed and its implementation. **Chapter 4** "Results and Evaluation" section shows the inferences and results obtained. Finally, **Chapter 5** "Conclusion" section highlights our contributions, and future plans to extend this work.

CHAPTER 2

LITERATURE REVIEW

This section gives an overview of the prior research work that has been done in this area. Typically, for ASD diagnosis standardized clinical tests are the only methods. They require hours of clinical examination and a huge medical cost for diagnosis. Although various techniques have been applied for ASD diagnosis like eye tracking techniques, brain imaging techniques, kinematic analysis, and so on. Numerous studies have used ML algorithms for either confirming a diagnosis or making an early ASD diagnosis.

Vakadkar et al., [1] proposed ML models for autism detection. The aim of this research was to ascertain whether a child is susceptible to autism during the early stages of development or not. They designed automated ASD prediction model to speed up the diagnosis as compared to the traditional methods. For this purpose, they applied SVM, RFC, NB, LR, and KNN to dataset compiled by Dr. Fadi Thabtah. This dataset was based on Q-CHAT-10 with 18 attributes and 1054 instances. Logistic regression gave the highest accuracy. The limitation of this study was that the dataset had a limited number of attributes and instances.

Erkan et al., [5] utilized three datasets (AQ-10-Adult, AQ-10-Adolescence, AQ-10-Child) from UCI database for their research. The aim of the study was to offer a simplified approach of ASD diagnosis at early stages. Authors evaluated the datasets using ML algorithms and found that RFC is more effective than SVM, and kNN for selected datasets. In this research, for every experiment, data were selected randomly 100 times to test the classification models. In this study, they found that the early identification of ASD is possible with huge dataset. Whereas, if the data sample is

larger then, the accuracy of diagnosis is also higher. So, the accuracy of ML based model would be dependent on the completeness of the data collected.

Thabtah et al., [32] suggested app for screening autism spectrum disorder using mobile technology is called ASDTests. This application contains tests in 11 different languages so that a large number audience can participate. The app's modules are designed to cater to distinct age groups, including toddlers, children, adolescents, and adults. Initially, this app served as a data collection tool and provides ASD diagnosis. Professionals can use it to help people or to let them know whether or not to pursue a fo rmal clinical diagnosis. AQ-10 and Q-CHAT are used for screening in application with visual aids.

Ruta et al., [33] an Italian clinical sample was used to validate the psychometric properties of the Q-CHAT questionnaire. Q-CHAT is specifically designed quantitative measure for autism instead of any other neurological disorder. For this study, a group of 315 children took part. They compared young autistic children (n = 139) to the DD (n = 50) and TD children (n = 126). All the statistics related to the 3 groups involved in the study were discussed as well. Q-CHAT scores were considerably higher in autistic group as compared to DD and TD groups.

Tartarisco et al., [34] used Machine learning algorithms were employed to investigate the accuracy of the Q-CHAT questionnaire in identifying autism in young children. In this study, they have used dataset of 265 children (n = 139 autistic children, n = 126 TD) collected by Ruta et al., [33] in 3 different Italian regions. The findings indicated that the Q-CHAT screening method has cross-cultural validity when used with an Italian sample. Researchers concluded that Q-CHAT can be utilized in primary care settings as it is a high performance, and easy to use tool. The study analyzed that SVM is the best performing ML model for their dataset.

Niedźwiecka et al., [35] used Q-CHAT to assess a sample of 1024 Polish children. This study aimed to identify ASD at an early age within a non-English speaking community. And to investigate the association of symptoms with age, gender, or ASD family history. The study involved four groups of contributors: typically

developing toddlers, toddlers reported by parents for ASD concerns, toddlers with delayed development and elder siblings with ASD who are at risk for autism. The results showed that the number of boys with ASD is higher than that of girls and the age is not associated with Q-CHAT score.

Farooqi et al., [36] have discussed the challenges faced during the data collection process in country like Pakistan where there is no tracking or reporting of ASD cases. The only government run Child Psychiatry Department in Pakistan, Mayo Hospital, Lahore, provided a sample of 100 people with ASD. Data was gathered on the basis of a questionnaire with 21 questions. After applying ML model to their dataset, authors concluded that first born and male are more effected by ASD. There was no physical difference between autistic and non-autistic children. Whereas genes also play an important role, 50% of children with ASD had some family member with cognitive disability history.

Jacob et al., [29] used AutoML based tools Just Add Data Bio (JADBio) in this study on dataset from UCI repository. The data was initially collected by Thabtah., [32] and is publicly available. In order to predict ASD, this study is the first to describe feature signatures and their importance in differentiating between classes.

Table 2-1 Reviewed Research Work

Ref.	**Year**	**Key Findings**	**Limitations**
[1]	2021	An automated model that can help medical professionals identify ASD was presented in this study. Used logistic regression to obtain an accuracy of 97.15% with 18 attributes and 1054 samples.	The limitations are unavailability of large and open-source datasets related to ASD.

[5]	2019	Three datasets for children, adolescents, and adults with ASD were employed to classify ASD using SVM, KNN, and RF algorithms. RF is a more effective method among SVM, and KNN for selected dataset.	Unavailability of large and complete dataset related to ASD.
[32]	2019	In this research, ASDTests app was proposed for data collection and assisting health professionals in ASD detection. Covered a large audience, available in 11 different languages. The study covered four distinct age groups: infants, children, adolescents, and adults. Initially, more than 1400 instances were gathered for analysis.	It was not possible to conduct feature analysis using the app. Scoring and rules for screening tools need to be replaced with ML models.
[33]	2019	Used Q-CHAT questionnaire method for ASD screening filled by parents.	Children were distributed unevenly among the three groups.

		The Q-CHAT questionnaire's diagnostic characteristics, score distribution, and external validity were all examined in the study. Dataset contains N = 126 TD children, n = 139 children with autism, and n = 50 children with DD.	Children in DD group were considerably younger than those in the other two groups. Children with TD have high Performance Developmental Quotients (PDQ) and it is possible that the study participants may not be representative of the general population.
[34]	2021	ML approach was used on dataset collected by Ruta et al., [33]. Q-CHAT with 25 items was evaluated. In Italian sample, cross-cultural validity of the machine learning results was determined. SVM was the most effective among RF, NB, LR, KNN.	The sample size was relatively small. The study did not include a comparison of the accuracy of Q-CHAT with other diagnosis tools, such as the M-CHAT.
[35]	2022	The study examined early ASD symptoms that are linked to the age, gender, and family history of toddlers.	The study was conducted on the sample of toddlers in Poland, which may limit generalizability of findings to other cultures or populations.

		The study used a sample of 1024 Polish toddlers, which included 585 typically developing toddlers, 252 toddlers in the ASD-concerns group, 67 toddlers in the ASD-sibling group, and 120 toddlers with delayed development. According to the study, boys scored higher than girls, indicating that boys are more likely to have ASD.	The study relied on parent report, which may be subject to bias or variability in interpretation. The overrepresentation of developmental concerns expressed by parents and developmental delays resulted in an overrepresentation of boys in the study. Significant age differences among the four subgroups. During the Q-CHAT evaluation, the ASD-concerns group did not have a confirmed diagnosis of ASD. Therefore, it was not feasible to determine the sensitivity and specificity of the tool or establish cut-off points to identify ASD in this group.
[36]	2021	Collected local dataset and then made use of SelectKBest, Variance threshold, and weights for feature selection.	The number of samples in dataset was limited with n = 100 ASD children and n = 100 normal developing children.

		ML technique	**Best performance**	Due to the small dataset, correlations between attributes were not strengthened. And the model was not reliable enough. Since the dataset used in the study was local, it may not be appropriate to apply the findings in other parts of the world without consulting with professionals.
		SVM	SelectKBest, weights	
		RFC	SelectKBest, Variance Threshold	
		XGB	Variance Threshold	
		Gradient Boosting	weights	

2.1. Tools

Python and R are mostly used for implementation of ML models. Various studies used Java programming language and WEKA for implementation.

AutoML has become a popular solution for minimizing the time and energy required for repetitive tasks in ML workflows, including data preparation, feature engineering, model choice, hyperparameter tuning, and analyzing prediction outcomes [37] [38]. The study discussed six commonly used AutoML frameworks which are Auto-Weka, AutoSKlearn, TPOT, Recipe, ATM, and SmartML and experiments were performed with different datasets in order to set a benchmark for AutoML [39].

Several open-source AutoML tools are available, including AutoWeka, H2O.ai AutoML, TPOT, AutoSklearn, and machineJS [37] [39]. We used TPOT open-source tool for our dataset. The Table 2-2 shows the various AutoML tools [37] [39] where (+) represent commercialized tools.

Table 2-2 Auto ML tools

S. No.	Tool	Open - Source	Cloud - Based	Input data sources		Supports	Techniques	Training Framework
				Spread sheet datasets	Image, test			
1	Auto Sklearn	Y	N	N	Y	Classification, Regression	Bayesian optimization + automated ensemble construction	Sklearn
2	Auto Keras	Y	N	Y	Y	CNN, RNN, LSTM for classification	Efficient neural architecture search with network morphism	Keras
3	Google Cloud AutoML (+)	N	Y	N	Y	CNN, RNN, LSTM for classification	Reinforcement learning with gradient policy upgrade	TensorFlow
4	Azure ML (+)	N	N	Y	Y	Classification, Regression	Probabilistic matrix factorization + Bayesian optimization	Sklearn
5	Auto			Y	N		Bayesian	Weka

	Weka						optimization	
6	H2O-AutoML (+)	Y	N	Y	N	Classification, Regression	Ensemble, random search, Bayesian search	H2O, XGBoost
7	Ludwig	Y		Y	Y	Supervised Learning	Ensemble, random search, Bayesian search, neural architecture search	
8	TPOT	Y	Y	Y	N	Classification, Regression	Ensemble, genetic algorithm, Bayesian optimization, Random search	Sklearn

CHAPTER 3

RESEARCH METHODOLOGY

3.1. Introduction

This chapter outlines a research framework that utilizes automated machine learning techniques for the detection of ASD in children. There are various steps involved in our research process. Initially data is collected, then pre-processed to refine it for model. Then data partitioning into training dataset and testing dataset. Then, model training using auto ML and verification of model generated by auto ML, and finally performance evaluation.

3.2. Proposed Methodology

Our research methodology involves 3 steps preprocessing, auto ML based model development, verification of model generated by auto ML, and performance evaluation. After collecting the data, preprocessing begins, and we eliminated missing values and encoded categorical attributes.

After preprocessing, the development process began by dividing the dataset into two parts: a training set and a test set.

Then we applied auto ML framework TPOT to our dataset and evaluated TPOT classifiers using performance metrics. The best pipeline generated by auto ML TPOTClassifer showed the best classifier with optimum parameters. For verification of

the pipeline generated by the TPOT the same classifier is trained manually with same parameters generated by auto ML TPOTClassifer. We performed three Experiments and results obtained from those Experiments were compared for evaluation. Experiment 1, 2, and 3 are explained further in Chapter 4. In our study we used precision, recall and F1-score for evaluation. Figure 3-1 shows our proposed auto ML based framework. Figure 3-2 shows the framework for manual ML.

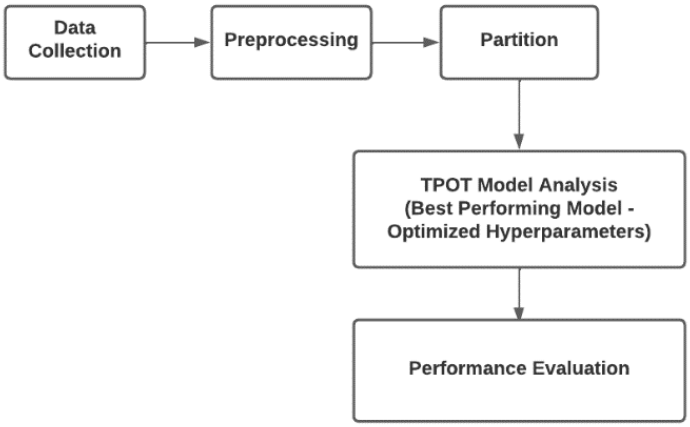

Figure 3-1 Proposed AutoML Framework – Experimentation

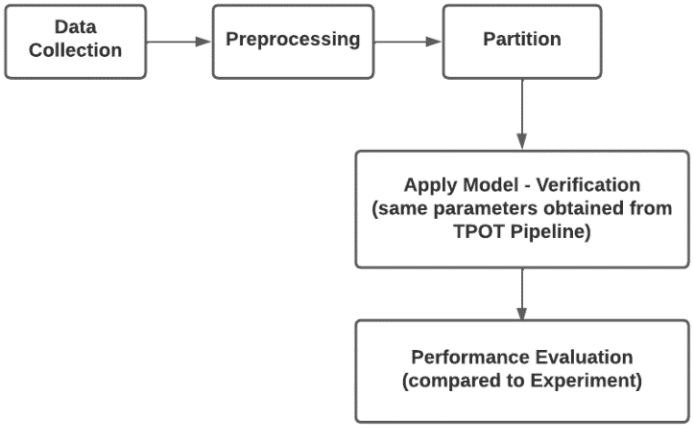

Figure 3-2 Proposed Manual ML Framework - Verification

3.3. Data Collection method

We collected data from several rehab centers in Pakistan. A number of samples were collected manually from the parents of children with autism. Then, all the data was stored in textual form in a Google sheet. It was a challenging process as there is no proper reporting and tracking of ASD cases in the country. To collect data for this study, a questionnaire was created based on the Q-CHAT screening method, as data had to be collected from scratch in both hard form and soft form. Google Form was used to collect data in soft form. The responses received via Google Form were downloaded in CSV format.

Q-CHAT was created by Allison et al., [40] to reduce the time required to fill the form enabling large population to fill the form. Initially, it consists of a set of 25 questions. Later Allison et al., proposed Q-CHAT-10 [41] which consist only 10 items. It enables a range of response categories. It is quick to administer as higher score signifies the autistic traits. This study based on Q-CHAT was approved by Cambridge Local Research Ethics Committee.

3.3.1. Q-CHAT-10

ASD condition can be minimized by means of therapy. For this early detection is desirable. For this purpose different screening methods are used. Allison et al., [40] developed Q-CHAT for ASD detection in children. Initially Q-CHAT was developed with 25 items. Later Allison et al., proposed Q-CHAT-10 [41] which consist only 10 items. The 10 best items for Q-CHAT-10 were selected on the basis of discrimination index (DI) for each item from the derivative sample. This shorter version of Q-CHAT showed the best results in the studies that is why we selected Q-CHAT-10 for our study. The Q-CHAT-10 also focuses on the most important questions for identifying ASD, as determined by statistical analysis of a larger set of questions.

In Table 3-1, for Question 1 to Question 9: if an option from columns C, D or E is chosen, one point should be scored for each related question. For Question 10: if an answer from columns A, B, or C is chosen, it should also be scored one point. Then these points will be added. If the score is more than 3/10, then multi-disciplinary assessment may be recommended by healthcare professionals for the child. Q-CHAT-10 means Q-CHAT with 10 questions. In Table 3-1. Q-CHAT-10 is shown below:

Table 3-1 Q-CHAT-10

		A	B	C	D	E
1	Does your child look at you when you call his/her name?	Always	Usually	Sometimes	Rarely	Never
2	How easy is it for you to get eye contact with your child?	Very easy	Quite easy	Quite difficult	Very difficult	Impossible
3	Does your child point to indicate that s/he wants something? (e.g. a toy that is out of reach)	Many times a day	A few times a day	A few times a week	Less than once a week	Never
4	Does your child point to share interest with you? (e.g. pointing at an interesting sight)	Many times a day	A few times a day	A few times a week	Less than once a week	Never
5	Does your child pretend? (e.g. care for dolls, talk on a toy phone)	Many times a day	A few times a day	A few times a week	Less than once a week	Never
6	Does your child follow where you're looking?	Many times a day	A few times a day	A few times a week	Less than once a week	Never
7	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)	Always	Usually	Sometimes	Rarely	Never
8	Would you describe your child's first words as:	Very typical	Quite typical	Slightly unusual	Very unusual	My child doesn't speak
9	Does your child use simple gestures? (e.g. wave goodbye)	Many times a day	A few times a day	A few times a week	Less than once a week	Never
10	Does your child stare at nothing with no apparent purpose?	Many times a day	A few times a day	A few times a week	Less than once a week	Never

Table 3-2 shows the description of the dataset. Dataset Variable A1-A10 refers to the Questions 1-10 shown in Table 3-1.

Table 3-2 Dataset variable description

Dataset Variable	Data Type	Attribute Description
A1	Binary (0, 1)	The response is based on the screening technique used
A2	Binary (0, 1)	The response is based on the screening technique used
A3	Binary (0, 1)	The response is based on the screening technique used
A4	Binary (0, 1)	The response is based on the screening technique used
A5	Binary (0, 1)	The response is based on the screening technique used
A6	Binary (0, 1)	The response is based on the screening technique used
A7	Binary (0, 1)	The response is based on the screening technique used
A8	Binary (0, 1)	The response is based on the screening technique used
A9	Binary (0, 1)	The response is based on the screening technique used
A10	Binary (0, 1)	The response is based on the screening technique used
Age_Mons	Number	Child's age in months
Sex	String	Male/Female
Jaundice	Boolean (Yes/No)	Whether the child was born with Jaundice
Family_mem_with_ASD	Boolean (Yes/No)	Any family member diagnosed with ASD

Who completed the test	String	Parent, caregiver, medical staff, clinician
Qchat-10-Score	Integer	Final score based on the scoring function
Class/ASD Traits	Boolean	The class label shows the presence of ASD traits, a score of "0" indicates the absence of such traits and a score of "1" indicates their presence.

3.3.2. Our Dataset Samples

After collecting data from multiple rehab centers, we transformed it as discussed in Chapter 3.3.1. for each question from A1-A10. Our dataset samples after applying transformation are shown in Table 3-3.

Table 3-3 Dataset Samples

Case_No	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**	**A10**	**Age_Mons**	**Sex**	**Jaundice**	**Family_mem_with_ASD**	**Who completed the test**
1	1	1	1	1	1	1	1	0	1	1	48.0	male	no	no	clinician
2	1	1	0	0	1	1	1	0	1	1	36.0	male	no	no	clinician
3	1	0	0	1	1	0	1	1	1	0	36.0	Male	no	no	caregiver
4	1	1	0	1	1	1	1	1	1	0	42.0	Male	no	no	caregiver
5	1	0	0	0	1	1	1	0	0	1	21.6	female	no	no	caregiver

3.4. Data Pre-processing

The process of transforming raw or noisy data to make it suitable for training and analysis is called data pre-processing. This step involves cleaning the data to remove any inconsistencies or errors.

Firstly, we checked for missing values in our data, and we removed them. Since our dataset had categorical attributes, we encoded the data attributes into binary form 0 and 1. Sex attributes having 2 classes (male/female) have been encoded. Jaundice to 0 and 1 for no and yes respectively. ASD Class/Traits to 0 and 1 for non-Autistic and autistic respectively.

We eliminated the irrelevant attributes, i.e. “Case_No” and “Who completed the test”.

3.5. Exploratory Data Analysis (EDA)

In this section, the description of techniques used to analyze the data, as well as the insights gained from the analysis are discussed. EDA helps to identify potential problems in the data and select appropriate methods to resolve them.

3.5.1. Correlation analysis

A correlation matrix shows the correlation coefficients between variables in the form of a table. The correlation coefficient is a statistical measure that indicates the strength and direction of the linear association between two variables.

- Size and color of the correlation coefficients: Larger and darker coefficients indicate stronger correlations.

- Positive and negative correlations: Positive correlations imply that variables tend to change together, while negative correlations indicate that variables tend to change in opposite directions.

The problems with highly correlated features in a dataset are overfitting, which can cause poor performance on new data, and difficulties in interpreting the individual impact of each feature on the model's predictions.

In Figure 3-4, A1, A2, A6, A8, and A9 have high correlation with Qchat-10-Score. So, Qchat-10-Score was dropped from dataset before being fed into the model. This was done to enhance the model's ability to generalize to new data by simplifying it and reducing redundancy. Correlation among attributes of dataset is shown below in Figure 3-3:

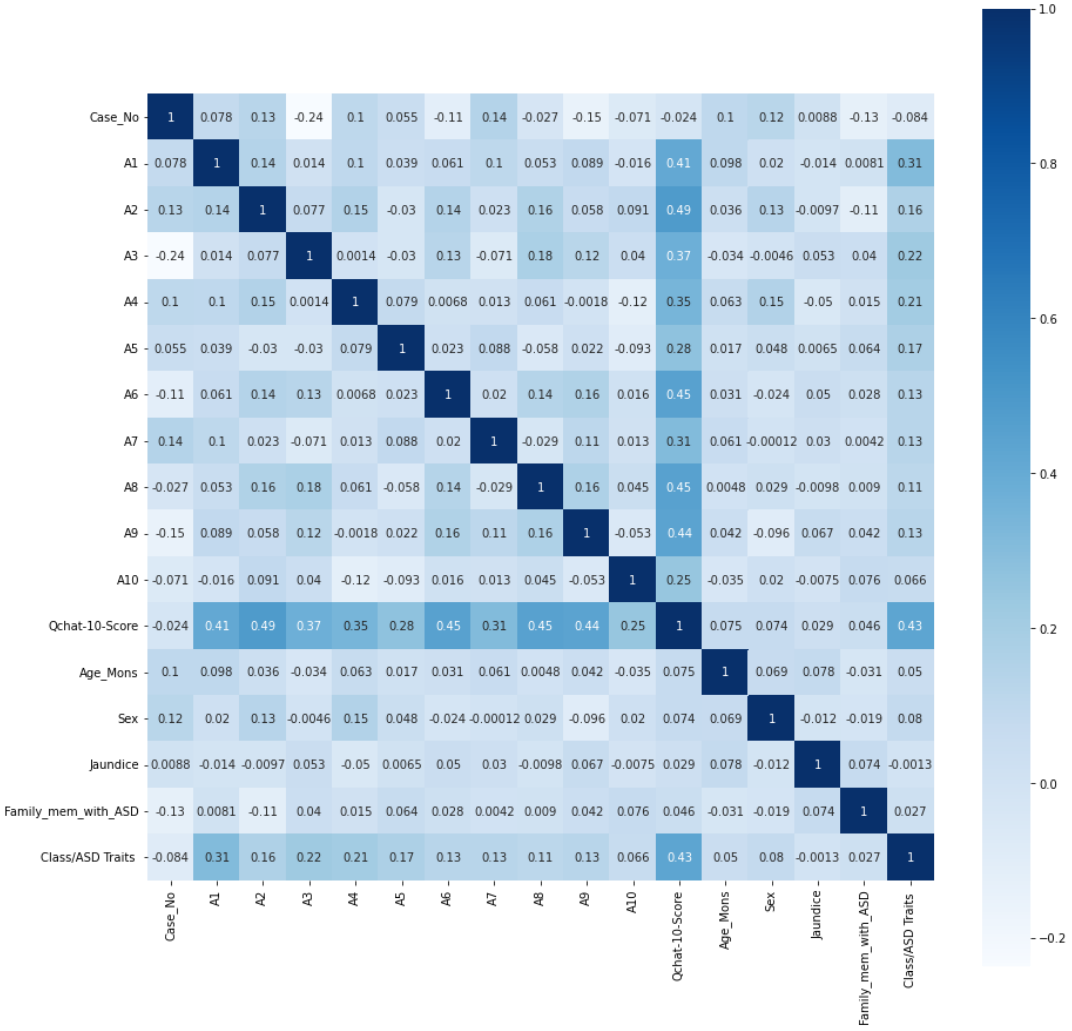

Figure 3-3 Correlation matrix of dataset

3.5.2. Visualizing Jaundice occurrence in males and females

Jaundice is nearly 2 times in males than in females as shown in Figure 3-4. This sample studies shows that Jaundice is not connected to Autism.

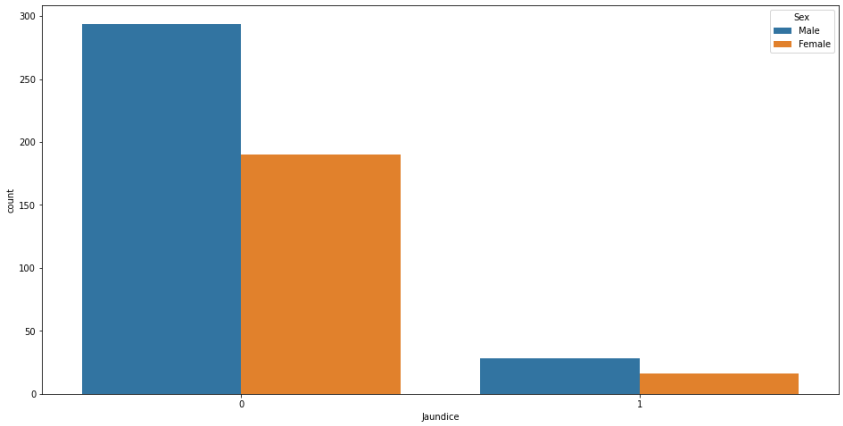

Figure 3-4 Visualization of Jaundice occurrence in males and females

3.5.3. Visualizing the Q-chat 10 score according to the gender

In general, Figure 3-5, demonstrates that the Q-Chat male responses have more positive answers than female. That shows that males are more prone to autism than females.

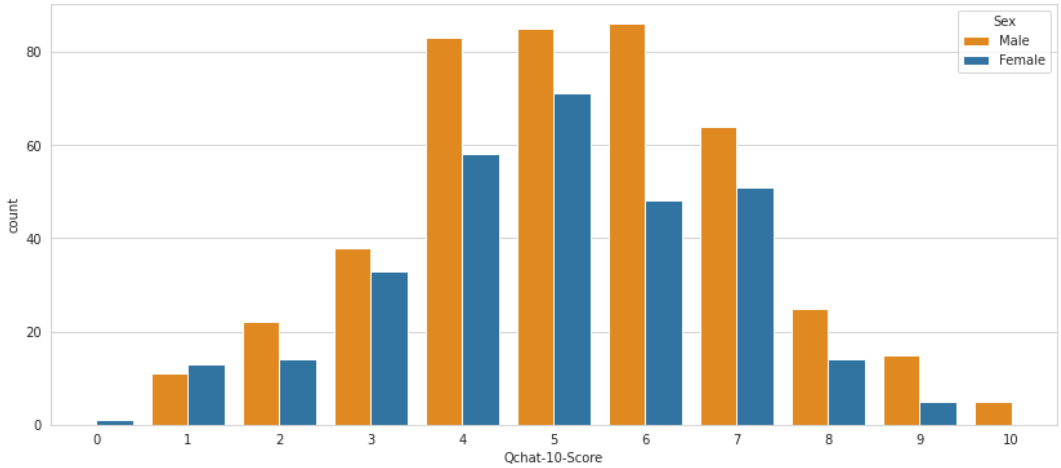

Figure 3-5 Visualization of the Q-Chat-10 score according to the gender

3.6. Train/ Test split

The train-test split approach is used for estimation of the performance of ML algorithms that are applicable to prediction-based algorithms. The dataset was split into two portions, with 80% designated for the training set and the remaining 20% for the test set. To train the auto ML tool, 80% of the data (n = 593) was used as the training set. The testing dataset comprising the remaining 20% of the data (n = 149) was reserved to assess the model's accuracy and efficiency on unseen data. Through a random division of data into training and testing sets, we were able to evaluate whether our model was overfitting or underfitting.

3.7. Model development phase

Finding the correct ML algorithm is not an easy task. We used auto ML for our dataset. To create a good model, we need to know the problem very well, the variables (instances), prepare the data, and test different parameters.

3.7.1. Applying Auto ML tool

The need for ML specialists is greater than the supply. To close this gap, progress has been made in creating user-friendly machine learning (ML) software that both novices and professionals may use. The machine learning workflow involves a substantial portion of automatically training and tuning multiple models within a user-defined time limit. The goal of auto ML is to relieve data scientists from the burden of tedious and time-consuming operations (such as designing machine learning pipelines and optimizing hyper parameters) so they may focus more effectively on tasks that are

considerably more challenging to automate. Auto ML solutions aim to automate some or all steps of the ML process which includes three common stages shown below in Figure 3-6:

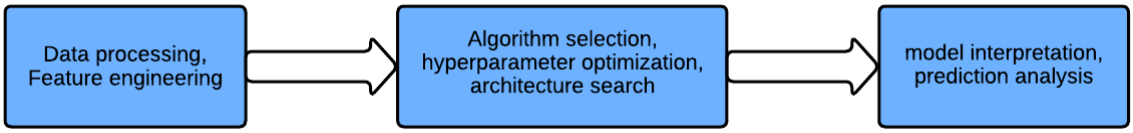

Figure 3-6 Working of auto ML

3.7.1.1. Tree-based Pipeline Optimization Tool (TPOT)

TPOT is an AutoML (Automated Machine Learning) library that automates the process of selecting and tuning machine learning models, which is coded in Python. TPOT is both free and open source. It utilizes a tree-based approach to optimize pipelines for improved performance and efficiency. Using genetic programming principles, TPOT identifies the most suitable machine learning pipeline for a given dataset to achieve optimal performance. TPOT is built upon the scikit-learn, its source code resembles that of scikit-learn. Figure 3-7 shows the TPOT framework, provided in the documentation of TPOT.

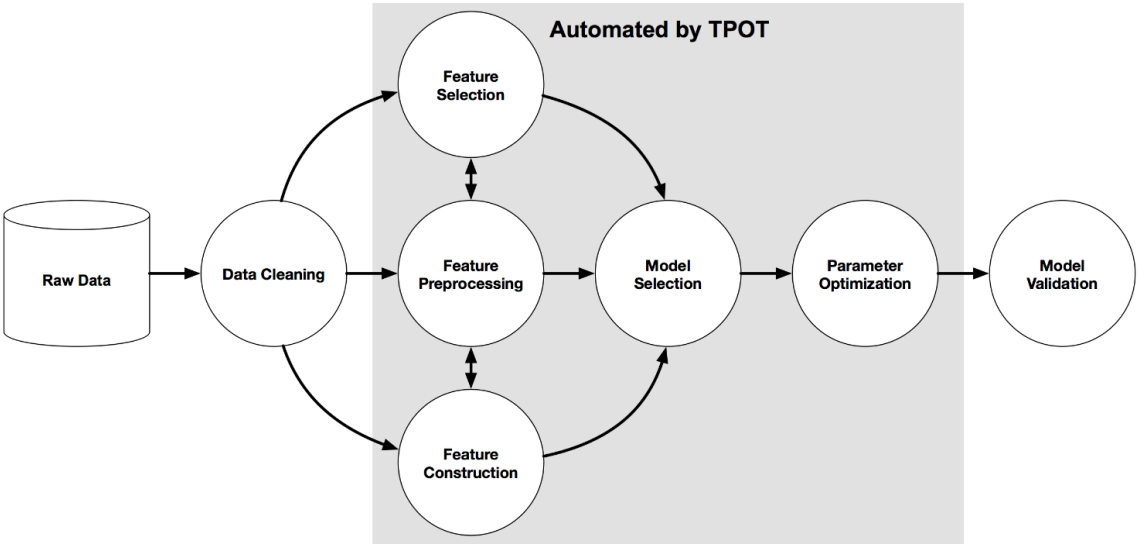

Figure 3-7 TPOT framework

For our dataset, we used TPOT open-source AutoML python library. For this purpose, we utilized google Colaboratory. Google Colaboratory, short form google colab, is a cloud platform for accessing free computing resources including GPUs. It is a hosted Jupyter notebook service. We installed TPOT library in colab by using following command as first step:

```
!pip install tpot
```

We imported the libraries needed in our implementation. In the next step, we imported our dataset into Pandas Data frame. Then the target was selected that is "Class/ASD Traits" in our dataset. After that data was divided into separate training and testing sets. In the next step, we defined the hyperparameters for TPOTClassifier, shown in Chapter 4 for each experiment performed. TPOTClassifier hyperparameters optimized are the following:

- generations: Number of iterations performed by the genetic algorithm. The default value is 100. TPOT evaluate population_size + generations x offspring_size in total. By default, offspring_size is equivalent to the population_size.
- population_size: Number of individual pipelines generated and evaluated during optimization process. The default value is 100.
- verbosity: Its value can be set to 0, 1, 2, and 3.
    - 0 → Print nothing.
    - 1 → Minimum information.
    - 2 → More information with progress bar.
    - 3 → Everything is printed, and progress bar will be visible.
- n_jobs: Number of processes running in parallel. The default value is 1.

```
[ ]  tpot = TPOTClassifier(generations=5,
                           population_size=100,
                           verbosity=2,
                           n_jobs=-1)
```

The final step is to apply fit () method and passed training data as arguments to the method. We performed this experiment with different hyperparameters for TPOTClassifier. It is further explained in Chapter 4 for each experiment performed.

CHAPTER 4

RESULTS AND EVALUATION

4.1. Performance Evaluation Measures

We evaluated performance measures using accuracy, confusion matrix, precision, recall, confusion matrix, f1-score, AUC-ROC analysis, and Precision-Recall curve.

4.1.1. Confusion matrix

It is a chart that gives a summary of accurate and inaccurate predictions or true and false classifications generated by a classification model for binary classification tasks.

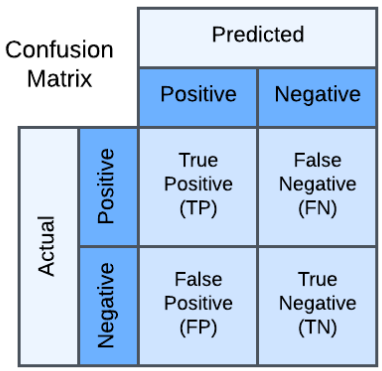

Figure 4-1 Confusion matrix

- True Positive: It refers to the instances where the positive values predicted by the model match the actual positive values accurately.
- False Positive: It denotes the instances where the model predicted positive values, but the actual values turned out to be negative.
- True Negative: It refers to the instances where the model correctly predicted negative values that align with the actual negative values.
- False Negative: The instances where the model forecasted negative values, but the actual values were positive.

4.1.2. Accuracy

The accuracy of a model is a measure of how well it performs at predicting the correct output or classification for a given output. In ML, accuracy is typically calculated by dividing the model's accurate predictions by its total predictions and multiplied by 100 to express the results as a percentage. The mathematical formula for accuracy metric is shown in Equation 1 and 1.1:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}\ x\ 100 \qquad Equation\ 1$$

$$Accuracy = \frac{TP + TN}{TP + TN\ TP + FP + TN + FN} \qquad Equation\ 1.1$$

4.1.3. Precision

It is the capacity of the model to categorize the positive values correctly. To calculate, you divide the total number of predicted positive values by the total number of actual positive values. The mathematical formula for precision metric is given as:

$$Precision = \frac{TP}{TP + FP} \qquad Equation\ 2$$

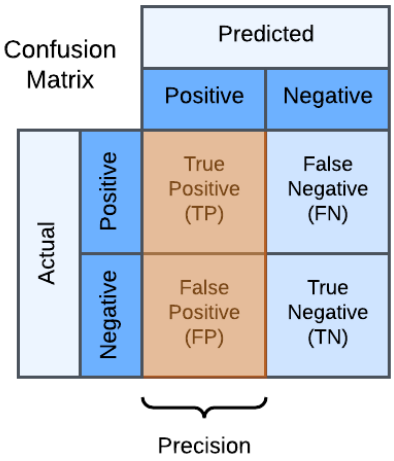

Figure 4-2 Precision

4.1.4. Macro-average precision

To obtain the macro average precision, you need to calculate the precision for every class and then find the mean of those precision scores. This metric gives equal weight to each class, regardless of their size or prevalence in the dataset. Mathematically, the macro average precision can be expressed as shown in Equation 3:

$$Macro\ Average\ Precision = (1/n) * \sum(precision\_i) \qquad Equation\ 3$$

In the above formula, "n" represents the total number of classes, and "precision_i" corresponds to the precision value of the i-th class.

4.1.5. Recall

The accuracy of a model to correctly predict positive values is known as recall. It is computed by dividing the total true positive values by the total number of positive values that are actually true. Recall is also referred to as sensitivity or true positive rate (TPR). The mathematical formula for recall metric is given below in Equation 4:

$$Recall = \frac{TP}{TP + FN} \qquad Equation\ 4$$

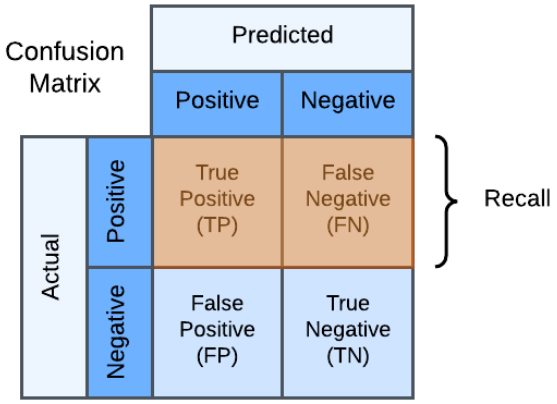

Figure 4-3 Recall

4.1.6. Macro-average recall

To determine the macro-average recall, you need to calculate the recall score for every class, which is the proportion of actual positive instances correctly predicted by the model in that class. Then, take the average of these recall scores across all classes. Mathematically, the macro-average recall can be expressed as shown in Equation 5:

*Recall_macro = (1/n) * sum(recall_i)* *Equation 5*

In the above equation, "n" refers to the total number of classes, and "recall_i" represents the recall score associated with the i-th class.

4.1.7. F1-Score

The F1 score measures the harmonic average of precision and recall. It proves to be useful when you want to consider both precision and recall in a combined metric. The mathematical formula for F1-score metric is expressed in Equation 6:

$$F1 = \frac{2 * TP}{2 * TP + FP + FN}$$ *Equation 6*

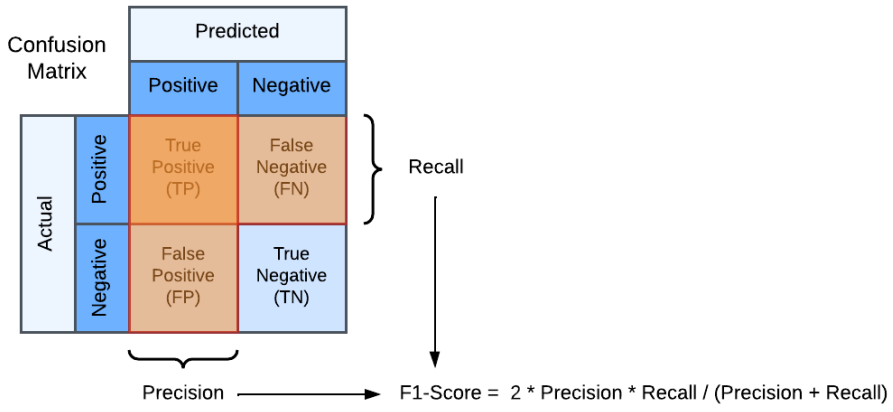

Figure 4-4 F1-Score

4.1.8. Macro-average F1-Score

To obtain the macro-average F1 score, you need to compute the F1 score for each class separately and then calculate the average of these scores. All classes are considered equally important, regardless of their frequency or size in the dataset.

The macro average F1-score is computed by determining the average of F1-scores for all classes as shown in Equation 7:

Macro Average F1-score = (F1-score_1 + F1-score_2 + ... + F1-score_n) / n
Equation 7

In the above equation, "n" represents the total number of classes present in the dataset.

4.1.9. AUC-ROC analysis

The AUC-ROC curve is a widely used metric for evaluating binary classification models based on the relationship between true positive and false positive rates at various probability thresholds. AUC-ROC values range from 0 to 1, with higher values indicating better model performance in classifying the two classes.

The ROC curve displays the probability predictions of a binary classification model by demonstrating the balance between its true positive and false positive rates at different threshold levels. The AUC value of the ROC curve measures the model's ability to distinguish between the classes, with higher values indicating better separability. Overall, the ROC curve provides predictive insights, while the AUC value quantifies the model's discriminative power.

The ROC curve represents the Sensitivity (TPR) plotted against 1-Specificity (FPR). On the y-axis there is TPR and on the x-axis there is FPR.

$$Specificity = \frac{TN}{TN + FP} \qquad Equation\ 8$$

$$FPR = 1 - Specificity \qquad Equation\ 9$$

$$FPR = \frac{FP}{FP + TN} \qquad Equation\ 9.1$$

4.1.9.1. Micro average ROC

It computes the ROC curve and AUC by aggregating the true positive, false positive, and false negative rates across all classes into a single aggregate curve, rather than computing recall separately for each class.

Micro average ROC is useful when the classes are imbalanced, and you want to prioritize the overall accuracy of the model as it gives equal importance to both classes.

4.1.9.2. Macro average ROC

It computes the ROC curve and AUC by computing the ROC curve and AUC for each individual class separately, and then averaging the results across all classes, giving equal weight to each class regardless of its size.

Macro average ROC is useful when you want to assess the model's performance across all classes equally.

4.1.10. Precision-Recall curve

It is a graphical illustration of the relationship between precision and recall for various thresholds. The graph displays precision on x-axis and recall on y-axis.

4.1.11. Micro average precision recall curve

Micro-average Precision-Recall curves are calculated by considering the individual predictions of each instance in the dataset, regardless of the class they belong to. The global precision and recall values are computed across all classes by considering the total true positives, false positives, and false negatives.

The micro-average is used to summarize the overall performance of the binary classification model. The micro-average calculates the precision and recall for each instance separately and then aggregates them to compute the overall precision and recall for the model. The micro-average precision and recall values are calculated as follows:

Micro-average precision = total true positives / (total true positives + total false positives) *Equation 10*

Micro-average recall = total true positives / (total true positives + total false negatives) *Equation 11*

4.2. Experimentation and Results

In this chapter, we go through the findings of our proposed auto ML based model. We performed three experiments with different parameters using TPOT for our dataset. Then, we verified pipeline generated by TPOT using parameters and evaluated the models using evaluation matrix discussed in Chapter 4.1. Following are the experiments performed with different parameters.

4.2.1. Experiment 1

In this experiment, we took generations=5 and population_size=10, random_state=1, while keeping the other parameters as default.

The best pipeline given by TPOT with Gradient Boosting Classifier, includes the parameters presented in Table 4-1 below:

Table 4-1 Parameters for Gradient Boosting Classifier

TPOT Best Pipeline Parameters	Values
learning_rate	0.1
max_depth	7
max_features	0.05
min_samples_leaf	16
min_samples_split	18
n_estimators	100
subsample	0.55

4.2.1.1. Classification Report

For the Non-Autistic class, the precision is 0.63, meaning that 63% of the instances predicted as non-Autistic are actually non-Autistic. The recall is 0.56, indicating that 56% of the actual non-Autistic instances are correctly predicted as non-Autistic. The F1-score for non-Autistic is 0.59, which is the harmonic mean of precision and recall for this class.

For the Autistic class, the precision is 0.83, indicating that 83% of the instances predicted as Autistic are actually Autistic. The recall is 0.87, meaning that 87% of the

actual Autistic instances are correctly predicted as Autistic. The F1-score for Autistic is 0.85, which is the harmonic mean of precision and recall for this class.

The accuracy of the model is 0.77, representing the ratio of accurately classified instances to all instances.

The macro-average F1-score is 0.72, which is the average F1-score across both classes weighted equally. The weighted average F1-score is 0.77, which is the mean F1-score of both classes, weighted by the number of instances in each class.

Overall, the model has better performance on the Autistic class than on the non-Autistic class, as evidenced by the higher precision, recall, and F1-score for the Autistic class. However, the performance of the model on the non-Autistic class is still reasonable, with a precision of 0.63 and a recall of 0.56. Classification report for Experiment 1 is given below in Table 4-2:

Table 4-2 Classification report of Experiment 1

Experiment 1	Precision	Recall	F1-Score	Support
Non-Autistic	0.63	0.56	0.59	43
Autistic	0.83	0.87	0.85	106
macro avg	0.73	0.71	0.72	149
weighted avg	0.77	0.78	0.77	149

4.2.1.2. Confusion Matrix

TP: The model correctly identified 92 positive instances.

TN: The model accurately predicted 24 negative instances as negative (class 0) when the actual label was also negative.

Figure 4-5 below shows the confusion matrix for Experiment 1:

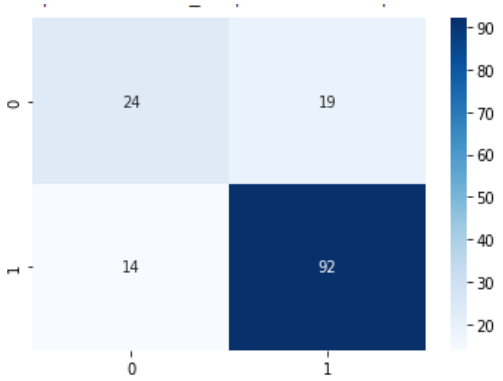

Figure 4-5 Confusion matrix for Experiment 1

4.2.1.3. Precision-Recall Curve

In Figure 4-6, the reported micro-average precision-recall of 0.86 indicates a good balance between identifying true positives (individuals with autism) and minimizing false positives (individuals without autism). In addition, the reported precision-recall values of 0.69 and 0.90 for the non-autistic and autistic classes, respectively, further support the idea that the model performs better at identifying individuals with autism compared to those without the condition.

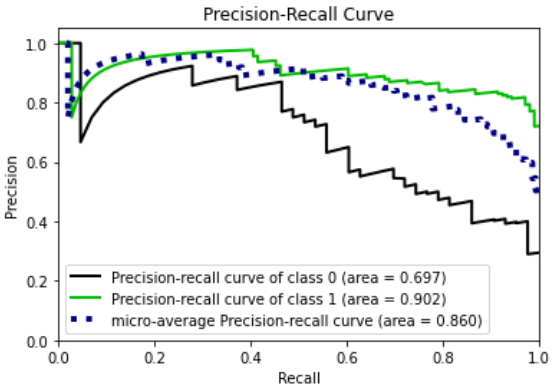

Figure 4-6 PR Curve for Experiment 1

4.2.1.4. ROC curve and AUC-ROC

In Figure 4-7, the micro average of the ROC curve is 0.87, which suggests that the overall model performance is good. The macro average of the ROC curve is 0.83, which indicates that the model is performing well for both classes.

The ROC curve for the non-autistic class has an area of 0.82, which indicates that the model has good performance in identifying non-autistic individuals. Similarly, the ROC curve for the autistic class also has an area of 0.82, suggesting that the model is performing well in identifying autistic individuals. Overall, the model seems to be performing well in distinguishing between autistic and non-autistic individuals, with similar performance for both classes.

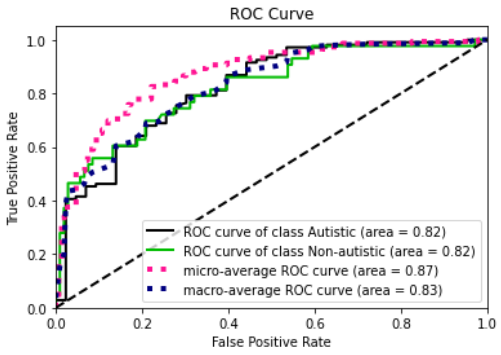

Figure 4-7 ROC Curve for Experiment 1

4.2.1.5. Verification of Experiment 1

To verify Experiment 1, we manually trained the same classifier which was produced by auto ML. For this purpose, the parameters given by TPOT pipeline are used for the Gradient Boosting Classifier, as auto ML showed that Gradient Boosting Classifier is best classifier. The results from verification of Experiment 1 are shown in

Table 4-3. Similar results were obtained as produced by auto ML. Similarly, we got the same visualizations for verification model as Experiment 1.

Table 4-3 Classification Report - Verification of Experiment 1

Experiment 1	Precision	Recall	F1-Score	Support
Non-Autistic	0.63	0.56	0.59	43
Autistic	0.83	0.87	0.85	106
macro avg	0.73	0.71	0.72	149
weighted avg	0.77	0.78	0.77	149

4.2.2. Experiment 2

In this experiment, we took generations=5 and population_size=100, random_state=1, while keeping the other parameters as default.

The best pipeline given by TPOT is an ensemble model, where the GaussianNB classifier is acting as the base model and the MLPClassifier is acting as the meta-model. In this pipeline, the Binarizer transformer is used as a preprocessing step to convert the input data into binary values, which are then passed to the GaussianNB classifier. The GaussianNB classifier produces probabilities, which are then used as input to the MLPClassifier. The MLPClassifier combines the output from the GaussianNB classifier with its own learned weights to produce the final prediction.

The ensemble is optimized using TPOTClassifier to find the best combination of hyperparameters for both the GaussianNB classifier and the MLPClassifier. The parameters of best pipeline are shown in Table 4-4 below:

Table 4-4 Parameters for Ensemble Model

TPOT Best Pipeline Parameters	Values
threshold	0.0
alpha	0.001
learning_rate_init	0. 001

4.2.2.1. Classification Report

For the non-Autistic class, the precision is 0.62, indicating that 62% of the instances predicted as non-Autistic are actually non-Autistic. The recall is 0.60, meaning that 60% of the actual non-Autistic instances are correctly predicted as non-Autistic. The F1-score for non-Autistic is 0.61, which is the harmonic mean of precision and recall for this class.

For the Autistic class, the precision is 0.84, indicating that 84% of the instances predicted as Autistic are actually Autistic. The recall is 0.85, meaning that 85% of the actual Autistic instances are correctly predicted as Autistic. The F1-score for Autistic is 0.85, which is the harmonic mean of precision and recall for this class.

The accuracy of the model is 0.78, which is the proportion of correctly classified instances out of all instances.

The macro-average F1-score is 0.73, which is the average F1-score across both classes weighted equally. The weighted average F1-score is 0.78, which is the average F1-score across both classes weighted by the number of instances in each class.

Overall, the model has better performance on the Autistic class than on the non-Autistic class, as evidenced by the higher precision, recall, and F1-score for the Autistic class. However, the performance of the model on the non-Autistic class is slightly lower, with a precision of 0.62 and a recall of 0.60. Classification report for Experiment 2 is given below in Table 4-5:

Table 4-5 Classification report of Experiment 2

Experiment 2	Precision	Recall	F1-Score	Support
Non-Autistic	0.62	0.60	0.61	43
Autistic	0.84	0.85	0.85	106
macro avg	0.73	0.73	0.73	149
weighted avg	0.78	0.78	0.78	149

4.2.2.2. Confusion Matrix

TP: The model correctly predicted 90 samples as positive (class 1) when the actual label was also positive.

TN: The model correctly predicted 26 samples as negative (class 0) when the actual label was also negative.

Figure 4-8 below shows the confusion matrix for Experiment 2:

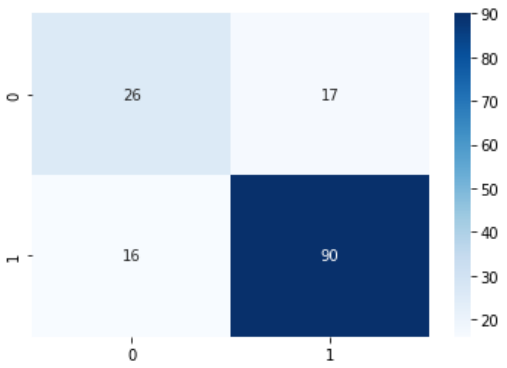

Figure 4-8 Confusion matrix for Experiment 2

4.2.2.3. Precision-Recall Curve

In the Figure 4-9, the PR for the Autistic class is 0.900, indicating that when the model predicts an instance as Autistic, it is correct 90% of the time. The PR for the Non-Autistic class is 0.713, which means that when the model predicts an instance as non-Autistic, it is correct 71.3% of the time.

In summary, the PR curve analysis shows that the model has better performance on the Autistic class, as evidenced by the higher PR score. However, the model's performance on the non-Autistic class is still reasonable, with a PR score of 0.713. The overall performance of the model is good, as indicated by the micro-average PR score of 0.852.

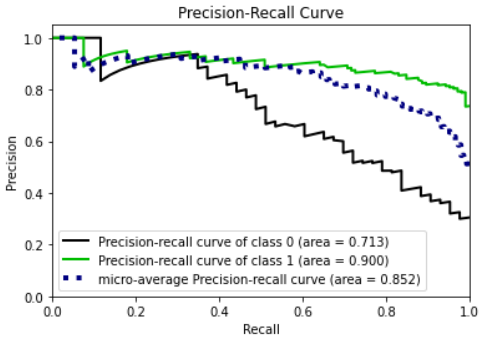

Figure 4-9 PR Curve for Experiment 2

4.2.2.4. ROC curve and AUC-ROC

In Figure 4-10, the ROC curve analysis shows that the model has similar performance on both classes, as evidenced by the equal ROC scores for the Autistic and Non-Autistic classes. The micro-average ROC score of 0.86 indicates that the model's overall performance is good, with high true positive and low false positive rates. The

macro-average ROC score of 0.82 shows that the model's performance is consistent across both classes.

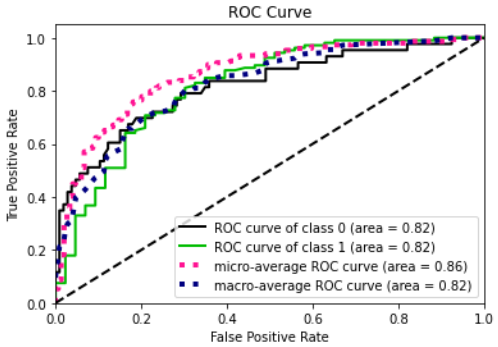

Figure 4-10 ROC Curve for Experiment 2

4.2.2.5. Verification of Experiment 2

To verify Experiment 2 we manually implemented the model. For this purpose, the parameters given by TPOT pipeline are used for the for the model's implementation along with TPOT Classifier. The results from verification of Experiment 2 are shown in Table 4-6. Similarly, we got the same visualizations for verification model as the experiment.

Table 4-6 Classification Report - Verification of Experiment 2

Experiment 2	Precision	Recall	F1-Score	Support
Non-Autistic	0.62	0.60	0.61	43
Autistic	0.84	0.85	0.85	106
macro avg	0.73	0.73	0.73	149
weighted avg	0.78	0.78	0.78	149

4.2.3. Experiment 3

In this experiment, we took generations=10 and population_size=100, random_state=1, while keeping the other parameters as default.

The best pipeline given by TPOT is an ensemble of BernoulliNB and RandomForestClassifier, while optimizing the hyperparameters using TPOTClassifier. In this pipeline, the BernoulliNB classifier is used as the base model, and the RandomForestClassifier is used as the meta model in the ensemble. The parameters for best pipeline given by TPOT are shown in Table 4-7:

Table 4-7 Parameters for Best Pipeline

TPOT Best Pipeline Parameters	Values
alpha	1.0
fit_prior	False
bootstrap	True
criterion	gini
max_features	0.05
min_samples_leaf	1
min_samples_split	4
n_estimators	100

4.2.3.1. Classification Report

For the Non-Autistic class, the precision is 0.68, indicating that 68% of the instances predicted as non-Autistic are actually non-Autistic. The recall is 0.53, meaning that 53% of the actual non-Autistic instances are correctly predicted as non-

Autistic. The F1-score for non-Autistic is 0.60, which is the harmonic mean of precision and recall for this class.

For the Autistic class, the precision is 0.83, indicating that 83% of the instances predicted as Autistic are actually Autistic. The recall is 0.90, meaning that 90% of the actual Autistic instances are correctly predicted as Autistic. The F1-score for Autistic is 0.86, which is the harmonic mean of precision and recall for this class.

The accuracy of the model is 0.78, which is the proportion of correctly classified instances out of all instances.

The F1-score for each class was averaged to calculate the macro-average F1-score, resulting in a value of 0.73. On the other hand, the weighted average F1-score, which considers the number of instances in each class, was found to be 0.78.

Overall, the model has better performance on the Autistic class than on the non-Autistic class, as evidenced by the higher precision, recall, and F1-score for the Autistic class. However, the performance of the model on the non-Autistic class is still reasonable, with a precision of 0.68 and a recall of 0.53. Classification report for Experiment 3 is given below in Table 4-8:

Table 4-8 Classification report of Experiment 3

Experiment 3	**Precision**	**Recall**	**F1-Score**	**Support**
Non-Autistic	0.68	0.53	0.60	43
Autistic	0.83	0.90	0.86	106
macro avg	0.75	0.72	0.73	149
weighted avg	0.78	0.79	0.78	149

4.2.3.2. Confusion Matrix

TP: The model correctly predicted 95 samples as positive (class 1) when the actual label was also positive.

TN: The model correctly predicted 23 samples as negative (class 0) when the actual label was also negative.

The confusion matrix for Experiment 3 is presented in Figure 4-11.

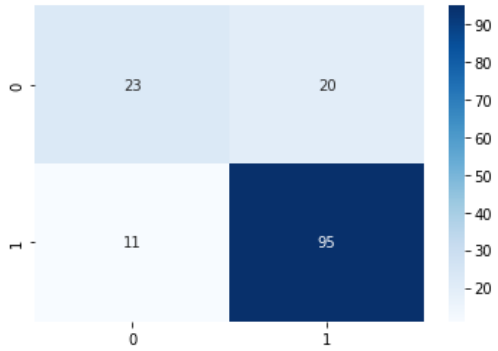

Figure 4-11 Confusion matrix for Experiment 3

4.2.3.3. Precision-Recall Curve

In the Figure 4-12, The PR for the Autistic class is 0.892, indicating that when the model predicts an instance as Autistic, it is correct 89.2% of the time. The PR for the non-Autistic class is 0.728, which means that when the model predicts an instance as non-Autistic, it is correct 72.8% of the time.

The PR curve analysis shows that the model has better performance on the Autistic class, as evidenced by the higher PR score. However, the model's performance on the non-Autistic class is still reasonable, with a PR score of 0.728. The micro-average PR score of 0.861 indicates that the model's overall performance is good, with a slight improvement compared to Experiment 2.

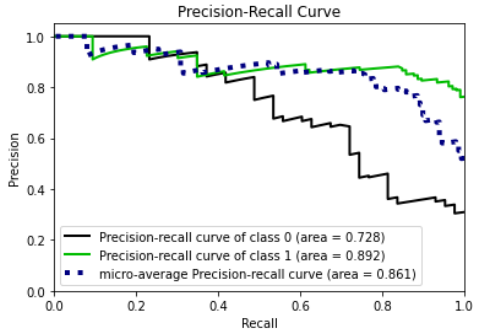

Figure 4-12 PR Curve for Experiment 3

4.2.3.4. ROC curve and AUC-ROC

In Figure 4-13, the ROC curve analysis shows that the model has similar performance on both classes, as evidenced by the equal ROC scores for the Autistic and Non-Autistic classes. The micro-average ROC score of 0.87 indicates that the model's overall performance is good, with high true positive and low false positive rates. This value represents the overall performance of the model in terms of correctly identifying all instances regardless of the class label. The macro-average ROC score of 0.81 shows that the model's performance is consistent across both classes.

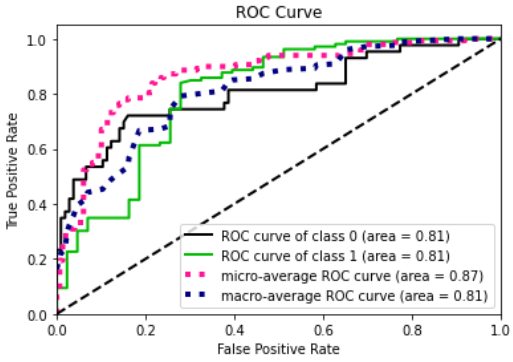

Figure 4-13 ROC Curve for Experiment 3

4.2.3.5. Verification of Experiment 3

To verify Experiment 3 we manually implemented the model. For this purpose, the parameters given by TPOT pipeline are used for the model's implementation along with TPOT Classifier. The results from verification of Experiment 3 are shown in Table 4-9. Similarly, we got the same visualizations for verification model as the experiment.

Table 4-9 Classification Report - Verification of Experiment 3

Experiment 3	Precision	Recall	F1-Score	Support
Non-Autistic	0.68	0.53	0.60	43
Autistic	0.83	0.90	0.86	106
macro avg	0.75	0.72	0.73	149
weighted avg	0.78	0.79	0.78	149

4.3. Comparison between Experimented Classifiers

We compared the results of models based on the same auto ML tool. The results are displayed for each experiment and its verification in terms of evaluation metric.

We implemented TPOT to find the best ML model and hyper parameters. We verified the models given by TPOT by manually processing those hyper parameters. The performance is evaluated in terms of evaluation metrics. TPOT achieves accuracy of 79%, precision 75%, recall 73%, and F1-score 73%. Our results show that performance is increasing as the number of pipelines increase in the experiment.

In experiment 1, the precision and recall for the non-autistic class are lower than those for the autistic class, indicating that the model has difficulty correctly identifying non-autistic individuals.

In experiment 2, the precision, recall, and F1-score for both classes are slightly higher than those in experiment 1, indicating better performance overall. Experiment 2 has the highest overall F1-score among all three experiments, which indicates that it has the best balance between precision and recall for both the autistic and non-autistic classes. The F1-score for the autistic class is 0.85, which means that the model correctly identifies 85% of the autistic cases while minimizing false positives. The F1-score for the non-autistic class is 0.61, which indicates that the model correctly identifies 61% of the non-autistic cases while minimizing false positives. Therefore, based on the F1-score, experiment 2 is the best-performing model among the three experiments.

In experiment 3, the precision and F1-score for the non-autistic class are the highest among the three experiments, while the recall for the non-autistic class is the lowest. This suggests that the model performs well in correctly predicting non-autistic individuals but may miss some cases. The precision, recall, and F1-score for the autistic class are similar to those in experiment 2.

Overall, experiment 2 and experiment 3 appear to have better performance than experiment 1.

Table 4-10 Comparison between Experiments - F1-Score

F1-Score	**Experiment 1**	**Experiment 2**	**Experiment 3**
Non-Autistic	0.59	0.61	0.60
Autistic	0.85	0.85	0.86
macro avg	0.72	0.73	0.73
weighted avg	0.77	0.78	0.78

4.4. Discussion and Comparative analysis

As mentioned in Chapter 2, studies using machine learning algorithms represent the state of the art in autism detection. Our results are compared to the several studies with several important distinctions. Currently, quantitative analysis of results is not possible due to the distinction of our approach followed. Table 4-11 below shows our work and some previous work regarding Autism Detection.

Table 4-11 Comparative Analysis

Author	**Dataset Size** (instances)	**Best Model**	**Accuracy**
Vakadkar et al., [1]	1054	LR	97.15%
Erkan et al., [5]	1100	RF	Almost 100%
Thabtah et al., [32]	1100	LR	97.94%
Tartarisco et al., [34]	315	SVM	95%
Experiment 1	742	GB Classifier	78%
Experiment 2	742	GaussianNB and MLP Classifier	78%
Experiment 3	742	BernoulliNB and RFC	79%

As we can see, many researchers got different results with different sizes of datasets. The results can be improved if we could use more data to train and test the model.

Our dataset had more autistic male than female with imbalance dataset of 742 samples. Similarly, study by Ruta et al., [33] showed more autistic male than female with dataset of 315 instances. In study [1], more males are prone to autism as compared to the females. It also showed that jaundice count is higher in male as compared to female. We plotted gender distribution graph of ASD observed in males and females. It

can be inferred that ASD is more commonly found in males than females as indicated in Figure 3-5.

The approach we followed, have not been applied by any other researcher. So, quantitative analysis of performance is not possible. Results show that model have high performance on autistic class than non-autistic class, this is due to class imbalance in dataset.

CHAPTER 5

CONCLUSION

This chapter discusses the conclusion of our work on Autism detection at early ages using auto ML TPOT. In addition, we discussed future work to extend this work.

5.1. Conclusion

In a nutshell, the purpose of thesis was to formulate a model for ASD diagnosis using auto ML to our dataset. To improve accuracy of ASD diagnosis in children and make the process faster than traditional methods. For this purpose, data of children was collected using a survey based on Q-CHAT-10. We applied auto ML TPOT to our dataset for ASD prediction and evaluated the outcomes of this technique. We used different evaluation metrics: precision, recall, F1-score, and AUC-ROC curves to evaluate the performance of our models. Auto ML TPOT shows accuracy of 79%, precision 75%, recall 73%, and F1-score 73%. The results showed that model perform better on autistic class than non-autistic class.

5.2. Future work

In the presented work, auto ML is applied to predict autism in children. There is still potential for further expansion of this work based on the contributions that have been made. The following are the potential areas for future work:

- We performed experiments using auto ML TPOT library. As a future work, we can develop auto ML model using other auto ML libraries for our dataset.
- As a future work, we can develop and compare results of two or more auto ML models based on different auto ML libraries used to develop those models.

REFERENCES

[1] K. Vakadkar, D. Purkayastha, and D. Krishnan, ‘Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques’, *SN Comput Sci*, vol. 2, no. 5, Sep. 2021, doi: 10.1007/s42979-021-00776-5.

[2] A. Sharma and D. P. Tanwar, *Deep Analysis of Autism Spectrum Disorder Detection Techniques.* 2020.

[3] A. S. Teja, A. S. Abhay, D. Mounika, and M. V. Pujitha, ‘Autism Spectrum Disorder Detection Techniques’, in *2022 International Conference on Communication, Computing and Internet of Things, IC3IoT 2022 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/IC3IOT53935.2022.9767927.

[4] L. Zwaigenbaum, J. A. Brian, and A. Ip, ‘Early detection for autism spectrum disorder in young children’, *Paediatrics and Child Health (Canada)*, vol. 24, no. 7, pp. 424–432, Oct. 2019, doi: 10.1093/pch/pxz119.

[5] U. Erkan and D. N. H. Thanh, ‘Autism Spectrum Disorder Detection with Machine Learning Methods’, *Current Psychiatry Research and Reviews*, vol. 15, no. 4, pp. 297–308, Nov. 2019, doi: 10.2174/2666082215666191111121115.

[6] V. Zope, T. Shetty, M. Dandekar, A. Devnani, and P. Meghrajani, ‘ML based Approaches for Detection and Development of Autism Spectrum Disorder: A Review’, in *International Conference on Sustainable Computing and Data Communication Systems, ICSCDS 2022 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 79–84. doi: 10.1109/ICSCDS53736.2022.9761040.

[7] Q. Tariq, J. Daniels, J. N. Schwartz, P. Washington, H. Kalantarian, and D. P. Wall, ‘Mobile detection of autism through machine learning on home video: A development and prospective validation study’, *PLoS Med*, 2018.

[8] K. S. Omar, P. Mondal, N. S. Khan, Md. R. K. Rizvi, and M. N. Islam, *A Machine Learning Approach to Predict Autism Spectrum Disorder*. Dhaka: IEEE, 2019.

[9] A. Roman-Urrestarazu *et al.*, ‘Autism screening and conditional cash transfers in Chile: Using the Quantitative Checklist (Q-CHAT) for early autism detection in a low resource setting’, *Autism*, vol. 25, no. 4, pp. 932–945, May 2021, doi: 10.1177/1362361320972277.

[10] by S. Jack Damico, N. Müller, M. J. Ball, and P. A. Prelock, ‘The Handbook of Language and Speech Disorders, Second Edition. Edited Autism Spectrum Disorders’, 2021.

[11] I. A. Ahmed *et al.*, ‘Eye Tracking-Based Diagnosis and Early Detection of Autism Spectrum Disorder Using Machine Learning and Deep Learning Techniques’, *Electronics (Switzerland)*, vol. 11, no. 4, Feb. 2022, doi: 10.3390/electronics11040530.

[12] S. Raj and S. Masood, ‘Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques’, in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 994–1004. doi: 10.1016/j.procs.2020.03.399.

[13] A. A. Hemu, R. B. Mim, M. M. Ali, M. Nayer, K. Ahmed, and F. M. Bui, ‘Identification of Significant Risk Factors and Impact for ASD Prediction among Children Using Machine Learning Approach’, in *2022 2nd International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies, ICAECT 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICAECT54875.2022.9808043.

[14] H. S. Nogay and H. Adeli, ‘Machine learning (ML) for the diagnosis of autism spectrum disorder (ASD) using brain imaging’, *Rev Neurosci*, vol. 31, no. 8, pp. 825–841, Dec. 2020, doi: 10.1515/revneuro-2020-0043.

[15] N. Zaman, J. Ferdus, and A. Sattar, ‘Autism Spectrum Disorder Detection Using Machinelearning Approach’, in *2021 12th International Conference on Computing Communication and Networking Technologies, ICCCNT 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ICCCNT51525.2021.9579522.

[16] P. McCarty and R. E. Frye, ‘Early Detection and Diagnosis of Autism Spectrum Disorder: Why Is It So Difficult?’, *Seminars in Pediatric Neurology*, vol. 35. W.B. Saunders, Oct. 01, 2020. doi: 10.1016/j.spen.2020.100831.

[17] M. D. Hossain, M. A. Kabir, A. Anwar, and M. Z. Islam, ‘Detecting autism spectrum disorder using machine learning techniques: An experimental analysis on toddler, child, adolescent and adult datasets’, *Health Inf Sci Syst*, vol. 9, no. 1, Dec. 2021, doi: 10.1007/s13755-021-00145-9.

[18] M. Bala, M. H. Ali, M. S. Satu, K. F. Hasan, and M. A. Moni, ‘Efficient Machine Learning Models for Early Stage Detection of Autism Spectrum Disorder’, *Algorithms*, vol. 15, no. 5, May 2022, doi: 10.3390/a15050166.

[19] M. M. Rahman, O. L. Usman, R. C. Muniyandi, S. Sahran, S. Mohamed, and R. A. Razak, ‘A review of machine learning methods of feature selection and classification for autism spectrum disorder’, *Brain Sciences*, vol. 10, no. 12. MDPI AG, pp. 1–23, Dec. 01, 2020. doi: 10.3390/brainsci10120949.

[20] Md. M. Rahman, O. L. Usman, R. C. Muniyandi, S. Sahran, S. Mohamed, and R. A. Razak, ‘A Review of Machine Learning Methods of Feature Selection and Classification for Autism Spectrum Disorder’, *Brain Sci*, 2020, doi: 10.3390/brainsci10120949.

[21] R. Romero-García, R. Martínez-Tomás, P. Pozo, F. de la Paz, and E. Sarriá, 'Q-CHAT-NAO: A robotic approach to autism screening in toddlers', *J Biomed Inform*, vol. 118, Jun. 2021, doi: 10.1016/j.jbi.2021.103797.

[22] F. Thabtah and D. Peebles, 'A new machine learning model based on induction of rules for autism detection', *Health Informatics J*, vol. 26, no. 1, pp. 264–286, Mar. 2020, doi: 10.1177/1460458218824711.

[23] M. Alsuliman and H. H. Al-Baity, 'Efficient Diagnosis of Autism with Optimized Machine Learning Models: An Experimental Analysis on Genetic and Personal Characteristic Datasets', *Applied Sciences*, 2022.

[24] K. S. Tejaswi, K. Meghavarshini, and P. Nivedhitha, 'Autism Prediction using ML Algorithms', in *2022 1st International Conference on Computational Science and Technology (ICCST)*, IEEE, Nov. 2022, pp. 1–6. doi: 10.1109/ICCST55948.2022.10040300.

[25] J. Zeidan *et al.*, 'Global prevalence of autism: A systematic review update', *Autism Research*, vol. 15, no. 5. John Wiley and Sons Inc, pp. 778–790, May 01, 2022. doi: 10.1002/aur.2696.

[26] V. Vishal, A. Singh, Y. B. Jinila, C. Kavitha, S. P. Shyry, and J. Jabez, 'A Comparative Analysis of Prediction of Autism Spectrum Disorder (ASD) using Machine Learning', in *2022 6th International Conference on Trends in Electronics and Informatics, ICOEI 2022 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 1355–1358. doi: 10.1109/ICOEI53556.2022.9777240.

[27] M. Marlow, C. Servili, and M. Tomlinson, 'A review of screening tools for the identification of autism spectrum disorders and developmental delay in infants and young children: recommendations for use in low- and middle-income countries', *Autism Research*, vol. 12, no. 2. John Wiley and Sons Inc, pp. 176–199, Feb. 01, 2019. doi: 10.1002/aur.2033.

[28] Z. A. Taha Ahmed and M. E. Jadhav, ‘A Review of Early Detection of Autism Based on Eye-Tracking and Sensing Technology’, in *Proceedings of the 5th International Conference on Inventive Computation Technologies, ICICT 2020*, Institute of Electrical and Electronics Engineers Inc., Feb. 2020, pp. 160–166. doi: 10.1109/ICICT48043.2020.9112493.

[29] S. G. Jacob, M. M. B. A. Sulaiman, and B. Bennet, ‘Feature Signature Discovery for Autism Detection: An Automated Machine Learning Based Feature Ranking Framework’, *Comput Intell Neurosci*, vol. 2023, pp. 1–14, Jan. 2023, doi: 10.1155/2023/6330002.

[30] S. Islam, T. Akter, S. Zakir, S. Sabreen, and M. I. Hossain, ‘Autism Spectrum Disorder Detection in Toddlers for Early Diagnosis Using Machine Learning’, in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, CSDE 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020. doi: 10.1109/CSDE50874.2020.9411531.

[31] D. Stevanović, ‘Quantitative Checklist for Autism in Toddlers (Q-CHAT): A psychometric study with Serbian Toddlers’, *Res Autism Spectr Disord*, vol. 83, May 2021, doi: 10.1016/j.rasd.2021.101760.

[32] F. Thabtah, ‘An accessible and efficient autism screening method for behavioural data and predictive analyses’, *Health Informatics J*, vol. 25, no. 4, pp. 1739–1755, Dec. 2019, doi: 10.1177/1460458218796636.

[33] L. Ruta *et al.*, ‘Validation of the Quantitative CHecklist for Autism in Toddlers (Q-CHAT) in an Italian clinical sample of young children with Autism and Other Developmental Disorders’, *Frontiers in Psychiatry*, vol. 10, no. JUN. Frontiers Media S.A., 2019. doi: 10.3389/fpsyt.2019.00488.

[34] G. Tartarisco *et al.*, ‘Use of machine learning to investigate the quantitative checklist for autism in toddlers (Q-CHAT) towards early autism screening’, *Diagnostics*, vol. 11, no. 3, Mar. 2021, doi: 10.3390/diagnostics11030574.

[35] A. Niedźwiecka and E. Pisula, ‘Symptoms of Autism Spectrum Disorders Measured by the Qualitative Checklist for Autism in Toddlers in a Large Sample of Polish Toddlers’, *Int J Environ Res Public Health*, vol. 19, no. 5, Mar. 2022, doi: 10.3390/ijerph19053072.

[36] N. Farooqi, F. Bukhari, and W. Iqbal, ‘Predictive Analysis of Autism Spectrum Disorder (ASD) using Machine Learning’, in *Proceedings - 2021 International Conference on Frontiers of Information Technology, FIT 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 305–310. doi: 10.1109/FIT53504.2021.00063.

[37] A. Truong, A. Walters, J. Goodsitt, K. Hines, C. B. Bruss, and R. Farivar, ‘Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools’, *IEEE*, Aug. 2019, doi: 10.1109/ICTAI.2019.00209.

[38] M. Bahri, F. Salutari, A. Putina, and M. Sozio, ‘AutoML: state of the art with a focus on anomaly detection, challenges, and research directions’, *International Journal of Data Science and Analytics*, vol. 14, no. 2. Springer Science and Business Media Deutschland GmbH, pp. 113–126, Aug. 01, 2022. doi: 10.1007/s41060-022-00309-0.

[39] H. Eldeeb, M. Maher, O. Matsuk, A. Aldallal, R. Elshawi, and S. Sakr, ‘AutoMLBench: A Comprehensive Experimental Evaluation of Automated Machine Learning Frameworks’, Apr. 2022, [Online]. Available: http://arxiv.org/abs/2204.08358

[40] C. Allison *et al.*, ‘The Q-CHAT (Quantitative CHecklist for Autism in Toddlers): A normally distributed quantitative measure of autistic traits at 18-24 months of age: Preliminary report’, *J Autism Dev Disord*, vol. 38, no. 8, pp. 1414–1425, Sep. 2008, doi: 10.1007/s10803-007-0509-7.

[41] C. Allison, B. Auyeung, and S. Baron-Cohen, ‘Toward brief “red flags” for autism screening: The short Autism Spectrum Quotient and the short Quantitative

Checklist in 1,000 cases and 3,000 controls', *J Am Acad Child Adolesc Psychiatry*, vol. 51, no. 2, pp. 202-212.e7, 2012, doi: 10.1016/j.jaac.2011.11.003.