processes

MDPI

Article

Research on Industry Data Analytics on Processing Procedure of Named 3-4-8-2 Components Combination for the Application Identification in New Chain Convenience Store

You-Shyang Chen [1,*], Chien-Ku Lin [2], Jerome Chih-Lung Chou [3], Ying-Hsun Hung [3,*] and Shang-Wen Wang [3]

[1] College of Management, National Chin-Yi University of Technology, Taichung 411, Taiwan
[2] Department of Business Management, Hsiuping University of Science and Technology, Taichung 412, Taiwan
[3] Department of Information Management, Hwa Hsia University of Technology, New Taipei City 235, Taiwan
* Correspondence: yschen@ncut.edu.tw (Y.-S.C.); sean@go.hwh.edu.tw (Y.-H.H.)

Abstract: With the rapid economic boom of Asian countries, the president of Country-A has made great efforts to reform in recent years. The prospect of economic development is promising, and business opportunities are emerging gradually, depicting a prosperous scene; accordingly, people's livelihood consumption also has changed significantly. The original main point of consumption for urban and rural people was the old and traditional grocery store with poor sanitation, but due to the economic improvement, the quality of consumption has also improved, and convenience stores are gradually replacing grocery store. However, convenience store management involves performance, logistic, competition, and personnel costs. Both whether the store can create a net profit and evaluate and select a new store will be important keys that significantly influence business performance. Therefore, this study attempts to use the industry data analysis method for highlighting a concept of processing an experience procedure of named 3-4-8-2 components combination in two stages. First, in the data preprocessing stage, this research considers 22 condition attributes and two types of decision factors, that include net profit and new store selection, and use both techniques of attribute selection and data discretization through the analysis and prediction of data mining tools. Next, in the experiment execution stage, three well-known classifiers (Bayes net, logistic regression, and J48 decision tree) with past good performance and four models (without preprocessing, with attribute selection, with data discretization, and with attribute selection and data discretization) are used for eight different experiments through two data verification methods (percentage split and cross-validation). Conclusively, three key results are identified from empirical analysis: (1) It is found that the prediction accuracy of the J48 decision tree classifier is relatively high and stable among the three classifiers in this study; at the same time, the J48 decision tree can yield comprehensible knowledge-based rules to instruct interested parties. (2) The results of this study show that the important attributes for the net profit decision attribute include the store type, POS number, and cashier number, while the important attributes for the new store selection include the store type and cashier number. (3) There is a difference in the selection of important attributes. Furthermore, four key valuable contributions are addressed from the empirical results, including academic contributions, enterprise contributions, application contributions, and management contributions. It is expected that the direction of store layout expansion can be found and identified through this study, but there are still many risks hidden behind the considerable business opportunities that need to be carefully managed.

Keywords: industry data application; chain convenience store; store expansion; data mining tools

Citation: Chen, Y.-S.; Lin, C.-K.; Chou, J.C.-L.; Hung, Y.-H.; Wang, S.-W. Research on Industry Data Analytics on Processing Procedure of Named 3-4-8-2 Components Combination for the Application Identification in New Chain Convenience Store. *Processes* **2023**, *11*, 180. https://doi.org/10.3390/pr11010180

Academic Editor: Xiong Luo

Received: 12 October 2022
Revised: 30 December 2022
Accepted: 3 January 2023
Published: 6 January 2023

1. Introduction

The International Monetary Fund (IMF) estimated that the economic growth rate of Country-A in 2017 and 2018 can reach 6.8% and 6.9%, and even more than 7% in 2020. The annual remittance of overseas workers of Country-A to their hometown is about USD 25

to 30 billion. In addition, the booming information industry has created a middle class with strong consumption ability. At the same time, the per capita income of Country-A will increase to USD 4200, so Country-A is about to enter a stage of rapid growth in the consumer market, and domestic consumption opportunities are expected [1]. In this rapid growth stage, the retail market has also undergone drastic changes, with the emergence of many department stores and shopping malls in metropolitan and suburban areas, and consumers have gradually changed their consumption patterns. With the increasing income of people in Country-A, the international chain convenience store operators sense the business opportunities. Bright and clean convenience stores are replacing traditional grocery stores on the street. In response to such changes, how do the chain convenience store operators lay out and seize the market quickly?

Thus, this study takes the chain convenience store B in Country-A as the object and analyzes the influence of decision of location selection on business benefit for store expansion of the chain convenience store B in Country-A. Due to the improvement of politics and public security as well as the positive economic development in recent years, people's income is increasing, which leads international chain convenience store operators to find business opportunities. Convenience stores are gradually replacing the traditional grocery stores. Along with the chain convenience store operators' certainty about Country-A people's rising disposable income, as well as a large local and untapped retail market, the situation has encouraged the chain convenience store enterprises to actively expand their stores from city to city in the country, with the hope of changing the habits of people still shopping in traditional grocery stores. The local climate, traffic, population distribution, public security, and other factors affect the decision-making of convenience stores in store expansion. We plan to use the benefits of industry data analysis to explore which conditions are the factors that affect convenience store expansion. It is expected that the empirical results of this study can provide convenience store operators with the direction to think about their business decision making.

This study takes the chain convenience store B in Country-A as the object, referred to as "SEP" in short. The number of store expansions since 2014 has entered a rapid outbreak stage. Since 2016, about 300 stores have been added on average every year, and the total number of stores has reached 2386, among which 1866 are in Islands C, 906 are located in Metropolitan Area D, 331 in Islands E, and 189 in Islands F. Franchised stores (FSs) account for 54% of all stores, while the remaining 46% are retail stores, far outpacing the total number of other supermarkets in Country-A, such as MINISTOP, Family Mart, LAWSON, and Circle K stores [2]. In order to enhance and understand the study context described for readers, it is necessary to define the FS in this study first. The FS is a chain store or branch store assisted by the headquarters to guide technology and operation management and owned by a local operator; the headquarters charge a certain percentage of royalties and guidance fees. The headquarters and the franchisees will jointly operate and share profits to create a win-win situation. Two types of FS are FS1 and FS2. The former needs to prepare their own stores to join and pay rent, and the company and the franchisee jointly invest to reduce the business risk of the franchisee's business. The latter is a license chain, and the company provides the cost of storefront and all equipment decoration and entrusts the franchisee to operate full-time.

This study uses the store data of the SEP company in 2018 for its research, hoping to use the real industry data mining and analysis method as the prediction tool of chain convenience store expansion. The purpose of this study is as follows: (1) To analyze whether the store operation will create a net profit by means of industry data. (2) To evaluate and select a new store (site) for expansion from specific areas by means of industry data. (3) To study the prediction accuracy of Bayes net, logistic regression, and decision tree classifiers. (4) To study the influence of new chain convenience store selection in data preprocessing of machine learning on the prediction accuracy. (5) To study the influence of random sample validation and mixed cross-validation on the accuracy of machine learning.

The remaining structure of the paper is as follows: Section 2 consists of a literature review, including some main issues such as the chain convenience store and its applications and industry data mining. Section 3 describes the proposed mixed classification model research steps and examples. Section 4 comprises the analysis of empirical results, and Section 5 contains our conclusions with several topics, such as the discussion, research findings, research contributions, research limitations, and future research.

2. Literature Review

This study mainly uses the industry data mining technique for analysis, to understand how the chain convenience store B in Country-A makes the location selection decisions and quickly occupies the market. This section introduces the international chain convenience store and its applications, industry data mining, attribute selection technique, data discretization technique, and classification algorithm.

2.1. Chain Convenience Store and Its Applications

In order to establish the operation management system, the chain retail convenience stores usually establish a set of business performance measurement systems that can be followed, to evaluate the current stores and select new stores in the future to improve business performance and competitive advantage. The store location is considered to be one of the most important factors in the business expansion of convenience stores because it is a long-term decision [3]. In store-based retailing, a good location is a key factor in attracting customers to the store and can sometimes even make up for a mediocre retail expansion application. Therefore, a good geographical location can bring a strong competitive advantage because the geographical location is considered to be one of the "unique" elements of retail operation and cannot be imitated by competitors. The importance of store location to retailers should not be underestimated. Why is the store location an important decision for retailers? First, the location is often one of the most influential considerations in a customer's store selection decision. For example, a working couple can easily decide to shop at the store closest to their bus stop on their way home from work [4]. The selection of store location usually requires the retailer to make extensive decisions [5], as there are many factors to be considered. These include the size and character of the surrounding population, the level of competition, the mode of transport, the parking spaces and nearby shops, the cost and duration of rent, and legal restrictions. After the store location has been selected, it should be regularly surveyed in detail. A thorough analysis of the location can provide multiple advantages to the retailer. It includes an understanding of the population size and characteristics, economic fundamentals, competitive dynamics, availability of store location, regulations, and labor availability in the area. According to the traffic flow at the store location [6], one of the most important factors affecting store sales is the number of vehicles and pedestrians passing through the location (traffic flow). Retailers often use traffic statistics to evaluate the attractiveness, thus optimizing the store performance [7]. Geographic information system (GISs) are an important support system for location research and trading area analysis. These software systems combine digital maps with key location data to map trading area characteristics, such as demographic data, customer purchase data, and competitor locations [8]. The site-specific applicability of the store location evaluation technique is based on the retailer's features (such as retail type, merchandise, and pricing strategy, etc.) and is influenced by a number of factors that need to be investigated. To guide the retail location decisions and to assess or predict the potential sales or profitability of retail stores in specific regions, districts or locations, a variety of technique assessment websites have been developed. Thus, the store location determines the potential demand for a particular location, as well as other factors that affect the potential sales and profitability [9]; in particular, the other factors may include transportation costs between customers and locations (e.g., distance or travel time) and characteristics of competitors (e.g., pricing of goods or company image). Regardless of the type of chain store, it is a basic requirement to be close to customers, such as opening in

shopping malls and stations, and location should save customers shopping time. It is also necessary to understand the shopping psychology of customers. For example, customers want to buy all the necessities of daily life at one time. The distribution centers of general chain stores usually adopt the method of unified procurement and centralized supply. This is in exchange for volume discounts and reduced procurement costs. Reasonable planning of transportation routes can reduce transportation costs and obtain economies of scale. Correct product positioning can help companies gain a deep understanding of consumer needs. Formulating a marketing strategy can better serve the target market and establish a good image of the product in the minds of consumers and gain enough target consumers through favorable purchase channels and innovative services.

An analysis of the previous literature on the factors that affect the exhibition store can help the research obtain the relevant important attributes and decision-making attributes required by the convenience store chain. In addition to good products and services to open a new store, the most important thing is to set the evaluation and selection of a new location. It is not only necessary to choose a good location, but also essential to understand the analysis of competitors and peers. Given the above past literature review, we can more accurately understand the key attributes of exhibition stores. It is very important to grasp the market positioning according to the attributes and then to formulate commodity strategies. At the same time, it is also significant to understand consumer needs and to win business opportunities from a business model perspective.

2.2. Industry Data Mining

Industry data are invaluable resources for mining valuable knowledge and information from originally unstructured data to create numerous benefits, such as discovering valuable insights from massive data in advance, lowering expensive time-consuming resources and avoiding laborious work for companies, and helping interested parties gain insights for better decision-making. Given the above, it is clear that data mining from industry (e.g., the healthcare industry) is a core part of data analytics and a key discipline in the data science field. For example, Santos-Pereira et al. [10] indicated that several works have suggested the use of data mining tools, to fill the gap in large amounts of complex data, for the challenging and necessary retrieval of knowledge for the successful coverage of healthcare data mining requirements. Particularly in the big data era at present, data mining makes it possible to more easily make better decisions through some advanced data mining techniques. Generally, big data structure from new data sources is too large or too complex to be processed and is unavailable in general to past ordinary computing devices [11]; data size is its first characteristic. As such, regarding the data size for big data structure, it is of relative concern across time and a variety of industries; for example, a total of more than 1 gigabyte of data was generally considered as big data to the available computing power of past decades. However, today it can make sense to consider a big data structure as containing at least 1 terabyte of data. Simply speaking, big data are data with the well-known three V's and consist of greater variety, increased volume, and higher velocity; another characteristic is that these big data cannot be processed by traditional data processing methods or tools [11].

More importantly, the general definition of big data shall at least contain wide properties of the "6 V's" in data management in recent years: volume, velocity, variety, value, veracity, and variability and complexity. In greater detail: (1) Volume represents the bigness of data, which is reported in multiple terabytes. (2) Velocity represents the generated rate of data and the analyzed speed from devices, such as sensors or smartphones. (3) Variety is the data heterogeneity from a given dataset, and the heterogeneity may have various categories, such as structured data, semi-structured data, and unstructured data. (4) Value defines the various types of valuability of an attribute from large volumes of data, and the original form of big data received and analyzed is either of a low value or a high value. (5) Veracity refers to the untrustworthiness inherent in data sources. (6) Variability is the data variation in its flow rate, and complexity represents the data generated from a great

number of sources. Furthermore, statistics and data mining techniques for industry data analysis include significance testing, classification, regression/prediction, cluster analysis, association rule learning, anomaly detection, and visualization. The statistical analysis provides the scientific reasoning for the procession from data to knowledge and then to action, which is crucial to data analysis from industry [11]. In [11], the authors emphasized the interest in developing suitable and efficient data analysis methods to leverage massive volumes of heterogeneous data, such as video formats, audio, and unstructured text. In addition, industry data is also a term to describe amounts of data with complexity-based and variability-based features and requires advanced IT (or machine learning) techniques to capture, store, distribute, manage, and analyze the data given. Industry data mining covers topics such as fundamental questions, classification, clustering, trends, bias analysis, next-generation database system, and applied case studies, with the contributors including researchers from academia, government, and industry [12]. In the past decade, there has been an explosion in the generation and collection of data, e.g., the widespread use of bar codes for most commercial products and the computerization of many commercial and government transactions, which has left us awash with data and in desperate need of new techniques and tools that can help automate the transformation of data into useful knowledge.

Given the above, it is understandable that the significance of industry data is not only seen in the production and mastery of useful data information but also in the professional processing of valuable data. In the era of continuous emerging big data, data visualization becomes easier and clearer with a systematic view. For example, industry data are applied in marketing research to analyze the population of each district, the consumption power of each district, the income level of each district, etc. Further, they can estimate the number of visitors and products that may be purchased in each region and calculate the potential turnover. This research is based on the above concepts and practices and studies the key factors of the new exhibition store. It is of urgent need to use data mining technology in specific industries to find out the key factors for the success of opening a new store and use rational data to evaluate the benefits.

2.3. Attribute Selection Technique

The "attribute selection" strategy for data preprocessing has been proven to be effective for all kinds of data mining and data preparation (especially high-dimensional data) and machine learning problems. The goals of attribute selection include building simpler and easier-to-understand models, improving data mining performance, and preparing clean and understandable data. The recent growth of industry data mining has presented some substantial challenges and opportunities to select the feature. It has been widely used in many research fields, such as statistical pattern analysis, machine learning [13], and data mining [14]. Armanfard et al. [15] explained that there are three reasons for using the attribute selection technique: to simplify the model, to be easy to understand, to shorten the training time and to improve the universality. Chandrashekar and Sahin [16] explained that attribute selection helps in solving the problem of too much low-value data and too little high-value data and assists to reduce the computing time, improve the prediction performance, and better understand the machine learning or pattern recognition application. Attribute selection is to search all possible combinations of all attributes in the data set and find out the group of attributes with the best prediction effect. Selecting the best part from the original data and features with good identification ability can not only simplify the calculation but also helps in understanding the causal relationship of the problem [17], which is an important part of machine learning. Attribute selection has many advantages, including: (1) Data collection: reduce the resource costs and make the data clear and easy to see. (2) Data preprocessing: delete redundant attributes to make the calculation of model building more efficient and simplify the model. (3) Data interpretation: after the attribute selection method, the interference of prediction results is improved, the explanatory ability is enhanced, and the model derivation and knowledge mining are accelerated [18]. The

purpose of attribute selection is to find out the most relevant classification features, reduce the accuracy technique of dimensionally correct training sample classification, and select the important and effective condition data from the data sources.

In summary, attribute selection is the process of removing irrelevant features or attributes according to the specified feature measurement conditions to select the best features for analyzing data in a given data set. This attribute selection work is very important to deal with classification work for grouping problems; thus, this study aims to use attribute selection technology for data preprocessing to simplify the model experiment and improve model performance in advance.

2.4. Data Discretization Technique

Discretization is a basic preprocessing technology [19], and the most influential data preprocessing task, which aims to take the concise data as a category suitable for learning tasks and convert the digital attributes into discrete data, making it easier for experts to understand [20]. Advanced data mining algorithms are considered to be the correct and most useful processing methods [21]. There are two discretization methods: according to the expert's personal judgment, the attribute can be changed to the classification distance, which is convenient to understand the result, namely expert discretization; to ensure the correctness of numerical values, different equations are used to perform the automatic data cutting, namely data discretization. In the data mining research, due to the limitation of huge data and resources, expert discretization cannot show the whole picture of results, so the automatic discretization becomes the favorite of researchers. It can simplify the calculation and data structure research, reduce the time of finding the rules, and reduce the complexity of rules. Discretization is a continuous quantization process. Continuous values are the most common rules obtained by being short, compact, and accurate, and the results are easier to be checked, used, compared, and reused. The purpose of data discretization is to set several cutting points within a specific continuous numerical range (e.g., L, M, and H represent low, medium, and high, respectively). Another function of data discretization is its ability to reduce the number of knowledge rules generated, improve the performance of classifiers, and cut off the continuous attribute value of suitability more effectively and efficiently.

Generally speaking, the advantages of discretization can meet the needs of the classification algorithm. For example, algorithms such as Bayes net cannot directly use continuous variables, and continuous data can only enter the algorithm engine after being processed by the discretization technique. Therefore, it is appropriate to this study to preferentially use the discretization technique as a data preprocessing method before the data mining stage in this study.

2.5. Classification Algorithm

Data mining mainly includes classification, association analysis, and clustering. Whether it is classification, association analysis, clustering or other data mining tools, the advantages and disadvantages of a data mining method are mainly measured from three aspects: (1) analysis efficiency, (2) operation efficiency, and (3) result interpretation. So far, few data mining methods have absolute advantages in these three aspects; thus, we should understand a data mining tool from these three aspects and choose the appropriate analysis tool to perform the data mining according to our own needs.

Thus, the classification mining tools selected and used in this study include Bayes net, logistic regression, and decision tree, since they have been commonly used in past academic research with superior performance. For example, a past study application included research on predicting bank depositor's behavior [22], which developed the accuracy of a classifier on a real bank dataset to use telemarketing to predict the sales potential of customers ordering bank long-term deposit services; this past research was based on decision tree J48 to remove unnecessary feature models through dimensionality reduction. Moreover, past research [23] on analytical models determined job applicants.

In [23], the authors used the technology of supervised and unsupervised machine learning classifiers to solve the problem that the human resource management department of an enterprise finds suitable college graduates through employment status prediction, and their final research results showed that the accuracy of logistic regression is better than other algorithms. In particular, the above two studies also had used these three classifiers, and their classification accuracy rates have a good performance. Thus, based on the above reasons, the three techniques are selected and used for identifying applications of a new chain convenience store for the purpose of data mining works. In this study, the three classification algorithms used for data analysis are described below:

(1) Bayes net: a Bayes net definition includes a directed acyclic graph (DAG) and a set of conditional probability tables. Explanation of DAG: DAG has no ring, no turning back, never turning back, and just moving forward. DAG can be redrawn, so that all edges extend in the same direction and all points have a sequence. Each node in DAG represents a random variable, which can directly observe the variable or hide the variable, while the directed edge represents the conditional dependence between the random variables. Each element in the conditional probability table corresponds to a unique vertex in the DAG and stores the joint conditional probability of this vertex for all its immediate predecessors. The training of Bayes net is divided into the following two steps: (1) to determine the topological relationship between random variables and form the DAG, which usually requires the domain experts to complete, and in order to establish a good topological structure, it usually requires repeated operations and improvements; (2) to train the Bayes net: this step is to complete the construction of the conditional probability table. If the value of each random variable can be directly observed, the training in this step is intuitive, similar to Naive Bayes classification. However, there are hidden variable vertexes in Bayes net, so the training methods are complicated, such as the gradient descent method. Research on Bayes net has been applied in various fields, such as intelligence quotient [24], learning style [25], patient discharge [26], and cervical cancer [27]. Bayes net can combine data with expert knowledge judgment and not only has the ability to predict but also can perform calculations for uncertain problems, showing the correlation between variables. For example, Bayes net is used in the prediction of consumer review analysis [28]. The Bayes net was chosen for this study experiment because it performed well in various domains when applied to classification tasks from past studies.

(2) Logistic regression: logistic is similar to linear regression analysis, mainly discussing the relationship between dependent variables and independent variables. The dependent variable (Y) in linear regression is usually a continuous variable, but the dependent variable (Y) discussed in logistic regression is mainly a category variable, especially variables divided into two categories (e.g., yes or no, with or without, agreed or disagreed, etc.). Logic regression is a statistical method used to analyze data sets with binary dependent variables (binary system). It can be used to find a relationship between a dependent binary variable and one or more independent variables. Each independent variable is multiplied by the weight and summed. This result is added to the sigmoid function to find the result between 0 and 1. Values above 0.5 are treated as 1; values below 0.5 are treated as 0, and it is important to find out the best weight or regression coefficient. Therefore, optimization techniques are used to find the optimal regression coefficient and weight [29]. Logistic regression does not require much computing resources, so it is widely used. In particular, from the study of Demidenko [30], it is indicated that there is no consensus on the approach to calculate the computational power resource and data sample size with logistic regression. Thus, the problem of defining unknown data sample size with power is important in different industry fields, especially in cases of expensive measurements of industrial applications. In [30], a Wald-based power and data sample size formula had been derived for logistic regression and proposed to minimize the total data sample sizes within a case–control study in order to obtain a power given by

optimizing the ratio of control cases; as a result, the optimal control cases are equal to the square root of the alternative odds ratio. Moreover, Motrenko et al. [31] treated the parameters of a regression model as a multivariate variable and used the distance of parameter distribution functions on cross-validation datasets to measure the data sample size, and they supported an applied mathematics contribution to data mining and statistical learning. Interestingly and importantly, it is also a key issue to address and describe the problem of related resource consumption in actual analysis when the data size is unknown. For example, the World Wide Web has grown and collected new data from being an active industry platform and has arrived at immense speed to a big data, and this situation has spawned a specialized computing basis in the data paradigm, where the massive amount of given data must be processed within a reasonable response time in order to handle its high velocity; thus, this spurred several ideas for processing fast data interests [32]. In [32], the authors proposed and optimized the prototype of their so-called Raincoat query, and they demonstrated the significant contribution in increasing the performance of a Raincoat query; concurrently, they used the random sampling method to estimate cardinality and to store less data compared to histograms, and this method is especially important since the total data size is not known like with data stream structure, which can take advantage of less time to compute the estimate data. Regarding data streams of a big data framework [33], many techniques such as different sampling algorithms have developed to overcome the dilemma between resource consumption and unknown data size and some have shown superior performance in recent years. In particular, from the research of Cardellini et al. [34], it is indicated that dealing with unbounded dataflows, data stream application is typically long running and thus likely experiences varying workloads and working conditions over time. Looking back to logistic regression, it is easy to interpret, requires no adjustment of input features and is easy to normalize, and its output is a well-corrected predicted probability. For example, the application in predicting waiting and treatment times in emergency departments [35] has a good performance. The reason for choosing the logistic regression method in this study is that it also has excellent performance in many application fields of research, so it is adopted and also used as one target of the supervised classification tools.

(3) Decision tree: decision tree is the main application tool in the field of data mining, and its method can be verified and segmented from root in order. Each branch tree represents a verification result, and the leaf node displays the distribution state of target variables, which is finally presented in the form of the tree. Each path from root to leaf node can extract a decision tree rule, and the data can be classified into a tree structure through the selection of different variables and the designation of targets, and a classification system or prediction model with hierarchical structure can be presented [36]. A variety of decision tree algorithms have been developed in modern times, mainly including Chi-squared Automatic Interaction Detection (CHAID) [37], Classification and Regression Tree (CART) [38], Interactive Dichotomiser 3 (ID3) [39], and C4.5 [40]. Advantages of decision tree classification: the classification rules are easy to understand, the data processing time is not too long, and it can also process the string data in category. Disadvantages: it is difficult to process the continuous string data type, and the data type of time series needs to be discretized first. When there are too many data types, the error rate increases rapidly. The decision tree is easy to understand and easy to extract the characteristics of the rules, which can make the research results easier to interpret through visual graphics, such as with the classification and prediction of student grades [41]. According to the performance of the decision tree, it is suitably used for comparison with other classification methods in this study.

3. Research Methods

This section introduces the framework and research steps of the proposed mixed classification models and takes the chain convenience store B in Country-A as an example to illustrate the process of exploring the prediction model of chain convenience store location selection and store expansion through data mining techniques.

3.1. Research Framework for the Proposed Mixed Models

This study takes the chain convenience store B in Country-A as an example to study the new store location selection and analyze its new store expansion application, so as to understand the results of the country's retail stores. Based on the performance and location data of chain convenience store B, using the data mining software as an analysis tool, this study adopts the binary classification method to construct the decision tree analysis model and takes the data from January to December 2018 as the analysis target, combined with the local climate, demographic statistics, and business circle conditions. The prediction model is used for analysis, and the rules are found with Bayes net, logistic regression, and decision tree classifiers through random proportion sampling and mixed cross-validation, to assist the logistics personnel to make decisions for store expansions through the analysis method. The research framework is shown in Figure 1. Importantly and interestingly, this study adopts a concept of hybrid procedure of named 3-4-8-2 components combination to appropriately process an experience policy in order to run the proposed mixed model with empirical case. That is, for highlighting the combination procedure of 3-4-8-2 components experiment execution stage in detail, 3 good target classifiers (Bayes net, logistic regression, and J48 decision tree) and 4 different data preprocessing models (without preprocessing, with attribute selection, with data discretization, and with attribute selection and data discretization) are experienced into 8 different experiments by using 2 data-partitioning methods (percentage-split and cross-validation) for identifying the application of new chain convenience store selection.

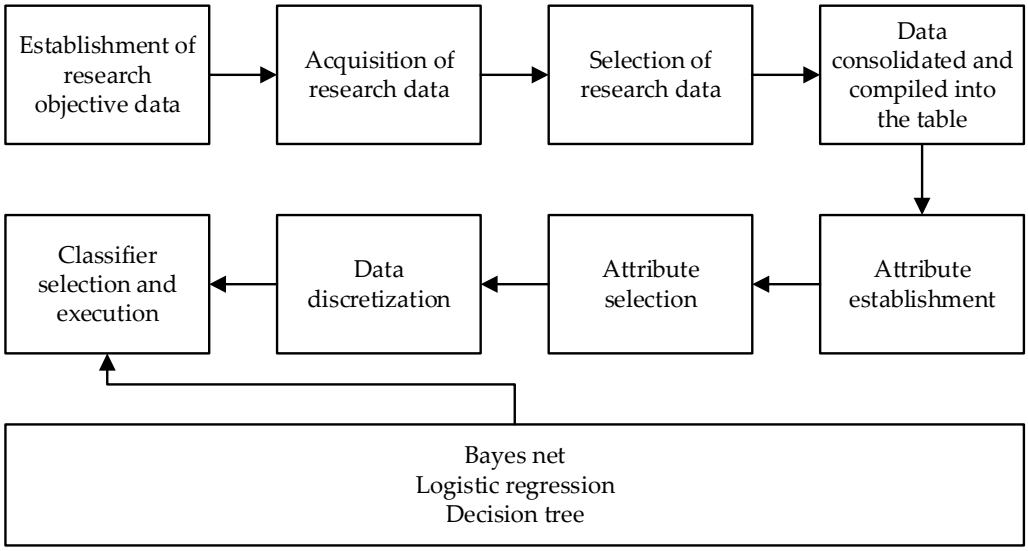

Figure 1. Research framework.

3.2. Mixed Classification Model Research Steps and Examples

In this study, the prediction model for the application on new chain convenience store location selection and new store expansion has the following eight key steps, and an actual example is used to experience in detail as follows:

Step 1: Establishment of research objective data. This study establishes the research direction and data collection by means of expert and field investigation. Through the data extraction and predictive analysis, this is an important step in data mining. Climate and regional population development are taken as the basis of data analysis for the location

selection of store expansion. Country-A's land is widely distributed, and since it is located in the subtropical areas, the topography of each region is greatly affected by climate. This also affects the decisions of store expansion, especially in the regions directly facing the possibility of typhoons and prone to be affected by flooding, landslides, and road collapse, which can lead to inaccessible logistics distribution; thus, a store in Islands C, Islands E, and Islands F is difficult to set up. In terms of regional population development, the Human Development Index (HDI) is a standard issued by UNDP since 1990 to measure the socio-economic development degree of countries, and countries are divided into four groups: very high, high, medium, and low. Only countries in the first group of "extremely high" are likely to become developed countries. The index values are calculated on the basis of life expectancy at birth, years of schooling (including the average and expected years of schooling), and per capita gross national income (GNI), and can be used worldwide for comparison between countries.

Step 2: Acquisition of research data. After confirming the collected data through local investigation and discussion with the IT department of chain convenience store B, the time interval from January to December 2018 is taken as the analysis range, and then the attributes of each data field are established.

Step 3: Selection of research data. The data selection is given priority to the existing store information; the data fields are the store type, zone, district, business circle, human development index, population, climate, population density coefficient, number of POS, number of cashiers, and monthly average daily performance in 2018 (a total of 22 conditional attributes), while the net profit and store expansion are the two types of decision attributes, and the recommendation for data feature of decisional attribute are made and provided by experts. For the net profit, there are two classes classified: Y (referring to have positive profit) and N (referring to negative profit–loss). Correspondingly, there are also two classes classified for new store selection (abbreviated as NSS in this study): Y refers to that this store can be selected for an expanded new-site location and N refers to do not select this site. In detail, Table 1 lists the information of 22 conditional features and two types of decisional attributes in order to identify and select applications of a new chain convenience store.

Table 1. Store information.

Item	Domain Name	Description	Attribute
01	Store Type	Store franchise type	Text
02	Zone	Large zone (divided by island)	Text
03	District_name	District name	Text
04	Cluster_name	Business circle	Text
05	HDI	Human development index	Number
06	Climate type	Climate coefficient	Number
07	Population	Population	Number
08	Population density	Population density coefficient	Number
09	POS	Cash register	Number
10	Cashier	Store staff	Number
11	PSD(JAN)	Average daily performance of January	Number
12	PSD(FEB)	Average daily performance of February	Number
13	PSD(MAR)	Average daily performance of March	Number
14	PSD(APR)	Average daily performance of April	Number
15	PSD(MAY)	Average daily performance of May	Number
16	PSD(JUN)	Average daily performance of June	Number
17	PSD(JUL)	Average daily performance of July	Number
18	PSD(AUG)	Average daily performance of August	Number
19	PSD(SEP)	Average daily performance of September	Number
20	PSD(OCT)	Average daily performance of October	Number
21	PSD(NOV)	Average daily performance of November	Number
22	PSD(DEC)	Average daily performance of December	Number
23	NSS	Evaluating the new store selection	Text
24	Profit	Evaluating the net profit (earning)	Text

Step 4: The data are consolidated and compiled into the table. The related data information for the store data set (original source of data) used is shown in Table 2. The raw

data collected from the Country-A data set has 2231 instances regarded as experimental targets. Afterwards, they are coded into suiting for experiencing and measuring the proposed model performance.

Table 2. Some date set (original source of data).

Store Type	Zone	District_Name	Cluster_Name	HDI	...	PSD(OCT)	PSD(NOV)	PSD(DEC)	NSS	Profit
CO	South Luzon	SOUTH FIVE	Residential	0.649	...	51,213	50,520	64,416	N	N
FC2	Central Luzon	CENTRAL TWO	School	0.649	...	89,316	86,267	97,059	Y	Y
FC3	Central Luzon	CENTRAL THREE	School	0.649	...	41,345	39,879	42,703	N	N
CO	Central Luzon	CENTRAL THREE	School	0.649	...	70,661	24,850	73,793	N	N
SA	Central Luzon	CENTRAL THREE	Transit	0.649	...	88,692	92,361	94,669	Y	N
...	...	...	...	...	...	...	...	...	...	...
CO	Central Luzon	CENTRAL TWO	Residential	0.649	...	80,697	82,987	88,594	Y	Y
CO	Visayas	WESTHERN VISAYAS	Transit	0.749	...	31,770	29,349	37,762	N	N
FC1	North Luzon	NORTH FOUR	Transit	0.749	...	33,321	32,958	39,839	N	N
CO	Visayas	WESTHERN VISAYAS	School	0.749	...	30,278	26,797	28,565	N	N
FC1	South Luzon	SOUTH SIX	Commercial	0.799	...	37,158	34,260	40,511	N	N

Step 5: Attribute establishment. The data are firstly standardized to facilitate the data mining software tools to analyze the climate and population density, etc., and the text is replaced by code names, as shown in Table 3. For example, the climate index is shown in Table 4. Population density is defined as the population density (PD) coded by expert recommendation: less than or equal to 10,000, more than 10,001 and less than or equal to 20,000, and more than 20,001 are distinguished by grading codes PDL (low), PDM (medium), and PDH (high), respectively. Net profit is judged by experts based on financial statements, and the grading codes are N (No profit–Loss) and Y (Profit). NSS is judged by the recommendation of experts based on a more complete and overall consideration of internal and external information and experience, and the grading codes are N (No selection) and Y (Selection).

Table 3. Code description.

Item	Domain Name	Description
01	Store type	Store franchise type
02	Zone	Divided by island
03	District_name	District name
04	Cluster_name	Business circle
05	HDI	Human development index
06	Climate type	Climate coefficient
07	Population	Population
08	Population Density	Population density coefficient
09	POS	Cash register
10	Cashier	Store staff
11~22	PSD (January–December)	Average daily performance of January to December
23	NSS	New store selection decision
24	Profit	Net profit (earning) decision

Table 4. Definition of climate index.

Climate Type	Grading Code
Two distinct seasons; dry from November to April and wet during the rest of the year.	I
There is no dry season; there is very significant rainfall from November to April and it gets wet for the rest of the year.	II
The seasons are not very obvious; relatively dry from November to April and wet during the rest of the year.	III
Rainfall is more or less evenly distributed throughout the year.	IV

Step 6: Attribute selection. Commonly used methods for general data mining analysis mainly include the classification, regression analysis, association rules, characteristics, change, and deviation analysis, respectively, from the different angles for data mining. This study will adopt the classification for analysis. The so-called classification analysis is to divide a set of data into several categories according to their similarity and difference. Its purpose is to make the similarity between the data in the same category as large as possible and the similarity between the data in different categories as little as possible. The attributes selection is to search all possible combinations of all the attributes in the data set and find the group of attributes with the best prediction effect. This study adopts the machine attribute selection, and the selection steps are explained in detail in Section 4.

Step 7: Data discretization. To discretize, simplify, and reduce the complexity of numerical data attributes, this study uses the automatic discretization in the data preprocessing method of the data mining package software.

Step 8: Classifier selection and execution. It is to select the suitable data attributes for analysis, select the suitable classification algorithms for different purposes, conduct the data mining to achieve the best prediction results, and evaluate the results once obtained. In this study, three well-known classifiers (Bayes net, logistic regression, and decision tree), which are commonly used in academic circles, are selected and used for the data mining to establish the prediction model. Firstly, the data sets are divided into the training and testing, and randomly divided into the training data sets and testing data sets for machine learning. It is a purpose of this study that the classification performance is evaluated to find out the most suitable model prediction method. Moreover, all the data are loaded into the analysis software for calculation. In the training and testing mode, the data mining software will randomly sample 67% of the data for training and then bring the remaining 33% into the test, and finally obtain the accuracy of the test data. The main flow chart of machine learning related to the proposed mixed models is shown in Figure 2.

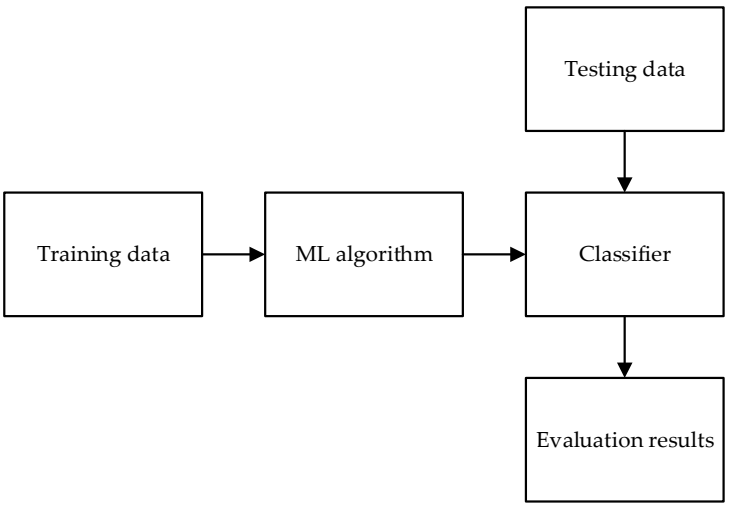

Figure 2. Flow chart of machine learning related for the proposed mixed models in this study.

4. Analysis of Empirical Results

Based on the research steps and examples of the mixed classification model established in the previous section, different experimental situations are constructed through the combination procedure of highlighting 3-4-8-2 components experience policy to further verify the contributions and research findings. Four main research steps are taken, and the experimental results and findings are described in two different decision attribute categories in each step.

4.1. Empirical Step Description

In this study, the data mining tools are used to analyze the data of classification models. This section performs the data analysis and records the results as follows: Step 1: Load the cleaned data. Step 2: Preprocess the data; verify by attribute selection and data discretization, respectively. Step 3: Execute the selected classifiers; in this study, the accuracy of Bayes net classifier, logistic regression, and decision tree are analyzed and compared. Step 4: Output the empirical results.

Step 1: Load the cleaned data. Since there is incomplete information or error situations in the obtained information, it needs to screen the information first for data cleaning. In this study, the data set used is first formatted in export text of a CSV file in Microsoft Excel, and then it is imported into a data mining tool in Weka package software, which was operated for processing data standardization analysis. Many data mining algorithms in Weka have various functions, such as calculate data similarity, execute classification, and do clustering, etc. For example, when calculating the similarity of data of multi-dimensional variables, the data units of each dimension are different, resulting in different full distances. If the unit is ignored and the similarity is calculated numerically, it will happen that the dimensional attribute with a larger full distance has a great influence on the similarity. In using Weka software, the default parameters of most algorithms are initially and automatically normalized. In order to overcome and standardize these problems of fair comparisons, this study adopts and uses all Weka's default parameters to normalize or standardize each dimension, which is a reasonable and reliable approach.

Step 2: Preprocess the data. The research direction is established through the advice of new chain convenience stores and retail experts, and the empirical results are compared. The research data range is the stores with complete performance of each convenience store in 2018, with a total of 2231 research data. This study takes 22 conditional attributes and two types of decision attributes as the data analysis objects: the conditional attributes include the store type, zone, district, business circle, human development index, climate coefficient, population, PD coefficient, number of POS, number of cashiers, and PSD (Per Store Daily) from January to December; net profit and NSS are the decision attributes, respectively. There are two main techniques used for the data preprocessing stage: (1) Attribute selection: data mining software adopts the supervised machine learning for attribute selection, and the selected conditions are judged directly by the machine. (2) Data discretization: data discretization is carried out also by the unsupervised machine learning with data mining tools.

Step 3: Execute the selected classifier. In this study, three classifiers, Bayes net, logistic regression, and J48 decision tree, are used for data mining, and models are established and evaluated after data analysis. The following two data verification (data-partitioning) methods are adopted:

(1) The data sets are divided into the training and testing data, and the optimal prediction model is found through machine learning techniques. The data mining tool randomly samples and trains 67% data and tests the remaining 33% data to verify the accuracy of the data analysis.
(2) K times (folds) of cross-validation: the training set is divided into K subsamples and one single subsample is reserved as the data to verify the model, and the other K-1 samples are used for training. Cross-validation is repeated for K times, once for each subsample, averaging the results of K times or using other combinations,

resulting in a single estimate. The advantage of this method lies in the repeated use of randomly generated subsamples for training and validation at the same time, with each result verified once. In this study, 10 times of mixed cross-validation are adopted and conducted.

Step 4: Output the empirical results. According to the collected data set, through the above experimental steps, the empirical results are produced, and a conclusion with meaningful information/knowledge is made.

According to the above research steps, highlighting components of various combinations of processing a called 3-4-8-2 experience procedure with clear definitions, including the data mining techniques (attribute selection and data discretization), classifiers (Bayes net, logistic regression, and J48–C4.5 algorithm), validation methods (percentage split and cross-validation), conditional attributes (including daily average performance), and decision attributes (net profit and NSS), are integrated into four types of models (such as A1~A4, B1~B4, . . . , G1~G4, and H1~H4), respectively, and eight experimental situations, as shown in Table 5.

Table 5. Eight experimental situations.

Experiment Number	Model Number	Attribute Selection	Data Discretization	Classifiers	Percentage Split	Cross-Validation	Daily Average Performance	Protfit	NSS
1	A1			V	V		V	V	
	A2	V		V	V		V	V	
	A3		V	V	V		V	V	
	A4	V	V	V	V		V	V	
2	B1			V		V	V	V	
	B2	V		V		V	V	V	
	B3		V	V		V	V	V	
	B4	V	V	V		V	V	V	
3	C1			V	V		V		V
	C2	V		V	V		V		V
	C3		V	V	V		V		V
	C4	V	V	V	V		V		V
4	D1			V		V	V		V
	D2	V		V		V	V		V
	D3		V	V		V	V		V
	D4	V	V	V		V	V		V
5	E1			V	V			V	
	E2	V		V	V			V	
	E3		V	V	V			V	
	E4	V	V	V	V			V	
6	F1			V		V		V	
	F2	V		V		V		V	
	F3		V	V		V		V	
	F4	V	V	V		V		V	
7	G1			V	V				V
	G2	V		V	V				V
	G3		V	V	V				V
	G4	V	V	V	V				V
8	H1			V		V			V
	H2	V		V		V			V
	H3		V	V		V			V
	H4	V	V	V		V			V

4.2. Empirical Results

Due to the length limitation, only the decision trees of Experiment 1 to Experiment 4 and three rules are listed here. The research steps of Experiment 5 to Experiment 8 are similar to Experiment 1 to Experiment 4, so they are omitted.

(1) Experiment 1: net profit is the decision attribute, including the conditional attribute of daily average performance. 67% data are randomly sampled for training, and the remaining 33% data are for testing.

Model description:

Model A1: without attribute selection and without data discretization. Model A2: with attribute selection (Store type, Cashier, APR, MAY, JUN, AUG, and DEC). Model A3: with data discretization. Model A4: with attribute selection and with data discretization. After the experiments, the comparison table and statistical graph of Models A1-A4 net profit decision are shown in Table 6 and Figure 3, respectively.

Table 6. Models A1-A4 net profit decision (comparison table).

Model	Bayes Net (%)	Logistic Regression (%)	J48 (%)
A1	88.9946	86.1413	91.0326
A2	90.2174	87.0924	90.4891
A3	89.5382	89.1304	91.0326
A4	90.7609	90.3533	91.0326

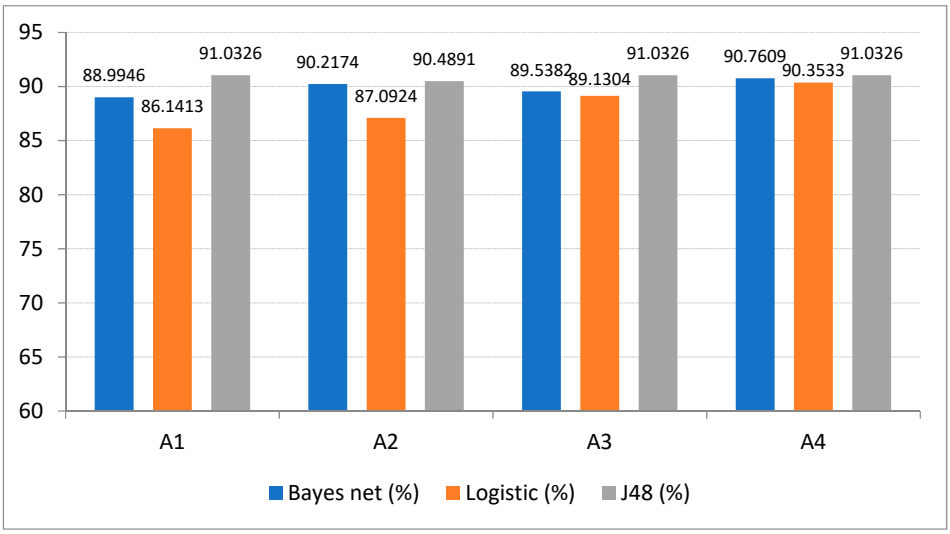

Figure 3. Models A1–A4 net profit decision (statistical graph).

Results of Experiment 1:

Bayes net classifier: Model A4 has the highest accuracy of 90.7609%, and Model A1 has the lowest accuracy of 88.9946%.

Logistic regression classifier: Model A4 has the highest accuracy of 90.3533%, and Model A1 has the lowest accuracy of 86.1413%.

J48 decision tree classifier: Model A2 has the lowest accuracy of 90.4891%, the other three models have an accuracy of 91.0326%. J48 decision tree takes Model A2 as an example. The decision tree is shown in Figure 4.

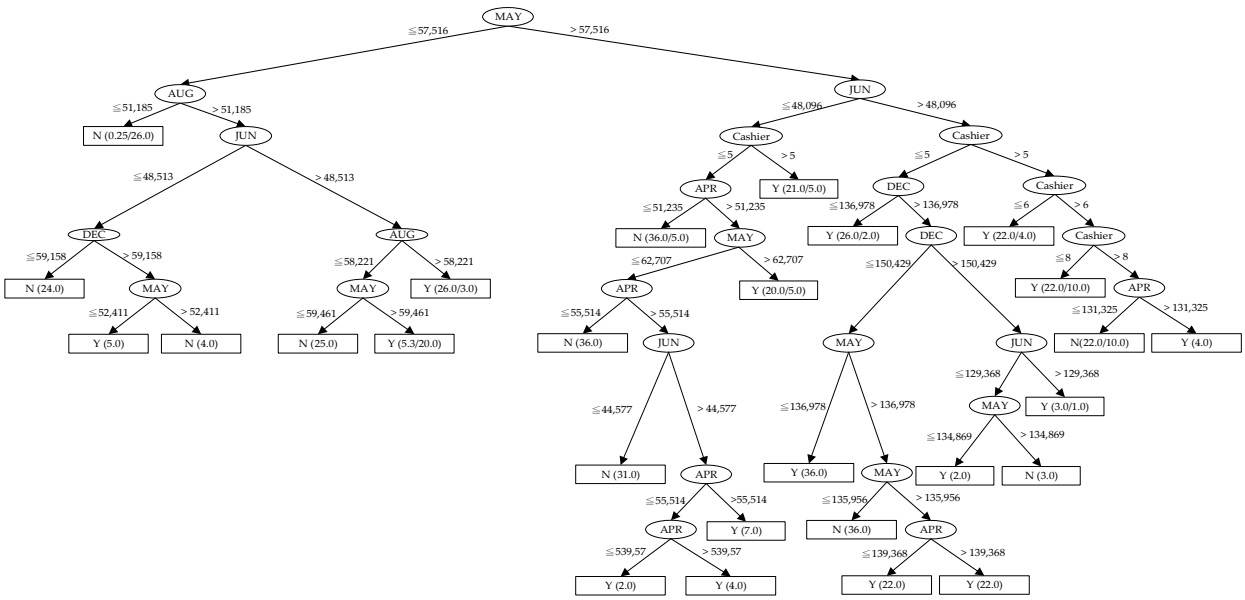

Figure 4. Model A2 J48 tree.

Rule description:

Rule 1: IF PSD MAY > 57,516 and JUN > 48,096 and 5 > Cashier ≤ 6 THEN Profit = Y.

Assuming that the daily performance is more than 57,516 in May and 48,096 in June, and the number of cashiers is more than 5 and less than or equal to 6, it is predicted that the store in this area will create a net profit.

Rule 2: IF PSD MAY > 57,516 and JUN > 48,096 and Cashier > 8 and APR > 131,325 THEN Profit = Y.

Assuming that the daily performance is more than 57,516 in May and 48,096 in June, and the number of cashiers is more than 8, and the daily performance is more than 131,325 in April, it is predicted that the store in this area will create a net profit.

Rule 3: IF PSD MAY > 57,516 and JUN > 48,096 and Cashier ≤ 5 and DEC ≤ 136,978 THEN Profit = Y.

Assuming that the daily performance is more than 57,516 in May and 48,096 in June, and the number of cashiers is less than or equal to 5, and the daily performance is less than or equal to 136,978 in December, it is predicted that the store in this area will create a net profit.

According to the analysis results of J48 decision tree, after attribute selection is included, the condition of the number of cashiers is added to the important attributes, and the store performance in April, May, June, and August is an important factor to predict whether the store in this area will create a net profit.

In Experiment 1, it is found that the J48 decision tree classifier has a higher average accuracy, and Model A1 logistic regression has the lowest accuracy of 86.1413%. In terms of the model, the accuracy of the model without attribute selection or data discretization is lower, while the accuracy of the model with attribute selection is generally higher. J48 decision tree has the highest average accuracy.

(2) Experiment 2: profit is the decision attribute, including the conditional attribute of daily average performance, by cross-validation (10-fold mixed cross-validation). Model description:

Model B1: without attribute selection and without data discretization. Model B2: with attribute selection (Cashier, APR, MAY, JUN, AUG, and DEC). Model B3: with data discretization. Model B4: with attribute selection and with data discretization. The comparison table and statistical graph of Models B1-B4 net profit decision are shown in Table 7 and Figure 5.

Table 7. Models B1-B4 net profit decision (comparison table).

Model	Bayes Net (%)	Logistic Regression (%)	J48 (%)
B1	89.1528	87.0013	90.8113
B2	90.5424	86.7772	90.2734
B3	89.9148	89.6459	89.6459
B4	90.9458	90.5424	90.7665

Results of Experiment 2:

Bayes net classifier: Model B4 with attribute selection and data discretization has the highest accuracy of 90.9458%, and Model B1 without preprocessing has the lowest accuracy of 89.1528%.

Logistic regression classifier: Model B4 with attribute selection and data discretization has the highest accuracy of 90.5424%, and Model B1 without preprocessing has the lowest accuracy of 89.1528%.

J48 decision tree classifier: Model B1 has the highest accuracy of 90.8113%, and Model B2 has the lowest accuracy of 90.2734%. J48 decision tree takes Model B1 as an example. The decision tree is shown in Figure 6.

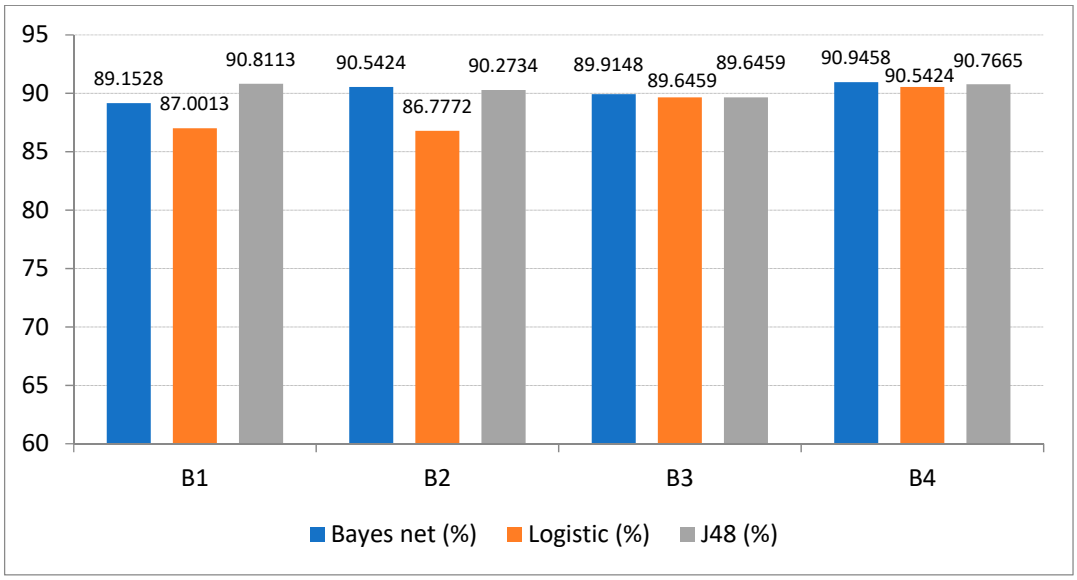

Figure 5. Models B1-B4 net profit decision (statistical graph).

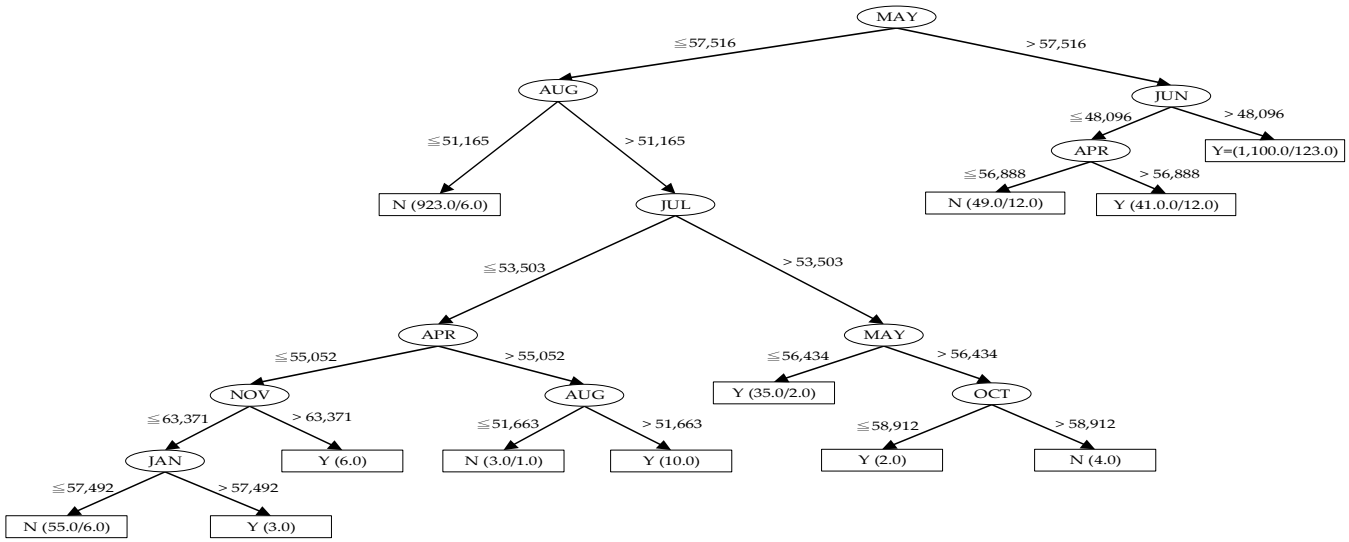

Figure 6. Model B1 J48 tree.

Rule description:

Rule 1: IF PSD MAY > 57,516 and JUN > 48,096 THEN Profit = Y.

Assuming that the daily performance is more than 57,516 in May and 48,096 in June, it is predicted that the store in this area can create a net profit.

Rule 2: IF PSD MAY > 57,516 and JUN ≤ 48,096 and APR > 56,888 THEN Profit = Y.

Assuming that the daily performance is more than 57,516 in May, less than or equal to 48,096 in June, and more than 56,888 in April, it is predicted that the store in this area can create a net profit.

Rule 3: IF PSD MAY ≤ 57,516 and AUG > 51,165 and JUL > 53,503 and OCT ≤ 58,912 THEN Profit = Y.

Assuming that the daily performance is less than or equal to 57,516 in May, more than 51,165 in August, more than 53,503 in July, and less than or equal to 58,912 in October, it is predicted that the store in this area can create a net profit.

According to the analysis results of J48 decision tree, based on the tree structure, the important attributes include PSD in January, April, May, August, and November, which are the important factors to predict whether the store in this area can create a net profit.

In Experiment 2, it is found that Model B1 (without preprocessing) has the highest accuracy through cross-validation, while Model B2 preprocessed with attribute selection has the lowest accuracy. In this experiment, it is found that the cross-validation method has no significant improvement in the accuracy after data preprocessing, whether it is with attribute selection or data discretization. However, compared with the other two classifiers, the accuracy of J48 is relatively stable and high.

(3) Experiment 3: NSS is the decision attribute, including the conditional attribute of daily average performance. 67% data are randomly sampled for training, and the remaining 33% data are for testing.

Model description:

Model C1: without attribute selection and without data discretization. Model C2: with attribute selection (Cashier, MAR, APR, MAY, JUN, SEP, and DEC). Model C3: with data discretization. Model C4: with attribute selection and with data discretization. The accuracy of Models C1~C4 is calculated by three classifiers: Model C1 (Bayes net = 93.6054%, Logistic regression = 93.0612%, J48 = 94.5578%); Model C2 (Bayes net = 94.6939%, Logistic regression = 93.3333%, J48 = 94.6939%); Model C3 (Bayes net = 93.3333%, Logistic regression = 93.8776%, J48 = 94.2857%); Model C4 (Bayes net = 93.8776%, Logistic regression = 94.6939%, J48 = 94.5578%). The statistical graph of Models C1-C4 new store selection decision is shown in Figure 7.

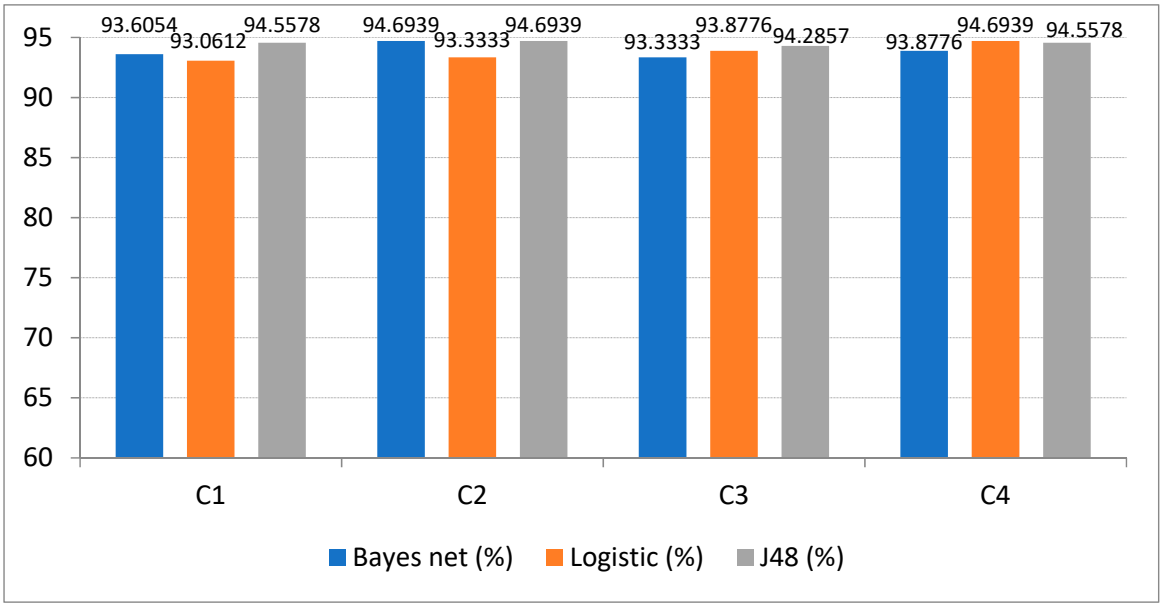

Figure 7. Models C1-C4 new store selection decision (statistical graph).

Results of Experiment 3:

Bayes net classifier: Model C2 has the highest accuracy of 94.6939%, and Model C3 has the lowest accuracy of 93.3333%.

Logistic regression classifier: Model C4 has the highest accuracy of 94.6939%, and Model C1 has the lowest accuracy of 93.0612%.

J48 decision tree classifier: Model C2 has the highest accuracy of 94.6939%, and Model C3 has the lowest accuracy of 94.2857%. J48 decision tree takes Model C2 as an example. The decision tree is shown in Figure 8.

Rule description:

Rule 1: IF PSD MAY > 57,795 and JUN > 48,096 THEN NSS = Y.

Assuming that the daily performance is more than 57,795 in May and more than 48,096 in June, it is predicted that this area (site) can be selected as a new store.

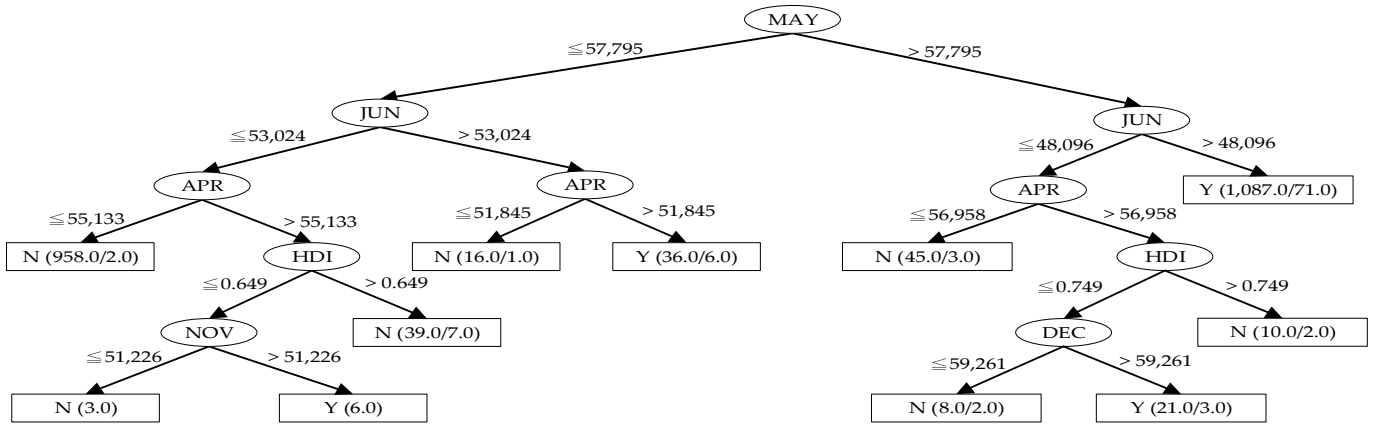

Figure 8. Model C2 J48 tree.

Rule 2: IF PSD MAY $>$ 57,795 and JUN $\leq$ 48,096 and APR $>$ 56,958 and HDI $<$ 0.749 and DEC $>$ 59,261 THEN NSS = Y.

Assuming that the daily performance is more than 57,795 in May, less than or equal to 48,096 in June, and more than 56,958 in April, and the human development index is less than 0.749, and the daily performance is more than 59,261 in December, it is predicted that this area (site) can be selected as a new store.

Rule 3: IF PSD MAY $\leq$ 57,795 and JUN $>$ 53,024 and APR $>$ 51,845 THEN NSS = Y.

Assuming that the daily performance is less than or equal to 57,795 in May, more than 53,024 in June, and more than 51,845 in April, it is predicted that this area (site) can be selected as a new store.

According to the analysis results of J48 decision tree, after attribute selection is included, the conditions of PSD and HDI are added to the important attributes. Based on the tree structure, the decision conditions of store performance in April, May, June, November, and December and HDI are the important factors to predict whether this area (site) can be selected as a new store.

In Experiment 3, it is found that the J48 decision tree classifier has a higher average accuracy, and the logistic regression classifier has the lowest accuracy of 93.3333%. In terms of the model, the accuracy of the model without attribute selection is lower, while the accuracy of the model with attribute selection is generally higher. J48 decision tree has the highest average accuracy.

(4) Experiment 4: NSS is a decision attribute, including the conditional attribute of daily average performance, by cross-validation (10-fold mixed cross-validation).

Model description:

Model D1: without attribute selection and without data discretization. Model D2: with attribute selection (Cashier, MAR, APR, MAY, JUN, SEP, and DEC). Model D3: with data discretization. Model D4: with attribute selection and with data discretization. The accuracy of Models D1~D4 is calculated by three classifiers: Model D1 (Bayes net = 93.3603%, Logistic regression = 93.0911%, J48 = 92.9116%); Model D2 (Bayes net = 94.1229%, Logistic regression = 93.6294%, J48 = 93.7192%); Model D3 (Bayes net = 92.8219%, Logistic regression = 92.9116%, J48 = 93.2257%); Model D4 (Bayes net = 93.2705%, Logistic regression = 93.5397%, J48 = 93.4948%). The statistical graph of Models D1-D4 new store selection decision is shown in Figure 9.

Results of Experiment 4:

Bayes net classifier: Model D2 has the highest accuracy of 94.1229%, and Model D3 has the lowest accuracy of 92.8219%.

Logistic regression classifier: Model D2 has the highest accuracy of 93.6294%, and Model D3 has the lowest accuracy of 92.9116%.

J48 decision tree classifier: Model D2 has the highest accuracy of 93.7192%, and Model D1 has the lowest accuracy of 92.9116%. J48 decision tree takes Model D1 as an example. The decision tree is shown in Figure 10.

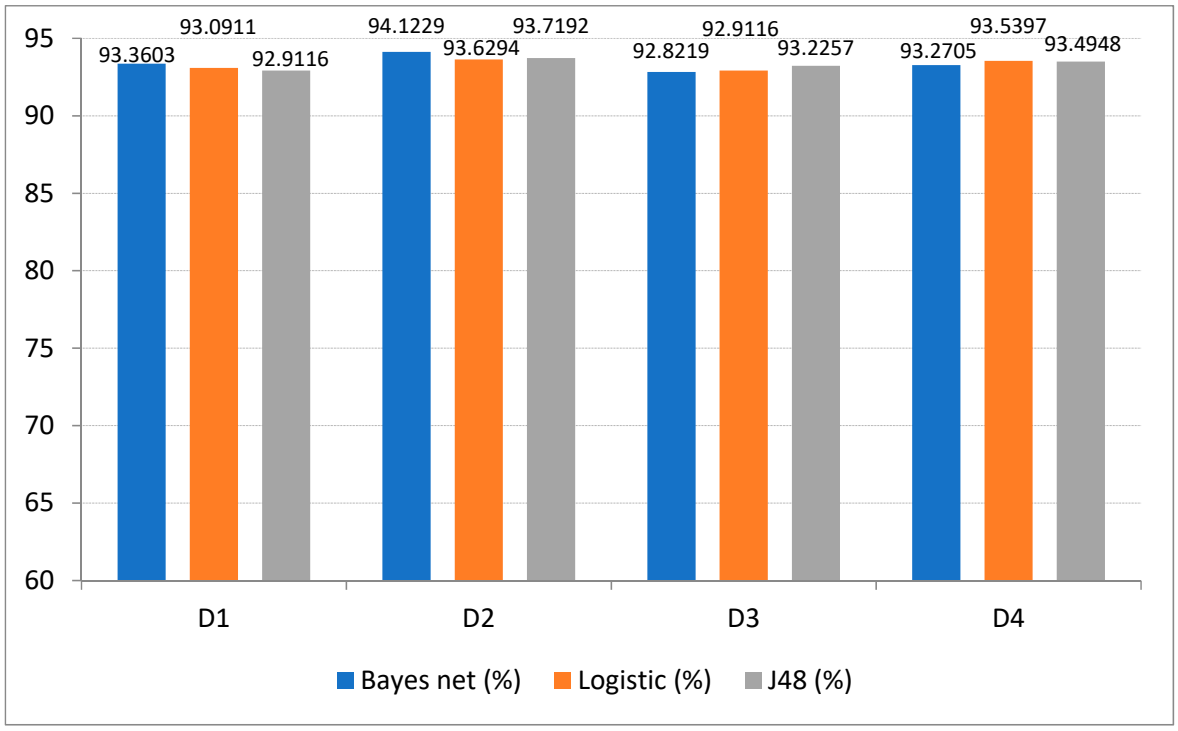

Figure 9. Models D1-D4 new store selection decision (statistical graph).

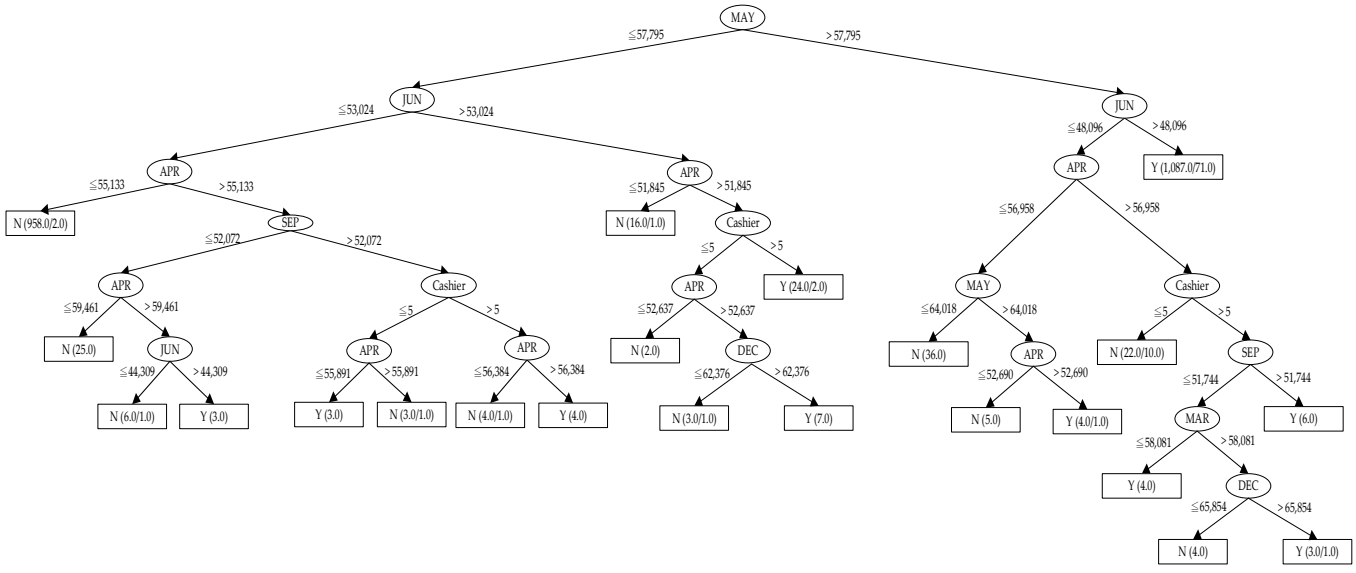

Figure 10. Model D2 J48 tree.

Rule description:

Rule 1: IF PSD MAY > 57,795 and JUN > 48,096 THEN NSS = Y.

Assuming that the daily performance is more than 57,795 in May and more than 48,096 in June, it is predicted that this area (site) can be selected as a new store.

Rule 2: IF PSD MAY > 57,795 and JUN ≤ 48,096 and APR > 56,958 and Cashier > 5 and SEP > 51,744 THEN NSS = Y.

Assuming that the daily performance is more than 57,795 in May, less than or equal to 48,096 in June, and more than 56,958 in April, and the number of cashiers is more than 5, and the daily performance is more than 51,744 in September, it is predicted that this area (site) can be selected as a new store.

Rule 3: IF PSD MAY $>$ 57,795 and JUN $\leq$ 48,096 and APR $>$ 56,958 and Cashier $>$ 5 and SEP $\leq$ 51,744 and MAR $>$ 58,081 and DEC $>$ 65,854 THEN NSS $=$ Y.

Assuming that the daily performance is more than 57,795 in May, less than or equal to 48,096 in June, and more than 56,958 in April, and the number of cashiers is more than 5, and the daily performance is less than or equal to 51,744 in September, more than 58,081 in March, and more than 65,854 in December, it is predicted that this area (site) can be selected as a new store.

According to the analysis results of J48 decision tree, the important attributes include the conditions such as PSD and HDI. Based on the tree structure, the decision conditions of store performance in April, May, June, September, November, and December, and HDI are the important factors to predict whether this area (site) can be selected as a new store.

In Experiment 4, it is found that Bayes net classifier has the highest accuracy, and the model with attribute selection generally has higher accuracy.

4.3. Comparison of Experiment 1 to Experiment 8

This study takes profit to represent whether the store can create a net profit or loss and NSS to represent whether this area can be selected as a new store as the decision attributes. These two decision attributes are suggested and provided by experts, and the analysis of these two data is combined with the average daily store performance from January to December 2018 as the conditional attributes, to analyze whether the condition has influence distinguished by the with/without performance condition. Based on the above data collation results, the optimal accuracy of two groups in each experiment is listed, respectively, with the statistical data of net profit as decision attribute and NSS as decision attribute, as shown in Tables 8 and 9. The bar graph with net profit as decision attribute and NSS as decision attribute is as shown in Figures 11 and 12.

Table 8. Statistical table of profit as decision attribute.

Model	Decision attribute	Performance Condition	Test Method	Bayes Net (%)	Logistic (%)	J48 (%)
A1	Profit	Y	Random Sampling	88.9946	86.1413	91.0326
A4	Profit	Y	Random Sampling	90.7609	90.3533	91.0326
B1	Profit	Y	Mixed	89.1528	87.0013	90.8113
B4	Profit	Y	Mixed	90.9458	90.5424	90.7665
E1	Profit	N	Random Sampling	80.0272	78.8043	79.2120
E3	Profit	N	Random Sampling	80.0272	79.2120	79.2120
F1	Profit	N	Mixed	78.7987	79.4263	78.7539
F3	Profit	N	Mixed	78.7987	79.0677	78.7539

Table 9. Statistical table with NSS as decision attribute.

Model	Decision Attribute	Performance Condition	Test Method	Bayes Net (%)	Logistic (%)	J48 (%)
C1	NSS	Y	Random Sampling	93.6054	93.0612	94.5578
C2	NSS	Y	Random Sampling	94.6939	93.3333	94.6939
D1	NSS	Y	Mixed	93.3603	93.0911	92.9116
D2	NSS	Y	Mixed	94.1229	93.6294	93.7192
G1	NSS	N	Random Sampling	82.4490	81.6327	81.2245
G2	NSS	N	Random Sampling	81.2245	81.2245	81.2245
H1	NSS	N	Mixed	80.5294	81.3818	80.7537
H2	NSS	N	Mixed	80.7088	80.3948	80.7537

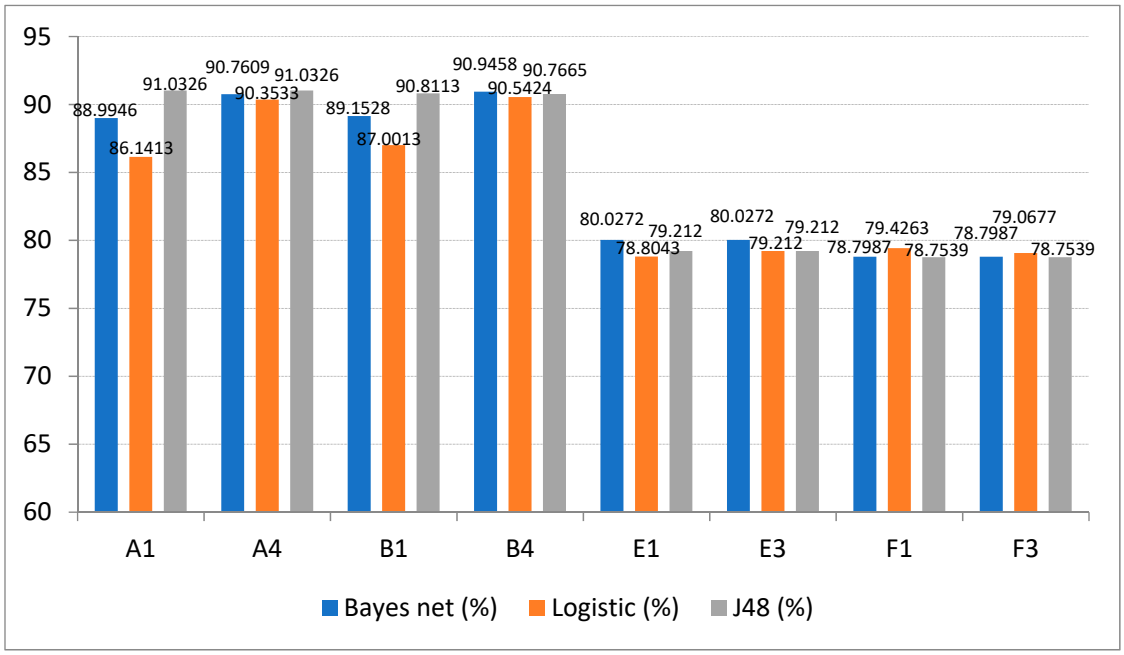

Figure 11. Bar graph with a net profit as decision attribute.

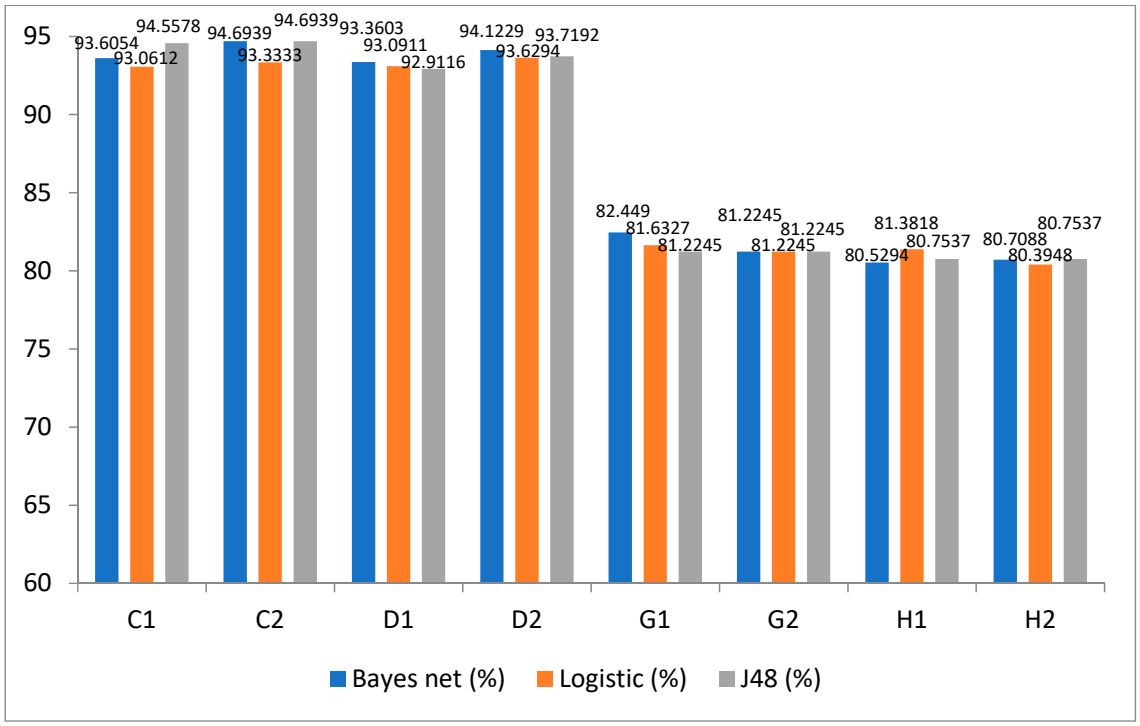

Figure 12. Bar graph with NSS as decision attribute.

(1) When to predict whether the store will create a net profit, the mixed test method has a high accuracy, while whether the performance condition (PSD) is added has a great impact, with an accuracy difference of about 10%. The accuracy of the classifier based on J48 decision tree is relatively high and stable. In addition, those with data preprocessing such as attribute selection and data discretization will also have higher accuracy than those without data preprocessing.

(2) When to predict whether this area (site) can be selected as a new store, the results are slightly different from whether the store will create a net profit. The accuracy

of the test method is higher with the random sampling method, and whether the performance condition (PSD) is added has a great impact, with an accuracy difference of about 10%. The accuracy of classifiers selected in this experiment does not differ greatly. In addition, the accuracy with data preprocessing of attribute selection is higher than that without data preprocessing. The only considerable difference is that the data mining software can perform the attribute selection in the experiment but not the data discretization, resulting in poor reference for the prediction, which needs to be further studied by interested people in the future.

5. Conclusions

In the process of drawing the empirical conclusions, the empirical results of this study are obtained. The conclusions are divided into five categories, namely the discussion, research findings, research contributions, research limitations, and future research, as described below.

5.1. Discussion

The decision-making process of retail stores for identifying and confirming the study results usually follows a systematic process with two insights of managerial and technical perspectives for further achieving valuable discussion. First, this process can be broadly discussed and described in a variety of determinants for effectively identifying applications of the new chain convenience store in the five-core importance for managerial perspectives, as follows:

(1) Market selection: Considering an area or site with potential for a new store location is an important and interesting issue. It is also valuable that identifying business districts with specific functions or features can be measured based on stores, goods, service content, and demographics. Afterwards, the general selection conditions for the market of a new chain convenience store are convenient transportation, obvious landmarks, convenient parking, easy access for consumers, and targeting specific consumer groups as the primary goal. However, exploring and discussing the variables influencing market selection is a considerable task because there are too many other factors to be addressed, particularly for unpredictable events. For example, the COVID-19 epidemic is accompanied with social distancing regulations, which restrict consumers' original shopping habits and challenge the supermarket connection with consumers. In order to minimize the contact between people, it is suggested that the location of supermarkets can be oriented towards community and delivery services. Supermarket exhibition stores are not limited to commercial areas, but they should be located in residential areas as much as possible. The purpose is to be close to consumers because the rental cost is relatively lower than that in commercial areas, and the number of stores can be more. The manpower and logistics of supermarkets may be affected by COVID-19. Automated robotics and Internet of Things (IoT) solutions can track goods and arrange shipments. Potential benefits of autonomous robots include increased efficiency, reduced errors, and improved safety. Facing the risk of uncertainty and the continuous change of online and offline consumption habits via Internet technology, it is expected that the operating mode of supermarkets can be changed more flexibly.

(2) Regional analysis: It is a key concern to select the potential best area for the new store within the selected area, e.g., investigate whether there is sufficient population, whether there is a target audience, etc. If you can know who lives in the surrounding area or how many people work nearby before you create a new store, it is possible and necessary to predict the turnover.

(3) On-site assessment: It is also a topic that can be discussed to take into account factors of on-site assessment, such as the number of competitors in the region, style, and popularity, etc. In this way, it is possible that the competition level of the region in

the future can be estimated to adjust the brand product mix, product pricing, and service model.

(4) Market positioning: The data mining techniques used for industry in this study are organized to predict whether a chain of identifying new convenience stores will be profitable, and whether an area (site) can be selected as a new store. Behind the data experiment processing, we can understand the potential demand and potential store capacity of the new store location. The scale of the new store can be huge, and the products can be various, but it is impossible to meet the needs of all consumers. Each store has a certain market scope and a specific consumer group. If the target market positioning is wrong, the store will not be able to have consumer groups and will be eliminated and excluded by the market. Thus, the market positioning is a vital part of marketing expansion. In particular, the new expansion of chain convenience stores is a long-term investment, and it is not as flexible as the marketing of manufacturing companies. Especially when the COVID-19 epidemic affects the global economy, the challenges of international convenience stores will be even more severe than that of past times.

(5) Comparative studies of this paper and next-future paper: First, although the study has used past classification algorithms to construct the proposed mixed models, we still have a significantly scientific advantage and have comparative studies of "three well-known classification algorithms" (i.e., Bayes net, logistic regression, and J48 decision tree) in this paper when compared to actuarial literature. That is, we use a combination procedure of highlighting and processing a named 3-4-8-2 components experience (i.e., three above well-known classifiers with past good performance and four models (without preprocessing, with attribute selection, with data discretization, and with attribute selection and data discretization) are used for eight different experiments, through two data verification methods (percentage split and cross-validation)) to address the issue of a new chain convenience store. Furthermore, these classification algorithms used had a well-known and popular method with superior performance in a variety of application fields; thus, they are selected as the research focuses and research objects of the study. In the entire research plan, we have two phased objectives and interests: near-term (present paper) and later-term (next research). First, the present paper aims to use the past classifiers and compare them in the short term. Next, we will upgrade the subsequent research to capture and model some sophisticated new algorithms or state-of-the-art techniques to develop a robust model for further identifying applications of the new chain convenience store over the long term; at the same time, the performance of the proposed present and future new models can be compared and differentiated in order to study new issues related from the study concerns and future empirical results addressed.

Second, this process can be further discussed in three directions for identifying this application issue for technical perspectives, which can be further explored in subsequent research, as follows:

(1) About computer resources used for the data analysis: It is a valuable issue to discuss that the computer resources used for the data analysis interests have some key software/hardware elements, such as OS, CPU, RAM, HDD, GPU, and software for data analysis, etc. The quality of them will have positive influences on the data analysis processing and performance; thus, to explore the better combinations of these components is a priority concern and a valuable issue to explore in future research.

(2) About the trade-off problem between data analysis speed and analysis accuracy: Four literature cases are addressed and studied in this problem: (a) According to the study of Fujiwara and Casanova on network simulation issues [42], they indicated and faced a trade-off problem that packet-level network simulators can enable higher accuracy for simulation, but they consume disappointingly long times; conversely, although some simulation networks are developed for a higher level to enable fast simulation, they lose accuracy performance. (b) In Guiard and Rioul [43], they focused

on discovering and strengthening the roots of the problem for the speed/accuracy trade-off for the sharp concern of human-computer interaction (HCI) work, and they proposed a method, which may have the help of HCI practitioners in obtaining from their experienced data more trustworthy and more comprehensive information on the superlative achievements of design options to evaluate the related resource-allocation strategy. (c) Lu et al. [44] proposed a framework of feature fusion deep learning, a residual neural network (ResNet) combined with attention modules, with limited computing resources to balance the accuracy and speed trade-off problem with the empirical results of the high performance of a promising video-based urban traffic crash detection system. As a result, they achieved a higher detection accuracy of 87.78% as well as an acceptable detection speed (FPS > 30 with GTX 1060). (d) Norman and Bobrow [45] analyzed the performance effect for limited processing resources. Their key principles have a limited process in its performance effect, either limits in the available processing resources, such as memory or processing effort, or limits in the available data quality. From the experimental results, they showed competition among processes affects a resource-limited process, but not a data-limited one. Consequently, based on the above four examples, it is clear that data analysis speed and analysis accuracy are exactly a trade-off dilemma issue and worth discussing and examining; thus, it is also a key goal to design some related techniques or methods of interest in future work.

(4) About the problem of too much low-value data and too little high-value data from Chandrashekar and Sahin [16]: Although the process of attribute selection and modeling tools or algorithms will actively select or discard attributes according to their practicability for data analysis, attribute selection will be dependent on different characteristics of the given data, and thus it cannot practically produce highly consistent standard values for addressing the low-value data and high-value data in this study.

5.2. Research Finding

Some research findings, which can be used and referenced for interested parties, are yielded from the main empirical results of the study, as follows:

(1) In this study, three classifiers, Bayes net, logistic regression, and decision tree are used to predict whether the store will create a net profit and whether this area (site) can be selected as a new store. A set of rules are developed as reference elements according to the empirical results. This is an important purpose of this study, and this provides effective knowledge-based references to academicians and practitioners.

(2) In this experiment, the store performance is considered for comparison, and it is also found that if PSD data are not included, the accuracy of prediction would decrease by about 10%. In addition, the store size, such as the number of cashiers and POS as well as population development index, would also affect the prediction of new store expansion or profit and loss. It is obvious that current store management can be a reference for the evaluation of future store expansion in this area.

(3) In this experiment, we found that climate and population are not valuable for the analysis of data mining software when selecting the attributes. Empirically, we also found that if the data is not preprocessed, the accuracy will be about 10% lower than if the data is preprocessed. Moreover, it is expected that this research can reduce the time and cost of new store selection and opening in an enterprise's operation, and then help the enterprise to obtain the maximum benefits and business opportunities.

5.3. Research Contribution

Conclusively, four key contributions are yielded and addressed from the study results, including academic contributions, enterprise contributions, application contributions, and management contributions, as follows:

(1) Academic contributions: In this study, four models, namely without preprocessing, with attribute selection, with data discretization, with attribute selection and data

discretization, are discussed, respectively. The classifiers used are Bayes net, logistic regression, and decision tree for data analysis. Within the scope of this study, the data analysis shows that the performance of the decision tree classifier is stable, and the average accuracy is high. Moreover, to model such a hybrid approach by machine learning techniques for assessment and selection applications of new chain convenience stores is rarely seen from past studies; thus, this study owns significant research interest.

(2) Enterprise contributions: Due to the different culture and climate factors, the new store expansion application is not necessarily universal. Country-A is in the role of a developing country in ASEAN member countries, so the successful experience of Country-A's chain convenience store B is appropriate as the research object, hoping to help enterprises desiring to develop new chain convenience stores in Country-A.

(3) Application contributions: Although modeling the mixed model from methodology views is not the key objective of this study, the core applications of the modeled methods are also key, intensified by the impressive results, to benefit the challenges of future application issues. This study provides a good bellwether in the field of data mining for new chain convenience store applications.

(4) Management contributions: This study offers trees-based knowledgeable rules as practical managerial directions for different purposes of interested parties and also contributes helpful management references to discover the related chain convenience store information from the study experiences.

5.4. Research Limitation

The scope of this study is limited to Country-A, where the store performance is prone to fluctuations due to factors, such as the gap between urban and rural wealth, the difference between the north and the south climate, the wide land, and the disparity of population density. Therefore, the scope of this study is not applicable to identify new store expansion applications in other countries.

5.5. Future Studies

Although this study has some excellent performance, it still has some space for improvement for subsequent research in the following four directions:

(1) As there are many factors that determine whether the store will create a net profit and this area (site) can be selected as a new store, such as store rent, personnel costs, water and electricity expenses, and taxes, which are not included in the scope of this study, the prediction reference may be inaccurate, and it is expected to be further explored in depth in the future.

(2) This proposed mixed method can be applied to the data analysis of different industries. For example, the commodities of fresh food stores include fresh food, daily necessities, etc., which will be affected by external environment changes of geographical location, population structure, regional attributes, weather, and seasons.

(3) The demand from different stores is different. Through data mining tools, it is possible to predict the demand of store merchandise and discuss the competition pattern among stores and the research on key factors of influencing demand.

(4) We will further build state-of-the-art classification techniques to construct a robust model for further re-identifying applications of new chain convenience store issues in our next research.

Author Contributions: Conceptualization, S.-W.W.; methodology, S.-W.W. and Y.-S.C.; software, Y.-H.H.; validation, J.C.-L.C. and C.-K.L.; formal analysis, J.C.-L.C.; investigation, S.-W.W. and Y.-S.C.; resources, S.-W.W.; data curation, S.-W.W.; writing—original draft preparation, S.-W.W.; writing—review and editing, Y.-S.C. and C.-K.L.; visualization, Y.-H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Science and Technology Council of Taiwan for grant numbers NSTC 110-2410-H-167-017 and 111-2221-E-167-036-MY2.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Department of Economic and Social Affairs. Available online: https://www.un.org/development/desa/zh/about/desa-divisions/population.html (accessed on 20 January 2022).
2. Statista. Available online: https://www.statista.com/ (accessed on 23 January 2022).
3. Davies, R.L.; Rogers, D. *Store Location and Store Assessment Research*; John Wiley & Sons Inc.: Hoboken, NJ, USA, 1984.
4. Jaravaza, D.C.; Chitando, P. The role of store location in influencing customers' store choice. *J. Emerg. Trends Econ. Manag. Sci.* **2013**, *4*, 302–307.
5. Reynolds, J. Retail location analysis: An annotated bibliography. *J. Target. Meas. Anal. Mark.* **2005**, *13*, 258–266. [CrossRef]
6. Levy, M.; Weitz, B.A.; Beitelspacher, L.S. *Retailing Management*, 8th ed.; McGraw Hill: New York, NY, USA; Irwin: Huntersville, NC, USA, 2012.
7. Wood, S.; Reynolds, J. Leveraging locational insights within retail store development? assessing the use of location planners' knowledge in retail marketing. *Geoforum* **2012**, *43*, 1076–1087. [CrossRef]
8. Church, R.L.; Murray, A.T. *Business Site Selection, Location Analysis, and GIS*; John Wiley & Sons: Hoboken, NJ, USA, 2009; pp. 209–233.
9. Wieland, T. Market area analysis for retail and service locations with MCI. *R. J.* **2017**, *9*, 298–323. [CrossRef]
10. Santos-Pereira, J.; Gruenwald, L.; Bernardino, J. Top data mining tools for the healthcare industry. *J. King Saud Uni.-Comput. Inform. Sci.* **2022**, *34*, 4968–4982. [CrossRef]
11. Gandomi, A.; Haider, M. Beyond the hype: Big data concepts, methods, and analytics. *Int. J. Inf. Manag.* **2015**, *35*, 137–144. [CrossRef]
12. Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. *Advances in Knowledge Discovery and Data Mining*; AAAI Press: Menlo Park, CA, USA, 1996; Volume 21.
13. Greener, J.G.; Kandathil, S.M.; Moffat, L.; Jones, D.T. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* **2022**, *23*, 40–55. [CrossRef]
14. Hamdi, A.; Shaban, K.; Erradi, A.; Mohamed, A.; Rumi, S.K.; Salim, F.D. Spatiotemporal data mining: A survey on challenges and open problems. *Artif. Intell. Rev.* **2022**, *55*, 1441–1488. [CrossRef]
15. Armanfard, N.; Reilly, J.P.; Komeili, M. Local feature selection for data classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 1217–1227. [CrossRef]
16. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [CrossRef]
17. Cui, P.; Athey, S. Stable learning establishes some common ground between causal inference and machine learning. *Nat. Mach. Intell.* **2022**, *4*, 110–115. [CrossRef]
18. Xu, Y.; Sun, Y.; Ma, Z.; Zhao, H.; Wang, Y.; Lu, N. Attribute selection based genetic network programming for intrusion detection system. *J. Adv. Comput. Intell. Intell. Inform.* **2022**, *26*, 671–683. [CrossRef]
19. Noering, F.K.D.; Jonas, K.; Klawonn, F. Improving discretization based pattern discovery for multivariate time series by additional preprocessing. *Intell. Data Anal.* **2021**, *25*, 1051–1072. [CrossRef]
20. Chen, Q.; Huang, M.; Wang, H.; Xu, G. A feature discretization method based on fuzzy rough sets for high-resolution remote sensing big data under linear spectral model. *IEEE Trans. Fuzzy Syst.* **2021**, *30*, 1328–1342. [CrossRef]
21. Jane, V.A. Survey on IoT data preprocessing. *TURCOMAT* **2021**, *12*, 238–244.
22. Safarkhani, F.; Moro, S. Improving the accuracy of predicting bank depositor's behavior using a decision tree. *Appl. Sci.* **2021**, *11*, 1–13. [CrossRef]
23. Awujoola, O.; Odion, P.O.; Irhebhude, M.E.; Aminu, H. Performance evaluation of machine learning predictive analytical model for determining the job applicants employment status. *Malays. J. Sci.* **2021**, *6*, 67–79. [CrossRef]
24. Cooke, R.M.; Joe, H.; Chang, B. Vine regression with Bayes nets: A critical comparison with traditional approaches based on a case study on the effects of breastfeeding on IQ. *Risk Anal.* **2022**, *42*, 1294–1305. [CrossRef]
25. Hidayat, N.; Afuan, L. Naïve Bayes for detecting student's learning style using Felder-Silverman index. *JUITA J. Inform.* **2021**, *9*, 181–190. [CrossRef]
26. Gramaje, A.; Thabtah, F.; Abdelhamid, N.; Ray, S.K. Patient discharge classification using machine learning techniques. *Ann. Data Sci.* **2021**, *8*, 755–767. [CrossRef]
27. Suman, S.K.; Hooda, N. Predicting risk of Cervical Cancer: A case study of machine learning. *Int. J. Stat. Manag. Syst.* **2019**, *22*, 689–696. [CrossRef]
28. Kannan, R.P. Prediction of consumer review analysis using Naive Bayes and Bayes Net algorithms. *Turk. J. Com. Math. Edu. (TURCOMAT)* **2021**, *12*, 1865–1874.
29. Manogaran, G.; Lopez, D. Health data analytics using scalable logistic regression with stochastic gradient descent. *Int. J. Adv. Intell. Paradig.* **2018**, *10*, 118–132. [CrossRef]

30. Demidenko, E. Sample size determination for logistic regression revisited. *Stat. Med.* **2007**, *26*, 3385–3397. [CrossRef] [PubMed]
31. Motrenko, A.; Strijov, V.; Weber, G.W. Sample size determination for logistic regression. *J. Comput. Appl. Math.* **2014**, *255*, 743–752. [CrossRef]
32. Stenersen, S.R.; Grønnbeck, K.O. Continuously adapting continuous Queries for Data Streams in Raincoat. Master's Thesis, Institutt for Datateknikk og Informasjonsvitenskap, Trondheim, Norway, 2013.
33. El Sibai, R.; Chabchoub, Y.; Demerjian, J.; Kazi-Aoul, Z.; Barbar, K. Sampling algorithms in data stream environments. In Proceedings of the 2016 International Conference on Digital Economy (ICDEc), Carthage, Tunisia, 28–30 April 2016; pp. 29–36.
34. Cardellini, V.; Lo Presti, F.; Nardelli, M.; Russo, G.R. Runtime adaptation of data stream processing systems: The state of the art. *ACM Comput. Surv.* **2022**, *54*, 1–36. [CrossRef]
35. Ataman, M.G.; Sarıyer, G. Predicting waiting and treatment times in emergency departments using ordinal logistic regression models. *Am. J. Emerg. Med.* **2021**, *46*, 45–50. [CrossRef]
36. Lee, C.S.; Cheang, P.Y.S.; Moslehpour, M. Predictive analytics in business analytics: Decision tree. *Adv. Decis. Sci.* **2022**, *26*, 1–29.
37. Kee, L.; Huynh, M.; Xanthos, P.; Davids, C.; James, L. The determinants of student attrition in an undergraduate sport and exercise science degree. *J. Sport. Sci. Edu.* **2022**, *7*, 7–16.
38. Huang, K.L.; Chen, M.H.; Hsu, J.W.; Tsai, S.J.; Bai, Y.M. Using classification and regression tree modeling to investigate appetite hormones and proinflammatory cytokines as biomarkers to differentiate bipolar I depression from major depressive disorder. *CNS Spectr.* **2022**, *27*, 450–456. [CrossRef]
39. Jeiad, H.A.; Ameen, Z.J.; Mahmood, A.A. Employee performance assessment using modified decision tree. *J. Eng. Technol.* **2018**, *36*, 806–811. [CrossRef]
40. Riandari, F.; Sihotang, H.T.; Gautama, R.; Ramen, S. Student graduation value analysis based on external factors with C4.5 Algorithm. *J. Mantik* **2022**, *6*, 2228–2235.
41. Al Karim, M.; Ara, M.Y.; Masnad, M.M.; Rasel, M.; Nandi, D. Student performance classification and prediction in fully online environment using Decision tree. *AIUB J. Sci. Eng.* **2021**, *20*, 70–76. [CrossRef]
42. Fujiwara, K.; Casanova, H. Speed and accuracy of network simulation in the Simgrid framework. In Proceedings of the 1st International ICST Workshop on Network Simulation Tools, Nantes, France, 22 October 2007. [CrossRef]
43. Guiard, Y.; Rioul, O. A mathematical description of the speed/accuracy trade-off of aimed movement. In Proceedings of the 2015 British HCI Conference, Lincoln, UK, 13–17 July 2015; pp. 91–100. [CrossRef]
44. Lu, Z.; Zhou, W.; Zhang, S.; Wang, C. A new video-based crash detection method: Balancing speed and accuracy using a feature fusion deep learning framework. *J. Adv. Transp.* **2020**, *2020*, 8848874. [CrossRef]
45. Norman, D.A.; Bobrow, D.G. On data-limited and resource-limited processes. *Cogn. Psychol.* **1975**, *7*, 44–64. [CrossRef]