DATA ANALYSIS OF CUSTOMER BEHAVIOR ON E-COMMERCE WEBSITES FOR PERSONALIZED MARKETING STRATEGIES.

BY

Sarthok Biswas Shetu
ID No: 222-17-530

This Report Presented in Partial Fulfillment of the Requirements for the Degree of MS in Management Information Systems (MIS)

Supervised By
Dr. Sheak Rashed Haider Noori,
Professor & Head,
Department of Computer Science and Engineering

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
February 2024

APPROVAL

This Thesis titled "**DATA ANALYSIS OF CUSTOMER BEHAVIOR ON E-COMMERCE WEBSITES FOR PERSONALIZED MARKETING STRATEGIES**", submitted by **Sarthok Biswas Shetu** to the Department of Computing and Information System, Under CSE, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of MS in Management Information System and approved as to its style and contents. The presentation was held on 03 February Saturday, 2024.

BOARD OF EXAMINERS

Professor Dr. Sheak Rashed Haider Noori — **Chairman**
Professor and Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Zahid Hasan — **Internal Examiner**
Associate Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Sazzadur Ahamed — **Internal Examiner**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Firoz Ahmed — **External Examiner**
Professor
Department of ICE
University of Rajshahi

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Dr. Sheak Rashed Haider Noori, Professor & Associate Head, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Professor Dr. Sheak Rashed Haider Noori
Professor & Head
Department of CSE
Daffodil International University

Submitted by:

Sarthok Biswas Shetu
ID: 222-17-530
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to Almighty God for His divine blessing which makes me possible to complete the final year project/internship successfully.

This accomplishment and completion of this report would not have been viable withoutthe contribution of some accommodative people, who gave their valuable time from their busy schedule to guide me in doing my project work. I sincerely remind their suggestion and guidance with full respect. I would like to extend my heartiest thanks and express my sinceregratitude to all those who directly or indirectly contributed to the completion of the report. Atthe beginning, I would like to convey my deepest gratitude and appreciation to my honorable supervisor **Dr. Sheak Rashed Haider Noori**, Professor & Head, Department of Computer Science and Engineering who has helped me with information, ideas and direction and the close guidance and tremendous support.

I would like to express my gratitude to other faculty members and the staffs of CSE department of Daffodil International University.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

Online shopping is a big deal nowadays, and many people love it. But there's this common issue – folks often put stuff in their online cart but not buy anything. It's a puzzle for online shops. This thesis dives into understanding why people do this and aims to predict their behavior using data mining. The main goal is to determine what customers might do when visiting an online store. We're using WEKA to analyze data we collected from an online shop. We split the data into two sets – one with all the details and another more focused based on what the computer thinks is important. We then tried seven different methods in WEKA to see which is best at predicting what customers might do. The idea is to find ways to encourage people who leave the website without buying anything to return and maybe make a purchase. The results of this study give helpful tips that can work for various online shops. We suggest rules and decision trees to predict and encourage customer behaviors, especially for those who leave the website early. The chosen method, which is good at predicting stuff, gives a solid foundation for making intelligent decisions in marketing and improving the user experience. This research aims to help businesses use Watson's data as an example to understand better and respond to what customers do when shopping online.

TABLE OF CONTENTS

LIST OF FIGURES

LIST TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

The advent of modern technologies has significantly transformed people's attitudes towards shopping. Online shopping is increasingly preferred over traditional shopping [1]. The accessibility of the Internet has led to a significant increase in the use of online tools and technologies by corporations in the 21st century for various objectives. Presently, there is a substantial amount of data stored in a data repository, and many corporations have embraced e-commerce [2]. Furthermore, the task of identifying clients has gotten more challenging due to the growing number of customers, products, and rivals, as well as the decreased response time compared to previous circumstances. Furthermore, a multitude of diverse factors contribute to the complexity of customer relationships [3]. Hence, firms acknowledge that it is imperative to effectively meet and exceed their consumers' needs and expectations in a timely manner. Furthermore, the response process for it is being shortened. It is no longer feasible to take action after indicators of client dissatisfaction are noticed [3].

Within this particular setting, research in the field of e-commerce must ascertain the factors that motivate individuals to make purchases, while also comprehending consumer behaviours. Identifying client behaviour and the elements that drive purchases is a challenging and intricate task. Analyses have increasingly centred around data mining techniques [1].

Data mining is the process of extracting valuable and essential information from extensive datasets using certain algorithms. Data mining, a widely used technology, is used across numerous fields to efficiently evaluate and analyse enormous datasets [4].
Finding patterns to better understand customer preferences and behavior is the main goal of data mining when it comes to online buying. Hence, several data mining approaches, including classification, clustering, association rules, and sequential patterns, have been effectively employed [1].

The primary objective of this project is to utilise a data categorization method to forecast client behaviour throughout their visits to the Watsons e-commerce channel. Owing to prevailing

patterns, a significant number of clients exhibit a preference for online purchasing. Within the group of customers that use online channels, a portion of them make purchases of the products found in their market basket. Conversely, a portion of users exit the website without making a transaction. Predicting how first-time visitors to the Watsons website will behave is the primary goal of the research. For data categorization purposes, the business makes use of demographic information and historical data on customer website visits. By comparing tried-and-true techniques, we may find the best data classification strategy for maximum predictive power. At last, the results are sent to the company, and a decision-support system is suggested to foretell the customer's behavior on their first website visit.

The study commences with a comprehensive examination of existing literature on data mining in Section 2. Section 3 provides an overview of the company's information and defines the problem based on the current application. Section 4 entails the intricate procedures of the technique. Computational findings are derived from the company's data and presented in Section 5. Section 6 provides the final findings and outcomes of this investigation.

1.2 Background Information on the Topic

In recent years, the landscape of retail has undergone a profound transformation, significantly influenced by the advent of modern technologies. One notable shift is the increasing preference for online shopping over traditional brick-and-mortar retail experiences. This shift is attributed to the widespread accessibility of the Internet, prompting corporations to leverage online tools and technologies for various objectives. E-commerce has become a pivotal component of business strategies, with a plethora of data stored in repositories, capturing the intricacies of customer interactions and transactions.

Watsons, a prominent player in the global beauty and personal care sector, has not only established a vast network of physical retail outlets but has also embraced the e-commerce platform to cater to the evolving preferences of consumers. With over 6300 stores across 11 diverse regions, Watsons recognizes the importance of understanding and meeting the expanded needs and expectations of its customers, both in physical stores and online.

However, the challenge lies in deciphering customer behavior in the dynamic e-commerce environment, where the growing number of customers, products, and competitors, coupled with reduced response times, adds layers of complexity to the task. Watsons, like many other e-commerce entities, faces the challenge of retaining customers who visit the website but leave without making a purchase. Understanding the factors that motivate customers to make a purchase and predicting their behavior during the initial visit to the e-commerce channel is crucial for enhancing customer satisfaction and optimizing business outcomes.

In response to these challenges, this thesis embarks on a comprehensive exploration of data mining techniques applied to e-commerce, specifically focusing on Watsons' e-commerce channel. The objective is to employ data classification methods to forecast and understand customer behavior during their initial interaction with the Watsons website. By leveraging historical data on customer website visits and demographic information, the study aims to identify the optimal data classification approach with the highest predictive accuracy. The findings of this research not only contribute to the field of e-commerce analytics but also provide actionable insights for Watsons to enhance its decision-making processes and tailor its online platform to better serve customer needs.

1.3 Objectives of the Thesis

Customer Behavior Prediction

- The objective is to utilize data mining techniques to forecast client behavior when they first access the Watsons e-commerce platform.
- To identify patterns and factors that influence customers in making a purchase or leaving the website without completing a transaction.

Data Classification Methodology

- To evaluate and compare the performance of various data classification algorithms, including Functions, Trees, Bayes, Lazy, Rules, Meta, Misc, in predicting customer behavior.

In order to find the best data classification technique that maximizes accuracy for dataset 1 and dataset 2 (which contains selected characteristics).

Attribute Selection and Data Set Creation

- To perform attribute selection procedures and create two distinct datasets (dataset 1 and dataset 2) to assess the impact of attribute inclusion on the predictive accuracy of the classification algorithms.
- To analyze and interpret the outcomes of attribute selection methods, considering the relevance of attributes in predicting customer behavior.

Rule and Tree Generation

- To generate rules and decision trees based on the selected data classification algorithms, particularly focusing on the most accurate approaches identified for dataset 1 and dataset 2.
- To provide actionable recommendations and insights derived from the rules and trees, aiding Watsons in understanding and responding to customer behavior on their e-commerce platform.

Application of Findings to Business Strategy

- To deliver practical recommendations to Watsons, drawing upon the results and discoveries from the data mining analysis, with a focus on improving the e-commerce customer experience.
- To propose strategies for customer retention and engagement, utilizing the insights gained from the data classification models, ultimately contributing to the enhancement of Watsons' online business strategies.

Enhancing Decision Support System

- To explore possibilities for improving the accuracy of class prediction in data mining classification by considering additional attributes or refining existing ones for future use.
- To suggest the development of a decision support system that can accurately predict and respond to customer behavior, providing real-time insights for effective decision-making within the e-commerce domain.

1.4 Significance of the Research

In the dynamic landscape of e-commerce, understanding and predicting customer behavior is pivotal for sustainable growth and customer satisfaction. This research holds significant importance for various stakeholders, offering valuable insights and potential improvements in several key areas:

- Empowers Watsons with actionable data to make informed decisions, aligning business strategies with customer preferences in the competitive e-commerce market.
- Aims to elevate the online shopping experience by tailoring services based on anticipated customer behaviors, ultimately increasing customer satisfaction and loyalty.
- Provides a roadmap for efficient resource allocation, enabling Watsons to streamline operations, optimize marketing efforts, and reduce costs.
- Positions Watsons as an industry leader by staying ahead in the realm of customer-centric e-commerce, outperforming competitors through data-driven strategies.
- Adds to the body of knowledge in e-commerce analytics, serving as a reference for academia and contributing to the discourse on data mining in the context of online retail.

1.5 Limitations of the Study

While this research strives to contribute valuable insights into customer behavior prediction in the context of Watsons' e-commerce platform, it is important to recognize the limitations that may affect the generalizability and extent of the findings.:

- Accuracy relies on available data; missing or incomplete information may affect model precision.
- Model effectiveness depends on selected algorithms; other choices may yield different results.
- Study may not capture the full complexity of human decision-making processes.
- Unaccounted external influences (market changes, economic shifts) could impact customer behavior.
- Success of suggested rules depends on organizational dynamics and market conditions.
- Adherence to ethical standards regarding customer privacy is acknowledged; ongoing considerations apply.

CHAPTER 2
REVIEW

2.1 Literature Review

The proliferation of big data and its associated applications has resulted in an exponential growth in the volume of information accessible to contemporary societies. To ascertain default risks with confidence, the banking and insurance sectors employ big data analytics and data mining. The master's thesis utilized a dataset consisting of 22,745 samples and 14 characteristics [5]. The dataset was acquired from the Turkish Statistical Institution. The primary objective of this project is to identify the most effective classification method that can precisely assess the default risk of individuals according to their attributes. An evaluation of several classification algorithms was conducted, encompassing metrics such as precision statistics, recall, Roc curve, Naive Bayes, J48, Logistic Regression, Random Forest, and Multi-Layer Perceptron. We have identified the optimal classification method. The data were analyzed using WEKA, an application for data mining [5].

Diabetes Mellitus is a prominent global medical issue. The prevalence of diabetes is rapidly increasing, causing significant deterioration in human, economic, and social well-being. For various clinical investigations, it is recommended to utilise data mining approaches to enhance the treatment of diabetes. Typically, all of these conventional systems rely on a solitary classifier or simple combinations. Lately, there have been extensive endeavours to enhance the precision of these systems by employing a collection of classifiers. In their study, researchers in reference [6] utilised risk factors associated with diabetes mellitus to categorise individuals based on their disease condition. The Adaboost and Bagging techniques were employed, and the J48 (c4.5) decision tree was utilized as the foundation for their investigation. Three distinct ordinal adult teams function within the broader Canadian Primary Care Sentinel Surveillance system to categorize these cases. The results indicate that, in terms of overall efficacy, the Adaboost ensemble method surpasses both the Bagging method and the individual J48 decision tree [6].

Anaklı [7] provides a concise overview of data mining applications. Currently, KDD databases and datamining are widely employed across several domains to collect vast amounts of

information. Data mining is the process of extracting significant and implicit patterns or information from vast quantities of data [8]. Classification and prediction-based data mining techniques are typically the most effective [8, 9]. The domains of control and quality improvement (QI) may each derive distinct advantages from the discoveries made through data mining. The investigation conducted by Estudillo et al. [10] explored the potential of data mining techniques in enhancing the quality of literature. Huang [11] suggested a decision tree approach for identifying the foremost determinants that influence the proportion of defective products. In order to illuminate particular quality issues, Kang [12] proposes the implementation of integrated machine learning methodologies. Fan et al. (2013) effectively integrated process knowledge, quality control, and data analysis in their research. Categorical data values for the future have been predicted through the utilization of classification algorithms. Statistical algorithms, decision tree algorithms, artificial neural network-based systems, and other methodologies comprise the majority of the categorization techniques utilized in the literature [7].

Due to the fact that data mining attempts to resolve a wide variety of problems, it can be categorized accordingly. Classifying objects based on preexisting categories, clustering objects by similarities, extracting association rules from transactions, locating data outliers, and predicting a continuous dependent variable are all challenges in this context. This segment offers concise overviews of numerous issue categories. Data miners utilize an array of foundational strategies to construct models for data mining. Descriptive learning, also known as unsupervised learning, and predictive learning, which is occasionally referred to as supervised learning, are the two principal categories of data mining techniques. Supervised learning is a well-established method for constructing models; it operates by comparing past input and output data to predict forthcoming values of the output variable in response to new input data. An output variable may be described by the term "dependent variable" if it describes the manner in which one or more "independent" factors influence that variable. In the realm of supervised learning, classification and regression represent the two primary categories. Throughout the learning process, monitoring must be implemented to achieve the intended outcomes. Unsupervised learning involves the estimation or deduction of the values of the input and output variables. Absence of a regulatory framework. The model is autonomously developed and assessed by the user. It is utilized to forecast forthcoming events or assess the current state of affairs. As an example of the numerous uses of unsupervised learning, consider association and cluster analyses.

Online purchasing has become an everyday occurrence for the majority of individuals. Collecting data regarding user preferences and behaviors is critical for tailoring e-commerce websites to the specific requirements of individual customers. It is possible to discern the preferences and activities of website visitors through the examination of server logs. A significant portion of the data analysis has been conducted utilizing data mining techniques, which classify individuals' behavior based on relatively consistent attributes and frequently disregard the sequential nature of their actions. Complicated behavioral patterns can therefore potentially be identified more effectively through the implementation of step-by-step monitoring during a session. In light of this concern, specifically in the evaluation of structured online commerce, this article presents a control approach based on a linear-temporal logic model. Online-published publications. Informed by the architecture of e-commerce, a standardized method of correlating log data simplifies the process of converting web logs into event records, which document user activity. Subsequently, specific behavioral patterns can be identified through the analysis of a multitude of pre-existing queries and the multitude of actions performed by a user throughout a session. Utilizing a Spanish e-commerce website as a case study from the actual world, the effectiveness of the proposed method was assessed. The research has provided us with insightful suggestions for enhancing the website's performance, allowing us to modify its layout [1].

A subfield of computer science, data mining encompasses both theoretical and practical investigations. The purpose of this study is to investigate, suggest, and enhance the application of contemporary methodologies, structures, and web mining tools to the acquisition of consumer data. By employing these methods, data mining techniques are able to extract valuable insights from online information, specifically user behavior data. In order to improve consumer service, the domains of online business and e-commerce have implemented data mining techniques. Web mining employs clustering algorithms to examine similar click-streams and overall site access patterns with the purpose of deducing e-commerce user behaviors. Present-day approaches to web session clustering primarily conceptualize sessions as collections of pages accessed within a specified time span. However, these algorithms neglect to consider the sequence of click-stream visits, a crucial element in detecting patterns within online sessions [3].

Particularly in the domain of e-commerce, the implementation of online data mining techniques is extremely consequential and has numerous potential uses. E-commerce, which refers to the use of the World Wide Web for commercial transactions, information exchange, and relationship building. Given the exponential growth of online purchasing and the substantial volumes of data generated by operational transactions over the past few years, the utilization of data mining techniques has assumed greater significance in the pursuit of uncovering and comprehending concealed consumer trends. A concise discussion of the applications of data mining in internet commerce can be found in reference [8]. Consumers can potentially share characteristics when categorized or clustered based on the degree of similarity in their browsing histories. By adopting this approach, we can enhance our comprehension of consumer behavior and deliver service that is more tailored and unique to each individual. The article examines the fundamental issues associated with e-commerce and explores the potential solutions that data mining methodologies could offer. We then proceed to the approach of association rules and discuss several application-relevant methods [8].

Already, a wide variety of structured and unstructured data categories are categorized as "big data." Internal operations, suppliers, markets, and the overall business environment are examples of these data sources. Specifically, opportunities arise for businesses that participate in electronic commerce (e-commerce). Data mining methods frequently employed in e-commerce consist of three fundamental algorithms—association, classification, and prediction—as stated in reference [2]. Utilizing three distinct data mining techniques for market segmentation offers an abundance of benefits, such as facilitating sales forecasting, analyzing the contents of purchasing carts, and assisting online retailers in managing client relationships. Examining the utilization of data mining in the context of online retail, specifically by analyzing structured and unstructured data gathered from diverse origins and stored in the cloud, is the principal objective of this study. The purpose of this inquiry is to illustrate the importance of data mining. This research investigates the various challenges associated with data mining, such as the identification of spiders, the transformation of data, the development of comprehensible models for business users, the adaptation to the data's dynamic dimensions, and the facilitation of business users' access to data transformation and model construction. Online retailers that have access to vast quantities of data are experiencing a significant surge in competitiveness, as indicated by the findings.

CHAPTER 3
PROBLEM DEFINITION

3.1 Company Information

Watsons, established in 1841, has maintained a leading position in the global beauty and personal care industry with more than 6300 retail locations in eleven countries, including Singapore, Malaysia, Taiwan, Macau, Thailand, China, the Philippines, Indonesia, Turkey, and Ukraine. With thirteen retail brands and over 14,000 locations in twenty-four countries, Watson Group is the largest health and cosmetics retailer in both Europe and Asia. The Watson Group, an enterprise specializing in a diverse range of retail formats and brands, possesses an extensive global presence. In addition to the 280 physical Watson locations in Turkey, customers may also place orders online.

3.2 Problem Definition

Online shopping has become increasingly favoured over traditional buying in recent years due to prevailing trends. E-commerce caters to users by providing limitless access to product catalogues, facilitating pricing comparisons, and delivering superior service quality. Consequently, numerous organisations now utilise the e-commerce platform as a means to efficiently engage in buying and selling transactions. Furthermore, organisations recognise that it is absolutely vital to acknowledge and promptly fulfil their customers' expanded demands and expectations. However, discerning clients' purchasing behaviour has become more challenging due to the growing quantity of customers, products, and competitors.

Watsons expresses concern regarding clients who abandon the website without making a purchase, as this is a significant challenge for the organisation. It is crucial for Watsons to identify the underlying factors that motivate customers to make a purchase, in order to have a comprehensive understanding of consumer behaviour. The primary objective of this project is to utilise a data classification method to forecast client behaviour when they are browsing the Watsons e-commerce channel.

CHAPTER 4
METHODOLOGY

4.1 Methodology of the study

WEKA 3.6 is a code-mining and machine learning application developed in New Zealand. It is occasionally referred to as the Waikato Knowledge Analysis Environment. Java, an additional programming language, is employed to enhance it. WEKA is open-source software (GNU General Public License-distributed). WEKA offers a wide range of algorithms for implementation, encompassing pre-processing, classification, and visualization functionalities. In particular, it provides comprehensive coverage of classification methodologies.

The data set must be submitted to WEKA for processing subsequent to its preparation. The file extensions supported by WEKA are XRFF, Attribute-Relation File Format (ARFF), CSV, C4.5, binary, and libsvm. URLs can also be used to access SQL databases and retrieve files from them. WEKA is capable of extracting data from a straightforward file format and is capable of processing both nominal and numeric attributes. Concurrently, data may be extracted from the database with the expectation that it will be converted to a file. WEKA provides an extensive array of classification methods, including Bayes, Functions, Lazy, Meta, Misc, Rules.

4.2 Proposed Solution Methodology

The flowchart of the methodology is illustrated in Figure 1. A detailed description of each stage is provided below:

- Organizational data retrieval: Watsons provides historical data extraction services encompassing a wide range of customer-specific details, including but not limited to gender, age, geographic location, average duration of visits, browser inclinations, purchase records, and browsing patterns. As a consequence of the data's enormous extent, a fraction of the data is eliminated.
- Excel lacks value management: Once the sample data has been selected, it is imported into the Excel spreadsheet. The list does not contain any duplicate or absent values.

- Attribute selection: All attributes are chosen in respect to the consumers' propensity to purchase a product, based on the information provided.
- In order to determine the dataset, two datasets are produced. The initial dataset comprises every attribute, while the subsequent dataset is derived from the results of attribute selection processes.

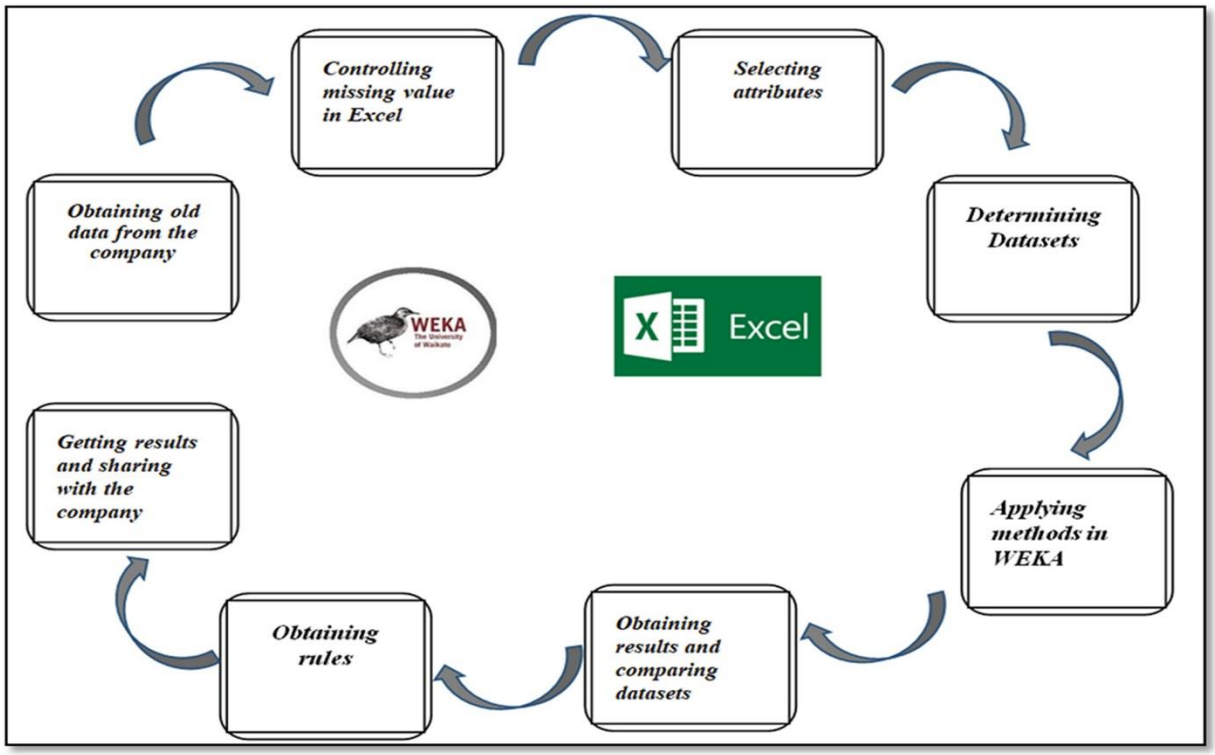

Fig. 1. Flowchart of the methodology

- All of the provided data are processed utilizing the selected methodologies via WEKA.
- Following their implementation in WEKA, the most accurate techniques from each category are selected. This enables the generation of results and the comparison of diverse datasets. By comparing Data Set 1 and Data Set 2, this article presents a thorough analysis of the overall results.
- In WEKA, to acquire rules, data sets 1 and 2 are independently processed.
- Execution and dissemination of information: Valuable recommendations are propelled for the organization in light of the results and understandings.

CHAPTER 5
IMPLEMENTATION AND RESULTS

The solution methodology described in Section 4.2 is implemented using the WEKA program.

5.1 Collection of Data

In the initial phase, the analysis period will be designated as March 19, 2018, to March 25, 2018. Utilize Watsons to obtain data regarding the demographics and purchasing patterns of online consumers at that particular juncture.

5.2 Controlling Missing Values

It is subsequently imported into Excel following the removal of some data. The list is subsequently purged of any duplicate or invalid information. By adhering to the prescribed protocols, 1317 data points were selected for examination within the WEKA software.

5.3 Data Set Description

The analysis of the data yielded the following seven variables: date, age, gender, city, browser type, average time spent on the site, and purchase status. Courses are further classified into two overarching categories: purchase and non-purchase.

The data set's description is presented in Table 1. The name, category, and quality of each attribute are explicitly specified. The prevalence of each attribute value is denoted in brackets around the category attributes enumeration, which contains every conceivable attribute value. To numerical properties are affixed values representing the maximum, minimum, mean, and standard deviation.

Table 1: Description data set

No	Attribute Name	Characteristics	
		Types	Properties
01	Sex	Categorical	Female (1051) Male (266)

02	Age	Categorical	18–24 (374) 25–34 (572) 35–44 (211) 45–54 (70) 55–64 (53) 65+ (37)
03	Day	Categorical	Monday (277) Tuesday (240) Wednesday (160) Thursday (160) Friday (160) Saturday (160) Sunday (160)
04	Average visit duration	Categorical	Min: 0 Max: 9440 Mean: 838.741 Std. Dev.: 1135.717
05	Browser type	Categorical	Chrome (334) Firefox (36) Safari (576) Internet explorer (92) Other (279)
06	City	Categorical	İstanbul (367) Ankara (159) İzmir (105) Adana (50) . . . Antalya (43)
07	Purchase or not	Categorical	Purchase (658) Not purchase (659)

5.4 Preprocessing

As part of the investigation's preparatory phase, 1317 data are transmitted to WEKA and eight selection procedures are implemented. The "selected attributes" of each technique are acquired through a thorough examination of their individual processes for every input data point. The results are generated through the implementation of the WEKA program. Following this, Table 2 presents the suggested datasets. Data set 2 is generated by exclusively considering the qualities that are present in each sequence, as opposed to data set 1, which encompasses all attributes. In each technique, the initial three parameters are selected. The final attributes of Data Set 2 are ascertained to be 5, 2, 3, 6, 1, and 7. Independent analyses of dataset 1 and dataset 2 are conducted in this study.

Table 2: Suggested data sets

Data sets	Attributes
Data sets 1	1, 2, 3, 4, 5, 6, 7
Data sets 2	5, 2, 3, 6, 1, 7
1: Sex 2: City 3: Browser type 4: Age 5: Average visit duration (second) 6: Day 7: Purchase or not	

5.5 Results

In this work, dataset 2 is analyzed independently from dataset 1 using the WEKA program. In order to determine the optimal classification method for datasets 1 and 2, we have developed and described seven distinct approaches in Section 4.1: Bayes, Functions, Lazy, Meta, Misc, Rules, and Trees. Confusion matrices and accuracy rates are also provided.

5.6 Results of Data Set 1

A comprehensive set of forty-one classification algorithms was evaluated for dataset 1.
In comparison to the alternatives, the Bayes method exhibits superior performance; BayesNet attained an accuracy rate of 86.8641%. When employing the Simple Logistic algorithm, the Lazy method achieves the highest accuracy of 88.2308%. In contrast, when utilizing the KStar algorithm, the Functions approach achieves an accuracy of 85.6492%. The Meta category achieved a noteworthy accuracy of 86.7122% through the utilization of the Attribute Selected Classifier and Multi Class Classifier Updateable. In comparison to the Misc approach, the Input Mapped Classifier demonstrates a remarkable accuracy rate of 49.8861%. Conversely, with regard to the Rules method, the JRip method attains an optimal accuracy rate of 86.8641%. At its zenith, the J48 algorithm attained an accuracy of 87.6693% by employing the Trees method. According to Tables 3 and 4, KStar demonstrates the highest accuracy rate of 88.2308% in dataset 1.

Table 3: Lazy Result for data set 1

Type	Method	Accuracy
Lazy	IBK	77.3728%
	KStar	88.2308%
	LWL	86.5604%

Table 4: Confusion matrix of KStar for data set 1

		Predicted classes	
		Purchase	Not Purchase
Actual classes	Purchase	586	72
	Not Purchase	83	576

0.8823 is its accuracy (586 + 576/1317).
In order to acquire a rate of error of 0.1177, deduct 0.8823 from 1.

5.7 Rules of Data Set 1

Given below are the JRip rules obtained, which have a significantly high level of correctness.

JRip Rules

Certainly! Here are the rules with more detailed explanations:

Rule 1: Average Visit Duration > 470 seconds

- Condition: If the customer's average visit duration is greater than 470 seconds.
- Classification: Classify as "purchase" with a ratio of 626.0/74.0.

Rule 2: Average Visit Duration > 220.67 seconds, Sex = Male, Browser Type = Safari

- Conditions: If the customer is male, using Safari, and has an average visit duration exceeding 220.67 seconds.
- Classification: Classify as "purchase" with a ratio of 24.0/2.0.

Rule 3: Average Visit Duration > 120 seconds, Sex = Male, Day = Monday

- Conditions: If the customer is male, shopping on a Monday, and has an average visit duration greater than 120 seconds.
- Classification: Classify as "purchase" with a ratio of 11.0/0.0.

Rule 4: Average Visit Duration > 102 seconds, Sex = Male, Day = Tuesday

- Conditions: If the customer is male, shopping on a Tuesday, and has an average visit duration surpassing 102 seconds.
- Classification: Classify as "purchase" with a ratio of 7.0/0.0.

Rule 5: Average Visit Duration > 225.4 seconds, Age = 55-64

- Conditions: If the customer's age is between 55 to 64 and the average visit duration is more than 225.4 seconds.
- Classification: Classify as "not purchase" with a ratio of 5.0/0.0.

Rule 6: Default Rule

- Classification: If none of the above conditions apply, classify as "not purchase" with a ratio of 644.0/61.0.

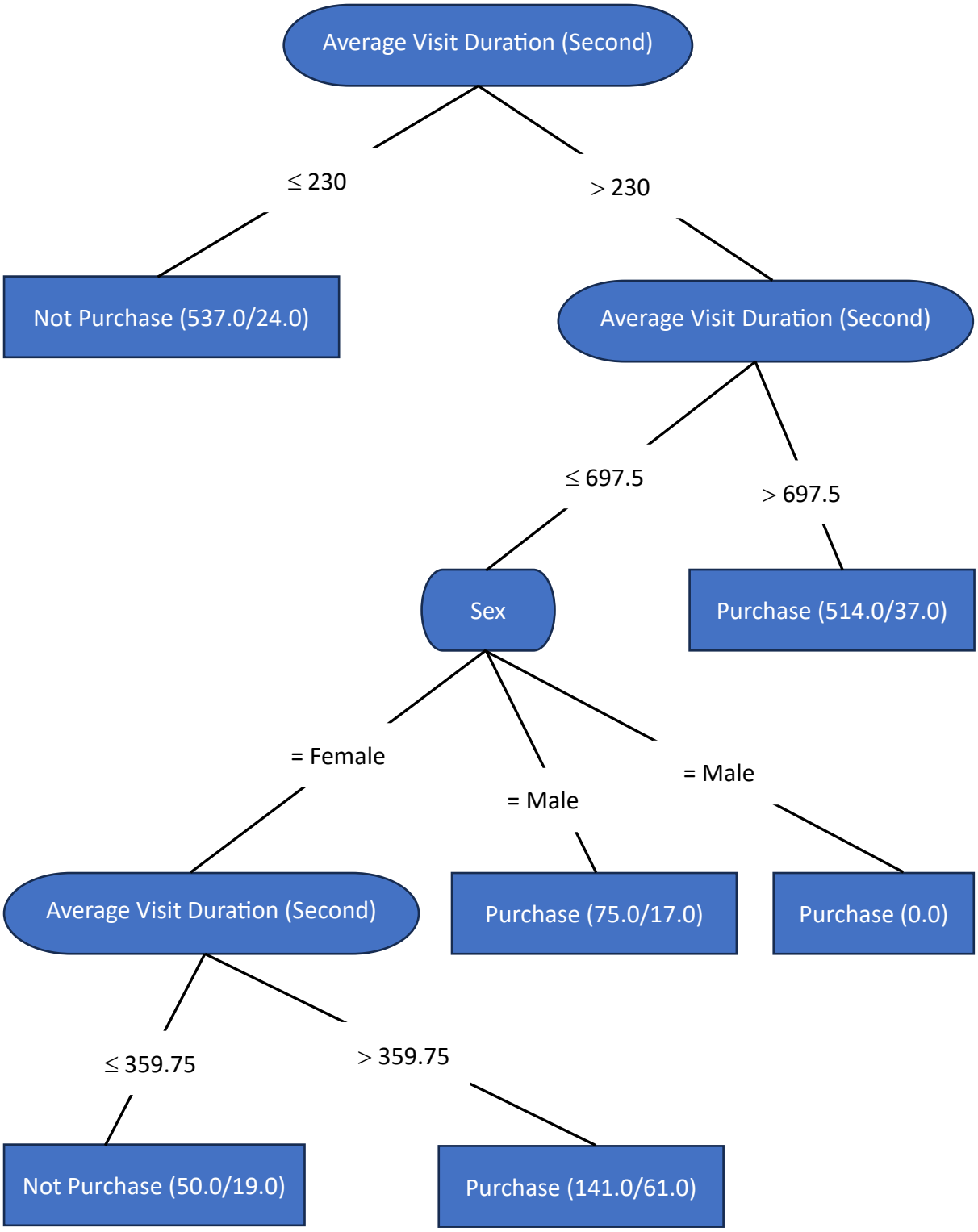

Fig. 2: Data set one's J48 Tree

Figure 2 illustrates the decision tree of the J48 algorithm. A customer is categorized as "not purchase" if their mean duration of visitation does not exceed 230 seconds. A "purchase" is defined as the average duration of a customer's visit surpassing 697.5 seconds.

5.8 Results of Data Set 2

A total of 41 categorization methods have been examined for data set 2.

The most effective approach for utilizing the Bayes technique is through the implementation of a Bayes Net, which achieves an accuracy rate of 86.8641%. The Functions approach achieved the maximum accuracy of 85.1177% using the Logistic algorithm, while the Lazy method achieved the highest accuracy of 87.4715% using the KStar algorithm. The classification method that achieves the highest accuracy for Meta is Classification Via Regression, with a value of 87.0919%. The accuracy of the Input Mapped Classifier for the Misc technique is 49.8861%, while the JRip method has the highest accuracy of 86.6363% for the Rules method. The Trees approach attains a peak accuracy of 87.6234% while employing J48. The J48 algorithm attained a peak accuracy rate of 87.6234% in dataset 2, as evidenced by the statistics presented in Table 5 and 6.

Table 5: Tree Result for data set 2

Type	Method	Accuracy
Trees	Decision stump	86.7122%
	Hoeffding tree	87.0159%
	J48	87.6234%
	LMT	85.4214%
	Random forest	84.9658%
	Random tree	79.2711%
	REP tree	85.4973%

Table 6: Confusion matrix for data set 2

		Predicted classes	
		Purchase	Not Purchase
Actual classes	Purchase	625	33
	Not Purchase	130	529

0.8762 is the precision (10.25 + 5.29/1317).

To determine the error rate of 0.1238 percent, deduct 0.8762 from 0.13.

In Figure 3, the decision tree generated by the J48 algorithm is illustrated. Males comprise the "Purchase" demographic, and their average session duration ranges from 230 to 697.5 seconds. Additional criteria for categorizing a female visitor as "not making a purchase" include the duration of her visit falling within the range of 230 to 359.75 seconds.

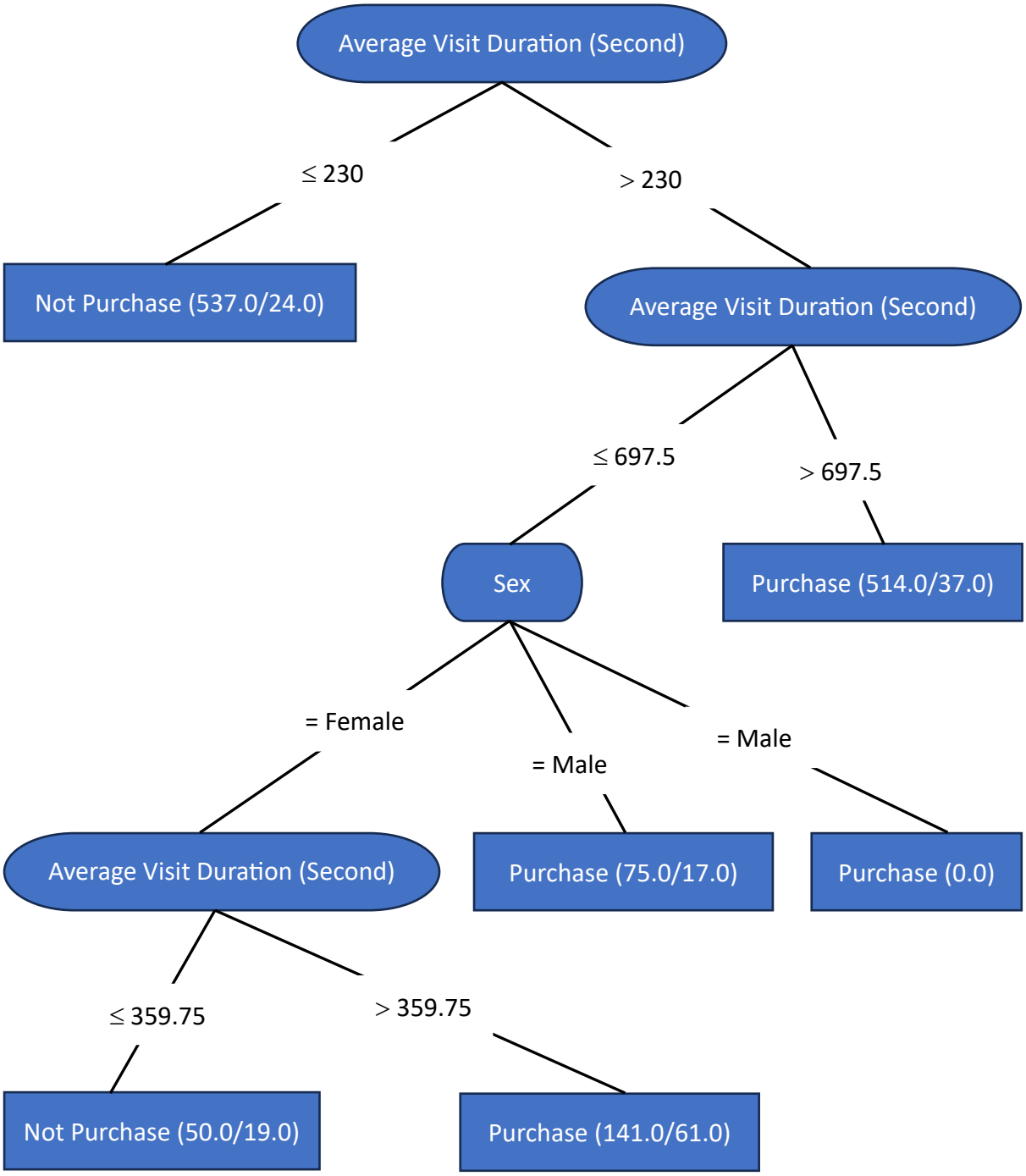

Fig. 3: Data set two's J48 Tree

5.9 Rules of Data Set 2

The principles obtained for JRip are as follows, in consideration of its exceptionally high accuracy:

JRip Rules:

Rule 1: Average Visit Duration > 248.5 seconds

- Condition: If the customer's average visit duration is more than 248.5 seconds.
- Classification: Classify as "purchase" with a ratio of 763.0/136.0.

Rule 2: Average Visit Duration > 166.67 seconds, Sex = Male

- Conditions: If the customer is male and the average visit duration is more than 166.67 seconds.
- Classification: Classify as "purchase" with a ratio of 18.0/7.0.

Rule 3: Default Rule

- Classification: If none of the above conditions apply, classify as "not purchase" with a ratio of 536.0/20.0.

5.10 Comparison of Results and Suggestions

In a thorough exploration of algorithmic performance, it is discernible that the KStar algorithm shines with an impressive accuracy rate of 88.2308% in dataset 1, showcasing its prowess in handling the intricacies of that specific dataset. On the other hand, J48 takes the lead in dataset 2, demonstrating its robust performance with the highest accuracy rate of 87.6234%. The subtle variations in accuracy rates between KStar and JRip algorithms underscore their comparable effectiveness, with both presenting reliable results. Notably, the KStar approach's exceptional accuracy, achieving the highest rate of 88.2308% in dataset 1, solidifies its standing as a formidable algorithm in this context.

Diving into the intricacies of rule-based classification, the JRip technique in Rules stands out as an exemplary performer, exhibiting exceptional accuracy across both datasets. This method crafts rules that prove adept at predicting class outcomes, establishing pivotal criteria for accurate classification. For instance, a customer with an average visit duration surpassing the

470-second threshold is definitively categorized as a "purchase." Similarly, individuals falling within the age bracket of 55 to 64, with an average visit duration exceeding 225.4 seconds, are unequivocally classified as "not purchase." Furthermore, male customers with an average visit duration exceeding 166.67 seconds are distinctly placed in the "not purchase" category. This level of granularity in rule formulation enriches the algorithm's interpretability and application.

Expanding the scope of algorithmic prowess, the J48 algorithm in Trees emerges as a prominent player, achieving the highest accuracy across both datasets. The decision tree generated by J48, as visually depicted in Figure 3, serves as an invaluable visual representation, unraveling additional layers of insight into the underlying rules governing the classification process.

Shifting focus towards an in-depth analysis of purchasing behavior, it becomes evident that an extended average visit duration, surpassing the 470-second mark, correlates positively with the likelihood of a purchase. This insight prompts a strategic recommendation: introducing a provision for free shipping on purchases exceeding a market basket value of 100 TL after a duration of 470 seconds, strategically capitalizing on the heightened probability of conversion during this timeframe. Additionally, a nuanced observation surfaces—customers aged between 55 and 64, despite exhibiting prolonged average visit durations, tend to exit the website without making a purchase. In response, a proposed strategy involves presenting targeted promotions or discounts within a specific duration of 225.4 seconds, aiming to incentivize this demographic to convert into making a purchase. This comprehensive approach not only deepens our understanding of algorithmic performance but also offers actionable insights for refining user behavior and steering purchasing decisions in a strategic direction.

CHAPTER 6
CONCLUSION

Watsons is concerned about unfinished-purchase consumers because this represents a significant issue for the business. Watsons must possess an in-depth understanding of consumer behavior and the factors that influence consumers to make purchases. The primary objective of this study is to forecast consumer behavior on the Watsons e-commerce platform through the implementation of a data classification strategy. Forecasting the behavior of initial visitors to the Watsons website constitutes the principal objective of the study. The Watsons data utilized in this study was gathered from March 19 to 25, 2018, as illustrated in Figure 1 of Section 4. Excel was utilized to cleanse the data by eliminating duplicate entries and missing data. Prior to implementing WEKA on data sets 1 and 2, seven variables were chosen for analysis: purchase status, date, age, sex, city, browser type, and average visit duration. There are two distinct categories of individuals: those who have engaged in purchasing activities and those who have refrained from doing so. We used different methods to sort the data, and the one that worked the best for each set was KStar for dataset1 and J48 for dataset2. KStar had a really good accuracy rate of 88.2308% for dataset1, while J48 had the highest rate of 87.6234% for dataset2. This means KStar was better at predicting in dataset1. We also got some rules from WEKA that were very accurate.

We have some suggestions for Watsons. If people leave the website without buying anything, we suggest offering them some recommendations to encourage them to come back.
Purchases increase when visitors spend over 470 seconds on the website. To boost sales, offer free delivery for orders over 100 TL after 470 seconds. Customers aged 55 to 64 may leave without buying, so consider quick promotions within 225.4 seconds to encourage purchases. To enhance predictions, add more details to the data. Create a system for better understanding and serving customers in the future.

REFERENCES

[1] Hernández S, Álvarez P, Fabra J, Ezpeleta J (2017) "Analysis of users behavior in structured e-commerce." websites. IEEE 5:11941–11958

[2] Ismail M, Ibrahim MM, Sanusi ZM, Nat M (2015) "Data mining in electronic commerce: benefits and challenges." Int J Commun Netw Syst Sci 8:501–509

[3] Satish B, Sunil P (2012) "Study and evaluation of user's behavior in e-commerce using data mining." Res J Recent Sci 1:375–387

[4] Jia W, Zhou H, Shi Y (2017) "Optimization of data-mining-based e-commerce enterprise marketing strategy." Revista de la Facultad de Ingenieraí U.C.V 32(13):273–278

[5] Çığşar B (2017) "Kredi risklerinde veri madenciliği sınıflandırma algoritmaları." Unpublished MS thesis, Çukurova Üniversitesi, Adana

[6] Perveen S, Shahbaza M, Guergachib A, Keshavjee K (2016) "Performance analysis of data mining classification techniques to predict diabetes." Procedia Comput Sci 82:115–121

[7] Anaklı Z (2009) "A comparison of data mining methods for prediction and classification types of quality problems." Unpublished. MS thesis, Middle East Technical University, Ankara 42 B. Altunan et al.

[8] Han J, Kamber M (2001) "Data mining: concepts and techniques. Morgan Kaufmann Publishers, Burlington"

[9] Rokach L, Maimon O (2006) "Data mining of improving the quality of manufacturing: a feature set decomposition approach." J Intell Manuf 17:285–299

[10] Astudill C, Bardeen M, Cerpa N (2014) "Editorial: data mining in electronic commerce - support vs. confidence." J Theor Appl Electron Commer Res 9. https://doi.org/10.4067/s0718-18762014000100001

[11] Huang H, Wu D (2005) "Product quality improvement analysis using data mining: a case study in ultra-precision manufacturing industry." In: Wang L, Jin Y (eds) FSKD 2005. LNAI, vol 3614, pp 577–580

[12] Kang B, Park S (2000) "Integrated machine learning approaches for complementing statistical process control procedures." Decis Support Syst 29:59–72

[13] Fan C, Guo R, Chen A, Hsu K, Wei C (2001) "Data mining and fault diagnosis based on wafer acceptance test data and in-line manufacturing data." In: IEEE, pp 171–174

DATA ANALYSIS OF CUSTOMER BEHAVIOR ON E-COMMERCE WEBSITES FOR PERSONALIZED MARKETING STRATEGIES.

ORIGINALITY REPORT

23%	21%	15%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	link.springer.com Internet Source	8%
2	"Proceedings of the International Symposium for Production Research 2018", Springer Science and Business Media LLC, 2019 Publication	6%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	4%
4	Submitted to Daffodil International University Student Paper	2%
5	www.coursehero.com Internet Source	1%
6	123dok.com Internet Source	<1%
7	dspace.bracu.ac.bd Internet Source	<1%
8	dspace.bracu.ac.bd:8080 Internet Source	<1%