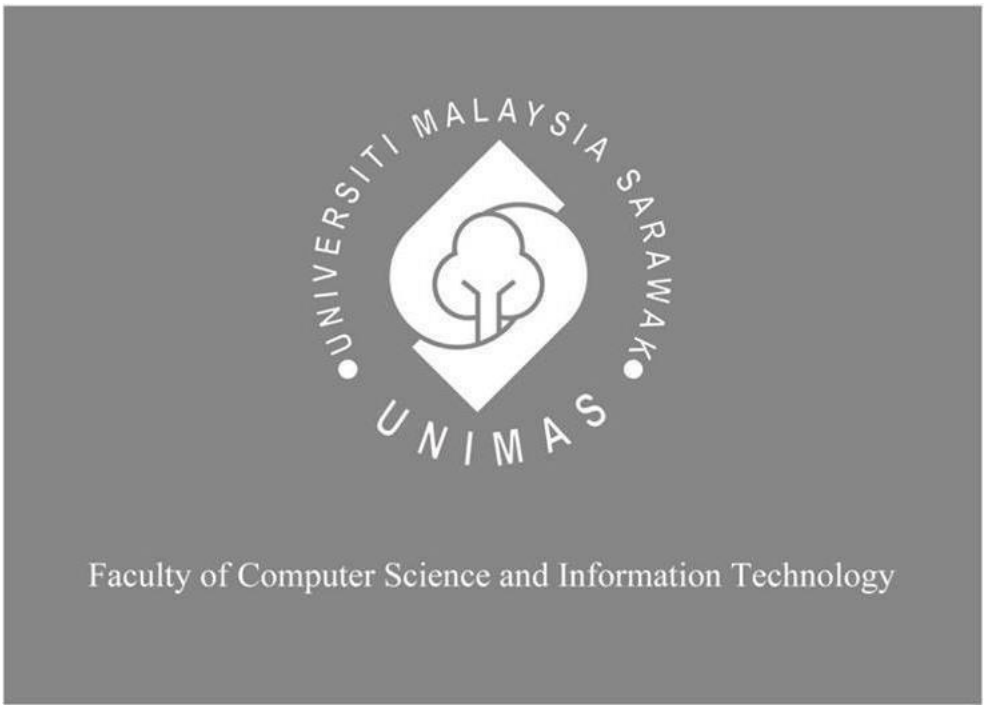


A COMPARATIVE STUDY OF MACHINE LEARNING MODELS FOR PREDICTING THE RISK OF ALZHEIMER'S DISEASE

Dexter Chai Tze Loong

Bachelor of Computer Science with Honours

(Information Systems)

2025

A COMPARATIVE STUDY OF MACHINE LEARNING MODELS FOR PREDICTING THE RISK OF ALZHEIMER'S DISEASE

DEXTER CHAI TZE LOONG

This project is submitted in partial fulfilment of the requirements for
the degree of Bachelor of Computer Science with Honours
(Information System)

Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK

2025

KAJIAN PERBANDINGAN MODEL PEMBELAJARAN MESIN UNTUK MERAMAL RISIKO PENYAKIT ALZHEIMER

DEXTER CHAI TZE LOONG

Projek ini merupakan salah satu keperluan untuk

Ijazah Sarjana Muda Sains Komputer dengan Kepujian

(Sistem Maklumat)

Fakulti Sains Komputer dan Teknologi Maklumat

UNIVERSITI MALAYSIA SARAWAK

2025

UNIVERSITI MALAYSIA SARAWAK

THESIS STATUS ENDORSEMENT FORM

TITLE **A COMPARATIVE STUDY OF MACHINE LEARNING MODELS FOR PREDICTING THE RISK OF ALZHEIMER'S DISEASE**

ACADEMIC SESSION: **2024/2025**

DEXTER CHAI TZE LOONG

(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

NO 4 POH MING PARK
FOOCHOW ROAD NO.1
93300 KUCHING SARAWAK

Date: **22 JULY 2025**

Date: **22 JULY 2025**

Note * Thesis refers to PhD, Master, and Bachelor Degree

** For Confidential or Restricted materials, please attach relevant documents from relevant organizations / authorities

DECLARATION

I, Dexter Chai Tze Loong, 79218, hereby declare that the project titled "A Comparative Study Of Machine Learning Models For Predicting Risk Of Alzheimer's Disease" is my original work. I have not plagiarized works from others or other sources in completing the project, except where references or acknowledgments are provided. Furthermore, no part of this work has been written by others on my behalf. I have diligently adhered to the academic guidelines and ethical standards set by the Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak (UNIMAS), in completing this project.

……. 20/07/2025

DEXTER CHAI TZE LOONG DATE

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to my project supervisor, Dr. Stephanie Chua Hui Li, for her invaluable guidance and encouragement, and to Prof. Dr. Wang Yin Chai, the Final Year Project Coordinator, for his support throughout this journey and guidance in the project's format. I am also thankful to Universiti Malaysia Sarawak (UNIMAS) and the Faculty of Computer Science and Information Technology (FCSIT) for providing a nurturing academic environment for me to successfully deliver the project. Finally, I am deeply grateful to my friends and family for their unwavering support, motivation, and understanding, which have been instrumental in the successful completion of this project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

ABSTRACT

The main objective of this project is to compare machine learning models to predict the risk of Alzheimer's disease. As the aging population continues to grow, the number of Alzheimer's disease patients is becoming increasingly concerning. This trend is often associated with unhealthy lifestyle patterns, which may exacerbate the risk. Thus, this project uses machine learning to predict Alzheimer's disease risk, enabling individuals to assess their risk and take proactive steps to reduce it. This project proposed a classification approach for this comparative study to predict the risks. This is implemented using the Knowledge Discovery in Databases (KDD) methodology with steps including data selection, pre-processing, transformation, data mining, and evaluation. The proposed prediction models are logistic regression, decision tree, support vector machine (SVM), CatBoost, XGBoost, random forest, and artificial neural network (ANN). In this study, the dataset collected from Kaggle underwent preprocessing involving data cleaning, feature selection, data engineering, and data standardization. It was then divided into training (80%) and testing (20%) sets using the holdout method. Next, these various models were compared using evaluation metrics derived, such as accuracy, precision, recall, F1 score, and sensitivity from the confusion matrix to identify the best-performing models. From this study, the best-performing model is the XGBoost model. It shows the highest performance among all the machine learning algorithms with accuracy at 95.3%, precision at 95.5%, recall at 94.3%, and F1-score at 94.9%.

ABSTRAK

Objektif utama projek ini adalah untuk membandingkan model-model pembelajaran mesin bagi meramalkan risiko penyakit Alzheimer. Dengan pertambahan populasi warga emas, bilangan pesakit Alzheimer semakin membimbangkan. Trend ini sering dikaitkan dengan corak gaya hidup yang tidak sihat, yang boleh meningkatkan risiko tersebut. Oleh itu, projek ini menggunakan pembelajaran mesin untuk meramalkan risiko penyakit Alzheimer, membolehkan individu menilai tahap risiko mereka dan mengambil langkah proaktif untuk mengurangkannya. Projek ini mencadangkan pendekatan klasifikasi bagi kajian perbandingan ini untuk meramalkan risiko. Ia dilaksanakan menggunakan metodologi Penemuan Ilmu dalam Pangkalan Data (Knowledge Discovery in Databases - KDD) yang merangkumi langkah-langkah seperti pemilihan data, pra-pemprosesan, transformasi, perlombongan data, dan penilaian. Model-model ramalan yang dicadangkan ialah regresi logistik, pokok keputusan, mesin vektor sokongan (SVM), CatBoost, XGBoost, hutan rawak (random forest), dan rangkaian neural buatan (ANN). Dalam kajian ini, set data yang dikumpulkan dari Kaggle telah menjalani pra-pemprosesan melibatkan pembersihan data, pemilihan ciri, kejuruteraan data, dan penyeragaman data. Data tersebut kemudian dibahagikan kepada set latihan (80%) dan ujian (20%) menggunakan kaedah holdout. Seterusnya, pelbagai model ini dibandingkan menggunakan metrik penilaian seperti ketepatan (accuracy), kepersisan (precision), kepekaan (recall), skor F1, dan sensitiviti yang diperoleh daripada matriks kekeliruan untuk mengenal pasti model dengan prestasi terbaik. Daripada kajian ini, model yang menunjukkan prestasi terbaik ialah model XGBoost, dengan ketepatan sebanyak 95.3%, kepersisan 95.5%, kepekaan 94.3%, dan skor F1 sebanyak 94.9%.

CHAPTER 1: INTRODUCTION

1.1 Introduction

Dementia is defined as the loss of memory, difficulty processing thoughts, speech impairments, and other cognitive challenges that are significant enough to disrupt daily life. The most common cause of dementia is Alzheimer's disease, which is classified as a brain disorder caused by the build-up of proteins in the brain, leading to brain cell death and loss of brain function (Alatrany et al., 2024). According to Kasani, there are over 50 million people diagnosed with Alzheimer's disease, and this number is set to reach 152 million by 2050 (Kasani et al., 2021). He also mentioned that individuals with a history of cognitive impairment or mental illness are at a higher risk of developing Alzheimer's disease, with 32% of them likely to progress to Alzheimer's within five years.

To make matters worse, Alzheimer's disease is often irreversible and has no known cure, despite significant resources being invested in finding one.(Korczyn, 2022). However, studies have shown that only a small percentage of Alzheimer's disease cases are genetically linked; most are influenced by environmental factors. Thus, the risk of developing Alzheimer's can be reduced through early intervention, which includes adopting a healthy lifestyle, a balanced diet, quality sleep, and regular exercise (Korczyn, 2022). For example, studies in Finland found that people with a healthy and active lifestyle had a 60% lower risk of developing Alzheimer's compared to those who remained sedentary. (Arab & Sabbagh, 2010). But, to effectively reduce the risk of Alzheimer's disease, action must begin as early as possible, ideally in midlife but not in old age.

Thus, it is essential to predict any early signs or symptoms linked to Alzheimer's disease and act on them promptly. This awareness can help individuals address lifestyle and health

conditions, such as engaging in regular physical activity, following a balanced diet, having strong social connections, and engaging in cognitive activities that stimulate the brain, like puzzles or learning new skills. By practicing a healthy lifestyle early, individuals may improve their overall brain health, potentially reducing the risk of Alzheimer's disease.

This project looks at predicting the risk of Alzheimer's disease using a machine learning approach. With this predictive tool, individuals can assess their risk of Alzheimer's disease and take proactive measures, such as adopting a healthier lifestyle, to reduce their chances of developing this disease.

1.2 Problem Statement

There are over 50 million people diagnosed with Alzheimer's disease, and this number will increase to 152 million by 2050 (Kasani et al., 2021). As the global aging population continues to rise, particularly in China, Japan, and other developing countries, the likelihood of being diagnosed with this disease also increases. Additionally, many individuals who did not or cannot adopt a healthy lifestyle in midlife have a higher chance of being diagnosed with Alzheimer's disease. This may be attributed to factors such as constant work pressure, night shifts, poor diet, and excessive smoking. Furthermore, the lack of regular medical check-ups and late detection contribute to the issue. People often lack awareness about Alzheimer's disease and neglect routine check-ups. Unfortunately, when severe symptoms appear, effective solutions are often unavailable. Thus, our project aims to predict the risk of Alzheimer's disease, so that proactive measures and early intervention can be taken to reduce the risks of Alzheimer's disease.

1.3 Scope

This project's scope is to use machine learning models to predict the risk of Alzheimer's disease. It utilizes datasets related to Alzheimer's disease from the Kaggle platform to obtain raw data and employs Python libraries for data pre-processing. Then, machine learning algorithms will be implemented for model learning to predict Alzheimer's disease. To evaluate the model, several model evaluation metrics such as accuracy, precision, recall, and accuracy will be used to assess the model's performance and determine the best-performing model from the rest.

1.4 Objectives

The objectives of this project are:

- To select the appropriate features for predicting the risk of Alzheimer's disease. This enables the model to focus on the most effective predictors for accurate predictions.
- To compare the performance of the machine learning model in predicting Alzheimer's disease. The comparison will help understand which model generalizes best across the diverse patient data.
- To determine the best machine learning model for predicting Alzheimer's disease. The best model will be used to develop a prediction system.

1.5 Methodology

The methodology used in this project for predicting Alzheimer's disease through machine learning follows the Knowledge Discovery in Databases (KDD) approach. Knowledge Discovery in Databases (KDD) is defined as the process of finding useful knowledge from data (Fayyad,1996). This process involves five key steps that the project follows: selection, pre-processing, transformation, data mining, and evaluation. Figure 1.1 depicts the process flow of the Knowledge Discovery in Databases (KDD) from the selection phase to the evaluation phase.

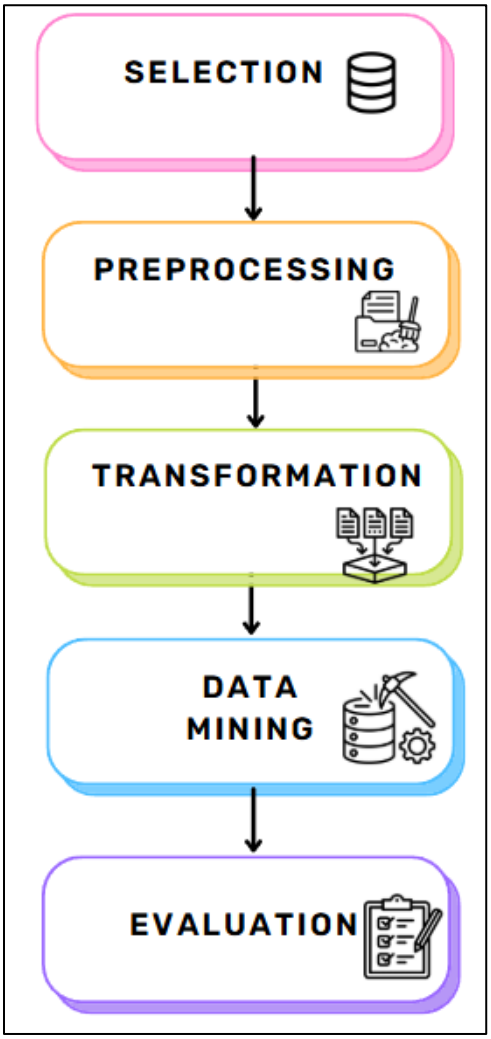


Figure 1.1 Knowledge Discovery in Databases (KDD) Process

a) Selection

In this step, data attribute is identified and selected based on their relevance, quality, impact, and availability to the project's objectives. This is essential for the later stages of data mining, as it forms the basis for data modelling (Rohit, 2024).

b) Pre-processing

In this step, the collected data undergoes cleaning and pre-processing. This involves removing missing, low-quality, and noisy data, and performing data discretization to improve the quality of data for modelling (Rohit, 2024).

c) Transformation

In this step, the data are consolidated and aggregated to prepare for data mining algorithms. The data is transformed into attributes, features, and functions more suitable for data mining. (Rohit, 2024).

d) Data Mining

In this step, machine learning algorithms are applied to extract meaningful patterns from the processed data, which are used for the project's predictions. Various algorithms and techniques can be used to uncover these patterns (Rohit, 2024).

e) Evaluation

In the final step, the patterns obtained from the data mining process are evaluated using appropriate metrics and visualized in various formats, such as pie charts,

histograms, bar graphs, and line graphs. This representation improves the evaluation of effectiveness (Rohit, 2024).

1.6 Significance of the Project

This project aims to predict the risk of Alzheimer's disease using various machine learning techniques, compare their techniques, and determine their accuracy in predicting the disease. This project also studies the significance of various attributes such as age, lifestyle, and illness that could contribute to Alzheimer's disease. Thus, we can classify the risk of Alzheimer's disease based on our prediction. With this project, people can utilize the project's prediction to predict individuals' risk, thus becoming more aware of this disease, taking proactive action against it, and practicing a healthy lifestyle.

1.7 Project Schedule

The project schedule is used to monitor task completion, including start dates, end dates, and durations. Additionally, this project uses a Gantt chart for a visual representation of the timeline and progress. Table 1.1 shows the project schedule's deliverables recorded with start dates and end dates.

Table 1.1 The Table of Project Schedule

Deliverables	Start Date (mm/dd/year)	End Date (mm/dd/year)	Duration (days)
Submission of an approved Brief Proposal	10/10/2024	10/19/2024	9
Feedback and Comments from Examiners	10/19/2024	11/14/2024	26
Submission of Full Proposal	11/10/2024	11/14/2024	4
Chapter 1 Submission	11/10/2024	11/21/2024	11
Chapter 2 Submission	11/21/2024	12/13/2024	22
Chapter 3 Submission	12/13/2024	1/5/2025	23
Submission of FYP 1 Report for Review	1/5/2025	1/17/2025	12
Submit Video Presentation for FYP 1	1/17/2025	1/19/2025	2
Amendment For FYP 1	1/17/2025	2/19/2025	33
FYP 1 Report Submission after Amendment	2/10/2025	2/19/2025	9
Submission of Revised FYP Structure	3/24/2025	4/01/2025	8
Chapter 4 Submission	4/01/2025	5/15/2025	44
Chapter 5 Submission	5/15/2025	5/31/2025	16
Chapter 6 Submission and Abstract for Paper	5/15/2025	5/31/2025	16
Submission of FYP 2 Final Report for Review	5/31/2025	6/10/2025	10
Presentation for FYP 2	6/10/2025	6/15/2025	5
Amendment For FYP 2	6/10/2025	7/04/2025	19
Submission Of Final Report	6/30/2025	7/28/2025	28

The Gantt chart shown in Figure 1.2, with a clear visualisation of the project's timeline and progress from the beginning to the end.

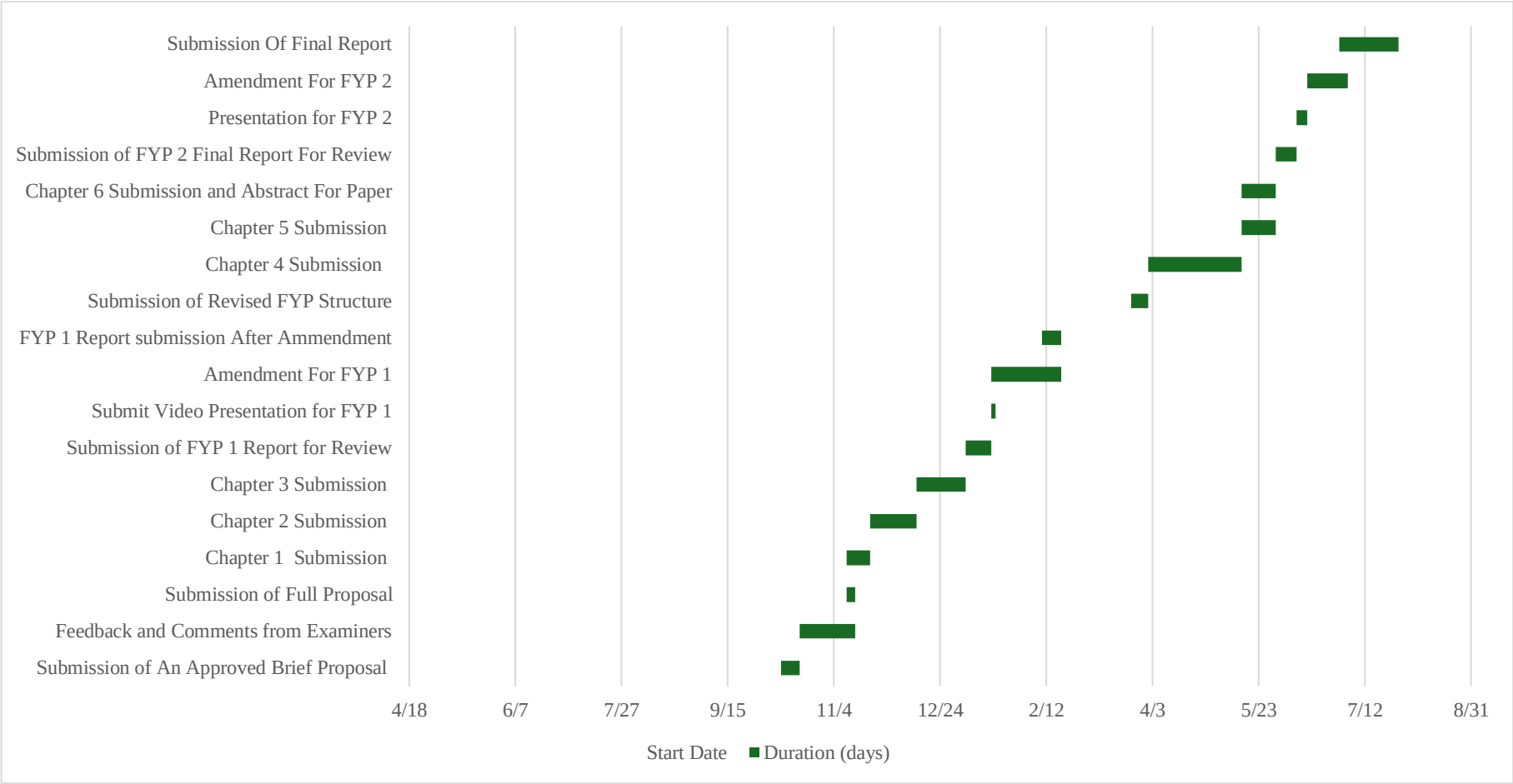


Figure 1.2 The Gantt Chart of the Project

1.8 Expected Outcome

The expected outcome of this project is to predict the risk of Alzheimer's disease using various machine learning models, along with a comparative evaluation. The project aims to identify the best model with the highest evaluation score, which will then be demonstrated using user input for predictions.

1.9 Report Outline

This report is structured into five chapters, each providing detailed discussions on the research issue and its objectives. In the first chapter, the introduction elaborates the project background, problem statement, objectives, methodology of the project, scope of the project significance, along with the project schedule, and its expected outcomes.

Next, the second chapter on literature review introduces the relevant literature from articles, books, and online resources related to the project. Additionally, it includes a comparative analysis of the related works, discussing their research and the machine learning algorithms used.

The third methodology chapter provides a detailed discussion of the approaches used in this project. It focuses on the Knowledge Discovery in Databases (KDD) process, outlining the steps in detail. Also, evaluation techniques such as the confusion matrix are outlined too. These KDD steps cover the data collection (selection), preprocessing, transformation, data mining, and evaluation methods.

For the fourth chapter on data mining and implementation, the chapter discusses the implementation of the model, which includes data selection, data pre-processing, data

transformation, data mining, and the evaluation of the model. Additionally, this chapter also includes details on the implementation of the application and deployment in Hugging Face.

The fifth chapter on results and analysis discusses the evaluation results of the machine learning model and its comparative analysis with various evaluation metrics. The results from the evaluation will be explained in detail and offer in-depth insight. Next, the best-performing machine learning model will be determined and selected from the analysis.

For the last chapter on conclusion and future work, the project discusses the project summary, contribution, potential areas for future work, and the limitations of the project to identify opportunities for further improvement.

1.10 Summary

This chapter provides an overview of Alzheimer's disease, outlining its prevalence and explaining how this project uses machine learning to predict the risk of developing the disease, which may aid in preventing people from getting the disease. Additionally, this project also summarizes the methodology used, Knowledge Discovery in Databases (KDD), highlighting the significance of the project and its expected outcomes in predicting the risk of Alzheimer's disease.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter introduces the background of Alzheimer's disease, discusses the common data mining process used for diagnostic prediction, presents the machine learning algorithms, and presents the tools used in this project. This chapter will also review the relevant literature from articles, books, and online resources. Additionally, it includes a comparative analysis of the associated works, discussing their research and the machine learning algorithms used. The research paper reviewed focuses on related work in analysing and predicting Alzheimer's disease using Artificial Intelligence (AI) approaches, including machine learning, deep learning, or a combination of both.

2.2 Background Knowledge of Alzheimer's Disease

Alzheimer's disease is a progressive disease caused by the abnormal build-up of proteins in the brain, such as amyloid plaques and tau, which is a common cause of dementia. This build-up leads to the progressive death of brain cells, resulting in brain shrinkage and, ultimately, a loss of brain function to carry out daily tasks (An, 2022). This is shown in Figure 2.1, which depicts the progression of a shrinking brain due to Alzheimer's disease.

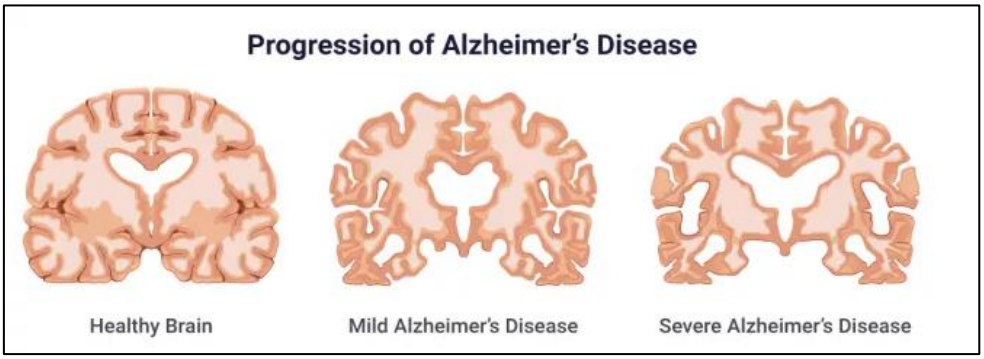


Figure 2.1 Progression of Alzheimer's Disease (The Helper Bees, 2024)

Symptoms of Alzheimer's disease most commonly appear in people over the age of 60 and can progress from early to later stages. Early symptoms include forgetfulness and confusion, while later symptoms may involve gradual loss of speech and physical abilities like sitting, walking, and swallowing (Kavitha et al., 2022).

The causes are believed to be due to a combination of factors, including genetics, health conditions, lifestyle, and environmental influences. For example, in Kornblith's research, he found an association between race and ethnicity in the risk of developing Alzheimer's disease. His findings show that the risk of Alzheimer's disease is higher among older Black and Hispanic participants, which could be attributed to the genetic difference among races (Kornblith et al., 2022). However, the cause of Alzheimer's disease is mostly influenced by environmental factors rather than genetics (Korczyn, 2012), such as smoking, alcohol consumption, diet, sleep quality, and various health conditions.

Thus, by incorporating environmental factors and common lifestyle traits, we can use a machine learning approach to predict the risk of Alzheimer's disease. This predictive approach enables early detection, allowing individuals to assess risk and take preventive actions. With such a tool, people can adopt healthier lifestyles, potentially reducing their likelihood of developing the disease.

2.3 Data Mining Process

The data mining process is defined as the use of advanced analytics techniques to identify patterns and relationships in large data sets, to uncover useful information (Gillis, 2024). Many data mining processes can be used for prediction projects, but the most common ones are Knowledge Discovery in Databases (KDD) and the Cross-Industry Standard Process for Data Mining (CRISP-DM).

2.3.1 Knowledge Discovery in Databases (KDD)

Knowledge Discovery in Databases (KDD) is defined as the process of finding useful knowledge from data(Fayyad,1996). This process involves five key steps that the project follows: selection, pre-processing, transformation, data mining, and evaluation.

a) Selection

In this step, data attribute is identified and selected based on their relevance, quality, impact, and availability to the project's objectives. This is essential for the later stages of data mining, as it forms the basis for data modelling (Rohit, 2024).

b) Pre-processing

In this step, the collected data undergoes cleaning and pre-processing. This involves removing missing, low-quality, and noisy data, and performing data discretization to improve the quality of data for modelling (Rohit, 2024).

c) Transformation

In this step, the data are consolidated and aggregated to prepare for data mining algorithms. The data is transformed into attributes, features, and functions more suitable for data mining (Rohit, 2024).

d) Data Mining

In this step, machine learning algorithms are applied to extract meaningful patterns from the processed data, which are used for the project's predictions. Various algorithms and techniques can be used to uncover these patterns (Rohit, 2024).

e) Evaluation

In the final step, the patterns obtained from the data mining process are evaluated using appropriate metrics and visualized in various formats, such as pie charts, histograms, bar graphs, and line graphs. This representation improves the evaluation of effectiveness (Rohit, 2024).

2.3.2 Cross-Industry Standard Process for Data Mining (CRISP-DM)

The CRISP-DM methodology includes a structured approach for planning a data mining project. It is an effective and robust methodology widely used across various domains (Smart Vision Europe, 2024). This process involves six key phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. Figure 2.2 shows the process flow of the Cross-Industry Standard Process for Data Mining from the business understanding phase to the deployment phase.

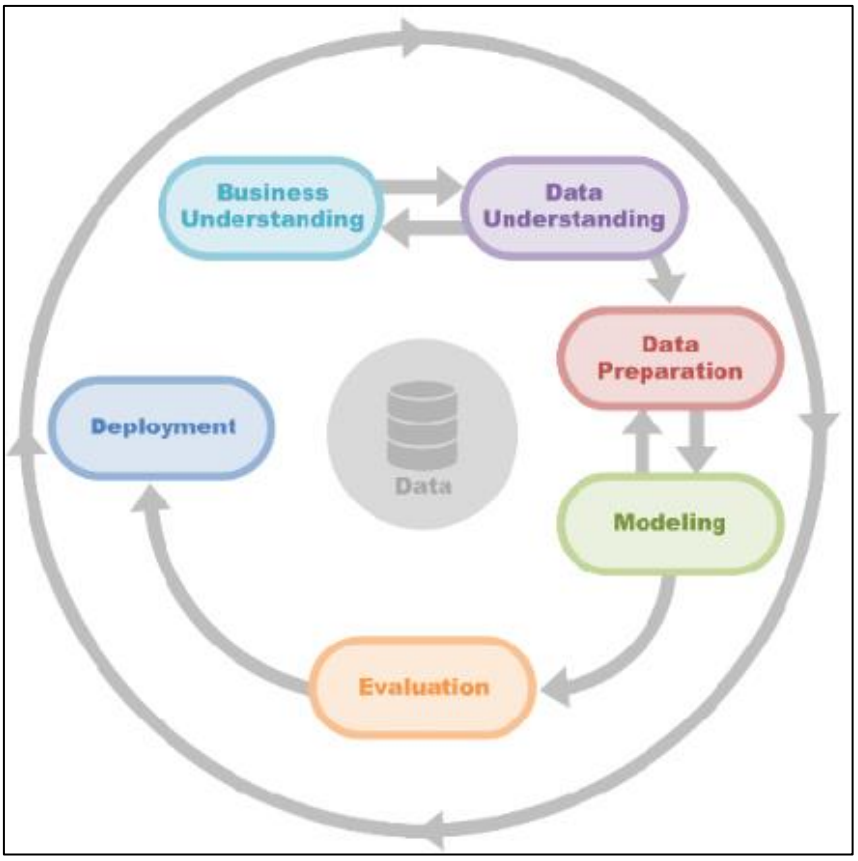


Figure 2.2 Cross-Industry Standard Process for Data Mining (CRISP-DM) Diagram (Tounsi et al., 2020)

a) Business Understanding

In this phase, important factors and objectives are determined. The goal is to understand the purpose of the project and define the objectives, create a project plan, and establish success criteria (Smart Vision Europe, 2024).

b) Data Understanding

In this phase, the data listed is acquired and loaded into specific tools for data understanding. The acquired data is described in terms of its format, quantity, and fields. It is then evaluated to determine whether it meets the requirements of the project (Smart Vision Europe, 2024).

c) Data Preparation

In this phase, the data is cleaned, pre-processed, derived, and merged for analysis. The preparation of the data depends on the data mining goals and aims to improve data quality for the next phase of the project (Smart Vision Europe, 2024).

d) Modelling

In this phase, machine learning algorithms are applied to create models. A test design is also developed to assess the model's quality, accuracy, and validity when the algorithm is used. The models are then interpreted based on domain knowledge and evaluated by domain experts (Smart Vision Europe, 2024).

e) Evaluation

In this phase, the evaluation is not just about the accuracy of the model, but the extent to which the model meets the business objectives and goals in the first phase. The evaluation phase can also be done on real applications (Smart Vision Europe, 2024).

f) Deployment

In this phase, a deployment plan and monitoring and maintenance plan are developed by summarising the procedure to create the model and the procedure to maintain it. Finally, a final report is produced for the entire project (Smart Vision Europe, 2024).

In this project, Knowledge Discovery in Databases (KDD) is used. This methodology will be further discussed in Chapter 3.

2.4 Machine Learning Algorithms

Machine learning is a subset of artificial intelligence (AI) that focuses on the learning of computer systems. And with machine learning algorithms, we can use them to predict, discover, and find patterns in data. For example, by using past data as input, machine learning algorithms can predict the likelihood of an occurrence based on patterns and relationships identified in the previous data (Lev, 2024). For this research, supervised learning approaches are used. In supervised learning, external supervision is required for the algorithms to learn, and the datasets need to be labelled (Mahadevkar et al., 2022). The various supervised learning algorithms used in this project are logistic regression, decision tree, support vector machine (SVM), CatBoost, XGBoost, random forest, and artificial neural network (ANN).

I. Logistic Regression

Logistic regression is a technique used to develop a predictive model by taking an input variable (independent variables) and checking whether it can predict a dichotomous (binary) dependent variable (Messaoud et al., 2020). When fitting the data into the logit function, it will produce probability values ranging from 0 and 1. This helps predict and classify datasets. The logit function formula is as shown in Equation 2.1, with Y calculating the probability.

$$\boldsymbol{Y = Logit(P) = Ln(P/(1-P)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n + \epsilon}$$

Equation 2.2 Logistic Model Probability (Messaoud et al., 2020)

II. Decision Tree

The decision tree is a representation of a hierarchical data structure. It is arranged such that the data structure follows a form of decision sequences for predicting and classifying the result (Messaoud et al., 2020). In a decision tree, a node's set is divided into root nodes, internal nodes (nodes with descendants), and terminal nodes (nodes with no descendants). When an input is in the tree, it is tested by internal nodes, and the decisions are made in the terminal nodes. It is important to note that the selection of attributes is important for placing each node in the tree, as it helps ensure the constructed tree remains small. This reduces the complexity, makes the tree more efficient, and allows for easier predictions.

III. Support Vector Machine (SVM)

The support vector machine (SVM) algorithm is used for the classification of qualitative variables by finding a hyperplane that best separates the data into two classes. The hyperplane is positioned to maximize the distance, or margin, from the closest data points in each class, ensuring optimal separation (Messaoud et al., 2020).

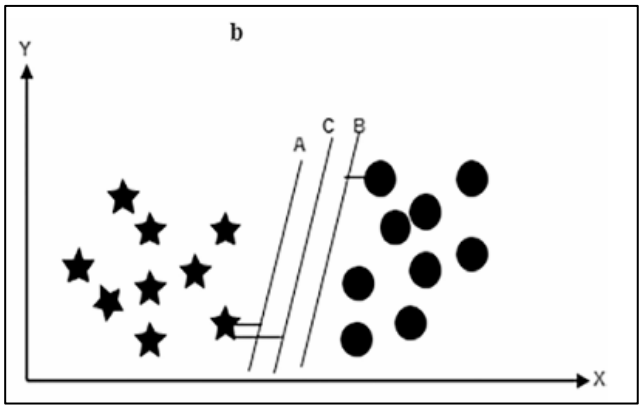


Figure 2.3 Hyperplane of the SVM Algorithm(Messaoud et al., 2020)

For example, Figure 2.3 above shows two classes of circles and stars with lines (hyperplane) A, B, and C. The distance between the nearest class and the hyperplane is called the margin. According to the SVM principle, the optimal hyperplane is such that the margin is highest, that is, the maximum distance between the nearest class to the line. Thus, the hyperplane C is the optimal hyperplane.

IV. CatBoost

CatBoost, also known as Categorical Boosting, was released in 2017 and is a gradient boosting algorithm based on decision trees. It uses a unique boosting algorithm designed to handle categorical features effectively. By applying a permutation method with an ordering principle, it combines categorical features into new features to enhance the effectiveness of the model. This algorithm is particularly useful for tasks involving complicated datasets and data that are heterogeneous (An, 2022). Figure 2.4 illustrates the classification process of the CatBoost algorithm, where features are combined and weighted to predict the target values.

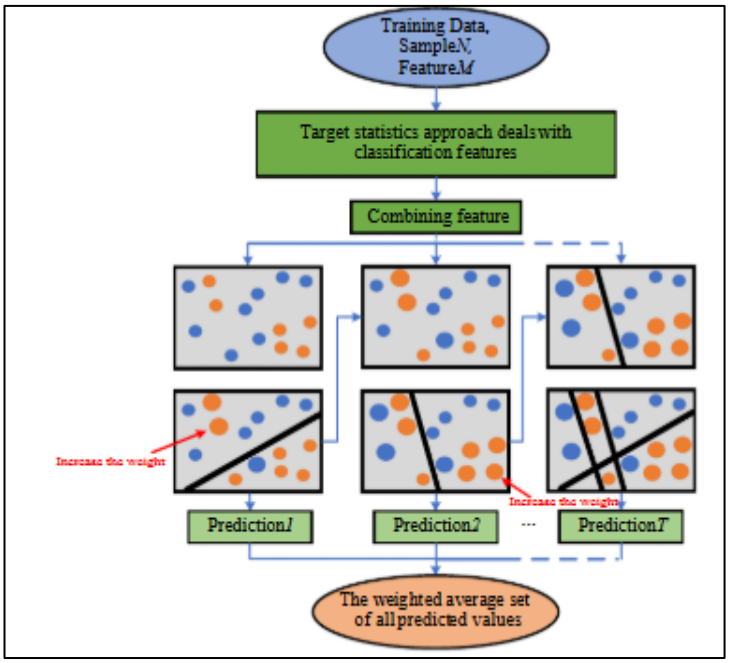


Figure 2.4 Flow Chart of CatBoost Algorithm (Chang et al., 2023)

V. Extreme gradient boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a parallel tree-boosting algorithm designed for fast and accurate performance across a wide range of datasets. XGBoost is also an ensemble learning algorithm focusing on decision trees and leveraging gradient-boosting techniques. It combines the predictions of multiple weak models to generate stronger and more reliable predictions, making it particularly effective for structured data and various prediction problems (Arif Ali et al., 2023). Figure 2.5 below illustrates the process involved in the XGBoost Algorithm classification. It shows how many trees are combined and ensembled to produce a final result.

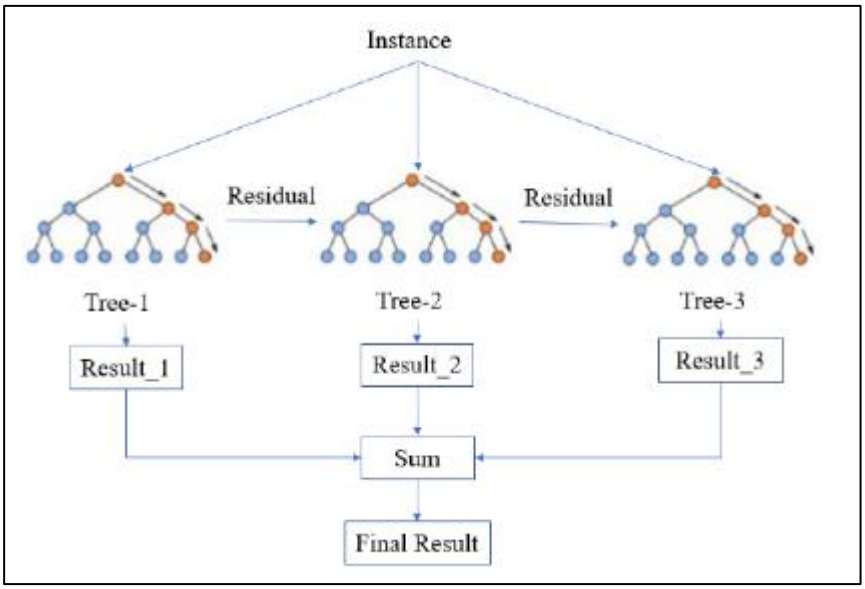


Figure 2.5 XGBoost Classification Structure (Wang et al., 2021).

VI. Random Forest

Random Forest is developed by merging the output of several decision trees and produces a single result (Sruthi, 2024). This is a powerful and effective algorithm that is widely used for classification and regression problems. Random Forest follows the principle of ensemble learning, where multiple decision trees are combined by voting or by averaging. This can improve the prediction accuracy of the model. Figure 2.5 illustrates the classification process of the Random Forest algorithm, where multiple decision trees are averaged to produce a result that balances the outcomes of the individual trees.

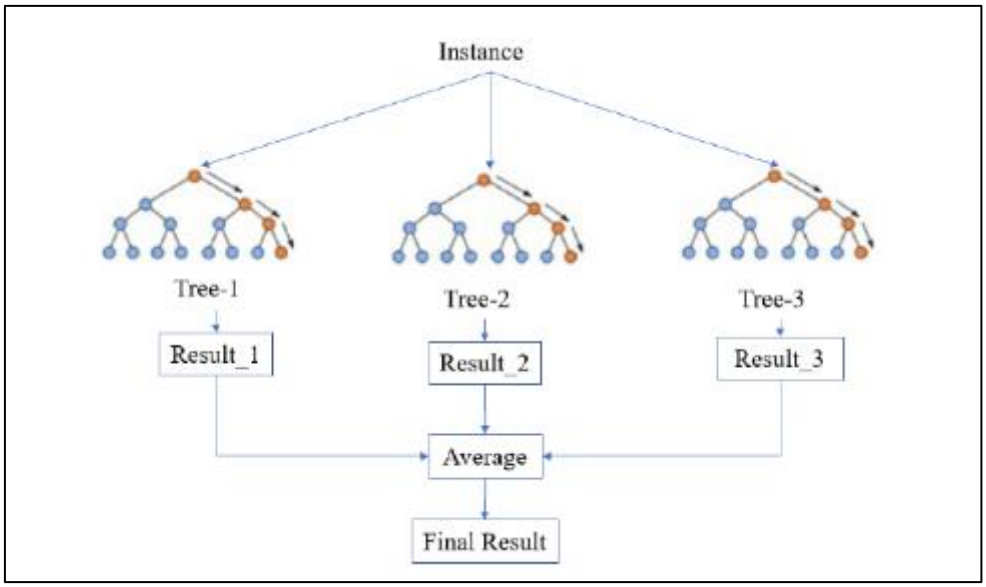


Figure 2.6. Random Forest ClassifierStructure (Wang et al., 2021)

VII. Artificial Neural Network (ANN)

The neural network classifier is the representation of human biological neurons with a high level of complexity, using various types of neurons with different activation functions. This creates statistical reasoning based on its learning from past data (Messaoud et al., 2020). In Figure 2.7, the neural network is trained using input-target pairs. When a new input is presented, the network generates a prediction based on the learned weights. A neural network typically consists of an input layer, one or more hidden layers, and an output layer (Messaoud et al., 2020).

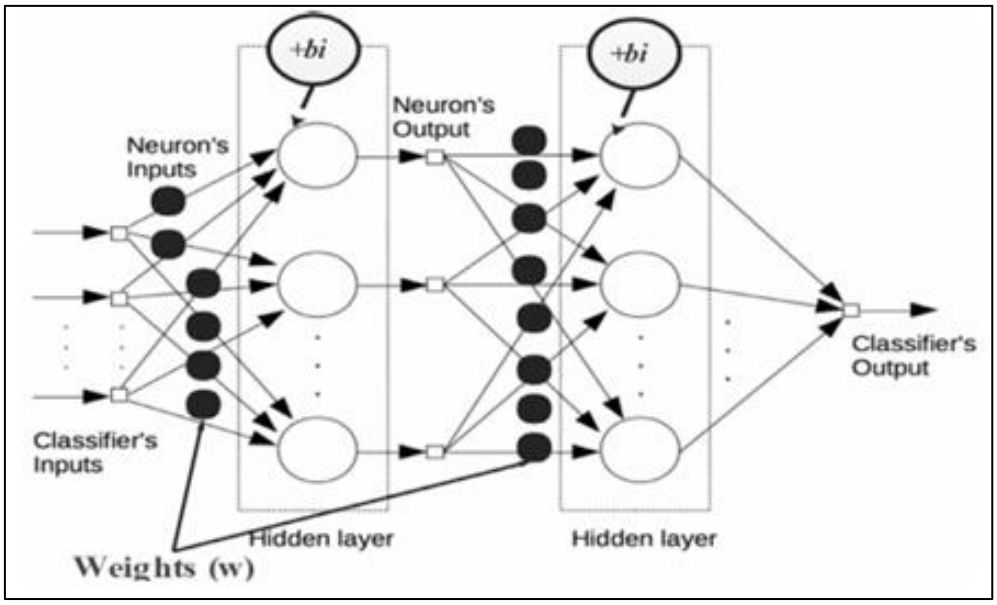


Figure 2.7 Neural Network Classifier Model (Messaoud et al., 2020)

2.5 Related Works

This section reviews past works and research in the domain of Alzheimer's disease from 2021 to 2024 with a three-year time frame. There have been several studies working on Alzheimer's disease prediction using machine learning. Thus, in this section, these related papers will be reviewed, and the findings and results will be presented in this section.

Pranao et al.(2022) conducted the study to develop a computer-based system for the early prediction of Alzheimer's disease. He points out the importance of early detection to reduce the progression of neurodegenerative diseases, particularly Alzheimer's disease. In this research, 6400 MRI images were sourced from Kaggle, with 80% of the data for training (5,121 images) and 20% for testing (1,279 images). These images consisted of four classes: Mild, Moderate, Very Mild, and Non-Demented. This research uses four classifier algorithms for their model training: support vector machine (SVM), logistic regression (LR), random forest (RF), and XGBoost. Among the models evaluated, XGBoost showed the best performance, achieving the highest metrics in both binary and multiclass, with 73% in Binary and 67% in Multiclass.

The limitation of this paper is its moderate accuracy compared to other literature reviewed in this project. Future work could focus on refining models, data enhancement, and incorporating additional factors to improve the accuracy of Alzheimer's disease prediction.

Next, Alatrany et al. (2024) worked on the study to develop a machine learning model that can classify the subjects into different stages of Alzheimer's disease, such as normal control (NC), mild cognitive impairment (MCI), and Alzheimer's disease (AD). The author conducted five experiments that aimed to evaluate the machine learning algorithms' performance. In the last experiment, only the best-performing classifiers will be selected for further testing, after other algorithms have been filtered out in previous experiments. The dataset used for this research was collected from the National Alzheimer's Coordinating Centre (NACC), comprising 169.408 records and 1,024 features. This research uses four classifier algorithms: support vector machine (SVM), random forest (RF), K-nearest neighbours (KNN), and naïve Bayes (NB). Among the models evaluated, SVM is the best-performing model, achieving the highest metrics, but the RF model is close too. This paper is limited by the implementation of feature correlation analysis, which can lead to information loss, reduced precision, and an incomplete assessment of the model's performance. Thus, in this project, this limitation can be reduced by lowering the correlation threshold to help retain valuable features.

Kasani et al.(2021) researched to show the advantages of machine learning-based computer-aided diagnosis (CAD) in classifying subjects into three categories: cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD). Then, his research goals were to identify the most accurate machine learning model. The datasets used in this research are from the Korean Brain Aging Study for Early Diagnosis and Prediction of Alzheimer's Disease (KBASE). In KBASE, 336 individuals were classified as CN (n=173), MCI (n=89), and AD (n=74). All subjects aged 55 to 90 years old, where AD is most likely to

occur. This research uses eight machine learning algorithms: support vector machine (SVM), naïve Bayes (NB), extreme gradient boosting (XGBoost), decision tree (DT), logistic regression, (LR), random forest, (RF), bootstrap aggregation (Bagging) and adaptive boosting (AdaBoost). For evaluation, the best-performing model is the XGBoost classifier, yielding an accuracy of 82.09%. Despite showing satisfactory results, this paper is limited by a small dataset of only 336 individuals from KBASE and a lack of demographic diversity, as most of the participants are of Korean ethnicity. This could lead to overfitting and a lack of generalisability.

Dixit et al.(2024) conducted their research to study and compare the performance of machine learning and deep learning algorithms with medical images. The datasets used in this research are sourced from Kaggle, with 7,839 neuroimaging samples. This research uses four machine learning algorithms: random forest, logistic regression, support vector classifier (SVM), and multi-layer perceptron (MLP). They also use another five deep learning algorithms for comparison: EfficientNetB3, EfficientNetV2S, MobileNetV2, NASNetMobile, and ResNet50. Among the models evaluated, for machine learning algorithms, random forest is the best-performing algorithm with 92.58% accuracy, while for deep learning, the best-performing is the ResNet50 with 99.75% accuracy.

Next, Alatrany et al.(2023) conducted another research with a different approach. They apply machine learning techniques to predict Alzheimer's disease for late-onset cases. The author also wants to identify the highest accuracy level of the machine algorithm in this prediction. The datasets used in this research are sourced from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, with 2,404 subjects after filtering unwanted data. This research uses five machine learning algorithms: logistic regression (LR), linear discriminant analysis (LDA), k-nearest neighbours (KNN), Gaussian naïve Bayes (GNB), and

support vector machine (SVM). After the evaluation, it was determined that linear discriminant analysis (LDA) performed the most efficiently, scoring the F1-score and accuracy of 1 (100%) in Dataset2 (subjects ≥65). A key limitation of this paper is the extremely high accuracy of 100% of LDA, which is likely unrealistic. This suggests potential overfitting and raises concerns about the model's robustness and generalizability to unseen data. Future work should focus on cross-validation techniques and regularization methods.

Kavitha et al.(2022) aimed to provide a timely intervention and treatment for Alzheimer's disease by using a machine learning approach in predicting it. The authors noted that early detection of Alzheimer's disease could potentially slow the disease's progression, as the prediction enables more effective management strategies early on. The dataset in this research uses the Open Access Series of Imaging Studies (OASIS) and the Kaggle dataset for modelling. This research uses five machine learning classifiers: support vector machine (SVM), random forest classifier, decision tree classifier, gradient boosting, and voting classifier. The results from this research show that the best-performing techniques are random forest and XGBoost, with an accuracy of 86.92% and 85.92%, respectively. The study also acknowledges its limitations and difficulty in identifying relevant attributes for Alzheimer's disease prediction. In future work, the authors will focus on improving the extraction of features to enhance the effectiveness of Alzheimer's disease prediction.

Kumar et al.(2024) conducted a comparative analysis between deep learning and machine learning methods for detecting Alzheimer's disease and determining their accuracy. Additionally, the study also aims to test both machine learning and deep learning approaches under conditions of data skewness and the increased complexity of interpreting diverse datasets, by comparing the performance of both methods. The research uses both machine learning and deep learning algorithms, such as random forest, logistic regression, decision trees,

support vector machine (SVM), convolutional neural network (CNN), and VGG16. After the evaluation, the best model using the "transfer learning approach" is the VGG16 model with 91% accuracy, while Random Forest at 88% accuracy with a machine learning approach. There are possible feature extraction limitations as this study uses Principal Component Analysis (PCA), which may remove subtle biomarkers in MRI scans. This can lead to information loss, affecting the prediction results.

Lastly, An(2022) published a study to analyse the performance of machine learning algorithms in predicting Alzheimer's disease using medical images of the brain. Additionally, the study also aims to identify the best-performing machine learning model in this research. The dataset is sourced from the Open Access Series of Imaging Studies (OASIS), which includes a cross-sectional dataset of 416 samples diagnosed with Alzheimer's disease (OASIS-1) and a longitudinal dataset of 373 MRI images diagnosed with Alzheimer's disease (OASIS-2). This research uses nine machine learning classifiers: logistic regression, support vector machine (SVM), naïve Bayes, decision tree, random forest, gradient boosting, AdaBoost, XGBoost, and CatBoost. After the evaluation, support vector machine, CatBoost, and decision trees are among the best-performing machine learning classifiers, ranging from 77-85% in OASIS-1 and 92-96% in OASIS-2. The author acknowledges the primary limitation of this study is the small sample size of the datasets due to the scarcity of data involving early-stage Alzheimer's disease.

2.6 Comparison of Related Works

This section will summarize the reviewed related works and present them in Table 2.1 below. The table will allosw for a comparison of the related works based on the machine learning algorithms used, a description of each study, the evaluation metrics used, and the results obtained.

Table 2.1 The Table of Comparison of Related Works.

Title	Author	Description	Machine / Deep Learning Algorithm	Evaluation Metrics	Results
Alzheimer's Disease Prediction Using Machine Learning Methodologies	Pranao, D. N., Harish, M. V., Dinesh, C., Sasikala, S., & Kumar, S. A.	• To find and predict Alzheimer's disease at an early stage • Total of 6400 MRI images were sourced from	• Support Vector Machine • Logistic Regression • Random Forest • XG Boost	• Accuracy • Precision • Recall • F1-score • Support	XG Boost showed the best performance in both binary and multiclass classification at 73% accuracy and

		Kaggle and images consisted of four classes Mild, Moderate, Very Mild, and Non-Demented. • Employs 80% training data (5,121 images) and 20% testing (1,279 images).			67% accuracy respectively.
An explainable machine learning approach for	Alatrany, A. S., Khan, W., Hussain, A., Kolivand, H.,	• To develop a machine learning model that can classify the	• Support Vector Machine • Random Forest	• Accuracy • Precision • Recall • F1-score	The Support Vector Machine is the best-performing model,

Alzheimer's disease classification	& Al-Jumeily, D.	subjects into different stages of Alzheimer's disease • Five experiments were conducted that aimed to evaluate the machine learning algorithms' performances. • Data contains 169.408 records and 1,024 features in the datasets.	• K- Nearest Neighbours • Naïve Bayes	• Mean • SD • P-value	achieving the highest metrics in all five experiments, but the Random Forest model is close, too.

An Evaluation of Machine Learning Classifiers for Prediction of Alzheimer's Disease, Mild Cognitive Impairment and Normal Cognition	Kasani, P. H., Kasani, S. H., Kim, Y., Yun, C. H., Choi, S. H., & Jang, J. W.	• To show the advantages of machine learning based computer-aided diagnosis (CAD) in classifying the subjects • To identify the most accurate machine learning model • Dataset used in this research is from the Korean	• Support Vector Machine • Naive Bayes • Extreme Gradient Boosting • Decision Tree • Logistic Regression • Random Forest • Bootstrap aggregation • Adaptive Boosting	• Accuracy • Precision • Recall • F1-score • CERAD neuropsychological test	The best-performing model is the XGBoost classifier yielding an accuracy of 82.09%, but the Bootstrap aggregation is performing as well.

		Brain Aging Study for Early Diagnosis and Prediction of Alzheimer's Disease (KBASE) with 336 individuals.			
Comparative Analysis of Deep Learning and Machine Learning Models for Alzheimer's and Parkinson's Disease	Dixit, M., Mishra, A. K., Kansal, I., Khullar, V., & Kumar, A.	• To study and compare the performances of machine learning and deep learning algorithms	• Random Forest • Logistic Regression • Support Vector Classifier • Multi-Layer Perceptron	• Accuracy • Precision • Recall • F1-score • Specificity • Validation • Loss	For machine learning algorithms, random forest is the best-performing algorithm with

Classification from Medical Images		• Datasets is sourced from Kaggle, with 7,839 neuroimaging samples in the datasets.	• EfficientNetB3 • EfficientNetV2S • MobileNetV2 • NASNetMobile • ResNet50	• G-mean	92.58% accuracy, while for deep learning, the best-performing is the ResNet50 with 99.75% accuracy.
Comparison of Machine Learning Algorithms for classification of Late Onset Alzheimer's disease	Alatrany, A. S., Hussain, A., Alatrany, S. S. J., Mustafina, J., & Al-Jumeily, D	• To use machine learning techniques to predict Alzheimer's disease in late-onset cases.	• Logistic Regression • Linear Discriminant Analysis • k-Nearest Neighbors	• Accuracy • Precision • Recall • F1-score	Linear discriminant analysis (LDA) performed the most efficiently, scoring the F1-score and accuracy of 1.

		• To identify the highest accuracy level of the machine algorithm	• Gaussian Naive Bayes • Support Vector Machine		
Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models	Kavitha, C., Mani, V., Srividhya, S. R., Khalaf, O. I., & Tavera Romero, C. A.	• To develop a system using a machine learning algorithm for early-stage Alzheimer's Disease prediction. • Dataset's in this research uses the	• Support Vector Machine • Random Forest Classifier • Decision Tree Classifier • Gradient Boosting • Voting Classifier	• Accuracy • Precision • Recall • F1-score	The best-performing techniques are random forest and XGBoost, with an accuracy of 86.92% and 85.92%, respectively.

		Open Access Series of Imaging Studies (OASIS) and the Kaggle datasets.			
Enhanced Alzheimer's Disease Prediction through Advanced Imaging: A Study of Machine Learning and Deep Learning Approaches	Kumar, N. K., Ambeth Kumar, V. D., Quamar, D., Reddy, B. S. K., Yogendra, R., & Poojitha, N.	• To do a comparative analysis between deep learning and machine learning methods for detecting Alzheimer's disease and to	• Random Forest • Logistic Regression, • Decision Trees • Support Vector Machine • Convolutional Neural Network • VGG16	• Accuracy • Precision • Recall • F1-score	The best model is the VGG16 model with 91% accuracy.

		determine their accuracy.			
Using CatBoost and Other Supervised Machine Learning Algorithms to Predict Alzheimer's Disease	An, J.	• To analyse the performance of machine learning algorithms in predicting Alzheimer's disease using medical images of the brain. • To identify the best-performing	• CatBoost • Logistic Regression • Support Vector Machine • Naive Bayes • Decision Tree • Random Forest • Gradient Boosting • AdaBoost • XGBoost	• Accuracy • Precision • Recall • F1-score • AURPC • AUROC	Support vector machine, CatBoost, and decision trees are among the best performing machine learning classifiers ranging from 77-85% in OASIS-1 and 92-96% in OASIS-2.

		machine learning model • The dataset is sourced from the Open Access Series of Imaging Studies (OASIS): Cross-sectional dataset of 416 samples diagnosed with Alzheimer's disease (OASIS-1) and a longitudinal			

		dataset of 373 MRI images			

Summary of Findings

The findings from the related works (2019-2024) reveal that various machine learning techniques have been employed for Alzheimer's prediction, including comparisons involving deep learning approaches. Most of these studies use imaging datasets, particularly MRI scans. This could be due to the recent trend that focuses on image processing. In contrast, this project uses a relational dataset consisting of tabular data with features and rows. A satisfactory result from these related works varied between support vector machine (SVM), XGBoost, and random forest (RF), frequently delivering good performance. However, this does not imply that other machine learning techniques or models are less accurate or ineffective. The effectiveness of a model depends on multiple factors rather than just the algorithm itself, which includes the nature of the dataset, the data's features, data size, and the specific problem being addressed, thus a holistic approach. Therefore, a holistic approach is essential in evaluating model performance, considering not only the choice of algorithm but also data preprocessing, feature engineering, and parameter tuning. Evaluation metrics are generally consistent across the studies and are derived from the confusion matrix. Similarly, this project also adopts the same evaluation metrics.

2.7 Tools and Technologies

i. Python

Python is one of the most popular programming languages and is considered the best programming language for Artificial Intelligence (AI) and data science projects (Toal, 2024). For this project, Python is used for the development of machine learning models and data pre-processing, using its libraries such as Scikit-learn, NumPy, Matplotlib, and Pandas.

a. Scikit-learn

Scikit-learn is a Python library that supports a wide range of machine learning algorithms, including both supervised and unsupervised (Nujan, 2018). This library is key to building the machine models in this project.

b. NumPy

NumPy is a Python library for scientific computing. It is used to process arrays and solve matrix problems (Nujan, 2018).

c. Matplotlib

Matplotlib is a Python library used to create data visualizations using charts such as bar charts, pie charts, or line charts (Nujan, 2018).

d. Pandas

Pandas is a Python library used for data analysis and data manipulation (Nujan, 2018). Pandas are used in the data pre-processing step to prepare and process the data.

ii. Kaggle

Kaggle is a community platform for discussing, publishing, and sharing work related to data science and artificial intelligence (Uslu, 2022). It offers rich and many open-source datasets for machine learning projects. In this project, all datasets related to Alzheimer's disease are obtained from Kaggle.

iii. Anaconda

Anaconda is a Python distribution with numerous features and pre-installed libraries that can simplify package deployment. It offers many data science libraries and packages, making it useful for developing machine learning models in this project.

iv. Jupyter

Jupyter is an open-source web application that is useful for this project. It allows code to be executed in chunks within sections called 'cells,' making the code more manageable and modular. A Jupyter extension is also readily available in VS Code.

v. Hugging Face

Hugging Face is an open-source platform that provides a wide range of pre-trained models and tools for natural language processing (NLP) and machine learning tasks. It offers the popular Transformers library, which simplifies the integration of advanced language models into projects. For this project, Hugging Face is valuable for accessing state-of-the-art models and accelerating model development and experimentation.

2.8 Summary

This chapter delves deep into the background of Alzheimer's disease, the data mining process, the machine learning algorithm, and the related works to the project. The commonly used data mining processes are thoroughly reviewed and explained. Furthermore, each machine learning approach is discussed in detail, particularly in supervised machine learning. Also, eight pieces of literature related to Alzheimer's disease prediction are reviewed and compared for analysis. This review focuses on the algorithms used, the datasets, the evaluation metrics, and their results, which will contribute to the progress of the project.

CHAPTER 3: METHODOLOGY

3.1 Introduction

This chapter discusses the methodologies used in this project. The methodology of Knowledge Discovery in Database (KDD) is used to evaluate the machine learning algorithms and datasets. In KDD, five key stages were implemented for this project: selection, pre-processing, transformation, data mining, and evaluation. Each of these stages is carried out and thoroughly explained in this chapter. In Figure 3.1 below, five key stages and their corresponding sub-stages are depicted in the graphic.

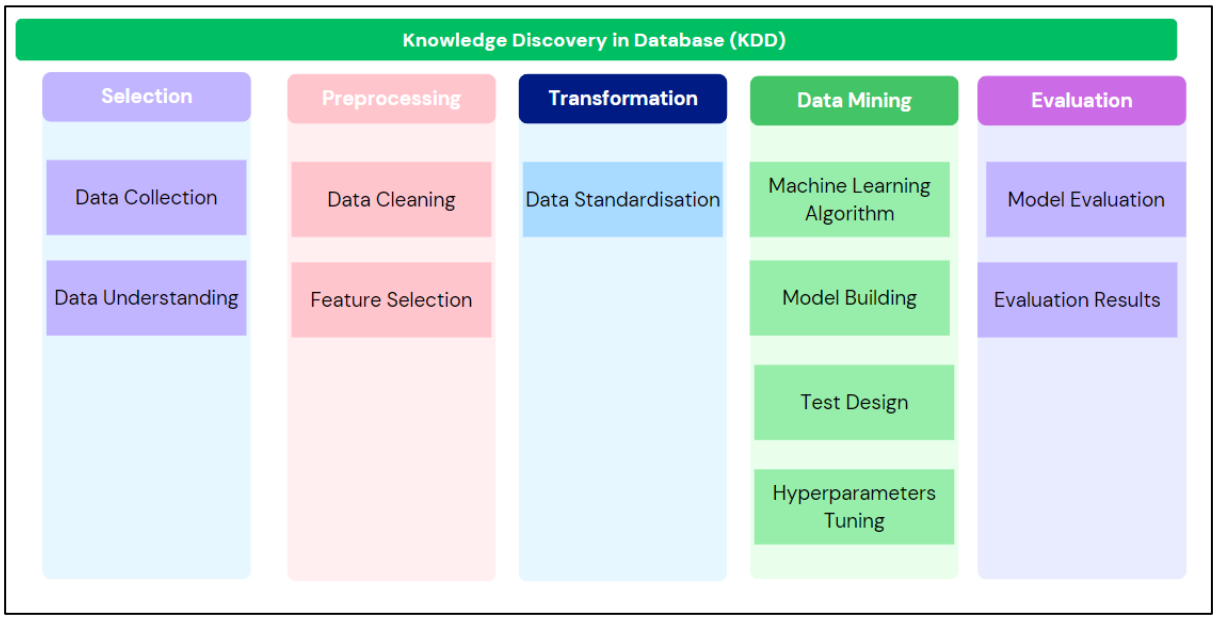


Figure 3.1 Knowledge Discovery in Database (KDD) Methodology Diagram

3.2 Knowledge Discovery in Databases (KDD)

Knowledge Discovery in Databases (KDD) is defined as the process of finding useful knowledge from data (Fayyad,1996). It involves processing, transforming, and refining data to uncover meaningful insights. The project follows five key phases: selection, pre-processing, transformation, data mining, and evaluation.

3.2.1 Selection

In this phase, datasets were sourced and collected from the Kaggle platform. The project uses the "Alzheimer's Disease Dataset," which contains extensive health records about patients' conditions and factors associated with Alzheimer's disease.

Next, the datasets are thoroughly examined and analysed for data understanding. There are a total of 2,149 patient records with an extensive 35 features of health information for the diagnosis of Alzheimer's disease. All these 2,149 records will be used in this study for the prediction model. The summary of data details is shown in Table 3.1 below.

Table 3.1 Features of the Dataset with Its Details

Features	Data Type	Description
Patient ID	Integer	A unique identifier is assigned to each patient.
Age	Integer	Patients' age in years old.
Gender	Integer	Gender of the patients, where 0 represents Male and 1 represents Female.
Ethnicity	Integer	The ethnicity of the patients is numbered as follows: 0: Caucasian 1: African American 2: Asian 3: Other

Education Level	Integer	The education level of the patients is numbered as follows: 0: None 1: High School 2: Bachelor's 3: Higher
BMI	Float	Body Mass Index of the patients
Smoking	Integer	Smoking status, where 0 indicates No and 1 indicates Yes
Alcohol Consumption	Float	Weekly alcohol consumption in units, ranging from 0 to 20
Physical Activity	Float	Weekly physical activity in hours, ranging from 0 to 10
Diet Quality	Float	Diet Quality Score ranges from 0 to 10, where it scales from unhealthy to healthy.
Sleep quality	Float	Sleep quality score, ranging from 0 to 10, where it scales from poor quality sleep to good quality sleep.

Family History Alzheimer’s	Integer	Family history of Alzheimer's Disease, where 0 indicates No and 1 indicates Yes.
Cardiovascular Disease	Integer	Cardiovascular disease, where 0 indicates No and 1 indicates Yes.
Diabetes	Integer	Diabetes, where 0 indicates No and 1 indicates Yes.
Depression	Integer	Depression, where 0 indicates No and 1 indicates Yes.
Head Injury	Integer	History of head injury, where 0 indicates No and 1 indicates Yes.
Hypertension	Integer	Hypertension, where 0 indicates No and 1 indicates Yes.
Systolic BP	Integer	Systolic blood pressure in mmHg measurement.
Diastolic BP	Integer	Diastolic blood pressure in mmHg measurement.
Cholesterol Total	Float	Total cholesterol in mg/dL measurement.
Cholesterol LDL	Float	Low-density lipoprotein cholesterol levels in mg/dL measurement.

Cholesterol HDL	Float	High-density lipoprotein cholesterol levels in mg/dL measurement.
Cholesterol Triglycerides	Float	Triglyceride levels in mg/dL measurement.
MMSE	Float	Mini-Mental State Examination score, ranging from 0 to 30. Lower scores indicate cognitive impairment.
Functional Assessment	Float	Functional assessment score, ranging from 0 to 10. Lower scores indicate greater impairment or lesser physical activity.
Memory Complaints	Integer	Memory complaints, where 0 indicates No and 1 indicates Yes.
Behavioural Problems	Integer	Behavioural problems, where 0 indicates No and 1 indicates Yes.
ADL	Float	Activities of Daily Living score, ranging from 0 to 10. Lower scores indicate greater impairment and a lack of ability to perform basic physical needs.

Confusion	Integer	Confusion, where 0 indicates No and 1 indicates Yes.
Disorientation	Integer	Disorientation, where 0 indicates No and 1 indicates Yes.
Personality Changes	Integer	Personality changes, where 0 indicates No and 1 indicates Yes.
Difficulty Completing Tasks	Integer	Difficulty completing tasks, where 0 indicates No and 1 indicates Yes.
Forgetfulness	Integer	Forgetfulness, where 0 indicates No and 1 indicates Yes.
Diagnosis	Integer	Diagnosis status for Alzheimer's Disease, where 0 indicates No and 1 indicates Yes.
Doctor In Charge	String	Confidential information about the doctor in charge.

3.2.2 Data Pre-processing

In this phase, data pre-processing techniques are used to improve data quality and enhance the performance of the prediction model. Pre-processing techniques such as data cleaning and feature selection are employed. The data are cleaned by removing irrelevant and incomplete data from the columns. For example, the features 'PatientID' and 'DoctorInCharge' are irrelevant to the prediction model and are therefore removed.

Next, feature selection is a technique used to select the best features for building a more optimized model (Aman, 2024). Feature selection helps reduce noise by eliminating insignificant features to the predictions, allowing the model to achieve optimal results with high accuracy and simplicity. This process also aids in reducing overfitting, leading to better performance. In feature selection, information gain techniques and correlation matrix are employed.

The information gain is calculated by using the entropy (self-information) of the dataset and then evaluating the reduction in uncertainty. In Equation 3.1 below, information gain can be represented by the difference between the entropy of information, S1, and the weighted average of the conditional entropies' pairs S1 and S2.

$$Gain(S_2) = H(S_1) - H(S_1 \mid S_2)$$

Equation 3.3 Formula of Information Gain (Kononenko & Kukar, 2007)

Next, the correlation matrix (correlation coefficient) is implemented. This technique compares the linear relationship between two or more variables and analyzes the correlation with the target. This technique uses the Pearson correlation coefficient formula to populate the correlation matrix with Pearson values, to create a graph-like plot in the matrix. A correlation coefficient close to +1 indicates a strong positive linear relationship, a coefficient close to -1 indicates a strong negative linear relationship, and a coefficient close to 0 suggests little to no linear relationship. This technique mapped the correlation, keeping highly correlated while removing low-correlation attributes. The formula of the Pearson's Correlation Coefficient is shown in Equation 3.2, where r, measures the strength and direction of a linear relationship between two variables x and y.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Equation 3.2 Formula of Pearson's Correlation Coefficient (Salomão, 2024)

3.2.3 Data Transformation

In the data transformation phase, the data is transformed into a more standardized and suitable format for the data mining phase. The data is transformed using feature scaling on features of different scales, such as the "BMI," "SystolicBP," and other features. They are standardized and scaled using the Z-score standardization. This is important because certain machine learning algorithms are sensitive to scale (A. Singh, 2024). For example, in support vector machines (SVM), different scales may also result in biases in locating the optimal hyperplane. Thus, data scaling can ensure features are contributed equally and avoid potential biases caused by differences in scales. The formula of Z-score standardization is shown in Equation 3.3.

$$z = \frac{x - \mu}{\sigma}$$

Equation 3.3 The formula of Z-Score Standardization (Pemasinghe, 2018)

Additionally, during this phase, feature engineering was performed to create composite features. A new feature, “Age Adjusted Memory,” was created by combining both features, “Age” and the “MMSE” score. This was done bsecause the ages of the patients in the dataset were all 60 years and above, making "Age" alone less significant. To address this, a new feature was created by taking the ratio of "Age" to the "MMSE" score. By forming this ratio, we aim to reintroduce the influence of age into the model, capturing how cognitive performance (as measured by MMSE) deteriorates relative to increasing age (Nagaratnam et al., 2022). By compositing it with the "MMSE" score, the new feature can improve the model's predictive performance.

3.2.4 Data Mining

In the data mining phase, several machine learning algorithms are employed for Alzheimer's disease prediction. The data mining techniques used are logistic regression, decision tree, support vector machine (SVM), CatBoost, XGBoost, random forest, and artificial neural network (ANN). Next, for the test design, this project uses a train-test split on the dataset, with 80% of the data used for training and 20% for testing, resulting in an 8:2 ratio. The models are built using the Python library scikit-learn, which contains all the machine learning algorithms, test design, and fold validation.

We also perform hyperparameter tuning to optimize model performance. Each parameter of the different algorithms is carefully tuned to ensure optimal settings, maximizing evaluation metrics. This is accomplished using Grid Search from Python's libraries.

3.2.5 Model Evaluation

In this phase, all models will be evaluated using metrics such as accuracy, precision, recall, F1 score, and sensitivity. The model with the best performance will then be selected for demonstration. To obtain these metrics, a confusion matrix is implemented to compute the metrics.

3.2.5.1 Confusion Matrix

A confusion matrix is a table used to evaluate the performance of machine learning classification algorithms. It helps visualize and summarize the performance of the algorithm (P. Singh et al., 2021). Figure 3.2 shows the diagram of a confusion matrix, where TP (True Positives) indicates the number of correctly predicted positive cases, FP (False Positives) refers to the number of negative cases incorrectly predicted as positive, FN (False Negatives) represents the positive cases incorrectly predicted as negative, and TN (True Negatives) indicates the number of correctly predicted negative cases.

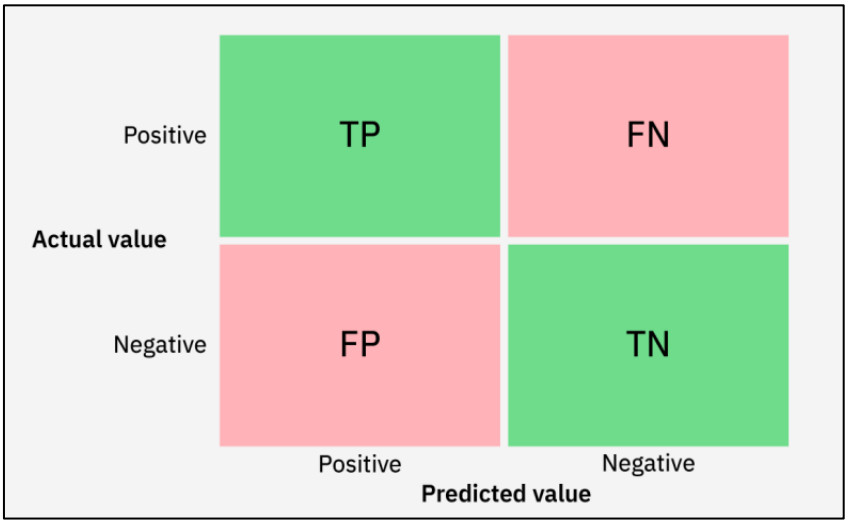


Figure 3.2 Diagram of Confusion Matrix (Murel, 2024)

Next, in Table 3.2 below, the general confusion matrix is adapted to fit the project's prediction context, where 'Positive' refers to predicted cases of Alzheimer's disease, and 'Negative' refers to predicted non-Alzheimer's cases.

Table 3.2 Confusion Matrix based on the Project's Prediction Definition

Predicted Value		
Positive Alzheimer's disease	Negative Alzheimer's disease	Actual Value
TP	FN	Positive Alzheimer's disease
FP	TN	Negative Alzheimer's disease

The confusion matrix in Table 3.2 shows the number of instances produced by the models and categorizes them into True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The definition is listed as follows:

i. True Positive (TP)

True positive (TP) represents the number of patient records that have been classified correctly as having Alzheimer's disease.

ii. True Negative (TN)

True negative (TN) represents the number of patient records that have been classified correctly as not having Alzheimer's disease.

iii. False Positive (FP)

False positive (FP) represents the number of patient records misclassified with the disease, but instead, they do not have the disease. FP is also known as a Type I error. (P. Singh et al., 2021).

iv. False Negative (FN)

False negative (FN) represents the number of patient records misclassified as not having Alzheimer's disease but are suffering from the disease. FN is also known as a Type II error (P. Singh et al., 2021).

Next, the performance and evaluation metrics are calculated based on the categories of the confusion matrix. These evaluation metrics are listed and defined in Table 3.3 below.

Table 3.3 Definition of The Evaluation Metrics and Its Formula

Evaluation Metrics	Description	Formula
Accuracy	Accuracy measures the proportion of total correct instances. It also measures how well a classification model predicts a condition (Valero-Carreras et al., 2023).	$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{N}}$ Where, N = TP + TN + FP + FN
Precision	Precision is used to evaluate the true correct instances from the relevant retrieved instances. It	$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$

	calculates the proportion between truly positive predictions and the set of all the real positive values (Valero-Carreras et al., 2023).	
Recall	Recalls measure all relevant instances where it's truly positive. It is calculated by the proportion between the true positive and the sum of the true positive with the negative.	$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$
F1-Score	F1- score measures the harmonic mean of precision and recalls finding a balance between both. Thus, combining both precision and recall.	$\text{F1} - \text{Score} = \frac{2.\text{Precision}.\text{Recall}}{\text{Precision} + \text{Recall}}$
Specificity	Specificity is used to calculate the proportions between negative instances and the set of all negative instances. It helps predict negative instances (Valero-Carreras et al., 2023).	$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$

3.3 Summary

In this chapter, the methodology for predicting Alzheimer's disease is described in detail. The approach follows the Knowledge Discovery in Databases (KDD) process, covering each phase from selection, pre-processing, transformation, data mining, and evaluation. Various pre-processing techniques like feature selection, data cleaning, and feasture scaling show the importance of preparing the data for effective data mining. In the data mining phase, we introduce the machine learning algorithms used in this project, including logistic regression, decision trees, support vector machines (SVM), CatBoost, XGBoost, random forests, and artificial neural networks (ANN). Lastly, we also introduce the method of evaluating the machine learning models, like the confusion matrix, together with the performance metrics: accuracy, precision, recall, F1-score, and specificity.

CHAPTER 4: DATA MINING AND IMPLEMENTATION

4.1 Introduction

This chapter discusses the implementation and a comparative analysis of several machine learning models in predicting the risk of Alzheimer's disease. The implementation was carried out using Visual Studio Code (VSCode) with the Jupyter extension and all the necessary libraries.

This chapter will also cover the details of implementing the Knowledge Discovery in Databases (KDD) methodology, including data selection, preprocessing, transformation, data mining, and evaluation. The seven machine learning algorithms used are logistic regression, decision tree, support vector machine (SVM), CatBoost, XGBoost, random forest, and artificial neural network (ANN). The model evaluation will be evaluated using metrics derived from the confusion matrix, such as accuracy, precision, recall, F1 score, and sensitivity.

Additionally, based on the evaluation, the best-performing machine learning model will be selected and incorporated into a web application. An application was developed using Gradio to create a web-based user interface (UI), allowing users to interact with the model, which is then deployed on Hugging Face.

4.2 Knowledge Discovery in Databases (KDD)

Knowledge Discovery in Databases (KDD) follows five key phases that are selection, pre-processing, transformation, data mining, and evaluation.

4.2.1 Selection

Datasets were collected from Kaggle, which contains extensive health records about patients' conditions and factors associated with Alzheimer's disease. This dataset is valuable for analyzing and understanding the factors associated with the disease.

The datasets are thoroughly examined and analysed for data understanding. There are a total of 2,149 patient records with an extensive 35 features of health information for the diagnosis of Alzheimer's disease. All these 2,149 records will be used in this study for the prediction model.

In our analysis, we found that most of the data and features examined show a moderately uniform distribution. For example, in Figure 4.1 below, features like the ADL (Activities of Daily Living) display a relatively even spread across the value range, with a slight tendency toward the higher or lower extremes. This suggests a balanced representation within the dataset, accompanied by mild skewness in certain features.

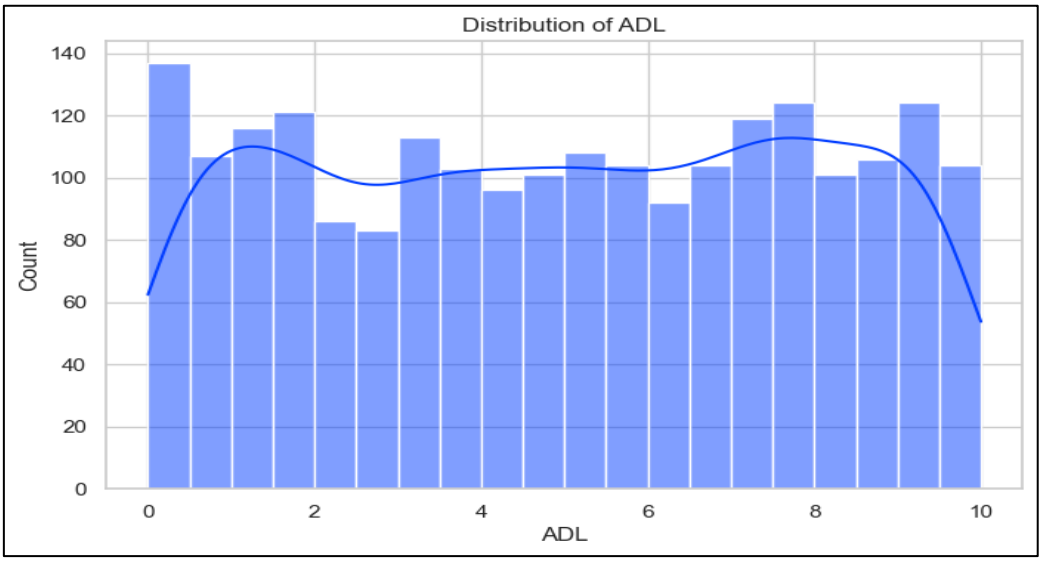


Figure 4.1 The Line Histogram Distribution of ADL Feature

However, an interesting observation was made in the MMSE (Mini Mental State Examination) feature, a bimodal distribution is observed. Which is indicated in Figure 4.2 by two distinct humps. This analysis suggests that the dataset contains two different groups that may be distinguishable from one another in the MMSE feature. For example, the diagnosis difference between individuals with normal cognitive function and those with Alzheimer's disease.

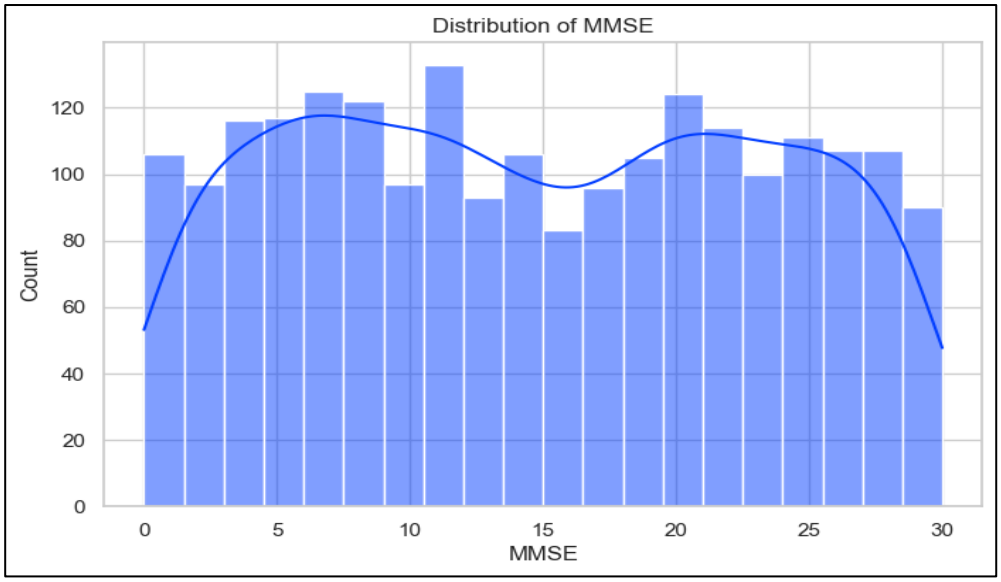


Figure 4.2 The Line Histogram Distribution of MMSE Feature

4.2.2 Pre-Processing

In this phase, data cleaning is performed to remove irrelevant features such as the 'PatientID' and 'DoctorInCharge' features. 'PatientID' feature is just a unique identifier, and 'DoctorInCharge' is the doctor assigned to the patient. In Figure 4.3 below, these irrelevant features are removed. This results in 33 features from the original 35 features.

```
# Remove 'PatientID' and 'DoctorInCharge' which are unrelated features to the prediction
df.drop(['PatientID', 'DoctorInCharge'], axis=1, inplace=True)
```

Figure 4.3 Code Snippet for Removing Irrelevant Features

Next, a feature selection technique is performed to identify the most relevant features for model prediction, helping to reduce the number of features and complexity. Two scoring methods are used for feature selection: information gain and Pearson correlation (correlation matrix).

For information gain, it is scored using the information gain ratio. A higher gain ratio indicates a more significant and relevant feature. A threshold parameter of 0.05 was set, meaning that features with a gain ratio lower than 0.05 will be removed, while those above the threshold will be retained for model training.

For the correlation matrix, it is scored using the Pearson correlation coefficient. A correlation coefficient close to +1 indicates a strong positive linear relationship, a coefficient close to -1 indicates a strong negative linear relationship, and a coefficient close to 0 suggests little to no linear relationship. A threshold parameter of 0.03 was set, meaning that features with a coefficient between negative 0.03 and positive 0.03 (-0.03< X<0.03) will be removed. Figure 4.4 shows the feature selection parameters with a gain ratio lower than 0.05 and a coefficient between negative 0.03 and positive 0.03.

```
# Feature selection parameters
selected_features = comparison_df[
    (comparison_df['Information_Gain'] > 0.05) |
    ((comparison_df['Correlation'] <= -0.03) | (comparison_df['Correlation'] >= 0.03))
]
```

Figure 4.4 Code Snippet of The Feature Selection Parameters

This feature selection process results in the top 15 significant features selected, while 18 less significant features are removed from the initial 33 features. The 18 lesser significant features removed are 'Alcohol Consumption', 'BMI', 'Cholesterol Total', 'Cholesterol Triglycerides', 'Confusion', 'Depression', 'Diastolic BP', 'Diet Quality', 'Difficulty Completing Tasks', 'Disorientation', 'Ethnicity', 'Forgetfulness', 'Gender', 'Head Injury', 'Personality Changes', 'Physical Activity', 'Smoking' and 'Systolic BP'.

To serve as a control for feature selection, this project also retained all features in a separate pre-processed dataset named “preprocessed_data_1.csv,” While features selected were saved as “preprocessed_data_2.csv”. Both datasets will be trained and tested separately, with “Modelling_1.ipynb” for the data with all features and “Modelling_2.ipynb” for the selected features dataset.

Figures 4.5 and 4.6 below show the Information Gain Ratio and Pearson Correlation, respectively. All the feature values have been calculated and obtained as part of the analysis.

The Information Gain Ratio from Figure 4.5, analysis revealed that certain features are particularly significant in predicting the risk of Alzheimer’s disease, where higher values indicate greater importance. For example, "Functional Assessment" scored 0.095, "ADL" scored 0.077, "MMSE" scored 0.066, and "Memory Complaint" scored 0.063. These values suggest that these features have a significant gain ratio. In contrast, features with a value of zero indicate little or no relevance in feature selection.

Consistent findings are also shown in Figure 4.6 through the Pearson Correlation Coefficient analysis. In this analysis, a positive coefficient suggests a positive linear relationship, a coefficient close to negative one reflects a strong negative linear relationship, and a coefficient near zero indicates little or no linear relationship. The results show a similar

trend to the Information Gain Ratio analysis, where "Functional Assessment" has a correlation of negative 0.364, "ADL" has negative 0.332, "MMSE" has negative 0.237, and "Memory Complaint" has positive 0.307.

These results indicate that the most significant features identified by both the Information Gain Ratio and the Pearson Correlation Coefficient also show notable correlations with the target outcome. Therefore, a combined feature selection approach was applied, using both methods to select the top 15 important features for Alzheimer's disease prediction. A threshold of 0.05 was set for the Information Gain Ratio, meaning only features with a value higher than 0.05 were selected. For the Pearson Correlation Coefficient, a threshold of 0.03 was applied, where features with a correlation coefficient between negative 0.03 and positive 0.03 were excluded.

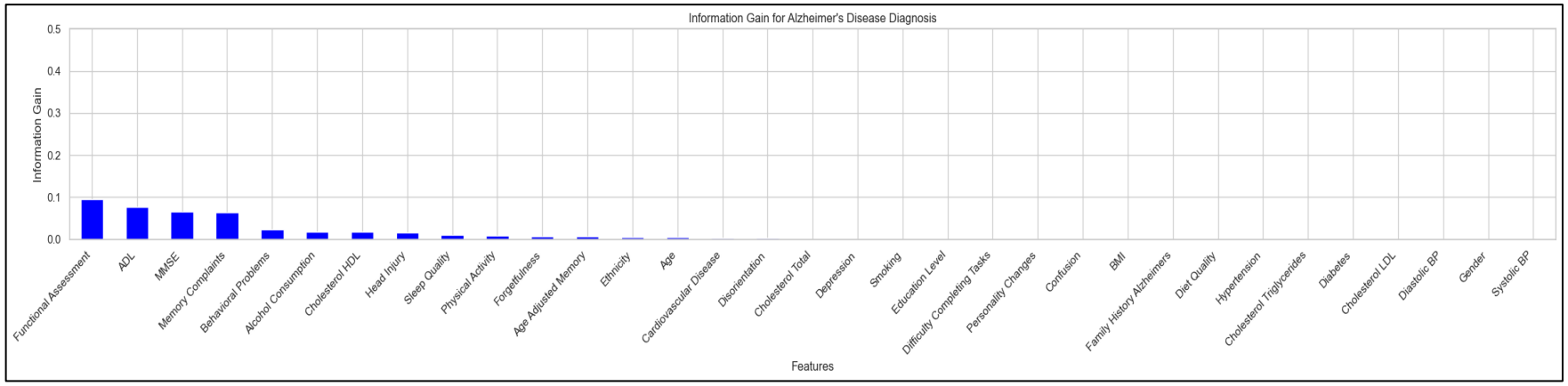


Figure 4.5 Information Gain Ratio Across the Features

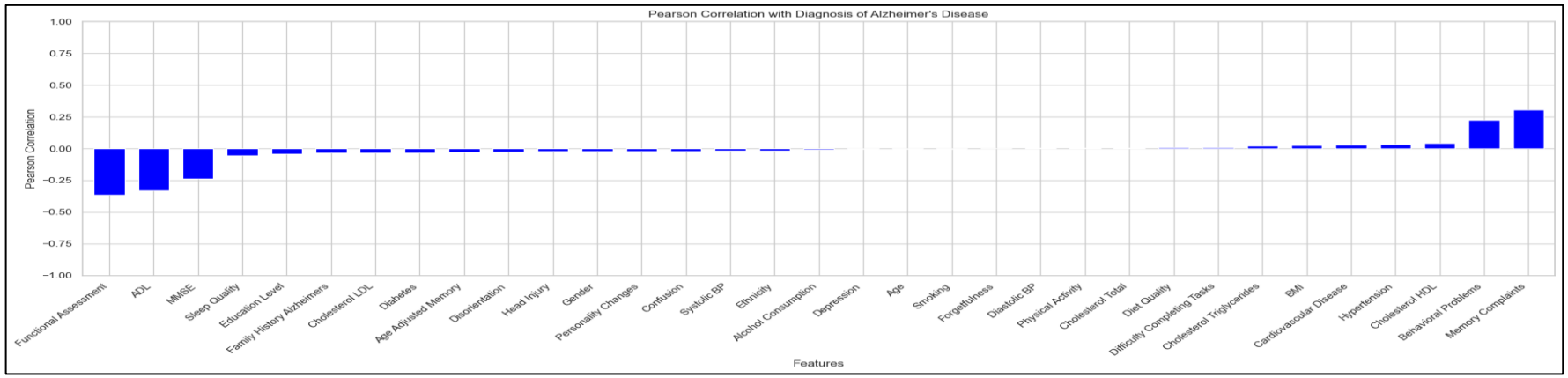


Figure 4.6 Pearson Correlation Coefficient Across the features

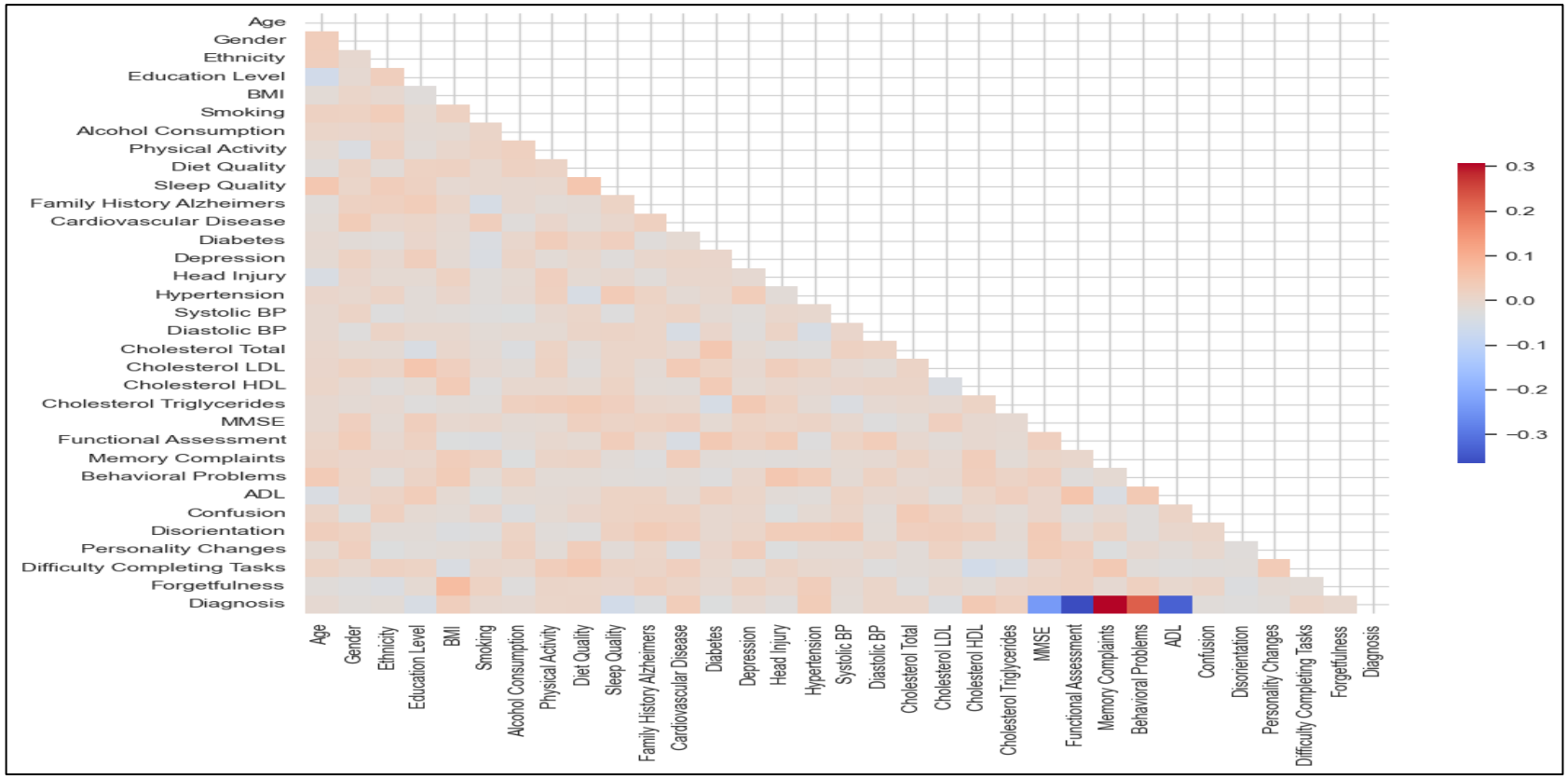


Figure 4.7 Correlation Matrix Across the Features

Figure 4.7 shows a triangular correlation matrix of all the features. The coefficient values are visualized using a heatmap, where redder shades indicate a stronger positive correlation with the target variable, and bluer shades indicate a stronger negative correlation. The figure shows that "Functional Assessment" displays the darkest blue shade, followed by "ADL" and then "MMSE," reflecting their strong negative correlations. In contrast, "Memory Complaint" appears with the darkest red shade, followed by a more orange shade observed for the "Behavioral Problems" feature, indicating their stronger positive correlations with the target variable.

4.2.3 Data Transformation

In the data transformation phase, data standardization (feature scaling) and data engineering are implemented. As illustrated in Figure 4.8 snippet, continuous features like 'BMI' and 'Age', which exist on different scales are standardized using Z-score normalization, unlike categorical features that do not require such scaling.

```
# Feature Scaling of non-categorical values
columns = ['Age', 'BMI', 'Alcohol Consumption', 'Physical

#Z score standardization
standard_scaler = StandardScaler()
df[columns] = standard_scaler.fit_transform(df[columns])
```

Figure 4.8 Code Snippet for Feature Scaling

Next, feature engineering is implemented to create a composite feature by combining both features. For example, in Figure 4.9, a snippet shows the features 'Age' and 'MMSE' score. A new feature, named 'Age-Adjusted Memory,' was created by taking the ratio of 'Age' to the 'MMSE' score to capture the relative significance of age in cognitive performance. This derived feature was then added to the dataset.

```
# Composite Feature (Feature Engineering)- due to samp
df['Age Adjusted Memory'] = df['MMSE'] / df['Age']
```

Figure 4.9 Code Snippet Feature Engineered of 'Age Adjusted Memory'

4.2.4 Data Mining

In the data mining phase, the test design used is the hold-out method through data splitting. For Table 4.1, a total of 2,149 records were used for data mining, with the dataset split into 80% for the training set and 20% for the test set.

Table 4.1 Data distribution of the Alzheimer's disease dataset

Class Label	**Train Dataset (80%)**	**Test Dataset (20%)**	**Total**
Alzheimer's Disease	1119	280	**1399**
No Alzheimer's Disease	600	150	**750**
Total	**1719**	**430**	**2149**

From the dataset, it was observed that there is a slight imbalance in class distribution. While this imbalance is present, it does not warrant the need for oversampling techniques, as the difference is not extreme. The " No Alzheimer's Disease" class represents 64.6%, while the "Alzheimer's Disease" class accounts for 35.4%, resulting in a 65/35 split.

Another reason for not addressing the imbalance is the concern that oversampling methods such as SMOTE (Synthetic Minority Oversampling Technique) artificially increase the number of minority-class instances and can potentially introduce synthetic patterns that do not accurately reflect the actual dataset. As shown in Figure 4.10 below, this can increase the risk of overfitting the classifier on false data where the class label is falsely classified as members of the minority class (false positive), leading to inaccurate prediction and lower accuracy (Alkhawaldeh et al., 2023).

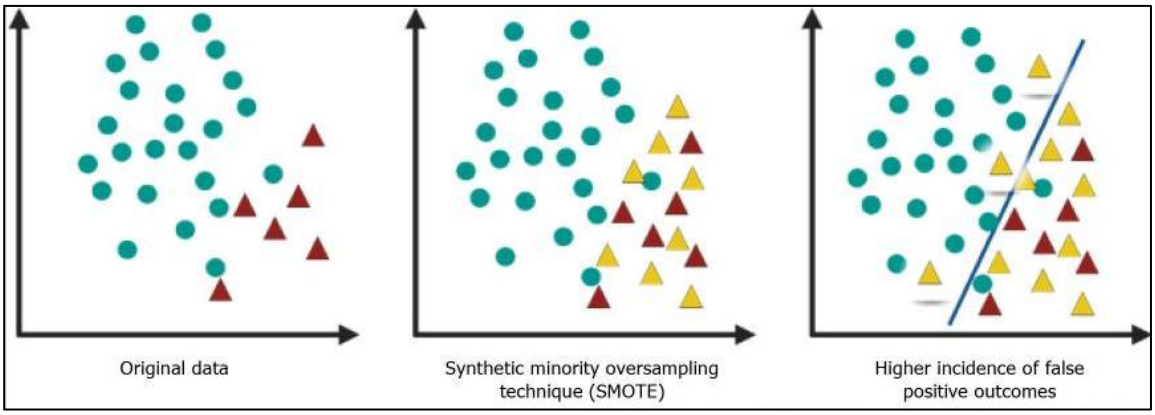


Figure 4.10 The Synthetic Minority Sample After SMOTE Application (Alkhawaldeh et al., 2023)

Next, for hyperparameter tuning, this project uses Grid Search. Grid Search is a scikit-learn library that systematically evaluates all possible combinations of specified hyperparameter values to identify the optimal configuration for model performance. This helps select the best model parameters for each of the algorithms based on the scoring metric. The machine learning algorithms applied are logistic regression, decision tree, support vector machine (SVM), CatBoost, XGBoost, random forest, and artificial neural network (ANN). Table 4.2 below outlines the range of hyperparameter values used to evaluate and select the best-performing configuration for each algorithm.

Table 4.2 Hyperparameter Configuration Settings of The Algorithms

Machine Learning Algorithm	Hyperparameter	Value Specified for Grid Search
Logistic Regression	C	[0.1, 1, 10]
Decision Tree	max_depth	[3, 5, 7, 12, None]
Support Vector Machine	C	[0.1, 1, 10]
	gamma	[0.1, 1, 'scale', 'auto']
CatBoost	iterations	[50, 100, 200]
	learning_rate	[0.01, 0.1, 1]
XGBoost	n_estimators	[50, 100, 200]
	learning_rate	[0.01, 0.1, 1]
	max_depth	[3, 5, 7]
Random Forest	n_estimators	[50, 100, 200]
	max_depth	[3, 5, 7, 12, None]
Artificial Neural Network	model__optimizer	['adam', 'rmsprop']
	model__activation	['relu', 'tanh']
	model__hidden_units	[32,64,128]
	model__hidden_units1	[16,32,64]
	epochs	[50]
	batch_size	[16,32]

4.2.5 Model Evaluation

In this phase, the performance of all the selected machine learning models will be tested and evaluated with the two datasets. The models evaluated are logistic regression, decision tree, support vector machine (SVM), CatBoost, XGBoost, random forest, and artificial neural network (ANN). The performance of each model will be evaluated using appropriate evaluation metrics such as accuracy, precision, recall, F1 score, and sensitivity. The model results and evaluation will be discussed in the next chapter.

4.3 Application Deployment

An application was developed to test the trained model for predicting the risk of Alzheimer's disease in individuals. The application was built using Python, with the Gradio library and other supporting libraries. The XGBoost model with feature selection dataset was selected for prediction, as it shows the best performance metrics among the models evaluated in the next chapter. The trained XGBoost model is subsequently loaded into the application, as illustrated in Figure 4.11 snippet.

```
# Load trained model
model = joblib.load("xgboost_new_model.pkl")
```

Figure 4.11 Code Snippet of XGBoost Model Loading

Next in the deployment stage, the developed application is published and hosted in Hugging Face. Hugging Face is a data science platform that allows users to deploy and test machine learning models. The application hosted on Hugging Face functions as a fully functional web application. Users can interact with the predictive model by inputting the details to receive the prediction results. The Hugging Face application interface is shown in Figures 4.12 and 4.13 below, where users input their health-related information and click the 'Predict Alzheimer's Disease' button to obtain the results.

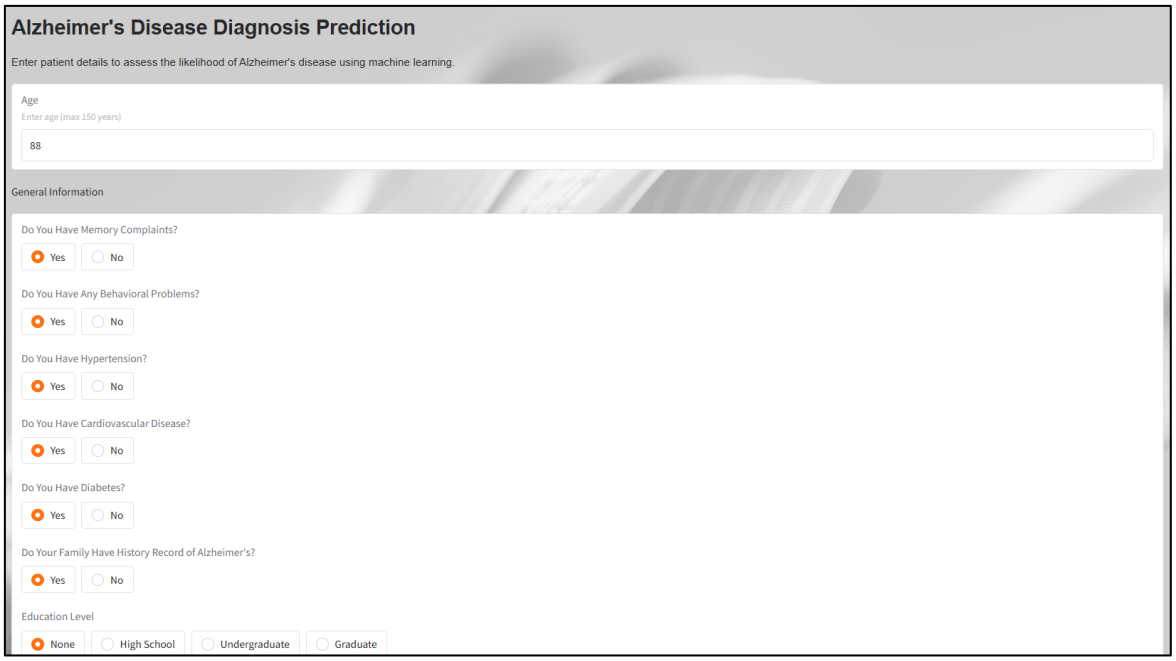


Figure 4.12 Hugging Face Application Interface 1

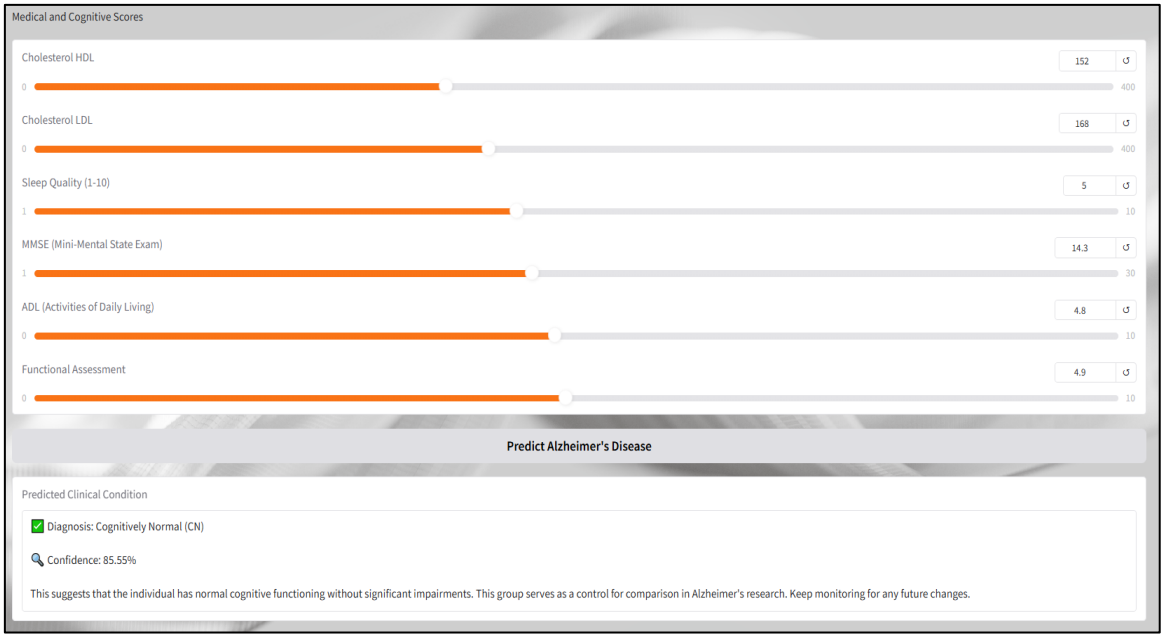


Figure 4.13 Hugging Face Application Interface 2

4.4 Summary

This chapter provides an overview implementation process of Knowledge Discovery in Databases (KDD), and the data mining phase is discussed in detail. It covers the data selection and preprocessing to transformation, mining, and evaluation. Preprocessing techniques such as correlation matrix analysis and information gain were applied to identify relevant features. Data transformation involved data engineering and scaling to prepare the dataset for modelling. And finally, the data mining process and evaluation for the seven models. Next, this chapter also discusses the application deployment. The prediction application was developed using Python and deployed on Hugging Face for users to interact with the trained model.

CHAPTER 5: RESULTS AND ANALYSIS

5.1 Introduction

This chapter will focus on discussing and analysing the results obtained from the seven machine learning models used in the prediction of Alzheimer's disease datasets. The seven machine learning algorithms used are logistic regression, decision tree, support vector machine (SVM), CatBoost, XGBoost, random forest, and artificial neural network (ANN). The analysis will be based on model evaluation using metrics derived from the confusion matrix, including accuracy, precision, recall, F1 score, and sensitivity.

Additionally, a comparative analysis will be conducted to determine the best-performing machine learning models. Key findings and patterns observed during the evaluation will also be examined to provide deeper insights into each model's performance.

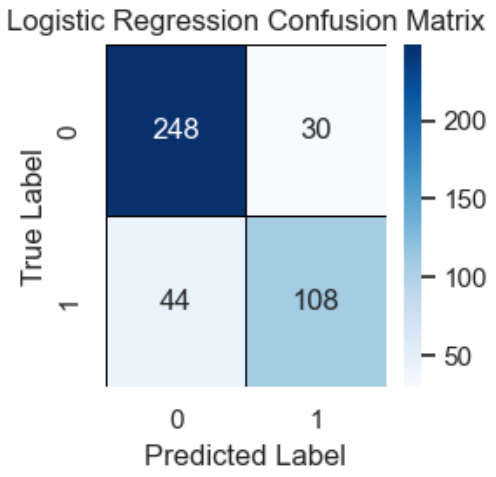
5.1 Model Evaluation

In model evaluation, all selected machine learning models will be tested and evaluated to obtain the evaluation metrics results. The performance of each model will be evaluated using appropriate evaluation metrics such as accuracy, precision, recall, F1 score, and sensitivity. Then, these results will be recorded and analysed to compare the effectiveness of each algorithm. Based on the comparison, the best-performing model will be selected for deployment to Hugging Face, where it will be integrated into the application. The machine learning algorithms are logistic regression, decision tree, support vector machine (SVM), CatBoost, XGBoost, random forest, and artificial neural network (ANN).

I. Result for Logistic Regression

Table 5.1 shows the performance of the Logistic Regression model, along with the confusion matrices for Dataset 1 (with all features) and Dataset 2 (with selected features). The model's performance is then evaluated using the evaluation metrics.

Table 5.1 Performance of Logistic Regression Model

All Features (33) – Dataset 1	**Feature Selection (15) – Dataset 2**
	

Evaluation Metrics	**All Features (33) – Dataset 1**	**Feature Selection (15) – Dataset 2**
Accuracy	82.8%	83.7%
Precision	81.6%	82.6%
Recall	80.1%	81.4%
F1-Score	80.8%	81.9%
Specificity	89.2%	89.5%

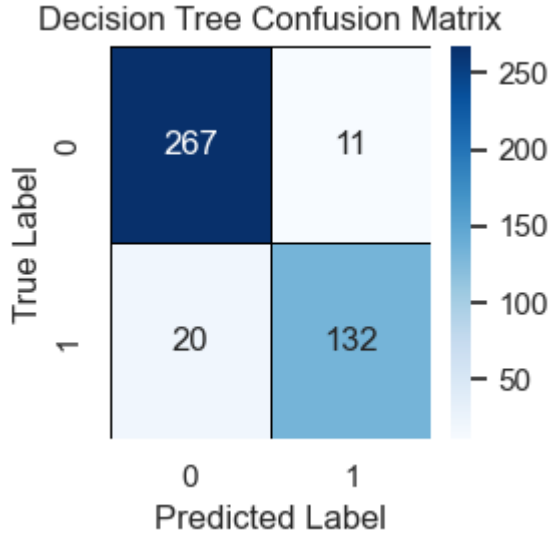
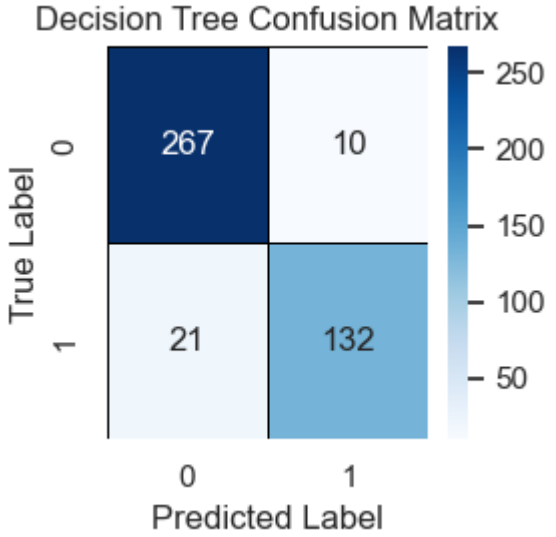
Table 5.1 above displays the confusion matrix with counts of true positives, true negatives, false positives, and false negatives. It also summarizes the performance metrics of the Logistic Regression algorithm, including accuracy, precision, recall, F1-score, and specificity for the two datasets used.

Based on the results, Dataset 2 with 15 features selected shows a higher accuracy (83.7%), a higher precision (82.6%), a higher recall (81.4%), a higher F1-score (81.9%) and a higher specificity (89.5%). Hence, a better performing Logistic Regression model.

II. Result for Decision Tree

Table 5.2 shows the performance of the Decision Tree model, along with the confusion matrices for Dataset 1 (with all features) and Dataset 2 (with selected features). The model's performance is then evaluated using the evaluation metrics.

Table 5.2 Performance of Decision Tree Model

All Features (33) – Dataset 1	**Feature Selection (15) – Dataset 2**
	

Evaluation Metrics	All Features (33) – Dataset 1	Feature Selection (15) – Dataset 2
Accuracy	92.8%	92.8%
Precision	92.7%	92.8%
Recall	91.4%	91.3%
F1-Score	92.0%	92.0%
Specificity	96.0%	96.4%

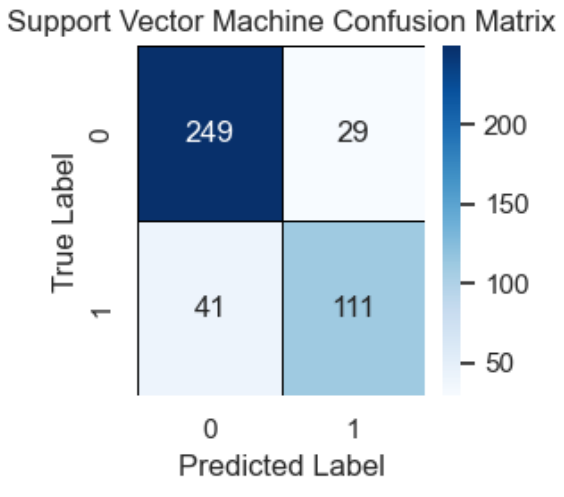
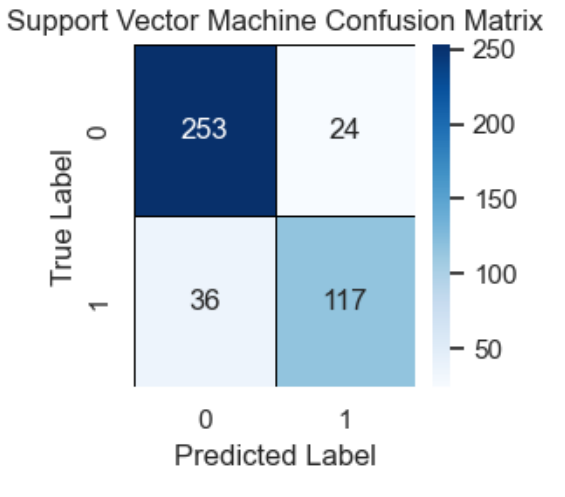
Table 5.2 above displays the matrices with counts of true positives, true negatives, false positives, and false negatives. It also summarizes the performance metrics of the Decision Tree algorithm, including accuracy, precision, recall, F1-score, and specificity for the two datasets used.

Based on the results, Dataset 2 with 15 features selected shows a same accuracy (92.8%) as Dataset 1, a higher precision (92.8%), a slightly lower recall (91.3%), a same F1-score (92.0%) and a higher specificity (96.4%). Hence, a better performing Decision Tree model after comparing the overall metrics.

III. Result for Support Vector Machine (SVM)

Table 5.3 shows the performance of the Support Vector Machine (SVM) model, along with the confusion matrices for Dataset 1 (with all features) and Dataset 2 (with selected features). The model's performance is then evaluated using the evaluation metrics.

Table 5.3 Performance of Support Vector Machine (SVM) Model

All Features (33) – Dataset 1	**Feature Selection (15) – Dataset 2**
	

Evaluation Metrics	All Features (33) – Dataset 1	Feature Selection (15) – Dataset 2
Accuracy	83.7%	86.0%
Precision	82.6%	85.3%
Recall	81.3%	83.9%
F1-Score	81.9%	84.5%
Specificity	89.6%	91.3%

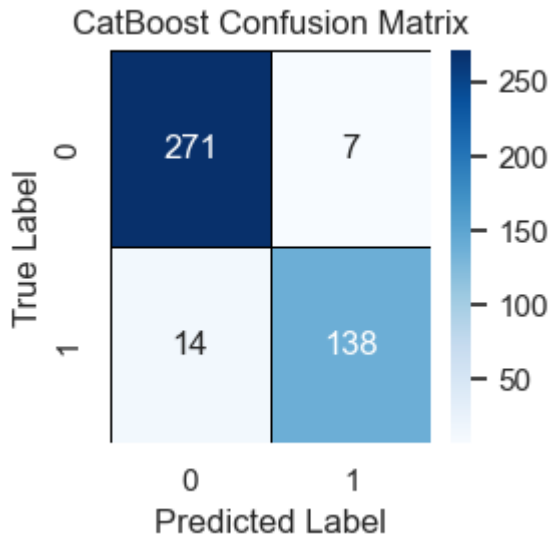
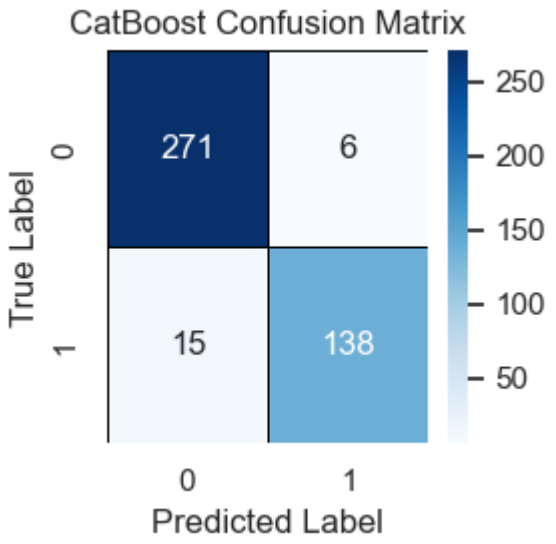
Table 5.3 above displays the matrices with counts of true positives, true negatives, false positives, and false negatives. It also summarizes the performance metrics of support vector machine algorithm, including accuracy, precision, recall, F1-score, and specificity for the two datasets used.

Based on the results, Dataset 2 with 15 features selected shows a higher accuracy (86.0%), a higher precision (85.3%), a higher recall (83.9%), a higher F1-score (84.5%) and a higher specificity (91.3%). Hence, a better performing support vector machine model.

IV. Result for CatBoost

Table 5.4 shows the performance of the CatBoost model, along with the confusion matrices for Dataset 1 (with all features) and Dataset 2 (with selected features). The model's performance is then evaluated using the evaluation metrics.

Table 5.4 Performance of CatBoost Model

All Features (33) – Dataset 1	**Feature Selection (15) – Dataset 2**
	

Evaluation Metrics	All Features (33) – Dataset 1	Feature Selection (15) – Dataset 2
Accuracy	95.1%	95.1%
Precision	95.1%	95.3%
Recall	94.1%	94.0%
F1-Score	94.6%	94.6%
Specificity	97.5%	97.8%

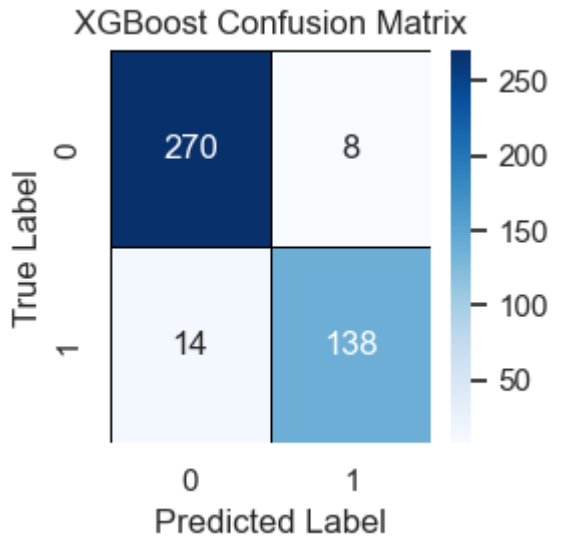
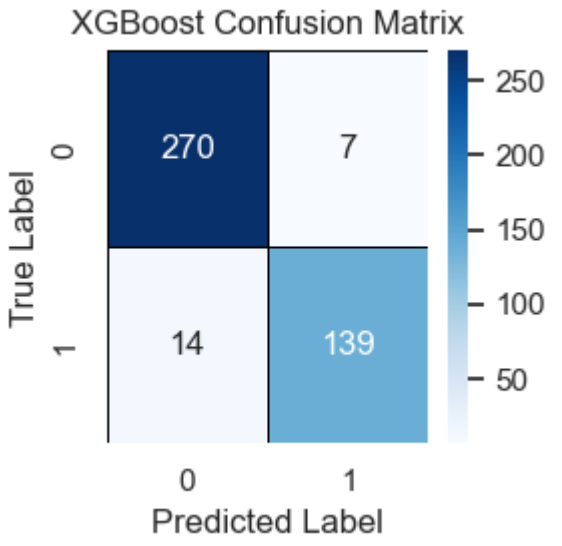
Table 5.4 above displays the matrices with counts of true positives, true negatives, false positives, and false negatives. It also summarizes the performance metrics of the CatBoost algorithm, including accuracy, precision, recall, F1-score, and specificity for the two datasets used.

Based on the results, Dataset 2 with 15 features selected shows the same accuracy (95.1%) as Dataset 1, a higher precision (95.3%), a slightly lower recall (94.0%), a same F1-score (94.6%) and a higher specificity (97.8%). Hence, a better performing CatBoost model after comparing the overall metrics.

V. Result for XGBoost

Table 5.5 shows the performance of the XGBoost model, along with the confusion matrices for Dataset 1 (with all features) and Dataset 2 (with selected features). The model's performance is then evaluated using the evaluation metrics.

Table 5.5 Performance of XGBoost Model

All Features (33) – Dataset 1	**Feature Selection (15) – Dataset 2**
	

Evaluation Metrics	**All Features (33) – Dataset 1**	**Feature Selection (15) – Dataset 2**
Accuracy	94.9%	95.3%
Precision	94.8%	95.5%
Recall	94.0%	94.3%
F1-Score	94.4%	94.9%
Specificity	97.1%	97.5%

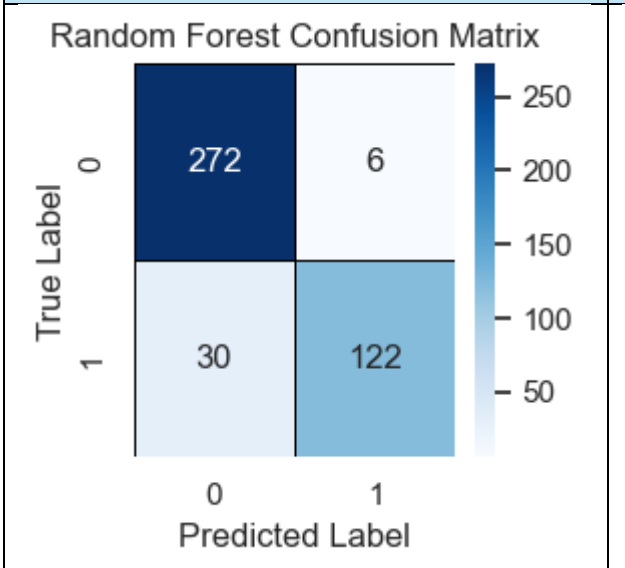
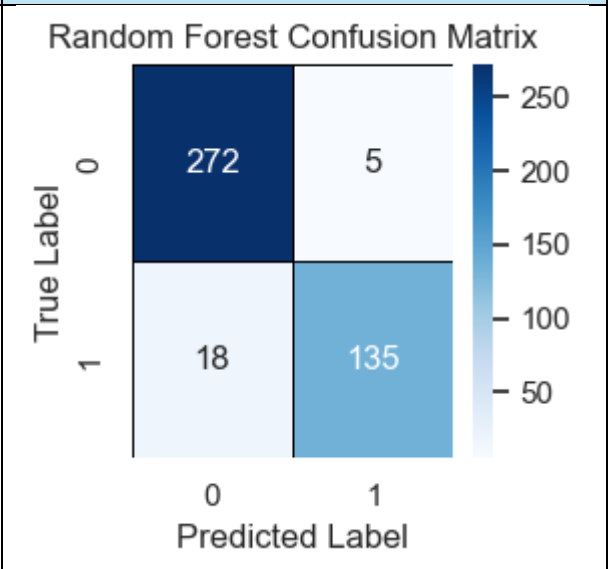
Table 5.5 above displays the matrices with counts of true positives, true negatives, false positives, and false negatives. It also summarizes the performance metrics of the XGBoost algorithm, including accuracy, precision, recall, F1-score, and specificity for the two datasets used.

Based on the results, Dataset 2 with 15 features selected shows a higher accuracy (95.3%), a higher precision (95.5%), a higher recall (94.3%), a higher F1-score (94.9%) and a higher specificity (97.5%). Hence, a better performing XGBoost model.

VI. Result for Random Forest

Table 5.6 shows the performance of the Random Forest model, along with the confusion matrices for Dataset 1 (with all features) and Dataset 2 (with selected features). The model's performance is then evaluated using the evaluation metrics.

Table 5.6 Performance of Random Forest Model

All Features (33) – Dataset 1	**Feature Selection (15) – Dataset 2**
	

Evaluation Metrics	All Features (33) – Dataset 1	Feature Selection (15) – Dataset 2
Accuracy	91.6%	94.4%
Precision	92.7%	94.9%
Recall	89.1%	92.9%
F1-Score	90.5%	93.8%
Specificity	97.8%	98.2%

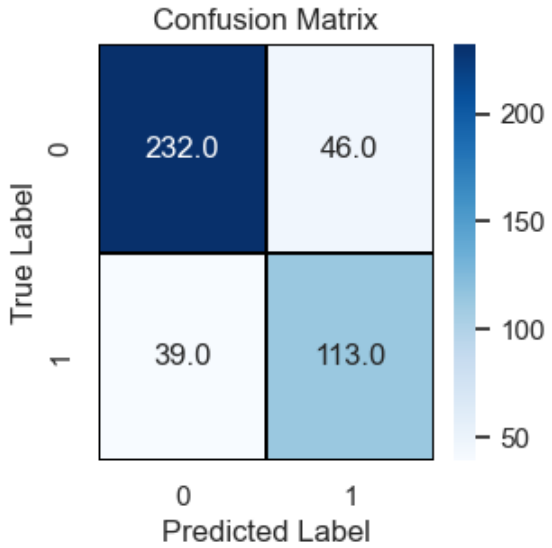
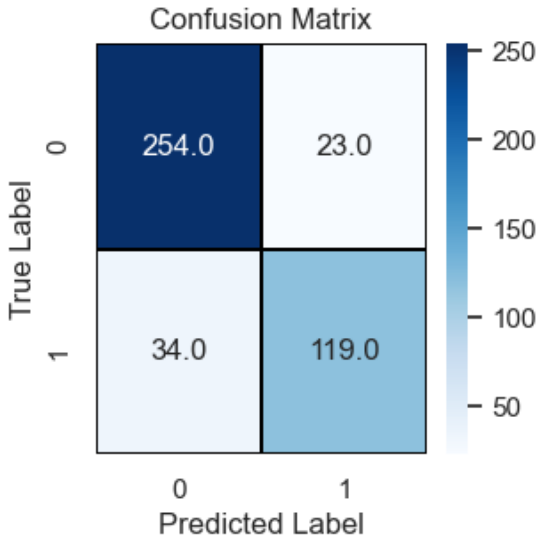
Table 5.6 above displays the matrices with counts of true positives, true negatives, false positives, and false negatives. It also summarizes the performance metrics of the random forest algorithm, including accuracy, precision, recall, F1-score, and specificity for the two datasets used.

Based on the results, Dataset 2 with 15 features selected shows a higher accuracy (94.4%), a higher precision (94.9%), a higher recall (92.9%), a higher F1-score (93.8%) and a higher specificity (98.2%). Hence, a better performing random forest model.

VII. Result for Artificial Neural Network (ANN)

Table 5.7 shows the performance of the Artificial Neural Network (ANN) model, along with the confusion matrices for Dataset 1 (with all features) and Dataset 2 (with selected features). The model's performance is then evaluated using the evaluation metrics.

Table 5.7 Performance of Artificial Neural Network (ANN) Model

All Features (33) – Dataset 1	**Feature Selection (15) – Dataset 2**
	

Evaluation Metrics	**All Features (33) – Dataset 1**	**Feature Selection (15) – Dataset 2**
Accuracy	80.2%	86.7%
Precision	78.3%	86.0%
Recall	78.9%	84.7%
F1-Score	78.6%	85.3%
Specificity	83.5%	96.2%

Table 5.7 above displays the matrices with counts of true positives, true negatives, false positives, and false negatives. It also summarizes the performance metrics of the artificial neural network algorithm, including accuracy, precision, recall, F1-score, and specificity for the two datasets used.

Based on the results, Dataset 2 with 15 features selected shows a higher accuracy (86.7%), a higher precision (86.0%), a higher recall (84.7%), a higher F1-score (85.3%) and a higher specificity (96.2%). Hence, a better-performing artificial neural network model.

5.2 Comparative Analysis

Previously, the evaluation conducted on machine learning algorithms like logistic regression, decision tree, support vector machine (SVM), CatBoost, XGBoost, random forest, and artificial neural network (ANN) shows that Dataset 2 with feature selection performed better overall. Thus, Dataset 2 is used for further analysis. In Table 5.7, machine learning algorithms will now be compared against each other to obtain comparative results and determine the best-performing model.

Table 5.8 Comparison of Evaluation Results Among the Models

	Accuracy (%)	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**Specificity (%)**
Logistic Regression	83.7	82.6	81.4	81.9	89.5
Decision Tree	92.8	92.8	91.3	92.0	96.4
Support Vector Machine	86.0	85.3	83.9	84.5	91.3
CatBoost	95.1	95.3	94.0	94.6	97.8
XGBoost	95.3	95.5	94.3	94.9	97.5
Random Forest	94.4	94.9	92.9	93.8	98.2
Artificial Neural Network	86.7	86.0	84.7	85.3	96.2

From the results obtained, the XGBoost model shows the highest performance among all the machine learning algorithms with accuracy at 95.3%, precision at 95.5%, recall at 94.3%, and F1-score at 94.9%. XGBoost is an ensemble learning algorithm that combines multiple models to generate stronger and more reliable predictions (Arif Ali et al., 2023)., resulting in the best-performing model in this Alzheimer's disease prediction problem. This is further supported by related works from Pranao et al.(2022), Kasani et al.(2021), and Kavitha et al.(2022). However, the random forest model achieved the highest specificity at 98.2%, slightly better than XGBoost, at a specificity of 97.5%. This shows that the random forest model was slightly more effective in identifying individuals without Alzheimer's disease, as specificity measures the model's ability to correctly classify negative (No Alzheimer's disease) cases. Nonetheless, the overall performance evaluation metric suggested XGBoost is still the best-performing machine learning model. Therefore, the XGBoost model is selected to be incorporated into the application

Another consistent finding is the strong performance of the CatBoost and random forest models, which shows a result comparable to the XGBoost model. These close results in performance can be attributed to the fact that all three algorithms, CatBoost, random forest, and XGBoost use ensemble learning techniques. Random forest uses bootstrap aggregating, where multiple models are trained on different bootstrapped samples and their outputs are combined. While XGBoost and CatBoost use gradient boosting, where models are built sequentially to correct previous errors and combined into a strong classifier. However, a minute difference in their results may be due to their ensemble type, optimization, or the nature of the dataset. Overall, these models show a robust performance and are comparable to XGBoost in this project.

Next, an interesting finding is the comparatively weaker performance of the artificial neural network models. Despite being a deep learning algorithm, it did not perform well compared to other machine learning models in the table. This could be because deep learning algorithm generally requires large datasets to effectively generalize and may not learn well in a small or medium-sized dataset. Another reason is that deep learning algorithms are less effective in tabular or structured data. From the observation, artificial neural network models are outperformed by tree-based models such as XGBoost, Random Forest, and CatBoost, possibly because they cannot handle categorical variables well.

Apart from that, the least performing model is logistic regression. This is shown by a lower accuracy, precision, recall, F1 score, and specificity relative to other machine learning models. Logistic regression assumes a linear relationship between the input features and the log-odds of the output (prediction). This assumption limitation, when facing a non-linear or a more complex dataset such as medical data like Alzheimer's disease data, could contribute to its poorer performance compared to other models.

Overall, all seven machine learning models performed excellently in predicting Alzheimer's disease. The Logistic Regression model recorded the lowest accuracy at 83.7%, while support vector machine and artificial neural network achieved accuracies in the 86% range. Next, the decision tree, random forest, Catboost, and XGBoost achieved significantly higher accuracies well-above the 92% accuracy score. These comparisons provide insights into the strengths and weaknesses of each model in predicting Alzheimer's disease. In Figure 5.1, the model performance is visualized using a bar graph to provide clearer insights and facilitate better comparison.

Figure 5.1 Comparison of The Machine Learning Models

5.3 Summary

This chapter discusses the evaluation results and comparative analysis of the machine learning algorithm in detail. The comparative analysis was conducted on the seven machine learning algorithms, including logistic regression, decision tree, support vector machine (SVM), CatBoost, XGBoost, random forest, and artificial neural network (ANN). This comparison helps identify the strengths and weaknesses of each model, and the best machine learning model is selected from this analysis. It is found that XGBoost is the best-performing machine learning algorithm.

CHAPTER 6: CONCLUSION AND FUTURE WORKS

6.1 Introduction

In this chapter, a conclusion is drawn from this comparative study of machine learning on the prediction of Alzheimer's disease. The primary objectives of this research were to identify the relevant features for the prediction, evaluate and compare the performance of various machine learning algorithms, and determine the best-performing model for the prediction of Alzheimer's disease. This chapter will provide a summary of the project's outcomes, highlight contributions, discuss potential areas for future work, and outline the limitations of the study to identify opportunities for further improvement.

6.2 Contributions

This project has made several significant contributions to the field of Alzheimer's disease prediction. The project successfully developed a high-performing prediction model for Alzheimer's disease detection. Then, the project indicates that XGBoost model shows the best performance across the evaluation metrics, including high accuracy, recall, specificity, precision, and F1 score. Besides, an application was also built using the model, allowing users to interact with and use the trained model.

Next, this project provides a comprehensive analysis and comparison of models in Alzheimer's disease prediction. A total of seven models were evaluated, including Logistic Regression, Decision Tree, Support Vector Machine (SVM), CatBoost, XGBoost, Random Forest, and Artificial Neural Network (ANN). This comparative analysis provides valuable insights into the strengths and weaknesses of each model, allowing future research to refer to this study as a reference for better research on projects related to these studies.

Furthermore, this project contributes to the growing body of research on the application of machine learning in healthcare, particularly in the prediction of Alzheimer's disease. The results and insights gained from this study can be a valuable reference for future research efforts. This can support the development of more efficient prediction models in this field.

6.3 Limitation

One of the main limitations of this project is the lack of diversity in the age group of the dataset. The Alzheimer's disease dataset used is sourced from Kaggle, which includes data only for individuals above the age of 60. This exclusion of younger age groups can limit prediction capability for early-onset Alzheimer's disease or Alzheimer's in younger age groups.

Another limitation is the relatively average sample size of only 2,149 medical records. While the dataset is not small, it is also not large. To ensure stronger generalizability, a larger dataset is preferred. A larger and more diverse dataset can potentially improve the model's effectiveness and robustness.

6.4 Future Works

Future work can address the limitations of this project by utilizing larger datasets or implementing real-time data collection from patients or individuals. This would help gather more comprehensive and accurate information. Thus, improving the reliability and generalizability of the models.

Additionally, future work can explore the use of deep learning models for Alzheimer's disease prediction by using different types of datasets, such as medical images, X-rays, or mammograms, instead of tabular data. This would enable an image-based diagnostic approach and allow researchers to evaluate the effectiveness of deep learning techniques in finding patterns associated with Alzheimer's disease from medical imagery.

Furthermore, future research can also focus on the development of hybrid models by combining multiple algorithms to create more robust and accurate predictive systems. For instance, integrating traditional machine learning models like Random Forest with deep learning architectures such as Convolutional Neural Networks (CNNs) could improve predictive performance and generalizability.

6.5 Summary

In conclusion, this project presented a comparative study of machine learning models for the prediction of Alzheimer's disease. This chapter summarizes the project's key outcomes, highlights its contributions, discusses potential areas for future work, outlines the study's limitations, and identifying opportunities for further improvement and research.

Additionally, project objectives such as selecting the appropriate features, comparing the performance of the machine learning model, and determining the best machine learning model for predicting Alzheimer's disease have all been successfully achieved.

Lastly, this research has made significant contributions to the field of Alzheimer's disease prediction by providing valuable insights into the application of machine learning in the medical domain. However, the limitations of this project should also be acknowledged, as they present opportunities for further enhancement and exploration isn future works.

REFERENCES

Alatrany, A. S., Khan, W., Hussain, A., Kolivand, H., & Al-Jumeily, D. (2024). An explainable machine learning approach for Alzheimer's disease classification. *Scientific Reports 2024 14:1*, *14*(1), 1–18. https://doi.org/10.1038/s41598-024-51985-w

Alexander S. Gillis. (2024). *What is data mining? | Definition from TechTarget*. https://www.techtarget.com/searchbusinessanalytics/definition/data-mining

Alkhawaldeh, I. M., Albalkhi, I., & Naswhan, A. J. (2023). Challenges and limitations of synthetic minority oversampling techniques in machine learning. *World Journal of Methodology*, *13*(5), 373. https://doi.org/10.5662/WJM.V13.I5.373

Aman. (2024). *Feature Selection in Machine Learning - Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/

An, J. (2022). Using CatBoost and Other Supervised Machine Learning Algorithms to Predict Alzheimer's Disease. *Proceedings - 21st IEEE International Conference on Machine Learning and Applications, ICMLA 2022*, 1732–1739. https://doi.org/10.1109/ICMLA55696.2022.00265

Arab, L., & Sabbagh, M. N. (2010). Are certain life style habits associated with lower Alzheimer disease risk? *Journal of Alzheimer's Disease : JAD*, *20*(3), 785. https://doi.org/10.3233/JAD-2010-091573

Arif Ali, Z., H. Abduljabbar, Z., A. Tahir, H., Bibo Sallow, A., & Almufti, S. M. (2023). eXtreme Gradient Boosting Algorithm with Machine Learning: a Review. *Academic Journal of Nawroz University*, *12*(2), 320–334. https://doi.org/10.25007/AJNU.V12N2A1612

Chang, W., Wang, X., Yang, J., & Qin, T. (2023). An Improved CatBoost-Based Classification Model for Ecological Suitability of Blueberries. *Sensors*, *23*(4). https://doi.org/10.3390/S23041811/S1

Fayyad, P.-S. S. (1996). *KDD Process/Overview*. Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, Menlo Park, CA. https://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html

Kasani, P. H., Kasani, S. H., Kim, Y., Yun, C. H., Choi, S. H., & Jang, J. W. (2021). An Evaluation of Machine Learning Classifiers for Prediction of Alzheimer's Disease, Mild Cognitive Impairment and Normal Cognition. *International Conference on ICT*

Convergence, *2021-October*, 362–367. https://doi.org/10.1109/ICTC52510.2021.9620780

Kavitha, C., Mani, V., Srividhya, S. R., Khalaf, O. I., & Tavera Romero, C. A. (2022). Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models. *Frontiers in Public Health*, *10*, 853294. https://doi.org/10.3389/FPUBH.2022.853294/BIBTEX

Kononenko, I., & Kukar, M. (2007). Measures for Evaluating the Quality of Attributes. *Machine Learning and Data Mining*, 153–180. https://doi.org/10.1533/9780857099440.153

Korczyn, A. D. (2012). Why have we failed to cure Alzheimer's disease? *Journal of Alzheimer's Disease : JAD*, *29*(2), 275–282. https://doi.org/10.3233/JAD-2011-110359

Kornblith, E., Bahorik, A., Boscardin, W. J., Xia, F., Barnes, D. E., & Yaffe, K. (2022). Association of Race and Ethnicity With Incidence of Dementia Among Older Adults. *JAMA*, *327*(15), 1488–1495. https://doi.org/10.1001/JAMA.2022.3550

Lev, C. (2024). *What is Machine Learning? Guide, Definition and Examples*. https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML

Mahadevkar, S. V., Khemani, B., Patil, S., Kotecha, K., Vora, D. R., Abraham, A., & Gabralla, L. A. (2022). A Review on Machine Learning Styles in Computer Vision - Techniques and Future Directions. *IEEE Access*, *10*, 107293–107329. https://doi.org/10.1109/ACCESS.2022.3209825

Messaoud, S., Bradai, A., Bukhari, S. H. R., Quang, P. T. A., Ahmed, O. Ben, & Atri, M. (2020). A survey on machine learning in Internet of Things: Algorithms, strategies, and applications. *Internet of Things*, *12*, 100314. https://doi.org/10.1016/J.IOT.2020.100314

Murel, J. (2024). *Create a confusion matrix with Python - IBM Developer*. https://developer.ibm.com/tutorials/awb-confusion-matrix-python/

Nagaratnam, J. M., Sharmin, S., Diker, A., Lim, W. K., & Maier, A. B. (2022). Trajectories of Mini-Mental State Examination Scores over the Lifespan in General Populations: A Systematic Review and Meta-Regression Analysis. *Clinical Gerontologist*, *45*(3), 467–476. https://doi.org/10.1080/07317115.2020.1756021;WGROUP:STRING:PUBLICATION

Nujan, N. I. (2018). *Top 5 Machine Learning Libraries in Python | by Nasir Islam Sujan | Towards Data Science*. https://towardsdatascience.com/top-5-machine-learning-libraries-in-python-e36e3e0e02af

Pemasinghe, S. (2018). *The Z-score and cut-off values - Sajeewa Pemasinghe*. https://sajeewasp.com/the-z-score/

Rohit Sharma. (2024, January 25). *KDD Process in Data Mining: What You Need To Know? | upGrad blog*. https://www.upgrad.com/blog/kdd-process-data-mining/

Salomão, A. (2024). *Pearson Correlation: Understanding the Math Behind Relationships - Mind the Graph Blog*. https://mindthegraph.com/blog/pearson-correlation/

Singh, A. (2024). *Feature Scaling In Machine Learning: What Is It?* https://www.appliedaicourse.com/blog/feature-scaling-in-machine-learning/

Singh, P., Singh, N., Singh, K. K., & Singh, A. (2021). Diagnosing of disease using machine learning. *Machine Learning and the Internet of Medical Things in Healthcare*, 89–111. https://doi.org/10.1016/B978-0-12-821229-5.00003-3

Smart Vision Europe. (2024). *Crisp DM methodology - Smart Vision Europe*. https://www.sv-europe.com/crisp-dm-methodology/

Sruthi. (2024). *Random Forest in Machine Learning*. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

The Helper Bees. (2024). *Alzheimer's and the Importance of Brain Health - The Helper Bees*. https://www.thehelperbees.com/families/healthy-hive/alzheimers-and-the-importance-of-brain-health/

Toal, R. (2024). *What is Python Used For? 7 Practical Uses for Python | Code Institute Code Institute Global*. https://codeinstitute.net/global/blog/what-is-python-used-for/

Tounsi, Y., Anoun, H., & Hassouni, L. (2020). CSMAS: Improving Multi-Agent Credit Scoring System by Integrating Big Data and the new generation of Gradient Boosting Algorithms. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3386723.3387851

Uslu, Ç. (2022). *What is Kaggle? | DataCamp*. https://www.datacamp.com/blog/what-is-kaggle

Valero-Carreras, D., Alcaraz, J., & Landete, M. (2023). Comparing two SVM models through different metrics based on the confusion matrix. *Computers & Operations Research*, *152*, 106131. https://doi.org/10.1016/J.COR.2022.106131

Wang, W., Chakraborty, G., & Chakraborty, B. (2021). Predicting the risk of chronic kidney disease (Ckd) using machine learning algorithm. *Applied Sciences (Switzerland)*, *11*(1), 1–17. https://doi.org/10.3390/APP11010202