HYBRID MODELLING USING DECISION TREE AND ORDERED REGRESSION: AN APPLICATION TO HEALTH SCIENCES RESEARCH

HAZIK BIN SHAHZAD

UNIVERSITI SAINS MALAYSIA

2024

HYBRID MODELLING USING DECISION TREE AND ORDERED REGRESSION: AN APPLICATION TO HEALTH SCIENCES RESEARCH

by

HAZIK BIN SHAHZAD

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

January 2024

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to my thesis advisor, Professor Dr. Wan Muhamad Amir Bin W Ahmed, for his unwavering guidance and support throughout this research journey. His open-door policy and patience were instrumental whenever challenges arose. His profound knowledge and dedication served as a constant source of motivation, and his insightful comments and encouragement propelled me forward. His mentorship not only enriched my research but also pushed me to explore diverse perspectives, contributing to the depth of my work. I am truly fortunate to have had him as my advisor, and I extend my sincere thanks.

Additionally, I owe a debt of gratitude to Dr. Anas for guiding me in choosing USM for my academic pursuits. The research environment at USM facilitated interactions with exceptional peers, fostering an atmosphere of learning and critical thinking.

A special note of appreciation goes to my sister, Maira. Her unwavering belief in me, countless late-night discussions, and willingness to lend a helping hand during hard times was instrumental in my ability to navigate the hurdles of this academic journey.

I extend my deepest gratitude to my wife, Sadia, whose boundless love, and patience have been my pillars of strength. Her steadfast support played a pivotal role in the completion of this thesis. I am grateful for her unwavering presence throughout this journey. Thank you for being by my side.

Lastly, my heartfelt appreciation goes to my parents, whose relentless support and encouragement have been my driving force throughout my academic journey. Their unwavering belief in me and their constant encouragement have made this achievement possible. Thank you for always being there for me.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

AID	Automatic Interaction Detection
CART	Classification and Regression Tree
CDSS	Clinical Decision Support Systems
CHAID	Chi-squared Automatic Interaction Detection
CI	Confidence Interval
CIT	Conditional Inference Trees
DM	Diabetes Mellitus
FN	False Negative
FP	False Positive
GBT	Gradient Boosting Trees
HAI	Hybrid Accuracy Index
ICU	Intensive Care Unit
ID3	Iterative Dichotomiser 3
IPS	Institut Pengajian Siswazah
LRT	Likelihood Ratio Test
ML	Machine Learning
MLE	Maximum Likelihood Estimation
NPV	Negative Predictive Value
OR	Odds Ratio
PD	Pre-Diabetic
PPV	Positive Predictive Value
R^2_{CU}	Cragg-Uhler's R-Squared
SRGH	Self-Rated Gum Health
SRTH	Self-Rated Tooth Health
THAID	Theta Automatic Interaction Detection
TN	True Negative
TP	True Positive
USM	Universiti Sains Malaysia

LIST OF APPENDICES

MODEL HIBRID MENGGUNAKAN PEPOHON KEPUTUSAN DAN REGRESI ORDINAL: SUATU APLIKASI DALAM PENYELIDIKAN SAINS KESIHATAN

ABSTRAK

Dengan meningkatnya kompleksiti data kesihatan, terdapat keperluan untuk teknik pemodelan ramalan yang lebih canggih dan integratif. Tesis ini membentangkan metodologi hibrid baharu yang mengintegrasikan pepohon keputusan dan regresi ordinal dengan menggunakan sintaks R. Objektif kajian ini merangkumi pembangunan kaedah hibrid, pengukuran tahap keberkesanan serta kecekapan, pengesahan terhadap prestasi menerusi analisis pengelasan peramalan, dan pengoptimuman terhadap nilai anggaran parameter untuk pentaabiran statistik. Metodologi hibrid ini menggunapakai kaedah pepohon keputusan, dipermudahkan secara visualisasi bagi mengenal pasti faktor yang berpengaruh dalam pembentukan model peramalan. Kaedah pensampelan semula menerusi bootstrap meningkatkan lagi tahap keteguhan set data serta pembentukan model regresi ordinal. Pengenalan indeks ketepatan hibrid meningkatkan lagi tahap kebolehtafsiran. Metodologi hibrid ini diaplikasikan dalam dua senario sains kesihatan. Kes Kajian I, memberi fokus kepada peramalan terhadap kekerapan memberus gigi di kalangan pelajar. Manakala bagi Kes Kajian II, ia meramalkan status diabetes dengan menggunakan penunjuk kesihatan oral. Kajian ini memperkenalkan kaedah hybrid yang menghasilkan hasil berangka berserta visualisasi grafik, meningkatkan ketepatan dan kecekapan anggaran parameter. Penemuan kajian ini menyumbang kepada pembangunan pendekatan inovatif dalam mengubah pemodelan ramalan dalam penjagaan kesihatan, menyumbang kepada metodologi penyelidikan masa depan untuk membuat keputusan yang lebih tepat.

HYBRID MODELLING USING DECISION TREE AND ORDERED REGRESSION: AN APPLICATION TO HEALTH SCIENCES RESEARCH

ABSTRACT

With the increasing complexity of healthcare data, there is a need for more advanced and integrative predictive modelling techniques. This thesis presents a novel hybrid methodology integrating Decision Trees and Ordinal Regression using the R-syntax. The study objectives include the development of the hybrid method, measuring its efficacy and efficiency, validating its performance through predictive classification analysis, and optimising parameter estimates for optimised statistical inferences. The hybrid methodology uses decision trees, facilitated by visualisation tools, to identify influential factors that shape the model's predictions. The bootstrap resampling method boosts the data set's resilience and facilitates the development of an ordinal regression model. The introduction of the hybrid accuracy index enhances interpretability. The hybrid methodology is employed in two health sciences scenarios. In Case I, it predicts the frequency of toothbrushing among students, and in Case II, it predicts diabetic status using oral health indicators. This study introduces a hybrid method that generates numerical results along with graphical visualisation, enhancing the accuracy and efficiency of the parameter estimates. The findings of this study contribute to the development of an innovative approach to transforming predictive modelling in healthcare, contributing to future research methodologies for more precise decision-making.

CHAPTER 1

INTRODUCTION

1.1 Overview of the Chapter

This chapter serves as a comprehensive summary of the thesis, beginning with an introduction to the study's background. It further explores the research problem and motivation behind the investigation. It provides the rationale for integrating decision trees and ordinal regression methodologies, providing a solid basis for the study. The conceptual framework is then introduced, offering a flowchart that outlines the entire research process.

Moving forward, the chapter presents the research hypothesis and outlines both the general and specific goals of the study. It also delves into the study's scope and highlights its contribution in line with the research objectives. The chapter concludes with a candid discussion of the study's limitations, recognising potential constraints that may impact the research outcomes. Additionally, it offers a comprehensive description of the thesis organisation, providing readers with a clear understanding of how the analysis was conducted and structured throughout the thesis.

1.2 Background of the Study – A Review of Statistical Modelling

Statistical applications have become indispensable in nearly every scientific discipline, ranging from economics and education to biology and health sciences. In the field of health sciences, statistical methods, often referred to as biostatistics, play a crucial role in analysing complex data and drawing meaningful conclusions. These statistical techniques are vital for inferring relationships between variables and making predictions related to health outcomes (Chowdhry, 2023).

Regression analysis, a fundamental statistical method, has a rich history and has been widely used to explore relationships between variables and forecast future values (Montgomery *et al.*, 2021). Its development dates back to the work of Legendre and Gauss in 1805, and it has evolved over time to become a cornerstone of statistical research (Sheynin, 2020). In the context of health sciences research, regression analysis has been used to study various aspects of medical and biological data. However, with the increasing complexity of health data, there is a growing need for more advanced and hybrid approaches to prediction and classification modelling (Khan and Yairi, 2018).

Previous research has already shown promising results with hybrid methodologies that combine multiple approaches to improve model accuracy and predictability. Ahmad *et al*. (2016) conducted a study demonstrating the effectiveness of a hybrid statistical methodology, while Adnan *et al*. (2023) successfully applied a hybrid approach by combining linear regression models with a multilayer perceptron (Ahmad *et al.*, 2016; Adnan *et al.*, 2023). These studies highlight the potential of integrating or combining different statistical techniques to achieve better results than standalone methods. Building upon these prior findings, the current study proposes a novel integration of decision tree analysis with ordered logistic regression for health sciences. By combining these methodologies, this study seeks to advance predictive modelling in the health sciences domain. The hybrid approach is expected to improve model accuracy, provide interpretable predictions, and guide evidence-based decision-making in healthcare settings. The findings of this thesis will contribute to the field of health sciences research and offer valuable insights for researchers and practitioners in developing robust and accurate prediction models. This innovative approach is

anticipated to contribute valuable insights to health sciences research and support the development of robust prediction models.

1.3 Research Motivation

This study addresses health research within the fields of medical and dental sciences, where traditional methods like percentage estimation and regression modelling have been prevalent (Lemon *et al.*, 2003). However, these conventional methods often assume linearity, independence, and normality, which may not reflect the realities of healthcare data. They struggle with multidimensional data, outliers, categorical data, and imbalanced datasets (Adnan and Akbar, 2019). Moreover, the complex, nonlinear associations found in healthcare data, arising from various factors, often elude simple linear models, especially when dealing with numerous variables (Hosmer Jr *et al.*, 2013; Chatterjee *et al.*, 2019).

In the realm of predictive modelling and data analysis, the continuous pursuit is to enhance accuracy, interpretability, applicability, and generalisability of methods. Predictive modelling involves creating statistical models to forecast future outcomes based on historical data (Toma and Wei, 2023). However, predictive models alone are sensitive, can overfit, and lack interpretability (Walter and Estey, 2020). In response, the integration of various methodologies has emerged as a potential solution. Transitioning from these challenges, this research aims to develop, validate, and optimise a hybrid modelling approach. This hybrid approach seamlessly blends decision tree analysis and ordered logistic regression, to achieve prediction classification by leveraging their respective strengths. By exploring this integration using advanced statistical tools like R syntax, the study strives to present a more

efficient, accurate, and versatile predictive modelling approach, contributing to the advancement of research methodologies.

The integration of statistical techniques, particularly decision tree analysis and ordinal logistic regression, offers the potential for a comprehensive grasp of health dynamics. Decision trees adeptly capture non-linear relationships, while ordinal logistic regression excels in handling ordinal outcomes with interpretable coefficients (Song and Ying, 2015; James *et al.*, 2013). By combining these components, the hybrid methodology harnesses their strengths while alleviating individual limitations, resulting in enhanced predictive model accuracy, interpretability, and robustness. Emphasising concise explanations and adhering to the principle of parsimony, this study aims to provide valuable insights into the complex interplay of variables within health sciences.

1.4 Rationale of the Study

In the domain of predictive modelling and classification, both decision tree analysis and ordered logistic regression offer distinct strengths and advantages. Decision trees are well-known for their ability to capture complex non-linear relationships and interactions among predictors, making them valuable tools for handling intricate data patterns (Kotsiantis, 2013). On the other hand, ordered logistic regression excels in modelling ordinal outcomes, where the response variable has an inherent ordering or hierarchy, normally found in many health scenarios (Tutz, 2022).

The rationale behind combining decision tree analysis with ordered logistic regression lies in the complementary nature of these methodologies. By integrating them, the strengths of both approaches can be leveraged along with mitigating their individual limitations, leading to more accurate, interpretable, and robust predictive

models. Decision tree algorithms struggle with ordered categorical outcome variables as they do not effectively capture the hierarchical relationships among outcome classes. Instead, they treat each category as independent, potentially leading to a loss of ordinal information and generating splits that don't align with the natural order of categories (Cao-Van and De Baets, 2003). In contrast, ordered logistic regression models the ordinal nature of the response variable, providing interpretable coefficient estimates that offer valuable insights into the direction and magnitude of predictor effects.

Moreover, the hybrid methodology can aid in better understanding the underlying relationships between predictors and outcomes, enabling evidence-based decision-making, and guiding targeted interventions in various research domains. The study of combining decision tree analysis with ordered logistic regression contributes to advancing predictive modelling techniques, ultimately benefiting the fields of healthcare, social sciences, and other areas where ordered outcomes play a significant role.

1.5 Conceptual Framework

This study's conceptual framework aimed to develop an extensive methodology by integrating decision trees with ordinal logistic regression. The framework comprises three key parts: theoretical concepts, data analysis and subsequent statistical inferences and results (Figure 1.1). The theoretical concepts involve understanding decision trees and ordinal logistic regression methodologies. Model building involves creating algorithms and incorporating accuracy metrics. The data analysis involves two distinct cases. Case I focuses on dental hygiene practices, and Case II identifies diabetes through the assessment of oral symptoms. It progresses

through decision tree diagram development, classification prediction, bootstrapping, and ordinal regression analysis. The results phase involves interpreting predictions, statistical inferences, and the final outcomes, culminating in comprehensive thesis writing.

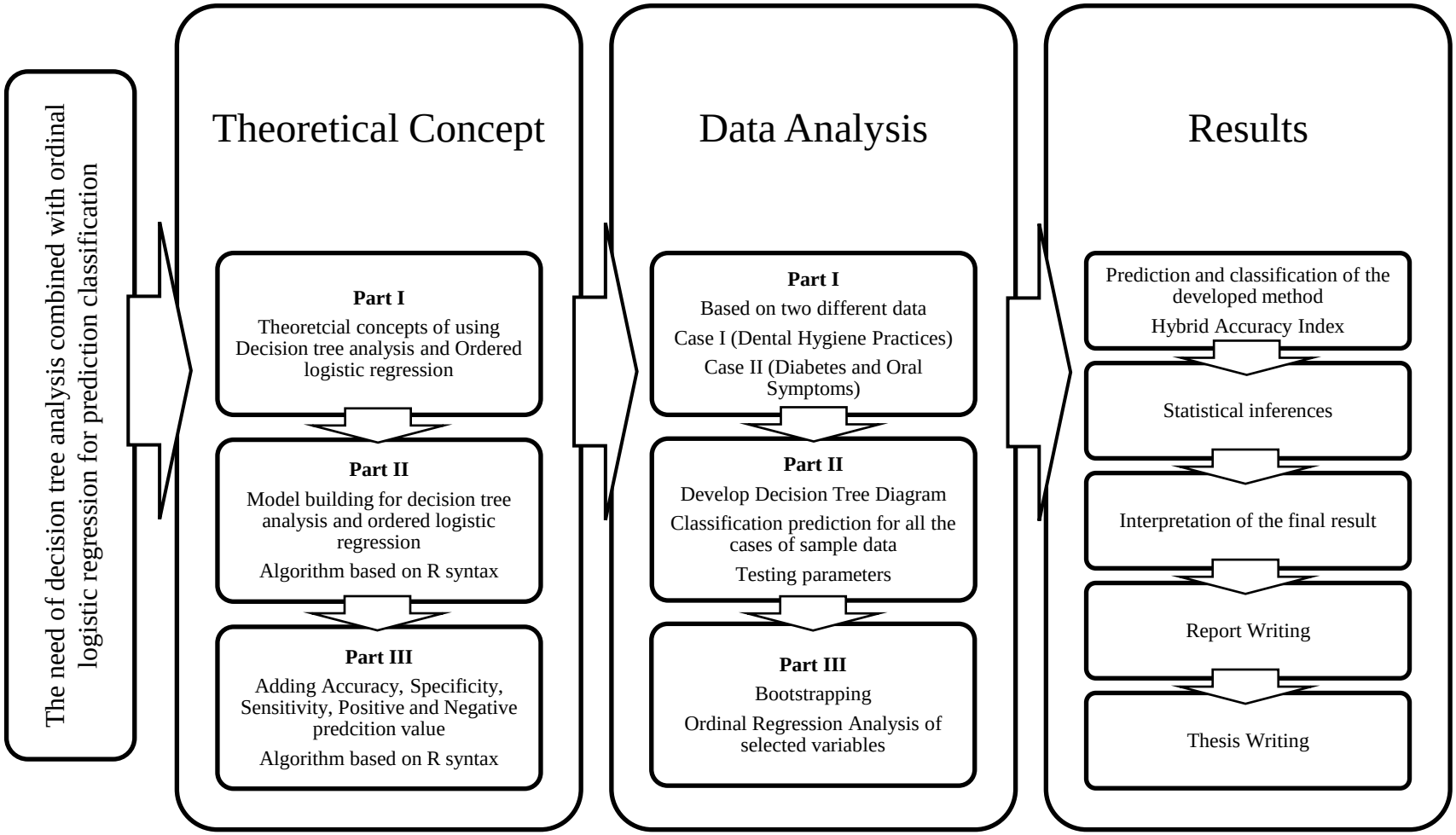

Figure 1.1: Conceptual framework

1.6 Research Hypothesis

This research is vital for the fundamental development of statistical computing methodology. The following hypotheses are proposed:

a) The integration of decision tree and ordered logistic regression using the R syntax will result in a novel and effective approach for predictive modelling.
b) The new model will improve the level of analysis in terms of the hybrid efficiency and efficacy.
c) The hybrid model will accurately predict or classify outcomes when tested with sample data.
d) Optimising the parameter estimates of the hybrid method will lead to improved statistical inferences.

1.7 Research Questions

These research questions serve as guiding principles for the study, leading to a comprehensive investigation of the integration of decision trees and ordinal regression methodologies and their potential impact on predictive modelling with ordinal outcome variables. The research will address the following key research questions:

a) How can a new hybrid method combining decision tree and ordered logistic regression be developed using the R syntax?
b) How can the new model contribute to enhancing the level of analysis in terms of the hybrid efficiency and efficacy?
c) How can the hybrid method's validity be established through prediction classification on sample data?
d) How can the parameter estimate of the developed hybrid method be optimised to enhance statistical inferences?

1.8 Research Objectives

By addressing the objectives, the study aims to contribute to the advancement of statistical computing methodology and its applications in various research domains, particularly in the context of healthcare and medical sciences. This section outlines the specific research objectives that drive the study, addressing the identified research gap.

1.8.1 General Objective

To develop and elucidate integration of decision tree and ordered logistic regression algorithm using the R language for prediction and classification in health sciences research.

1.8.2 Specific Objective

The following specific research objectives will guide the study in achieving its overarching goal of exploring and validating the hybrid method.

a) To develop a new hybrid method combining decision tree and ordered logistic regression using the R syntax.
b) To evaluate the efficiency and efficacy of the hybrid model in enhancing the level of analysis.
c) To elucidate and validate the hybrid method through prediction classification using sample data.
d) To optimise parameter estimates in the developed hybrid method to enhance statistical inferences.

1.9 Scope of the Study

This study focuses on the development and validation of a hybrid method using decision trees and ordered logistic regression for prediction and classification. The

scope of the research includes exploring the integration to address prediction problems in healthcare scenarios. The study aims to enhance the accuracy and interpretability of predictive models while mitigating the limitations associated with standalone decision tree and ordinal logistic regression approaches. The research would involve real-world datasets from health sciences, providing practical insights and applications. The study's scope is developed using R programming language for implementing the hybrid method. The evaluation and validation of the hybrid model will be performed through rigorous testing using prediction classification techniques.

1.10 Contribution of the Study

Statistical computing in health sciences research is propelled forward by this study through the integration of multiple techniques and harnessing computational tools like R programming. The study outlines the value of hybrid methodologies over individual techniques and lays the groundwork for future statistical model advancements. The hybrid method of combining decision tree analysis with ordered logistic regression, aims at improving predictive accuracy and interpretability, especially for ordinal outcome variables. By capturing non-linear relationships through decision trees and addressing ordinal outcomes via ordered logistic regression, the methodology refines predictions in healthcare contexts. This greatly improves precision, thereby enabling more informed decision-making within healthcare settings.

This study also makes significant contributions by identifying and validating key predictors, thereby enhancing the understanding of health determinants, and guiding potential interventions. This study showcases the practicality and effectiveness of the proposed hybrid model through rigorous prediction classification and validation, demonstrating its robustness in making accurate predictions and

classifications based on complex real-world data. It also expands insights into the relationships between predictor variables and health outcomes by examining parameter estimates. By merging these attributes into a transparent hybrid model, researchers gain a more comprehensive approach to interpreting models, thereby progressing the field's capacity to generate meaningful insights. Furthermore, this study enhances the consistent evaluation of various models, ensuring rigorous guidelines in biometry modelling and fostering reliable outcomes.

Additionally, the introduction of the Hybrid Accuracy Index (HAI) and the integration of diverse techniques streamlines data analysis, simplifying complex model evaluation for researchers and healthcare practitioners. This research demonstrates the potential for cost and time savings in data analysis, offering more resource-efficient approaches in the face of escalating healthcare expenditures. By utilising real-world health sciences datasets, the study ensures practical relevance, directly influencing areas such as disease diagnosis, treatment selection, prognosis assessment, and public health planning.

1.11 Organisation of the Thesis

The thesis is structured to address research objectives related to combining decision trees and ordinal regression. It offers a logical and coherent progression to ensure a comprehensive understanding. Here's a concise overview:

This thesis comprises six chapters. Chapter 1 serves as the introduction, establishing context, objectives, rationale, scope, methodology, significance, and limitations. This foundational chapter sets the stage for the entire thesis. Chapter 2 provides a thorough literature review of decision trees and ordinal regression, examining their strengths, limitations, and applications. Chapter 3 introduces the

methodology, detailing data selection, study design, and hybrid model creation. A flowchart is presented to illustrate the proposed statistical modelling process. It elaborates on the specific algorithm employed and outlines the steps involved in the hybrid modelling process. Chapters 4 and 5 present model results and their interpretation, discussing their implications and strengths. Chapter 6 concludes with findings, recommendations, and future research directions.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview of the Chapter

With the significance of statistical analysis in mind, this literature review comprehensively explores theoretical foundations, model assumptions, diagnostic tools, and practical considerations related to decision trees and ordinal regression. The last sections of the chapter present concluding thoughts by emphasising the need for a hybrid methodology to address challenges and maximise predictive modelling accuracy and interpretability across health sciences research.

2.2 Introduction to Decision Tree

Decision trees, belong to the family of supervised learning algorithms, where models are trained using labelled data (Soni and Varma, 2020). These models predict the classes or outcomes of new data and are typically evaluated using separate test datasets (Ramesh *et al.*, 2022). Decision trees are regarded as highly powerful tools for achieving classification and prediction objectives (Kantardzic, 2011). Their intuitive structure allows decisions to be reached using a set of features or attributes, making them popular in decision analysis (Luna *et al.*, 2019). The simplicity, interpretability, and ability to capture complex relationships within the data have contributed to their popularity (Dumitrescu *et al.*, 2022).

Decision trees can be visualised as flowcharts or tree-like structures, making them easy to interpret (Gilmore *et al.*, 2021; Roy and Shree, 2023). The internal nodes represent decision points, and the branches correspond to possible outcomes (Sharma and Kumar, 2016). The leaf nodes represent the final predictions or outcomes (Song and Ying, 2015). This transparency allows for effective decision support and

communication, enabling both experts and non-experts to understand the underlying logic and reasoning (Gilpin *et al.*, 2018). Figure 2.1 shows the basic structure of a decision tree (Chiu *et al.*, 2016).

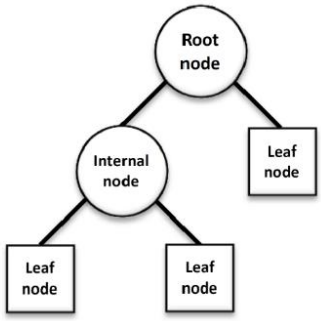

Figure 2.1: Basic structure of decision tree (Chiu *et al.*, 2016, p.187)

Compared to other algorithms, decision trees involve relatively less complex mathematical calculations, making them accessible to beginners in Machine Learning (ML) and data science (Lian *et al.*, 2020; Biehler and Fleischer, 2021). The construction process involves recursive partitioning, where the algorithm selects the best features to split the data based on specific metrics, creating homogeneous subsets that are more predictable or classifiable (Chandrasekar *et al.*, 2017). Different algorithms may use different metrics, depending on the problem and implementation (Zacharis, 2018).

Decision trees are also versatile, capable of handling categorical and numerical features, and capturing complex decision boundaries and non-linear relationships (Sharma and Kumar, 2016). They have the capability to manage missing values and outliers, thereby minimising the necessity for extensive data preprocessing (Zahin *et al.*, 2018; Shamrat *et al.*, 2021). Decision trees serve not only as tools for analysing past data but also for making predictions about the future. Their ability to learn from existing patterns and decision rules allows them to be applied to fresh, uncharted data,

enabling the forecasting of outcomes, identification of trends, and accurate classifications.

2.2.1 History of Decision Trees

The origins of decision trees can be traced back to the ancient Greek philosopher Porphyry, who introduced the Porphyrian tree, the oldest documented classification tree dating back to the third century (Lima, 2011). In the beginning of the 20th century, Frank P. Ramsey developed the first methodology of decision analysis, while Ronald Fisher's paper on discriminant analysis in 1936 influenced many in the field (Ramsey, 1931; Fisher, 1936; Insights, 2022). These early origins highlight the intuitive and powerful visual display capabilities of decision trees, making them a useful tool for interpreting results. Modern decision trees have their roots in the research conducted by Dr. William Belson in the 1950s, who conducted a nationwide audience survey for the British Broadcasting Corporation (Belson, 1956). Belson introduced a further refinement that involved the differential assessment of nested subclass predictors (Belson, 1959).

Building upon Belson's work, Morgan and Sonquist found decision trees as a substitute to regression in analysing survey data. The first decision tree regression was invented in 1963 as part of the AID (Automatic Interaction Detection) project by Morgan and Sonquist (Morgan and Sonquist, 1963). Decision trees also found application in management decision-making, with John F. Magee publishing articles on their use in capital investments and decision making in 1964 (Magee, 1964).

In 1966, the decision tree model was further explored by Hunt in the fields of psychology and programming (Hunt *et al.*, 1966). Messenger and Mandell developed the first classification tree for the THAID project (Theta Automatic Interaction Detection) in 1972 (Messenger and Mandell, 1972). In 1974, Jerome Friedman and

Richard Olshen from Stanford and Leo Breiman and Charles Stone from Berkeley worked on the Classification and Regression Tree (CART) algorithm, revealed in 1977 (Friedman, 1977). In 1984, the first official publication using a CART software marked a significant milestone as it revolutionised the world of algorithms and remains one of the most widely used methods for decision tree data analysis (Breiman *et al.*, 1984).

John Ross Quinlan played a significant role in the field, introducing the concept of trees with multiple answers. He created ID3 (Iterative Dichotomiser 3) in 1986, which was later upgraded to C4.5 to address its shortcomings. C4.5 gained recognition as a top algorithm in data mining (Quinlan, 2014). Despite advancements, algorithms such as CART and C4.5 rely on biased systems. The more recent algorithms, such as the Conditional Inference Trees (CIT) developed by Hothorn, Hornik, and Zeileis in 2006, offer an unbiased approach by selecting predictors based on statistical hypothesis testing, mitigating variable selection bias (Hothorn *et al.*, 2006).

Decision trees have a rich history that spans centuries. From the ancient Porphyrian tree to modern algorithms like CART and CIT, decision trees have evolved to become a powerful tool for interpreting results, aiding decision-making, and analysing data in various domains.

2.2.2 Decision Tree Structure

Before diving into the specifics of decision trees, it is important to understand the general structure of decision trees. Decision trees follow a hierarchical structure, similar to other tree concepts, depicted in Figure 2.2, and consist of three types of nodes: the root node, internal nodes, and leaf nodes (Chen *et al.*, 2009).

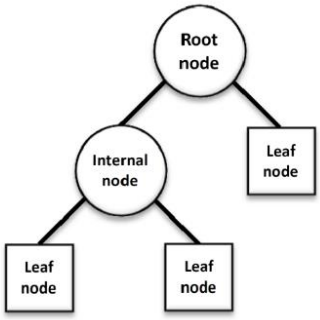

Figure 2.2: Types of nodes in decision tree (Chiu *et al.*, 2016, p.187)

- Root Node: The root node is the starting point and entry into the decision tree. It represents the entire dataset and initiates the decision-making process by making the first split based on a selected attribute. It has no incoming edges but can have zero or more outgoing edges. The outgoing edges from the root node lead to either internal nodes or leaf nodes. The branches from the root node guide the flow of data through the subsequent nodes in the tree. The root node typically represents an attribute or feature of the decision tree model and acts as the initial point of decision-making in the tree (Rastogi and Shim, 2000).
- Internal Nodes: Internal nodes, also known as decision nodes or test nodes, are the intermediate nodes in the decision tree that appear after the root node or other internal nodes. Each internal node represents a particular feature attribute or condition that is used to split the data. These nodes play a crucial role in the decision-making process by providing conditions for further branching. An internal node is connected to one incoming edge and has a minimum of two outgoing edges. The incoming edge represents the flow of data from the parent node, while the child nodes represent the subsets of data after the splitting. The number of outgoing edges corresponds to the distinct values or ranges of values for the selected attribute (Song and Ying, 2015).

- Leaf Nodes: Leaf nodes, also known as terminal nodes or decision nodes, are the bottom most elements of the decision tree. Leaf nodes have one incoming edge, either from an internal node or directly from the root node. This incoming edge represents the path that the data takes to reach the leaf node. However, leaf nodes do not have any outgoing edges since they represent the end of the decision tree's branches (Rokach and Maimon, 2005). They represent the final outcomes or classifications of the tree model, mostly containing a definitive prediction or classification (Fakhari and Moghadam, 2013).

The decision tree's structure guides the data flow from the root node, through various internal nodes, until reaching the corresponding leaf node that holds the final prediction (Kotsiantis, 2013). The hierarchical arrangement of nodes enables decision trees to effectively model complex decision-making processes and provide interpretable outcomes (Elmachtoub *et al.*, 2020).

2.2.3 Types of Decision Trees and their Branching

2.2.3(a) Univariate Decision Trees

Univariate decision trees involve splitting tree nodes based on one variable at a time, resulting in a tree that branches univariately. Univariate decision trees focus on finding the most discriminatory feature for making predictions or classifications (Bertsimas and Dunn, 2017). They employ a top-down, greedy approach, recursively splitting nodes based on the selected feature until a stopping criterion is met. They are easily understandable and provide interpretable models that can be easily understood by researchers and domain experts. Univariate decision trees offer flexibility and are suitable for both classification and regression tasks (Podgorelec *et al.*, 2002). Figure 2.3 depicts a univariate decision tree where each node corresponds to a single variable being split (Nanfack *et al.*, 2022).

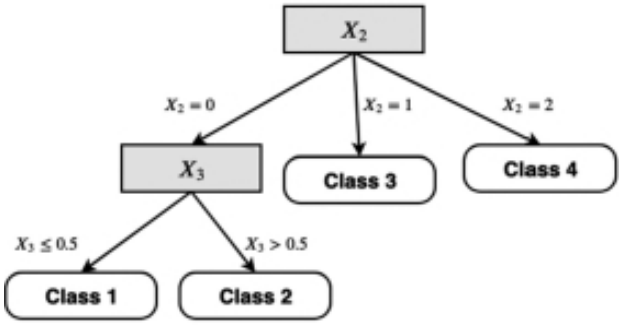

Figure 2.3: Univariate decision tree (Nanfack *et al.*, 2022, p. 201:3)

2.2.3(b) Multivariate Decision Trees

Multivariate decision trees extend the concept of traditional decision trees by considering multiple input variables simultaneously when making splitting decisions (Brodley and Utgoff, 1995; Kretowski and Czajkowski, 2018). They assess the joint information provided by multiple features, allowing for the capture of complex interactions and relationships between variables (Pei *et al.,* 2016). In contrast to traditional trees that assess individual attributes, the internal nodes of multivariate decision trees conduct tests on combinations of attributes. By considering the combined information from multiple features, these trees can potentially provide more detailed and accurate predictions or classifications (Pei *et al.,* 2016). Figure 2.4 depicts a multivariate decision tree where the node corresponds to multiple variables being split together (Nanfack *et al.*, 2022).

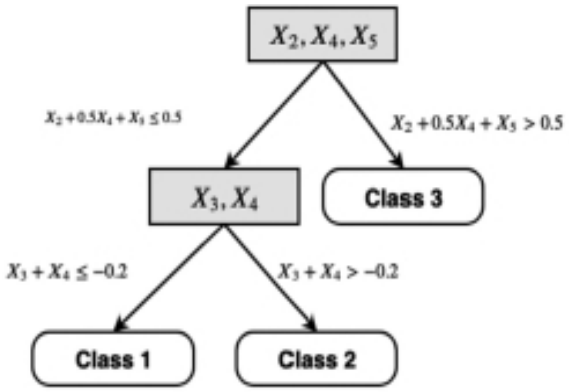

Figure 2.4: Multivariate decision tree (Nanfack et al., 2022, p. 201:3)

2.2.3(c) Types of Branching in Decision Trees

As decision trees traverse through the data, they make a series of decisions to reach a final prediction or classification. One crucial aspect of decision trees is their branching mechanism, where nodes split into multiple branches to handle different attribute values. These branching strategies play a pivotal role in how decision trees analyse data and make predictions.

The different types of branching in decision trees are the following:

- Discrete Valued Branching: In this type of branching, the decision tree splits based on categorical attribute values. Each attribute value forms a separate branch, resulting in splits representing the number of unique attribute values (Abspoel *et al.*, 2020). In Figure 2.5, the categorical attribute "Brush Size" has values "Small", "Medium", and "Large", and the decision tree will create three branches, one for each attribute.
- Continuous Valued Branching: Continuous valued branching is used when the attribute has continuous numeric values. The node is split into two intervals based on the test criteria, creating two branches: one for values below the threshold and another for values above the threshold (Abspoel *et al.*, 2020). In Figure 2.5, for the continuous numeric attribute "Age", the decision tree splits it based on a threshold of 12, creating one branch for individuals younger than and equal to 12 and another for individuals older than 12.
- Binary Discrete-Valued Branching: In this type of branching, the selected attribute is split into two branches using binary values. This type of branching is often used when dealing with binary attributes (Cha and Tappert, 2009). In Figure 2.5, the binary attribute "Flossing" has values "Yes" and "No", and the decision tree will create two branches, one for each.

- Attribute Value Grouping: Attribute value grouping involves merging similar or related attribute values into a single branch. This grouping simplifies the decision tree structure and can improve accuracy, particularly when dealing with attribute values that have a small number of instances (Pedrycz and Sosnowski, 2005; Yang, 2019). In Figure 2.5, the categorical attribute "Toothache" has values "Never", "Occasionally", and "Regularly", and the decision tree merges "Never" and "Occasionally" into one branch and keep "Regularly" as a separate branch to provide better predictive power or simplicity.

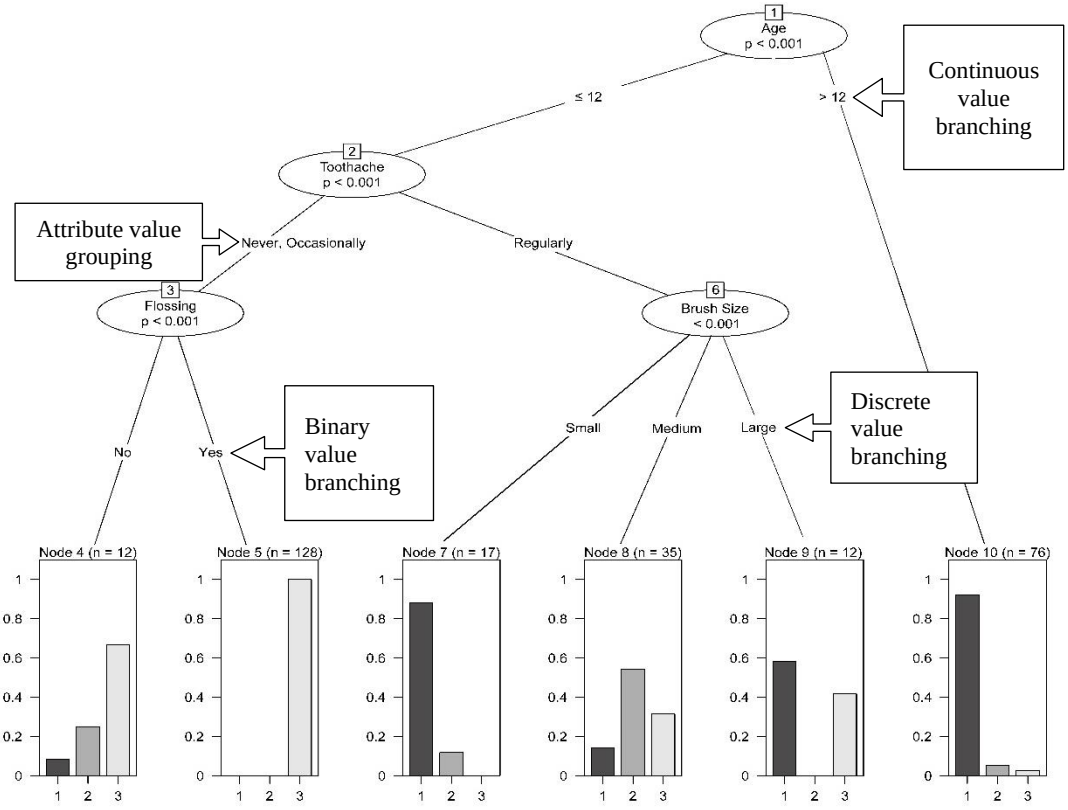

Figure 2.5: Branching types in decision trees

2.2.3(d) Choosing Between Univariate and Multivariate Decision Trees

Univariate decision trees focus on evaluating one feature at a time and provide interpretable models, while multivariate decision trees consider multiple features simultaneously to capture complex interactions and relationships. Univariate decision

trees are often simpler and more interpretable, making them suitable for scenarios where the relationships between variables are straightforward (Costa and Pedreira, 2023). On the other hand, multivariate decision trees are preferred when there are complex relationships and interactions between variables that need to be accurately captured (Yıldız and Alpaydın, 2000). The decision to employ either univariate or multivariate decision trees should be determined by the complexity and interpretability requirements of the specific problem domain.

2.2.4 Key Learnings of Decision Tree

The development and evolution of decision trees can be traced back to their pre-computational origins and their subsequent adaptation to mechanical calculators using Hollerith punch cards (De Ville, 2013). As the era of digital computers emerged, decision trees continued to prove their effectiveness as tools for discovering interactions among inputs and characterising target fields (Barrett, 2022). Unlike regression models, decision trees also have the ability to uncover interactions as the tree grows, eliminating the need for prior specification (De Ville, 2013).

Decision trees learn from training data to create a model that can make predictions or decisions based on observed patterns and relationships in the data (Ramesh *et al.,* 2022). Key learnings of decision trees include selecting features of importance, decision rules, relationships and interactions, hierarchical structure, generalisation, and the concept of overfitting and pruning (Kuhn *et al.*, 2014). Firstly, decision trees learn the relative importance or relevance of different features in making predictions. They evaluate split criteria to identify features that provide the most discriminatory power and contribute significantly to the decision-making process (Barros *et al.,* 2011). Secondly, decision trees learn the optimal decision rules that divide the input space into subsets or regions. These decision rules are based on the