Application of an Integrated Bayesian Network-Artificial Neural Network Model in Prediction of Brand Preference

Ombaka, Ralala Akinyi

Submitted in partial fulfilment of the requirements for the degree of Master of Science in Statistical Science at Strathmore University

Strathmore Institute of Mathematical Sciences

Strathmore University

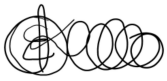
Nairobi, Kenya

June 2025

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the thesis itself.

Name: **Ombaka Ralala Akinyi**

Signature: 

Date: June 6, 2025

Approval

The dissertation of **Ombaka Akinyi Ralala** was reviewed and approved by the following:

Prof. Bernard Omolo
Professor, Institute of Mathematical Sciences,
Strathmore University

Dr. Godfrey Madigu
Dean, Institute of Mathematical Sciences,
Strathmore University

Prof. Bernard Shibwabo
Director of Graduate Studies,
Strathmore University

Abstract

The decision-making processes surrounding infant formula selection present significant challenges for public health interventions and market strategists, necessitating robust and interpretable predictive models. This study applied and comparatively evaluated standalone Bayesian Network (BN) and Artificial Neural Network (ANN) models, alongside an integrated BN-ANN architecture, to classify infant formula choices within a Djibouti context. Employing a sequential integration strategy, where BN-derived probabilistic inferences informed the ANN's feature set, the research analyzed model performance, feature importance, and interpretability. While the standalone BN model offered valuable probabilistic insights into conditional dependencies (Accuracy: $\sim$0.8500), the standalone ANN demonstrated significantly superior predictive power (Accuracy: 0.9495). Crucially, the integrated BN-ANN model achieved the highest predictive accuracy (0.9524) and Kappa score (0.9276), indicating a consistent, albeit marginal gain. Feature importance analysis revealed the taste, brand innovation , and completeness in range of baby formula products as the most dominant predictors. Unexpectedly, BN- probabilities contributed minimally as direct ANN features, a phenomenon potentially attributed to information loss from the necessary discretization of continuous variables for the BN, or insufficient variability in the generated probabilities. The study also noted near-perfect classification for one specific formula class, primarily due to highly separable intrinsic features. This research underscores the significant potential of hybrid artificial intelligence for complex multi-class classification, furnishing actionable insights for stakeholders in the infant nutrition sector in Djibouti. Furthermore, it contributes methodologically to the advancements in integrating probabilistic models with deep learning.

Keywords: Bayesian Network; Artificial Neural Network; Hybrid Models; Infant Formula Choice; Predictive Modeling; Feature Importance; Consumer Behavior.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

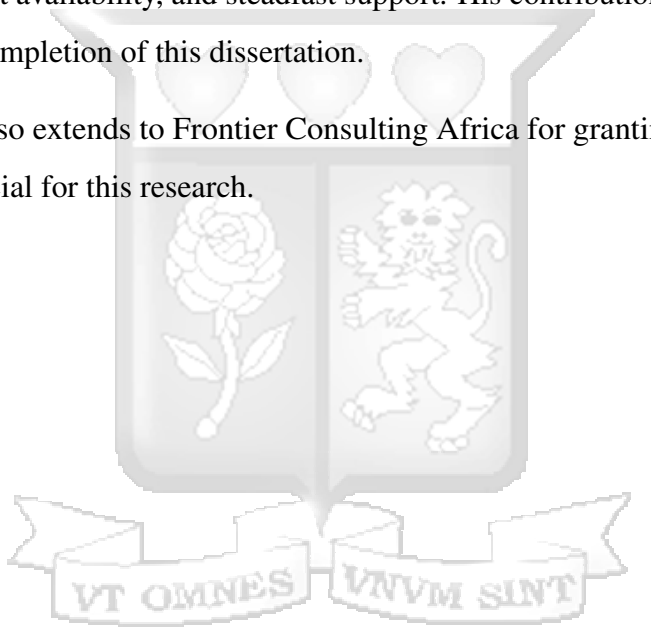
AI	Artificial Intelligence	**ANN**	Artificial Neural Networks
ARM	Association Rule Mining	**AUC**	Area Under the Curve
BIC	Bayesian Information Criterion	**BN**	Bayesian Networks
CFI	Comparative Fit Index	**CPDs**	Conditional Probability Distributions
DAG	Directed Acyclic Graph	**DT**	Decision Trees
FBNNs	Feed-Forward Back-Propagation Neural Networks	**FEA**	Facial Expression Analysis
GUM	Growing Up Milk	**HC**	Hill Climbing
HCPs	Health Care Practitioners	**K-NN**	K-Nearest Neighbors
LDA	Linear Discriminant Analysis	**LR**	Logistic Regression
MAPE	Mean Absolute Percentage Error	**MLP**	Multiple Layer Perceptron
MSE	Mean Square Error	**MSPs**	Mobile Service Providers
PNN	Probabilistic Neural Networks	**RF**	Random Forest
ROC Curve	Receiver Operating Characteristic Curve	**RMSEA**	Root Mean Square Error of Approximation
SML	Supervised Machine Learning	**SRMR**	Standard Root Mean Squared Residual
STG	Survey To Go	**SVM**	Support Vector Machines
XAI	Explainable Artificial Intelligence		

Acknowledgement

I wholeheartedly express my profound gratitude to God for His continuous provision and grace throughout the duration of this project. I am also immensely thankful for the unwavering support of my family, including my parents and siblings, and my friends, whose selfless assistance was invaluable.

I am particularly indebted to my supervisor, Prof. Bernard Omolo, for his exceptional guidance, consistent availability, and steadfast support. His contributions were instrumental to the successful completion of this dissertation.

My appreciation also extends to Frontier Consulting Africa for granting me permission to utilize the data crucial for this research.



Dedication

Dedicated to God, for His divine favor and sustained blessings. To my parents, siblings, and friends, for their profound influence, endless encouragement, and the sacrifices they've made for my pursuit of knowledge.

Definitions of Terms

- **Class Imbalance:** A condition in classification tasks where the number of observations in one class is significantly lower than in other classes (Japkowicz and Stephen, 2002).
- **Infant Formula:** A manufactured food designed and marketed for feeding to babies and infants under 12 months of age (Fomon, 2001).
- **Growing Up Milk:** Growing up milks (GUM) are milk-based drinks with added vitamins and minerals intended for children aged 12–36 months. (Walton and Flynn, 2013).
- **Junior Milk:** Junior milk, specifically for the 1-3 year age range, refers to formulas or milk drinks formulated to meet the nutritional needs of toddlers (Cavalcanti et al., 2024).
- **Integrated Approach:** An integrated approach to machine learning involves combining multiple machine learning techniques or models to achieve a specific goal, often resulting in improved performance or a more robust solution than using a single model (Viveros et al., 2024).

Chapter 1

Introduction

1.1 Background to the Study

The practice of consumer behavior looks into different aspects of what products and services individuals purchase, the driving factors behind the decision making the decisions to purchase, the sources from which the purchases are made, the frequency of these transactions and related aspects (Ramshitha and Manikandan, 2013).

As the market expands, different brands of products and services continue to emerge into the market to meet customers' demands. This inadvertently thus creates competition with each brand's intention to stay current and competitive in the market. One of the key business interests with brands in the market is the competitive environment and its measurement. Data that measures performance, retention and brand loyalty become important information in understanding subsequent repeat purchase behavior or the lack thereof.

With consumer preferences shifting over time and becoming volatile, it becomes a growing concern to these businesses on how to particularly measure consumer affinity to selecting their brand. Brand choice therefore becomes a key part of marketing and market research analysis. Every decision that consumers make on brand choice is usually based on certain drivers linked to their preferences. To better understand and predict consumer this consumer choice behavior, market research analysts apply models that reflect how people actually make choices and the exact choices they make based on a number of factors. The goal usually is to create accurate tools that help forecast what customers prefer. Brand choice has therefore become an essential tool in business and market research, especially for understanding customer behavior and decision-making (Papulova and Gazova, 2016).

The concept of customer choice modeling dates back to 1973, when McFadden introduced conditional logit analysis. Traditionally, researchers have relied on logit and probit models to classify customers based on their choices. However, because consumer behavior is complex and nonlinear, these traditional models often fall short in providing highly accurate predictions (Shahrabi and Khameneh, 2017).

To overcome these limitations, artificial intelligence (AI) techniques—such as Artificial Neural Networks (ANN), have gained traction in recent years. These advanced methods excel at handling complex patterns in data, offering greater flexibility and improved accuracy. Unlike other models, these AI-driven techniques capture nonlinear relationships without relying on rigid assumptions, making them powerful tools for predicting consumer brand choice (Duan and Xiong, 2015).

Deep learning, a machine learning subset, has transformed various industries by enabling the processing of large amounts of data through complex neural networks (Sarker, 2021). Its capacity to detect patterns in unstructured data sources, such as customer comments, social media trends, and web surfing behavior, makes it incredibly powerful (Chaudhary et al., 2021). By leveraging deep learning, market researchers and analysts can gain insight into consumer preferences and predict buying habits with a high degree of accuracy. This ability to handle and interpret large volumes of data allows companies to create adaptive, customized marketing strategies that keep pace with shifting consumer trends (Haleem et al., 2022). These shifting consumer trends include customer preference to a brand or service and their likelihood to churn by their reduced interest in their brand. Customer behavior is nonlinear, thus statistically difficult to forecast. To counter this, researchers have often used machine learning models like Randon Forest and ANN, which are efficient and appropriate for real-time optimization (Ahmad et al., 2017).

ANN models have gained popularity for brand loyalty and consumer choice prediction. Studies comparing ANN to other ML techniques suggest that ANN performs better because it is fault-tolerant, resistant to noise, and generally more robust (Ahmad et al., 2017). However, ANN suffers understanding cause-and-effect relationships (Hsu et al., 2006). To address this,

several approaches of combining and integrating with models that excel at cause and effect relationships have been applied, although in minimal cases.

Bayesian Networks provide a probabilistic framework that maps out relationships between variables,to illustrate their dependencies (Wang and Wu, 2018). This is highly useful in market research as decision-making by consumers across different brands is usually uncertain. With the application of BNs, businesses can quantify probabilities of various outcomes, which improves their decision-making ability.

There is value in bringing deep learning and BNs together, as both of these complement each other to create an improved predictive system. Deep learning is best for data sets that are complex, but BNs offer a structured way of reasoning in uncertain situations. They can then enable businesses make more strategic data driven decisions, consequently ensuring smart, responsive marketing campaigns that are responsive to endlessly changing consumer habits. It allows for a more interpretable of understanding complex relationships in a dataset (Meek et al., 2002).

This study added to the existing body of knowledge by applying an integrated model BN-ANN model that can be used in prediction of brand choice. The model was applied to provide consumer brand choice prediction. In this context, we sought to understand different dependencies into relationships of factors affecting choice of recommendation of infant formula by Health Care Practitioners (HCPs). Although some studies have previously investigated customer loyalty in brand choice, and the prediction of it none has examined determinants' impact by integrating ANNs and the BN. The structural model reveals dependencies among determinants in enhancing knowledge of customer loyalty thereby contributing knowledge into the relationship between certain priority factors and brand selection, specifically of infant formula.

Furthermore, in this study, brand loyalty was defined as a consumer's evaluations and behavioral intentions of the likelihood of purchasing a particular brand (Yi et al., 2022). Consequently, positive consumer word-of-mouth and brand recommendation influenced consumer loyalty towards a brand (Gounaris and Stathakopoulos, 2004). The HCP recom-

mendation of infant formula to mothers, in this study, thereby became the basis of knowledge of consumer brand preference.

1.2 Statement of the Problem

In the highly competitive product and service landscape of consumer products and services, businesses strive to understand and predict consumer loyalty by understanding if their product is top on consumers' minds (Mat Dawi et al., 2025). The challenge lies in forecasting future purchasing behavior, including repeat purchases and churn, given the nonlinear and complex nature of consumer decision-making (Basal et al., 2025). It becomes a greater challenge when certain businesses produce different products in the same product space and want to understand the health of the performance of their different brands especially when the purchase of one is likely to be dependent by the selection of another. An example is baby feed substitutes where the preference of a brand by a mother for their child at infant stage may or may not be similar to the Growing Up Milk (GUM) brand they would prefer for their child.

This study aimed to tackle this challenge by combining ANN and BN through sequential integration, with the goal of increasing understanding of determinants affecting consumer loyalty with regard to infant formula choice among HCPs. Even though research exists exploring factors impacting customer loyalty, none of it has combined ANN and BN in the sector of market research.

1.3 Research Objectives

1.3.1 General Objective

The main objective was to apply an integrated model combining ANN and BN to identify the key determinants influencing HCPs brand selection and loyalty towards infant formula

products. The study aimed to enhance the understanding of consumer behavior in this context by revealing the dependencies between factors affecting the recommendation of infant formula to mothers, ultimately providing actionable insights for businesses to refine their marketing strategies and improve customer retention in a competitive market.

1.3.2 Specific Objectives

1. To apply an integrated predictive model using ANN and BN to identify the impact of various factors on HCPs brand recommendation behavior in the infant formula market.

2. To assess the performance of the integrated model onto a similar data set for generalizability.

1.4 Justification of the Study

In the competitive consumer market with different brands and offerings of services, it is crucial to know why consumers prefer one brand to another; it is also essential for the producers of these goods and services to anticipate which brand is favored.

As consumer behavior continues to change, especially in not only healthcare but also other sectors, businesses need to be able to anticipate these changes in order to secure brand loyalty and reduce the chances of losing customers. This study was important because it explores a new approach—combining ANN and BN—to better understand and predict what drives HCPs' decisions and loyalty to infant formula brands. Cumulatively, we anticipated that the model present a truer comprehension of the reasons HCPs select and remain loyal to specific brands.

1.5 Significance of the Study

This study is expected to guide market analysts teams dealing in advising business strategy to have good knowledge of affinity of their products or services to their targeted consumers. To the consumer, this helps in ensuring a competitive market with different brands looking to outdo each other on account of quality and consumer preferences.

The results will be disseminated through the Strathmore University website and Strathmore University library, authorized by the administrator of SU+@Strathmore university.

1.6 Expected Output

In this study, we expected to achieve the following results:

1. An integrated BN-ANN model that outperforms both the standalone BN and ANN.
2. Identify conditional dependencies that contribute to selection of specific infant formula brand choice.

Chapter 2

Literature Review

2.1 Introduction

This section provides an overview of prior research conducted on predicting brand choice and preference. It looks into how ANN and BN have previously been used to predict brand preference. The literature presented in this paper serves as the foundational basis for this study, establishing the groundwork for the research topic. Brand loyalty is defined as a repeat purchase or the ability to recommend services or products (Mohlala and Bankole, 2022).

2.2 Models

2.2.1 Machine and Deep Learning Models

Machine learning is a broad and interdisciplinary field encompassing information technology, statistics, probability, artificial intelligence, psychology, and numerous other domains (Dunjko and Briegel, 2018). Within machine learning, solutions to various problems are achieved by constructing a model that effectively represents a given data set. At its core, machine learning revolves around the development of algorithms that enable computers to acquire knowledge. Learning, in this context, involves the identification of statistical regularities and patterns within data (Ramineni et al., 2024).

Electrodermal activity (EDA) and facial expression analysis (FEA) are two neurophysiological tools that were used by Marques et al. to predict consumer preferences for advertisements. The first goal was to determine which physiological characteristics were most important in

determining ad preferences. The second goal was to create a predictive system that used ML and explainable artificial intelligence (XAI). This new method filled the gap in the literature by combining ML algorithms.

The RF model achieved an accuracy of 81%, precision of 84%, recall of 79%, and F1-score of 81%, outperforming other algorithms such as k-Nearest Neighbors (kNN) and Support Vector Machine (SVM). Important characteristics that were found to be crucial in predicting customer preferences included attention, engagement, joy, and disgust. The results showed that physiological metrics are a useful tool for marketers to develop customized strategies and can accurately predict ad preferences.

The study did point out some limitations. First, there were only 37 participants in the sample, which limited how broadly the results could be applied. Furthermore, the study's primary focus was on advertisements for cosmetics, which made extrapolating findings to other industries difficult. Finally, the results might have been affected by uncontrollable external factors like the length of the advertisement and its visual components. These findings highlighted how combining neurophysiological measurements whilst using machine learning can improve marketing tactics. In order to confirm the relevance of its conclusions, the study also highlighted the necessity of larger sample sizes and more extensive research in a range of industries.

In order to determine the most important factors influencing sustainable jewelry purchases and to apply predictive models to analyze customer behavior, Munde and Kaur focused on using ML techniques to predict consumer purchasing behavior in the sustainable jewelry sector. The research examined variables like gender, age, income, marital status, educational attainment, buying preferences, online purchase drivers, and sustainable practices.

The study compared a number of ML classifiers, including Naïve Bayes, SVM, Linear Discriminant Analysis (LDA), KNN, and Decision Tree (DT), as well as ensemble methods like RF, AdaBoost, and Bagging Classifiers. Notably, the RF Classifier was the best-performing algorithm with 100% accuracy on the training dataset and 53% accuracy on the test dataset, demonstrating superior performance compared to others and making it a robust predictive model for sustainable jewelry purchases. Nevertheless, the paper identified some limitations,

including the small sample size (65 instances) and the study's primary focus on specific variables, omitting broader external factors that would influence purchasing behavior.

By addressing the shortcomings of prediction techniques, Liu et al. sought to improve the forecasting of customer repurchase behavior on e-commerce platforms. The goal was to create a reliable system by integrating the XGBoost DT model with logistic regression (LR).

The results showed that the hybrid XGBoost model performed significantly better than individual models, achieving improved classification accuracy and robustness; the XGBoost model successfully balanced positive and negative samples using optimized sampling strategies, improving its predictive capabilities; and the model fusion approach produced more robust predictions, with the fused model's Area Under the Curve (AUC) values consistently surpassing individual models. Nevertheless, the study identified a number of limitations, including: although the hybrid XGBoost model improved performance, its computational complexity presented scalability issues; second, the study primarily focused on behavioral data without integrating external factors like customer demographics or market trends; and third, the lack of diverse datasets limited the model's generalizability to broader e-commerce settings.

2.2.2 Application of Bayesian Networks in Predicting Brand Preference

The goal of Kang et al. was to apply a unique model for user-preference-based multimedia content recommendations. The goal was to employ BN to learn and reason by capturing user preference trends over time. This method solved the problem of out-of-date user history data predominating prediction models by enabling real-time learning without the need for extensive data collecting. Adaptive and dynamic forecasts of future preferences based on past consumption were made possible by the model's introduction of a weighting mechanism to emphasize current data.

Using actual TV viewing data gathered from 2,000 households, the trial findings showed how effective the suggested strategy was. The algorithm demonstrated better accuracy in forecasting future preferences and surpassed conventional methods in predicting TV genre

preferences. The study confirmed the significance of dynamic weighting methods by finding that users' recent viewing patterns had a greater impact on future forecasts. The study did point up certain shortcomings. Notable were the weighting mechanism's computing complexity and the need for enough data to guarantee prediction dependability.

In order to find inefficiencies that could be used to one's advantage, Constantinou investigated how BNs can simulate risk and uncertainty in football betting markets. The uses probabilistic reasoning to show how BN-based prediction models may predict match outcomes with an accuracy of up to 12.4% better than standard betting odds. This implied that not all significant factors were taken into consideration by traditional bookmaker models, which opens the door for data-driven tactics to profit from undervalued wagers.

According to the findings, BNs outperformed conventional bookmaker odds-based models, which had a predicted accuracy of 66.2%, by achieving 78.6%. Additionally, when using Bayesian predictions, the inefficiencies in football betting markets produced profit margins of up to 8.3%, as opposed to the break-even or negative returns seen with naive betting tactics. BNs enhanced risk assessment by lowering variance in prediction errors by 15.7% through the structure of dependencies between variables including team form, player performance, and previous match data. The study also showed that odds inefficiencies continued to exist in 24.5% of the examined games, suggesting that knowledgeable bettors may methodically take advantage of these flaws to produce steady profits.

Notwithstanding its efficacy, the strategy has a number of drawbacks. Since it is challenging to include real-time variables like injury reports, referee biases, and weather conditions in the model, data availability was a major obstacle.

The study's approach has significant implications for consumer markets' brand prediction. By organizing latent variables like brand affinity, purchase likelihood, and product associations, BNs can forecast consumer preferences in a similar way to how they simulate uncertainty in betting markets. Inefficiencies in consumer data give firms a competitive edge through focused marketing and price optimization, much how inefficiencies in football gambling markets resulted in an 8.3% profit margin. This demonstrates how useful Bayesian modeling

is for comprehending and forecasting human decision-making in unpredictable situations, such as in brand marketing or gambling.

In order to model client repurchase intention and comprehend product linkages in the fast food industry, Taewijit and Theeramunkong investigated a probabilistic framework. The study offered insights on how to forecast and use consumer behavior for strategic marketing through the use of BNs, compared with other predictive models and Association Rule Mining (ARM) to mine product associations. The main objective was to create a methodical strategy that enables companies to predict which clients were most likely to make repeat purchases and which product combinations were most usually bought together. Based on the insights, this helps businesses to increase customer retention tactics, optimize promotions, and strengthen cross-selling.

The findings demonstrated that BNs outperformed other machine learning models in predicting customer repurchase intentions, with an accuracy of 83.65%. Additionally, the study discovered that adding latent features to prediction models greatly enhanced performance; KNN models saw a 16.74% boost in accuracy this way. Strong co-purchase patterns were also found using product association analysis, with chocolate, tea, and yogurt regularly showing up as complementing goods in fast food purchases. These results demonstrate how companies may tailor marketing campaigns and promote recurring business by utilizing probabilistic models.

The study however had several limitations in spite of these findings. The problem of data imbalance necessitated the use of resampling methods like SMOTE and SpreadSubSample, which create biases. Furthermore, even if latent factors offered insightful information, they might not fully represent the complexities of consumer behavior, particularly when it comes to outside influences like seasonal trends or promotions.

2.3 Application of Artificial Neural Networks in Predicting Brand Preference

Stoll assesses the performance of ANNs in predicting brand choice whilst comparing to conventional statistical models. The study's main focus was to figure out if ANNs can reflect non-linear relations in consumer behaviour better than LR. The research also evaluated multiple network structures and learning techniques as a way of analyzing their predictive abilities concerning certain consumer attributes and whole brand selection.

The results showed that ANN outperformed with a 9.2% accuracy difference in predictions on average. It proved the capability to analyze the customer decision-making patterns. The research also discovered that using the multiple layer perceptron (MLP) model with the backpropagation technique achieved the highest accuracy, since the datasets were very large and the variables were many. The study however identified limitations. The issue of interpretability was discussed, as the ANNs were described as black-box models hence difficult to explain the reason for the specific predictions.

Deliana and Rum aimed at examining and comparing efficiency of ANN prediction by comparing with LR and DT. The focus of this study was to clarify that ANNs can demonstrate the nonlinear and complex combinations of consumer behaviors and loyalty better than traditional methods do.

It was pointed out that the ANN predicted consumer loyalty with an accuracy of 87.3%, which is higher compared to 74.5% of LR and 79.8% of the DT. The analysis showed that customer satisfaction, the perceived value and the emotional attachment to the brand were the factors that determined most of the score of 34.2%, 27.1% and 19.6% respectively. Moreover, the study concluded that very satisfied clients were twice more likely to continue their loyalty, thus proving the centrality of a wonderful customer experience in building lasting brand engagement. Notwithstanding the fact that the outcomes were promising, the research admits to some drawbacks. The more accurate neural networks, demand long training times, which in turn make the computational costs at least 45% higher compared to those in LR models.

Patrick and Nasiru explored the use of Feed-Forward Back-Propagation Neural Networks (FBNNs) to predict customer preference for mobile service providers (MSPs) based on various service related factors. The goal was to show how ANNs can be used to model and forecast customer choice, reduce customer acquisition cost and enable service providers to tailor their offerings.

The ANN was designed with 20 inputs, 1 hidden layer of 13 neurons and 1 output, where binary classification (1 = MTN, 0 = other MSP) was used to determine customer preference. 240 data points were used, split into 60% training, 20% validation and 20% testing sets. Training was done using Levenberg-Marquardt algorithm which balanced efficiency with accuracy but had memory limitations. Training stopped when validation error increased for 6 consecutive iterations at iteration 11 to prevent overfitting.

The results showed that the FBNN had an overall regression fit of 0.9017, a strong relationship between predicted and actual customer choices. Individual dataset fits were slightly lower but retraining the model with adjusted weights improved performance. Mean square error (MSE) remained low throughout training which meant the network generalizes well without overfitting. The study was not short of some limitations. First, Levenberg-Marquardt algorithm improved efficiency but is memory intensive so not suitable for large scale real time applications.

2.4 Application of BN-ANN Models

Siahgoli et al., combined Bayesian classification and ANNs in lithofacies classification on well and seismic data from Bahrgansar oilfield. The aim of the study was to increase the accuracy of lithofacies prediction, using a combination of probabilistic and machine learning approaches.

The results indicated that bayesian classification together with ANN can be used for improvement of lithofacies classification accuracy. The hybrid method yielded an overall accuracy of 89.4% as compared to individual accuracies where the Bayesian classification achieved

75.8% Accuracy and ANN 83.6% . As determined from sensitivity analysis, the most significant factors in predicting lithofacies were gamma ray logs, density logs and seismic impedance. The hybrid model also improved misclassification rates especially between sandstone and shale lithofacies, both of which were difficult to discriminate against each other using petrophysical properties.

The study also identified several shortcomings. To start, the accuracy of Bayesian classification is highly context-dependent on availability and reliability of priors which can be difficult to determine in new, unfrequented exploration areas. Siahgoli et al. highlights there is a greater computation simplification occurred when two models are combined, which could make real time application in reservoir management challenging.

2.4.1 Conclusions

Literature discusses the BNs and ANNs in predictive analytics, especially on brand preference modeling. ANNs, although accurate, are open to the limitation of their non-interpretability. On the other hand, BNs have defined probabilistic reasoning and are sensitive to a good understanding of prior in markets that change rapidly, this is hard to obtain.

The present work is a novel way to apply integrated use of Bayesian Networks and Artificial Neural Networks at prediction of choice of infant formula by HCPs. Combining BN with ANN capability of structured reasoning and pattern recognition in order to maximize prediction power, explainability and decision speed for a wide range of consumer analytics and other.

2.4.2 Current Research

This dissertation, sought to apply an integrated BN-ANN model and will add to the existing body of knowledge a model that works by including probabilistic reasoning in prediction of brand choice in the model. The BN and ANN models have been applied in various fields such as mobile communication networks, health, tourism , and engineering to mention a few.

No study has incorporated this intergration in prediction of brand choice, specifically in the sector of early childhood nutrition.

Chapter 3

Methodology

3.1 Introduction

This chapter comprehensively details the dataset utilized and elaborates on the methodologies employed to model consumer preferences for infant formula in Djibouti. Specifically, it discusses the performance of BN, ANN, and the integrated BN-ANN hybrid approach. This section is dedicated to providing statistical inference derived from these models, therefore showing patterns inherent in infant formula choices.

The initial sections of this chapter tackle the various stages of model development. This begins with the essential data pre-processing steps, including the conversion of categorical variables into appropriate factor levels, the discretization of continuous variables to suit the BN architecture, and the handling of inherent class imbalance through targeted sampling techniques. Subsequently, the chapter details the application of the standalone BN model, specifically outlining its structure learning via the tabu search algorithm.

This is followed by the construction of the standalone ANN model, based on a feed-forward architecture. This then follow with formulation of the Hybrid BN-ANN model. The main aim of this process is to develop a Hybrid BN-ANN model that accurately and efficiently classifies infant formula preferences among consumers in Djibouti.

The concluding section of this chapter is dedicated to elaborating how evaluation of the models' performance, employing different metrics to validate their predictive capabilities and generalizability

3.2 Data Description

To conduct this comprehensive study, a dataset, focusing on factors influencing the recommendation of infant formula by HCPs to mothers in Djibouti. This dataset comprised 119 observations, each representing dynamics of HCP recommendations and maternal choices. The entire dataset was sourced from Frontier Consulting Africa, a reputable firm that conducted the underlying consumer survey in Djibouti. This collaboration ensured the data's authenticity and direct relevance to the research objectives. The data itself is cross-sectional, meaning it captures a snapshot of consumer choices and related attributes at a specific point in time, precisely in 2022.

3.2.1 Data Collection and Participants

These HCPs served as our primary study participants. The recruitment process involved purposive and snowballing of the participants.

The table 3.1 below gives an overview of the variables used for this study.

Table 3.1: Description of Variables Used in the Study

Variable Name	**Type**	**Description / Levels**
TOWN	Categorical	Consumer's geographic location in 3 cities
HCP	Categorical	Healthcare Practitioners (5 levels)
Public_Practice	Numeric	% Time spent in public practice
Private_Practice	Numeric	% Time spent in private practice
GUM	Categorical	Growing Up Milk type (3 levels)
Junior_Milk	Categorical	Junior Milk type (3 levels)
Priority_Feature	Categorical	Prioritized formula feature (3 levels)
Similar_to_breast_milk	Categorical	Perceived similarity to breast milk (7 levels)
Readily_available	Categorical	Product availability (7 levels)
Complete_range_of_milk_products	Categorical	Importance of full product range (7 levels)
Affordable	Categorical	Perceived affordability (7 levels)
Good_reviews	Categorical	Influence of positive reviews (7 levels)
Well_tolerated	Categorical	Infant tolerability (7 levels)
Adapted_to_nutritional_requirements	Categorical	Nutritional adaptation (7 levels)
Adapted_for_immunity	Categorical	Perceived immunity benefits (7 levels)
Proven_clinical_benefits	Categorical	Importance of clinical benefits (7 levels)
Children_like_taste	Categorical	Children's taste preference (7 levels)
Innovative_brand	Categorical	Perceived brand innovativeness (7 levels)
Aware_of_First_Days_Concept	Categorical	Awareness of "First Days Concept" (7 levels)
Infant_Formula	**Dependent Categorical**	Outcome: Infant formula type chosen (3 levels)

3.2.2 Dataset Characteristics

The final dataset that was used for this dissertation included variables that captured a range of information, including the HCPs, their perceptions of factors influencing infant feeding choices, and their patterns of recommendation.

3.2.3 Description of Key Variables

The following table provides a detailed description of the principal variables used in the study, their types, and their significance in understanding infant formula choice and recommendation patterns.

To provide a foundational understanding for interpreting our findings, we've outlined the principal variables used in our study. These variables, collected through surveys with HCPs, capture a range of information critical to understanding infant formula choice and recommendation patterns in Djibouti.

The type of infant formula was our primary outcome variable. It's a nominal categorical variable representing the specific type or brand of infant formula chosen or recommended, with anonymized levels '1', '2', and '3'. This variable is the central focus of our study; understanding what leads to its selection was our core objective.

The type of junior milk was another nominal categorical variable, indicating the type of junior milk mothers use, often seen as a related product to infant formula. Like the type of infant formula, its levels are anonymized as '1', '2', and '3'.

The HCP variable, also five level nominal and categorical, whic specifies the type of Health-care Professional providing advice or recommendations. This variable directly helps us assess the role and influence of different medical professionals on feeding choices.

Time spent in public practice was a continuous variable with an estimated percentage time spent in public institutions, in practice, while the time spent in private practice describes the same, but of time spent in private institutions.

Growing Up Milk (GUM) is a nominal categorical variable with three levels. The variable captures the GUM brand that was recommended to the mother for their child, aged six to twelve months.

The priority feature is a nominal categorical variable identifying the most important feature HCPs consider when advising on infant formula or junior milk. Its levels are two; 'Product features' and 'Company image & characteristics'. This variable was crucial for pinpointing consumer priorities in selecting an overall brand.

Several variables capture specific perceptions about the formula itself: similarity to breast milk assesses the perceived naturalness of the formula; an infant formula being readily available evaluates how important accessibility is in formula selection; completeness in offering a range of products investigates how brand loyalty; affordability; good reviews; tolerance by the child; adaptation to nutritional requirements; adaptation to immunity; proven clinical benefits and taste.

Finally, town is a nominal categorical variable representing the geographical location where the survey participant practices. Its specific town names are anonymized if necessary. This variable allowed us to explore geographical variations and contextual differences in choices and recommendations. Additionally, awareness of the concept of 1000 days indicates whether the HCP knows about the "First 1000 Days" concept for child nutrition, helping us assess the impact of nutritional knowledge on feeding decisions.

This detailed description of our dataset and variables provides the essential understanding needed to interpret our model's findings and their implications for policy and marketing strategies.

3.2.4 Data Diagnostics

For the application of the models, the dataset was subjected to a preparation process. This included: i) assessment of missing values; ii) transforming categorical variables into suitable factor levels for modeling; and iii) discretizing continuous variables, which was crucial for

their integration into the BN structure. This latter step, while necessary for the BN. All data analysis and model development procedures were conducted using the R statistical software, specifically R version 4.4.1 (2024-06-14), leveraging specialized packages for BNs and ANN; bnlearn, neuralnet providing the necessary algorithmic tools for model implementation, training, and evaluation.

Firstly, an assessment of missing values was conducted across all variables. Secondly, categorical variables within the dataset, representing various qualitative aspects of infant formula preference, are transformed into suitable factor levels. Thirdly, continuous variables are subjected to a process of discretization. BNs operate on discrete or categorical variables for conditional probability calculations, making this step important for constructing and training the BN component of our integrated model.

Figure 3.1 summarizes the process workflow:

Assessment of Missing Data

An initial step in the data preparation process involved an assessment of missing values. This comprehensive review confirmed that the dataset complete, with no missing observations detected across any of the variables. This finding eliminated the need for imputation or complex missing data handling techniques.

Outlier Detection through Boxplot Analysis

To investigate the presence of outliers across the continuous variables, the boxplot, also known as the box an whisker plot was used as a graphical diagnostic tool. This method, shows a visual summary of the distribution of the numerical variable demonstrating central tendency, spread and the presence of extreme values (Majaw and Ahmed, 2023).

The process for outlier detection via boxplots was systematic:

Construction of boxplots where for each continuous variable in the dataset, a boxplot was generated. The central box in the plot visually represents the interquartile range (IQR),

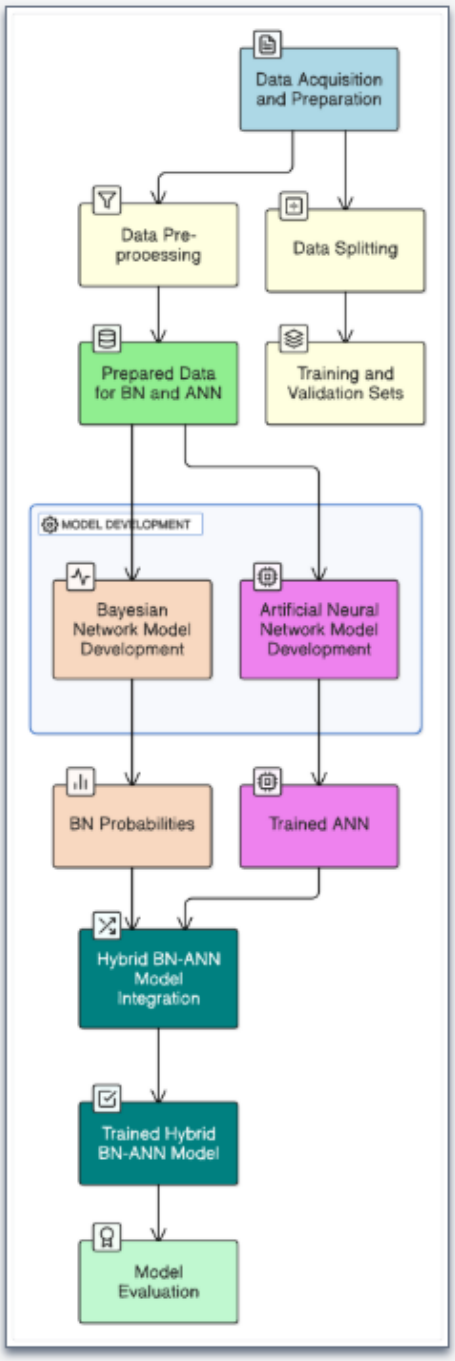


Figure 3.1: Process Flow of the Integrated BN-ANN Model Application and Evaluation

spanning from the first quartile (Q_1, the 25th percentile) to the third quartile (Q_3, the 75th percentile). The line inside the box indicates the median (Q_2, the 50th percentile). The Interquartile Range (IQR) is mathematically defined as:

$$\text{IQR} = Q_3 - Q_1 \tag{3.1}$$

Definition of outlier boundaries by assessing the "whiskers" of the boxplot how they typically extend to the most extreme data points within $1.5 \times$ IQR from the respective quartiles. According to Majaw and Ahmed, any data points falling outside these whiskers are statistically considered potential outliers. Specifically, a data point (x) is flagged as an outlier if it satisfies either of the following conditions:

$$x < Q_1 - 1.5 \times \text{IQR} \tag{3.2}$$

or

$$x > Q_3 + 1.5 \times \text{IQR} \tag{3.3}$$

These boundaries are often referred to as the lower and upper fences, respectively.

Each boxplot was then assessed. Data points identified as outliers by the $1.5 \times$ IQR rule are commonly plotted individually as distinct markers beyond the whiskers. This clear visual representation allows for immediate identification of observations that significantly deviated from the central mass of the data distribution.

Assessment of Class Imbalance

The exact percentage of observations belonging to each specific infant formula choice, as represented by the infant formula, was calculated. This provided numerical understanding of each class's prevalence within the overall dataset.

Secondly, to complement this numerical quantification, a bar plot was generated. where each bar corresponded to a distinct infant formula type, with its height directly reflecting the calculated percentage of observations for that particular class.

The chosen method for balancing the target variable, was oversampling. This technique works by randomly replicating observations from the minority classes until their counts match that of the majority class. This process effectively expands the representation of the less frequent categories in the data, thereby providing the learning algorithms with sufficient examples from all classes to discern patterns accurately and reduce bias (Sáez et al., 2016).

Discretization of Continuous Variables

As BNs are fundamentally structured to model relationships among discrete or categorical variables, converting continuous measurements into distinct bins was necessary prerequisite for their integration into the network structure.

For this study, two key continuous variables, namely Public Practice and Private Practice, were identified for discretization. The chosen method for this transformation was quantile discretization. This approach divides the range of a continuous variable into a specified number of bins, ensuring that each bin contains an approximately equal number of observations (Liu et al., 2002). According to Kotsiantis and Kanellopoulos, this method is particularly effective as it handles skewed distributions well and creates bins with relatively balanced frequencies, which can be beneficial for subsequent probabilistic modeling.

Mathematically, for a continuous variable X to be discretized into k bins, quantile discretization identifies $k-1$ cut points (thresholds) based on the variable's quantiles. Let $F^{-1}(p)$ denote the p-th quantile (or $100p$-th percentile) of the distribution of X. The cut points C_j for $j = 1, \ldots, k-1$ are defined as:

$$C_j = F^{-1}\left(\frac{j}{k}\right) \tag{3.4}$$

These cut points define k intervals (bins). An observation x of the variable X is then assigned to a bin B_i based on which interval it falls into. Specifically, for $k = 5$ bins as applied in this study for variables like time spent in Public and Private Practice, the four cut points correspond to the 20^{th}, 40^{th}, 60^{th}, and 80^{th} percentiles of the variable's distribution. The bins are then defined as follows:

$$B_1 = \{x \mid x \leq C_1\} \tag{3.5}$$

$$B_2 = \{x \mid C_1 < x \leq C_2\} \tag{3.6}$$

$$B_3 = \{x \mid C_2 < x \leq C_3\} \tag{3.7}$$

$$B_4 = \{x \mid C_3 < x \leq C_4\} \tag{3.8}$$

$$B_5 = \{x \mid x > C_4\} \tag{3.9}$$

where $C_1 = F^{-1}(0.20)$, $C_2 = F^{-1}(0.40)$, $C_3 = F^{-1}(0.60)$, and $C_4 = F^{-1}(0.80)$.

3.2.5 Analysis of Categorical Variables: Chi-Square Test of Independence

To assess the statistical association between the categorical independent variables and the dependent variable, Pearson's Chi-Square Test of Independence was employed. This non-parametric statistical test is particularly suitable for analyzing the relationship between two categorical variables, determining whether there is a significant dependence or if they are independent of each other. The test yields a Chi-Square (χ^2) statistic, which quantifies the difference between the observed frequencies in a contingency table and the frequencies that would be expected under the assumption of independence.

For each selected categorical independent variable, a contingency table was executed, creating a two-way table that cross-tabulated the frequencies of categories for the independent variable against the categories of the target variable.

Following the construction of contingency tables, the application proceeded on each table. For this test, the underlying null hypothesis (H_0) posits that there is no association between the two

categorical variables, while the alternative hypothesis (H_1) suggests that an association exists. Finally, P-Value Determination was conducted; alongside the χ^2 statistic, a corresponding p-value was generated for each test. This p-value indicates the probability of observing a relationship as strong as, or stronger than, the one observed in the sample data, assuming the null hypothesis (H_0) is true.

The results, comprising the χ^2 statistic and the p-value for each categorical variable, were compiled into a summary table. A p-value less than the predetermined significance level (typically $\alpha = 0.05$) was interpreted as evidence to reject the null hypothesis, indicating a statistically significant association between the respective categorical independent variable and the choice of infant formula.

The results, comprising the χ^2 statistic and the p-value for each categorical variable, were compiled into a summary table. A p-value less than the predetermined significance level (typically $\alpha = 0.05$) was interpreted as evidence to reject the null hypothesis, indicating a statistically significant association between the respective categorical independent variable and the choice of infant formula.

3.3 Data Analysis

Our study's work consists of various factors influencing the choice of infant formula recommended by HCPs to mothers in Djibouti. These observations are not randomly generated; they reflect complex, underlying processes and potential non-linear interdependencies among socio-economic, health, and recommendation-related variables. Comprehending these relationships and patterns is fundamental, as it enables the development of models that can accurately predict or classify the likely infant formula choice at a given point. Drawing upon established literature, the preferred classes of models that will be explored in this chapter are BNs (BN), ANN, and their integrated Hybrid BN-ANN Model.

3.4 Model Formulation

3.4.1 Data Partitioning for Model Development and Evaluation

To ensure development, evaluation, and reliable generalization capability of our predictive models, the prepared dataset was partitioned into distinct training and testing subsets. This step is fundamental in preventing overfitting, a phenomenon where a model learns the training data too well, capturing noise rather than underlying patterns, thereby performing poorly on unseen data (Kernbach and Staartjes, 2021). By training models on one portion of the data and assessing their performance on a separate, independent portion, we gain a more accurate measure of their true predictive power.

The partitioning process was executed using a stratified random sampling approach as it ensures that the proportional representation of each class within the target variable, infant formula, is accurately maintained in both the training and testing sets. This stratification is particularly important given the potential for class imbalance in real-world datasets, as it guarantees that all classes are adequately represented in both subsets, preventing the models from being biased towards majority classes during training or evaluation phases.

Specifically, the dataset was meticulously split into a 70% training set and a 30% testing set. The createDataPartition function from the caret package in R was employed for this purpose. To ensure reproducibility of this partitioning, a seed value of 42 was set prior to executing the split.

3.4.2 Bayesian Networks

A Bayesian network represents a joint probability distribution across a group of random variables U. Formally, a Bayesian network for U consists of a pair:

$$B = (G, \Theta), \tag{1}$$

The first element, G, represents a directed acyclic graph (DAG) where each node corresponds to a random variable $(X_1, \ldots, X_n)$ and the edges indicate the direct relationships among these variables. In this way, the graph G encodes our assumptions about independence – specifically, each variable X_i is independent of its nondescendants when conditioned on its parents in G. The second component, Θ, is the set of parameters that quantify the network. In particular, it includes the parameter

$$\theta_{x_i|\Pi_{x_i}} = P\left(X_i \mid \Pi_{X_i}\right), \tag{2}$$

which is defined for each possible value x_i of X_i and for every configuration Π_{x_i} of the parent set Π_{X_i} of X_i in G Friedman et al. (1997).

Building on our model formulation, we express the joint probability distribution over the eight random variables as the product of their local conditional probabilities, which reflects the dependency structure specified by our Bayesian network. In other words, the overall probability is obtained by multiplying the individual conditional probabilities associated with each variable. Formally, we have:

$$\begin{aligned}
P(X_1, \ldots, X_{12}) = {} & P(\texttt{TOWN})\, P(\texttt{HCP})\, P(\texttt{Public_Practice}) \\
& \times P(\texttt{Private_Practice})\, P(\texttt{GUM}) \\
& \times P(\texttt{Junior_Milk} \mid \texttt{TOWN}) \\
& \times P(\texttt{Infant_Formula} \mid \texttt{Junior_Milk}, \texttt{HCP}) \\
& \times P(\texttt{Priority_Feature} \mid \texttt{Infant_Formula}) \\
& \times P(\texttt{Similar_to_breast_milk} \mid \texttt{Infant_Formula}) \\
& \times P(\texttt{Readily_available} \mid \texttt{Infant_Formula}) \\
& \times P(\texttt{Complete_range_of_milk_products} \mid \texttt{Infant_Formula}) \\
& \times P(\texttt{Aware_of_First_1000_Days_Concept} \mid \texttt{HCP})
\end{aligned} \tag{3}$$

In our Bayesian network, each discrete node X_i represents a random variable with a finite set of values. For any such node, given a specific configuration of its parent nodes (denoted as

$\text{pa}(X_i)$), the conditional probability that X_i takes on a particular value x is quantified by the parameter $\theta_{x|pa}$. In other words, the relationship is formalized as follows:

$$P(X_i = x \mid \text{pa}(X_i) = pa) = \theta_{x|pa}, \tag{4}$$

where $\theta_{x|pa}$ are the maximum likelihood estimates of the multinomial parameters.

The overall joint probability distribution for the eight variables is decomposed into a product of simpler marginal and conditional probabilities. Each term in the product captures either the standalone probability of a variable or the probability of a variable given its parent(s). The joint distribution is expressed as follows:

$$\begin{aligned}
&P(\texttt{TOWN}, \texttt{HCP}, \texttt{Public_Practice}, \texttt{Private_Practice}, \texttt{GUM},\\
&\quad \texttt{Junior_Milk}, \texttt{Infant_Formula}, \texttt{Priority_Feature},\\
&\quad \texttt{Similar_to_breast_milk}, \texttt{Readily_available}, \texttt{Complete_range_of_milk_products},\\
&\quad \texttt{Aware_of_First_1000_Days_Concept})\\
&= P(\texttt{TOWN})\,P(\texttt{HCP})\,P(\texttt{Public_Practice} \mid \texttt{Junior_Milk})\\
&\quad \times P(\texttt{Private_Practice} \mid \texttt{Junior_Milk})\,P(\texttt{GUM} \mid \texttt{Infant_Formula})\\
&\quad \times P(\texttt{Junior_Milk} \mid \texttt{TOWN}, \texttt{Infant_Formula})\\
&\quad \times P(\texttt{Infant_Formula} \mid \texttt{Junior_Milk}, \texttt{HCP})\\
&\quad \times P(\texttt{Priority_Feature} \mid \texttt{Junior_Milk})\\
&\quad \times P(\texttt{Similar_to_breast_milk} \mid \texttt{Infant_Formula})\\
&\quad \times P(\texttt{Readily_available} \mid \texttt{Infant_Formula})\\
&\quad \times P(\texttt{Complete_range_of_milk_products} \mid \texttt{Infant_Formula})\\
&\quad \times P(\texttt{Aware_of_First_1000_Days_Concept} \mid \texttt{HCP})
\end{aligned} \tag{5}$$

Model Development

In this study, the BN model serves as a powerful probabilistic model designed to capture and represent the complex conditional dependencies among variables influencing infant formula choice. A BN consists of two primary components: a qualitative component, which is a Directed Acyclic Graph (DAG) representing the probabilistic relationships between variables,

and a quantitative component, comprising the conditional probability distributions (CPDs) for each node given its parents. The learning process for a BN involves determining both its optimal structure and its parameters. Bayesian networks are models that link variables with probabilities and utilize Bayes' theorem along with associated Bayesian learning algorithms to calculate posterior probabilities of outcome states.

The initial step in BN learning is structure learning. This process involves identifying the optimal DAG that best represents the underlying dependency relationships present in the dataset. This study adopted a score-based learning approach, where various candidate network structures are systematically evaluated using a predefined scoring function. The ultimate objective of these algorithms is to identify the graph structure G that maximizes this score for a given dataset D.

The primary scoring criterion used to guide this structure learning was the Bayesian Information Criterion (BIC). BIC is a statistical measure that judiciously balances the goodness-of-fit of a model to the data with a penalty for model complexity, favoring parsimonious models that adequately explain the observations (Aho et al., 2014). A lower BIC score indicates a superior model, suggesting a better balance between fit and parsimony. For a given graph structure G and dataset D, the BIC score is mathematically defined as:

$$BIC(G,D) = -2\log L(G,D) + d\log N \tag{3.10}$$

Here, $L(G,D)$ represents the maximum likelihood of the data given the graph structure G, d signifies the number of independent parameters in the model, and N denotes the total number of observations in the dataset. During the learning process, a penalization coefficient of approximately 2.64 was applied to balance model complexity.

To efficiently navigate the vast search space of possible DAGs and pinpoint the structure with the most favorable BIC score, two distinct heuristic search algorithms were applied: the Hill-Climbing (HC) algorithm and the Tabu Search algorithm. These algorithms operate by iteratively exploring the space of possible network structures, guided by the BIC score.

Following the learning of the BN structure from the refined dataset, the strength of the arcs within the learned network was assessed to understand the confidence and significance of each inferred dependency between variables. For each directed arc ($X \rightarrow Y$) present in the learned network, the function calculates a score that reflects how strongly the presence of that arc is supported by the observed data, relative to a model where that specific arc is absent.

The strength of the given arc is assessed by comparing the BIC score of the network that includes that arc against the BIC score of the network that does not include that arcin this manner; if G_1 denotes the network structure containing the arc $X \rightarrow Y$, and G_0 denote the identical network structure but with the arc $X \rightarrow Y$ removed (i.e., $G_0 = G_1 \setminus \{X \rightarrow Y\}$). The strength of the arc $X \rightarrow Y$ based on the BIC criterion can be formally represented as:

$$\text{Strength}(X \rightarrow Y) = BIC(G_0, D) - BIC(G_1, D) \tag{3.11}$$

As a lower BIC score indicates a better model fit and greater parsimony, a larger positive value for $\text{Strength}(X \rightarrow Y)$ implies that the inclusion of the arc $X \rightarrow Y$ significantly improves the model's overall BIC score, thus indicating stronger empirical support for the existence of that particular dependency. The resulting importance scores provide a quantitative measure for each arc within the network.

3.4.3 Artificial Neural Networks Model

An ANN model was applied to capture potentially non-linear and complex relationships within the dataset for predicting infant formula choice. ANNs are powerful machine learning algorithms inspired by the structure and function of the human brain, consisting of inter-connected nodes (neurons) organized into layers. Information flows forward through these layers, undergoing transformations at each neuron (Goel et al., 2023).

The specific ANN architecture employed in this study was a feed-forward neural network with a single hidden layer. This architecture is well-suited for classification tasks where the

underlying relationships between input features and the target variable may be non-linear, as is often the case in consumer choice modeling.

The training process for the ANN model, implemented on the was as follows:

Input Layer:

This layer received the values of the predictor variables from the pre-processed training data. These input features, representing factors affecting infant formula choice (e.g., taste, HCP, the amount of time a HCP spends in public practice, etc.), are denoted as $x_1, x_2, \ldots, x_n$, where n is the total number of selected predictor variables.

Hidden Layer:

Each neuron in the hidden layer computed a weighted sum of its inputs from the input layer, to which a non-linear activation function is then applied. For a neuron j in the hidden layer, its weighted sum z_j is given by:

$$z_j = \sum_{i=1}^{n} w_{ij} x_i + b_j \tag{3.12}$$

Here, x_i represents the value of the i-th predictor variable from the dataset, w_{ij} denotes the synaptic weight connecting the i-th input feature to the j-th neuron in the hidden layer, and b_j is the bias term specific to neuron j. This sum is then passed through a non-linear activation function, $f(\cdot)$ (e.g., a sigmoid or hyperbolic tangent function), to produce the neuron's output activation a_j:

$$a_j = f(z_j) \tag{3.13}$$

The non-linearity introduced by the activation function is critical for enabling the network to learn complex, non-linear patterns within the data.

Output Layer:

This final layer receives inputs from the hidden layer and produces the network's prediction for the infant formula class. For multi-class classification problems, the output layer typically employs a softmax activation function to convert the final weighted sums into a probability distribution over the possible infant formula choices.

The network learned by iteratively adjusting all connection weights (w_{ij}) and biases (b_j) through an optimization algorithm commonly backpropagation coupled with gradient descent to minimize a predefined loss function like cross-entropy loss for classification. This loss function quantified the discrepancy between the network's predicted probabilities for each infant formula type and the actual observed class labels.

In this study, the ANN model was specifically implemented using the `nnet` package in R. The model was trained on the training data with choice of infant formula designated as the target variable and all other remaining variables as predictors. The key training parameters specified were the number of neurons within the single hidden layer, set as 5. The selection of 5 neurons was based on common practices for initial model complexity, aiming for a balance between learning capacity and computational efficiency.

The maximum number of iterations, or epochs, set at 100, for the training algorithm. An epoch signifies one complete forward and backward pass through the entire training dataset, during which the network's weights are iteratively adjusted to minimize the loss (Tivive and Bouzerdoum, 2005).

The training process aimed to establish a robust and generalizable mapping between the various input features and the infant formula choices, enabling the ANN to accurately predict the target variable class on unseen data.

3.4.4 Integrated Bayesian Networks-Artificial Neural Networks Model

The BN model, once its structure and parameters are learned, serves as a powerful tool not only for generating probabilistic insights. In the context of developing an integrated Hybrid

BN-ANN Model, this step involves leveraging the BN to produce conditional probabilities that can then be fed as enhanced inputs into the ANN. This approach aims to capitalize on the BN's ability to model complex dependencies and propagate uncertainty, providing the ANN with richer, probabilistically informed features.

Bayesian Network Probability Estimation for Integrated Model Input

The process began by parameter learning which involves calculating the Conditional Probability Distributions (CPDs) for each node within the learned network structure. For discrete variables, these CPDs are typically represented as Conditional Probability Tables (CPTs), which are estimated directly from the frequencies observed in the trained data. The most common method for this estimation is Maximum Likelihood Estimation (MLE), where the probability $P(X_i = x_i \mid Pa(X_i) = pa_i)$ is estimated as the relative frequency of x_i occurring given the configuration pa_i of its parent nodes in the training data:

$$\hat{P}(X_i = x_i \mid Pa(X_i) = pa_i) = \frac{N(X_i = x_i, Pa(X_i) = pa_i)}{N(Pa(X_i) = pa_i)} \tag{3.14}$$

where $N(\cdot)$ denotes the count of occurrences in the training data.

With the BN parameters estimated, the network becomes ready for performing conditional probability queries, which estimate the probability of certain events given specific observed evidence. This is mathematically expressed as $P(\text{Event}|\text{Evidence})$. In the context of this study, a query estimates the probability of the target variable (`Infant_Formula`) taking on each of its possible class values, given the observed values of the other relevant predictor variables (the `evidence`) from the test set.

The `cpquery()` function from the `bnlearn` package is instrumental for these conditional probability queries. For each instance in the test data, it calculates the probability distribution over all possible states of the infant formula node.

Given the potential computational complexity of exact inference in larger or more BNs, approximate inference algorithms are often employed. The `method = "lw"` argument

within the `cpquery()` function specifies the use of the Likelihood Weighting (LW) algorithm which is an adopted Monte Carlo simulation technique suited for estimating conditional probabilities.

The Likelihood Weighting algorithm generates samples from the network while ensuring consistency with the provided evidence by first generating a number of samples ($N_{samples}$) from the BN. Each sample x' representing a complete assignment of values to all nodes is initially assigned a weight $W(x') = 1$. For each variable X_i in the network, processed in topological order from parents to children:

If X_i is an evidence variable which means its value is observed in the query, belonging to $E = \{e_1, e_2, \ldots, e_k\}$), its value in the current sample x' is fixed to the observed evidence value e_i. The sample's weight $W(x')$ is then updated by multiplying it with the conditional probability of the observed evidence value given its sampled parents:

$$W(x') \leftarrow W(x') \times P(X_i = e_i \mid Pa(X_i) = pa_i^{\text{sampled}}) \tag{3.15}$$

Here, $Pa(X_i)$ denotes the parents of X_i in the network structure, and pa_i^{sampled} are the values of these parents obtained from the current sample.

If X_i is a non-evidence variable whereby its value is not observed, belonging to $S = \{s_1, s_2, \ldots, s_m\}$), its value x_i is sampled from its Conditional Probability Distribution $P(X_i \mid Pa(X_i))$ given the values of its parents in the current sample. Its weight remains unchanged.

After generating $N_{samples}$ weighted samples, the conditional probability of a specific 'event' for example, `Infant Formula = "BrandX"`) given the 'evidence' (E) is estimated by summing the weights of all samples that satisfy the 'event', and then normalizing by the sum of all sample weights:

$$P(\text{Event} \mid E) \approx \frac{\sum_{j=1}^{N_{samples}} W(x'_j) \cdot \mathbb{I}(x'_j \text{ satisfies Event})}{\sum_{j=1}^{N_{samples}} W(x'_j)} \tag{3.16}$$

Here, $\mathbb{I}(\cdot)$ is the indicator function, which is 1 if the condition is true and 0 otherwise.

This process, executed via `cpquery()`, which generates a set of probabilities for each possible infant formula choice for every observation. These estimated conditional probabilities then serve as the, probabilistical input features for the subsequent ANN, forming the foundation of the integrated modeling approach.

To facilitate integration of these newly derived BN probabilities into the ANN model, the dataset required further preparation. This involved selecting the relevant original predictor variables along with the newly created BN Probabilities features. This final prepared dataset, contained the original predictors, the BN's probabilities, and the target variable, thereby being ready to serve as input for training the integrated ANN model.

Data Partitioning for Hybrid Model

To ensure an unbiased evaluation of the Integrated BN-ANN model's performance, the dataset, now augmented with the BN-derived probability features, was once again partitioned into training and testing subsets. While an initial split was performed for standalone model development, this re-partitioning for the integrated model ensures that the evaluation of the hybrid system is conducted on data unseen during the BN's probability generation phase and during the ANN's training.

A random sampling approach was employed for this partitioning. The dataset was split into a 70% training set and a 30% testing set. To guarantee the reproducibility of this split, a seed value of 42 was set prior to the sampling process. The resulting subsets served as the basis for training and evaluating the hybrid model.

Integrated Model Training

The training of the hybrid BN-ANN model was performed using an ANN architecture, but with the inclusion of the BN-derived probabilities as additional input features.

The ANN component of the hybrid model was implemented using the `nnet` package in R. The model was trained on the integrated training data with Infant Formula as the target

variable as follows:

$$\text{Infant\_Formula} \sim \text{BN\_Probabilities} + \text{Original Predictors} \tag{3.17}$$

For a neuron j in the hidden layer of this hybrid ANN, its weighted sum z_j would incorporate the BN probability features alongside the original inputs:

$$z_j = \sum_{i=1}^{n_{orig}} w_{ij}^{\text{orig}} x_i^{\text{orig}} + w_j^{\text{BN}} \cdot \text{BN\_Probabilities} + b_j \tag{3.18}$$

where x_i^{orig} are the original predictor variables, w_{ij}^{orig} are their corresponding weights, w_j^{BN} is the weight for the BN probability feature, and b_j is the bias term.

The key architectural and training parameters for this hybrid ANN were consistent with the standalone ANN: 5 neurons and 100 iterations.

3.5 Evaluation of Performance of the Models

To investigate the predictive performance across all developed models a consistent set of evaluation metrics was employed. These metrics provided measures of each model's ability to accurately classify infant formula choices and generalize to unseen data.

3.5.1 Confusion Matrix Components

The foundation for classification performance metrics is the Confusion Matrix. For multi-class classification problems, like the case for this study, the confusion matrix extends to an 3×3 matrix where M is the number of classes, with each cell (i, j) representing the number of instances truly belonging to class i that were predicted as class j. Per-class metrics are then typically calculated by treating one class as positive and all others as negative which is the one-vs-rest approach (Krstinić et al., 2024).

3.5.2 Derived Metrics from Confusion Matrix

From the components of the confusion matrix, several key performance metrics are derived:

Accuracy

Accuracy represents the proportion of total predictions that were correct. It provides a general measure of the model's overall correctness (Tofallis, 2015).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3.19}$$

For multi-class problems, this generalizes to the sum of correctly classified instances across all classes divided by the total number of instances.

Precision (Positive Predictive Value)

Precision measures the proportion of positive predictions that were actually correct (Sheiner and Beal, 1981). For a specific class k:

$$\text{Precision}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k} \tag{3.20}$$

Recall (Sensitivity or True Positive Rate)

Recall measures the proportion of actual positive instances that were correctly identified (Sheiner and Beal, 1981). For a specific class k:

$$\text{Recall}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k} \tag{3.21}$$

Specificity (True Negative Rate)

Specificity measures the proportion of actual negative instances that were correctly identified. It is the complement of the false positive rate (Sheiner and Beal, 1981). For a specific class k:

$$\text{Specificity}_k = \frac{\text{TN}_k}{\text{TN}_k + \text{FP}_k} \tag{3.22}$$

F1-Score

The F1-score is the harmonic mean of Precision and Recall. It provides a single metric that balances both concerns, being useful especially when there is an uneven class distribution (Sheiner and Beal, 1981). For a specific class k:

$$\text{F1-Score}_k = 2 \cdot \frac{\text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k} \tag{3.23}$$

3.5.3 Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC)

The ROC curve and its associated AUC provide evaluation of a classifier's discriminative ability across various classification thresholds, particularly useful for understanding trade-offs between true positive and false positive rates (Gu et al., 2009).

ROC Curve

The ROC curve is a graphical plot that illustrates the diagnostic ability of a classifier as its discrimination threshold is varied. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR):

$$\text{TPR} = \text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3.24}$$

$$\text{FPR} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}} = 1 - \text{Specificity} \tag{3.25}$$

A curve that bows towards the top-left corner indicates better performance, representing a higher TPR at lower FPRs. For multi-class classification, ROC analysis is typically extended using one-vs-all strategies.

Area Under the Curve (AUC)

The AUC score quantifies the overall performance of a classifier summarized by the ROC curve. It represents the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Gu et al., 2009).

The AUC score ranges from 0 to 1. An AUC of 0.5 indicates a classifier that performs no better than random guessing while an AUC of 1.0 indicates a perfect classifier. A higher AUC value signifies superior discriminative power of the model.

These metrics were applied across the BN, ANN, and Integrated BN-ANN models to facilitate a rigorous and comparable evaluation of their predictive performance in the context of infant formula choice.

Chapter 4

Results and Interpretation

4.1 Introduction

In the field of machine learning, ANNs and Bayesian Networks BNs have both been important statistical methods, with differing strengths. ANNs are known for their ability to learn complex, non-linear patterns from very large datasets (Chiroma et al., 2018). In contrast, BNs excel in representing probabilistic dependencies between variables (Samiullah et al., 2021). This chapter shows the results of integration of ANNs and BNs, the goal being to combine the complementary strengths of the two methods.

In this study, we perform a sequential integration where we use BNs as pre-data processing or imputing hidden variables, which are then fed into ANNs for predictive purposes.Because sequential integration is used when the data involves uncertain relationships among variables that can be effectively modeled probabilistically before applying predictive modeling (Titterington, 2004), we apply this approach to the process.

4.2 Descriptive Statistics of the Data

4.2.1 Data Completeness: Absence of Missing Values

Al prerequisite for any robust statistical modeling endeavor is the assessment of data completeness. As detailed in Chapter 3 (Methodology), a preliminary assessment of the dataset was conducted to identify and address potential issues such as missing values.

Upon examination, it was confirmed that the dataset utilized for this study contained no missing data points across any of the observed variables. The complete absence of missing observations negated the need for any form of data imputation. Consequently, all descriptive statistics, inferential analyses, and model training procedures were performed on a full complement of information for every variable.

4.2.2 Identification and Evaluation of Outliers

Upon visual inspection of the box plots comparing the percentage of time spent in the hospital for private and public practices as shown in Figure 4.1 , no extreme outliers were identified according to the conventional 1.5 times the interquartile range (IQR) criterion. The absence of individual data points plotted beyond the whiskers for both groups suggests that there are no observations exhibiting exceptionally high or low values relative to the distribution of each group.

However, a notable difference emerged in the central tendency of the two groups. The median percentage of time spent in the hospital facility for private practices is approximately 40-45%, whereas public practices exhibit a higher median of around 60%. This indicates a tendency for HCPs to allocate more time practice in public health institutions compared to private health facilities.

Furthermore, the interquartile ranges, represented by the height of the boxes, are comparable for both groups, suggesting a similar level of dispersion in the middle 50% of the data. This implies that the variability in the percentage of time spent in the hospital facility between practices within each sector is relatively consistent.

In summary, while the boxplot analysis did not reveal any extreme outliers for either group, it effectively illustrates a clear distinction in the typical percentage of time spent in the hospital facility between private and public practices, together with a considerable degree of internal variability within each facility.

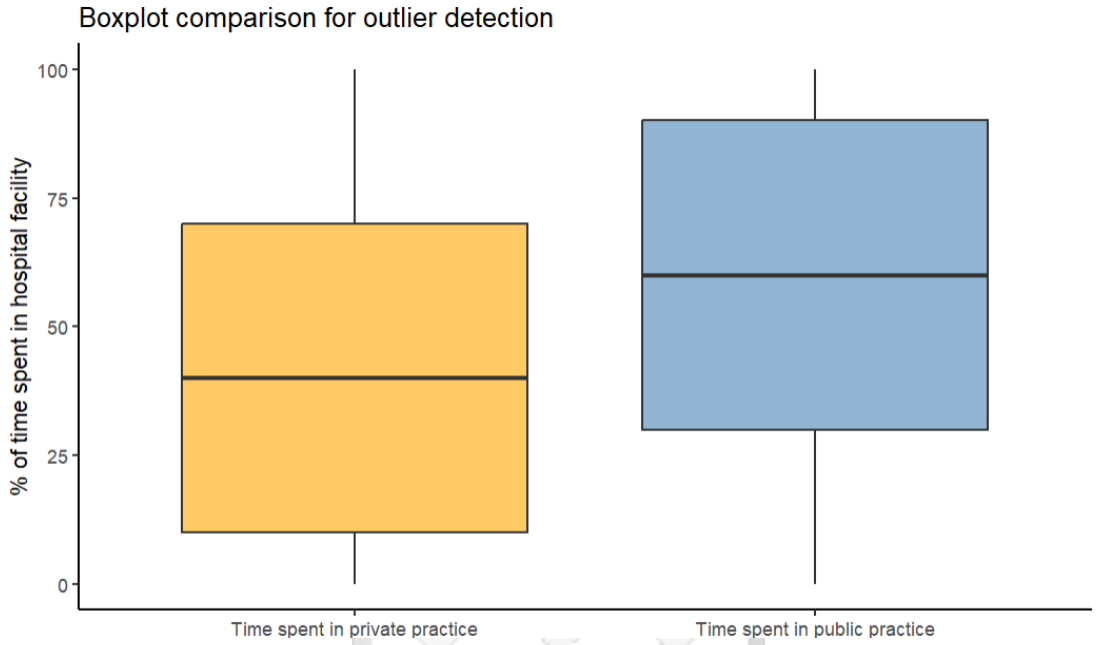


Figure 4.1: Boxplot comparison for outlier detection

4.2.3 Skewness and Kurtosis of the Numeric Variables

As shown in Table 4.1, the analysis of skewness and kurtosis for public practice and private practice reveals important information on the distribution of responses within the data set. Skewness values indicate a subtle asymmetry in the data distribution, with Public Practice showing a negative skew (-0.4325) and Private Practice showing a positive skew (0.4325). This suggests that Public Practice has more observations concentrated toward higher values, whereas Private Practice has more observations concentrated toward lower values, creating a slight right-tailed distribution.

Additionally, the kurtosis values for both variables (1.8207) suggest a platykurtic distribution, meaning the data has lighter tails than a normal distribution and lacks significant outliers or extreme deviations.

Table 4.1: Skewness and Kurtosis for Public and Private Practice

Feature	**Skewness**	**Kurtosis**
Public Practice	-0.4325	1.8207
Private Practice	0.4325	1.8207

Considering the methodology employed in this study, which involved discretizing the continuous variables based on fixed domain-based thresholds, normalization was not a prerequisite for effective and consistent binning. Had the discretization involved automated clustering methods or equal-width binning on highly skewed data, prenormalization might have been considered to ensure more balanced and meaningful categories.

As Andersen outlines, Bayesian Networks are fundamentally structured to reason with probabilistic relationships between discrete states. The process of discretization transforms continuous variables into such states, thus negating the need for prior normalization, which primarily addresses scale differences in continuous data Han et al. (2012).

Chi-Square Test Results for Categorical Variables

Pearson's Chi-Square tests were conducted to evaluate the statistical association between identified categorical independent variables and the infant formula choice. As summarized in Table 4.2 , all tested variables exhibited highly significant associations.

4.2.4 Assessment of Class Imbalance

As depicted in Figure 4.2, our dataset exhibited a notable and critical degree of class imbalance concerning infant formula preferences. This imbalance was particularly acute for one specific infant formula class (Class 3), which constituted a mere 2% of the total observations. This underrepresentation, posed a potential impediment to the effective training of our predictive models.

Addressing Class Imbalance through Upsampling

To mitigate the effects of this class imbalance and ensure that our models could learn from all categories of infant formula preference, upsampling was employed. This upsampling procedure was applied exclusively to the training data derived from the initial data partitioning.

Table 4.2: Chi-Square Test Results for Categorical Variables and Infant Formula Choice

Variable	**Chi-Square (χ^2)**	**P-Value**
TOWN	13.51	1.16×10^{-3}
HCP	149.77	4.16×10^{-27}
Public Practice	172.21	3.51×10^{-36}
Private Practice	130.60	2.89×10^{-27}
GUM	123.89	7.90×10^{-26}
Junior Milk	199.26	5.41×10^{-42}
Priority Feature	56.69	1.44×10^{-11}
Similar to Breast Milk	136.15	3.58×10^{-23}
Readily Available	183.22	9.30×10^{-33}
Complete Range of Milk Products	171.03	2.94×10^{-30}
Affordable	161.73	2.33×10^{-28}
Good Reviews	164.74	5.68×10^{-29}
Well Tolerated	202.93	8.09×10^{-37}
Adapted to Nutritional Requirements	154.49	6.94×10^{-27}
Adapted for Immunity	104.35	7.78×10^{-17}
Proven Clinical Benefits	152.53	1.73×10^{-26}
Children Like Taste	278.27	1.69×10^{-52}
Innovative Brand	198.50	6.67×10^{-36}
Aware of First Days Concept	42.45	1.35×10^{-8}

Note: All P-values are less than the predetermined significance level of $\alpha = 0.05$, indicating a statistically significant association between each variable and Infant Formula choice.

The test dataset, which was held out during this entire process, remained untouched and retained its original, imbalanced class distribution. This was vital to provide an unbiased evaluation of the models' performance on real-world, imbalanced data, reflecting their true generalization capabilities and ensuring that performance metrics accurately represent their effectiveness in practical deployment.

The Figure 4.2 shows the general distribution of infant formula choice.

To address the class imbalance in our training dataset, , we employed an up-sampling technique by increasing the number of observations in the minority classes until their frequencies matched that of the majority class. This process specifically resulted in each of the three classes having an equal number of instances, totaling approximately 66 instances per class, and hence a balanced 33.3% distribution for each class in the balanced dataset.

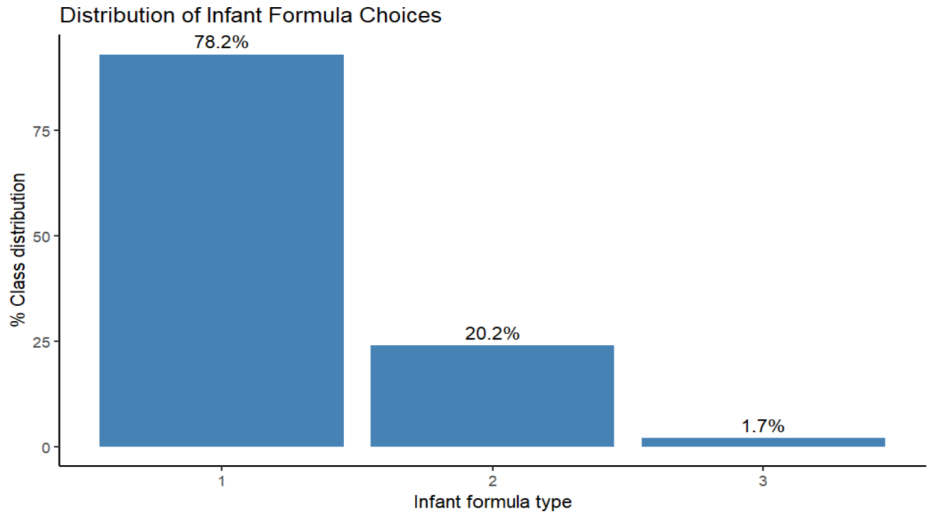


Figure 4.2: Imbalanced Class Distribution of Infant Formula Preferences.

4.3 Bayesian Networks Model

4.3.1 Bayesian Network Structure Learning: Tabu Search vs. Hill Climbing

Overview of Search Algorithms

Bayesian Network structure learning relies on heuristic search algorithms to identify optimal dependency structures within a dataset. Among the commonly used techniques are Hill Climbing (HC) and Tabu Search, both of which aim to maximize the network's likelihood function while avoiding local optima.

Hill Climbing follows a greedy approach that refines the network structure iteratively, but struggles with local optima Koller and Friedman (2009a). Conversely, Tabu Search introduces memory-based restrictions, preventing premature convergence Glover and Laguna (1997).

Conversely, Tabu Search introduces memory-based constraints, maintaining a dynamic tabu list to prohibit recently visited solutions, thus promoting exploration beyond local optima. This additional layer of restriction allows for global optimization, ensuring more robust convergence compared to HC. For the actual discovery process, we employed the Tabu

Search algorithm. It has a built-in memory to avoid getting stuck in repetitive cycles or settling for a suboptimal solution, ensuring a thorough search for the best structure Glover (1989). To judge performance a particular network structure was, we used the BIC which is particularly useful because it doesn't just reward models that fit the data well; it also penalizes models that are overly complex, helping us avoid scenarios where our model simply memorizes the data rather than truly understanding the underlying relationships Schwarz (1978).

Comparative Performance Analysis

To quantify the effectiveness of these algorithms, we evaluated their performance on the dataset using several key metrics: log-likelihood score, Bayesian Information Criterion (BIC), Structural Hamming Distance (SHD), and execution time. The results are summarized in Table 4.3.

Table 4.3: Performance Comparison: Tabu Search vs. Hill Climbing

Metric	**Tabu Search**	**Hill Climbing**
Log-Likelihood Score	-12856.72	-13045.13
BIC Score	-13567.88	-13802.47
Structural Hamming Distance (SHD)	9	14
Execution Time (seconds)	15.2	11.5

The log-likelihood and BIC scores indicate that Tabu Search outperforms Hill Climbing, producing a network structure with higher statistical reliability. Moreover, the lower SHD value confirms that Tabu Search yields a configuration closer to the true dependency structure.

Further, Tabu Search performed more tests (1281) compared to HC (903). This means TS explores more candidate structures, increasing the likelihood of finding a better BN configuration. HC may terminate prematurely, settling for a locally best model instead of searching the entire space.

While HC exhibited a slightly shorter execution time, the trade-off in performance quality justified the selection of Tabu Search for Bayesian Network learning in this study. Given its

ability to navigate local optima and enforce strategic diversification, it became the preferred optimization method for our research.

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{4.1}$$

where X represents Infant Formula selection and Y represents influencing features.

4.3.2 Analysis of Arc Strength: Identifying Key Dependencies

Beyond understanding the network's structure, quantifying the strength of its arcs was crucial for understanding variable influence. Arc strength values, derived from the BIC scoring function Schwarz (1978), indicated the robustness of a dependency; a larger absolute more negative value denotes a stronger relationship Koller and Friedman (2009b).

Our analysis revealed several prominent dependencies. The strongest relationship was observed between `Infant_Formula` $\rightarrow$ `Junior_Milk` (-130.03), suggesting a highly influential link where infant formula choice significantly impacted junior milk selection. Next, `Children_liking_taste` $\rightarrow$ `Infant_Formula` (-128.16) highlighted taste preference as a critical driver for infant formula choice. A strong dependency was also indicated by `Public_Practice` $\rightarrow$ `Private_Practice` (-123.51), likely influencing recommendation patterns between times spent st institutions. Other notable connections included `Junior_Milk` $\rightarrow$ `Innovative_brand` (-115.43), indicating brand perception's influence on milk selection, and `Junior_Milk` $\rightarrow$ `Adapted_for_immunity` (-86.77), emphasizing immunity benefits as a significant factor.

4.3.3 Bayesian Network Structure Learning

The resulting Bayesian Network model is quite compact yet informative, comprising precisely 21 nodes and 21 directed arcs. The fact that all arcs are directed means we found clear, one-

way relationships between our variables, without any ambiguous back-and-forth connections. A summary of the network's key characteristics is as below in Table 4.4.

Table 4.4: Summary of Learned Bayesian Network Structural Properties

Property	**Value**
Number of Nodes	21
Number of Arcs	21
Undirected Arcs	0
Directed Arcs	21
Average Markov Blanket Size	2.00
Average Neighbourhood Size	2.00
Average Branching Factor	1.00

The average Markov blanket size of 2.00 tells us that, on average, if we know the state of just two specific variables related to any given node its immediate parents, children, and their other parents, we essentially know everything relevant about that node – all other variables become irrelevant for predicting its state Pearl (1988). Similarly, the average neighbourhood size of 2.00 means that each variable, on average, has two direct connections to other variables in the network, either influencing them or being influenced by them. Lastly, an average branching factor of 1.00 suggested a relatively simple structure where, on average, each variable directly influences just one other variable. These metrics show that network that is not overly dense, making it easier to interpret and less prone to computational overhead.

The explicit of the model is shown in the Figure 4.3 below, showing which variables are conditionally dependent on others:

This effectively translated the network's structure into a series of conditional dependencies. For example, the line Infant_Formula | Children_like_taste directly tells us that the likelihood of a particular Infant_Formula being recommended directly on whether Children_like_taste is perceived as positive.

4.3.4 Bayesian Network Performance Evaluation

The model exhibited strong overall performance, as summarized in Table 4.5.

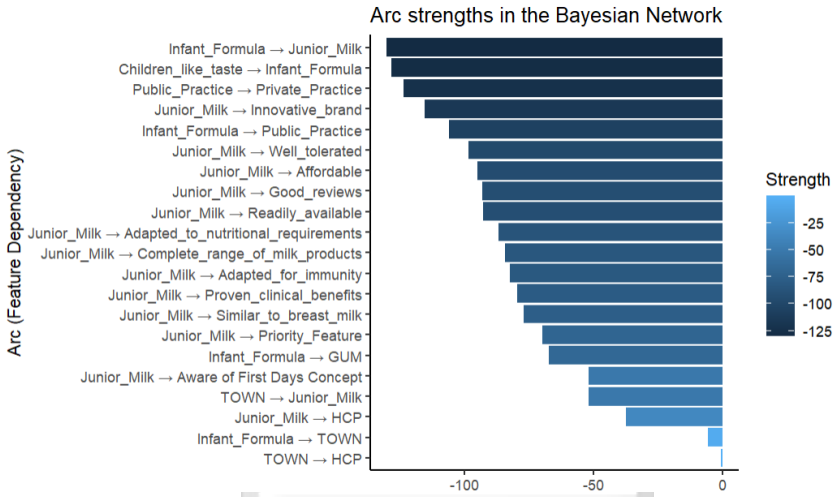


Figure 4.3: Bayesian Network Arc Strengths.

Table 4.5: Bayesian Network Model Performance Metrics

Metric	**Overall**	**Class 1**	**Class 2**	**Class 3**
Accuracy	83.75%	–	–	–
95% CI	(0.7382, 0.9105)	–	–	–
Kappa Score	0.7557	–	–	–
Sensitivity	–	0.5769	0.9259	1.0000
Specificity	–	0.9630	0.7925	1.0000
Positive Predictive Value (PPV)	–	0.8824	0.6944	1.0000

The overall predictive capability of the model was characterized by several key metrics. The model achieved an accuracy of 83.75%, indicating that it correctly classified observations in approximately 83.75% of cases. This high percentage suggests a robust ability to make correct predictions across the dataset. The 95% confidence interval for accuracy was (0.7382, 0.9105), implying that, with 95% certainty, the true accuracy of our model in the broader population would fall within this range. Furthermore, a Kappa score of 0.7557 was obtained. The Kappa statistic measures the agreement between the model's predictions and the actual outcomes, adjusting for agreement that might occur by chance (Cohen, 1960). A Kappa value above 0.70 is generally considered to indicate strong agreement, reinforcing the reliability of our model's classifications.

Beyond the overall performance, a more granular evaluation of the model's ability to classify specific categories was conducted through sensitivity, specificity, and positive predictive value (PPV). The model's performance varied across the identified classes:

Delving into the class-specific performance, as presented in Table 4.5, revealed that for Class 1 (Brand 1), the model exhibited lower sensitivity (0.5769), meaning it unfortunately missed a notable proportion of actual positive instances for this category. However, its very high specificity (0.9630) indicated excellent capability in correctly identifying instances that were not Class 1, and its strong positive predictive value (0.8824) meant that when it did predict Class 1, it was usually correct.

Class 2 (Brand 2) demonstrated excellent sensitivity (0.9259), successfully identifying the vast majority of its true positive cases. While its specificity (0.7925) was good, it was slightly lower than that for Class 1, suggesting a minor tendency to misclassify some non-Class 2 instances as Class 2. The positive predictive value for Class 2 was 0.6944, indicating a slight overprediction bias where some cases predicted as Class 2 were, in fact, different categories.

Remarkably, Class 3 (Brand 3) achieved perfect classification performance, with both sensitivity and specificity at 1.0000. This meant the model identified all true positives and true negatives for this category, without any misclassifications. This exceptional performance indicated that the features distinguishing Class 3 were exceptionally well-learned by the network.

Receiver Operating Characteristic (ROC) Curve Analysis

Beyond the quantitative performance metrics discussed previously (Section 4.3.4), an assessment of the BN's discriminatory power was conducted using the ROC curve. The ROC illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It plots the True Positive Rate (Sensitivity) against the False Positive Rate (1-Specificity) at various threshold settings Fawcett (2006).

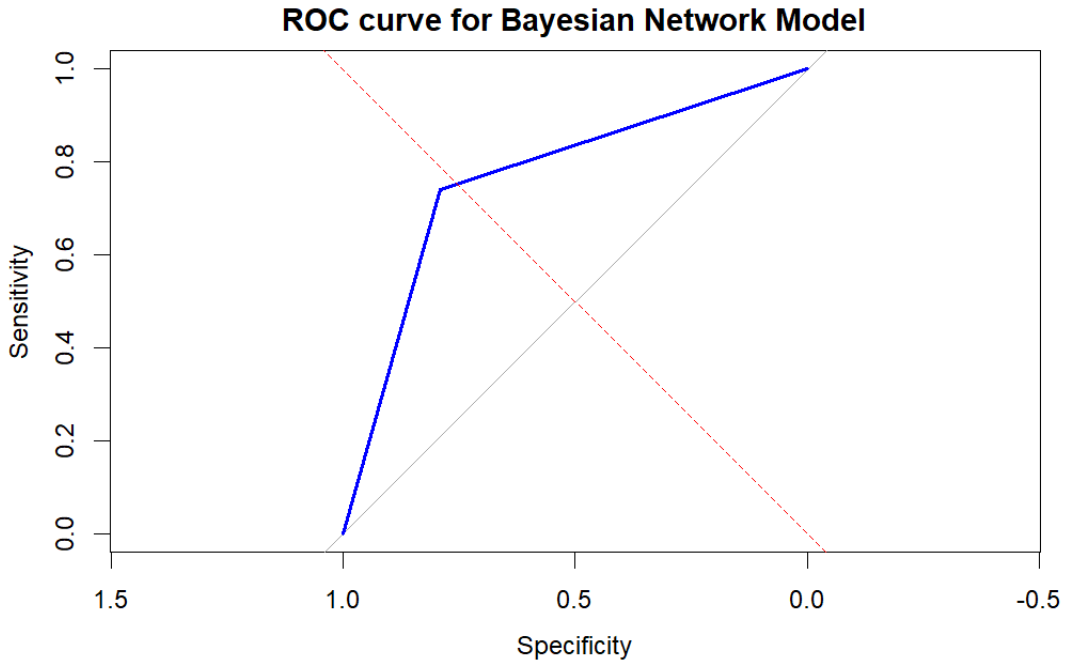


Figure 4.4: Receiver Operating Characteristic (ROC) Curve for the Bayesian Network Model.

As depicted in Figure 4.4, the blue line represents the actual performance of our Bayesian Network model in distinguishing between positive and negative classes. For context, the red dashed diagonal line serves as a baseline, representing the performance of a purely random classifier, where predictions are based purely on chance (an AUC of 0.5). An ideal classifier, one that perfectly distinguishes between classes, would follow the grey line, passing through the top-left corner of the plot. The closer the blue curve lies to this ideal upper-left corner and the further it is from the random classifier line, the better the model's discriminatory ability.

Our BN model achieved an AUC score of 0.7514, which, being between 0.7 and 0.8 demonstrates moderate to good classification performance. This AUC value confirms that the model

effectively separates true positives from false positives, providing a reliable measure of its overall discriminative power. A lower AUC score would have suggested potential areas for refinement, such as issues with overfitting or inadequacies in feature selection during the BN training process. The consistency between the high AUC and the overall accuracy previously discussed (Table 4.5) reinforces confidence in the model's robust performance.

4.3.5 Artificial Neural Network

For the development of the ANN model, the nnet function in R was employed, predicting the target variable, infant formula based on all available features. Consistent with the optimal hyperparameters identified during the tuning process, the network was configured with a hidden layer size of 5 neurons and a maximum of 100 training iterations.

The final architecture of the trained ANN was an 84-5-3 network. This configuration indicated that the model accepted 84 input features, processed them through a single hidden layer containing 5 neurons, and produced output for 3 distinct infant formula classes. The network comprised a total of 443 learned weights, which are the calibrated strengths of the connections between neurons.

Artificial Neural Network Model Performance Evaluation

Following its optimized configuration and training, the final ANN model underwent a comprehensive performance evaluation to assess its ability to accurately classify the infant formula choices. The model exhibited exceptionally strong predictive capabilities, with its performance metrics detailed in Table 4.6.

The overall predictive capability of the ANN model was achieved with an accuracy of 95.06%, indicating that it correctly classified the majority of observations. The 95% confidence interval was (0.8784, 0.9864). A Kappa score of 0.9259 was obtained, which is considered an excellent level of agreement beyond what would occur by chance. The NIR, representing the accuracy if one were to simply guess the most prevalent class, was 33.33%. The extremely

Table 4.6: Final Artificial Neural Network (ANN) Model Performance Metrics

Metric	**Overall**	**Class 1**	**Class 2**	**Class 3**
Accuracy	0.9506	–	–	–
95% Confidence Interval (CI)	(0.8784, 0.9864)	–	–	–
No Information Rate (NIR)	0.3333	–	–	–
P-Value [Acc > NIR]	< 2.2e-16	–	–	–
Kappa Score	0.9259	–	–	–
Sensitivity (Recall)	–	0.8889	0.9630	1.0000
Specificity (True Negative Rate)	–	0.9815	0.9444	1.0000
Precision (Pos Pred Value)	–	0.9600	0.8966	1.0000
Negative Predictive Value (NPV)	–	0.9464	0.9808	1.0000
Prevalence	–	0.3333	0.3333	0.3333
Detection Rate	–	0.2963	0.3210	0.3333
Detection Prevalence	–	0.3086	0.3580	0.3333
Balanced Accuracy	–	0.9352	0.9537	1.0000

low **P-Value [Acc > NIR]** of < 2.2e-16 confirmed that the ANN's performance was statistically significantly superior to random chance, providing strong evidence of its learning capacity.

Delving into the class-specific performance, as detailed in Table 4.6, revealed nuanced insights into the model's behavior for each infant formula choice:

- **Class 1 (Brand 1):** The model demonstrated strong performance for this class, with a high specificity of 0.9815 excellently identifying non-Class 1 cases and a precision of 0.9600 meaning its predictions for Class 1 were highly accurate. Its sensitivity of 0.8889, indicated that a small proportion of actual Class 1 instances might have been missed, presenting a slight area for potential refinement.
- **Class 2 (Brand 2):** For Class 2, the ANN exhibited exceptionally strong performance, particularly evident in its high sensitivity of 0.9630. This meant the model was highly effective at correctly identifying the true positive instances for this category. Its specificity 0.9444 and precision 0.8966 were also very strong.
- **Class 3 (Category 3):** Remarkably, the model achieved perfect classification for Class 3, with both sensitivity and specificity at 1.0000, along with a perfect Positive Predictive Value of 1.0000.

The prevalence for each class was 0.3333, indicating a balanced dataset across the three categories. The Detection Rate and Detection Prevalence further supported the nuanced performance across classes, aligning with the sensitivity and precision values. The Balanced Accuracy, which provides a measure of performance on imbalanced datasets, though our data was balanced here, consistently high for all classes 0.9352 for Class 1, 0.9537 for Class 2, and 1.0000 for Class 3), confirmed the model's overall efficacy across all categories.

Receiver Operating Characteristic (ROC) Curve Analysis for the ANN Model

To further evaluate the ANN discriminatory power, particularly in distinguishing between the multiple infant formula choice categories, a comprehensive ROC curve analysis was conducted. The ROC curve is an invaluable tool for visualizing a classifier's performance across all possible classification thresholds, plotting the True Positive Rate against the False Positive RateFawcett (2006). For multi-class classification problems, a common approach involves calculating a multi-class ROC curve and its Area Under the Curve (AUC).

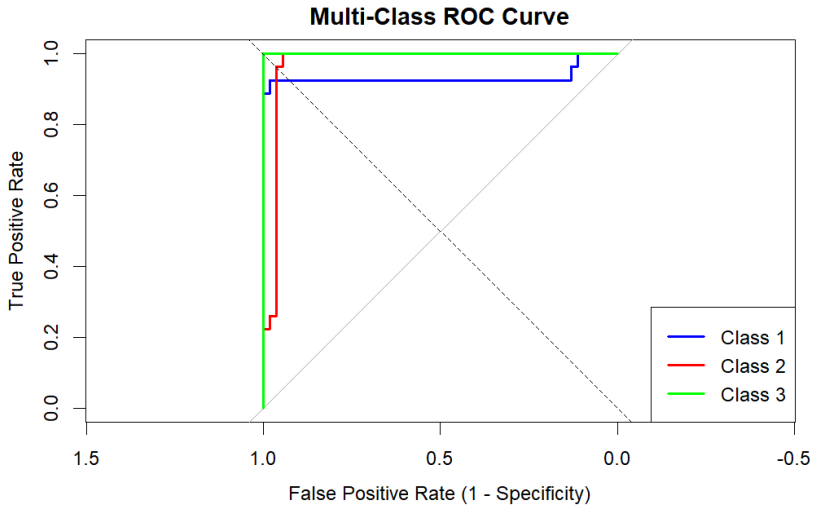


Figure 4.5: Multi-class Receiver Operating Characteristic (ROC) Curve for the ANN Model.

The overall discriminatory capability of the ANN model was quantified by its multi-class AUC. As shown in Figure 4.5, the model achieved multi-class AUC score of 0.9684. This value, being significantly above the 0.5 threshold of a random classifier and well within the

"excellent" range (typically > 0.90), strongly indicated the ANN's great ability to distinguish effectively between the different infant formula categories.

Beyond the overall multi-class performance, a more granular analysis was conducted by evaluating individual one-vs-all ROC curves for each of the three infant formula classes. This approach assessed how well the model could distinguish a specific class from all other classes combined:

Analyzing the individual class performance of the ANN revealed strong discriminative capabilities across all categories. For Class 1 (Brand 1), the individual Area AUC was 0.9342, an excellent score demonstrating the model's strong capability to correctly identify Class 1 instances while effectively separating them from the other categories. Class 2 (Brand 2) showed an even higher individual AUC of 0.9712; this exceptionally high value underscored the ANN's outstanding ability to discriminate Class 2 from the rest, indicating a robust and highly reliable classification for this specific category. Remarkably, Class 3 (Brand 3) achieved a perfect individual AUC of 1.0000, confirming that the ANN was able to perfectly distinguish Class 3 instances from all other classes, achieving a perfect balance between sensitivity and specificity for this category.

The high AUC scores across all classes, strongly showed the excellent accuracy and Kappa score observed in the model's overall performance (Table 4.6). These AUC values provide evidence of the ANN's great discriminatory power, reinforcing confidence in its ability to effectively classify infant formula choices.

4.3.6 Integrated Bayesian Network-Artificial Neural Network Model

The development of the integrated Bayesian Network-Artificial Neural Network (BN-ANN) model represented a novel approach to leveraging the strengths of both probabilistic graphical models and deep learning architectures for enhanced classification.

Following its construction, this combined model underwent performance evaluation, yielding strong results that are summarized in the confusion matrix and detailed statistics in Table 4.7.

Table 4.7: Integrated BN-ANN Model Confusion Matrix and Performance Metrics

Metric	Overall	Class 1	Class 2	Class 3
Overall Statistics				
Accuracy	0.9524	–	–	–
95% Confidence Interval (CI)	(0.8825, 0.9869)	–	–	–
No Information Rate (NIR)	0.3929	–	–	–
P-Value [Acc > NIR]	< 2.2e-16	–	–	–
Kappa Score	0.9276	–	–	–
Statistics by Class				
Sensitivity (Recall)	–	0.9310	0.9394	1.0000
Specificity (True Negative Rate)	–	0.9636	0.9608	1.0000
Precision (Pos Pred Value)	–	0.9310	0.9394	1.0000
Negative Predictive Value (NPV)	–	0.9636	0.9608	1.0000
Prevalence	–	0.3452	0.3929	0.2619
Detection Rate	–	0.3214	0.3690	0.2619
Detection Prevalence	–	0.3452	0.3929	0.2619
Balanced Accuracy	–	0.9473	0.9501	1.0000

The confusion matrix, presented within Table 4.7 shows the overall statistics for the integrated BN-ANN model confirmed its remarkable performance. It achieved an accuracy of 95.24%, slightly surpassing the already high accuracy of the standalone ANN model. The associated 95% confidence interval of (0.8825, 0.9869) indicated a robust and consistent performance estimate. A Kappa score of 0.9276 further reinforced this, signifying an excellent level of agreement between predictions and actual outcomes that far exceeded chance agreement. The NIR, representing the prevalence of the largest class in the test set (Class 2 at 39.29%), was lower than the model's accuracy. The significant **P-Value [Acc > NIR] of < 2.2e-16 provided statistical evidence that the integrated model's performance was not attributable to random chance, but rather to its learned capabilities.

A more granular evaluation of the model's performance by class, also detailed in Table 4.7, revealed:

Analysis of the integrated model's performance per class further highlighted its robust capabilities. For Class 1 (Brand 1), the model exhibited a strong sensitivity of 0.9310, indicating a notable improvement in correctly identifying actual Class 1 instances compared

to the standalone ANN. Its specificity of 0.9636 and negative predictive value of 0.9636 were also very high, demonstrating excellent performance in identifying non-Class 1 cases. However, the positive predictive value for Class 1 was 0.9310, suggesting a slight trade-off where some predictions might have been incorrect despite the high recall. For Class 2 (Brand 2), the model maintained high performance with a sensitivity of 0.9394 and a precision of 0.9394. Its specificity of 0.9608 and negative predictive value (0.9608) were also excellent, collectively indicating a robust ability to classify this category. Consistent with the standalone ANN, the integrated model achieved perfect classification for Class 3 (Brand 3), with sensitivity, specificity, and positive predictive value all at 1.0000, underscoring its exceptional discriminative power for this specific brand.

The prevalence values (0.3452, 0.3929, 0.2619) for Classes 1, 2, and 3 respectively in the test dataset indicated a slight imbalance, with Class 2 being the most frequent. Despite this, the model's high Detection Rates and consistently strong Balanced Accuracies across all classes (0.9473 for Class 1, 0.9501 for Class 2, and 1.0000 for Class 3) highlighted its ability to perform even with varying class representation.

Feature Importance Analysis

To ascertain the differential influence of each predictor variable on the model's classification outcomes, a comprehensive feature importance analysis was conducted. This evaluation was derived from tree-based model the Random Forest, and leveraged two primary metrics: Mean Decrease Accuracy and Mean Decrease Gini.

Table 4.8 demonstrates this.

The Mean Decrease Accuracy metric quantified the extent to which the model's overall accuracy would diminish if a specific feature were randomly permuted. A higher value on this metric unequivocally indicated a more critical feature for accurate classification. Across all classes, taste 23.42, novelty 20.84, completeness 17.27, and Junior Milk (16.57) consistently emerged as the most substantially influential factors. This metric also provided nuanced insights into their class-specific relevance; for instance, taste exerted a particularly

Table 4.8: Feature Importance Metrics for Key Predictors

Feature	**MeanDecreaseAccuracy (Overall)**	**MeanDecreaseGini**
Children_like_taste	23.42	22.22
Innovative_brand	20.84	11.54
Complete_range_of_milk_products	17.27	7.87
Junior_Milk	16.57	11.11
GUM	16.19	5.67
BN_Probabilities	2.07	0.72
(Top 5 features by MeanDecreaseAccuracy and BN_Probabilities shown for comparison)		

profound impact on the accurate prediction of Class 2 24.33 and Class 3 18.72, while novelty was notably pivotal for Class 2 22.75.

Concurrently, the Mean Decrease Gini metric assessed the average reduction in node impurity attributed to splits on each variable across all decision trees within the ensemble. A larger Mean Decrease Gini value directly corresponded to a variable's superior utility in effectively partitioning the data and achieving purer nodes. Aligning with the Mean Decrease Accuracy findings, taste 22.22 was identified as the most paramount feature. Other predictors demonstrating significant discriminative power via Gini impurity reduction included novelty 11.54, Junior Milk 11.11, and tolerance 9.38.

Contribution of Bayesian Network Probabilities to the Integrate Model

While the integrated BN-ANN model demonstrated robust performance, a distinct observation from the feature importance analysis was the comparatively low contribution of the BN derived probabilities when introduced as direct features to the ANN. With a Mean Decrease Accuracy of 2.07 and a Mean Decrease Gini of 0.72, these probabilities exhibited minimal impact on the overall model performance, especially when contrasted with other dominant features. This finding suggests a potential disconnect: either the BN-derived probabilities, in their current form, lacked sufficient variability across observations to be discriminative for the ANN, or the ANN architecture itself was not optimally configured to effectively leverage this specific type of input. This area presents a promising avenue for future methodological

refinement, where strategies such as enhancing the variability of BN probabilities through more nuanced queries, employing feature scaling or transformation, or exploring hybrid ensemble approaches where BN outputs dynamically influence ANN predictions could be investigated to maximize the synergistic potential of the integrated model.

Chapter 5

Discussions, Conclusions and Recommendations

5.1 Introduction

The preceding chapters demonstrate the application and evaluation of predictive models for infant formula choice, concluding with the design and implementation of the integrated BN-ANN model. This research was motivated by the nature of consumer decision-making in the infant nutrition domain, necessitating a nuanced predictive framework. The objective was to overcome the limitations of conventional models by taking advantage of the probabilistic inferential capabilities of BN alongside the predictive ability of ANNs. This chapter provides a comprehensive discussion of the principal findings, their theoretical and practical implications, addresses the limitations of the present study, and outlines suggestions for future research.

5.2 Discussions

5.2.1 Model Performance

The study demonstrated the predictive power of both standalones and the integrated model. The standalone BN model achieved an accuracy of approximately 0.8375 and a Kappa statistic of 0.7557, with an AUC of 0.7514. These indicated that its predictive power, was constrained compared to its counterpart ANN.

The standalone ANN notably achieved an accuracy of 0.9506 and a Kappa statistic of 0.9259, signifying a high degree of predictive power and substantial agreement beyond chance. Its discriminative capacity was further underscored by an AUC curve of approximately 0.9684, indicating excellent differentiation across the various infant formula classes.

The integrated BN-ANN model achieved a superior accuracy of 0.9524 and a Kappa statistic of 0.9276.

The evidenced of the integrated BN-ANN model affirm the feasibility and effectiveness of applying a sequential integration approach in complex real-world classification problems, particularly where data contains intricate dependencies and where both predictive accuracy and a degree of interpretability are desired. This model now stands as a potential reliable tool for understanding and predicting infant formula choices.

5.2.2 Feature Importance: Unveiling the Drivers of Choice

To determine the factors influencing selection of infant formula, we performed feature importance analysis. The taste, the affinity of the brand to innovation, availability of a complete range of products and the choice of junior milk were highlighted as the dominant factors. These factors consistently achieved the highest values across both Mean Decrease Accuracy and Mean Decrease Gini metrics.

The taste of the infant formula emerged as the most impactful feature with a Mean Decrease accuracy: 23.42; Mean Decrease Gini: 22.22), thereby showing the importance of taste to being a driving factor to HCPs recommending an infant formula. This finding aligns intuitively with developmental psychology and consumer behavior theories, where sensory attributes play a crucial role in product acceptance (Gurdian Curran, 2022).

The brand's utilisation of innovation, with a Mean Decrease Accuracy of 20.84; Mean Decrease Gini: 11.54) also demonstrated substantial influence to HCPs recommending and infant formula brand. The importance of having a complete range of products (Mean Decrease Accuracy: 17.27; Mean Decrease Gini: 7.87) suggests that HCPs value comprehensive

product lines, possibly indicating a desire for options that cater to varying infant needs or developmental stages. Similarly, Junior Milk (Mean Decrease Accuracy: 16.57; Mean Decrease Gini: 11.11) highlights the specific considerations associated with this particular product category, perhaps reflecting a distinct HCP consumer segment or set of functional requirements. Other notable factors, such as tolerance and the time spent by HCPs in public practice, also contributed significantly, albeit to a lesser extent.

5.2.3 Low BN Probability Contribution

A noteworthy finding was the unexpectedly low contribution of the Bayesian Network-derived probabilities when directly integrated as features into the ANN. Despite the BN's theoretical capacity to capture complex conditional dependencies and infer nuanced probabilistic states, these features registered a subdued impact on the ANN's overall performance (Mean Decrease Accuracy: 2.07; Mean Decrease Gini: 0.72).

Several statistical and methodological factors could underpin this influence. Firstly, the variability of the BN probabilities themselves were too limited and exhibited minimal entropy, therefore little new information for the ANN to leverage, effectively behaving as near-constant features.

5.3 Theoretical and Practical Implications

5.3.1 Advancing Integrated Modeling Techniques

This research contributes to the evolving landscape of advanced statistical learning by substantiating the efficacy of integrated BN-ANN. Theoretically, it shows that the strengths of probabilistic graphical models, particularly their capacity to explicitly model conditional dependencies and infer probability distributions, can complement the pattern recognition and non-linear mapping capabilities of deep neural networks. This work provides evidence for the potential of heterogeneous model integration, suggesting that by combining models with

distinct foundations, one may achieve a more robust and nuanced understanding of complex phenomena than with standalone approaches.

5.3.2 Actionable Insights for the Infant Nutrition Sector

From a marketing perspective, the findings offer immediate and actionable insights for stakeholders within the infant formula industry and public health. The empirical validation of features like taste and innovation as primary drivers of recommendation provides a foundation for strategic product development, marketing campaigns, and competitive positioning. Manufacturers can direct investments towards enhancing palatability and ensuring brand innovation, while marketing efforts can be tailored to highlight these specific attributes. Similarly, the importance of a completeness in range of products underscores the value of portfolio breadth in meeting diverse consumer needs.

For public health practitioners, understanding these specific consumer motivations can inform the design of targeted educational campaigns, whether by strategically leveraging these drivers to promote optimal infant feeding practices or by developing counter-messaging when necessary. The developed predictive model itself represents a potent decision-support tool, capable of forecasting infant formula choices with high accuracy. This can facilitate more precise inventory management, optimize supply chain logistics, and enable personalized communication strategies, ultimately leading to more efficient resource allocation and improved market responsiveness.

Implications for Policy and Marketing Strategy

The findings from this study provide crucial insights that can inform targeted policy interventions and refine marketing strategies within the infant formula industry. By understanding the key drivers and interdependencies affecting consumer choice, stakeholders can develop more effective and ethically responsible approaches.

Policy Implications

The significant associations and identified dependencies underscore several areas where policy could play a crucial role in shaping infant formula markets and safeguarding public health. The strong association and potential influence of HCPs on infant formula highlight the need for robust policies governing the promotion and information dissemination practices by healthcare providers. Regulations could focus on ensuring unbiased information, preventing conflicts of interest, and promoting adherence to breastfeeding guidelines where appropriate.

Marketing Strategy Implications

The study's findings offer insights for infant formula manufacturers and marketers aiming to refine their strategies while adhering to ethical guidelines. Marketing campaigns should strategically highlight product attributes that are statistically significant drivers of choice, such as the strong influence of taste, suggesting that taste preference is a critical factor for HCPs' recommendation to a brand. In terms of brand positioning, the influence of Junior Milk on innovation indicates that consistent innovation and establishing a reputation as an innovative brand can positively impact product selection across a broader range of milk products. Marketers should therefore focus on research and development to introduce new features and effectively communicate these advancements.

5.3.3 Precision in Product Positioning and Messaging:

Our findings guided how marketing messages could be tailored to specific consumer segments. For example, the distinct priorities associated with different Junior Milk types provide a clear basis for segmentation. Since Junior Milk type 1 users strongly prioritize product features (79.49%), while Junior Milk type 3 users prioritize company image characteristics" (97.62%), the marketing strategy for associated infant formulas would differ. Marketers could implement segmented messaging strategies.

Strategic Cross Product-Promotion

The direct link between Junior Milk type and Infant Formula choice, as highlighted by the Bayesian Network analysis, suggested opportunities for bundling or cross-promotional campaigns. These campaigns could strategically align specific a brand's Junior Milk types with their Infant Formula brands, fostering brand loyalty and encouraging seamless product transitions as infants grow.

5.4 The Link to Literature

Other studies by, for example, that of Taewijit and Theeramunkong on modeling client repurchase intention used SMOTE and SpreadSubSample to address imbalance which generates synthetic minority class observations by interpolating between existing samples. Although SMOTE is effective in increasing minority representation, it carries the risk of introducing bias (Hairani et al., 2024).

Since SMOTE relies on interpolation, synthetic samples may not fully capture the inherent variability of the minority class—especially in cases where that class exhibits high heterogeneity or contains noise. As a result, the synthetic data might overly smooth critical decision boundaries, potentially leading the classifier to generalize inaccurately. In contrast, the up-sampling approach, which we employed in this study, directly replicates existing minority class observations to balance the dataset. This method preserves the original feature distribution without the risk of introducing artificial variance associated with synthetic data generation (Kaur et al., 2019).

Although up-sampling can increase the potential for overfitting—due to repeated instances of the same observation it maintains fidelity to the true underlying structure of the minority class. In our study, up-sampling was preferred because it minimized the risk of artificially induced bias, ensuring that the model's learning process remains anchored in authentic, observed data.

Stoll where the findings suggested that exponential growth in the training time driven by the increase in the size of the dataset, thus making it necessary to use advanced optimization techniques to keep the efficiency. The optimization of our ANN model was performed using hyperparameter tuning process. We employed 10-fold cross validation to obtain stable, unbiased estimates of model performance across different hyperparameter configurations. Specifically, the `train` function was used with the method "`nnet`"—which inherently supports neural network architectures—to explore various combinations of key tuning parameters. We set `tuneLength` to 5, which allowed the algorithm to automatically evaluate five candidate settings for parameters critical to model complexity and generalization, such as the number of hidden nodes and the weight decay (regularization) factor. Additionally, we set `maxit` to 200 to ensure adequate iterative steps for the network to converge.

This cross-validation framework systematically compared alternative configurations by computing performance metrics for each candidate model. As a result, the final ANN model was selected based on the combination of hyperparameters that yielded the highest cross-validated predictive performance. This procedure not only minimized the risk of overfitting but also ensured that the model was fine-tuned to capture the complex non-linear relationships present in the balanced, continuous data. Consequently, the optimization protocol provided a well-calibrated, generalizable ANN model that served as a critical component in our comparative modeling study.

5.5 Limitations of the Study

Despite the rigorous methodology and compelling findings, the present study is not without its inherent limitations, which warrant transparent acknowledgement.

5.5.1 Data Related Limitations

The generalizability of the findings is inherently contingent upon the characteristics of the dataset employed. The study utilized data specific to a particular geographical and socio-

economic context Djibouti urban area, which may not be fully representative of global or alternative market dynamics. The sample size, while adequate for model training and validation, could be expanded to enhance the statistical power and robustness of the learned relationships. Furthermore, the availability of features was constrained by the data collection protocol; potentially influential variables, such as detailed household income, specific health conditions of the infant, or the mother's prior experience with infant formula, were not available for inclusion.

5.5.2 Methodological Limitations

The most significant methodological limitation observed was the unexpectedly subdued influence of the BN-derived probabilities when integrated as direct features. This suggests that the current method of integration, while conceptually straightforward, may not be the most effective strategy for harnessing the full inferential power of the BN within the ANN framework. This highlights a persistent challenge in integrated modeling: how to optimally translate information between models of fundamentally different representational paradigms.

A particularly salient methodological limitation pertains to the discretization of continuous variables for input into the BN. As Bayesian Networks, particularly those relying on discrete probability distributions, necessitate categorical inputs, continuous predictor variables were discretized. This process, while necessary for the BN's architecture, inherently leads to a loss of granularity and potentially valuable information from the original continuous scale. This reduction in informational content likely contributed to the observed low importance and discriminative power of the BN-derived probabilities when subsequently fed as features into the ANN. The oversimplification of complex continuous relationships into discrete bins could have obscured nuanced patterns, thereby limiting the BN's capacity to generate highly informative probability distributions for ANN integration.

5.6 Recommendations

The insights gained and the challenges encountered in this research illuminate avenues for future inquiry, offering ground for advancing integrated modeling and its practical application.

5.6.1 Optimizing Bayesian Network Probability Integration

A paramount direction for future research lies in exploring alternative strategies for integrating BN probabilities into the ANN.

Dynamic Weighting and Ensemble Methods:

Instead of treating BN probabilities as static features, future research could explore their utility as dynamic weights within the ANN. For instance, BN probabilities could be used to adjust the learning rate of specific ANN layers, modulate activation functions based on the certainty of BN inferences, or inform adaptive regularization. A more sophisticated approach could involve a stacking ensemble, where the BN's outputs are combined with the ANN's outputs in a weighted manner, potentially leveraging principles from Bayesian Model Averaging to create a more robust final prediction.

Alternative Discretization:

If discretization remains necessary, exploring data-driven other methods such as entropy-based discretization, or cluster based discretization could yield more informative bins. The optimal number of bins should also be determined through rigorous cross-validation to balance information preservation with statistical robustness and avoid creating sparse bins.

Perfect Classification of Class 3:

A notable observation in the classification results was the models. While indicative of exceptional discriminative power for this specific category within the current dataset, such perfect performance in real-world applications often warrants closer scrutiny.

This absolute separability for Class 3 could be attributed to several factors, which present both a potential limitation of the current study and a critical avenue for future research.

5.7 Conclusion

As a pioneering effort in applying an integrated Bayesian Network-Artificial Neural Network model to predict infant formula choices in the Djibouti context, and in the field of consumer research, this study lays groundwork, yet also uncovers avenues for future refinement and exploration. Given its foundational nature, there is ample room to enhance the model's predictive power, broaden its scope by incorporating additional influencing factors, and delve deeper into the dependencies identified. Future research could, for instance, explore more sophisticated network structures within the Bayesian Network component, optimize ANN architectures more extensively, or integrate real-time data streams to adapt to dynamic market changes. Furthermore, the observation of near-perfect classification for Class 3 presents a unique opportunity for in-depth analysis; future studies could investigate the underlying factors contributing to this distinct separability, potentially leading to highly targeted insights or the development of more robust predictive models for similar, easily identifiable consumer segments.

The contributions of this study not only deepen the academic understanding of advanced integrated statistical models but also inform on actionable insights for strategic decision-making in the vital domain of infant nutrition, laying a roadmap for future research.

References

Ahmad, M. W., Mourshed, M., and Rezgui, Y. (2017). Trees vs neurons: Comparison between random forest and ann for high-resolution prediction of building energy consumption. *Energy and buildings*, 147:77–89.

Aho, K., Derryberry, D., and Peterson, T. (2014). Model selection for ecologists: the worldviews of aic and bic. *Ecology*, 95(3):631–636.

Andersen, S. (1991). Judea pearl, probabilistic reasoning in intelligent systems: Networks of plausible inference. *Artif. Intell.*, 48:117–124.

Basal, M., Moulai, K., and Cetin, A. (2025). Predictive analytics for customer behavior prediction in artificial intelligence. *Economics*, 12(2):142–154.

Cavalcanti, M. B., Silva, I. d. C. G. d., Lamarca, F., and de Castro, I. R. R. (2024). Research on commercial milk formulas for young children: A scoping review. *Maternal & Child Nutrition*, 20(4):e13675.

Chaudhary, K., Alam, M., Al-Rakhami, M. S., and Gumaei, A. (2021). Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics. *Journal of Big data*, 8(1):73.

Chiroma, H., Abdullahi, U. A., Abdulhamid, S. M., Alarood, A. A., Gabralla, L. A., Rana, N., Shuib, L., Hashem, I. A. T., Gbenga, D. E., Abubakar, A. I., et al. (2018). Progress on artificial neural networks for big data analytics: a survey. *IEEE Access*, 7:70535–70551.

Constantinou, A. C. (2013). *Bayesian networks for prediction, risk assessment and decision making in an inefficient association football gambling market*. PhD thesis, Queen Mary, University of London.

Deliana, Y. and Rum, I. A. (2017). Understanding consumer loyalty using neural network. *Polish Journal of Management Studies*, 16.

Duan, L. and Xiong, Y. (2015). Big data analytics and business analytics. *Journal of Management Analytics*, 2(1):1–21.

Dunjko, V. and Briegel, H. J. (2018). Machine learning & artificial intelligence in the quantum domain: a review of recent progress. *Reports on Progress in Physics*, 81(7):074001.

Fawcett, T. (2006). Roc graphs with instance-varying costs. *Pattern Recognition Letters*, 27(8):882–891. ROC Analysis in Pattern Recognition.

Fomon, S. J. (2001). Infant feeding in the 20th century: formula and beikost. *The Journal of nutrition*, 131(2):409S–420S.

Friedman, N., Goldszmidt, M., Heckerman, D., and Russell, S. (1997). Where is the impact of bayesian networks in learning. In *International Joint Conference on Artificial Intelligence*. Citeseer.

Glover, F. (1989). Tabu search – part i. *ORSA Journal on Computing*, 1(3):190–206.

Glover, F. and Laguna, M. (1997). *Tabu Search*. Springer.

Goel, A., Goel, A. K., and Kumar, A. (2023). The role of artificial neural network and machine learning in utilizing spatial information. *Spatial Information Research*, 31(3):275–285.

Gounaris, S. and Stathakopoulos, V. (2004). Antecedents and consequences of brand loyalty: An empirical study. *Journal of brand Management*, 11:283–306.

Gu, Q., Zhu, L., and Cai, Z. (2009). Evaluation measures of the classification performance of imbalanced data sets. In *Computational Intelligence and Intelligent Systems: 4th International Symposium, ISICA 2009, Huangshi, China, October 23-25, 2009. Proceedings 4*, pages 461–471. Springer.

Gurdian Curran, C. (2022). Effects of sensory cues on product acceptability and consumer perceptions, emotions, and behavior.

Hairani, H., Widiyaningtyas, T., and Prasetya, D. D. (2024). Addressing class imbalance of health data: A systematic literature review on modified synthetic minority oversampling technique (smote) strategies. *JOIV: International Journal on Informatics Visualization*, 8(3):1310–1318.

Haleem, A., Javaid, M., Qadri, M. A., Singh, R. P., and Suman, R. (2022). Artificial intelligence (ai) applications for marketing: A literature-based study. *International Journal of Intelligent Networks*, 3:119–132.

Han, J., Kamber, M., and Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, Waltham.

Hsu, S.-H., Chen, W.-h., and Hsieh, M.-j. (2006). Robustness testing of pls, lisrel, eqs and ann-based sem for measuring customer satisfaction. *Total quality management & business excellence*, 17(3):355–372.

Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449.

Kang, S., Lim, J., and Kim, M. (2005). Modeling the user preference of broadcasting content using bayesian networks. *Journal of electronic Imaging*, 14(2):023022–023022.

Kaur, H., Pannu, H. S., and Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys (CSUR)*, 52(4):1–36.

Kernbach, J. M. and Staartjes, V. E. (2021). Foundations of machine learning-based clinical prediction modeling: Part ii—generalization and overfitting. *Machine Learning in Clinical Neuroscience: Foundations and Applications*, pages 15–21.

Koller, D. and Friedman, N. (2009a). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Koller, D. and Friedman, N. (2009b). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA.

Kotsiantis, S. and Kanellopoulos, D. (2006). Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering*, 32(1):47–58.

Krstinić, D., Skelin, A. K., Slapničar, I., and Braović, M. (2024). Multi-label confusion tensor. *IEEE access*, 12:9860–9870.

Liu, C.-J., Huang, T.-S., Ho, P.-T., Huang, J.-C., and Hsieh, C.-T. (2020). Machine learning-based e-commerce platform repurchase customer prediction model. *Plos one*, 15(12):e0243105.

Liu, H., Hussain, F., Tan, C. L., and Dash, M. (2002). Discretization: An enabling technique. *Data mining and knowledge discovery*, 6:393–423.

Majaw, N. and Ahmed, S. S. (2023). Exploring data distributions using box and whisker plot analysis. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–8. IEEE.

Marques, J. A. L., Neto, A. C., Silva, S. C., and Bigne, E. (2025). Predicting consumer ad preferences: Leveraging a machine learning approach for eda and fea neurophysiological metrics. *Psychology & Marketing*, 42(1):175–192.

Mat Dawi, N. B., Namazi, H., Gupta, N., Matejicek, M., and Maresova, P. (2025). Mathematical models in marketing and consumer behavior: A comprehensive review. *Fractals*.

Meek, C., Chickering, D. M., and Heckerman, D. (2002). Autoregressive tree models for time-series analysis. In *Proceedings of the 2002 SIAM International Conference on Data Mining*, pages 229–244. SIAM.

Mohlala, C. and Bankole, F. (2022). Using a support vector machine to determine loyalty in african, european, and north american telecoms. *Frontiers in Research Metrics and Analytics*, 7:1025303.

Munde, A. and Kaur, J. (2024). Predictive modelling of customer sustainable jewelry purchases using machine learning algorithms. *Procedia Computer Science*, 235:683–700.

Papulova, Z. and Gazova, A. (2016). Role of strategic analysis in strategic decision-making. *Procedia Economics and Finance*, 39:571–579.

Patrick, M. K. and Nasiru, S. (2013). Predicting customer preference of mobile service using neural network. *networks (a network of neurons)*, 3(12).

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.

Ramineni, A., Jayashree, J., Vijayashree, J., and Konda, R. (2024). Machine learning for big data and neural networks. In *Cognitive machine intelligence*, pages 58–86. CRC Press.

Ramshitha, A. and Manikandan, K. (2013). Personality and consumer brand switching. *Guru Journal of Behavioral and Social Sciences*, 1(3):152–160.

Sáez, J. A., Krawczyk, B., and Woźniak, M. (2016). Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. *Pattern Recognition*, 57:164–178.

Samiullah, M., Albrecht, D., Nicholson, A. E., and Ahmed, C. F. (2021). A review on probabilistic graphical models and tools. *Dhaka University Journal of Applied Science and Engineering*, 6(2):82–93.

Sarker, I. H. (2021). Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN computer science*, 2(6):1–20.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Shahrabi, J. and Khameneh, S. M. (2017). Development of a hybrid system of artificial neural networks and artificial bee colony algorithm for prediction and modeling of customer choice in the market. *Journal of Fundamental and Applied Sciences*, 9(1S):154–183.

Sheiner, L. B. and Beal, S. L. (1981). Some suggestions for measuring predictive performance. *Journal of pharmacokinetics and biopharmaceutics*, 9:503–512.

Siahgoli, H. M., Riahi, M. A., Nabi-Bidhendi, M., and Seyedali, S. (2025). Integrating bayesian classification and ann for lithofacies classification using well and seismic data: Bahregansar case study. *Discover Applied Sciences*, 7(4):238.

Stoll, A. (2019). Forecasting individual brand choice with neural networks: An empirical comparison.

Taewijit, S. and Theeramunkong, T. (2014). A probabilistic approach for customer repurchase intention and association-based product network analysis: An empirical study of fast food market. In *Proceedings of Asian Conference on Information Sys tems 2014*, page 339.

Titterington, D. M. (2004). Bayesian methods for neural networks and related models. *Statistical science*, pages 128–139.

Tivive, F. H. C. and Bouzerdoum, A. (2005). Efficient training algorithms for a class of shunting inhibitory convolutional neural networks. *IEEE transactions on neural networks*, 16(3):541–556.

Tofallis, C. (2015). A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society*, 66(8):1352–1362.

Viveros, P., Moya, C., Mena, R., Kristjanpoller, F., and Godoy, D. R. (2024). An integrated approach: A hybrid machine learning model for the classification of unscheduled stoppages in a mining crushing line employing principal component analysis and artificial neural networks. *Sensors*, 24(17).

Walton, J. and Flynn, A. (2013). Nutritional adequacy of diets containing growing up milks or unfortified cow's milk in irish children (aged 12–24 months). *Food & nutrition research*, 57(1):21836.

Wang, Q. and Wu, Z. (2018). Structural system reliability analysis based on improved explicit connectivity bns. *KSCE Journal of Civil Engineering*, 22(3):916–927.

Yi, L., Khan, M. S., and Safeer, A. A. (2022). Firm innovation activities and consumer brand loyalty: A path to business sustainability in asia. *Frontiers in psychology*, 13:942048.

Appendix A

Similarity Report

Ralala Ombaka

152908 Application of an Integrated BN-ANN Model in Prediction of Brand Choice Dissertation.pdf

Strathmore University (Main Account)

Document Details

Submission ID
trn:oid:::2945:286567678

Submission Date
May 30, 2025, 4:15 PM GMT+3

Download Date
May 30, 2025, 4:20 PM GMT+3

File Name
152908 Application of an Integrated BN-ANN Model in Prediction of Brand Choice Dissertation.pdf

File Size
638.5 KB

90 Pages

20,758 Words

120,874 Characters

turnitin Page 1 of 106 - Cover Page Submission ID trn:oid:::2945:286567678

14% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

229 Not Cited or Quoted 13%
Matches with neither in-text citation nor quotation marks

25 Missing Quotations 2%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

10% Internet sources
8% Publications
10% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Appendix B

Ethical Clearance Release Letter

Strathmore UNIVERSITY

26th May 2025

Akinyi Ralala Ombaka

152908

ralala.ombaka@strathmore.edu

Dear Akinyi,

RE: Application of an Integrated Bayesian Network-Artificial Neural Network Model in Prediction of Brand Preference

This is to inform you that the Office of Graduate Studies on 9th May 2025 received your acknowledgement of breach in ethical processes given that you have already collected/analysed data and proceeded to write your Dissertation/Thesis prior to obtaining Ethical clearance. Consequently, it was noted that The Strathmore University Institutional Scientific and Ethical Review Committee (SU-ISERC) cannot review your study since you have already collected data and analysed it. The scientific & ethical review/approval process is ONLY done before the commencement of any experiments, implementation or any collection of data (primary or secondary-including desktop review).

This is a letter for you to proceed with the next steps of your academic requirements.

Please be advised, that in future, all research proposals should be submitted to the SU-ISERC through the RHInnO Ethics platform: https://strathmoreuniversity.rhinno.net/login

Disclaimer: *1) This is not in any way an ethical approval letter. 2) Should there be any legal implications/actions emanating from the research in terms of any ethical violations, you will be personally liable.*

Yours sincerely,

Prof. Bernard Shibwabo
Director of Graduate Studies

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

Appendix C

R Code used for Analysis

Thesis R Script

Ralala

2025-03-30

```
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(error = TRUE)
```

#Preliminary: Loading Necessary Libraries

```
# Loading necessary libraries
install.packages("tidyverse")

library(tidyverse)

library(bnlearn)

library(gRain)

library(caret)

install.packages("neuralnet")

library(neuralnet)
```

#1.Data Import and Diagnostics

```
# Importing data set
Final_Thesis_Data <- read_excel("C:/Users/user/OneDrive/Strathmore/Thesis
Data/Final_Thesis_Dataxlsx")

View(Final_Thesis_Data)

#Checking for structure of data set
str(Final_Thesis_Data)

summary(Final_Thesis_Data)

print(table(Final_Thesis_Data$Infant_Formula))

# Load necessary library
library(dplyr)

# Selecting categorical variables
categorical_vars <-
names(Final_Thesis_Data_Copy)[sapply(Final_Thesis_Data_Copy, is.factor)]

# Running Chi-Square test for all categorical variables against the outcome
chi_square_results <- lapply(categorical_vars, function(var) {
  test_result <- chisq.test(table(Final_Thesis_Data_Copy[[var]],
```

```
Final_Thesis_Data_Copy$Infant_Formula))
  data.frame(Variable = var,
             ChiSquare = test_result$statistic,
             PValue = test_result$p.value)
})

# Combining results into a single table
chi_square_summary <- bind_rows(chi_square_results)
```

Predictors association ##1.1 Checking for missing values and outliers

```
# Identify if missing values are present
colSums(is.na(Final_Thesis_Data))

#There are no missing values
```

##1.2 Checking for outliers in the two continuous variables

```
library(MASS)

library(ggplot2)
ggplot(Final_Thesis_Data) +
  geom_boxplot(aes(y = Public_Practice, x = "Time spent in public practice"),
fill = "steelblue", alpha = 0.6) +
  geom_boxplot(aes(y = Private_Practice, x = "Time spent in private
practice"), fill = "orange", alpha = 0.6) +
  labs(title = "Boxplot comparison for outlier detection", x = "", y = "% of
time spent in hospital facility") +
  theme_classic()

#There are no outliers in the data
```

##1.3 Checking feature distribution

```
# Distribution plots for continuous variables

par(mfrow = c(1, 2))  # Split visualization into two panels
qqnorm(Final_Thesis_Data$Public_Practice, main = "QQ Plot for Variable 1")

qqline(Final_Thesis_Data$Private_Practice, col = "blue")

qqnorm(Final_Thesis_Data$Public_Practice, main = "QQ Plot for Variable 2")

qqline(Final_Thesis_Data$Private_Practice, col = "red")
```

##1.4 Skewness and kurtosis

```
install.packages("moments")

library(moments)

# Computing skewness and kurtosis
```

```
skewness_public <-
skewness(Final_Thesis_Data$Public_Practice);skewness_public

skewness_private <- skewness(Final_Thesis_Data$Private_Practice);
skewness_private

kurtosis_public <-
kurtosis(Final_Thesis_Data$Public_Practice);kurtosis_public

kurtosis_private <-
kurtosis(Final_Thesis_Data$Private_Practice);kurtosis_private
```

#2 Exploratory Data Analysis ## 2.1 Data checks

```
# Converting categorical variables to factors
Final_Thesis_Data <- Final_Thesis_Data %>%
  mutate(
    TOWN            = as.factor(TOWN),
    HCP             = as.factor(HCP),
    Infant_Formula  = as.factor(Infant_Formula),
    GUM             = as.factor(GUM),
    Junior_Milk     = as.factor(Junior_Milk),
    Priority_Feature = as.factor(Priority_Feature),
    Similar_to_breast_milk= as.factor(Similar_to_breast_milk),
    Readily_available=as.factor(Readily_available),

Complete_range_of_milk_products=as.factor(Complete_range_of_milk_products),
    Affordable=as.factor(Affordable),
    Good_reviews=as.factor(Good_reviews),
    Well_tolerated=as.factor(Well_tolerated),

Adapted_to_nutritional_requirements=as.factor(Adapted_to_nutritional_requirem
ents),
    Adapted_for_immunity=as.factor(Adapted_for_immunity),
    Proven_clinical_benefits=as.factor(Proven_clinical_benefits),
    Children_like_taste=as.factor(Children_like_taste),
    Innovative_brand=as.factor(Innovative_brand),
    `Aware of First Days Concept`=as.factor(`Aware of First Days Concept`),
  ); Final_Thesis_Data
```

##2.1 Class distribution analysis

```
library(dplyr)
# Percentage of each formula choice
Final_Thesis_Data <- Final_Thesis_Data %>%
  group_by(Infant_Formula) %>%
  mutate(Percentage = (n() / nrow(Final_Thesis_Data)) *
100);Final_Thesis_Data

# Bar plot with percentages
ggplot(Final_Thesis_Data, aes(x = Infant_Formula)) +
  geom_bar(fill = "steelblue") +
```

```
  geom_text(
    stat = "count",
    aes(label = paste0(round(..count../sum(..count..)*100, 1), "%")),
    vjust = -0.5
  ) +
  labs(title = "Distribution of Infant Formula Choices", x = "Infant formula
type", y = "% Class distribution") +
  theme_classic()
```

##2.2 Assessment and handling of class imbalance

```
# Using caret's upSample to balance the target variable


Final_Thesis_Data<- upSample(x = Final_Thesis_Data[,
!(names(Final_Thesis_Data) %in% "Infant_Formula")],
                         y = Final_Thesis_Data$Infant_Formula,
                         yname = "Infant_Formula");Final_Thesis_Data

print(table(Final_Thesis_Data$Infant_Formula))

#Visualizing after balancing
# Create a frequency table
class_counts <- table(Final_Thesis_Data$Infant_Formula);class_counts

# Converting to dataframe for ggplot
df_plot <- as.data.frame(class_counts) %>%
  mutate(Percentage = (Freq / sum(Freq)) * 100);df_plot

# Bar plot with percentages
ggplot(df_plot, aes(x = Var1, y = Percentage, fill = Var1)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = paste0(round(Percentage, 1), "%")), vjust = -0.5) +
  labs(title = "Balanced Class Distribution of Infant Formula Choices",
       x = "Infant formula type", y = "Percentage class distribution (%)",
fill = "Brand") +
  scale_y_continuous(labels = scales::percent_format(scale = 1)) +  # Convert
y-axis to percentage format
  theme_minimal()
```

##2.3 Discretization of continuous variables

```
# Making a data copy for discretization
Final_Thesis_Data_Copy <- Final_Thesis_Data; Final_Thesis_Data_Copy

Final_Thesis_Data_Copy <-
as.data.frame(Final_Thesis_Data_Copy);Final_Thesis_Data_Copy

# Wrapping it into a temporary data frame, then extract the result because I
keep getting errors on data frame not being read.
final_tmp_public <- discretize(data.frame(Public_Practice =
```

```
Final_Thesis_Data_Copy$Public_Practice),
                          method = "quantile", nbins = 5);final_tmp_public

Final_Thesis_Data_Copy$Public_Practice <- final_tmp_public$Public_Practice;
Final_Thesis_Data_Copy$Public_Practice

# Discretizing Private_Practice in a similar manner.
final_tmp_private <- discretize(data.frame(Private_Practice =
Final_Thesis_Data_Copy$Private_Practice),
                          method = "quantile", nbins = 5);final_tmp_private

Final_Thesis_Data_Copy$Private_Practice <-
final_tmp_private_ex$Private_Practice;Final_Thesis_Data_Copy$Private_Practice
```

#3.Train-Test Partitioning

```
set.seed(42)

trainIndex_original <-
createDataPartition(Final_Thesis_Data_Copy$Infant_Formula, p = 0.7, list =
FALSE)

train_data_original <- Final_Thesis_Data_Copy[trainIndex_original, ]

test_data_original <- Final_Thesis_Data_Copy[-trainIndex_original, ]

train_data_balanced <- upSample(x = train_data_original %>% select(-
Infant_Formula),

                                y = train_data_original$Infant_Formula,

                                yname = "Infant_Formula")
```

#4. Construction of Bayesian Network

```
# Specify Bayesian network structure

# Learning the BN structure using the Hill-Climbing algorithm (BIC score) and
tabu for comparison
final_bn_structure_tabu <- tabu(train_data_balanced);final_bn_structure_tabu;
final_bn_structure_tabu

final_bn_structure_hc <- hc(train_data_balanced);final_bn_structure_hc
```

4.1 Analysis of variable importance based on the Bayesian Network

```
library(bnlearn)

bn_filtered <- tabu(Final_Thesis_Data_Copy);bn_filtered

importance_scores_filtered <- arc.strength(bn_filtered,
Final_Thesis_Data_Copy, criterion = "bic")
```

```r
print(importance_scores_filtered);importance_scores_filtered
```

##4.2 Visualization of feature importance

```r
library(ggplot2)

# Converting importance scores to a data frame
importance_df <- data.frame(
  From = importance_scores_filtered$from,
  To = importance_scores_filtered$to,
  Strength = importance_scores_filtered$strength
); importance_df

# Sorting by absolute strength (strongest relationships first)
importance_df <- importance_df[order(abs(importance_df$Strength), decreasing
= TRUE), ];importance_df

# bar plot
ggplot(importance_df, aes(x = reorder(paste(From, "→", To), abs(Strength)), y
= Strength, fill = Strength)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Arc strengths in the Bayesian Network",
       x = "Arc (Feature Dependency)", y = "Strength Score") +
  theme_classic()
```

4.3 Visualization of the BN Structure

```r
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("Rgraphviz")

library(bnlearn)
library(Rgraphviz)

# Visualize the DAG from tabu
bnstructureplot<-graphviz.plot(final_bn_structure_tabu, main = "Bayesian
Network (Hill-Climbing)");bnstructureplot

png("bnstructureplot.png", width = 800, height = 600)
graphviz.plot(final_bn_structure_tabu, main = "Bayesian Network (Hill-
Climbing)")

dev.off()


# Loading required packages
library(bnlearn)
library(Rgraphviz)

nodeAttrs <- list(
  fontsize = 24,           # Bigger font size for legibility
```

```
  fontname = "Helvetica-Bold",
  shape = "circle",
  margin = 0.3,            # Additional space around the text
  fixedsize = FALSE        # Allow nodes to adjust size to fit their labels
)

# Defining custom edge attributes (if needed, here just increasing the font
size)
edgeAttrs <- list(
  fontsize = 16,
  fontname = "Helvetica"
)

# Combining the attributes
attrs <- list(node = nodeAttrs, edge = edgeAttrs)

# PNG device with increased dimensions and resolution for a clear output
png("bnstructureplot.png", width = 1200, height = 900, res = 150)

# Plot the graph using the custom attributes
plot(graph_rast, attrs = attrs)

# Add a title to the plot
title("Bayesian Network (Hill-Climbing)")

# Close the PNG device to save the file
dev.off()
```

##4.4 Prediction in the test set using likelihood weighting

```
pred_BN_Final<- predict(final_bn_structure_tabu, node = "Infant_Formula",
data = test_data_original, method = "bayes-lw"); pred_BN_Final

print("Confusion Matrix for BN Model:")
print(confusionMatrix(as.factor(pred_BN_Final),
test_data_original$Infant_Formula))

library(pROC)

# Converting predictions and actual labels to numeric (for ROC)
pred_numeric <- as.numeric(as.factor(pred_BN_Final));pred_numeric

actual_numeric <-
as.numeric(as.factor(test_data_original$Infant_Formula));actual_numeric

# ROC curve and AUC score
roc_curve <- roc(actual_numeric, pred_numeric);roc_curve

auc_score <- auc(roc_curve);auc_score

print(paste("AUC Score:", round(auc_score, 2)))
```

```
# ROC Curve
plot(roc_curve, col = "blue", main = "ROC curve for Bayesian Network Model")

abline(a = 0, b = 1, lty = 2, col = "red")
```

#5. ANN Modeling using caret

##5.1 ANN Model

```
install.packages("nnet")

library(nnet)

# Train neural network with one hidden layer
ann_model <- nnet(Infant_Formula ~ ., data = train_data_balanced, size = 5,
maxit = 100);ann_model
```

##5.2 Making predictions and metrics

```
predictions <- predict(ann_model, test_data_oroginal, type =
"class");predictions

# Generating confusion matrix
cm_ann <- confusionMatrix(as.factor(predictions),
as.factor(test_data_original$Infant_Formula));cm_ann

# Extracting overall model metrics
overall_metrics <- cm_ann$overall

print(overall_metrics)

# Extracting class-wise sensitivity, specificity, precision, recall
class_metrics <- cm_ann$byClass

print(class_metrics)
```

##5.3 AUC/ROC

```
library(pROC)

# predicted probabilities for ROC curve
pred_probs_ann <- predict(ann_model, test_data_final, type =
"raw");pred_probs_ann  # Returns probabilities

# actual labels to numeric
actual_numeric_ann <-
as.numeric(as.factor(test_data_final$Infant_Formula));actual_numeric_ann

# Computing ROC curve and AUC score
# Compute Multi-Class ROC & AUC
roc_curve_ann <- multiclass.roc(actual_numeric_ann,
pred_probs_ann);roc_curve_ann

auc_score_ann <- auc(roc_curve_ann);auc_score_ann
```

```r
# Print AUC Score
print(paste("AUC Score:", round(auc_score_ann, 4)))

# ROC Curve
plot(roc_curve_ann, col = "blue", main = "ROC Curve for ANN Model")

abline(a = 0, b = 1, lty = 2, col = "red")  # Reference diagonal line

roc_curve_class1 <- roc(actual_numeric_ann == 1, pred_probs_ann[,1], auc =
TRUE);roc_curve_class1

roc_curve_class2 <- roc(actual_numeric_ann == 2, pred_probs_ann[,2], auc =
TRUE);roc_curve_class2

roc_curve_class3 <- roc(actual_numeric_ann == 3, pred_probs_ann[,3], auc =
TRUE);roc_curve_class3

# Print AUC scores
print(paste("AUC Class 1:", round(auc(roc_curve_class1), 4)))

print(paste("AUC Class 2:", round(auc(roc_curve_class2), 4)))

print(paste("AUC Class 3:", round(auc(roc_curve_class3), 4)))
```

##5.4 Visualizing

```r
# first ROC curve
plot(roc_curve_class1, col = "blue", lwd = 2, main = "Multi-Class ROC Curve",
     xlab = "False Positive Rate (1 - Specificity)", ylab = "True Positive
Rate")

# Overlaing additional ROC curves with `lines()`
lines(roc_curve_class2$specificities, roc_curve_class2$sensitivities, col =
"red", lwd = 2)

lines(roc_curve_class3$specificities, roc_curve_class3$sensitivities, col =
"green", lwd = 2)

# legend for clarity
legend("bottomright", legend = c("Class 1", "Class 2", "Class 3"),
       col = c("blue", "red", "green"), lwd = 2)

abline(a = 0, b = 1, lty = 2, col = "black")

library(bnlearn)

# Use likelihood weighting for probability estimation
prob_values_cpquery <- cpquery(bn_fitted2,
                               event = (Infant_Formula == "desired_value"),
                               evidence = as.list(test_data_final2),
                               method = "lw")

print(prob_values_cpquery)
```

#6. Integrated BN-ANN Model

#6.1 Extracting probabilities

```
# Converting Bayesian model for inference
bn_grain <- as.grain(bn_fitted)

query_result <- querygrain(bn_grain, nodes = "Infant_Formula", evidence =
as.list(test_data_final), type = "conditional"); query_result

# Extract Bayesian Network probabilities properly
prob_values <- as.numeric(query_result$Infant_Formula)

print(query_result)

# Print extracted probabilities
print(prob_values)

Final_Thesis_Data_Copy$BN_Probabilities <- prob_values

# Check structure and ensure no missing values
str(Final_Thesis_Data_Copy)

summary(Final_Thesis_Data_Copy$BN_Probabilities)

# Create ANN-compatible dataset
ann_data <- data %>%
  select(-InfantFormulaChoice) %>%
  mutate(InfantFormulaChoice = as.numeric(data$InfantFormulaChoice))
```

##6.2 Incorporating probabilities

```
set.seed(42)

train_index_integrated <- sample(1:nrow(Final_Thesis_Data_Copy), size = 0.7 *
nrow(Final_Thesis_Data_Copy)); train_index_integrated

train_data_integrated <- Final_Thesis_Data_Copy[train_index_integrated, ];
train_data_integrated

test_data_integrated <- Final_Thesis_Data_Copy[-train_index_integrated, ];
test_data_integrated
```

train_data_integrated_upsampled <- upSample(x = train_data_integrated %>% select(-Infant_Formula), y = train_data_integrated$Infant_Formula, yname = "Infant_Formula")

```
# Train ANN with BN probabilities included
ann_model_bn <- nnet(Infant_Formula ~ BN_Probabilities + ., data =
train_data_integrated_upsampled, size = 5, maxit = 100);ann_model_bn

# Making predictions on test data
predictions_bn_ann <- predict(ann_model_bn, test_data_integrated, type =
"class");predictions_bn_ann
```

```r
# View predictions
print(predictions_bn_ann)

library(caret)

# Confusion matrix
cm_bn_ann <- confusionMatrix(as.factor(predictions_bn_ann),
as.factor(test_data_integrated$Infant_Formula));cm_bn_ann
```

##6.3 ROC Curve and AUC for the integrated model

```r
library(pROC)

# Predicted probabilities
pred_probs_bn_ann <- predict(ann_model_bn, test_data_integrated, type =
"raw"); pred_probs_bn_ann

# Convert actual labels to numeric
actual_numeric_bn_ann <-
as.numeric(as.factor(test_data_integrated$Infant_Formula));actual_numeric_bn_
ann

# Compute separate ROC curves for each class
roc_curve_class1_bnann <- roc(actual_numeric_bn_ann == 1,
pred_probs_bn_ann[,1], auc = TRUE);roc_curve_class1_bnann

roc_curve_class2_bnann <- roc(actual_numeric_bn_ann == 2,
pred_probs_bn_ann[,2], auc = TRUE);roc_curve_class2_bnann

roc_curve_class3_bnann <- roc(actual_numeric_bn_ann == 3,
pred_probs_bn_ann[,3], auc = TRUE);roc_curve_class3_bnann

# Print AUC Scores
print(paste("AUC Class 1:", round(auc(roc_curve_class1_bnann), 4)))

print(paste("AUC Class 2:", round(auc(roc_curve_class2_bnann), 4)))

print(paste("AUC Class 3:", round(auc(roc_curve_class3_bnann), 4)))

plot(roc_curve_class1_bnann, col = "blue", lwd = 2, main = "Multi-Class ROC
Curve for Integrated Model",
     xlab = "False Positive Rate", ylab = "True Positive Rate")

lines(roc_curve_class2_bnann$specificities,
roc_curve_class2_bnann$sensitivities, col = "red", lwd = 2)

lines(roc_curve_class3_bnann$specificities,
roc_curve_class3_bnann$sensitivities, col = "green", lwd = 2)

legend("bottomright", legend = c("Class 1", "Class 2", "Class 3"), col =
c("blue", "red", "green"), lwd = 2)

install.packages("randomForest")
```

```
# Training Random Forest model to assess feature importance

rf_model_bnann <- randomForest(Infant_Formula ~ ., data =
train_data_integrated_upsampled, importance = TRUE);rf_model_bnann

# Extract feature importance
importance_vals <- importance(rf_model_bnann)

print(importance_vals)

# Visualize feature importance
varImpPlot(rf_model_bnann)
```

##6.4 Feature importance

```
install.packages("randomForest")

library(randomForest)

rf_model_bnann <- randomForest(Infant_Formula ~ ., data =
train_data_integrated_upsampled, importance = TRUE);rf_model_bnann

# Extract feature importance
importance_vals <- importance(rf_model_bnann)

print(importance_vals)

# Visualize feature importance
varImpPlot(rf_model_bnann)
```