A COMPARATIVE ANALYSIS OF FEATURE SELECTION TECHNIQUES USING AUTOMATED MACHINE LEARNING TOOLS

Rana Tuqeer Abbas
01-241221-007

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

Department of Software Engineering

BAHRIA UNIVERSITY ISLAMABAD

August 2024

Approval For Examination

Scholar's Name: Rana Tuqeer Abbas Registration No. 01-241221-007

Program of Study: __MS (Software Engineering)__

Thesis Title: A Comparative Analysis of Feature Selection Techniques Using Automated Machine Learning Tools.

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 ___ 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, Rana Tuqeer Abbas hereby state that my MS thesis titled "A Comparative Analysis of Feature Selection Techniques Using Automated Machine Learning Tools" 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: Rana Tuqeer Abbas(01-241211-007)

Date: ______________________________

Plagiarism Undertaking

I, Rana Tuqeer Abbas, solemnly declare that research work presented in the thesis titled "A Comparative Analysis of Feature Selection Techniques Using Automated Machine Learning Tools" 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: Rana Tuqeer Abbas(01-241211-007)

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 want to acknowledge my parents, who supported me and. This thesis is dedicated to my parents, who have always supported me through challenges in life and prayed for my success. I am truly thankful for the support in every aspect whether that is financial, emotional, or mental.

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

AutoML (Automated Machine Learning) is the field that seeks to automate the process of developing machine learning models. AutoML is created to boost productivity and efficiency by automating as much of the process that occurs when machine learning is applied, which streamlines the workflow from data preprocessing to model deployment, especially as it considered important for feature selection process. In this study, we use two popular AutoML frameworks, TPOT and KNIME, to compare numerous feature selection methods. Feature selection is a crucial step in machine learning pipeline, as it involves identifying the most relevant features that improve models ability. Effective feature selection can improve model accuracy, reduce overfitting, and enhance interpretability by focusing the key attributes. In this study, we used the autism spectrum disorder (ASD) dataset which is collected from multiple rehabilitation centres in Pakistan, our goal is to determine which features offer the best model for the diagnosis of autism spectrum disorder (ASD). TPOT and KNIME both demonstrated their capability in identifying ASD, achieving impresive accuracy rates of 85.23% and 83.89%, respectively. The evaluation metrics precision, recall, and F1-Score, among others—verified the models' reliability as well. The proposed frameworks and their feature selection methods enhanced the overall approach of the model in addition to identifying important features which have a strong impact on the model. Using these AutoML frameworks not only optimised the feature selection process, but also greatly reduced the amount of time required for diagnosis. This study demonstrates how AutoML approaches, and feature selection techniques can be used to improve model efficiency, which will help with early detection and improve outcomes for children with ASD and their families.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ABBREVIATIONS

AutoML	-	Automated Machine Learning
ASD	-	Autism Spectrum Disorder
ML	-	Machine Learning
TPOT	-	Tree-based Pipeline Optimization Tool
KNIME	-	Konstanz Information Miner
WHO	-	World Health Organization
AHPS	-	Algorithm And Hyperparameter Space
ADOS	-	Autism Diagnostic Observation Schedule
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
NAS	-	Neutral Architecture Search
PCOS	-	Polycystic Ovary Syndrome
RFC	-	Random Forest Classifier
ROC	-	Receiver Operating Curve

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Machine learning (ML) has slowly invaded every part of our life, and the beneficial impact has been amazing but Automated Machine Learning (AutoML) is emerging as a way to speed the integration of ML into additional applications and real-world scenarios. Machine learning has become an important tool in a variety of fields, including healthcare, finance, image classification, and fraud detection. To increase the efficiency and effectiveness of machine learning models, feature selection approaches are used to determine the most important features to model. The selection of appropriate features is an important stage in the machine learning pipeline since it has a direct impact on model performance and interpretability. In recent years, Automated machine learning tools have increased its importance because they can automate model selection, hyperparameter tuning, and feature selection.

Automated machine learning (AutoML) systems have improved the model generation process by features selection techniques and hyperparameters automatically. Large, complex datasets can be handled by AutoML approaches, which can also effectively choose key features for building models [1] [2]. A crucial phase in the automated machine learning process is feature selection. Finding the most pertinent characteristics for result prediction is its goal, especially when working with high-dimensional data. Feature selection improves the readability of models and helps avoid over-fitting by lowering the number of dimensions [3].

The retention of features with barely the duplication and a significant connection to the predicted target variable is the fundamental goal of feature selection [4]. AutoML technologies have made machine learning tasks like feature selection easier recently. This paper investigates the performance of several feature selection methods in AutoML systems. Through the automation of processes like feature selection and model training, JADBIO is an AutoML platform that facilitates the development of predictive models, particularly for biological and biomedical data. The quantity and complexity of the dataset being utilized determine the approach used by the JADBIO algorithm and hyperparameter space (AHPS) [5]. When working with data that contains several features, supervised unsupervised and semi-supervised approaches should be used side by side [6]. Now a days mostly domains, including sentiment analysis [7], intrusion detection [8], diseases diagnosis [9][10], and stock price prediction [11], are impacted by feature selection process. It improves prognostic accuracy in the medical field for diseases like Autism Spectrum Disorder (ASD) [13] and Polycystic Ovary Syndrome (PCOS) [12]. Autism Spectrum Disorder (ASD) is a neurodevelopmental disease characterised by difficulties in speech, behaviour, and social relationships. Early detection and intervention are critical to improve outcomes for children with ASD. Specifically, In our research we used different AutoML frameworks to detect ASD in a dataset gathered from various rehabilitation centres in Pakistan.

Formal diagnosis of ASD is an extensive process in which the average wait period for ASD detection in United Kingdom is over 3 years . 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 [14] [15]. Early action to enhance language and communication skills, as well as the overall children with autism [16] [17]. ASD diagnosis can be made at an early age of 18 – 24 months. The formal diagnosis of ASD is frequently delayed until the age of 4 years due to variation in symptoms . There are various screening methods for ASD such as AQ, SCQ, and M-CHAT, CARS-2, and STAT [18]. Early diagnosis leads to early treatment which in return enhance the life of those who have ASD. Data from various screening methods have been used to detect autism with the help of machine learning (ML) technique. For screening in ASD children's, Q-CHAT 10 is used for this purpose [19] [20].

1.1.Motivation

The goal of this study is to automatically generate machine learning models including data preprocessing, algorithm selection and hyperparameter optimisation which represent workarounds for malaria diagnosis using PJ materials. The aim of AutoML is to optimize the machine learning workflow saving time and energy for researchers or practitioners which can be dedicated to more complex activities such as problem definition, interpretation of results. Feature selection - as the name suggests, is used to select only those features from the data we have that has more relevance or impact on our model prediction. Model will wisely chose and select the features so that we have better results, less overfitting and a good model. We can additionally speed up the process of getting to production-ready, high performing machine learning models by adding feature selection into our AutoML pipeline. And this will ultimately lead to better and smarter results for all types of applications including healthcare diagnostics.

1.2.Research Gap

Machine learning methodologies has made significant progress, but more exploration is needed on AutoML techniques and feature selection methods. While various features have been utilized in existing studies but there is limited research on systematically ranking of these features using Automated Machine Learning (AutoML) that contribute the most to accurate ASD detection.

1.3.Problem Statement

Feature selection is crucial for constructing robust and efficient machine learning models. It identifies the most relevant features, enhancing forecast accuracy while reducing model complexity. Autism Spectrum Disorder (ASD) cannot be diagnosed through a medical test, making diagnosis challenging. The current clinical process for ASD detection is complex and time-consuming. To address this issue, an optimized model for ASD detection can be developed using Automated Machine Learning (AutoML) tools that focus on the most relevant features. This approach not only enhances the accuracy of the diagnosis process but also simplifies and speeds up the detection of ASD.

1.4.Research Objectives

The objective of this study to eliminate human participation in autism prediction by utilizing Automated Machine Learning techniques. The study's goal is to use AutoML approaches to determine whether a youngster is prone to ASD, which can help with early diagnosis. This may result in better treatment for children with ASD at an early stage of the disorder.

1.5.Research Questions

RQ 1: How TPOT framework will be compared with KNIME data analytics tool for ASD detection?
RQ 2: How to systematically rank features and determine their importance?

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 and KNIME. We evaluated our models using various metrics, such as precision, recall, F1-score, and AUC-ROC curves.
- We conducted a comparison between the models generated by KNIME and TPOT, finding that TPOT gives better accuracy results. This indicates that the TPOT workflow gives the best model based on our evaluation as compared to KNIME.

1.7.Outline of this thesis

The organization of this paper is as follows: **Chapter 1** “Introduction” section includes introduction of 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 to extend this work.

CHAPTER 2

LITERATURE REVIEW

Automated Machine Learning (AutoML) methods have gained importance because they aim to automate processes in Machine Learning (ML) pipelines such feature selection and hyperparameter optimization. Automated Machine Learning (AutoML) systems improve machine learning workflows by automating processes like data preprocessing, algorithm selection, and hyperparameter tuning. Feature extraction is the most important part of AutoML, which indentifies and extract relevant features.

Automated machine learning (AutoML) aims to develop optimal machine learning solutions based on a problem description, task type, and datasets. It might relieve data scientists of the time-consuming manual tuning process and allow domain experts to use off-the-shelf machine learning systems without extensive knowledge. Feature selection is an important phase in the machine learning process since it allows you to discover the most relevant features that contribute to a model's prediction performance. Automated machine learning tools have been increasingly popular in recent years, providing a mechanism to automate the feature selection process and other parts of building a model [1] [2].

High-dimensional analysis of data is a significant challenge in machine learning, and feature selection provides an effective approach for overcoming this problem [3]. As long as class labels are accessible, feature selection algorithms in supervised learning tasks aim to optimise some function of predicted accuracy. One of the main principles behind feature selection is the idea that "a good feature subset is one that contains features that are highly correlated with the class but uncorrelated with each other" [4].

This study [5] shows how JADBIO an AutoML tool which use the Algorithm and HyperParameter Space (AHPS) technique to extract features, customizing its methodology based on the dataset's dimensions and size. Even with a small number of records, this method ensures accurate and efficient predictive and diagnostic model generation by identifying the most significant features from high-dimensional data. This research [6] investigated the compatibility of five commonly used feature selection approaches in data mining research for sentiment analysis using an online movie review dataset. Their research focused on comparing feature selection and machine learning algorithms for sentiment analysis, which shed insight on the efficacy of various feature selection methods.

In this study [7] used a comparative analysis of supervised learning techniques and the Fisher Score feature selection algorithm to detect intrusions, introducing a knowledge of the role of feature selection in improving the performance of supervised learning techniques for intrusion detection. To automate the diagnosis of Polycystic Ovary Syndrome (PCOS), Kuzhippallil and Josephadopted in [8] explained various classification algorithms and a hybrid feature selection strategy to minimise the number of features. Understanding the importance of feature selection in automating disease diagnosis has been aided by this study. Furthermore, Khagi, Kwon, and Lama et al., [9] hence they work on liver disease prediction by using various methods of feature selection along with classification models. This research gave more light about the effect of feature selection techniques on disorders classification model. This research [10] demonstrated the use of automated machine learning to estimate yield and biomass in three broadacre crop types using high spatial resolution hyperspectral data has shown a promising improvement through feature selection. Li et al., [11] also presented a summary of the use of deep learning networks in predicting stock prices and stressed that selecting good features for useful models to get better prediction results is essential. Proposed a reliable and interpretable methodology to automatically evaluate credit fraud detection, including feature selection and compelling ML techniques, emphasizing the significance of feature selection techniques in developing effective fraud detection systems [12].

Autism Spectrum Disorder (ASD) is a neurological disorder that affects social interaction, communication, and behaviour. Early detection and diagnosis of ASD are critical for prompt intervention and better results to overcome this situation. In recent years, automated machine learning approaches have gained popularity as an effective approach for detecting ASD. Raj and Masood et al., [13] investigated on the analysis and detection of autism spectrum disorder using machine learning techniques. Their results showed a high accuracy rate of 70.22% in ASD detection, confirming the usefulness of automated machine learning in this medical field. D. Peebles et al., [15] discusses the issues of ASD screening, emphasising the need for increased diagnostic accuracy and speed. A new ML technology which is Rules Machine Learning not only detects autism symptoms but also generates interpretable rule sets for doctors and carers.

M. Tomlinson & M. Marlow et al ., [18] M. Tomlinson & M. Marlow et al., [17] emphasise the importance of affordable, brief screening tools for DD and ASD in children in LMICs, where resources are limited. They suggest 10 promising techniques but note challenges due to resource constraints and cultural differences, emphasizing the need for ongoing research and collaboration. Ruta et al. [21] employed an Italian clinical sample to validate the psychometric features in Q-CHAT questionnaire, Quantitative measure established solely for autism, not any other neurological illness. This study involved 315 youngsters. They compared young autistic children (n = 139) with DD (n = 50) and TD children (n = 126). All of the statistics for the three study groups were also discussed. Q-CHAT scores were greater in group of autistic people as compare to DD & TD groups.

In this study [22], result showed that boys were more likely than girls to have ASD, while age has no effect on Q-CHAT scores. Farooqi et al., [23] noted the issues encountered during the data gathering procedure in countries such as Pakistan, where there is no tracking or reporting of ASD cases. Thabtah., [25] initially collected that data, which is now publicly available. The feature signatures and their importance in differentiating between classes to predict autism are described for the first time in this work. M. Cerrada *et al.*, [26] The systems eventually came to different decision tree techniques; H2O favoured ensembles of XGBoost models, whereas TPOT produced various kinds of stacked models. According to the convergence of both AutoML systems, pipelines focusing on very similar subsets of features across all problems can handle numerous problems in this domain with up to 90% accuracy using a relatively small collection of 10 common features. M. A. Moni et al., [27] ASD in Bangladesh identified 8 key features out of 23 for diagnosing autism in children aged 16-30 months using J48 decision tree.

However, the literature review identifies certain knowledge gaps. While previous research has thoroughly explored the use of feature selection techniques in various domains such as disease diagnosis, fraud detection, and stock price prediction etc, there is an a lack of comprehensive analysis that compares the performance of various feature selection techniques across multiple tools. Overall, the literature review provides a comprehensive understanding of the current study of feature selection techniques using automated machine learning tools, emphasizing the need for further research to overcome existing knowledge gaps and advance the field.

Table 2-1 Reviewed Research Work

Ref.	Year	Key Findings	Limitations
[1]	2022	The study identified economic attributes indicating logistics performance, with PCA and Elastic-net providing the best key features, and the ANN model being the most effective predictor.	Focusing mainly on economic attributes in feature selection may overlook other factors, limiting the logisitic performance.
[5]	2023	Used AutoML and feature ranking to find nonclinical signals for early autism detection, obtaining approximately 90% Mathew's coefficient and 95% balanced accuracy. AutoML techniques showed to be more adaptable and easier to apply than deep learning, which obtained a maximum accuracy of 92.7%.	The study's reliance on specific datasets may limit it's ability to generalize to larger populations.
[7]	2018	The study used AutoML on the CICIDS2017 dataset and Fisher Score algorithm to select the best features for detecting DDoS attacks. Classification was performed using Support Vector Machine(SVM),K-Nearest Neighbour (KNN), and Decision Tree (DT) algorithms, achieving success rates of 99.97%, 57.76%, and 99%, respectively.	The study limited its application by focussing only on DDoS attacks, and its feature reduction increased KNN performance but decreased SVM accuracy. To overcome these constraints, large-scale data processing and deep learning will be used in future research.
[9]	2019	Through the use machine learning algorithms, manual feature ranking, and convolutional neural network (CNN) feature extraction techniques, the study was able to identify Alzheimer's disease from brain MRI images with a high classification accuracy (98–99%). Compared to using CNN solely feature selection strategies enhanced classification performance.	The high classification accuracy might be limited to the specific dataset. Future work should explore generalizability to other datasets and refine feature selection techniques.

[12]	2021	Using a dataset containing 39 features, created a machine learning model to automate the diagnosis of Polycystic Ovary Syndrome (PCOS). The model's performance was enhanced by the hybrid feature selection method; the Support Vector Machine with a Linear kernel (Linear SVM) achieved the maximum recall (80.6%), accuracy (91.6%), and precision (93.6%).	The study's findings are based on a specific dataset of 541 subjects, which may limit their use to bigger sample size.
[15]	2020	The research offers a new approach to machine learning, known as Rules Machine Learning (RML), which enhances ASD screening's expected accuracy, sensitivity, & specificity.	Due of their rarity, the study excludes toddler-related cases and faces issues with imbalanced datasets.
[23]	2023	Models which used to detect ASD in this paper: Logistic Regression (LR) and : Support Vector Machine (SVM), getting 81% accuracy in adults with SVM and 98% accuracy in children with LR. It emphasizes how machine learning has the capability to support accurate & timely diagnosis of ASD in a range of age groups.	This study does not provide a comprehensive screening method optimized for ASD detection and relies on existing datasets, which may limit generalizability. Future research should explore transfer-learning models or deep learning techniques which enhance accuracy & assess severity of ASD.
[25]	2022	Using AutoML systems (H2O DAI and TPOT), the researchers achieved great classification accuracy more than 96% in detecting gearbox faults (fractured teeth, pitting, and cracking). AutoML models frequently outperform hand-tuned ones. Time-domain statistical features were very informative, and features had effect on accuracy, simplifying the modelling procedure. Both AutoML systems identified shared characteristics that were consistently useful across all failure situations.	The study did not explore the use of other AutoML platforms, such as Neural Architecture Search (NAS) or Bayesian techniques like Auto-Sklearn. Furthermore, the results are specific to the datasets and failure modes examined, necessitating additional validation before applying the findings to other mechanical systems or operational settings.

CHAPTER 3

RESEARCH METHODOLOGY

3.1. Introduction

This chapter describes study framework that uses Automated Machine Learning (AutoML) techniques to extract key features for model creation. Our research methodology process consists of several steps. Data is initially gathered and then pre-processed before being used for creating the model. After that data then partitioned for training & testing datasets. Finally model training with Auto ML, followed by model verification, performance evaluation and then feature extraction.

3.2. Proposed Methodology

The organised approach utilised in this study to create and assess AutoML models for diagnosing autism spectrum disorder (ASD) and feature selection as shown in Figure 3-1. Approach begins with collection of data, which gathers crucial information about ASD symptoms and related traits. This data is then partitioned to form training and testing datasets, ensuring that the models can be properly trained and validated. Two AutoML platforms, TPOT and KNIME, are used to automatically construct and optimise machine learning pipelines. These systems simplify the model development process by automating feature selection, model selection, and hyperparameter tweaking. The top models from TPOT and KNIME are then compared to determine the best model based on performance criteria.

We performed two experiments and results obtained from those experiments were compared for evaluation, experiment 1 and 2 are explained further in Chapter 4. In our study precision, recall and F1-score used for evaluation metrics.

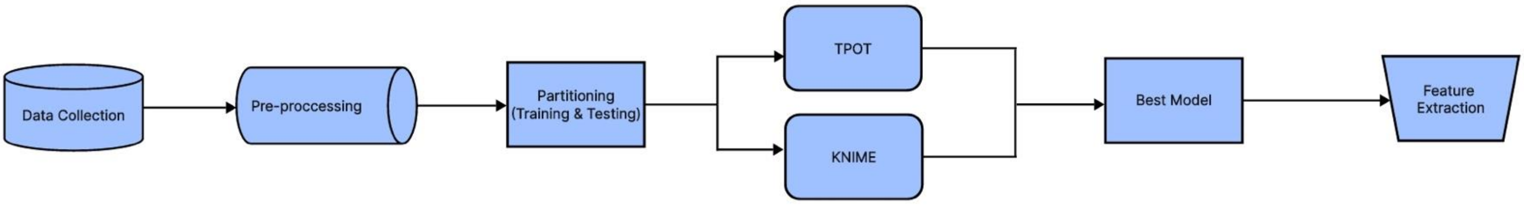


Figure 3-1 Proposed AutoML Framework

3.3. Data Collection Method

We collected information from a few Pakistani Rehabilitation Centers. A lot of samples were gathered straight from the parents of autistic children. After that, the information was put into a spreadsheet and stored in text file format. The lack of efficient re-reporting and tracking of autism cases across the nation made the task challenging. Since data for this study needed to be gathered from scratch in both hard and soft formats, a questionnaire based on the Q-CHAT screening technique was developed. Data was gathered in a soft form using Google Form. The responses gathered using Google Forms were downloaded in CSV format. Each response was then categorized based on predefined criteria, framing the problem as a classification problem.

Allison et al. [29] proposed Q-CHAT to minimize the time required to fill out the form, allowing a huge population to do so. Initially, it comprises of 25 questions. Allison et al. later proposed Q-CHAT-10 [30], a questionnaire with only ten questions. It provides a variety of response categories. It is quick to give because a higher score indicates autistic features. The Cambridge Local Research Ethics Committee accepted this study, which used Q-CHAT.

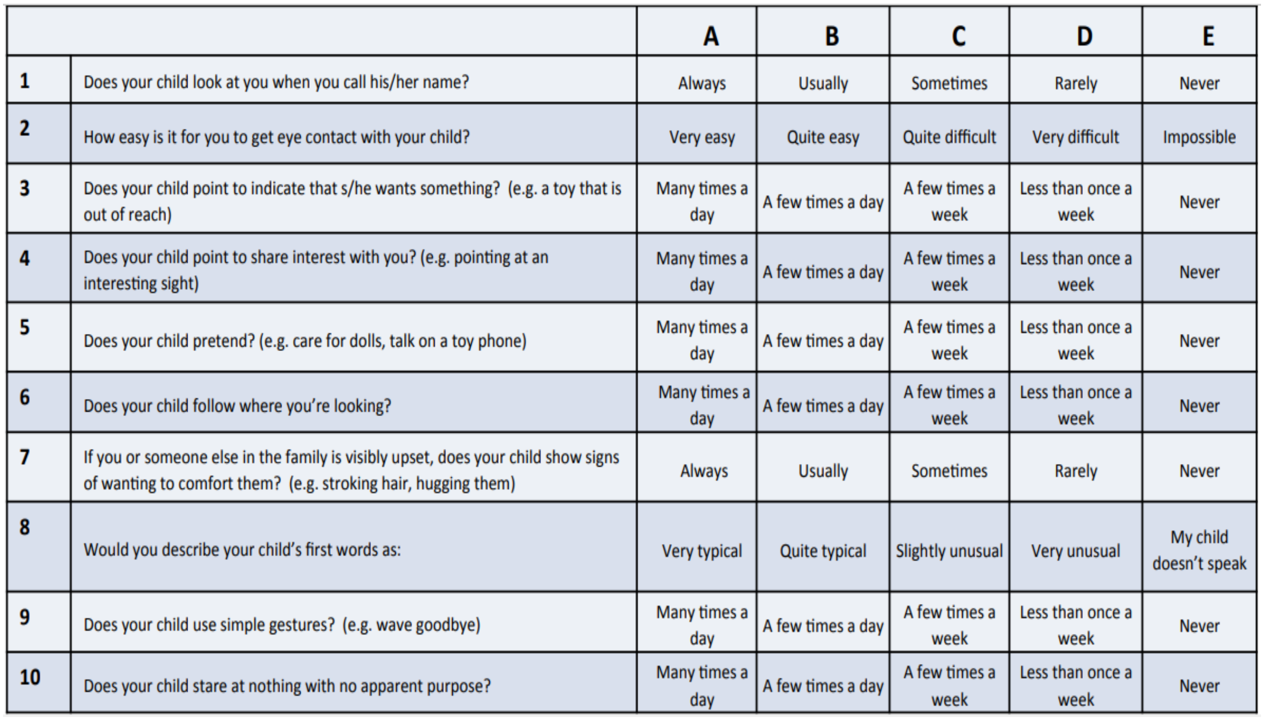
3.3.1. Q-CHAT-10

Therapy can help reduce the symptoms of ASD. Early detection is therefore preferred. To achieve this, various screening techniques are applied. Q-CHAT was created by Allison et al. [29] to identify ASD in children. Q-CHAT was first created with twenty-five pieces. Later, Q-CHAT-10, which consists of just 10 questions, was proposed by Allison et al., [30].

Based on the discrimination index (DI) for every item in the derivative sample, the top ten items for Q-CHAT-10 were chosen. We used Q-CHAT-10 for our study since the shorter form of the test produced the greatest outcomes in the trials. Based on statistical analysis of a broader collection of questions, the Q-CHAT-10 also emphasizes the most crucial questions for diagnosing ASD.

When choosing a response from columns C, D, or E in Table 3-1, one point should be given for each question that is connected to the selected answer. Regarding Question 10, each selected response from columns A, B, or C gets one point as well. These points will thereafter be included. Healthcare providers may suggest a multidisciplinary assessment for the child if the score is higher than 3/10. Q-CHAT-10 is short for Q-CHAT with ten questions. In 3-1 Table. 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 and Description

Dataset Variable	Data Type	Attribute Description
A1	Binary (0, 1)	Outcome depends on the screening method which is used
A2	Binary (0, 1)	Outcome depends on the screening method which is used
A3	Binary (0, 1)	Outcome depends on the screening method which is used
A4	Binary (0, 1)	Outcome depends on the screening method which is used
A5	Binary (0, 1)	Outcome depends on the screening method which is used
A6	Binary (0, 1)	Outcome depends on the screening method which is used
A7	Binary (0, 1)	Outcome depends on the screening method which is used
A8	Binary (0, 1)	Outcome depends on the screening method which is used
A9	Binary (0, 1)	Outcome depends on the screening method which is used
A10	Binary (0, 1)	Outcome depends on the screening method which is used
Age_Mons	Number	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	Int	Final score based on the scoring function
Class/ASD Traits	Boolean	A score of "0" indicates the lack of ASD traits, while a score of "1" indicates their existence. The class name reflects the presence of certain characteristics.

3.3.2. Our Dataset Samples

We transformed the data after collecting it from several rehabilitation centers, as mentioned in Section 3.3.1, for each of the questions from A1 to A10. Table 3-3 displays our dataset samples following transformation.

Table 2-3 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 an unprocessed or noisy dataset into a format suitable for analysis and training is called data pre-processing. Cleaning up the data at this stage involves removing any errors or inconsistencies. Initially, we searched our data for any missing numbers and eliminated them. The group features in our dataset were converted to the binary values 0 and 1. Two groups of sex traits—male and female—have been encoded. For yes and no, jaundice is set to one and zero, respectively. ASD Class/Traits are set to 1 for people with autism and 0 for people without autism. Unused attributes like "Case_No" and "Who completed the test" have been eliminated.

3.5. TPOT: Tree-based Pipeline Optimization Tool

TPOT is a library for automating procedure for selecting the best Machine Learning model and hyperparameters, saving the time and improving the results. Instead of manually testing alternative models and configurations for data, TPOT uses genetic programming to explore a variety of Machine Learning pipelines and select the one that is most suited to your individual dataset. Tree-based Pipeline Optimisation (TPOT) represents a pipeline model using a binary decision tree structure. This covers data preparation, algorithm modelling, hyperparameter tuning, model selection and feature selection as shown in figure 3-2.

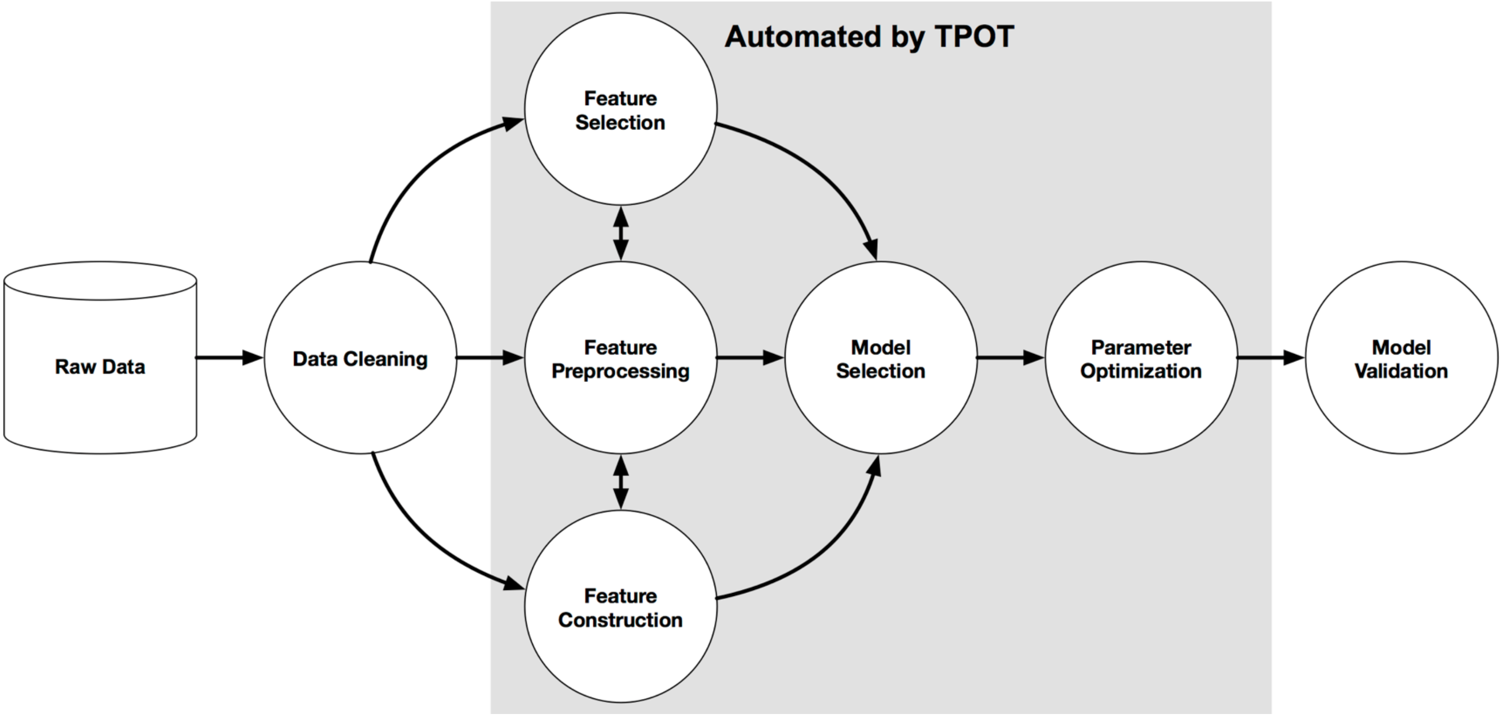


Figure 3-1 TPOT Pipeline

How does TPOT Model Selection Work?

- **Initial:** Started with generating an initial set of diverse machine learning pipelines. It includes different machine learning algorithms, preprocessing steps, and feature selection techniques.
- **Evaluation:** These pipelines are evaluated based on their performance using a predefined scoring metric, such as accuracy. The performance of each pipeline is assessed on a validation set.
- **Selection:** The top-performing pipelines are selected to form the next generation. This step makes sure that the best pipeline kept for further process.

- **Crossover:** By combining elements of many pipelines from the present generation, TPOT builds new pipelines. Crossover refers to this process, in which biological evolution is equivalent to genetic recombination.
- **Mutation:** Intended to try new pipelines, TPOT applies some random changes for the pipeline. This mutation adds a little randomness maybe change hyperparameter values, add preprocessing steps or even swap algorithms.
- **Iteration:** Reptation is simply the process that applies selection, crossover and mutation for multiple generation. As TPOT continues to iterate, it will slowly refine the pipelines until they start improving.
- **Best Pipeline:** After completing a set number of generations, TPOT identifies the best-performing pipeline from the final generation. This optimal pipeline is then recommended for deployment on the dataset.

The ability of TPOT to automate the complete machine learning workflow is one of the main reasons why AutoML uses it. Conventional model selection and hyperparameter adjustment can be difficult and complicated. By automating feature selection procedures, preprocessing stages, algorithm selection, and hyperparameter tuning, TPOT solves this problem and enables users of various skill levels to utilise advanced machine learning techniques. Moreover, TPOT is made to integrate easily with scikit-learn, one of the most widely used Python machine learning libraries. This compatibility allows users to integrate TPOT into their existing machine learning workflows easily. It is capable of handling various types of data and task, including as regression, classification, and even certain unsupervised learning situations. By streamlining the model development process and enabling the discovery of high-performing pipelines, TPOT accelerates the deployment of machine learning solutions, thereby fostering innovation and efficiency across various domains.

3.6. KNIME (Konstanz Information Miner)

KNIME is an open-source platform for analysis of data, reporting, and automation that makes machine learning workflow creation and implementation easier. KNIME's straightforward, user-friendly interface enables users to visually build data workflows as can find in figure 3-3 and making it suitable for both beginners and specialists in data science.

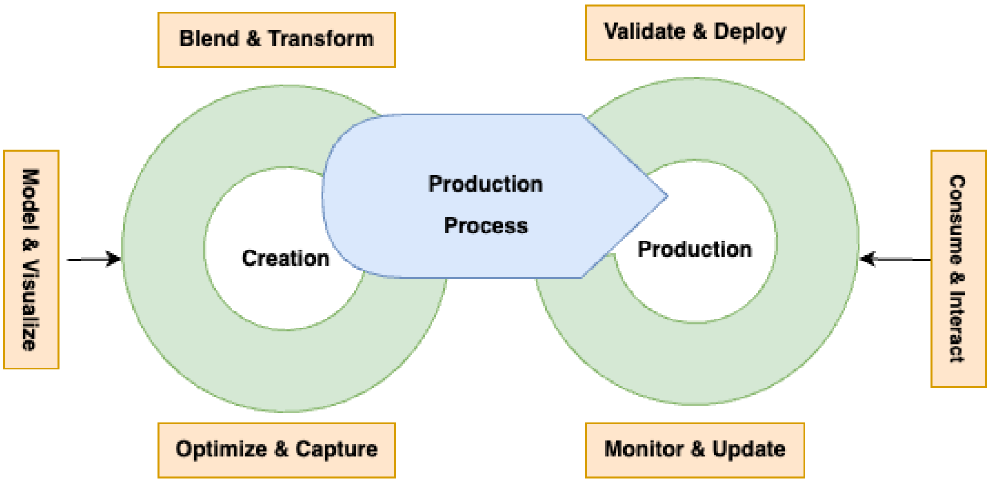


Figure 2-3 KNIME Working

KNIME begins by importing data via nodes like the CSV Reader. Preprocessing steps such as cleaning, normalization, and transformation are then applied. After that data is splitted into training & testing sets by Partitioning node, usually with an 80-20 ratio for robust evaluation. AutoML capabilities in KNIME are handled by nodes that automate model selection and hyperparameter tuning, evaluating different models to find the best one. The Scorer node assesses model performance using metrics like accuracy, precision, recall, and F1-Score. Analysis for feature signature identifies some important features contributing to model accuracy. KNIME supports iterative optimization, allowing users to refine workflows by modifying pipelines and parameters.

After validation, KNIME allows you to deploy your model and make real time predictions. Given its end-to-end set of data ingestion, pre-processing, model selection/evaluation to deployment tools . KNIME provides an excellent platform for a fast and efficient development cycle making machine learning models. The flexibility of KNIME is particularly noteworthy as it has no issue handling large-scale data processing jobs as well small research questions equally easily. KNIME offers tools and connectors that support AutoML workflows in all kinds of computing settings, from local model deployment to cloud-based infrastructure. Also, factor in KNIME's strong integration capabilities with multiple big data platforms and databases which magnify its relevance to the enterprise. KNIME gives graphical interface for many complex tasks and can also be used by users who have less technical knowledge of coding but still allow user friendly interface.

CHAPTER 4

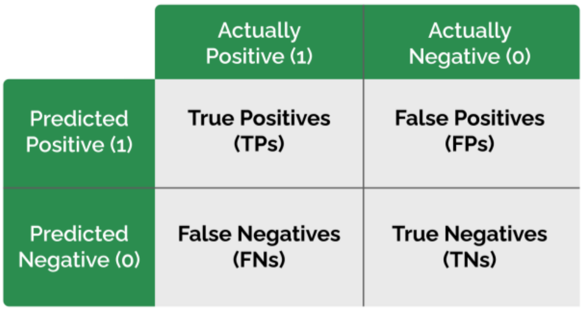
RESULTS AND EVALUATION

4.1. Performance Evaluation Measures

In our study, we compared performance of two popular AutoML tools, KNIME and TPOT, on our data set. We evaluated the results of models produced by these tools through various performance evaluation metrics such as accuracy which is the percentage of correct prediction over total, precision shows how many actual positives were actually positive; recall measures those predicted true in all real world occurrences and finally F1-score keeps weighted average between values obtained from precision & recall. In addition, we used the ROC-AUC statistic to assess the model's ability to differentiate between positive and negative classifications. Using these criteria, we assured a thorough evaluation of the models' efficacy, robustness, and reliability, giving us a clear picture of their performance on our dataset.

4.1.1. Confusion Matrix

It's actually a performance evaluation for a classification task in machine learning, with a possibility for two or more classes as the result. Four distinct combinations of the expected and actual values are shown in the figure.



	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 3-1 Confusion Matrix

True Positive: Predicted positive and it’s true.

False Positive: Predicted negative and it’s true.

True Negative: Predicted positive and it’s false.

False Negative: Predicted negative and it’s false.

4.1.2. Accuracy

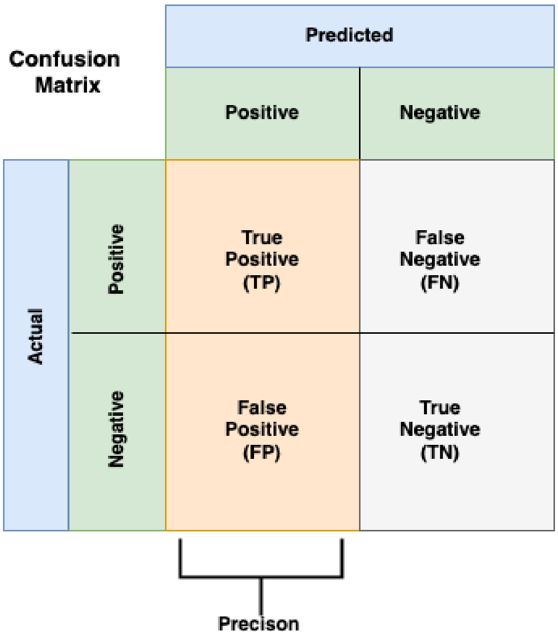
Accuracy refers to the proportion of correctly classified instancesout of all the instances in the datset. Accuracy is a metric that indicates how frequently a machine learning model accurately predicts an outcome. To calculate accuracy, divide the number of correct predictions by the total number of predictions. Mathematically accuracy is defined in Equation 1:

$$Accuracy = \frac{Number\ of\ correctly\ classified\ instances}{Total\ number\ of\ instances} \qquad Equation\ 1$$

4.1.3. Precision

In binary and multiclass classification tasks, precision is a metric used to assess a model's performance, especially in terms of positive class (or a class of interest). It calculates the percentage of true positive predictions—that is, correctly predicted positive cases—among all the instances the model predicts as positive. The mathematical formula for precision metric given as:

$$\mathit{Precision} = \frac{\mathit{True\ Positvie}}{\mathit{True\ positive{+}False\ Psitive}} \qquad \mathit{Equation\ 2}$$



Confusion Matrix		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Figure 4-2 Precision

4.1.4. Recall

Recall also knowns as sensitivity or True Positive Rate (TPR), is a metric used to evaluate the ability of a model to identify all positive instances, including those that are positive (True Positive) and those that are incorrectly predicted as negative (False Negative). The mathematical formula for recall metric is given below in Equation 3:

$$\mathit{Recall} = \frac{\mathit{True\ Positive}}{\mathit{True\ Psoitive{+}\ False\ Negative}} \qquad \mathit{Equation\ 3}$$

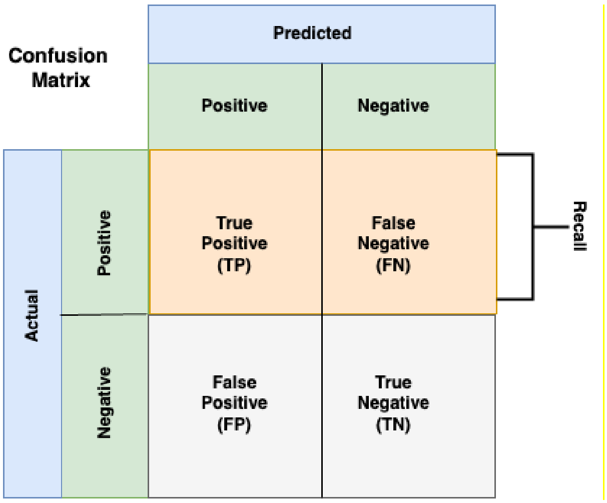


Figure 4-3 Recall

4.1.5. F1-Score

F-1 score is metric which is used to balance the precision and recall of a model into a single score. It is the harmonic mean of precision and recall , providing a single metric that captures both aspect of model’s performance. Mathematically F-1 score is expressed in Equation 4:

$$F1\ Score \ = \ 2\ x\ \frac{Precision\ x\ Recall}{Precision+Recall} \qquad \textit{Equation 4}$$

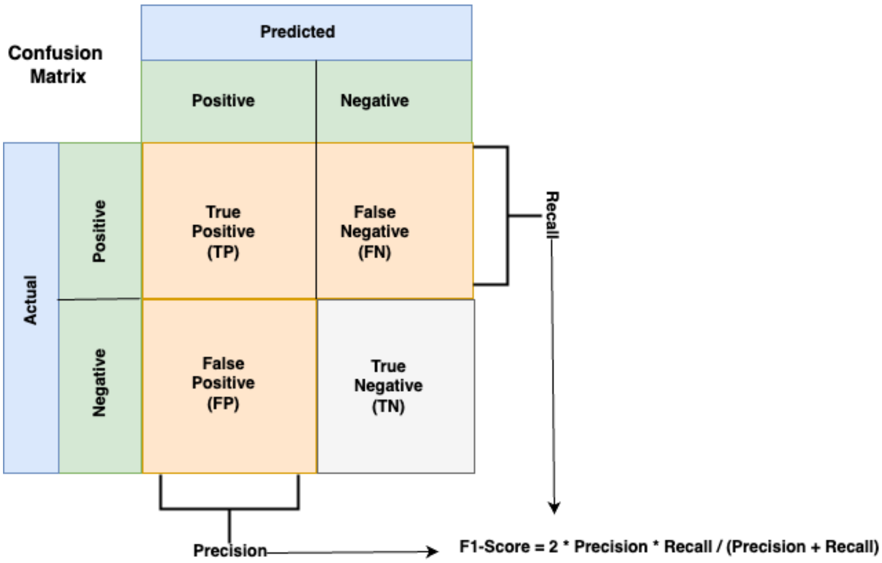


Figure 4-4 F1 Score

4.1.6. ROC Analysis

Receiver Operating Characteristic (ROC) analysis is a technique used for evaluating the performance of classification models. It involves plotting the ROC curve, which reflects the trade-offs between True Positive Rate (TPR) and False Positive Rate (FPR) at various threshold values. The curve is a graphical representation that shows the TPR (sensitivity or recall) on the y-axis and the FPR (1 - specificity) on the x-axis. The area under the ROC curve (AUC-ROC) is a single statistic that summarises the model's performance. It ranges from 0 to 1, with 1 indicating a perfect classifier, 0.5 representing a model with no discrimination capacity (random guessing), and values less than 0.5 indicating poor performance. ROC analysis is very useful for determining suitable thresholds, comparing multiple models, and evaluating models on imbalanced datasets because it is unaffected by class distribution.

$$Specificity = \frac{TN}{TN + FP}$$

Equation 5

$$TPR = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Equation 6

$$FPR = \frac{False\ Positives}{False\ Positives + True\ Negatives}$$

Equation 7

4.1.6.1. Micro average ROC

Rather than computing recall separately for each class, it aggregates the true positive, false positive, and false negative rates across all classes to produce a single aggregate curve that computes the ROC curve and AUC. Because it assigns equal balance to both classes, the micro average ROC is helpful when there is an imbalance between the classes and you want to focus on the overall accuracy of the model.

4.1.6.2. Macro average ROC

In multi-class classification, the macro average ROC approach is used to assess a model's performance by averaging the AUC scores and ROC curves for each class separately. By considering each class as a one-versus-all binary classification problem, the ROC curve and the Area Under the Curve (AUC) are calculated for each class independently in this method. Score of macro average ROC and macro average AUC are helpful for checking the performance of all classes consistently.

4.2. Experimentation and Results

In this part, we will discuss the results of our proposed AutoML based model for ASD detection. Using our dataset, we conclude two experiments with TPOT and KNIME. Each workflow is created to evaluate the best efficiency and accuracy of the model created by these AutoML tools.

4.2.1. Experiment 1 (Applying TPOT)

In the first experiment, we streamlined the development of machine learning pipeline by applying TPOT framework to our dataset. TPOT choose the best optimal model by applying evolutionary algorithm. The evaluation metrics described in section 4.1 used to verify and asses the best pipeline generated by TPOT.

For better results, we first split the dataset into training and testing with the ratio of 80-20. TPOT was then configured with specific parameters, including a set number of generations and population size, to thoroughly explore the solution space. We defined and configured the TPOT classifier with specific parameters:

- Generations: 10
- Population size: 200
- Scoring metric: Accuracy
- Verbosity: 2 (to provide detailed logs)
- Random state: 42 (for reproducibility)

Throughout the optimization process, TPOT generated and evaluated multiple pipelines over several generations. After completing 10 generations, TPOT identified the best pipeline, which included a Random Forest classifier as the optimal model. This model achieved an impressive accuracy of 85.23% on the testing dataset, demonstrating its robustness and effectiveness. The classification report of the optimal model is shown in figure 4-5.

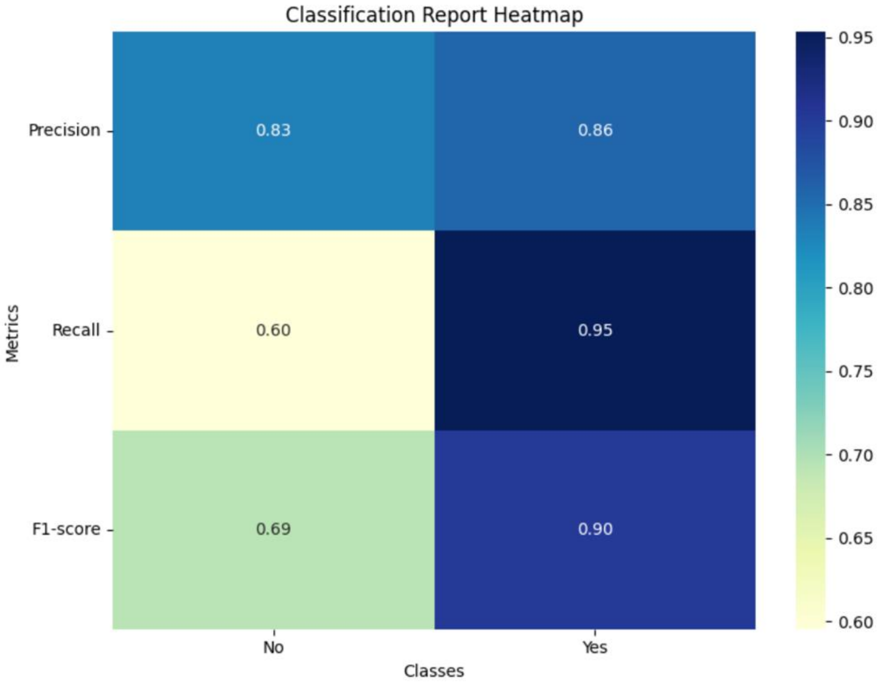


Figure 4-5 Classification Report of TPOT

The Random Forest classifier's performance of TPOT can be seen in Figure 4-6 confusion matrix. It demonstrates that there were 30 actual "No" instances, 25 were accurately predicted as "No" (true negatives), while 5 were incorrectly predicted as "Yes" (false positives). Of the 119 actual "yes" instances, 102 were true positives (TP), whereas 17 were false negatives (FN). Overall, the confusion matrix demonstrates that the classifier performs well in identifying the positive class.

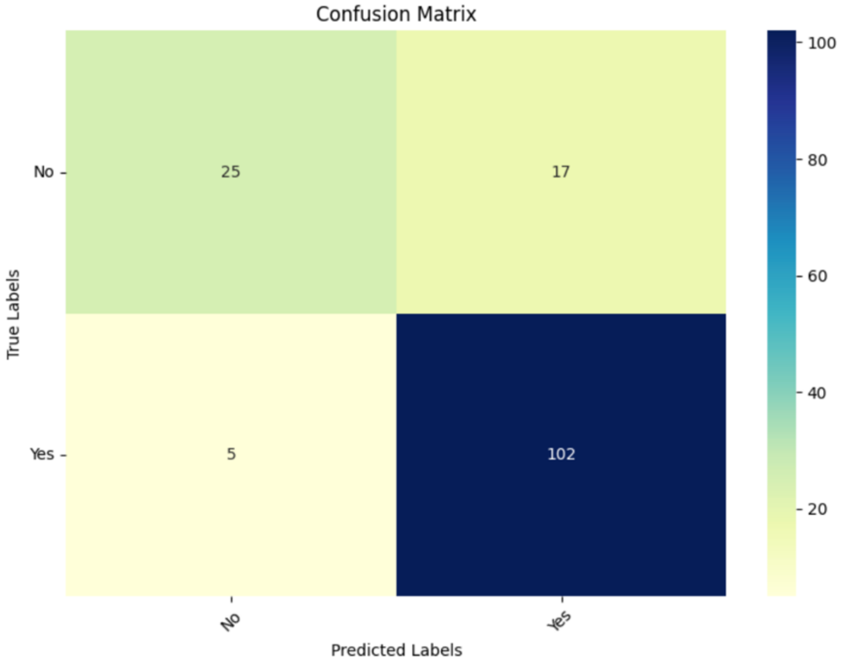


Figure 4-6 Confusion Matrix of TPOT

Receiver Operating Characteristic (ROC) curve of the model, which compares the True Positive Rate (TPR) to the False Positive Rate (FPR) in Figure 4-7, were used to assess the Random Forest classifier's performance. With an Area Under the Curve (AUC) of 0.86% on the generated ROC curve, the model is very capable of differentiating between the positive and negative classes.

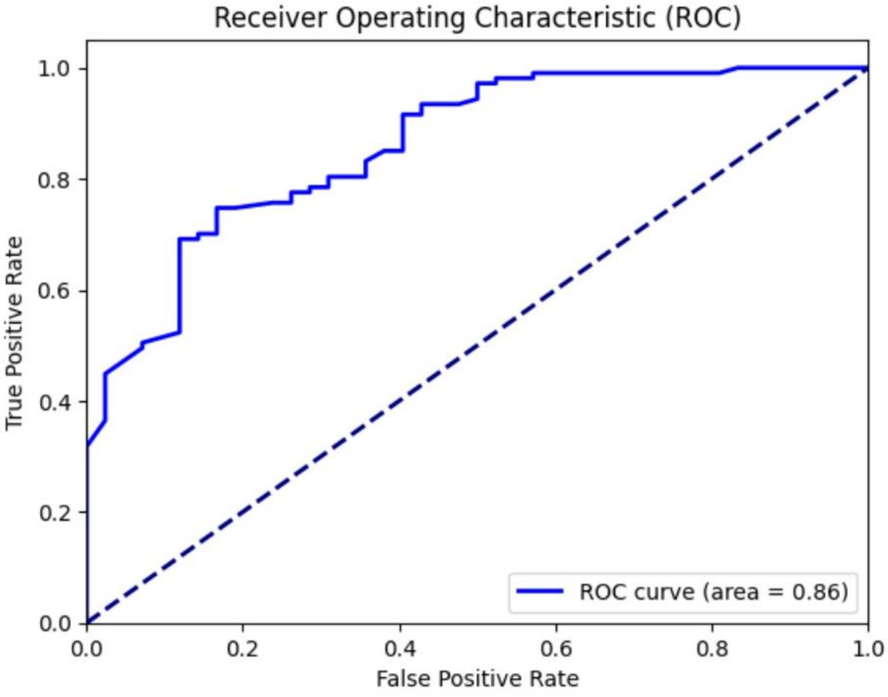


Figure 4-7 TPOT ROC Curve

4.2.1.1 Feature Selection using TPOT

Following the identification of the best model, we performed feature extraction to determine most significant & contributing featurees to the model's accuracy. The Tree-based Pipeline Optimisation Tool (TPOT) is utilised to select features during the automated machine learning (AutoML) process. TPOT uses genetic programming to optimise machine learning pipelines, which include feature selection algorithms. Specifically, TPOT used Recursive Feature Elimination (RFE) to select the most relevant attributes. RFE works by fitting a model recursively and deleting the weakest features according on their usefulness in predicting the target variable, until the optimal amount of features is found. TPOT uses cross-validation throughout the optimisation process to estimate the performance of each pipeline configuration. The pipelines with the highest accuracy are chosen, and their configurations are further optimised via crossover and mutation processes, resulting in new generations of pipelines with varied feature subsets. The analysis revealed that the "Q-CHAT 10 score" was the most important feature. The importance of various features is depicted in the figure 4-8 illustrating their relative contributions to the model's predictiveness.

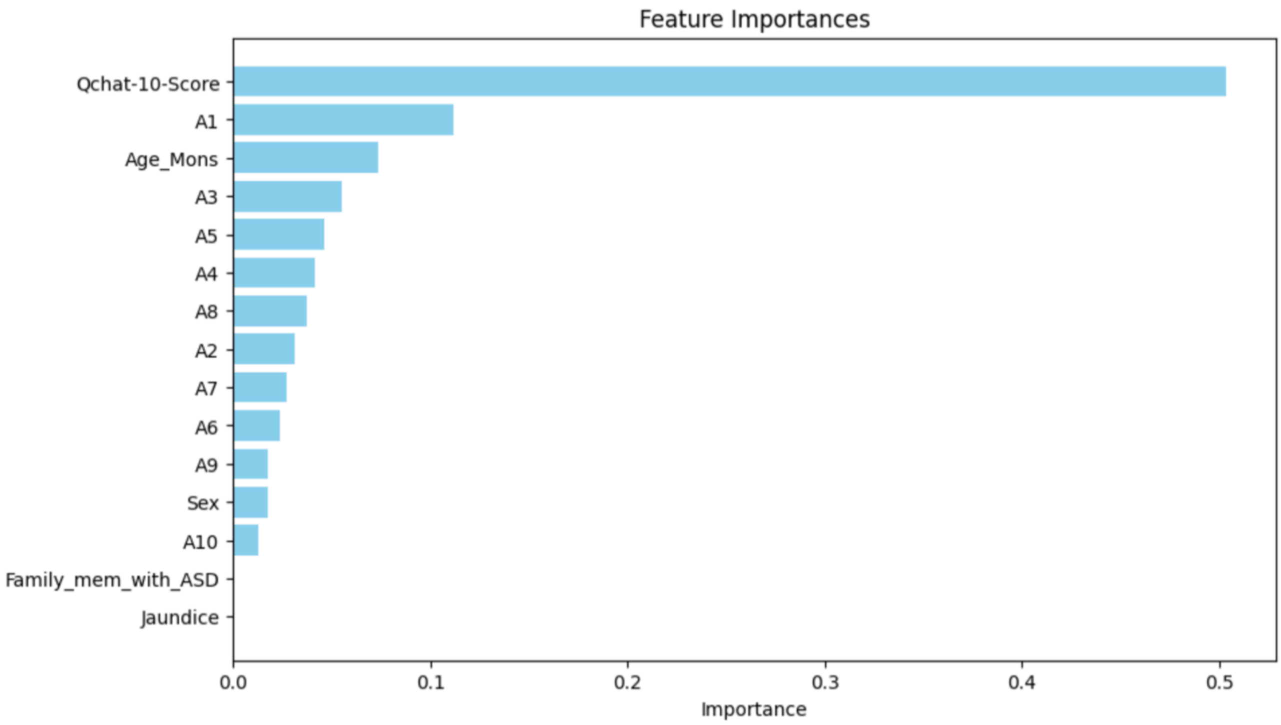


Figure 4-8 TPOT Feature Importance

The selected Random Forest classifier was subsequently evaluated using the testing data, confirming its high performance and reliability for ASD detection. This experiment highlighted TPOT's capability to automate the creation and optimization of machine learning pipelines, leading to the selection of a highly accurate model.

4.2.2. Experiment 2 (Applying KNIME)

In this experiment, we repeat the procedure from experiment 1 and verified our results using KNIME, east to use and adaptable AutoML platform. KNIME's comprehensive features and intuitive interface facilitated a streamlined and efficient workflow for our ASD detection task which is shown in figure 4-9.

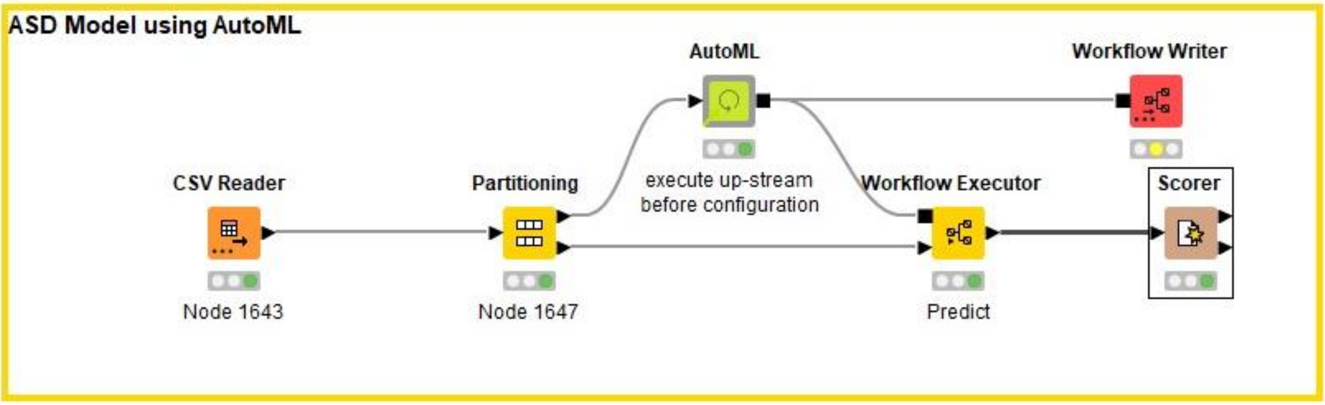
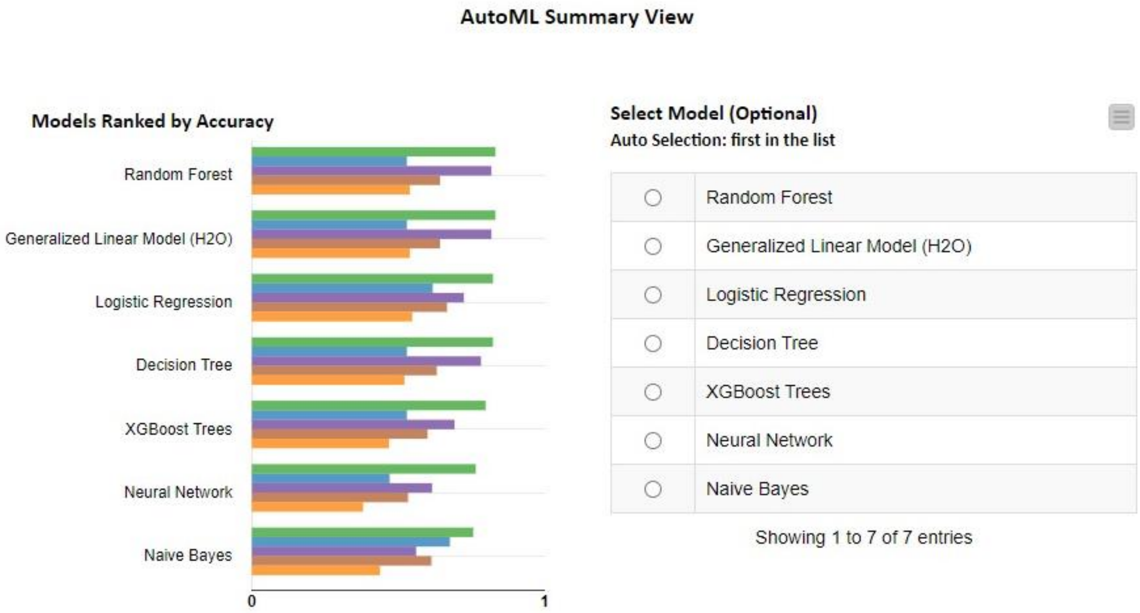


Figure 4-9 KNIME AutoML Workflow

We initiated the process by loading our dataset through the CSV Reader node, which allowed us to efficiently handle and preprocess our data. Next, we used the Partitioning node to split the dataset into training and testing sets with an 80-20 ratio, ensuring a robust evaluation framework. The core of our KNIME experiment was the AutoML node, which encompasses variety of machine learning models & automates selection and hyperparameter tuning process. This node systematically evaluated multiple models and configurations to identify the best-performing model for our dataset. After extensive analysis, the AutoML node selected the Random Forest classifier as the optimal mode which is mentioned in figure 4-10.



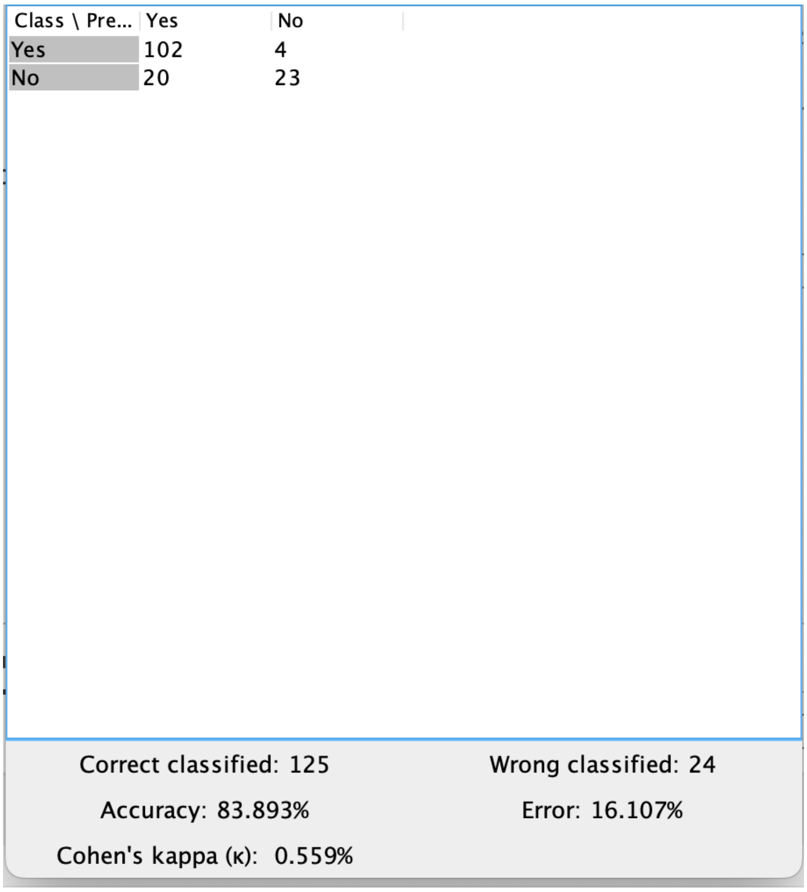
Select Model (Optional)
Auto Selection: first in the list

○	Random Forest
○	Generalized Linear Model (H2O)
○	Logistic Regression
○	Decision Tree
○	XGBoost Trees
○	Neural Network
○	Naive Bayes

Showing 1 to 7 of 7 entries

Figure 4-10 AutoML Summary View of KNIME

Finally, the Scorer node was employed to assess the accuracy of our model, providing a detailed performance evaluation and confirming the reliability of our results. Achieving an impressive accuracy of 83% on the dataset, below the confusion matrix can be seen in Figure 4-11.



Class \ Pre...	Yes	No
Yes	102	4
No	20	23

Correct classified: 125

Wrong classified: 24

Accuracy: 83.893%

Error: 16.107%

Cohen's kappa (κ): 0.559%

Figure 4-11 KNIME Confusion Matrix

In the figure 4-12 the ROC curve created by KNIME provides a visual picture of model's ability to distinguish between autistic and non-autistic classes. The ROC curve in Figure 4-12 shows the classification model's performance by displaying the True Positive Rate (sensitivity) vs the False Positive Rate (1-specificity) at various threshold values. The value 0.8 shown in the picture represents a specific threshold used to identify expected probabilities as positive or negative at that point on the curve. It is vital to understand that 0.8 is a classification threshold, which is also the Area Under the ROC Curve (AUC). The AUC is a single scalar value between 0 and 1 that summarises the model's overall performance, with values closer to 1 indicating stronger discriminatory ability. This graphical depiction improves understanding of model's predictive capabilities and adds vital insights to the findings.

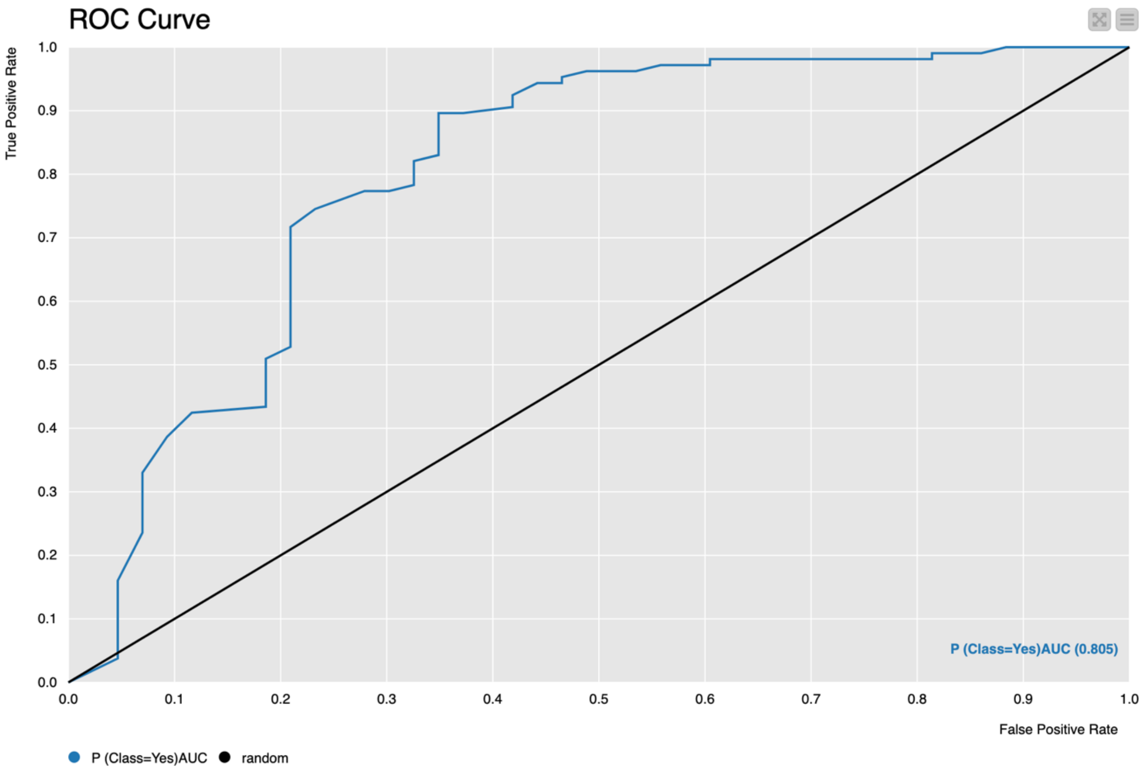


Figure 4-12 KNIME ROC Curve

4.2.2.1 Feature Selection Using KNIME

Following the model selection, The KNIME Analytics Platform was used to implement the feature selection procedure, as shown in Figure 4-13. The workflow is designed to identify the most relevant features using forward feature selection technique in KNIME. The workflow involves the following steps:

CSV Reader:

The dataset is initially loaded into the workflow using the CSV Reader node. This node reads the data from the csv file.

Feature Selection Loop Start (1:1):

Feature selection loop is used to start the feature selection process by selecting the specified method. This node allows model to use different combination of features and trained the model according to those sets.

Partitioninig:

This node tells about itself by its name. It is used to split the dataset into training and testing dataset. This make sure that model is trained on one subset of data and tested on the other subset.

Random Forest Learner:

This is the model which is used on our training dataset for each subset of features using the Random Forest Learner node. This node builds the model on the selected key features which are best for the model.

Random Forest Predictor:

This node is used on our testing dataset to check how much strong our model is in predicting positive label as positive and negative label as negative. This node generates predictions for the test instances based on the model built in the previous step using training dataset with labeled column.

Scorer:

Scorer node is used to evaluate the performance of the model. This node compares the predicted labels with the actual labels in the test set, calculating various performance metrics to assess the model's effectiveness for each subset of features.

Feature Selection Loop End:

The loop is concluded with the Feature Selection Loop End node. This node aggregates the scores from each iteration, providing a comprehensive evaluation of different feature subsets' performance.

Feature Selection Filter:

Finally, the Feature Selection Filter node selects the best subset of features based on the aggregated performance scores. This node identifies the feature combination that yields the highest model performance, optimizing the feature set for the Random Forest classifier.

This KNIME workflow effectively discovers the most relevant features, improving the model's predicted accuracy and stabilit

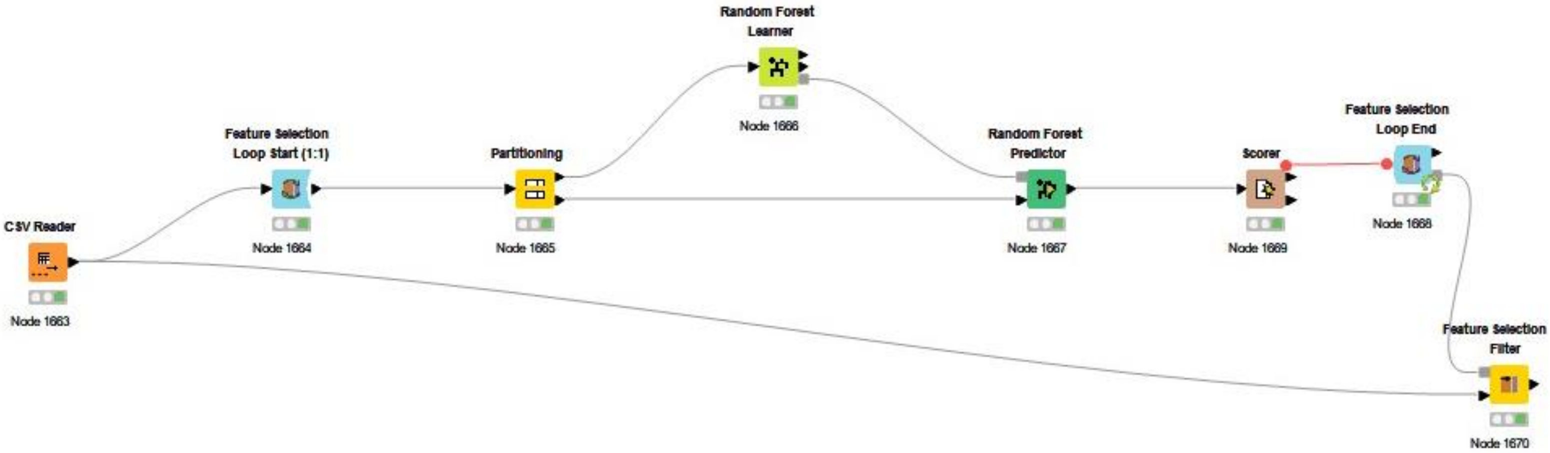
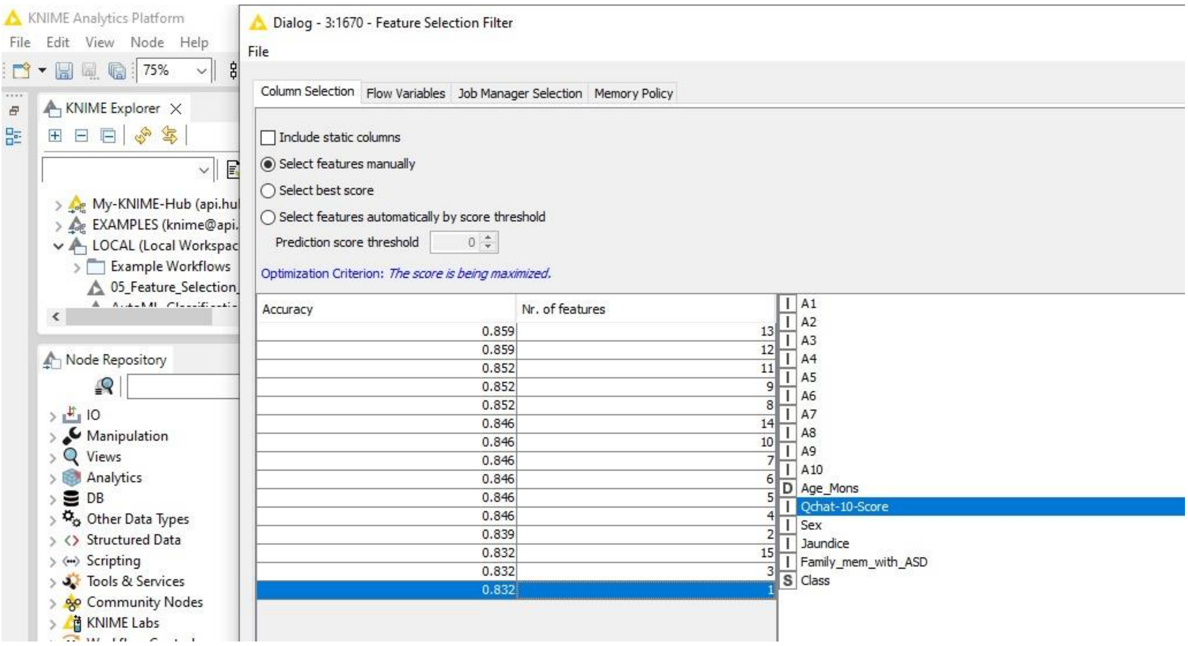


Figure 4-13 Random Forest Classifier Workflow KNIME

The analysis revealed that the "Q-CHAT 10 score" was the most influential feature, consistent with the findings from our TPOT which also gives the same attribute as a important key feature. The attached figure 4-14 illustrate the workflows and highlight the importance of various features.



Dialog - 3:1670 - Feature Selection Filter

□ Include static columns
● Select features manually
○ Select best score
○ Select features automatically by score threshold
Prediction score threshold 0

Optimization Criterion: *The score is being maximized.*

Accuracy	Nr. of features
0.859	13
0.859	12
0.852	11
0.852	9
0.852	8
0.846	14
0.846	10
0.846	7
0.846	6
0.846	5
0.846	4
0.839	2
0.832	15
0.832	3
0.832	1

I A1
I A2
I A3
I A4
I A5
I A6
I A7
I A8
I A9
I A10
D Age_Mons
I Qchat-10-Score
I Sex
I Jaundice
I Family_mem_with_ASD
S Class

Figure 4-14 KNIME Feature Importance

Using KNIME's features, we confirmed the efficacy of using AutoML tools to construct high-accuracy models, ensuring consistent and predictable performance for ASD detection. These findings add to the expanding body of evidence supporting the use of automated machine learning approaches in clinical settings, hence improving early detection and intervention for ASD.

4.3. Comparison and Discussions

In comparing experiment 1 and experiment 2, which utilized TPOT and KNIME, respectively, to automate the machine learning pipeline for ASD detection, several notable distinctions and similarities emerged. Both experiments aimed to optimize model performance and extract best features set through the use of AutoML tools. Experiment 1, employing TPOT, showcased its evolutionary algorithm-based approach to automatically design and optimize machine learning pipelines and gives best key features. TPOT discovered a Random Forest Classifier that achieved 85.23% accuracy rate on the testing dataset. On the other hand, experiment 2 performed by using KNIME, which is an open-source platform and well known for its adaptability & ease of use. Like experiment 1, experiment 2 also identified a Random Forest classifier as the optimal model, attaining an accuracy rate of 83% on KNIME. Both experiments employed a dataset split of 80-20 for training and testing to ensure robust evaluation frameworks but TPOT gives better results as compared to KNIME.

Table 4-1 Comparison between TPOT & KNIME

	TPOT	KNIME	Explanation
Model Accuracy	85.23%	83.89%	TPOT genetic programming approach gives better accuracy as it explores a broader range of model configurations pipelines and feature combinations optimizing them iteratively as compared to KNIME rule-based workflow.
Important Key Feature	Q-CHAT 10 score	Q-CHAT 10 score	Both techniques identified the most significant feature as having high predictive value for ASD identification.
Feature Signature Discovery	Q-CHAT 10 score A1 Age_Mons A3 A5 A4 A8 A2 A7 A6 A9 Sex A10 Jaundice	Q-Chat 10- score Age_mons A3 A6 A4 A1 A2 A10 Jaundice A7 A5 A8 A9 Sex	The features changes since each tool utilizes different algorithms and techniques for feature selection. TPOT's genetic programming approach uses iterative evolution and selection, but KNIME's randomized process leads feature selection in a different way.

The feature signatures produced by TPOT and KNIME vary principally because each AutoML tool uses different approaches for feature selection and model optimisation. TPOT uses genetic programming, an evolutionary technique for repeatedly evolving machine learning pipelines by choosing, modifying, and recombining features based on their performance. This repetitive approach gives better result than KNIME model. On the other hand, KNIME employs a rule-based and workflow-based approach that systematically leads the model building and feature selection process. Because of these methodological variations, both tools can consistently identify the most significant characteristic, but cross validation techniques selected by each tool may differ.

Additionally, feature extraction in both experiments highlighted the "Q-CHAT 10 score" as the most influential feature contributing to model accuracy. While experiment 1 relied on TPOT's automated pipeline generation and hyperparameter optimization, experiment 2 leveraged KNIME's AutoML node for model selection and tuning. Notably, despite differences in the underlying AutoML tools and workflows, experiment 1(TPOT) yielded better accuracy rate and identified consistent key features for ASD detection. These results give the efficiency of AutoML techniques used in clinical settings, offering consistent and reliable performance for early ASD detection.

This highlights AutoML's ability to standardise and improve diagnostic processes across a wide range of healthcare settings. The importance and reliability of TPOT for medical application is demonstrated by its capability to attain comparably high accuracy and identify important features. Using this AutoML approach, we can improve the results in medical terms and specially in our case for ASD patients. We can save their future by predicting this disorder in early stages so they can overcome this disorder to make their life’s better.

CHAPTER 5

CONCLUSION

In this chapter we will discuss the conclusion of our research on AutoML for autism detection at early ages.

5.1. Conclusion

Finally, this research emphasized the use of automated machine learning (AutoML) for Autism Spectrum Disorder (ASD) detection. In our approach, TPOT gave better results with 85.23% for the Random Forest Classifier model in diagnosis ASD compared to KNIME that resulted only into a 83.89 %accuracy Thus, a significant part of our analysis was dedicated to such feature selection and the recognized relevance incorporating "Q-CHAT 10 score" in both TPOT and KNIME workflow. This shows that automated feature selection is essential to enhance the accurateness of ASD diagnostic models. Through thorough investigation into the relevance of features, we detected the key factors in ASD diagnosis which will be helpful for understanding diagnostic qualities. The importance of automated machine learning methods is highlighted, and the time required to build models using feature selection may be minimized. Through this focus on feature selection process, our work contributes to the progress of ASD diagnosis and highlights how AutoML has the potential to revolutionize clinical practices.

5.2. Future Work and Limitation

One of the main challenges we encountered in our study was the limited size of dataset we used. This limitation may reduce the generalizability of our findings and the robustness of our models. Future work should involve the collection and analysis of larger and more diverse dataset to validate and enhance the performance of the AutoML framework for diagnosing ASD.

REFERENCES

[1] S. a. W. W. P. a. K. K. W. Jomthanachai, "An application of machine learning regression to feature selection: a study of logistics performance and economic attribute," *Neural Computing and Applications,* vol. 34, pp. 15781--15805, 2022.

[2] T. A. a. d. L. I. B. Abdallah, "Survey on Feature Selection," *arXiv preprint arXiv:1510.0289,* 2015.

[3] J. a. L. J. a. W. S. a. Y. S. Cai, "Feature selection in machine learning: A new perspective}," *Neurocomputing,* vol. 300, pp. 70--79, 2018.

[4] G. Roffo, "Feature selection library (MATLAB toolbox)," *arXiv preprint arXiv:1607.01327,* 2016.

[5] S. G. S. M. M. B. A. &. B. Jacob, "Feature signature discovery for autism detection: An automated machine learning based feature ranking framework.," *Computational Intelligence and Neuroscience,* vol. 2023, p. 6330002, 2023.

[6] A. a. D. S. Sharma, "A comparative study of feature selection and machine learning techniques for sentiment analysis," in *roceedings of the 2012 ACM research in applied computation symposium,* 2012.

[7] D. a. s. S. a. A. M. A. a. A. l. Aksu, Intrusion detection with comparative analysis of supervised learning techniques and fisher score feature selection algorithm, Springer, 2018, pp. 141--149.

[8] "Comparative analysis of machine learning techniques for indian liver disease patients," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS),* IEEE, 2020.

[9] B. a. K. G.-R. a. L. R. Khagi, "Comparative analysis of Alzheimer's disease classification by CDR level using CNN, feature selection, and machine-learning techniques," *International Journal of Imaging Systems and Technology,* vol. 29, pp. 297--310, 2019.

[10] M. a. T. T. a. A.-B. M. A. a. T. J. a. A. M. A. a. A. D. I. a. T. H. Mafarja, "Classification framework for faulty-software using enhanced exploratory whale optimizer-based feature selection scheme and random forest ensemble learning," *Applied Intelligence,* vol. 53, pp. 18715--18757, 2023.

[11] K.-Y. a. S. d. L. R. a. B. N. G. a. V. E. a. K. T. a. S. K. a. C. P. V. H. a. Y. M.-D. a. V.

A. a. S. K. Li, "Toward automated machine learning-based hyperspectral image analysis in crop yield and biomass estimation," *Remote Sensing,* vol. 14, p. 1114, 2022.

[12] Y. A. A. a. R. D. G. a. C. M.-Z. J. a. S. R. A. a. N. J. a. D. M. O. Adla, "Automated detection of polycystic ovary syndrome using machine learning techniques," in *2021 Sixth international conference on advances in biomedical engineering (ICABME),* IEEE, 2021.

[13] S. a. M. S. Raj, "Analysis and detection of autism spectrum disorder using machine learning techniques," *Procedia Computer Science,* vol. 167, pp. 994--1004, 2020.

[14] R. a. M.-T. R. a. P. P. a. d. l. P. F. a. S. E. Romero-Garc{\'\i}a, "Q-CHAT-NAO: A robotic approach to autism screening in toddlers," *Journal of Biomedical Informatics,* vol. 118, p. 103797, 2021.

[15] F. a. P. D. Thabtah, "A new machine learning model based on induction of rules for autism detection," *Health informatics journal,* vol. 26, pp. 264--286, 2020.

[16] J. a. F. E. a. S. J. a. I. A. a. D. M. S. a. S. S. a. Y. A. a. S. A. a. E. M. Zeidan, "Global prevalence of autism: A systematic review update," *Autism research,* vol. 15, pp. 778--790, 2022.

[17] V. a. S. A. a. J. Y. B. a. S. S. P. a. J. J. a. o. Vishal, "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),* IEEE, 2022.

[18] M. a. S. C. a. T. M. Marlow, "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, pp. 176--199, 2019.

[19] S. a. A. T. a. Z. S. a. S. S. a. H. M. I. Islam, "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),* IEEE, 2020.

[20] D. Stevanovi, "Quantitative Checklist for Autism in Toddlers (Q-CHAT): A psychometric study with Serbian Toddlers," *Research in Autism Spectrum Disorders,* vol. 2021, p. 101760, 83.

[21] L. a. C. F. a. A. G. M. a. A. F. a. L. E. a. M. R. a. C. C. a. C. N. a. C. V. a. T. N. a. o. Ruta, "Validation of the quantitative checklist for autism in toddlers in an Italian clinical sample of young children with autism and other developmental disorders," *Frontiers in psychiatry,* vol. 10, p. 488, 2019.

[22] A. a. P. E. Nied{\'z}wiecka, "Symptoms of Autism Spectrum disorders measured by the qualitative checklist for Autism in toddlers in a large sample of Polish toddlers," *International Journal of Environmental Research and Public Health,* vol. 19, p. 3072, 2022.

[23] N. a. B. F. a. I. W. Farooqi, "edictive analysis of autism spectrum disorder (ASD) using machine learning," in *2021 International Conference on Frontiers of Information Technology (FIT),* IEEE, 2021.

[24] F. Thabtah, "An accessible and efficient autism screening method for behavioural data and predictive analyses," *Health informatics journal,* vol. 25, pp. 1739--1755, 2019.

[25] M. a. T. L. a. H. D. E. a. C. Z. H. A. a. M. J. C. a. C. D. a. V. S. R. Cerrada, "AutoML for feature selection and model tuning applied to fault severity diagnosis in spur gearboxes," *Mathematical and Computational Applications,* vol. 27, p. 6, 2021.

[26] M. S. a. S. F. F. a. A. M. S. a. A. M. H. a. M. M. A. Satu, "Early detection of autism by extracting features: a case study in Bangladesh," in *2019 international conference on robotics, electrical and signal processing techniques (ICREST),* IEEE, 2019.

[27] A. a. K. P. a. C. T. Ranaut, "Identifying autism using EEG: unleashing the power of feature selection and machine learning," *Biomedical Physics \& Engineering Express,* vol. 10, p. 035013, 2024.

[28] J. a. L. A. E. a. B. M. a. D. S. Eldridge, "Robust features for the automatic identification of autism spectrum disorder in children," *Journal of neurodevelopmental disorders,* vol. 6, pp. 1--12, 2014.

[29] C. a. B.-C. S. a. W. S. a. C. T. a. R. J. a. P. G. a. B. C. Allison, "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," *Journal of autism and developmental disorders,* vol. 38, pp. 1414--1425, 2008.

[30] C. a. A. B. a. B.-C. S. Allison, "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," *Journal of the American Academy of Child \& Adolescent Psychiatry,* vol. 51, pp. 202--212, 2012.

Rana Touqeer

ORIGINALITY REPORT

16%	10%	11%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Avishek Saha, Dibakar Barua, Mahbub C. Mishu, Ziad Mohib, Sumaya Binte Zilani Choya. "Development of an Interactive Dashboard for Analyzing Autism Spectrum Disorder (ASD) Data using Machine Learning", International Journal of Information Technology and Computer Science, 2022 Publication	1%
2	Imrus Salehin, Md Shamiul Islam, Pritom Saha, S.M. Noman, Azra Tuni, Md Mehedi Hasan, Md Abu Baten. "AutoML: A systematic review on automated machine learning with neural architecture search", Journal of Information and Intelligence, 2023 Publication	1%