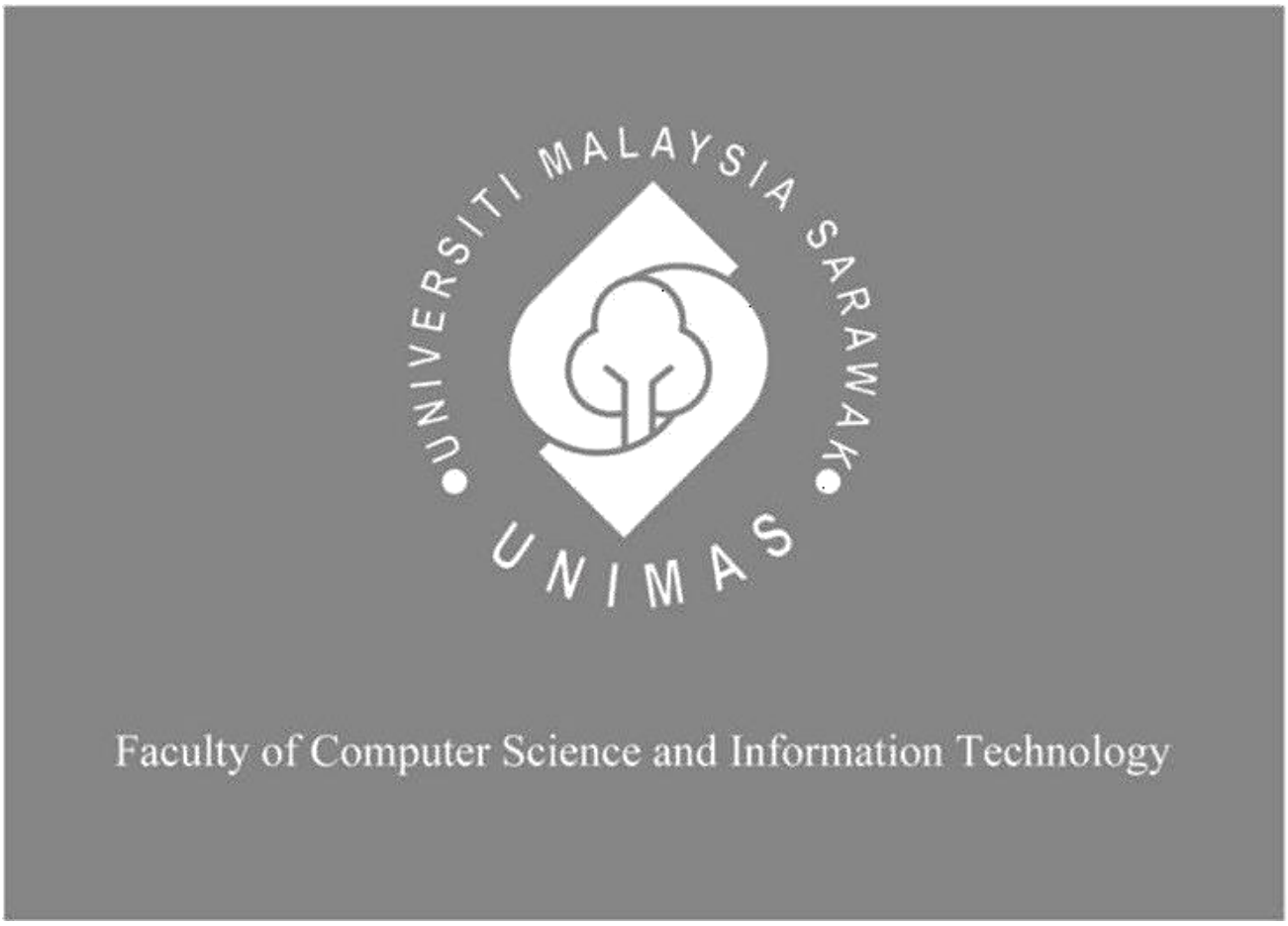



# *A COMPARATIVE STUDY OF MACHINE LEARNING MODELS FOR PREDICTION OF AUTISM SPECTRUM DISORDER USING SCREENING DATA*

**Yeap Ming Yue**

**Bachelor of Computer Science with Honours**
**(Computational Science)**

2023

# A COMPARATIVE STUDY OF MACHINE LEARNING MODELS FOR PREDICTION OF AUTISM SPECTRUM DISORDER USING SCREENING DATA

YEAP MING YUE

This project is submitted in partial fulfillment of the requirements for the degree of Bachelor of Computer Science with Honours

Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK
2023

**UNIVERSITI MALAYSIA SARAWAK**

# THESIS STATUS ENDORSEMENT FORM

**TITLE** A COMPARATIVE STUDY OF MACHINE LEARNING MODELS FOR PREDICTION OF AUTISM SPECTRUM DISORDER USING SCREENING DATA

**ACADEMIC SESSION:** 2022/2023

YEAP MING YUE

**(CAPITAL LETTERS)**

hereby agree that this Thesis* shall be kept at the Centre for Academic Information Services, Universiti Malaysia Sarawak, subject to the following terms and conditions:

1. The Thesis is solely owned by Universiti Malaysia Sarawak
2. The Centre for Academic Information Services is given full rights to produce copies for educational purposes only
3. The Centre for Academic Information Services is given full rights to do digitization in order to develop local content database
4. The Centre for Academic Information Services is given full rights to produce copies of this Thesis as part of its exchange item program between Higher Learning Institutions [ or for the purpose of interlibrary loan between HLI ]
5. ** Please tick ( √ )

| | | |
|---|---|---|
| ☐ | CONFIDENTIAL | (Contains classified information bounded by the OFFICIAL SECRETS ACT 1972) |
| ☐ | RESTRICTED | (Contains restricted information as dictated by the body or organization where the research was conducted) |
| ☑ | UNRESTRICTED | |

Validated by

(AUTHOR'S SIGNATURE)

(SUPERVISOR'S SIGNATURE)

Permanent Address

5-11-8, Lintang Macallum 2,
10300 George Town,
Pulau Pinang.

Date: 26 June 2023

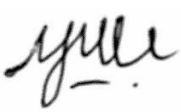
Date: 26 June 2023

Note * Thesis refers to PhD, Master, and Bachelor Degree
** For Confidential or Restricted materials, please attach relevant documents from relevant organizations / authorities

# DECLARATION OF ORIGINALITY

I hereby declare that this research together with all of its content is none other than that of my own work, with consideration of the exception of research-based information and relative materials that were adapted and extracted from other resources, which have evidently been quoted or stated respectively.

Signed,



……

YEAP MING YUE

Faculty of Computer Science and Information Technology 26 June 2023
Universiti Malaysia Sarawak.

## ACKNOWLEDGEMENT

Firstly, I would like to express my sincerest gratitude to my supervisor, Dr. Stephanie Chua Hui Li, for her invaluable assistance, guidance, suggestions, support, and supervision throughout my Final Year Project. Her expertise and feedback have been instrumental in shaping the project and ensuring its quality. Secondly, I wish to extend my deepest gratitude to my examiner for their valuable comments and feedback regarding my Final Year Project. I would also like to convey my gratefulness to Professor Dr. Wang Yin Chai, the Final Year Project coordinator, for providing guidelines and coordination throughout the conduction of the project.

Furthermore, I want to express my gratitude to the Faculty of Computer Science and Information Technology (FCSIT) at Universiti Malaysia Sarawak (UNIMAS) for providing me with the opportunity to gain valuable experience and learn new concepts throughout my studies. The resources and support provided by the faculty have greatly contributed to the success of this project.

I would like to thank my friends for their willingness to share their knowledge and provide moral support throughout the project. I want to express my heartfelt appreciation to my family and parents for their unwavering love, support, and encouragement. Their belief in me and constant motivation have been the driving force behind my achievements.

Lastly, I would like to extend my deepest gratitude to the respondents who participated in the Google Forms survey, providing valuable data for the prediction of autism spectrum disorder (ASD). Their contribution has played a crucial role in the success of this project, enabling a more comprehensive analysis and evaluation of the models.

**TABLE OF CONTENTS**

## LIST OF TABLES

# LIST OF FIGURES

## ABSTRACT

Autism spectrum disorder (ASD) is a neurological and developmental disorder that affects how people interact with others, communicate, learn, and behave. ASD prediction is difficult because the diagnostic factors may not be based solely on observation. The project focuses on using ASD screening data to predict ASD traits in adults. This project aims to predict ASD traits in adults based on screening data using a machine learning approach. This can help them decide whether to seek a medical practitioner. The project proposed using classification, which is one of the machine learning approaches to predict autism spectrum disorder. The proposed prediction models are Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours, Naïve Bayes, and Neural Network. The methodology adopted by the project is knowledge discovery in databases (KDD) to accomplish the needs of this project. The steps include domain understanding, data selection, data pre-processing, data transformation, data mining/modelling and model evaluation. The project will create a dataset based on AQ-10 adults questionnaire data that will facilitate future work in future work in predicting ASD in adults. Feature selection will be performed to find useful features in predicting ASD traits in adults. The performance of the classification models for ASD will be compared. Finally, the best classification model for ASD prediction was a model trained using the Support Vector Machine (SVM) algorithm.

# ABSTRAK

Gangguan spektrum autisme (autism spectrum disorder, ASD) ialah gangguan neurologi dan perkembangan yang menjejaskan cara orang berinteraksi dengan orang lain, berkomunikasi, belajar dan berkelakuan. Ramalan ASD adalah sukar kerana faktor diagnosis mungkin bukan hanya berdasarkan pemerhatian. Projek fokus pada menggunakan data saringan ASD untuk meramalkan sifat ASD pada orang dewasa. Ini boleh membantu mereka membuat keputusan sama ada untuk mendapatkan rawatan daripada doktor perubatan. Projek ini akan menggunakan klasifikasi yang merupakan salah satu teknik pembelajaran mesin untuk meramal ASD. Model ramalan yang dicadangkan ialah Regresi Logistik, Hutan Rawak, Mesin Vektor Sokongan (SVM), K-Jiran Terdekat, Naïve Bayes, dan Rangkaian Neural Buatan. Metodologi yang digunakan oleh projek ini ialah penemuan pengetahuan dalam pangkalan data (KDD) untuk mencapai objektif projek ini. Langkah tersebut termasuk pemahaman domain, pemilihan data, pra-pemprosesan data, transformasi data, perlombongan/pemodelan data dan penilaian model. Projek ini akan menghasilkan set data berdasarkan data soal selidik dewasa AQ-10 yang akan memudahkan kerja masa depan dalam meramalkan ASD pada orang dewasa. Pemilihan ciri akan dilakukan untuk mencari ciri yang berguna untuk meramal sifat ASD dalam kalangan orang dewasa. Prestasi model klasifikasi untuk ASD akan dibandingkan. Akhirnya, model klasifikasi terbaik untuk ramalan ASD ialah model yang dilatih menggunakan algoritma Mesin Vektor Sokongan (SVM).

# CHAPTER 1

# INTRODUCTION

## 1.1 Introduction

Autism spectrum disorder (ASD) is a disability in development caused by differences in the brain (Centers for Disease Control and Prevention, 2022). People with ASD usually have problems with limited or repetitive behaviours or interests, as well as communication skills and social engagement. Although the symptoms are easy to identify, a diagnosis of autism requires skilled medical professionals to supervise behavioural assessments that are measured according to the incidence of numerous symptoms that interfere with a person's capacity to talk, play, and create communication relationships. Depending on how serious the symptoms are, ASD can range from mild to severe (Hodges, Fealko, & Soares, 2020). Machine learning methods are being used on data sets related to ASD in order to identify valuable hidden patterns and create a predictive model for diagnosing its risk (Jalaja, Geetha, & Vivek, 2019). The main purpose of the project is to use machine learning techniques to learn models for the classification of autism spectrum disorder (ASD) using screening data. In this project, a significant feature set will be determined using a correlation matrix. A feature selection algorithm will also be applied to the feature set to determine the best set of features for learning the classification model. The feature set will then be used in six machine learning algorithms to learn models for the classification of ASD into ‘ASD trait’ and ‘No ASD trait’ (Thabtah, 2018). A comparative study will be conducted to determine the best classification model. This best model will be selected for deployment.

## 1.2 Problem Statement

Parents who are concerned that their child may be autistic can bring them to medical practitioners. Medical practitioners will conduct screening tests on toddlers and children to diagnose if they have ASD. Many times, diagnosis cannot be determined in one visit, and it involves multiple visits to the clinic for a period of time, sometimes up to a few years to finally get a definitive diagnosis (Saihi & Alshraideh, 2021). There are also teenagers and young adults who were not diagnosed with ASD from young and did not receive early intervention. The problem with ASD is that it is quite hard to diagnose as every child may progress through life at a different developmental speed. Sometimes, parents may also be unaware of certain ASD traits which they may think are normal in their child. Therefore, this project looks into using the machine learning approach to learn models for the classification of ASD based on past data available.

## 1.3 Scope

The scope of the study covers the prediction of ASD in adults who are 18 years and older. Six prediction models are built to be compared. The machine learning algorithms used for building the models are Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naïve Bayes, and Neural Network. The deliverables are a comparative study of classification models.

## 1.4 Objectives

The objectives of this project are:

1. To collect data using Google Forms to augment the dataset from Kaggle.
2. To determine the best feature set for ASD classification and build classification models for ASD using the machine learning approach.
3. To compare the performance of the classification models for ASD.

## 1.5 Methodology

The methodologies of this project used are knowledge discovery in databases (KDD) is as shown in Figure 1.1.

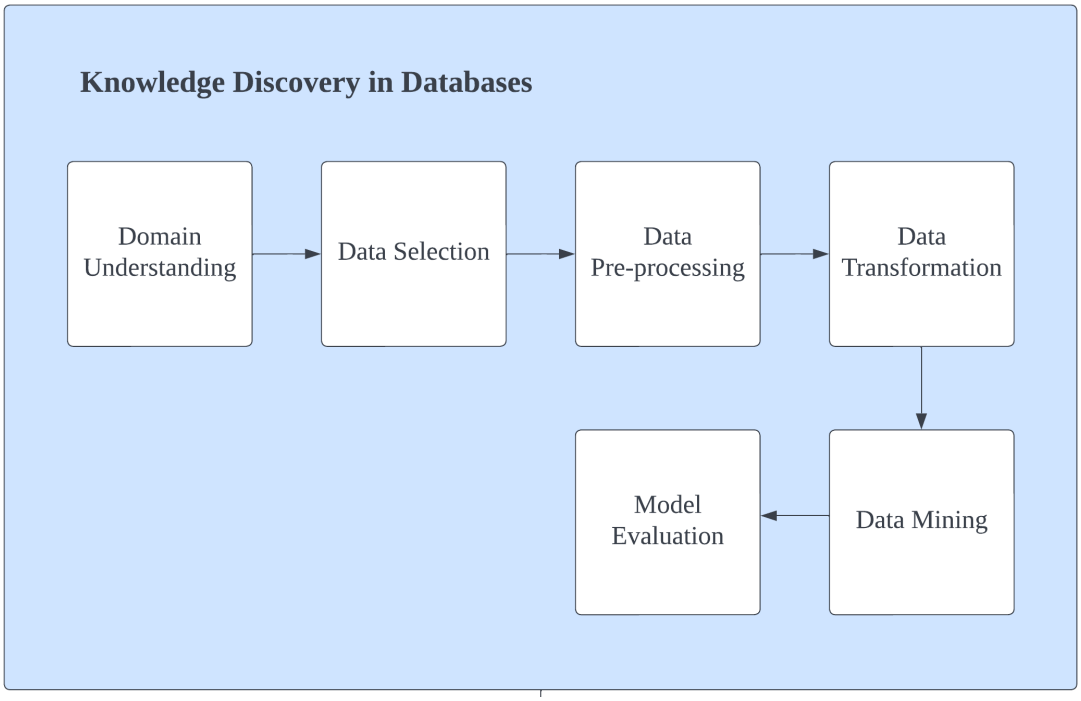



Figure 1.1 Flowchart for Project Methodology

### 1.5.1 Domain Understanding

Supervised machine learning is used to analyse the massive dataset of autism spectrum disorder (ASD) screening data. Classification is a data analysis task. It is a process of finding a model that describes and distinguishes data classes which are 'ASD Traits' and 'No ASD Traits'. It is a binary classifier as the classification problem has only two possible outcomes. Learning the patterns of human behaviours. Classification is the problem of identifying to which of a set of categories a new observation belongs, based on a training set of data containing observations and whose categories membership is known.

### 1.5.2 Data Selection

The next step after project understanding is to collect the data. Data will be collected using Google Forms to augment the dataset from Kaggle. The questions in Google Forms are based on Autism Spectrum Quotient 10 items (AQ-10) (Adult) from ASD Tests App. The data will be collected from adults to predict whether they have autism spectrum disorder traits.

### 1.5.3 Data Pre-processing

Data pre-processing techniques can be applied to the autism spectrum disorder screening data so that the data is suitable for learning classification models. Feature selections should also be done to ensure the features used in predictions can enhance the model's performance. The feature selection method is aimed at reducing the number of less useful features. The most important will use to build a model to predict the target variables which are 'ASD Traits' and 'No ASD Traits'.

### 1.5.4 Data Transformation

Before performing data mining, data transformation is a crucial data preprocessing technique that must be applied to the data in order to produce patterns that are simpler to analyze. This process transforms raw data into a format that makes it easier to conduct data mining and obtain strategic information.

### 1.5.5 Data Mining/Modelling

Using datasets to train the model using various machine learning algorithms. Training a model is required so that it can understand the various patterns, rules, and features. Construction of classification models using Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naïve Bayes and Neural Network algorithms. Different algorithms are used to build a classifier by making the model learn using the training set available. The model has to be trained for the prediction of accurate results.

### 1.5.6 Model Evaluation

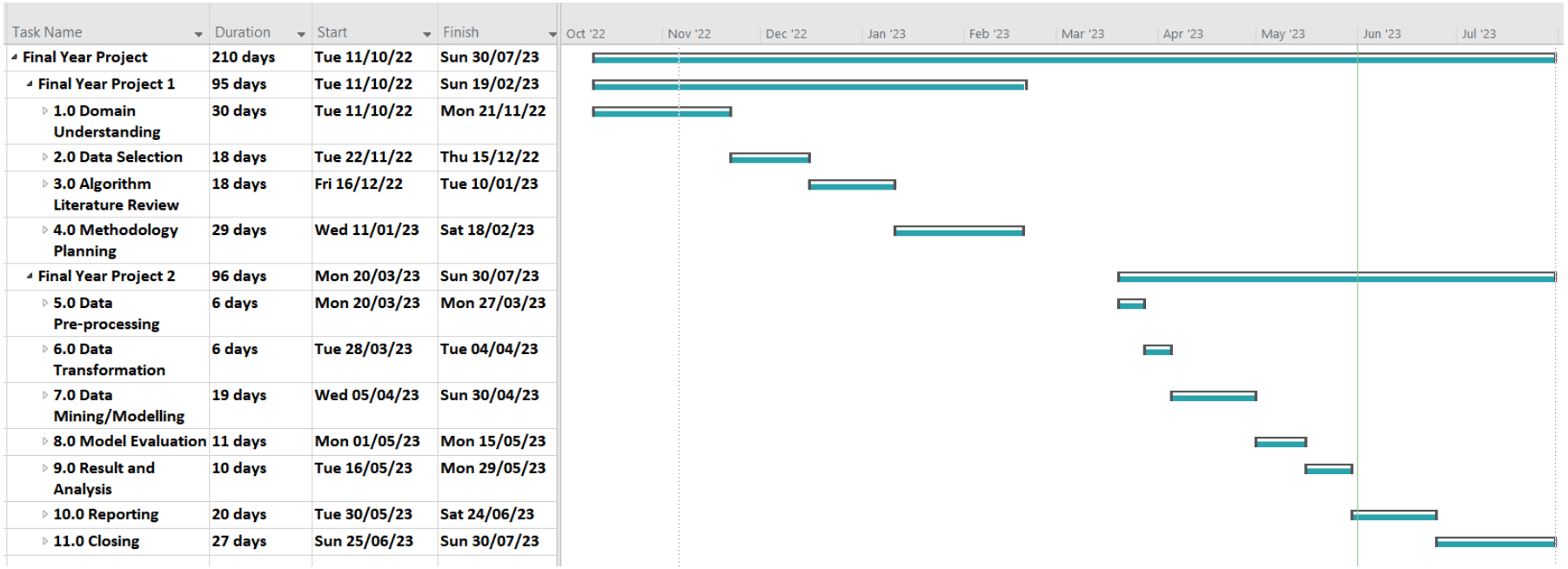
After the model is ready, it will be evaluated using metrics like accuracy, precision, recall, F-measure and area under the curve. The best model will be chosen for deployment.

## 1.6 Significance of Project

This research holds significant implications for the field of ASD diagnosis. The outcomes of this study have the potential to greatly benefit medical practitioners in their future analyses of ASD. By shedding new light on the diagnosis of ASD, this research will provide valuable insights that can enhance the understanding and treatment of individuals on the autism spectrum. The findings from this study will be instrumental in supporting mental health organizations' efforts to increase awareness and knowledge surrounding various concerns related to ASD. Furthermore, the comprehensive analysis presented in this research will yield valuable information for future studies aimed at developing improved methods of diagnosing ASD.

## 1.7 Project Schedule

Figure 1.2 shows the Gantt Chart planned for Final Year Project.



| Task Name | Duration | Start | Finish |
|---|---|---|---|
| Final Year Project | 210 days | Tue 11/10/22 | Sun 30/07/23 |
| Final Year Project 1 | 95 days | Tue 11/10/22 | Sun 19/02/23 |
| 1.0 Domain Understanding | 30 days | Tue 11/10/22 | Mon 21/11/22 |
| 2.0 Data Selection | 18 days | Tue 22/11/22 | Thu 15/12/22 |
| 3.0 Algorithm Literature Review | 18 days | Fri 16/12/22 | Tue 10/01/23 |
| 4.0 Methodology Planning | 29 days | Wed 11/01/23 | Sat 18/02/23 |
| Final Year Project 2 | 96 days | Mon 20/03/23 | Sun 30/07/23 |
| 5.0 Data Pre-processing | 6 days | Mon 20/03/23 | Mon 27/03/23 |
| 6.0 Data Transformation | 6 days | Tue 28/03/23 | Tue 04/04/23 |
| 7.0 Data Mining/Modelling | 19 days | Wed 05/04/23 | Sun 30/04/23 |
| 8.0 Model Evaluation | 11 days | Mon 01/05/23 | Mon 15/05/23 |
| 9.0 Result and Analysis | 10 days | Tue 16/05/23 | Mon 29/05/23 |
| 10.0 Reporting | 20 days | Tue 30/05/23 | Sat 24/06/23 |
| 11.0 Closing | 27 days | Sun 25/06/23 | Sun 30/07/23 |



Figure 1.2 Overall Gantt Chart for Final Year Project

## 1.8 Expected Outcome

The outcome of this project is a comprehensive comparative study of six different machine learning models used for classifying autism spectrum disorder (ASD) traits. The study aims to evaluate and compare the performance of these models in accurately predicting the presence of ASD traits based on screening data. The selected best-performing models will then be utilized to demonstrate the prediction of ASD traits.

## 1.9 Thesis Outline

### 1.9.1 Introduction

Chapter 1 provides a general idea about the project. It includes the problem statement, objectives, project scope, methodology, significance of the project, project schedule, expected outcome, and thesis outline.

### 1.9.2 Literature Review

Chapter 2 discusses a bit of background and the domain of autism spectrum disorder (ASD). Besides, the literature review also covers related work in machine learning and how computer science has been specifically used in explaining this problem.

### 1.9.3 Methodology

Chapter 3 describes the methodology used in resolving the problem. This section also explains the machine learning algorithms proposed for the predictive models, setup of the training environment and training process. The overall processes will act as guidelines for the whole project.

### 1.9.4 Data Mining Implementation

Chapter 4 provides a detailed description of the implementation of the data mining methodology, which included the process of feature extraction and modelling. Additionally, the chapter includes step-by-step instructions on how to execute the Python code necessary to implement the methodology.

### 1.9.5 Result and Discussion

Chapter 5 explains the results of the machine learning model trained for ASD prediction. A comparative study will be carried out to evaluate the predictive models. The best model will be chosen for implementation.

### 1.9.6 Conclusion and Future Works

Chapter 6 consists of a summary of the project, its limitations, future works, and a conclusion.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Introduction

The ability to collect and access healthcare data is becoming more powerful due to technological advancements in the form of hardware and computer-based applications for healthcare information. Knowledge discovery in databases (KDD) has been widely used in the medical field such as diagnosis prediction. Analysis of recorded medical data records may aid in the discovery of hidden features and patterns that could greatly improve our understanding of disease development and treatment interventions (Abdulkadium, Shekan, & Hussain, 2022).

Data mining and analytics approaches may be utilised to predict autism spectrum disorder (ASD) by employing previous patient information and screening data. Many people use data processing and analytics to extract useful data from information. Data mining is becoming a critical area of healthcare that is used to find undiscovered data in healthcare databases and use analytics to predict illnesses (Victoire, Ramalingam, Naresh, Nasimudeen, & Jaya Kumar, 2021). In this study, supervised machine learning which is classification will be used to classify autism spectrum disorder (ASD) screening data into two classes which are ‘ASD Traits’ and ‘No ASD Traits’.

This chapter will discuss the background of knowledge discovery in databases (KDD), machine learning algorithms, related works, comparison of related works and tools used in this project.

## 2.2 Knowledge Discovery in Databases (KDD)

Knowledge discovery in databases (KDD) is known as finding relevant knowledge from a set of data. Researchers in the fields of artificial intelligence, databases, statistics, machine learning, pattern recognition, knowledge acquisition for expert systems, and data visualisation may be interested in it (Priyadharsini & Thanamani, 2014). Extraction of knowledge from data in the setting of huge databases is the objective of the KDD process. This is accomplished by employing data mining techniques to extract and identify what is considered knowledge following the requirements of measures and thresholds. Data preparation and selection, data cleansing, the integration of prior information about data sets, and the interpretation of precise solutions from the observed results are all steps in the widely used data mining technique. The KDD process is presented in Figure 2.1. Knowledge discovery techniques have primarily been employed in academic environments over the past few years (Soundappan & Sugumar, 2017).

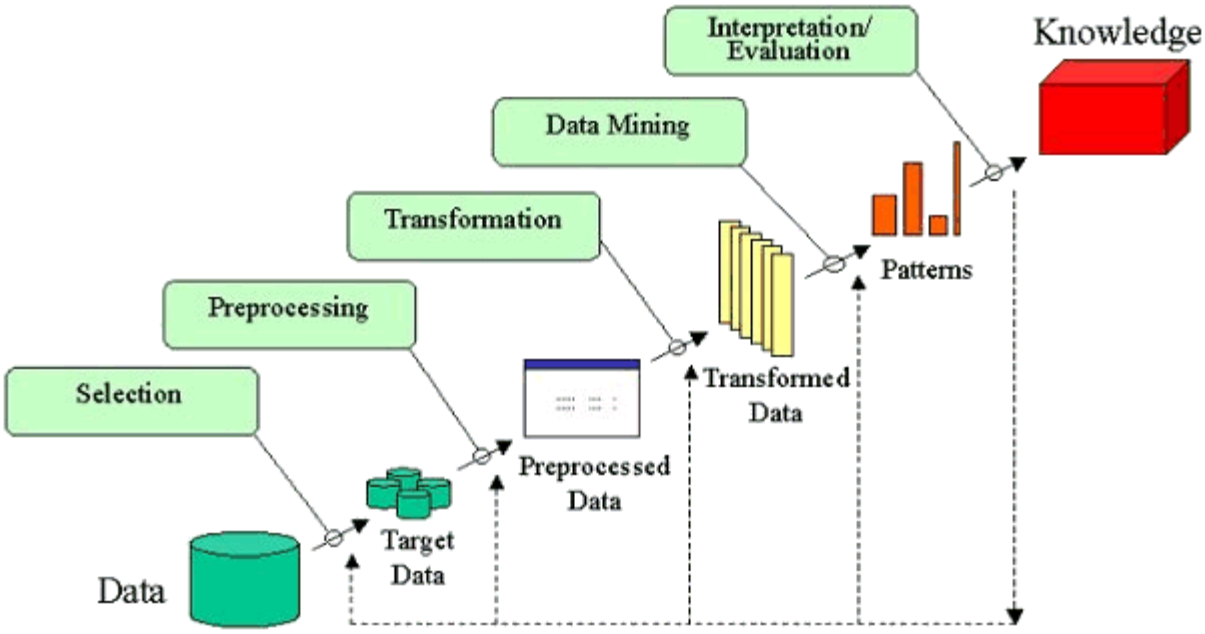



Figure 2.1 The Knowledge Discovery in Databases (KDD) Process *(Tomar & Agarwal, 2014)*.

## 2.3 Machine Learning Algorithms

Machine learning is a technique that enhances system performance by using computational techniques to learn from data. Data is the primary source of knowledge in computer systems, and the primary goal of machine learning is to create learning algorithms that create models from data. Researchers created a model that can anticipate the results of fresh observations by providing machine learning algorithms with experience data. The results learned from the data are generally referred to as a "model" (Zhou, 2021).

There are a few main categories of machine learning algorithms: supervised, unsupervised, and reinforcement learning. These categories are based on the learning technique, the types of data they input and produce, and the types of problems they solve. A few hybrid strategies and other widely used techniques provide an organic expansion of machine learning issue types (Sah, 2020).

Supervised learning is a type of machine learning that uses labelled datasets. These datasets are developed to train or "supervise" algorithms so they can correctly categorise data or predict outcomes. The model may evaluate its correctness and improve over time using labelled inputs and outputs (Delua, 2021).

Machine learning algorithms are used in unsupervised learning to examine and cluster unlabelled datasets. These algorithms are referred to as "unsupervised" because they identify hidden patterns in data without human assistance (Delua, 2021).

Popular supervised machine learning algorithms are logistic regression, random forest, support vector machine algorithm and Naïve Bayes algorithm (Tavasoli, 2022). The supervised machine learning algorithms used in this project are Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours, Naïve Bayes and Neural Network.