United Arab Emirates University
Scholarworks@UAEU

Dissertations | Electronic Theses and Dissertations

11-2022

INTEGRATING KANO MODEL WITH DATA MINING TECHNIQUES TO ENHANCE CUSTOMER SATISFACTION

Khaled Abdulla Ali Al Rabaiei
930710357@uaeu.ac.ae

Follow this and additional works at: https://scholarworks.uaeu.ac.ae/all_dissertations

Part of the Computer Engineering Commons

Recommended Citation

Al Rabaiei, Khaled Abdulla Ali, "INTEGRATING KANO MODEL WITH DATA MINING TECHNIQUES TO ENHANCE CUSTOMER SATISFACTION" (2022). *Dissertations*. 189.
https://scholarworks.uaeu.ac.ae/all_dissertations/189

This Dissertation is brought to you for free and open access by the Electronic Theses and Dissertations at Scholarworks@UAEU. It has been accepted for inclusion in Dissertations by an authorized administrator of Scholarworks@UAEU. For more information, please contact mariam_aljaberi@uaeu.ac.ae.

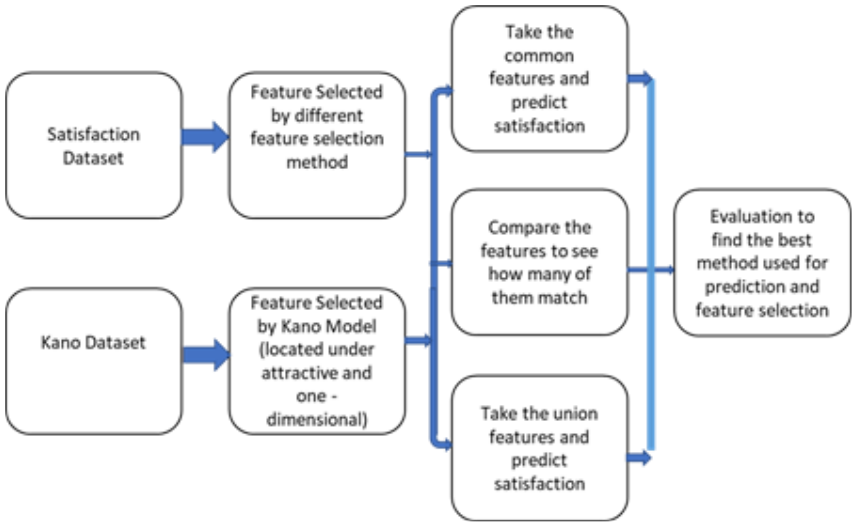
جامعة الإمارات العربية المتحدة
United Arab Emirates University

DOCTORATE DISSERTATION NO. 2022: 38

College of Information Technology

INTEGRATING KANO MODEL WITH DATA MINING TECHNIQUES TO ENHANCE CUSTOMER SATISFACTION

Khaled Abdulla Ali Al Rabaiei



November 2022

United Arab Emirates University

College of Information Technology

INTEGRATING KANO MODEL WITH DATA MINING TECHNIQUES TO ENHANCE CUSTOMER SATISFACTION

Khaled Abdulla Al Rabaiei

This dissertation is submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Informatics and Computing

Under the Supervision of Dr. Fady Alnajjar

November 2022

Declaration of Original Work

I, Khaled Abdulla Al Rabaiei, the undersigned, a graduate student at the United Arab Emirates University (UAEU), and the author of this PhD dissertation entitled "*Integrating Kano Model with Data Mining Techniques to Enhance Customer Satisfaction*", hereby solemnly declare that this dissertation is my own original research work that has been done and prepared by me under the supervision of Dr. Fady Alnajjar at the UAEU. This work has not been previously presented, published, or formed the basis for the award of any academic degree, diploma, or a similar title at this or any other university. Any materials borrowed from other sources (whether published or unpublished) and relied upon or included in my dissertation have been properly cited and acknowledged in accordance with appropriate academic conventions. I further declare that there is no potential conflict of interest with respect to the research, data collection, authorship, presentation, and/or publication of this dissertation.

Student's Signature:  Date: 08/12/2022

Copyright © 2022 Khaled Abdulla Al Rabaiei
All Rights Reserved

Advisory Committee

1) Advisor: Fady Alnajjar
   Title: Associate Professor
   Department of Computer Science and Software Engineering
   College of Information Technology

2) Co-Advisor: Amir Ahmad
   Title: Associate Professor
   Department of Computer Science and Software Engineering
   College of Information Technology

3) Member: Mohammad Masud
   Title: Associate Professor
   Department of Computer Science and Software Engineering
   College of Information Technology

Approval of the Doctorate Dissertation

This Doctorate Dissertation is approved by the following Examining Committee Members.

1) Advisor (Committee Chair): Fady Alnajjar

Title: Associate Professor

Department of: Computer Science and Software Engineering

College of Information Technology

Signature: __ Fady Alnajjar ____ _____ Date: 06/12/2022

2) Member: Zouheir Trabelsi

Title: Professor

Department of information system and systems and security

College of Information Technology

Signature: __ Zouheir _____ Date: 08/12/2022

3) Member: Jose Lopez Berengueres

Title: Associate Professor

Department of Computer Science

College of Information Technology

Signature: Date: 06/12/2022

4) Member (External Examiner): Khaled Shaalan

Title: Professor

Department of: Computer Science

College of Information Technology -The British University in Dubai (BUiD), UAE

Signature: K.Shaalan Date: 06/12/2022

This Doctorate Dissertation is accepted by:

Dean of the College of Information Technology: Professor Taieb Znati

Signature: ___Taieb Znati_____ Date: ___15/12/2022_____

Dean of the College of Graduate Studies: Professor Ali Al-Marzouqi

Signature _____Ali Hassan_____ Date ___02/03/2023_____

Abstract

The business world is becoming more competitive from time to time; therefore, businesses are forced to improve their strategies in every single aspect. So, determining the elements that contribute to the clients' contentment is one of the critical needs of businesses to develop successful products in the market. The Kano model is one of the models that help determine which features must be included in a product or service to improve customer satisfaction. The model focuses on highlighting the most relevant attributes of a product or service along with customers' estimation of how these attributes can be used to predict satisfaction with specific services or products. This research aims at developing a method to integrate the Kano model and data mining approaches to select relevant attributes that drive customer satisfaction, with a specific focus on higher education. The significant contribution of this research is to improve the quality of United Arab Emirates University academic support and development services provided to their students by solving the problem of selecting features that are not methodically correlated to customer satisfaction, which could reduce the risk of investing in features that could ultimately be irrelevant to enhancing customer satisfaction. Questionnaire data were collected from 646 students from United Arab Emirates University. The experiment suggests that Extreme Gradient Boosting Regression can produce the best results for this kind of problem. Based on the integration of the Kano model and the feature selection method, the number of features used to predict customer satisfaction is minimized to four features. It was found that either Chi-Square or Analysis of Variance (ANOVA) features selection model's integration with the Kano model giving higher values of Pearson correlation coefficient and R^2. Moreover, the prediction was made using union features between the Kano model's most important features and the most frequent features among 8 clusters. It shows high-performance results.

Keywords: Customer satisfaction; data mining; feature selection; The Kano model.

Title and Abstract (in Arabic)

تكامل نموذج كانو مع تقنيات استخراج البيانات لتعزيز رضا العملاء

الملخص

أصبح عالم الأعمال أكثر قدرة على المنافسة من وقت لآخر، وبالتالي ، تضطر الشركات إلى تحسين استراتيجياتها في كل جانب. من المعروف أن أي شركة تقدم منتجات أو خدمات بناءً على توقعات عملائها من المرجح أن تحقق النجاح في السوق. لذلك ، فإن عملية انتقاء السمات (features selection) التي تساهم في إرضاء العملاء هي إحدى الاحتياجات الحاسمة للشركات من أجل تطوير منتجات ناجحة في السوق. يعد نموذج كانو(Kano model) أحد النماذج التي تساعد في انتقاء السمات(features) التي يجب تضمينها في منتج أو خدمة لتحسين رضا العملاء. يركز النموذج على إبراز السمات الأكثر صلة لمنتج أو خدمة إلى جانب تقدير العملاء لكيفية استخدام وجود هذه السمات للتنبؤ بالرضا عن خدمات أو منتجات معينة. يهدف هذا البحث إلى تطوير طريقة لدمج نموذج كانو وأساليب التنقيب في البيانات (data mining) لتحديد السمات ذات الصلة التي تحفز رضا العملاء ، مع التركيز بشكل خاص على التعليم العالي. تتمثل المساهمة الكبيرة لهذا البحث في تحسين جودة خدمات الدعم والتطوير الأكاديمي المقدمة من جامعة الإمارات العربية المتحدة لطلابهم من خلال حل مشكلة اختيار الميزات التي لا ترتبط بشكل منهجي برضا العملاء ، مما قد يقلل من مخاطر الاستثمار في الميزات التي يمكن أن تؤدي في النهاية تكون غير ذات صلة بتعزيز رضا العملاء. تم جمع بيانات الاستبيان من 646 طالب وطالبة من جامعة الإمارات العربية المتحدة. تشير التجربة إلى أن تعزيز التدرج الشديد ((extreme gradient boosting ينتج أفضل النتائج لهذا النوع من المشاكل. استنادًا إلى التكامل بين نموذج كانو وطريقة اختيار السمه ، تم تقليل عدد السمات المستخدمة للتنبؤ برضا العملاء إلى أربع سمات. لقد وجد أن تكامل ايا من نموذج كاي التربيعي chi-square او تحليل التباين انوفا(ANOVA) مع نموذج كانو(Kano model) يعطي معاملات ارتباط بارسون (Pearson correlation coefficient) أعلى وقيم ار2 (R^2) (أعلى. تم إجراء تجارب إضافية لاختبار فائدة التكامل بين نموذج كانو واستخراج البيانات. كانت نتائج تجربة التكامل أظهرت أداء عاليا ، لكنها غير موثوقة بسبب العدد الكبير من السمات المستخدمة في عملية التنبؤ. علاوة على ذلك ، تم إجراء التنبؤ باستخدام سمات الاتحاد بين أهم سمات نموذج كانو Kano model والميزات الأكثر شيوعًا بين 8 مجموعات. وقد أظهرت نتائج عالية الأداء.

مفاهيم البحث الرئيسية: رضا العملاء. بيانات التعدين؛ اختيار ميزة؛ نموذج كانو.

Acknowledgments

Firstly, I'd like to convey my heartfelt thanks to my adviser, Dr. Fady Alnajjar, for his unwavering support during my Ph.D. research and study plus his patience, inspiration, and vast knowledge. His advice was always invaluable during the research and writing of this thesis. For my Ph.D. study, I could not have asked for a greater adviser and mentor.

Apart from my advisor, I'd like to express my gratitude to the rest of my thesis committee: Dr. Amir Ahmad, and Dr. Mohammad Masud, not only for their insightful comments and encouragement but also for the challenging questions that prompted me to broaden my research to include multiple perspectives.

Last but not least, I'd wish to express my gratitude to my family, particularly my mom, my wife, my daughters, and my sons, for their spiritual support during the composition of my thesis and my life in general.

Dedication

To my beloved parents and family

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AdaBoost	Adaptive Boosting
AHP	Analytic Hierarchy Process
ANOVA	Analysis of Variance
A-Kano	Analytical Kano
ADTree	Alternating Decision Tree
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
CHI	Chi-Square
CC	Correlation Coefficient
CFS	Correlation-based Feature Selection
CR	Customer Requirements
CS	Customer Satisfaction
DT	Decision Trees
DBSCAN	Density-based Spatial Clustering of Applications with Noise
EM	Expectation Maximization
EFA	Exploratory Factor Analysis
XGB	Extreme Gradient Boosting
FMCG	Fast Moving Consumer Goods
GMM	Gaussian Mixture Models

HoQ	House of Quality
IGA	Importance Grid Analysis
IPA	Important Performance Analysis
IG	Information Gain
IDS	Intrusion Detection Systems
Lasso	Least Absolute Shrinkage and Selection Operator
LR	Logistic Regression
ML	Machine learning
MAD	Mean Absolute Difference
MAE	Mean Absolute Error
MI	Mutual Information
NCSI	National Customer Satisfaction Index
OR	Odds Ratio
PRCA	Penalty-Reward Contrast Analysis
PCA	Principal Component Analysis
QFD	Quality Function Deployment
QoS	Quality of Service
R^2	R_Sequard
RF	Random Forest

RT	Random Tree
REP	Reduced Error Pruning
REPTree	Reduced Error Pruning Tree
RMSE	Root Mean Square Error
RSV	R-Square Value
SERVQUAL	Service Quality
SFFS	Supervised Forward Feature Selection
SVM	Support Vector Machine

Chapter 1: Introduction

1.1 Overview

This research focuses on the most known contributions in the literature of customer satisfaction prediction to enhance customer satisfaction by selecting the most important attributes. The new and developing markets are characterized by stiff competition, which necessitates robust strategies for the companies in an endeavor to meet ongoing customer requirements besides achieving and sustaining customer satisfaction. Therefore, organizations always strive to find and develop accurate protocols for ascertaining key parameters which affect customer satisfaction. Organizations that invest heavily in specific niche markets have been proven over the years to have better performance than those that follow generic production patterns. A significant number of studies have revealed that customer satisfaction is an imperative predictor of loyalty to a brand or service, which points out both the theoretical and practical value of studying customer satisfaction.

Organizations have started considering meeting customer needs in the changing business environment through market-oriented strategies. Organizations primarily focus on improving customer retention and satisfaction by investigating a good and encouraging association amidst customer satisfaction and loyalty (Anderson et al., 1994; Rust & Zahorik, 1993; Taylor & Baker, 1994). Managers must take measures to achieve customer satisfaction to ensure that utilizing these measures can benefit the organization to outperform competitors. Besides, it is a critical concern for managers to understand customer satisfaction dimensions. Several steps of customer experience were presented in the recent past (Meyer & Schwager, 2007). No one can deny the importance of analyzing customer experience in efficaciously realizing and

accomplishing organizational policy to ensure customer satisfaction (Martilla & James, 1977).

Some studies have revealed that the fundamental factor behind the behavior of customer purchases, accounting, and enhancement in customer numbers is customer satisfaction (Ittner & Larcker, 1998). Organizations thus aim to adopt strategies that satisfy their customers by analyzing and addressing their needs and demands (Yu, 2007). A few researchers have testified that customer satisfaction influences the growth of the customer base, which impacts organizational performance (Babakus et al., 2004; Yu, 2007).

Strategic integration of customer satisfaction into organizational operations is crucial, given its administrative importance and advantage. The presence of satisfaction requirements in administrative functions and products is essential to enhance customer satisfaction (Matzler et al., 1996), whose measurement is essential to understand its effects.

Various methodologies can be adopted to measure and explore customer satisfaction and its association with organizational operations. The methodologies' effectiveness, accuracy, and strength significantly determine customer satisfaction effects (Witell et al., 2013). To ensure and achieve customer satisfaction, organizations need to focus on customer needs because it is customer need that creates customer willingness to buy a product or service. The Kano model is one of the methods that help determine which features must be included in a product or service to improve customer satisfaction.

The Kano model could help managers better understand customer requirements (Avikal et al., 2020). The Kano model moves from a "more is always better" approach to a "less is more" approach, so adding one feature could be much better than adding

many features, which could have the opposite effect on enhancing customer satisfaction. On the other hand, clustering the customers into different segments using data mining techniques will allow the Kano model to improve satisfaction for each segment. Furthermore, comparing both approaches could support selection decisions and avoid removing attributes that could cause information loss (Au et al., 2012).

Data mining methods have made many advances in information processing and representation compared to traditional techniques. Various types of regression analysis used to assess Kano quality. Features According to the previous research, among all data collection techniques and surveys used, only those who used the direct classification method and kano questionnaire could categorize the features according to Kano's five categories (Attractive, performance, basic, indifferent, or reverse category) (Chen, 2012). A brief explanation of Kano's five categories is given here. The first category is Attractive. These characteristics are also called excitement requirements. They are quality characteristics that satisfy customers if present but do not make them unsatisfied when absent. Secondly, the category of must-be quality refers to characteristics that are also called basic requirements. In contrast, must-be quality characteristics define the opposite situation, so they would not satisfy customers when present but would make them feel dissatisfied when absent. Thirdly, one-dimensional quality characteristics cause customers to be satisfied when present but dissatisfied when absent. Fourthly, reverse quality characteristics improve customer satisfaction when absent and reduce it when present. Finally, indifferent quality characteristics do not affect customer satisfaction (Južnik & Kozar, 2017).

The main problem with existing data mining techniques is that the features selected do not represent the features correlated to customer satisfaction as domain knowledge. As aforementioned, the main contribution of this research is to solve the

problem of choosing elements that are not thoroughly correlated to customer satisfaction. The new model could reduce the risk of investing in features that could ultimately be irrelevant in enhancing customer satisfaction because it will exclude them.

This research has clarified how previous studies tried to assume customer satisfaction depending on data gathered through traditional questionnaires and online data collection. Feature selection techniques have been enforced to choose the most critical attributes to minimize dimensionality. Moreover, studies exploring the Kano model have applied the model without any integration with feature selection. The only combination was to group clients into different clusters, and then the Kano model was applied to draw out the users' requirements of each cluster. Nevertheless, to our knowledge, no study has proposed a model that combines the Kano model with feature selection to select and rank the most prominent attributes related to customer satisfaction, as presented in this proposal (Al Rabaiei et al., 2021).

1.2 Motivation

This research aims at developing a method to integrate the Kano model and data mining approaches to select relevant attributes that drive customer satisfaction with a specific focus on higher education. It also intends to apply data mining and feature selection techniques to predict customer satisfaction with a particular focus on the higher education field. Also, it intends to use data mining and feature selection techniques to predict customer satisfaction and check whether the chosen attributes can produce a similar prediction accuracy with all the details.

This kind of research requires datasets suitable for customer satisfaction analysis for both approaches: Kano Model and data mining techniques. However, since

there is no dataset available from previous research to satisfy both methods simultaneously, the intention is to conduct two types of surveys to meet both approaches. The total population sample will be drawn from United Arab Emirates University (UAE) University. Secondly, the data mining and feature selection techniques will be implemented, and their results will be compared with Kano's results. This approach could ultimately improve the selection of relevant attributes that drive customer satisfaction across different fields, including higher education and business, which may lead to enhanced customer loyalty and the market share of an institution or university.

1.3 Problem Statement

The main contribution of this research is to solve the problem of selecting features that are not methodically correlated to customer satisfaction. This could reduce the risk of investing in features that could ultimately be irrelevant to enhancing customer satisfaction. This research studies the degree of correlation between customer satisfaction and attributes; in the context of customer satisfaction, how can customer satisfaction be improved by integrating the Kano model with data mining techniques to select relevant attributes that drive customer satisfaction and reduce the risk of investing in features that could ultimately be irrelevant to enhancing customer satisfaction.

1.4 Research Questions

In this research, a proposed solution for the problem of selecting features that are not methodically correlated to customer satisfaction is presented. The following research questions are raised to address this problem and achieve the dissertation’s objectives. Figure 1 represents the research’s problem and subproblems.

1. How can integrating the Kano Model with Data Mining improve customer satisfaction?

2. Can the selected attributes achieve similar prediction performance as with all attributes?

3. How can irrelevant features to customer satisfaction be removed?

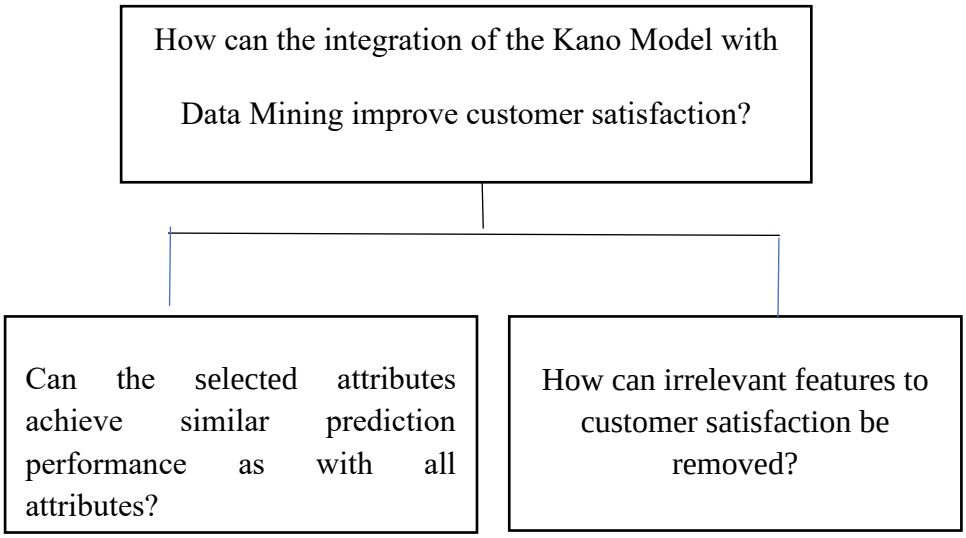


Figure 1: Research Problem

The Kano model has the advantage of classifying customer requirements into different categories (Attractive, performance, basic, indifferent, or reverse factors) (Aktepe et al., 2015). It could enhance the understanding of customer requirements. Therefore, integrating the Kano model with data mining techniques could enhance the process of selecting the aspects that are more significant for the clients' contentment. Moreover, the process could reduce the resources required to produce a particular product or service, consequently helping in efficient manufacturing.

1.5 Evaluation Criteria

Evaluation of the results will be carried out using a variety of performance assessment methodologies, for instance, the mean absolute error, root means square error, and R-square value (RSV) (Kazemi et al., 2015). These are the most widely used metrics while dealing with continuous variables, such as the one in question. It is possible to determine the correlation between the actual and estimated values of Y using the coefficient. Coefficients greater than one indicate the effectiveness of an approach. The accuracy of the prediction may be assessed using a specific error detection method. Root Mean Square Error (RMSE) is the sample standard deviation of disparities between planned values (y') and actual values (y) (y). As a rule, smaller mean absolute error numbers are attributed to higher performance. Similarly, Pearson correlation will be used to assess the overall connection between independent variables and their dependents. To measure the model's performance, R-square value is used. This statistic is also critical in the assessment of regression models. According to the authors, one of the most important measures for evaluating regression models is the R-square value. The R-square value is found in the range of 0 to 1. The higher the R-square value, the more accurate the models are in predicting the future. Also, one of the goals of the integration experiment was to find out the subset of attributes that can provide almost the same prediction accuracy as with all attributes besides knowing which attributes match between Kano and other feature selection methods (Amin et al., 2017).

1.6 Research Gap

To predict how to best improve customer happiness, this study examines both important contributions and research gaps. There have been many breakthroughs in

data processing and representation since the days of conventional approaches, but this study will demonstrate why the Kano categorization of feature classification has remained a challenge for data mining.

Data collection methods and surveys employed in the prior studies found that only the Kano questionnaire was able to identify characteristics according to Kano's five categories (Chen, 2012; Du et al., 2020). Data mining approaches now in use have a major flaw: the characteristic picked does not accurately reflect the attribute most closely associated with customer happiness. According to prior research, this study's key contribution is to overcome the issue of picking attributes that are not rationally tied to consumer pleasure. The new recommended strategy might lessen the danger of spending money on things that aren't going to have an impact on consumer happiness.

The Kano Model for Quality Improvement in Higher Education was used in earlier research to compare the existing situation with the ideal state of the quality indicators using a conventional survey (Xiong et al., 2021). Customers' needs were categorized into five categories, and the Kano Model was used to identify the most important features. To reduce the number of dimensions, feature selection approaches have been used. To top it all off, no feature selection approaches were used in the investigations looking into the Kano model. Clients were simply combined to form distinct clusters, and the Kano model was then used to extract the specific needs of each cluster's users.

According to the author, no research has yet produced a model integrating the Kano model with feature selection approaches to choose and rank the most significant qualities associated with customer satisfaction, as provided here (Al Rabaiei et al., 2021).

1.7 Methodology Statement

This research focuses on the notable contributions in the literature of customer satisfaction prediction to enhance customer satisfaction by selecting the most essential attributes. Though data mining methods have made numerous advances in information processing and representation as compared to traditional techniques, this research will show why they still have not resolved the problem of feature categorization according to the Kano categorization. This research will clarify how previous studies endeavored to assume customer satisfaction depending on data gathered through traditional questionnaires and online data collection. The proposed methodology is shown in Figure 2.

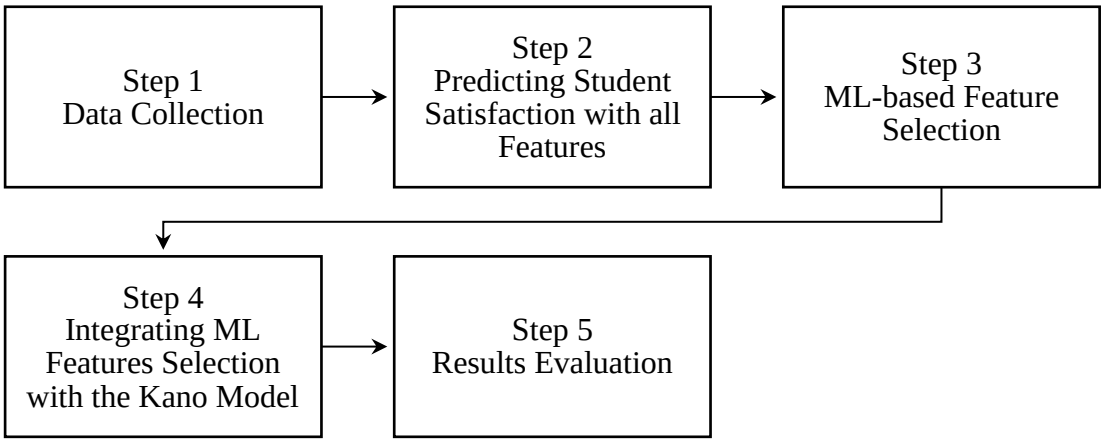


Figure 2: Proposed Methodology

Step 1: Data collection: The potential methods used in this problem are identified in the first stage. Here, the project combines two approaches; ML-based feature selection and the Kano model. At first, the questionnaire is developed based on the literature study findings. The developed questionnaire contains questions related to student satisfaction with the university. It has 38 questions. The questionnaire is shared with many students, and their answers are collected and saved

as the satisfaction dataset. The dataset contains 37 features and a student satisfaction rate on the liker scale (dependent variable). Then the second dataset, named the Kano dataset, is created based on the Kano model survey. Here, similar features will be coded according to the Kano model specification. The data used in the experiments had 37 attributes. It is difficult for university managers to concentrate on all the attributes.

Step 2: Predict student satisfaction using all the Features. At first, the prediction will be made using all the variables. Here, the prediction results will be used as a benchmark for comparing the results of the model developed after the feature selection model. As stated earlier, the primary intention is to attain closer results of the model with all features using a few selected features. For making predictions, a variety of machine learning models, like linear regression, logistic regression (Joshi et al., 2021), Decision Tree Regression, Random Forest Regression (RF), Adaptive Boosting (AdaBoost) Regression (Zhu et al., 2021), and Extreme Gradient Boosting (XGB) Regression, M5P, Random Tree and REPTree (Amin et al., 2017), are used.

Step 3: ML-based feature selection approaches like Correlation-based feature Chi-square, Mutual information, Least Absolute Shrinkage and Selection Operator (Lasso), Pearson, and Analysis of Variance (ANOVA) will be combined with the Kano-based feature selection approach (Sukarsa et al., 2021). According to several studies, the selected algorithms deliver better performance and are widely used for prediction problems (Park et al., 2013; Grömping, 2009; Joshi et al., 2021).

Step 4: Develop a new method to integrate data mining and the Kano model approaches to enhance the selection and ranking of the essential attributes to improve customer satisfaction. The selection technique can't categorize the features into five Kano categories, so the integration could reduce the risk of investing in features that

could ultimately be irrelevant to enhancing customer satisfaction. Here, various approaches like taking union among both datasets and taking standard features among the ML-based process as well as the Kano model, etc., have been tried. The selected approach takes the features of the Kano model and the machine learning approach. That will bring many features closer to the overall model results. Then the prediction will be made based on different ML algorithms, and the results will be tabulated and presented.

Step 5: Evaluation of the results will be done by different performance measures; correlation coefficient, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-Square value (Amin et al., 2017; Botchkarev, 2018). These evaluation techniques are the most popular metrics for continuous variables similar to our problem. Correlation coefficient finds the relationship between the predicted values, Y' and valid values Y. It can have values between –1 and 1. Higher values of correlation coefficient specify better performance of the regression methods. Mean Absolute Error was used to measure the closeness of the prediction to the eventual outcomes. Root Mean Square Error represents the sample standard deviation of the differences between predicted values (y') and observed values (y). The lower mean absolute error and root mean square Error indicate better performance. Also, the Pearson correlation will be used to find the overall correlation between the independent and dependent variables. The following measure used for evaluating the model is the R Square value. It is also one of the essential measures for evaluating the regression model (Amin et al., 2017; Botchkarev, 2019). It exactly shows the percentage of dependent variables measured by the model. According to the authors, the R square is one of the essential measures for evaluating regression models. R

Square value is founded between 0 and 1. The higher the R square value, the higher the models' performance.

The proposed research helps to find out the important features that have a maximum impact on the student satisfaction rate so that those few parameters to improve the student satisfaction rate can be focused on immediately. This paper uses the Kano Model and ML feature selection approaches to select the essential features that significantly impact student satisfaction. The research offers a method for mining students' happiness or satisfaction with the university based on significant features like lab facilities, dorms, teaching quality, etc. Primarily the author intended to sort out the major elements that affect the student's satisfaction with the university so that the universities can focus on these areas to improve student satisfaction. To achieve that, the Kona model is integrated with the ML techniques.

In literary research, data mining tools are often used to examine consumer pleasure. Data mining has been utilized to evaluate students' behaviour based on several variables, such as their usage of laboratory facilities.

1.8 Dissertation Contribution

This paper uses the Kano model and ML feature selection approaches to select the essential features that significantly impact student satisfaction. The paper offers a method for determining students' happiness or joy with the university based on features like lab facilities. The author primarily intended to sort out the major elements that affect the student's satisfaction with the university so that the universities can focus on those areas to improve student satisfaction. Figure 3 shows a high-level design of the integration between the kano model and data mining.

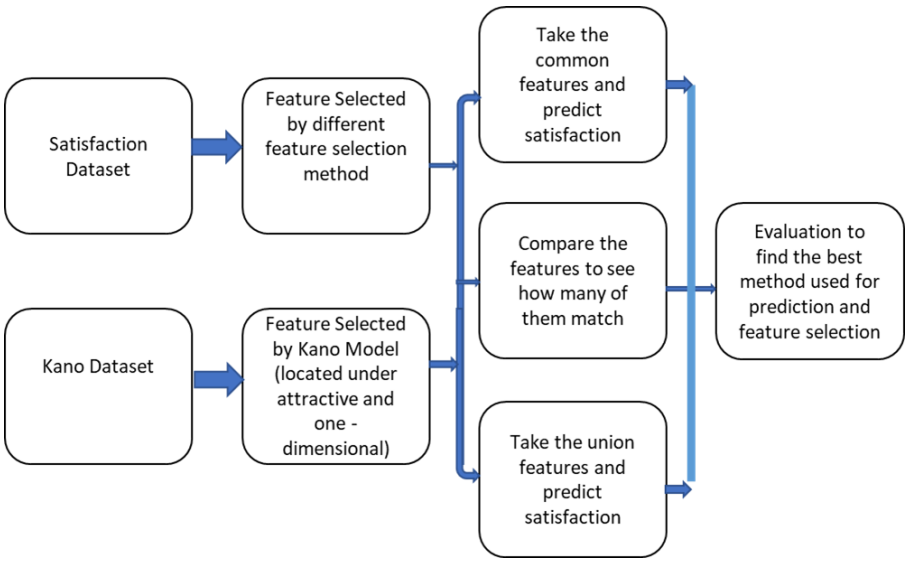


Figure 3: High Level Design of the Integration

1.9 Dissertation Structure

This research proposal is structured as follows. Chapter 2 discusses research on customer satisfaction methods and customer satisfaction prediction using data mining techniques. Chapter 3 will illustrate how integrating the Kano model and data mining could improve customer satisfaction. Chapter 4 presents the experiments. Chapter 5 covers the conclusion.

Chapter 2: Literature Review

Maintaining current clients, increasing market share, and increasing profit margins are all critical goals for any business. Corporations must go above and beyond to satisfy their customers' needs (Witell et al., 2013). It's safe to say that customer happiness is a critical factor in every company's success or failure (Kaya et al., 2018). To keep customers loyal, firms strive to fulfill and exceed the goals they have set for themselves. When a consumer is dissatisfied, it can cause a 'churn,' which can lead to the failure of the firm (Mikulić & Prebežac, 2011). An unhappy customer is a significant and tough challenge for every business. Customer retention is far more gratifying than the acquisition of new ones. Consequently, predicting consumer happiness has become a critical business idea. Conceptualization is garnering the attention of academics and corporations alike.

The Kano model can accurately categorize customer demands, such as Attractive, performance, basic, neutral, indifferent, or opposite aspects (Chen, 2012). Additionally, data mining algorithms take into account all possible combinations of patterns of interaction from all variables to rank characteristics (Zhao et al., 2019). Combining the two methods will allow you to reap the benefits of both. The five Kano classifications are briefly explained here. Attractive is at the top of the list. What makes customers happy if present but not unhappy is defined by the must be quality qualities (Južnik & Kozar, 2017). Customer satisfaction and dissatisfaction are caused by one-dimensional quality characteristics. Conversely, reverse qualities have the opposite impact. Apathetic qualities have no impact on client happiness. The Kano model can be used to better understand client needs. According to the Kano model, which advocates a "less is more" approach rather than a "more is better" philosophy, adding a single feature may be preferable to adding several, which could have the opposite

effect of increasing consumer happiness. In contrast, the Kano model will be able to improve customer satisfaction for each group by clustering customers into separate segments utilizing data mining methods. A comparison of the two approaches may also help in selecting and preventing the removal of features that could result in information loss (Avikal et al., 2020).

The Kano model is one of several practical techniques that managers may use to determine which product qualities are most important for customer satisfaction. Since the inception of this paradigm, academics and practitioners alike have shown an interest in it. There are theoretically five kinds of product characteristics that may be used as qualitative and quantitative aspects of a product (Zeinalizadeh et al., 2015).

Several customer satisfactions models, such as the Analytical Kano (A-Kano) model based on quantitative measures and fuzzy Kano approach and Kano model based on the classic conjoint analysis, may be utilized in research (Idris & Khan, 2017).

According to Hassan and Tabasum (2018) customer satisfaction and the fulfillment of customer needs may be linked via the Kano model, which uses both quantitative and qualitative methodologies. Using the Fuzzy Kano questionnaire, the most important factors of food quality were identified.

A wide range of scholars have used the Kano model to support their point of view in several ways. Consumer satisfaction, according to experts, is influenced by factors such as product and service quality as well as the availability of the product. According to Adjimi et al. (2019) an organization's capacity to meet customer expectations requires a deep awareness of what its customers want and expect from it. Noriaki Kano developed and released the Kano model more than three decades ago

with an intention to make it simpler for consumers to grasp the characteristics of a product or service while keeping the requirements of the customers in mind. This paradigm of social psychology, developed by Kano, has been around for quite some time. Researchers were able to establish three types of expectations that may be broken down in terms of service, and they were able to categorize them. It has been publicized that the fulfillment of the parameters listed above has a substantial impact on customer satisfaction. Thanks to a distinguishing characteristic, the creative design guide may reap the benefits of the categorization technique.

Using artificial intelligence-based algorithms to solve this kind of issue is perfect since they are meant to hunt for hidden qualities and commonalities to link clusters of data that have certain attributes (Olsen et al., 2014). In addition, they may forecast factors such as pricing, weather, and customer preferences. It is possible to categorize consumers into artificial intelligence groups based on the qualities that they have in common to foresee customer behavior (Othman et al., 2017).

The identification of patterns in vast amounts of data or data that has already been obtained by a corporation may help enhance the customer experience. It is expected that the size of this industry would develop greatly in the next few years as a consequence of the expansion of this industrial sector (Violante & Vezzetti, 2017). Customers may become unsatisfied as a consequence of the usage of data mining technologies. The key objective of this study is to choose a feature set and machine learning model that employs the fewest variables and features while reliably anticipating the result (He et al., 2016). This questionnaire seeks feedback about around 37 different facets of student satisfaction with the institution as a whole. So, for a college or university administration to increase student happiness, administration

ought to pay attention to all 37 characteristics of student happiness. The issue, however, is that it will take a significant amount of time and money to put into effect satisfiers, which, according to Gacto et al. (2019) are performance qualities. These characteristics contribute to the overall satisfaction of the client with the product or service. They are not required by the product in any way. Exciting features, also known as surprise components, were found, in Gao et al. (2018) to provide goods with a competitive advantage over their competitors' offerings. According to Chen et al. (2017) it’s not clear whether this feature is necessary for the product to perform properly. Customer satisfaction is directly impacted by these characteristics. A study by Hazra et al. (2016) indicated that buyers were happier with products that included just the essentials. Customer satisfaction is boosted by delighters and one-dimensional qualities, which give consumers the impression that they have the greatest product or service in hand. It also creates the impression that they are distinct from others (Xiong et al., 2021). For something to be classified as having an appealing quality, it must have certain features that enhance consumer happiness when they are present, but do not cause dissatisfaction when they are not present. Exciting needs, as they are often referred to, might be seen as little extras that make consumers happier but are not anticipated by them (Xiao et al., 2015). On the other hand, qualities that must be present are those that do not satisfy clients when present but make them unhappy when missing.

Predicting the behavior of customers’ unstructured data is well-suited for AI-based algorithms, which search for hidden (Farhadloo et al., 2016) features and commonalities to link clusters of data that have specific properties. Furthermore, these models are capable of forecasting price, weather conditions, and customer preferences. It is possible to create customer behavior predictions by segmenting consumers into

artificial intelligence groups since customers with similar traits are more likely to buy the same item (Bell & Mgbemena, 2017).

Marketers may enhance their service to potential and existing customers by detecting patterns in big data or data already collected by an organization. With the expansion of this industry, it is expected to grow much more in the years to come (Lin & Vlachos, 2018). Using data mining technologies is prone to have problems with customer satisfaction. The main idea here is to select the appropriate feature selection combination and ML model that predicts the maximum possible accuracy by using the minimum number of variables or features (Raschka et al., 2020). In this case, the developed questionnaire contains 37 features related to student satisfaction in the university. So, if the college or university management wants to increase the student satisfaction rate, it would need to concentrate on all 36 features. However, the problem is that it practically takes a lot of time as well as resources.

According to Ingaldi and Ulewicz (2019) performance attributes are also known as satisfiers. These attributes increase the customer's enjoyment of the product or service. They do not come under the basic requirements of the product. Madzík et al. (2019) revealed that Attractive attributes, also known as surprising elements, offer the uniqueness from the products of rivals and the competitive edge to the product.

According to Gupta and Shri (2018) customers do not know whether they want this feature or not for the functioning of the product. However, these attributes increase customer satisfaction directly. Turisová (2015) found that the basic features provide more satisfaction to the customers. However, along with the basic functioning features, the usage of delighters and one-dimensional attributes increase customer satisfaction because it makes the customers feel that they have the best product or service in hand. It also gives the feeling that they have something different from the common ones

(Shahin & Akasheh, 2017). The category of attractive quality refers to characteristics of a product that can improve customer satisfaction if they are present but do not make customers dissatisfied when absent. These characteristics, also called excitement requirements, can be observed as minor bonuses that make customers more satisfied but are not expected by the customers (Tontini, 2007). On the other hand, the category of must-be quality refers to those characteristics which would not make customers satisfied when present but would make them dissatisfied when absent.

Data mining was utilized to evaluate students' behavior based on several variables, such as their usage of laboratory facilities. Different research papers discussed how one-dimensional and delight features are related to satisfaction.

2.1 Customer Satisfaction

Customer satisfaction is an essential factor for the growth of a company or organization. The customer satisfaction prediction finds out the information about customer satisfaction and happiness with the service and products that the company sells to the customers. Different methods predict customer satisfaction after data analysis that aims to improve the services and quality of products.

2.2 Customer Satisfaction Predication

Generally speaking, in customer relationship management which deals with customer development, customer retention, customer attraction, and customer identification, the combination of data mining and machine learning has been widely used to investigate the relationship between different factors. So, machine learning is the mainly tested technique in this research that relates influential factors to customer satisfaction. A group of machine learning models have been tested to evaluate the best

model in our problem besides identifying the remarkable factors that can promote the education process efficacy.

2.3 Data Mining Algorithms for Prediction

Data mining is the process of transforming data from raw form into meaningful information. This era is the age of data, and its analysis is a must (Hand et al., 2007). Every institute can benefit from its data. Hospitals can detect trends of flu during winter. Search engines can choose the best places to put an advertisement. Stores can determine the most requested items in certain period. Hotels can detect the most relevant features that affect customer satisfaction (Han et al., 2011). The last example is major in our research project. Our main interest concerns the tools of data mining that can help in specifying the best features that contribute to customer satisfaction.

Data mining starts with preprocessing to understand and clean the data e.g., outlier detection for detecting the entries in the dataset that are not meaningful. In our dataset, this may be ratings of all 1 or all 10, which is not realistic or meaningful. Other preprocessing technique is association detection (AD). As an example, some features are highly correlated like gender and football playing, so having both features in the dataset will be misleading for any machine learning model. For this reason, only one of the highly correlated features stays in the dataset so that better results could be obtained (Dasu & Johnson, 2003).

For analyzing the survey data, a couple of ML models like Multi linear Logistic Regression, Decision Tree Regression, Random Forest Regression, AdaBoost Regression, XGB Regression, and Random Tree were used also some deep learning method like Multilayer Perception (MLP) and Convolutional Neural Network (CNN) were used. For tuning the model and improving the performance of the model, some

of the common feature selection methods like Pearson Correlation-based feature selection, Chi-Square-based feature selection, Mutual information Lasso feature selection, and ANOVA t-test based feature selection have been used (Pandey et al., 2020). The following subsections will provide a brief explanation of the most common data mining tools and prediction techniques. Moreover, examples on similar research projects that have used these tools will be provided. Lastly, a justification of using some of them in this project will be provided.

2.3.1 Decision Tree Regression

A decision tree makes the classification or regression models in tree form. It splits the data into progressively smaller subgroups or sets and develops a tree in a step-by-step manner. The result in the form of output is a tree that has leaf nodes and nodes of the decision. The two more branches of the decision tree show the attributes and values tested. Leaf node basically provides information about numerical value. Decision trees are capable of dealing with both category and statistical numeric information data. The best prediction about the attributes and features are obtained from the top root node. A different test, VI and VIF calculation, is performed to display the small value of the top variable (Tirenni et al., 2007). The performance has been improved as a result of the elimination of all of the unwanted elements.

In the case of complex interactions among the capability and the variable output, decision trees can be extremely useful. They also perform well when compared with other methods (algorithm) if there are lacking capabilities, if there is a mixture of specified and numerical information, and if there is a significant difference in the size of features among other situations (Tirenni et al., 2007).

Neural network is outperforming many ML models, especially when it comes to unstructured data such as images. However, with small structured, and tabular data, Decision trees (DT) based algorithms are still considered to be the best which direct us toward their usage especially with the nature of the data of this project (Luo et al., 2021).

Decision trees are widely used as a decision-making model to develop classification or regression models based on tree topologies. It progressively subdivides a dataset into smaller and smaller subgroups while simultaneously constructing a decision-making tree to represent the data) (De Caigny et al., 2018). The tree-shaped topology is mainly composed of three types of nodes.

The root node of a decision tree is the node at the top of the tree that corresponds to the best prediction where first branching-based numerical calculations take place. The second type of nodes deals with the inner nodes where another decision is to be made based on specific criteria. A decision node is composed of two or more branches, each of which represents a value for the feature being checked (Gokhan & Keceoglu, 2019). In the binary decision trees, each item is to be set under the right branch if it fulfills the criteria, otherwise it is set under the left branch. The branches might be more than two according to the data construction. The third node type is the leaf node which is found at the lowest level of the tree. The leaf node contains the final decision whether it is a specific class in classification problems or a specific value in regression problems. The final decision that is represented by the leaf node is according to different combinations of fulfilled criteria or requirements of the above nodes. Numerous tests, including multicollinearity test, VIF calculations, and IV calculations on variables, may be performed to narrow the field down to a small

number of top variables. Therefore, performance is enhanced since all the undesirable factors have been eliminated (Christa et al., 2022). Figure 4 shows a schematic view of decision tree architecture and how it works (Luo et al., 2021; Leonard, 2017).

Like the most of machine learning models, a dataset contains list of samples. Each sample has its own features, which are needed to construct the decision tree. Each sample's features then undergo a series of tests beginning with the root node, passing by the inner node. Each test divides the dataset into samples that share the same outcome in this test. This dividing process and testing keep going until final subset of samples is grouped and does not accept any mode divisions. Each final subset of samples represents a leaf node.

According to the research published by Choi et al. (2008), DT has many advantages that made us eager to test it for our problem. One of these important advantages is its usage in detection and prediction of customers' behaviors. In addition, it has the ability to extract models describing important data classes. DT is easy to understand and interpret as it can be summarized in a set of if-else statements, which makes it useful in the field of marketing to find out influential factors. For non-academic fields, such as marketing and business having people with machine learning not their domain of interest, visual representation of the model is an important criterion to illustrate how reasonable are your results.

Decision Tree models have been used in customer satisfaction similar problems. The research published by Choi et al. (2008) investigated the factors that influence customer satisfaction and loyalty of m-commerce and e-commerce. The authors endeavored to prove the essential influence of "content reliability" and "availability" in addition to "perceived price level of mobile Internet (m-Internet)" to

m-loyalty and m-satisfaction. They used decision tree to compare their proposed important features with the current e-commerce. Choi et al. (2008) used customer satisfaction as the target output, which in the language of machine learning is called label, and different customer satisfaction factors as the ml model input that are used in decision making. Their constructed DT was a binary tree that is built on binary splits on each node as shown in Figure 4 (Luo et al., 2021; Leonard, 2017). The best splits are determined based on entropy indexing. In addition, their labels are set to be binary; the customer is either satisfied, the label is to be 1, or unsatisfied, the label is to be 0. In these settings, they managed to pinpoint both the unique and parallel features of m-commerce.

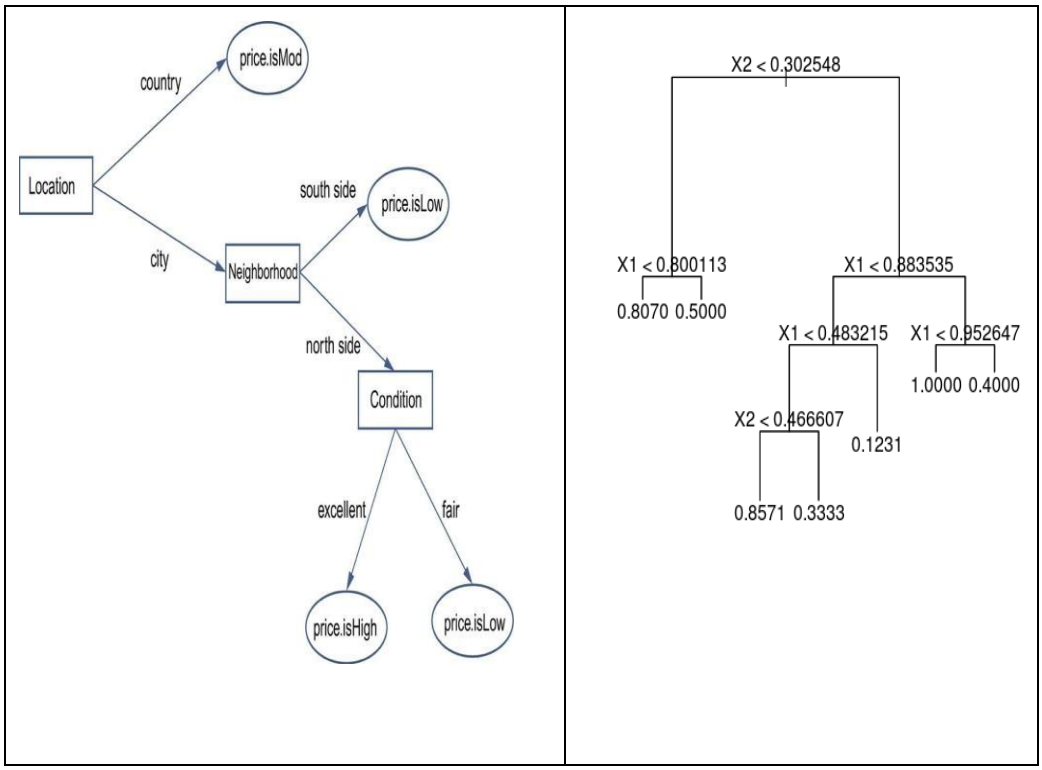


Figure 4: A Schematic View of Decision Tree Regressor Architecture

Another research published by Tama (2015) that was focused on fast-food industry used both DT and neural network as machine learning models. They have the same common purpose; identifying the essential factors that participate in customer satisfaction. They proposed a pipeline that resembles most of this area's pipelines with a special usage of DTs (Figure 5) (Tama, 2015). Both decision tree and neural network achieved more than 80% of predictive accuracy.

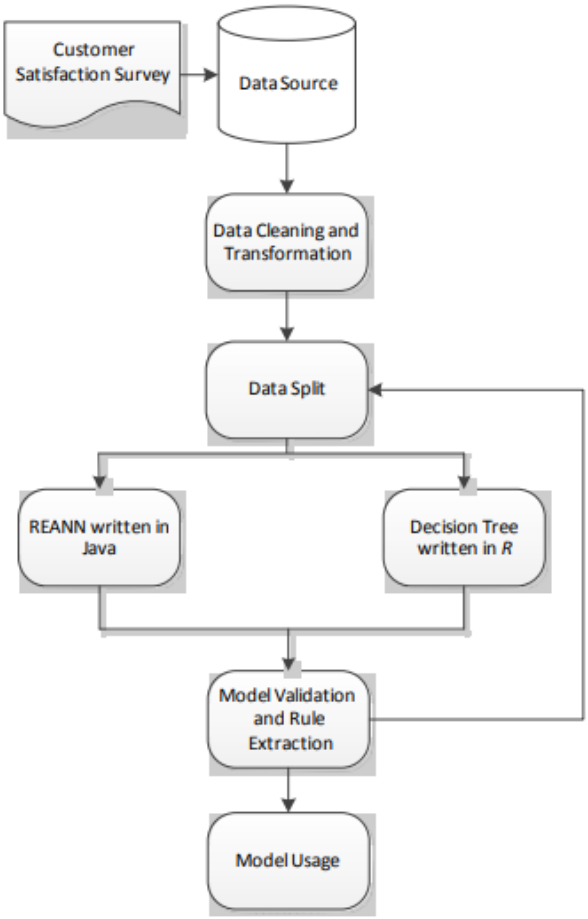


Figure 5: Diagram of Research Process

Moreover, one of the customer satisfaction problems being investigated is transport service quality. This problem also correlates with users' perception and expectations with the same data collection method as the one we used; customer

satisfaction survey. Tsami et al. (2018) achieved an accuracy of 89.5397% in one of their research projects related to ours that used DT model in classification. They built a DT that has 51 nodes and 26 leaves (end nodes). Figure 6 shows an example of a binary tree (Galimberti & Soffritti, 2011).

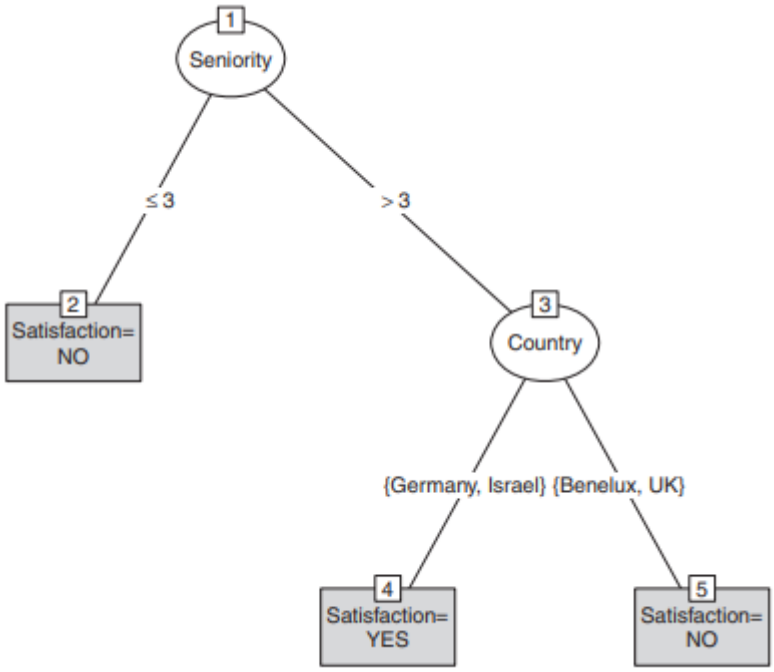


Figure 6: An Example of A Binary Tree

2.3.2 Multiple linear Regression

A statistical method known as multiple linear regression is used to describe the concurrent relationships between numerous variables and one continuous outcome. The estimating and inference processes, variable selection during model construction, and model fit evaluation are crucial elements in applying this technique. Regressions with categorical (grouping) variables, polynomial regressions, regressions with interactions between the variables, and distinct slopes models are specific instances that are also treated. The entire time, examples from microbiology are used (Eberly, 2007). It is assessing the link between variables that have a relationship between cause

and effect is regression analysis. it analyzes the relationship between a dependent variable and a single independent variable and to create a linear relationship equation between the two. Multilinear regression is the name given to regression models with one dependent variable and several independent variables (Uyanık & Güler, 2013). Equation (1) shows a linear regression formula in which $\hat{Y}$ is the predicted value of the response variable Y for a given value of the predictor variable X. The intercept b0 estimates the value of the response when the predictor is 0, and the slope b1 estimates the average change in the response for a unit change in the predictor. Equation (2) shows a multiple linear equation in which $\hat{Y}$ is the value of the response predicted to be on the regression plane with the best fit (the multidimensional generalization of a line). The intercept b0 is the reference position of the plane; it defines the value of Y when both X1 and X2=0. The regression coefficient b1 quantifies the sensitivity of Y to a change in X1, taking into account the effect of X2 on Y. b2 quantifies the sensitivity of Y to a change in X2, taking into account the effect of X1 on Y.

$$\hat{Y} = b_0 + b_1 X \quad (1)$$

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 \quad (2)$$

The multi–Linear Regression is used for solving Regression problems whereas Logistic Regression is used for solving the Classification problems. Under the umbrella of supervised learning, logistic regression is one of the powerful machine learning models. In the empirical research, logistic regression is a statistical technique that is often used to analyze categorical dependent variables. An individual's class (or category) may be predicted using the statistical method of logistic regression, which is based on one or more factors (x). Logistic regression is a transformed form of the

linear regression for the classification problems with logistic regression having range between 0 and 1 (Buyya et al., 2016). Besides, linear regression requires linear relationships between inputs and labels in contrary to logistic regression, which is considered as an advantage over linear regression for our problem of interest, because in logistic regression, nonlinear log transformation to odds ratio is applied in the first place. The Odd of event is a probability of an event taking place divided by the probability of an event not taking place. It is mentioned earlier that logistic regression has the range between 0 and 1. Sigmoidal shape (s-shaped), therefore, represents the probability curve on a binary scale. As an example, let's apply values −20 to 20 to the logistic function. The input values will be transferred to 0 and 1 as illustrated in Figure 7 (Belyadi & Haghighat, 2021).

Sometimes, variable is dependent and discrete. The logistic regression is the precise evaluation regression of behavior that may be performed (binary). A prediction evaluation is performed using logistic regression in the same way as it has done with all other regression analyses (Yi et al., 2019). When attempting to explain the information or relationship between a binary structured variable and one or even more variables that are independent and ordinal nominally c program language period, or ratio-stage in nature, logistic regression is employed to do so. A prediction evaluation is performed using logistic regression in the same way as it is done with all other regression analyses.

Logistic regression is considered as a statistical method of analysis variation on the basis of no or yes. Different attributes in customer satisfaction include quality of product, prices of the product, the quantity of product with the increase in price, and market values (Tirenni et al., 2007).

When it comes to reading logistic regressions, it might be tricky. Nonetheless, the Intellects gadget of statics makes it simple to finish the evaluation and, after that, translate the outcomes into unmistakable English by utilizing the incorporated translation.

Since it is simple to implement a broad range of applications, it may serve as a performance basis for several systems. As a result, each engineer should be acquainted with the ideas it contains (Hung et al., 2018). The often-used logistic model is the one with binary outcome. Multinomial logistic regression will be used, as our problem is multi-output class problem, which means that there are more than two discrete outcomes (Kwak & Clayton-Matthews, 2002). Logistic regression has a less complicated mathematical background than Multinomial, so it is better to explain logistic regression first in this context. The below equation explains the mathematical background of logistic regression model, which is represented by what is called logistic function (Belyadi & Haghighat, 2021). The following equation is used in case of a problem of binary output. Figure 8 shows a linear regression equation on a linear scale (left) and a logistic regression equation on a probability scale (Seufert, 2013).

$$logisticfunction = \frac{1}{1+e^{-x}} \qquad (3)$$

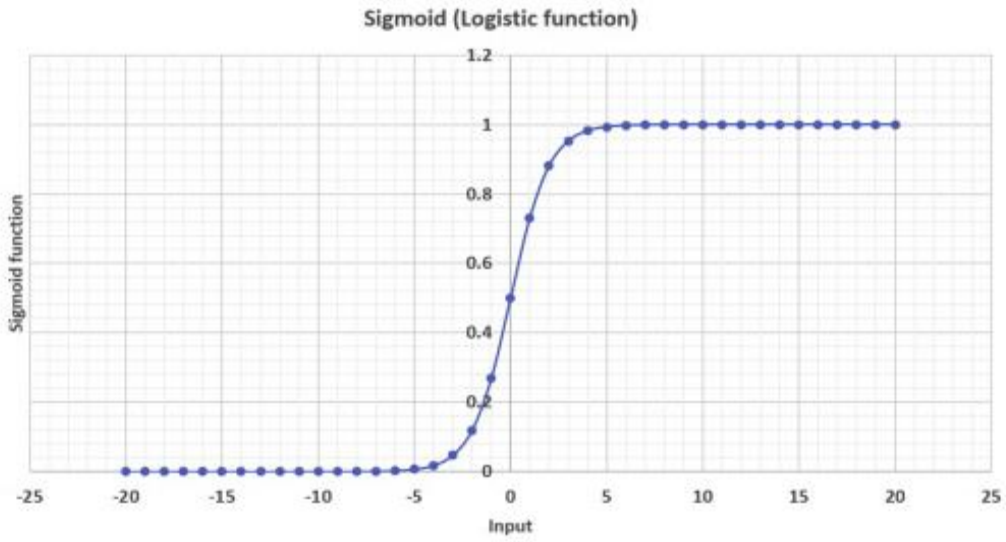


Figure 7: Logistic Regression Applied to A Range of −20 to 20

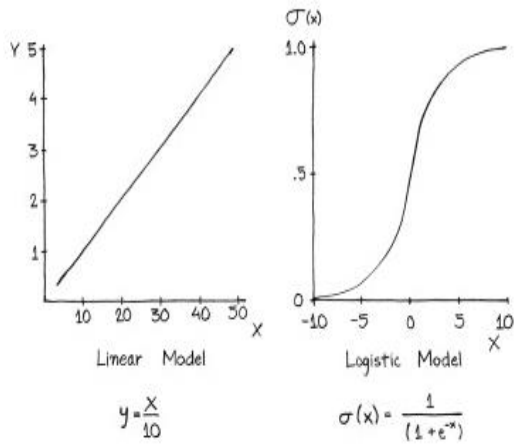


Figure 8: A Linear Regression Equation on A Linear Scale (left) and A Logistic Regression Equation on A Probability Scale (right)

2.3.3 Random Forest Regression

Random forest (RF) is an ensemble learning method used for the classification and regression (Južnik & Kozar, 2017). RF Regression is a supervised learning technique that makes use of a regression learning methodology to obtain its results (Gómez Fernández et al., 2022). Using ensemble learning, one may build a forecast

that is more accurate than a single model by combining predictions from multiple algorithms simultaneously (Iannace et al., 2019).

The Random Forest is construct, wherein the trees run parallel to one another, but do not meet one another at all. Random Forests are used to train decision trees since they build multiple decision trees at once and give the mean class for all the trees (Pekel, 2020).

Random forest means an assemblage of decision trees as illustrated in Figure 9 (Chapron et al., 2018). Each decision tree uses different samples and features in making its decision. Random sets of samples are generated. Then, each set is to be used for one DT. Finally, entire forest votes for the final decision. Hence, RF corrects decision tree defect of over-fitting. Moreover, RF has an advantage over decision tree. It is not just about constructing the different bootstrap samples of the data that are used in decision tree construction nor using the vote of many decision trees, rather it deals with the construction of RF's decision trees themselves. DT takes the decision of each node splitting based on the best split amongst all variables, while RF takes the decision of each node splitting on the basis of the best amongst a subset of predictors haphazardly (Boateng et al., 2020). This unexpectedly supports the fact that often RF performs very well as compared to numerous other classifiers, including support vector machine (SVM), discriminant analysis, and NNs, which avoids overfitting (Boateng et al., 2020).

RF specifically has been chosen to be investigated in this research because its results are interpretable, which means that the features used in decision making must be known. Lack of interpretability of many machine learning approaches is a strong limitation contrary of RF. Earlier random forest was used in customer satisfaction

problems many times, such as the research published by Baswardono et al. (2019) that is covering the classification of airlines' customer satisfaction. The authors conducted a comparative analysis between different decision tree algorithms, RF and C4.5. Both algorithms show quite similar accuracies, precisions, recalls, and area under the curve (AUC) that are considerable. The best accuracy they reached was 93%, which was achieved by RF after tuning the parameters. They built a system on the basis of RF with the intent of analyzing historical mobile data usage and profiles of the customer.

Another research conducted by Hu et al. (2018) recommended using RF in telecom promotion recommendation. The traditional methods that were used for offers' promotions depended merely on experiences and personal intuition. The authors of the paper suggested that consumption level and mobile data usage pattern of the customers are the important features that should be considered in decision making. Based on the researchers' proposed RF-system in this paper, they managed to improve accuracy from 80.36% to 93.36% as compared to the accuracy that was achieved by the traditional methods for offers' promotions. Given that their data has a quite massive number of samples, which was more than 500 thousand mobile data usage, their results are reliable enough for us to depend on in the process of choosing RF among our ML models.

The review article published by Boateng et al. (2020) confirmed that most researchers advocated RF as an easier and extensively utilized method, which recurrently achieves results with high precisions, and customarily quicker to implement. Besides, it was mentioned in the article that RF is insensitive to noise or overtraining and demonstrates the capability of dealing with the unbalanced data. All the previously mentioned research contributions strongly directed us toward using RF.

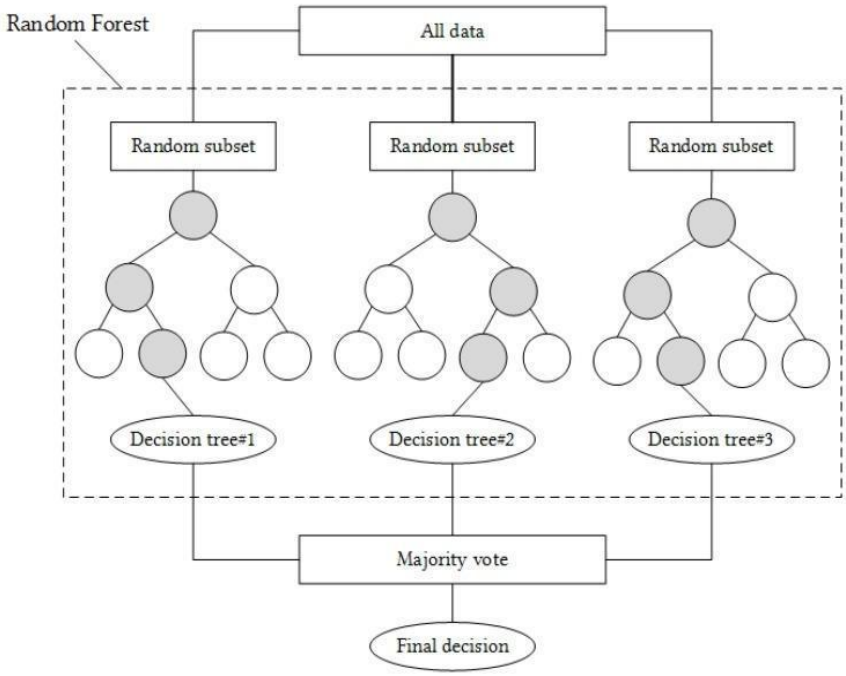


Figure 9: A Schematic View of Random Forests Architecture and How it Works

2.3.4 Adaptive Boosting (AdaBoost) Regression

AdaBoost develops and assembles itself mostly via the efforts of succeeding members that have been trained to correctly predict the appearance of certain data events (Xiong et al., 2021). Each new predictor is provided with a training package that includes progressively difficult examples that may be weighted or resampled as they go through the training process (Shahraki et al., 2020). It is a straightforward meta-estimator that begins by fitting an instance regressor to the original dataset, and then fits further regressor copies to the same dataset, but with the weights of the instances modified to account for the current prediction error (Koduri et al., 2019). Therefore, successive regressors lay emphasis on more complicated circumstances.

AdaBoost regression is a type of regression which is a primary effort of subsequent members that have been trained to accurately estimate the presence of specific factual events that AdaBoost builds and gathers itself within the natural course

of things. It is possible to reduce the influence of large datasets by using adaptive boosting (AdaBoost), which is used for cascading numerous decision trees (Tirenni et al., 2007). When a new predictor is introduced, he or she is given a new offer of education that contains progressively harder instances that can be weighed and resized as it proceeds through the process of learning.

AdaBoost is regarded as a reliable Meta estimator because it begins by fitting a specific case of a regression model with a distinguishable set of data, and afterward fits perfectly additional regressor duplicates with a similar set of data and with the strength of the times adjusted to compensate for the present forecasting of faults. As a result, successive regressors lay emphasis on circumstances that are more intricate (Yi et al., 2019).

Boosting is a repetitive predator algorithm. It mainly depends on creating prediction model of the training dataset, and then building a second model that rectifies the first one, followed by a third and fourth model. Each model rectifies the previous one until the model reaches stopping criteria that indicates good predictive capacity of the final mode (Zhang, 2004). Boosting is a general idea that resembles the idea of Random Forest. Random forest builds multiple DT, and then takes vote on them all to decide its classification. The same is valid for boosting. It takes the voting of multiple machine learning models. Each of them is a week predictor by itself, but their combinations increase the predictive capacity. For example, if, as a start, KNN model was created, and it achieved an accuracy of 80%, then this is followed by creating a DT model that achieved 75% accuracy, and finally a third model of SVM was created with a predictive accuracy of 85%. All the three models have a low predictive capacity. However, their combined voting is expected to show better results.

One type of boosting is called adaptive boosting that is used as an ensemble method. Commonly, it uses Decision Stumps as an algorithm. Decision Stumps is basically a DT with one split Figure 10 (Bohacik, 2014). Initially, it builds a model with equal weights to all samples, then it builds a second model with updated weights. The weights of samples are updated according to whether it was classified correctly in the first model or not. If it was correctly classified, it will be given higher weight to pay more attention to in the next model. These procedures keep going until reaching an acceptable margin of error. A schematic review of AdaBoost mechanism is illustrated in Figure 11 (Wang & Li, 2021).

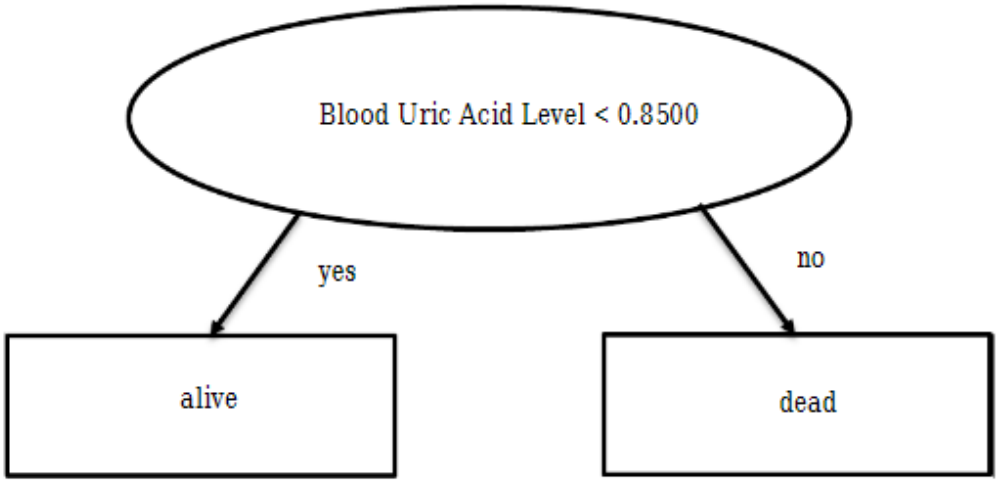


Figure 10: A Schematic View of Decision Stumps

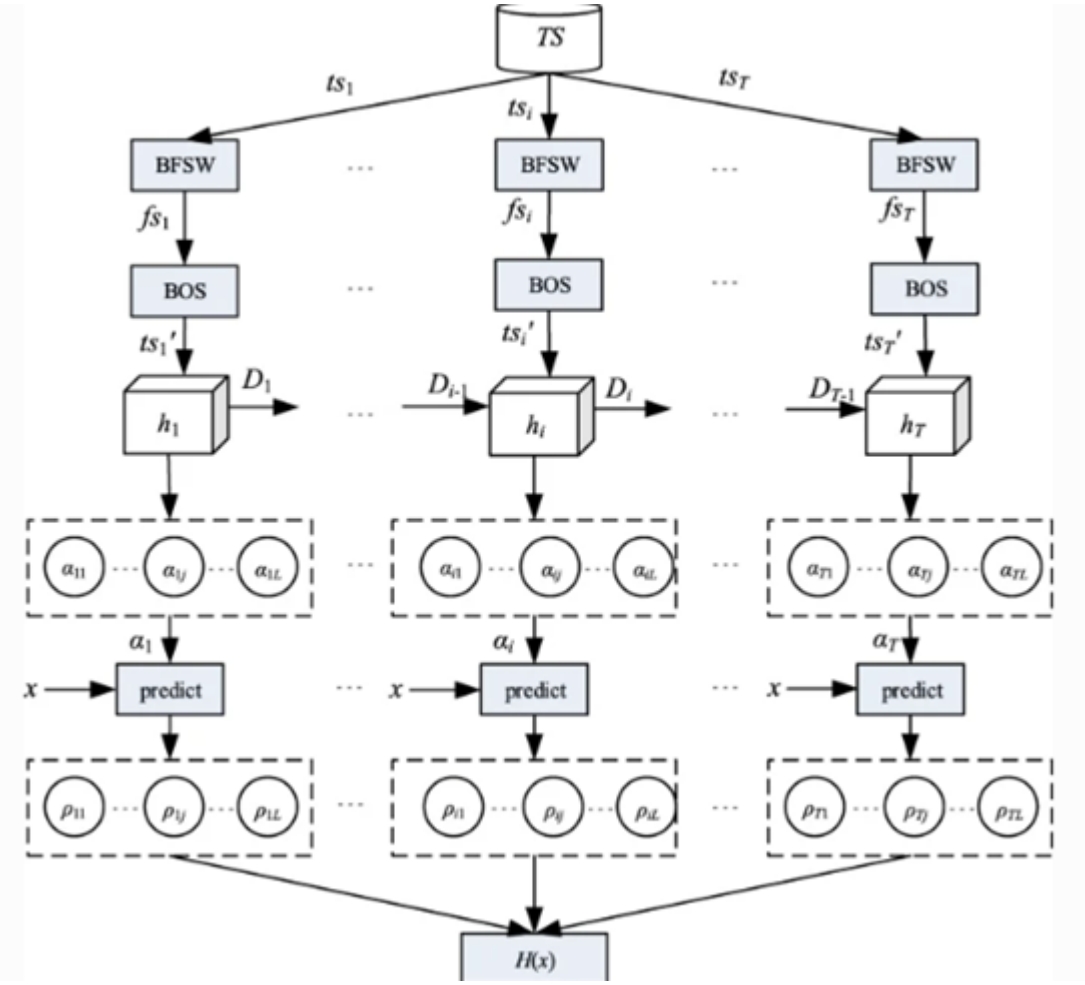


Figure 11: A Schematic Review of AdaBoost Mechanism

Consider an example of a dummy dataset that has a binary classification; sample is diseased or not. The decision in this data is to be made according to 3 attributes which are gender, age, and income. AdaBoost algorithm is composed of 7 stages, which are illustrated as following:

Stage 1: Each and every sample of our training dataset is to be assigned a weight value. Initially, all samples get an equal value. Given that N represents the number of samples, the initial weights are to be calculated using the following formula (Shrestha & Solomatine, 2006).

$$w(x_i, y_i) = \frac{1}{N}, i = 1,2, \dots n \qquad (4)$$

Stage 2: The second stage is measuring the classification dependencies on each attribute, meaning that how much each attribute contributes to the classification process. To proceed with this goal, a decision stump is to be built for each attribute, followed by the Gini index of each decision stump. The lower Gini index indicates better classification, so its corresponding decision stump will be the first.

Stage 3: It deals with measuring how accurately the model is built in classifying the samples using the total error that represents all the weights of the misclassified samples (Shrestha & Solomatine, 2006). The total error is to be integrated in the following formula that indicates the importance of the built decision stumps.

$$performanceofhestump = \frac{1}{2} log_e \,) \qquad (5)$$

Given that our weight is a fraction, the total error will always be between 0, perfect stump, and 1, bad stump.

Stage 4: as discussed before, the weights of each point is to be updated according to the classification accuracy. The wrongly classified points are to be given higher weight. The weights are to be updated using the following formula:

$$Newsampleweight = oldweight * e^{\pm Amountofsay(\propto)} \qquad (6)$$

where, alpha represents the performance of the model that was calculated in stage 2.

In **stage 5**, the data is to be modified according to the updated weights. The updated weights column is to be used to divide the data points into buckets. In stage 6, the dataset that will be used in the next model is to be created out of the original dataset with a higher probability and existence rate of the samples with higher weights.

The seventh and last stage of AdaBoost algorism deals with repeating all the above steps with the new formulated dataset in **stage 6.** Starting from assigning equal weights to the new dataset, followed by finding the best stump, then calculating the total error, and finally updating the weights and dataset, until a predefined error acceptance rate is reached.

AdaBoost is a commonly used algorithm of projects related to customer acceptance and satisfaction. One of these research projects is the project published by Wu et al. (2022). They combined the AdaBoost algorithm with principal component analysis (PCA) for e-commerce customer churn prediction. Customer churn is other face of customer satisfaction. Both affect the organization ‘revenues. To improve customer satisfaction, and hence customer retention, the attributes that participate in customer retention need to be identified. That is why projects that are working on customer churn are closely related to our problem. Zengyuan Wu’s project deals with e-commerce which causes their data to be high-dimensional and unbalanced. That is why they specifically integrated data pre-processing and ensemble learning. They used PCA to reduce the data dimensionality and used AdaBoost to minimize the effect of unbalanced data by cascading multiple decision trees. Their proposed model, PCA-AdaBoost model, achieved higher accuracy than all the evaluated models in literature; SVM, Logistic Regression, and the typical AdaBoost. The results they achieved are demonstrated in Table 1 below.

Table 1: The Performance of Different Models Against PCA-AdaBoost

Methods	**Overall accuracy**	**G-mean**	**Recall**	**Precision**
Logistic regression	0.9788	0.9831	0.9771	0.9793
SVM	0.6751	0.6335	0.8912	0.6197
AdaBoost	0.8737	0.9714	0.9766	0.9669
PCA-AdaBoost	0.9898	.9897	0.9917	0.9880

Their proposed PCA-AdaBoost model achieved higher accuracy than all other models. However, the typical AdaBoost achieved considerably close accuracies. As our dataset is not high dimensional, it has only 37 attributes, and we need to implement a model that is interpretable, so it was chosen to evaluate AdaBoost among our tested models.

Another research project conducted by Sabbeh (2018) proved that AdaBoost along with random forest outperform many other evaluated machine learning techniques, such as Decision Trees (DT), Discriminant Analysis, Naïve Bayesian, Support Vector Machines, Multi-layer perceptron, instance-based learning (k-nearest neighbors), and Logistic Regression. Sarah was working on customer retention. The dataset she used has more than 3000 samples, making her results quite reliable. Both ensemble learning techniques that were used, AdaBoost and RF, achieved almost the same accuracy which is 96% in comparison to the other models that achieved 94%, 90%, 88%, and finally 86.7% accuracy.

2.3.5 XGB Regression Random Tree

XGB is a highly successful regression technique for the development of controlled models that may be found in many applications (Sahin, 2020). It is possible that knowledge of its goal function (XGB), in addition to the basic learners, will aid in

determining the veracity of this claim. In the purpose function, there is a loss function as well as a regularization term that must be considered. The distance between the actual values and the model's predictions is shown by this parameter, which is also known as the gap between the observed and expected values. The reg: linear and reg: logistics functions are the most often encountered sources of XGB regression problems (Jangaraj et al., 2021).

Numerous systems use XGB regression, which is a successful regression technique for the development of management models that is particularly well-suited for this purpose. It is possible that those with a prior understanding of its principal function (XGB), as well as others who are just getting started, will be able to assist in determining the validity of this claim. While considering the motive feature, it is necessary to take into account both the loss characteristic and the regularization term (Yi et al., 2019). This factor is considered as a space among determined or anticipated figures is used to demonstrate the disparity between the actual values and the predictions made with the version. The most frequently occurring resources in XGB regression situations are the linear: reg and logistics: reg capabilities.

XGB is a gradient boosting machine learning algorithm that stands for Extreme Gradient Boosting. It is not only used for regression, classification purposes, but also for ranking problems (Li & Zhang, 2019). XGB is a decision tree ensemble learning algorithm that depends on many ML models in taking its final decision. Ensemble learning models use multiple algorithms, each of them makes its own judgment, and then voting is to be conducted to reach the final decision (Sagi & Rokach, 2018).

Evolutionally wise, XGB is a descendant of the great ancestor, DT. The species starts with DT, then Bagging is introduced. Passing by random forest, boosting and

gradient boosting, finally reached XGB (Sahin, 2020). XGB uses multiple DTs that are built in parallel not sequentially, such as the Gradient Boosting Decision Trees algorithm. Indeed, XGB is another implementation of gradient boosting, but with some upgrades at the level of both the algorithm and system. The enhancements that are related to the system are parallelization, tree pruning (Luiz de Freitas Vieira & Almeida Có, 1997), and hardware optimization, while the algorithmic optimization points are regularization, sparsity awareness (Nguyen et al., 2020), weighted quantile sketch (Dong et al., 2020), and cross-validation.

As mentioned earlier, XGB depends on the parallel constructed DTs, resulting into improvements in the algorithm performance. The second improvement is tree pruning. XGB does not depend on greedy approach in the stopping criterion. On the contrary, XGB uses the max depth approach, and then implements backward pruning which is described as depth first approach. This approach considerably improves the computational power. When it comes to the algorithmic enhancements, XGB has a special add point in avoiding overfitting using both LASSO (L1) and Ridge (L2) regularization. In addition, it uses both Shrinkage and Column Subsampling (Dong et al., 2020) to avoid over-fitting. Shrinkage resembles stochastic optimization that reduces the effect of each DT to grasp the attention of the newly formed DTs.

XGB outperformed many ML models in a research project conducted by Hota and Dash (2021) that focused on prediction of customer churn in telecom industry. They investigated multiple machine learning models in customer churn prediction. They compared the predictive capability of each of XGB, GradientBoost, AdaBoost, ANN, Logistic Regression, and Random Forest models. The dataset they worked on was large, consisting of 7043 samples. Each sample has 21 features that are correlated

with customer churn. Their features of interest are gender, age, dependents, services they have signed up for, contract information, payment methods, paperless billing, and monthly charges. Their results strongly direct us toward using XGB. Table 2 shows the accuracy reached by each of the tested machine learning models along with recall and precision values of each of them. XGB outperforms all the investigated 6 models.

Table 2: Models Analysis

ML Models	Accuracy	Precision	Recall	F1-Score
GradientBoost	80.41%	0.66	0.58	0.59
AdaBoost	80.59%	0.67	0.57	0.58
XGBoost	82.20%	0.66	0.45	0.56
ANN	79.98%	0.89	0.84	0.85
LR	80.29%	0.68	0.56	0.60
RF	81.10%	0.66	0.49	0.56

One of other research projects that used XGB in customer churn prediction was the research conducted by Abdelrahim Kasem Ahmad, Assef Jafar and Kadan Aljoumaa (Ahmad et al., 2019). They aimed to assist telecom companies in figuring out the factors that should be reduced or completely eliminated to avoid the churn of customers using machine learning approaches. They used Gradient Boosted Machine Tree “GBM”, Random Forest, Extreme Gradient Boosting “XGB”, and Decision Tree. The best accuracy was detected by using XGB tree model which achieved 93.301% accuracy. However, GBM achieved 90.89% accuracy, which is the second-best accuracy. The accuracies reached by the four algorithms are illustrated in Figure 12.

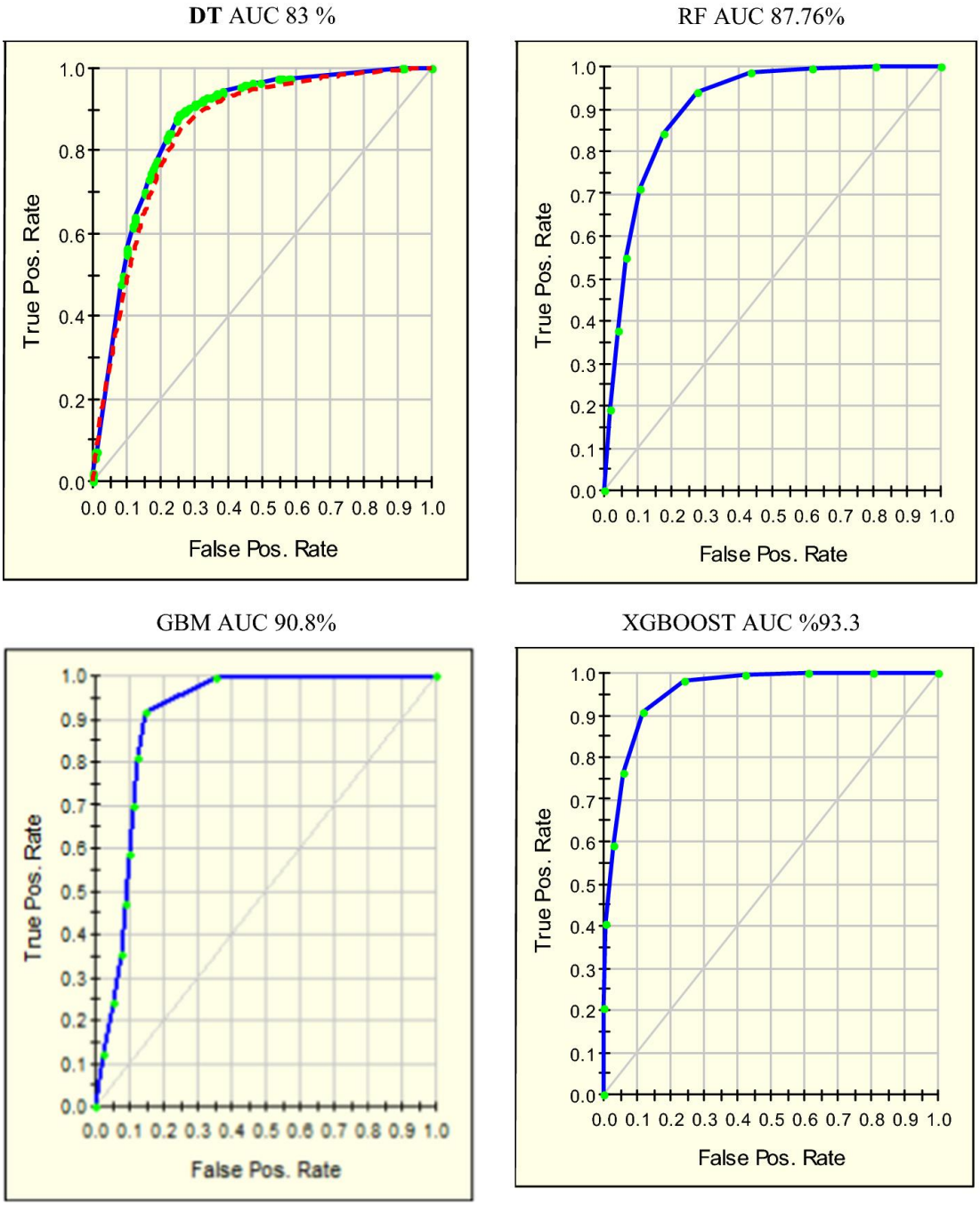


Figure 12: Accuracy Detected in Telecom Customer Churn for the Four Models.

Moreover, AL-Shatnwai and Faris (2020) used XGB as a ML model for customer retention in telecommunication sector (AL-Shatnwai & Faris, 2020). The dataset used was churn dataset. They evaluated multiple machine learning models;

RF, SVM, Logistic Regression, SCD, and XGB. XGB achieved the best accuracy as illustrated in Table 3.

Table 3: Detected Accuracies of All Investigated Machine Learning Models.

	Accuracy	Precision	Recall	F1 measure
RandomForest	0.955 (0.008)	0.936 (0.044)	0.743 (0.053)	0.827 (0.033)
SVM	0.827 (0.021)	0.351 (0.065)	0.225 (0.056)	0.272 (0.058)
XGboost	0.956 (0.009)	0.924 (0.052)	0.752 (0.057)	0.829 (0.052)
LogisticRegression	0.864 (0.016)	0.618 (0.154)	0.204 (0.042)	0.302 (0.052)
SGD	0.801 (0.193)	0.552 (0.208)	0.225 (0.275)	0.223 (0.120)

According to the analysis and above discussion for the prediction of customer satisfaction, decision tree regression is the best and most appropriate method as it provides the precise, accurate value of different attributes of customer satisfaction. Moreover, the XG5 boost regression tree is also similar to the decision tree that gives brief information about valuation and features that help improve customer satisfaction.

2.3.6 M5P

According to Quinlan (1986) the M5P tree is a decision tree learning that can be used to solve regression issues. The M5P tree approach applies linear function regression to the nodes terminal while fitting the linear, multivariate regression model to every domain through categorizing or splitting the total records area into several sub bands using the classification or division technique respectively (Arosha Senanayake & Joshi, 2021). The M5 tree approach, as opposed to discrete classes, is more appropriate for dealing with continuous elegance problems, and can handle commitments with extremely high dimensionality. In the statistics set, it is well-known for its piecewise recordings of each linear version, which are used to approximate nonlinear relationships in the statistics set. In M5P, three kinds of tree branches are

formed that are known as leaves, internal, and root nodes. These nodes are internally connected with each other with the help of branches.

In order to construct the tree, a selection-tree implementation plan is used. However, rather than maximizing the benefits of the records at every access point, dividing criteria are employed, which minimizes the inter variation in the elegance values down to each intermediate node (Arosha Senanayake & Joshi, 2021).

In the preprocessing steps for M5P model, two main stages are completed. First, binarization is applied to all enumerated attributes so that all node splits are binary. The second stage is considering handling the missing values. Usually, in these cases, the instances that have a missing value of one of its attributes are to be deleted. Another approach that is mostly followed when there is no luxury of deleting samples due to the small size of data is to impute the missing values using one of the different proposed techniques, such as imputing the average or most frequent value. In the case of M5P, a different technique is applied, which is called surrogate splitting.

When a missing value for the specific attribute-based split is found, surrogate splitting during training stage searches for the most correlated attribute and uses its corresponding value only for this sample. The alternative attribute is the most correlated attribute to the one that originally should be in use. That is how all missing values are imputed during training stage. However, the average value imputation technique is used during the testing phase. The missing value for a specific attribute is replaced by the average value of the training instances for that same attribute.

M5P was one of five machine learning models chosen by Geler et al. (2021) in their research about customers' assessments in food serving businesses (Geler et al., 2021). They compared the prediction capability of six different machine learning

models, namely SMO, RandF, RandT, REPT, M5P, and MP, in predicting customer satisfaction of restaurant and food services. In addition, they worked on finding the important attributes that contribute to customer satisfaction. They found that food taste, service, and environment are the most important three features that affect customer satisfaction. The four tested machine learning models did not show great difference in their results, instead they showed great similarity. Random Forest and MP models' range difference was between 0.12 and 0.23, while the other models showed difference not more than 0.09. Figure 13, shows the exact results of each of the six models regarding the three important features (Geler et al., 2021).

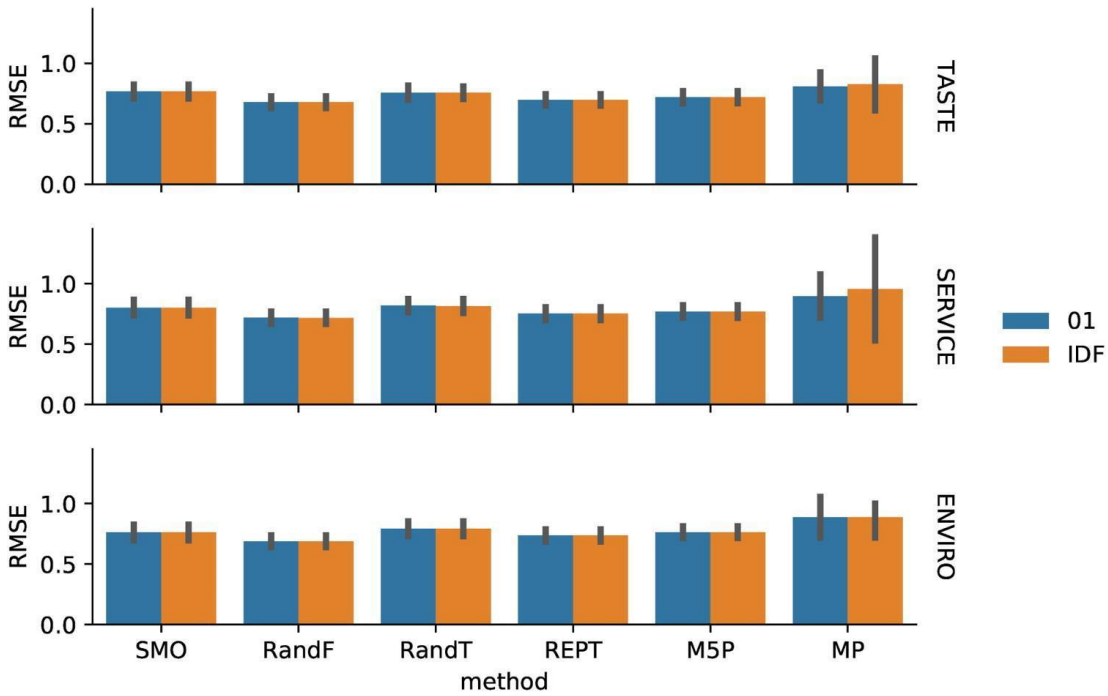


Figure 13: Graphical Representation of the Average Values and Standard Deviations of RMSE

2.3.7 Random Tree

Random tree is decision tree that does not search for all attributes on each split, rather it creates a random subset of attributes at each split. All tree-based models have

the advantage of high visualization ability and interpretability. Random tree is used in classification problems as well as regression problems. In training phase, the records are recursively split into groups with similar output field values. To generate sample data for tree model building, bootstrap is used with replacement. Moreover, random tree is a binary tree, which means that it creates a binary split for two sub-trees at each node. As our data is categorical with multiple class, each branch does not necessarily have only one category but a group of categories. The random tree is usually very large because there is no pruning in random tree, which means that it goes to the largest possible branching extension (Ullah et al., 2019).

Random Tree is used in customer satisfaction and customer churn problems. The research project conducted by Ullah et al. (2019) is focused on customer churn in telecom industry (Ullah et al., 2019). They were concerned with the reasons of customers churn and their behavior patterns. They used multiple machine learning models for classification predictions, such as AdaBoostM1 + Decision Stump, J48, Decision Stump, Random Forest (RF), Random Tree (RT), and Bagging + Random Tree Logistic Regression (LR). The best performing models were Random Forest and J48 with 88.63% accuracy, while Random Tree was ranked second with an accuracy of 84.34%.

2.3.8 Reduced Error Pruning Tree (REPTree)

REPTree is a decision tree with improvements for pruning stage that can work on classification as well as regression problems (Al Snousy et al., 2011). It depends on information gain / variance in the construction and Reduced Error Pruning (REP) for pruning (Elomaa & Kaariainen, 2001). In the reduced error pruning, complete sub-trees are to be pruned by replacing them with only one node.

REPTree achieved an accuracy of 98.39% in the research conducted by Al Snousy et al. (2011). Their research was concerned with microarray analysis for cancer diagnosis problem. They compared nine decision tree-based algorithms; Decision Stump, C4.5, Random Tree and REPTree, CART, Random Forests, AdaBoost (C4.5 and REPTree), Bagging (C4.5 and REPTree), and alternating decision ADTree.

Another research conducted by Boodhun and Jayabalan (2018) used REPTree. They were concerned about enhancing the risk assessment for life insurance companies. They used Correlation-Based Feature Selection and Principal Components Analysis for dimensionality reduction, Random Tree classifiers, REPTree, Artificial Neural Network, and Multiple Linear Regression as machine learning models. REPTree showed the highest prediction accuracy among all used models with the lowest mean absolute error (MAE) value of 1.5285.

2.3.9 Deep Learning

Deep learning is a concept of machine learning based on artificial neural networks. This makes it possible to manage unstructured data, including text, images, and documents. Deep learning models outperform shallow machine learning models and conventional data analysis techniques in many situations (Janiesch et al., 2021). According to Bailly et al. (2022) deep-learning models were able to perform well even without interaction terms, while machine-learning models were less affected by the dataset size and needed interaction terms to perform well. In summary, well-specified machine learning models outperformed deep learning models in the scenarios that were considered (Bailly et al., 2022). The most basic type of deep neural network is the multilayer perceptron (MLP). Multiple hidden layers make up the architecture of an MLP in order to capture more intricate associations seen in the training dataset. The

MLP is also known as a deep feedforward neural network (DFN) (Bisong, 2019). CNNs is one of the deep learning methods which are typically utilized to tackle challenging image-driven pattern recognition applications. (O'Shea & Nash, 2015).

The study conducted by Alnagar (2020) investigates the determinants of student satisfaction with e-learning and proposes a model to identify the factors that influence student satisfaction at Tabuk university using MLP artificial neural networks. (Alnagar, 2020). Also, a study done by Tariq et al. (2021) proposed predicting churned users through CNN. This paper aims to monitor customer behavior and make decisions accordingly.

2.4 Feature Selection

In machine learning, attribute selection has been perceived to be a preferred technique for selecting a subset of relevant features from high-dimensional data. According to a study, the Feature Selection Model is essential for analyzing the variability and how common the product is amongst other products in an organization’s portfolio. It proposes incorporating customer preference information into the model using sentiment analysis of user-generated product reviews (Adjimi et al., 2019).

Different feature selection methods have been used to discover the most important attributes among all the attributes of various brand measures. Principle Component Analysis (PCA), Correlation-based Feature Subset Selection, and Relief method have been discussed as attribute selection methods (Amir, 2017). Furthermore, feature selection algorithms such as Exploratory Factor Analysis (EFA) (Zeinalizadeh et al., 2015), feature-based transfer learning strategy, TFS supervised forward feature selection (SFFS), and Filter–Wrapper (Idris & Khan, 2017) were used.

In addition to this, balanced iterative reducing and clustering using hierarchies (BIRCH) have been used for customer segmentation (Hassan, 2018). K-means algorithm clustering was based on the loyalty level (Chou et al., 2011). Different feature selection techniques in text categorization have been discussed, like Information Gain (IG), Chi-Square (CHI), Correlation Coefficient (CC), and Odds Ratio (OR) (Zheng et al., 2004). To compare different feature selection techniques, different performance metrics like the number of features selected, a list of features, Classifier accuracy, and elapsed time can be used (Sheena et al., 2016). Feature selection could improve the performance of the prediction algorithms and reduce the memory storage requirements and computation time, which could reduce the computational costs for data analytics.

As mentioned before, the Kano model can categorize attributes into five different categories, which make the Kano model very popular models over the last 3 decades; thus, different approaches had been applied to explore asymmetric and non-linear relationships in the Kano model studies. A study conducted by Chang et al. (2009) specified that various methods have been used to classify quality attributes into five Kano categories like Penalty-Reward Contrast Analysis (PRCA), Importance Grid Analysis (IGA), direct classification method, and the moderated regression analysis. The study concluded that the Kano questionnaire remains the most appropriate classification method to identify Kano despite the fact that it is very complicated and not easy to be implemented (Chang et al., 2009).

2.4.1 Features Selection Types Techniques

Feature selection is a data mining tool that aims to select the most descriptive features for the target variable. In this process, a compact representation of the data is found. A small subset of the features might contain most of the information about the

data (Liu & Motoda, 2012). Most of the time, feature selection is used to reduce computations, but in our case, it will be used to select the top features that affect customer satisfaction in a business.

Informing the business providers about the most important features that directly affect customers is a huge gain. Due to the importance of feature selection methods in this research, a brief explanation of feature selection types will be presented, and the selected approaches will be highlighted.

2.4.1.1 Filter Techniques

In filtering method, the best features are chosen based on the correlation coefficient without the use of any machine learning techniques. Only statistical measures are used to determine the best features. Afterward, a machine learning model is applied to get the performance of those selected features (Pavya & Srinivasan, 2018).

Drawback of this method is the ignorance of in-between feature relations. In other words, the filter method gets only the correlation between each individual variable and the target variable. This might miss valuable information because some features are highly informative being together, but each one of them has much less significant information on its own (Wang et al., 2014). Some examples of filter methods are shown in Figure 14 (Suppers et al., 2018).

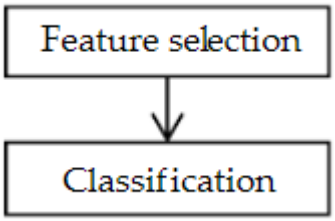


Figure 14: Filter Method in Feature Selection

2.4.1.2 Wrapper Techniques

Wrapper technique can get the best set of features instead of only important or relevant features. By training the model and observing the effect of adding or removing features, this technique can decide the most important group of features. This process might be computationally expensive, but it gets the optimal set of features. Nothing prevents getting the best set of features using relevant features only (Jović et al., 2015). Figure 15, shows the wrapper approach in feature selection. The wrapper method is divided into 3 categories (Suppers et al., 2018).

- **Forward Feature Selection**: This type starts with an empty set of features, then adds features one by one, and observes its effect on the model's final accuracy. After that, the best features are selected by continuously adding features and observing them (Jović et al., 2015).
- **Backward Feature Selection:** This type is exactly the inverse of the FFS. The model starts with all sets of features and eliminates the least important as it goes. The least important feature is the one that has the smallest effect on model accuracy after removal (Pavya & Srinivasan, 2018).
- **Heuristic Feature Elimination:** Similar to BFS, it recursively removes irrelevant features until ending up with the best set of features (Pavya & Srinivasan, 2018).

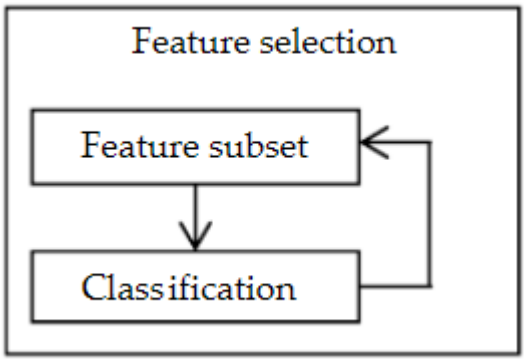


Figure 15: Wrapper Approach in Feature Selection

2.4.1.3 Embedding Technique

In the embedding method, the feature selection method is embedded into the machine learning algorithm and is optimized (Pavya & Srinivasan, 2018). Examples include Lasso L1 and ridge regression L2, in which a penalty is added for large coefficients. More details will be given in the description of each feature selection method. Figure 16, shows the embedding technique in feature selection (Suppers et al., 2018).

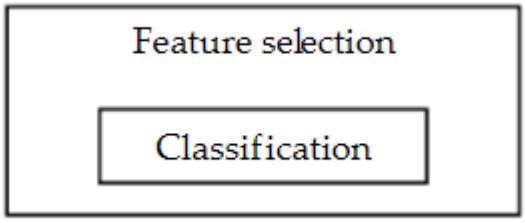


Figure 16: Embedding Technique in Feature Selection

In the following subsections, a group of feature selection methods will be reviewed. Some of them have been used in this research, and justification of the reasons behind using these methods will be provided.

2.4.2 Features Selection Techniques

2.4.2.1 Chi-Square test

This test examines the independency between two variables. Two variables are fully independent when the probability of both taking place at the same time is equal to the multiplication of each probability of occurrence:

$$P(XY) = P(X)P(Y) \qquad (7)$$

Particularly, it tests the correlation between feature values and the predicted classes. It can't only address the significance of the observed differences, but also provide detailed information about exactly which categories are responsible for the differences found (McHugh, 2013).

2.4.2.2 Mutual Information (MI)

In this feature selection technique, relevant features contain large information about the target class. This might be similar to correlation, but the difference is mutual information measure. The redundancy inside a random variable X is not just its correlation with the target. According to information theory, the amount of redundancy uncertainty inside a random variable is another representation of how much information this variable has (Sulistiani et al., 2019). According to Shannon, the amount of uncertainty inside random variable X can be measured by using entropy function $H(X)$, which is defined as:

$$H(X) = -\sum_{x \in X} \quad p(x) log \ p(x) \qquad (8)$$

Where, $p(x)$ is the marginal probability of x, which is the probability of event x to happen. Joint probability is the probability of two x and y events to happen at the same time $p(x, y)$. From joint probability, joint entropy appears. Joint entropy is a measurement of the uncertainty related to two variables. To measure joint entropy between two variables, the following formula is used:

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) log\ p(x, y) \quad (9)$$

Mutual information is the amount of information that both variables share (Sulistiani et al., 2019). The equation for calculating the mutual information is given below:

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) log \frac{p(x,y)}{p(x)p(y)} \quad (10)$$

For two random variables X and Y, if $I(X; Y) > 0$, the two variables contain some mutual information. If $I(X; Y) \leq 0$, the two variables have no relation. Mutual information can be used as a metric to state how variable/ Feature X is descriptive of variable/ label Y. Mutual information diagram is illustrated Figure 17 (Li et al., 2009).

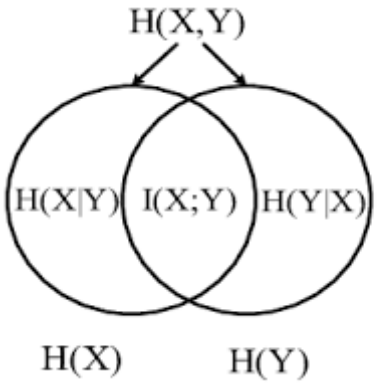


Figure 17: Mutual information diagram

In Sulistiani et al. (2019) mutual information (MI) with support vector machine had been used to build a classification model for customer loyalty. Fast-moving consumer goods (FMCG) is an important sector in business. This model tries to classify customers in FMCG into loyal or non-loyal customers using features selected from MI and SVM models. The method starts with data cleaning. Afterward, the feature selection method gets the most predicting 5 features out of 26 features. After selecting the features, the result of prediction using all features and the top 5 features is compared. Figure 18 illustrates the research methodology (Sulistiani et al., 2019).

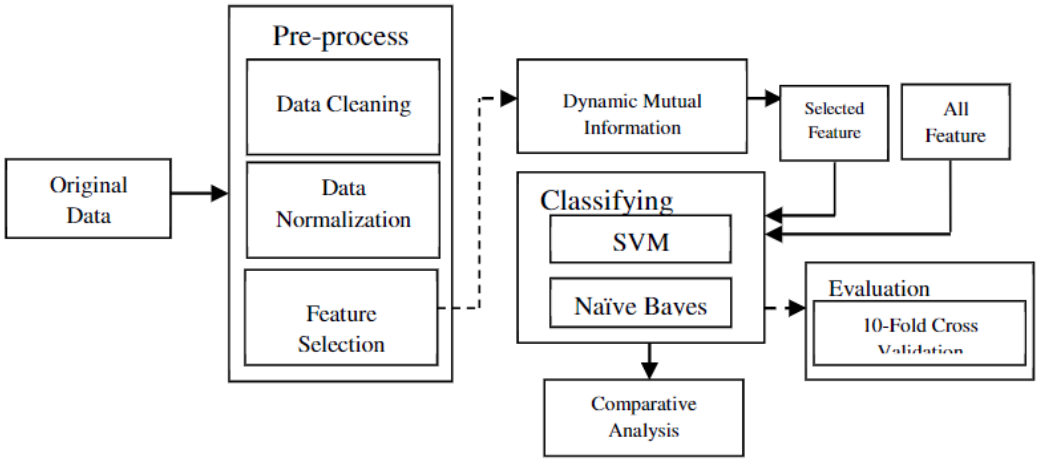


Figure 18: Research Methodology

The model gives 73.57% correct classification accuracy by using the best four features as compared to 76.42% correct classification accuracy by using all features. The researchers proved the effectiveness of using the DMI-SVM model for selecting the best features affecting customer loyalty.

2.4.2.3 Least Absolute Shrinkage and Selection Operator (Lasso)

The least Absolute Shrinkage and Selection Operator (Lasso), is a powerful method in feature selection and regularization. This method belongs to the embedded feature selection family because the feature selection module and the machine learning

evaluating model are present together. For a better understanding of Lasso, starting with a linear model is better. Linear model is the simplest form of prediction, wherein it is assumed that the target output is a linear combination of the input features as given in the following equation:

$$y_{pred} = a_0 + a_1 x_1 + \cdots + a_n x_n \qquad (11)$$

The best regression model minimizes a cost function using certain values of a_i. For simple regression model, the mean square error is used as a cost function:

$$\sum_{i=1}^{N_{\text{training}}} \left(y_{\text{real}}^{(i)} - y_{\text{pred}}^{(i)} \right)^2 \qquad (12)$$

One of the biggest problems in this simple model is the collinearity between features. If two features are correlated, the overall model variance increases. When model variance increases, the ability to generalize other data decreases significantly, and the process of feature selection losses its overall significance.

The lasso method, solves this problem by adding the penalty term L1 to the cost function. The idea behind the new term is to penalize and shrink the useless features' coefficients. Lasso performs a sort of automatic feature selection. If two features are highly correlated, they will increase the value of the cost function. Lasso penalizes the coefficient of one of them to make the important feature survive. Lasso shrinks the coefficient of the feature to 0 to eliminate the least important feature (Fonti & Belitser, 2017).

$$\frac{1}{2N_{\text{training}}} \sum_{i=1}^{N_{\text{training}}} \left(y_{\text{real}}^{(i)} - y_{\text{pred}}^{(i)} \right)^2 + \alpha \sum_{j=1}^{n} |a_j| \qquad (13)$$

LASSO method has some advantages that justify its usage in the feature selection process.

1. Features resulting from Lasso have good prediction accuracy because the process of shrinking and removing the coefficients reduces the variance without a large change in the bias. This is significantly useful when having a large number of features but a relatively small number of observations.
2. Lasso eliminates irrelevant variables that have small interpretation of the target variable.

Researchers in Wu et al. (2022) tried to determine the factors that affect customer satisfaction in online travel agencies during the pandemic. The lockdown affected these types of companies, and the competition for customer satisfaction was crucial. Using online surveys and Lasso for feature selection, researchers concluded that refund, promptness, and easiness are the top factors affecting customer satisfaction for OTAs.

2.4.2.4 Analysis of Variance (ANOVA)

ANOVA – (Analysis of variance) – is a statistical method to analyze the difference among means of several variables. ANOVA is a very powerful method in feature selection because it analyzes the relation between feature variance and predictor variance. Usually, ANOVA is used to select the best features from categorical features that predict continuous variables.

The algorithm measures the ratio between the variance between each group and the variance within the group. As the variance between the groups increases and variance within the group decreases, the prediction of the target variable is affected. This vague explanation will be clarified by the following example.

Assume, that a school wants to know if a guardian type affects the student's grade. Table 4, represents the student grade in 3 cases. If the variance between the case of a mother as a guardian and a father as a guardian increases, this is evidence that the type of guardian affects the grade of a student. On the other hand, if the variance within each group is small, it indicates that the type of guardian restricts the student's grade within a small range.

Table 4: Student's Grade in Different Types of Guardians

Father	Mother	Other
8	5	3
5	15	7
12	20	5
16	23	6
13	19	3
14	17	10
20	13	12
10	7	8
3	5	9

The previous two notes can be concluded as the following. If the ratio between groups' variance and within group variance increases, this feature affects the prediction of the target variable.

$$F = \frac{Variancebetweengroups}{Variancewithingroups} \qquad (14)$$

F-test score is used in ANOVA as a description of this ratio. Sum of square differences is a statistical measurement that describes the variance in a certain variable. The following equation describes the basic form of sum of squares.

$$\sum of\ squares = \sum_{i=0} \quad \left(X_i - \overline{X}\right)^2 \qquad (15)$$

$$X_i\ is\ the\ item\ number\ i \in theset$$

$$\underline{X}\ is\ the\ mean\ of\ all\ items \in theset$$

To get the F-Test ratio, two values are calculated, SSB and SSW. SSB is the sum of squares between all groups, and SSW is the sum of squares within each group.

$$SSB = \sum\left(g_i - \underline{X}\right)^2 \quad (16)$$

$$SSW = \sum(x_i - g)^2 \quad (17)$$

In the previous two equations, $\underline{X}$ is the grand mean which is the average of all average values between all groups. g_i is the average value for group number i. As the ratio SSB/SSW increases, the feature gives more information about the target variable.

In Jahanshahi et al. (2011) researchers addressed an important question regarding the automotive industry in India. They examined the relationship between customer service, product quality, customer satisfaction, and customer loyalty. The data was collected at different stages; for example, measuring customer satisfaction and loyalty at the beginning and after years of the buying process. Using ANOVA and regression analysis, the research concluded that there is a strong correlation between product quality and customer service level with customer satisfaction and loyalty.

2.4.2.5 Correlation

Correlation is a term used in statistics that represents a measurement of how variables are related/correlated to each other. A high positive correlation value between two variables means that when one variable increases, the other variable increases, for example $x = 3y$. High negative correlation means that if one variable increases, the other variable will decrease and vice versa. Having a small value of correlation between two variables implies that if one variable increases, there is no information available for the other variable (Doshi & Chaturvedi, 2014).

The previous discussion of correlation is the core of CFS. If two variables are highly correlated, then each variable is highly predictive of the other. CFS measures the efficacy of individual features in predicting the target. Choosing the features by using heuristics that filter the redundant and irrelevant features results in the least significant prediction results (Pavya & Srinivasan, 2018).

The equation used to rate the feature is given below:

$$F_s = \frac{N * r_a}{N + N(N-1) r_n} \qquad (18)$$

2.4.2.6 Pearson Correlation

Pearson's correlation is a type of similarity measure similar to normal correlation. Pearson's correlation is defined as the ratio between covariance and the standard deviation for two sets of data.

$$\rho_{X,Y} = \frac{cov\,(X,Y)}{\sigma_X \sigma_Y} \qquad (19)$$

Covariance describes the variability between two variables. Having high covariance value of two variables means that if one variable changes with large value, the other variable would change with significant value too.

For two variables $X \wedge Y$, r coefficient can be calculated as following:

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i-1}^{n} (y_i - \bar{y})^2}} \qquad (20)$$

Where n is the number of samples, and $\underline{x}$ and $\underline{y}$ represent the mean value for the two variables $X \wedge Y$. The larger the r coefficient, the more correlated the two variables.

Pearson correlation is an acceptable method for feature selection, but it has some disadvantages. For example, it can only describe linear relations between the two variables. It cannot handle complex non-linear relations. As with normal correlation, we cannot tell the difference between correlation and causation.

As shown in Figure 19, it can be noticed that there is a high correlation between Ice cream consumption and drowning. At the first glance, this could be understood mistakenly as If ice cream causes drowning. For sure, this is not correct. A famous concept in statistics highlights that “correlation does not imply causation”. The hidden independent variable in this case, is the summer season. Both the ice cream consumption and drowning cases increase in summer, and both are dependent variables.

It can be concluded that the researchers must have a good knowledge of the processed data to be able to find a valuable results from Pearson correlation.

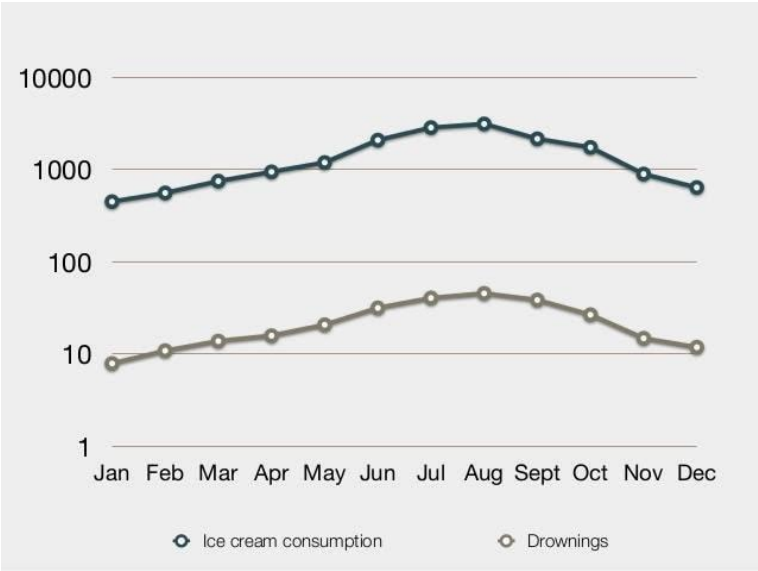


Figure 19: Relation Between Ice Cream and Drowning

.

2.4.2.7 Fisher's Score

Fisher score belongs to the filter method family in feature selection. The method gives a score to each feature based on certain algorithms and then selects the top-m features from a set of n features. The basic idea of the algorithm is to select features that span most of the data space. This could be achieved by selecting features with a distance between points in different classes as large as possible. The other criteria is to minimize the distance between points from the same class.

The Fisher's score is calculated by the following equation:

$$F(x^j) = \frac{\sum_{k=1}^{c} n_k \left(\mu_k^j - \mu^j\right)^2}{\left(\sigma^j\right)^2} \qquad (21)$$

Where μ_k^j corresponds to the mean of the $k-$thclass in the $j-$th features. $\sigma^j \wedge \mu^j$ represent the standard deviation and the mean for the $j-$th feature in the whole dataset (Gu et al., 2012).

After computing the Fisher's score for all the features, the $top-m$ features are selected. This algorithm might be computationally efficient but not optimal. The algorithm is suboptimal because it ignores the relation between features. Ignoring the relation between features will result in two problems. The first problem is the redundancy in information. If two features are highly correlated, but both of them describe the variability in the target variable, Fisher's score will include both of them in the selected $top-m$ features. The second problem is rejecting all features with small scores regardless of the effect of the relation between them. For features $a \wedge b$, each feature might have small score, but the combination of ab is highly valuable and descriptive.

An improved Fisher's score method was introduced in (Gu et al., 2012). This research proposed a generalized form of Fisher's score to solve the previously discussed two problems. Using linear programming, researchers could outperform the classical fisher score. It is a state-of-the-art method for feature selection on several benchmarks.

2.4.2.8 Variance Threshold

The amount of information the feature might contain is highly related to the variance of the feature. Variance is a statistical measure of spreading or dispersion in a set or a group of data. In the variance threshold method, the variance of each feature is measured, and features with a variance less than a certain value are removed. This value is called the variance threshold. Features with no zero-variance are removed as a baseline of the algorithm. The variance threshold can be considered as a feature eliminator rather than a feature selector (Ferreira et al., 2012).

The drawback of this method is that it ignores the relationship between the features and target variables. The selection/elimination of features depends only on the variance of each feature regardless of the relation between this feature and the target. For this reason, the variance threshold method is considered a preprocessing technique rather than a feature selection technique. For a large number of features, this method is used to remove all features with no or small variance.

The variance is calculated using the following formula:

$$\mathrm{var}\,(X_i) = \frac{1}{n}\sum_{j=1}^{n} \left(X_{ij} - \overline{X}_i\right)^2 \qquad (22)$$

Where, X_i is feature number i, and X_{ij} is instance j in this feature. $\underline{X}_i$ is the mean value for all instances in the feature.

In (Fida et al., 2021) researchers examined variance threshold as a feature selector for intrusion detection systems (IDS). IDS detects malicious attacks and separates them from normal attacks. It is a classification model for either malicious or normal traffic types. Due to the large number of features, the variance threshold is a suitable feature selection method to eliminate features that affect model performance. Combined with random forest, researchers achieved 76% classification accuracy.

2.4.2.9 Mean Absolute Difference (MAD)

Similar to the variance threshold, mean absolute difference method belongs to the dispersion measure feature selection family. In such types of feature selection methods, the feature selected is based on the amount of dispersion contained in it. This dispersion can be measured/interpreted in various terms, e.g., variance or mean absolute difference.

In MAD, the dispersion is measured based on the sum of the differences between all feature instances and the mean value. Similar to the variance threshold, if this value is larger than a certain threshold, then this feature is accepted.

MAD is calculated by using the following formula:

$$\mathrm{MAD_i} = \frac{1}{n}\sum_{j=1}^{n} \left|X_{ij} - \overline{X}_i\right| \qquad (23)$$

X_i is feature number i

X_{ij} is instance j in this feature.

$\underline{X}_i$ is the mean value for all instances in the feature.

n is the number of features.

The division by n is to normalize the value of the summation.

2.4.2.10 Dispersion Ratio

Another dispersion measure is the dispersion ratio. Simply, it is the ratio between arithmetic mean and geometric mean of a certain feature X. Arithmetic mean is the summation of values of all the features divided by the number of features. Geometric mean is the multiplication of all values of the feature power $\frac{1}{n}$, where n is the number of features. The following are the equations of AM and GM:

$$AM_i = \overline{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij} \qquad (24)$$

$$GM_i = \left(\prod_{j=1}^{n} X_{ij}\right)^{\frac{1}{n}} \qquad (25)$$

The ratio between AM and GM varies from 1 to infinity. AM is larger than GM. The value of R equals 1 if and only if all feature instances have the same value.

$$R_i = \frac{AM_i}{GM_i} \in \qquad (26)$$

Similar to all dispersion measures, as the value of R increases, the feature becomes more important.

2.4.2.11 Recursive Feature Elimination

This method uses a learning algorithm (e.g., linear regression), that assigns weights to each feature. In the beginning, the learning algorithm is trained on a set containing all features. The features are evaluated based on the coefficients of the learning algorithm or a feature importance estimator. The least important feature is eliminated from the set, and the training and elimination process is repeated.

The process terminates in two cases. Firstly, when the best m features are selected, where m is a user defined variable. Secondly, when the learning algorithm accuracy metric is below certain threshold T.

The relation between number of features and accuracy score is shown below in Figure 20 (Bengfort, 2020).

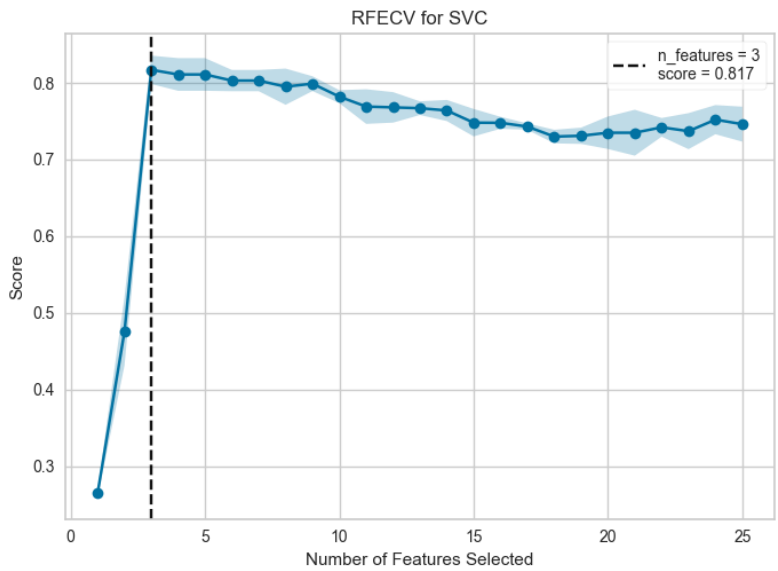


Figure 20: Number of Features with Accuracy Score

2.4.2.12 Correlation-based Feature Selection (CFS)

Correlation is a term used in statistics that represents a measurement of how a variable is related/correlated to each other. High positive value of correlation between two variables means that when one variable increases, the other increase too, for example $x = 3y$. High negative correlation means that if one variable increases, the other will decrease, and vice versa. Having a small value of correlation between two variables implies that if one variable increases, it states no information for the other variable (Doshi & Chaturvedi, 2014).

The previous discussion of correlation is the core of CFS. If two variables are highly correlated, then each variable is highly predicting of the other. CFS measures the efficacy of individual features in predicting the target. Choosing the features is

done by using heuristics that filter the redundant and irrelevant features that give the least significant prediction results (Pavya & Srinivasan, 2018).

The equation used to rate the feature is defined as:

$$F_s = \frac{N * r_a}{N + N(N-1) r_n} \quad (27)$$

2.5 Clustering

Different organizations and scientific sectors continue experiencing exponential growth in the amount of data at their disposal. To make productive use of such data, the data must be first categorized before the datasets can be explored; this categorization is done automatically using various tools, a process known as clustering (El Aissaoui et al., 2018). In generic terms, clustering refers to the process of classifying groups of different data objects as similar objects based on their closeness.

Clustering is a machine learning (ML) - based unsupervised algorithm that groups data points into clusters, in the process splitting data into various subsets. Every subset contains similar data; the subsets are referred to as clusters. This clustering represents a technical problem that needs to be overcome, and machine learning algorithms can solve the problem of grouping diverse data based on their closeness, making automatic clustering possible (Mittal et al., 2019). The techniques for automatic categorization of data groups (clustering) are useful in discovering and exposing the structure of a dataset (Novikov, 2019).

Liu et al. (2019) contended that large datasets require clustering in order to identify structures within the datasets that can be used as the basis for decision-making or to enable the identification of previously unknown groups within the data. Unusual observations in data that are distinct from other clusters can also be identified through

clustering, enabling the noise and outliers to be identified. Through clustering, data points belonging to the same cluster within a homogenous group can be summarized using a single representative cluster, thereby achieving a reduction in data (data volume) (Tang & Liao, 2021).

According to Novikov (2019) data points from a given cluster exhibit features that are similar, whilst data points from different clusters exhibit dissimilar features. Using ML algorithms to cluster data, the data points are segmented into various distinct groups from the data under interest, making clustering an unsupervised learning given the groups are not identified from known target classes.

Clustering, within the realms of data science, has several applications in a wide variety of industries and sectors- it can be applied in market research, data analysis, image processing, in search engines, and pattern recognition (Tang & Liao, 2021).

Clustering pertains to the ability of the clustering algorithm to scale approximately to the complexity as the amount of data objects is boosted in order of the algorithm (Liu et al., 2019). The outcomes of clustering should not only be comprehensible but interpretable and usable. Furthermore, the clustering algorithm should not be limited only to finding distance measurements (that have the tendency of discovering spherical small-sized clusters) but, be capable of finding arbitrarily shaped clusters (Liu et al., 2019). The clustering algorithm should be applicable to diverse data types, such as numeric, categorical, and binary data, as well as be sensitive to ‘noise’ in data; noise implies aspects such as missing, irrelevant, or incorrect data (Tang & Liao, 2021). In addition, high dimensionality implies that the clustering algorithm should be able to handle low-dimensional as well as high-dimensional data.

2.6 Application of Clustering in Customer Satisfaction

Data science principles provide a means by which customer satisfaction can be effectively measured; clustering algorithms can be applied in analyzing customer satisfaction by organizations. The clustering algorithm is a method that aids in segmenting customers; customer segmentation refers to the process of classifying customers with similar attributes into a single segment. Using the clustering algorithm, the customers can be understood better in the context of dynamic behaviors and static demographics (Krishnamurthy, 2011).

Customers with similar characteristics often interact in a similar way with businesses/organizations. Subsequently, a business can benefit from the clustering algorithms by developing marketing strategies that are tailored to each customer segment. Regarding data science, the clustering algorithm is termed an unsupervised algorithm for machine learning; to use the clustering algorithm for customer satisfaction classification, data on customers must first be prepared. There are a number of clustering algorithms that can be used in customer satisfaction applications, including k-means, Density-based Spatial Clustering of Applications with Noise (DBSCAN), Expectation maximization (EM), Clustering using GMM Gaussian Mixture Models (GMM), Mean-Shift Clustering, and Agglomerative Hierarchical Clustering (Schüller & Pekárek, 2018).

Abdi and Abolmakarem (2018) investigated the customer behavior mining framework through the use of clustering algorithms and classification techniques. The proposed customer behavior mining framework was applied to a telecom company using data mining techniques. Using the k-means clustering technique, a portfolio analysis was implemented on the data with previous customers grouped based on

socio-demographic factors. The cluster analysis was undertaken using two criteria; the number of services each customer in each group selected and the hours (number) of telecom services used by the customers. The analysis identified six customer groups with three attractiveness levels based on the results of analyzing the customer portfolio. The researchers undertook a second clustering devoted to customer behavior feature mining, and it was possible to predict the customers' churn behavior as well as the attractiveness of new customers (Abdi & Abolmakarem, 2018). The findings indicate that clustering can be used to scientifically gauge customer sentiment. From the findings, suitable tactics can be developed, based on customer attractiveness, to improve product offerings and develop tailor-made solutions for customers.

As per Zhang (2019) DBSCAN, a noise clustering algorithm, works through the distribution density of data points by identifying the data density degree and classifies data points within the distribution while identifying sporadic data points to be noise. Yang et al. (2021) applied the DBSCAN clustering algorithm in evaluating the capabilities of customer commissioners' fernet types of businesses by mining potential characteristics of categories and scoring the customer commissioners capabilities comprehensively under target categories using the entropy method. The authors were able to cluster the customer commissioners using the combined entropy and clustering scores into those with weak and strong capabilities, and the findings were applied to develop effective business training for the commissioners. EM-GMM works by assigning query points that maximize the posterior probability of the component to multivariate data points in the data, achieving flexible high-dimensional clustering (hard and soft) (Krishnamurthy, 2011). Sadewo et al. (2021) used the Mean Shift Clustering algorithm to maximize the total number of matches in a ride-sharing

application as a way of solving the matching problem (in ride-sharing). The use of the algorithm enabled more effective, easy, and better pairing of riders and drivers.

Agglomerative Hierarchical Clustering works using a bottom-up approach where every object is first considered as a leaf (single-element cluster), and at every step of the algorithm, the two most similar clusters are combined into nodes, which are new, bigger clusters. It has been demonstrated that hierarchical clustering, combined with linear regression and Ward's criterion in clustering, is able to be effective in partitioning of customers into segments based on their satisfaction (Schüller & Pekárek, 2018). Different customer preferences were inferenced using the agglomerative hierarchical clustering algorithm when used with linear regression and formed the basis for developing better, customer-focused solutions. Higher coefficients of determination were observed in linear models for the partitioned data compared to the whole market model (Schüller & Pekárek, 2018). The findings revealed that the ranks of customer satisfaction variables fluctuate significantly amongst the sectors. This is due to the fact that customers have various preferences.

Customer segmentation using clustering algorithms provides a number of benefits, according to Pascal et al. (2015) these include the ability for companies to develop marketing plans suited for each category of customers, provide business decision-support in situations fraught with high risks, such as developing credit relationships with clients and identifying services and products for each customer segment to help with demand forecasting and inventory management and to uncover useful details about customer associations with different product types.

2.7 Application of Feature Selection

In machine learning, the process of choosing a subset of pertinent characteristics to be used in the creation of a model is known as "feature selection." It is also referred to as variable subset selection, attribute selection, or variable selection (Kratsios & Hyndman, 2021). Techniques for feature selection are used for encoding inherent symmetries that exist within an input space, simplifying models so users can understand them better, shortening training times, improving the compatibility of data with learning model classes, and avoiding the dimensionality pitfall. When utilizing a feature selection strategy, the fundamental assumption is that the data includes some features that may be deleted with little to no information loss because they are irrelevant or redundant. It is possible for a single important feature to be redundant and yet have a strong correlation in the presence of another feature that is relevant (Kratsios & Hyndman, 2021). By removing unnecessary or redundant features, feature selection approaches are used to decrease the number of input variables. The list of features is then reduced to those that are most important to the ML model (Kratsios & Hyndman, 2021). In ML, a feature selection objective determines the most beneficial group of attributes that may be applied to create effective models of the phenomenon under study. Any algorithm's ability to anticipate outcomes requires effective feature representation; feature selection, which is used to increase the process' accuracy, is the most crucial phase in predictive ML (Hira & Gillies, 2015). By focusing on the most important variables and removing the redundant and unimportant ones, also improves the algorithms' ability to anticipate outcomes.

Finding the most pertinent features for successful prediction is the aim of feature selection; more predictive features provide users with more data points that may be utilized to anticipate the target with better results. The data points per region

become fewer as the number of features increases, making the feature space increasingly sparse. Significant-dimensional datasets frequently exhibit high sparsity, which poses a serious challenge for ML applications and results in the "curse of dimensionality" problem, alternately known as the "dimensionality curse" (Hira & Gillies, 2015). The term "curse of dimensionality" refers to the inability or failure of a model to recognize patterns and generalize from the training data due to the shallow feature space produced by the massive numbers of predictive features. In a model with a large feature space, there will be fewer data points for each region, which is problematic because models typically require adequate data point numbers per region to function satisfactorily. A model has a higher likelihood of fitting unusual observations that don't represent the population accurately if it is trained using data in a sparse feature space. The resulting model wouldn't generalize well and wouldn't perform well with new data. In machine learning, the amount of data needed to produce a reliable analysis increases exponentially as the dimensionality of the data increases (Hira & Gillies, 2015). Small oscillations in the data can be mistaken for significant variation by an overfitted model, which can result in classification errors. The term “curse of dimensionality” describes a number of phenomena that appear while organizing and analyzing data in high-dimensional contexts but do not exist in low-dimensional settings. Noisy characteristics can potentially make this challenge worse. Noise in a dataset refers to the variance error in a measured variable, which can be caused by measurement errors or random variation. Data that is noisy, which might be due to class or attribute noise, has a tendency to impact machine learning algorithms. To avoid needless complications in the inferred models and boost algorithmic effectiveness, noise should be eliminated as much as possible.

The feature selection approach is a crucial component of both machine learning and data mining. It is commonly used in the classification analysis of bioomics data as well as video, image, and text data; it is critical in the development of highly sensitive classification systems. There are two main categories of feature selection algorithms: wrapper and filter, depending on whether the feature selection process is independent of later models to train the learning process (Lu & Yuan, 2018). One area where feature selection has been applied within the context of ML is predicting the performance of students; Lu and Yuan (2018) evaluated a number of feature selection algorithms, including Relief, mRMR, AVC, and SVM-RFE, among others, and found variations in performance, with their proposed model, DPEFS (Optimized Ensemble Feature Selection Algorithm by Density Peaks), having a better feature selection for prediction performance both in the multi-class and binary class data. Ramaswami and Bhaskaran (2009) investigated various techniques of feature selection as applied in EDM (educational data mining) to determine the most relevant subset features that result in the highest prediction accuracy (in EDM) in terms of ROC (receiver operating characteristics) value and F-measure value through a comparative study. The ROC value compares the ability of the different selection techniques to predict a dichotomous outcome's specificity and sensitivity for a spectrum of values, while the F-measure evaluates the effectiveness of techniques used in feature selection. of the six algorithms for filter feature selection, namely GR (gain-ratio attribute evaluation), IG (information attribute evaluation), CB correlation-basic attribute evaluation) RF (relief attribute evaluation), SU (symmetrical uncertainty attribute evaluation), and CH (chi-square attribute evaluation). Applying the feature selection techniques to educational data (from India), Ramaswami and Bhaskaran (2009) established that, in terms of ROC values, the CB and IG techniques had the highest (RO) values. The

study also established variations in the F-measure of the different feature selection techniques, with CH, IG, and the SU techniques having the highest F-measures (Ramaswami & Bhaskaran, 2009). For example, it is found that the chi-square method ranks the following characteristics as important:

1. Area of residence of the students.
2. Father's occupation.
3. Means of transportation.
4. Private tuition.
5. Mother's income.
6. Type of school.

Then, by looking more closely at these characteristics, it can be clearly said that they are not directly or logically related to the student's performance. They may be related in some way to the student's performance, but to call them the most important characteristics responsible for the student's performance does not seem right. The neighborhood in which the student lives does not matter in this case. A student can perform well or poorly regardless of the neighborhood if the will and motivation are there. Similarly, the father's occupation is secondary in the performance evaluation. The father's profession cannot be directly blamed. A father can do whatever his career choice is. If he sends a child to an institution, his occupation cannot be judged. The type of transportation to school is also classified as an important characteristic, but there are other, more crucial factors than this. If the student can attend school, it does not matter how the student gets to school. Private tutoring is also ranked as an important item, but in today's digital world, all information is available on the Internet. If the school has better educational techniques, tutoring is not even necessary. A mother's income cannot be one of the most important characteristics,

there are many more important characteristics than this. The type of school can also be overcome by other important characteristics.

In sum, the findings of the investigation effectively confirm the well-known principle that prediction accuracy increases with the presence of fewer features. The student performance model's training phase and classification phase both show a reduction in construction costs and computing time, which aligns with the expected results. (Ramaswami & Bhaskaran, 2009).

The methods of feature selection have different performances and outcomes; filter feature selection techniques are faster and simpler compared to wrapper techniques; further, the filter techniques for selecting features are more model-agnostic, which means they have greater generalizability and so will not result in the overfitting of specific algorithms. Interpreting filter feature selection techniques is also quite simple since if a feature lacks any statistical association with the target, it is discarded. However, filter techniques also have their limitations, the most significant of which is their propensity to discard predictors that are useful, albeit weak predictors of targets on their own, but which, when used together with other predictors, can significantly add value to the model. According to Drotár et al. (2015) even though the wrapper strategy may produce superior results, more processing resources are needed. This is why a hybrid technique that mixes wrapper and filter methods has recently come into fashion. Selecting the filter method that provides the optimum relevance index for every case is one of the problems of filter techniques, and this is a difficult challenge to resolve. There are numerous indices for ranking and selection that result from various relevance evaluation techniques. Which filter provides the best relevance index for every case is one of the issues that must be addressed, and this is a difficult challenge to resolve. Numerous indices for selection and ranking are produced as a

result of various methods for evaluating significance (Drotár et al., 2015). Chen et al. (2020) examined the performance of four classifier methods: SVM (Support Vector Machines), LDA (Linear Discriminant Analysis), KNN (K-Nearest Neighbors), and RF (Random Forest). In order to choose the best classification technique based on each classifier's performance, the following feature selection techniques were combined: RFE, Boruta, and RF. The results showed that Random Forest was the most accurate classifier. Additionally, varImp() by RF emerged as the superior strategy for selecting features in all experiments using three distinct dataset methods when compared to RFE and Boruta. The RF technique results in significantly high prediction accuracy with very few features used; for instance, Chen et al. (2020) observed an accuracy in the prediction of EDM of 93.26% when using only six features; furthermore, the accuracy when using the RF technique was 98.57% when using 561 features. In addition, RF approaches are quite helpful and effective in identifying the key characteristics, so we shouldn't use every feature in the dataset. The literature review indicates that different methods for feature selection in EDM have different performances; some algorithms create the curse of dimensionality, which has an adverse effect on the accuracy of the predictive models. This has the effect of giving the wrong predictions because some feature selection techniques, particularly filter methods, have a propensity for discarding useful predictors; these predictors would result in greater prediction accuracy when combined with other features.

Chapter 3: Customer Satisfaction Improvement Methodologies

The significance of measuring customer satisfaction level or organizational retention level and customer base demands the utilization of a measuring tool and algorithm that help segment satisfaction level. This section focuses on reviewing important models that can determine these profound organizational performance components: customer retention and customer satisfaction.

3.1 The Kano Model

Shen et al. (2000) explained that complete awareness of customers' requirements, i.e., desires and anticipations, represents the critical and mandatory qualification for all those organizations that want to achieve customer satisfaction. Customer satisfaction is one of the most important tools to evaluate the quality of products and services. Almost two decades ago, Noriaki Kano conceptualized and presented an extremely beneficial diagram (Kano Model), as shown in Figure 21, to categorize the characteristics of a product or service bearing in mind how any product or service can fulfill the demands of the users. Kano's model is deeply entrenched in social psychology which is called the ''Motivator-Hygiene Theory'' by Frederick Hertzberg (Berger et al., 1993).

Kano (1984) differentiated three categories of service requirements that impact customer satisfaction in diverse ways when fulfilled. They include ''must be'' (basic) quality requirements, ''one-dimensional'' (performance) quality requirements, and ''attractive'' (excitement) quality requirements. The classification process might be advantageous for the innovative design guide as an outcome of novelty element.

The Kano model represents one of the practical methodologies used by managers to assess the most relevant product characteristics associated with customer

satisfaction (Sauerwein et al., 1996), and the method's effectiveness has captured researchers' growing interest (Witell et al., 2013). Theoretically, every qualitative and quantitative product characteristic can be classified into five categories (attractive, must-Be, one-dimensional, indifferent, reverse) (Lee & Huang, 2009).

Firstly, the category of attractiveness has characteristics that are referred to as Attractive requirements, and can be observed as minor bonuses that increase customer satisfaction but are not expected by the customers (Tontini, 2007). They refer to a product's characteristics that can improve customer satisfaction when they are present, but do not make customers dissatisfied when absent.

Secondly, the category of must-be quality refers to characteristics that are also called basic requirements, and are considered pre-requisite features that are disregarded and affect customer satisfaction only when absent (Tontini, 2007), so they would not satisfy customers when present, but would make them feel dissatisfied when absent.

Thirdly, the category of one-dimensional quality, also called performance requirements, affects satisfaction regardless of their presence and absence (Tontini, 2007). When present, they improve customer satisfaction, whereas their absence lead to less satisfaction.

Fourthly, the category of reverse quality attributes improves customer satisfaction when absent and reduces it when present. Finally, indifferent quality characteristics do not have a relevant contribution to customer satisfaction. Table 5, shows the 5 Kano categories.

Table 5: Kano Categories.

Kano Category	**Kano Code**
Must be (Basic)	1
One-dimensional (Performance)	2
Attractive (Excitement)	3
Indifferent	4
Reverse	5

The Kano model is based on three methodologies: questionnaires, evaluation tables, and result tables. The questionnaires are used to examine the element of service quality using a pair of functional and dysfunctional questions. Each question has five possible answers, dislike, like, must-be, neutral, and live-with (Meng et al., 2016). The evaluation table classifies each service quality element as one of the Kano categories for each respondent. The final Kano results have been recorded in the table. The observations are frequently sampled from the sample set of responses, and taken as the final element of service quality items.

The Kano model distinguishes various relationships between customer satisfaction and fulfillment of customer requirements (Violante & Vezzetti, 2017). This model primarily focuses on the qualitative analysis of curves. Various qualitative and quantitative methodologies have been proposed as an extension of the Kano model to understand and derive customer satisfaction accurately with less chance of errors. Various customer satisfaction models that were adopted in research include the analytical Kano (A-Kano) model that uses quantitative measures (Južnik & Kozar, 2017; Xu et al., 2009), the fuzzy Kano approach (Shokouhyar et al., 2017), the Kano method, which is based on the classical conjoint analysis model (Johnson et al., 2002), and the CS trust that combines quality of service (QoS) and customer satisfaction

prediction (Avikal et al., 2020). Violante and Vezzetti (2017) identified that the quantitative and qualitative Kano models could explain the association between customer satisfaction and fulfillment of customer requirements. Figure 21 shows The Kano model (adapted from Matzler & Hinterhuber, 1998).

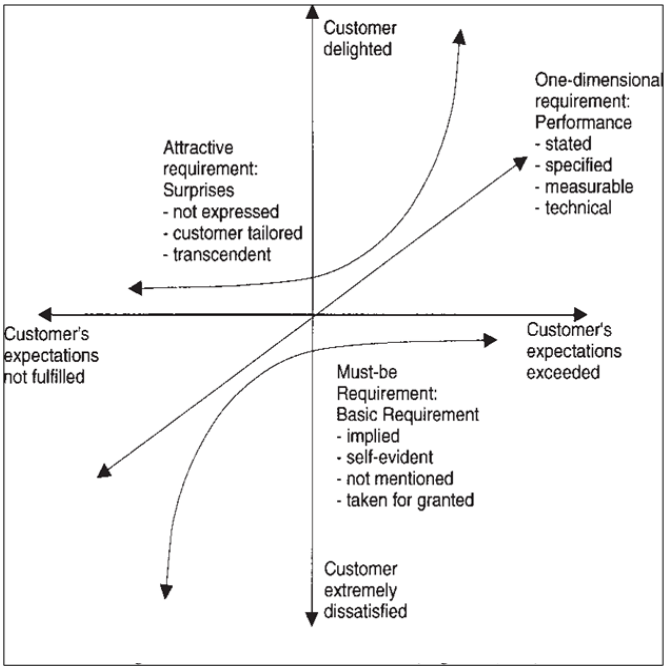


Figure 21: The Kano Model

The fuzzy Kano questionnaire was implemented to determine the most important food quality factors (Shokouhyar et al., 2017). A study Johnson et al. (2002) has utilized aggregate satisfaction measures leveraging marketing, sociological, psychological, and economic domains to assess cross-country and -industry customer satisfaction differences.

Bjertnaes et al. (2013) research discussed the quantitative and qualitative models' weaknesses and strengths, and reviewed an assessment framework that can identify the relationship between classification requirements and approaches. This framework helps select an appropriate methodology for judging customer satisfaction

(Bjertnaes et al., 2013). Various experiments have been performed to improve the Kano model's application due to its numerous deficiencies (Bjertnaes et al., 2013). This study proposes the A-Kano model, which primarily focuses on the analysis of customer needs. The Kano indices have been proposed following the principles of Kano for the incorporation of quantitative steps into customer satisfaction. This study has proposed two alternate methodologies, namely the Kano classifiers and the configuration index, to support the decision-making of a product design. The Kano classifiers are utilized for the tangible criteria to categorize customer needs. In contrast, the configuration index is utilized as a deciding factor for the product design. A product configuration's merit is justified by employing a Kano evaluator. Thus, producer capacity and customer satisfaction are leveraged.

Additionally, in research that revealed that the A-Kano model can optimize the relationship between producer capacity and customer satisfaction, Xu et al. (2009) presented a case study of automotive design focusing on dashboard. The Kano model employed by product designers to include products features demanded by users is one of the popular survey methodologies for user satisfaction (Xu et al., 2009).

The Kano model has certain drawbacks due to its tedious data analysis and user response processing, though it has several benefits. It is also more likely to be subjected to human errors (Atlason & Giacalone, 2018). One of the significant limitations of the Kano model is its inability to provide adequate quantitative results while reviewing the features leading to customer satisfaction.

Various customer satisfaction models can be adopted in research which can include the analytical Kano (A-Kano) model using quantitative measures (Xu et al., 2009), the fuzzy Kano approach (Shokouhyar et al., 2017), Kano method, which is based on the classical conjoint analysis model (Olsen et al., 2014), and CST rust that

combines the quality of service (QoS) and customer satisfaction prediction (Othman et al., 2017). One author (Violante & Vezzetti, 2017) identified the Kano model that uses quantitative and qualitative approaches, which could explain the association between customer satisfaction and customer requirement fulfillment (Violante & Vezzetti, 2017). A fuzzy Kano questionnaire was implemented to determine the most important factors in food quality (Shokouhyar et al., 2017).

Different scholars have used the Kano model to explain their viewpoints. The literature review reveals the following product and service quality features and their impact on customer satisfaction as mentioned by researchers. The summary of the literature review from different studies conducted on the Kano Model is presented in the Table 6.

Table 6: Literature Review Summary of Studies on the Kano Model

Kano Element Description	**Impact Measurement**	**Source**
Product and services (Quality)	Customer satisfaction	Kano et al. (1984)
Customers' preferences (Attractive, One-dimensional, Must-be, Indifferent, Questionable, Reverse.	Product development and customer satisfaction	Berger et al. (1993)
Product characteristics	Customer satisfaction	Sauerwein et al. (1996)
Service Attributes	Improvement in attributes	Huiskonen and Pirttila (1998)

Table 6: Literature Review Summary of Studies On the Kano Model (Continued)

Kano Element Description	**Impact Measurement**	**Source**
Reverse quality attributes	Improve customer satisfaction if absent and vice versa.	Tontini (2007)
Indifferent quality characteristics	No impact on customer satisfaction	Tontini (2007)
One dimensional product requirement	Competitive advantage	Witell et al. (2013)
Qualitative and quantitative product characteristics	Attractive, Must-Be, One-Dimensional, Indifferent, Reverse	Lee and Huang (2009)
A-Kano model	Optimize producer capacity and customer satisfaction	Xu et al. (2009)
Product features using Kano at design level	Customer needs	Xu et al. (2009)
Research on Kano elements	Effective role	Witell et al. (2013)
Quantitative and qualitative models' gaps	Customer satisfaction	Bjertnaes et al. (2013)
Experimental analysis	Improve Kano Application	Bjertnaes et al. (2013)

Table 6: Literature Review Summary of Studies On the Kano Model (Continued)

Kano Element Description	**Impact Measurement**	**Source**
One dimensional product requirements	Customer satisfaction	Redfern and Davey (2003)
Product Requirements (Must)	Customer Satisfaction	Busacca and Padula (2005)
Product unexpected features	Increase customer satisfaction	Tontini (2007)
Product with must be features	Customer dissatisfaction if not present	Tontini (2007)
Performance requirements	Affect customer satisfaction regardless of their presence or absence.	Tontini (2007)

3.2 Integration of Kano Model with Data Mining

This paper reviews data mining integration with the Kano model. The data mining model can predict customer satisfaction by employing a minimum number of customer attributes required with extremely accurate results. A correlation between the degree of these attributes and customer satisfaction can be analyzed (Gacto et al., 2019).

Thanks to this methodological approach, company market shares and customer loyalty can be enhanced, and risk can be reduced by avoiding investment in those attributes that are not directly linked to customer satisfaction maximization. The integration of the Kano model with the data mining approach is expected to enhance the limitations of previous standalone methodologies. Furthermore, organizational

performance transformation can be guaranteed and reinforced via effective customer satisfaction measurement (Yanan et al., 2018).

The Kano Model is identified to be utilized for a long period, and is also being utilized in recent years to support the process of determining the contentment of the clients. However, other techniques are identified to be supporting the process in a significant manner. One such technique is the process of data mining, and the study revolves around the identification of the fact that whether it is helpful to unite the functions of data mining techniques with the utilities of the Kano Model.

The utilization of the models of regression in relation to the process of determining the efficiency of the Kano Model is a subject of constant debate. Chang et al. (2009) have suggested the use of neural networks to support the functions of the Kano Model. In this way, the Kano Model can be utilized to determine the requirements of the clients in a differently.

Sufficient data of customers, products, and services have been collected and statistically analyzed overtime to provide insights into the field of business intelligence (Afshar, 2015). The Kano model has been used alongside other data mining approaches to complement the creation of customer satisfaction policies (Tontini, 2007). However, suggestions hold that the Kano model tends to favor user opinion based on quality attribute selection (Tontini, 2007). Moreover, most users seem to rate basic requirements with priority. Thus, a merger between the Kano model and data mining was proposed, with a view that such an approach would equally inform the best thresholds of a product and service's basic and innovative requirements.

Violante and Vezzetti (2017) criticized the Kano model, comparing it with the fuzzy Kano model. The study clarified that the fuzzy Kano model outperforms the traditional Kano model in determining customer satisfaction by analyzing their

appealing sentiments toward a product. However, the fuzzy Kano model has a disadvantage because it consists of open-ended questions that require considerable time and effort from interviewees, discouraging participants from responding (Violante & Vezzetti, 2017).

The Kano model has employed various regression analyses to evaluate the model's non-linear and asymmetric relationships (Cheng et al., 2019). Other researchers have criticized the effectiveness of those models that were conceptualized to assess the repetitions in order to evaluate the model's reliability in assessing the aspects of quality. The Kano model would be employed to extract users' inherent needs from the derived clusters (Hazra et al., 2016). The resulting model customized a website's content per user cluster and provided an improved newsfeed ideal for each user (Caballero,2017).

Shokouhyar et al. (2017) also proposed integration of the Kano model with data mining methodologies, i.e., K-mean clustering. The study examined a food industry company's response to a Kano questionnaire. The study revealed that combining the Kano model and K-mean clustering can provide better insights into their customer satisfaction policy (Shokouhyar et al., 2017). Based on the collected data, the study creates data clusters, and analyzes each cluster's needs.

Various studies promoted the integration of the Kano model and data mining methodologies. For example, Eid (2013) applied neural networks and the Kano method to the content recommendation in web Pearsonalization. Users are grouped into different clusters using artificial neural networks, and the Kano model is applied to extract the users' cluster requirements.

Moreover, Arefi et al. (2012) deployed the Kano model in higher education quality improvement by comparing quality indicators by using a traditional survey

between a current and an ideal situation. With student requirements related to higher education quality and those that influence students' satisfaction having been identified in the study, the Kano model was applied to cluster the requirements into five categories to determine the attributes that could increase customer satisfaction. Students' expectations and perceptions were compared and analyzed. The two-dimensional Kano model was employed, where indicators that had a noticeable negative gap were introduced in the model. Four categories of quality requirements were formed. The study measured the worst and best values and identified factors resulting in satisfaction and dissatisfaction (Okazaki et al., 2015).

Mikulić and Prebežac (2011) compared five approaches, namely the Kano model, "penalty–reward contrast analysis," "importance grid," qualitative data methodologies, and "direct classification." The study found that the Kano questionnaire and direct classification method are better than the other methods in classifying the attributes according to the Kano categorization. Wardy et al. (2014) explored the purchasing and customer satisfaction drivers of chicken eggs. Twenty egg product attributes influenced customer satisfaction. Principal component analysis (PCA1–PCA5) and the Kano model were applied. Furthermore, customer satisfaction in the food industry using K-means with the Kano model has been evaluated, where users were segmented into three clusters, whose needs have been categorized by the fuzzy Kano (Shokouhyar et al., 2017).

These previous studies aimed to predict customer satisfaction using data collected through traditional and online surveys. Feature selection methodologies have been applied to select and rank the most important attributes to reduce dimensionality. Furthermore, studies investigating the Kano model have applied the model without integrating any feature selection. The only integration that occurred was grouping

customers into different clusters, and then the Kano model was applied to extract the user requirements of each cluster. However, to the best of our knowledge, no studies to date have developed a model integrating the Kano model and feature selection to select and rank the most important customer satisfaction attributes as presented in this dissertation.

Bulk data analysis and finding the hidden correlations are considered data mining. This method can describe any existing organizational data, and predict future actions and situations. Customers' conversations, purchase records, and user data can be used to expand the understanding of customer needs as well as to provide insights into how to meet them, so data mining can also provide insights into customer interaction with the company.

By integrating data mining with the Kano model, customer data can be collected. Then, the Kano model can be applied to these large data for categorization of requirements into basic, performance, and Attractive. Kano integration with data mining needs a survey to collect information. However, the organization keeps customer records in database which can be directly fed into the Kano model for analysis of customer satisfaction attributes.

Several kinds of research have promoted the integration of the data mining approach and the Kano model. In (Okazaki et al., 2015), the research was conducted using data mining technique to explore customer engagement on Twitter. The identification of prosumers created social networks, and the study mainly aimed at the clear connection between customers' presumption and engagement. A data mining technique was employed in this regard. Tweets about IKEA were analyzed and experimented with as a sample. The produced and finalized algorithm was based on approximately 300 tweets, which were then applied to 4000 tweets to broaden the area

of research and findings. It was observed that satisfied and neutral customers disseminated objective statements, whereas dissatisfied customers disseminated subjective statements. A satisfied customer shared knowledge through which the presumption behavior was reflected (Okazaki et al., 2015).

The existence and continued use of the Kano model over the past three decades are indicative of the model's effectiveness in analyzing customer satisfaction (Eid, 2013). However, new approaches, such as data mining, have become popular. Thus, the following section will examine certain literature to elaborate on whether the use of data mining to complement the Kano model is a novel idea. Table 7, presents a literature review summary of studies on Kano-Data mining Integration.

Table 7: Summary of Studies On Kano-Data Mining Integration

Kano-Data Mining Integration	**Impact/Measurement**	**Source**
K-mean clustering,	Better insights into their customer satisfaction policy	Shokouhyar et al. (2017)
Data analysis	Provide insights into the field of business intelligence	Afshar (2015)
Kano-Data mining approaches	Creation of customer satisfaction policies	Tontini (2007).
Product and service's basic and innovative requirements	Customer satisfaction	Tontini (2007).
Fuzzy Kano model	Customer satisfaction by analyzing their appealing sentiments	Violante and Vezzetti (2017)
The worst and best values factors	Satisfaction and dissatisfaction	Okazaki et al. (2015)
Kano-Data mining integration	Customer engagement	Okazaki et al. (2015)

3.3 Integration Methods Key Finding

Different researchers have presented their specific viewpoints in respective studies as mentioned in the literature review throughout this report.

It shows how the integration with kano could enhance customer satisfaction in related to data mining A summary of key findings from these studies is presented below. Table 8: show summary of literature review.

- The integration of these tow techniques would help establish accurate customer requirements, needs, and expectations.
- These methodologies when integrated will help identify and prioritize products and services as per customer needs.
- Integration of these tow methodologies bestows priority furtherance in service attributes and technical necessities.
- Organizations who implement the integration of these tow methodologies will gain excellence in competition by understanding and incorporating customer needs in a better way, consequently leading to competitive advantage.
- It helps organizations to reveal marketing strategies.
- These will promote the customer excitement with aesthetic design features in products and services, which attract new customers and motivate existing customers.

Table 8: Summary of Literature Review

Exiting Data mining feature selection methods	**Proposed solution: integration kano model with Data mining**
Feature selection is practical when dealing with machine learning tasks and the problem of prediction. It removes the features with the lowest correlation with the target variable and removes features with high correlation with each other. This doesn't mean a large loss of information for the prediction model, but it might remove important features from the perspective of domain knowledge, so the feature selection doesn't reflect the importance of the features in reality. (Huang, 2015)	This research aims at developing a method to integrate the Kano model and data mining approaches to select relevant attributes that drive customer satisfaction and reduce the risk of investing in features that could ultimately be irrelevant to enhancing customer satisfaction.

This research study elucidates how an integrative approach of the KANO Model and data mining can be utilized for revamping the quality of products and services and augmenting customer satisfaction by transfiguring customer desires into engrossed product and service design.

Driving factors of statistical and data mining differ as being machine-driven and hypothesis-based respectively. Data mining focuses on important organizational links, whereas the statistical method tends to overlook certain links through association. Moreover, statistical modeling requires a trained researcher to determine

relevant models, whereas data mining involves the model's building process. Data mining does not ignore varying relationships in the independent and dependent variables. In contrast, the statistical method assumes a linear association between independent and dependent variables. Moreover, it is an economic-friendly method due to easy complex data management, user speed, friendliness, and performance compared with statistical methodologies. Therefore, data mining and statistical modeling differ although data mining consists of the application of automated statistical models that fit within their use (Magnini et al., 2003).

According to this research, there is a strong correlation between customer satisfaction and a variety of attributes. This model, in combination with data mining methods, is used to ascertain the most critical aspects that affect customer happiness besides limiting the risk of investing in characteristics that may not be significant in long term.

Chapter 4: Experiments

4.1 Experimental Setup

This section describes the experiment's methodology and provides an overview of the collected data. Also describe the equipment and detectors used. furthermore, describe the steps used to gather the data and do the experiments.

4.1.1 Dataset Setup

The datasets to be used in this research are collected by a questioner. The questionnaire consists of items designed to measure United Arab Emirates University students' satisfaction based on their standard question. The questionnaire consists of four questions. Here the first question is about the college details of the respondent. Then the second question is asked about the gender of the student. Then the third question has some of the sub-questions on this. The third question contains sub-questions that directly measure the service quality of the university like lab facilities, cleaning, and maintenance, etc. And the fourth question is overall satisfaction with the university. Here the questionnaire survey is shared with the university students through the survey platform named "Survey Monkey". This study could improve the quality of UAEU academic support and development services provided to their students in order to enhance the student's satisfaction level in related to the following area at UAEU:

- Academic and teaching quality
- Extra-curricular activities
- Library resources
- Academic advising
- Internship
- Registration process

- Campus facilities
- Social life
- Security
- Policies
- Research experience.
- Overall satisfaction

In this research, a survey was conducted involving students from the United Arab Emirates University. For this research, the sample was selected randomly from different colleges of United Arab Emirates University (UAE). It is found that nearly 14,387 students are studying in the UAE University. For ensuring the 95% confidence level and 5% margin of error, a minimum of 375 or more respondents are needed. It is calculated using the below-given formula.

$$\text{Sample}\ \frac{\frac{z^2 * p(1-p)}{e^2}}{1+\frac{z^2 * p(1-p)}{e^2 N}} \qquad (4.1)$$

Where N represents population size, e denotes Margin of error (percentage in decimal form), and z indicates z-score. So, a survey questionnaire was sent to nearly 1500 students. For reaching the respondents, online survey conducting tools were used. At the end of the data collection process, responses of 646 students were collected. Figure 22 show the sample distribution based on the college.

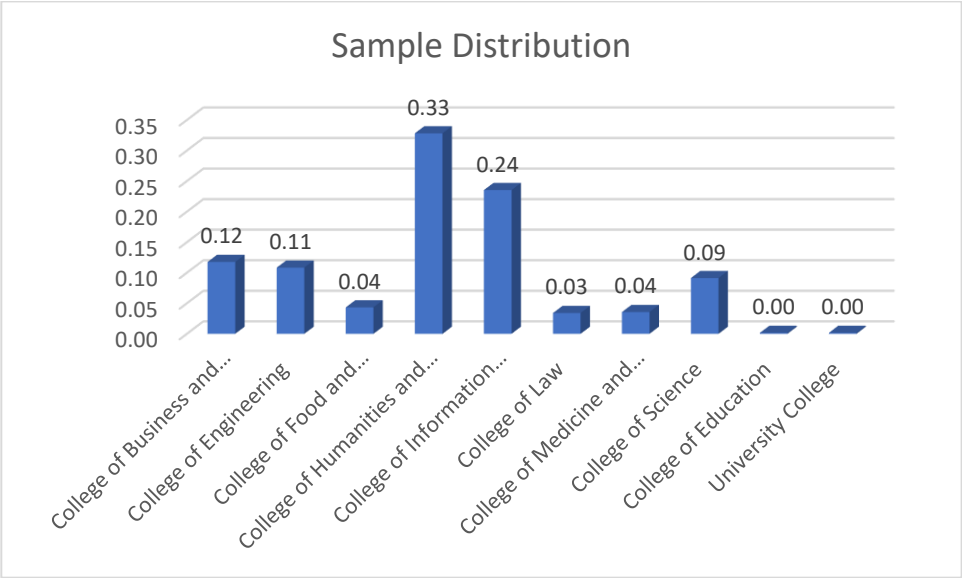


Figure 22: Sample Distribution Based On the College

4.1.2 Kano Model Dataset

There are 37 services and students were asked about what they think about each service. Then optional answers are shown in Table 9 below:

Table 9: Answers to Some Questions, Categorized by Kano

Kano Category	**Kano Code**
Must be (Basic): Their presence doesn't add to satisfaction but their absence cause dissatisfaction	1
One-dimensional (Performance): More is better and less is worst.	2
Attractive (Excitement): More is better (exceed customer expectation) but their absence doesn't cause dissatisfaction.	3
Indifferent: don't affect satisfaction	4
Reverse: More is worst and less is better	5

4.1.3 Student Satisfaction dataset

The satisfaction dataset is built based on Student satisfaction survey. Students were asked about their satisfaction Level about each service. The optional answer is from 1 (low) to 5 (Extremely):

1. Male/ female
2. Dormitory
3. Residence services and cleaning in the housing
4. Cleaning and hygiene on the campus
5. Modern equipment and decoration in the classrooms
6. Uncrowded classroom
7. Food dining hall services
8. The possibilities of doing lessons in the laboratories
9. Shopping services in school buildings
10. Student unions and clubs
11. Health services
12. The possibility of good communication with the teaching staff
13. The possibility of communicating with the administration
14. Transportation facilities on campus
15. How close the bus stations form classrooms?
16. How close the car parking?
17. Scholarships given by the university body
18. Shopping center on campus
19. Sports and entertainment facilities
20. Organizations of festivals, concerts, and celebrations
21. Advising unit and Tools

22. The Internship Experience
23. Information Technology Services
24. Online Registration Process
25. The Information in the E-services (Grades, Schedules, Payment Reports, etc.)
26. Organizing socio-cultural activities
27. Teaching quality
28. Curve grading system
29. The availability of internet in the campus
30. Organizing some courses with certificate
31. The libraries having got a rich data base
32. The range of Academic Majors
33. The security system on campus
34. University Policies and Regulations
35. Response to Complaints
36. Scooter
37. Online courses

The last Question is the about the overall satisfaction and it will be added to both dataset as a target column.

4.1.4 Pre-processing

Drop column with high missing value that have more than 640 missing values Also, there are column names having non-English characters, which were also removed in order to maintain meaningful column names.

We can omit the columns named Collector ID Respondent ID since identification numbers are irrelevant to satisfaction levels. We can omit the columns

named Collector ID Respondent ID since identification numbers are irrelevant to satisfaction levels. To ensure the data reliability, the reliability testing has been done using the SPSS version 26. The results show a Cronbach alpha (indicator) value of 0.938 for the whole dataset, which means that the dataset is fully reliable to perform further testing. Table 10 shows Satisfaction dataset and Table 11 shows Kano Dataset.

Table 10: Satisfaction Dataset

	Gender	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	...	Q29	Q30	Q31	Q32	Q33	Q34	Q35	Q36	Q37	target
0	0	1	1	1	5	5	5	3	4	5	...	5	5	5	5	4	4	5	5	5	4
1	1	4	5	5	5	5	3	3	3	3	...	5	3	5	5	4	4	3	3	5	4
2	1	5	5	5	5	5	5	3	4	5	...	5	5	5	5	5	5	5	5	5	5
3	1	1	2	5	4	1	2	3	3	4	...	4	4	2	2	3	1	3	4	1	3
4	0	3	3	5	3	3	2	3	2	2	...	3	3	2	2	4	3	2	3	4	4

Table 11: Kano Dataset

	Gender	Code2	Code3	Code4	Code5	Code6	Code7	Code8	Code9	Code10	...	Code29	Code30	Code31	Code32	Code33	Code34	Code35	Code36	Code37	target
0	0	1	1	2	1	1	1	5	5	5	...	1	3	1	2	1	1	1	5	5	5
1	0	1	1	1	1	1	1	1	1	1	...	1	1	1	1	1	1	1	1	1	5
2	0	4	4	4	3	2	2	1	2	1	...	1	1	1	3	1	1	2	1	1	4
3	0	1	1	1	1	1	1	2	1	1	...	1	1	1	2	1	1	1	1	1	4
4	1	2	4	2	3	1	3	3	3	3	...	3	4	4	4	4	4	4	4	4	3

4.1.5 Learning Models

In this case study, nine machine learning algorithms are used to evaluate the prediction performance: Linear Regression, Decision tree regressor, random forest regressor, AdaBoost Regressor, XGB Regressor., M5P Tree, Random Tree and Rep Tree Regression. These learning algorithms will be used because of their popularity in the recently published literature as well as their ranking as the most accurate data mining algorithms.

Model parameters are very important for model effects. Different model has different parameters, and each parameter has its meaning, so parameter calibration is indispensable. Sensitivity analysis can accurately explain which parameters are

sensitive in the study area, and these results can provide a basis for selecting the optimal parameter combination. When the optimal parameter combination is selected, the model's accuracy will be increase. So, parameter sensitivity analysis plays a complementary role to have the best model. The best model hyperparameters, or those that produce the most "correct" predictions, are found via grid search. (Lameski et al., 2015).

Analyzing the parameters of our model was very important because the chosen parameter's values introduced variability to the model's prediction of resulting dynamics. The first step in estimating model parameters was identifying sensitive parameters that impact model output. In general, this was a rapid way of getting a first look at the key parameters that drive model behavior. Class definitions can be changed depending on the behavior for which we wanted to identify these important parameters; for example, maximum depth, n_ estimators, learning rate.

4.1.6 Feature Selection Techniques

Feature selection techniques like ANOVA, Lasso, Chi-Square, Mutual Information, and Pearson are used to identify the best features out of 37 and compared the performance between models that are built based on top features.

4.1.7 Parameter Sensitivity

The most prevalent technique for creating an accurate classifier from the provided data is machine learning. In contrast, dynamical systems models provide further data about the time course of a system as opposed to just an outcome, but these models also include parameters that might be challenging to estimate given the limited data and model complexity. The quality of the link between a model's inputs and output, however, determines how effective a model is. This uncertainty needs to be

measured through sensitivity analysis before a model of this kind can be considered useful for prediction. This is a crucial step in the model-building process since it will inform the design of experiments, the assimilation of data, the estimation of parameters, and the complexity-reducing refining of models.

Regarding model impacts, model parameters are essential. Parameter validation is essential because various models contain unique parameters, and each parameter has a distinct meaning. The outcomes of sensitivity analysis can be used as a foundation for choosing the best parameter combination since they can identify which parameters are sensitive in the study region. The model's accuracy will rise after the ideal set of parameters is chosen. In order to have the best model, parameter sensitivity analysis is therefore complementary.

Our objective is to examine the model's parameters because the parameter values selected affect how accurately the model predicts the dynamics that will emerge. Finding sensitive parameters—those that have an impact on the model's output—is the first stage in estimating model parameters.

This is a pretty rapid technique to receive the first impression of the important variables that affect model behavior. Depending on the behavior for which we want to identify these crucial variables, we can alter the definition of the class. For instance, we can change the max depth, n estimators, learning rate, etc. (Al Nuaimi & Masud, 2020).

4.1.8 Parameters Setup

1. Number of groups / clusters (K) - for the K-Medoid baseline
2. Based on grid search, parameters for Learning algorithms – used default with the following parameters:

DecisionTreeRegressor:max_depth=15, criterion=mse

RandomForestRegressor : max_depth = 15, n_estimators = 500, criterion = mse

AdaBoostRegressor : n_estimators=500, base_estimator=DTR, learning_rate=0.01

XGB Regressor:max_depth=15, learning_rate=0.1, n_estimators=500, objective= reg:linear booster= gbtree .

4.1.9 Experiment Setup/ Hardware

The experiments will be conducted on a computer with Windows 10, 2.6 GHz CPU and 4 GB memory, Processor Intel(R) Core (TM) i7-6600U CPU @ 2.60 GHz 2.81 GHz, Installed RAM 8.00 GB (7.67 GB usable), System type 64-bit operating system, x64-based processor. The following resources was needed to conduct this research:

A. Technical resources:

We used good resources (high-speed processors) to handle model training while running the experiment.

- Google Collab
- API(s) from WEKA machine learning
- Survey Monkey

B. Administrative resources:

Publishing our research in well-known conferences/journals would support the research findings in the final examination.

4.1.10 Evaluation

Evaluation of the results will be carried out using a variety of performance assessment methodologies, for instance the mean absolute error, root means square error, and R-square value (Kazemi et al., 2015).

A. Classifier Performance Metrics

In this study, 10-fold cross-validation (CV) was employed, and k was assumed to be ten (k=10). Numerous studies using various learning algorithms on a large number of datasets demonstrated that 10 folds were approximately the optimal amount to provide the best estimate of error (Al Nuaimi & Masud, 2020).

4.2 Data Mining Experiments Dataset

As shown in Table 12, the categorization of all attributes according to kano model. it is found that the Kano model (one-dimensional and delight features) are 7, 13, 25, 27, 35, 12, 23, 31, and 36. These features can be considered as the most important features because it has an impact on customer satisfaction. The common features between Kano (one dimensional and delight features) and other feature selection methods are shown in Table 13.

Table 12: Kano Categorization

Symbols	**Questions**	**Satisfaction Level**
Code2	Dormitory	Basic
Code3	Residence services and cleaning in the housing	Basic
Code4	Cleaning and hygiene on the campus	Basic
Code5	Modern equipment and decoration in the classrooms: (projection machine, data machine, etc.)	Basic
Code6	Uncrowded classroom	Basic
Code7	Food dining hall services	Delight
Code8	The possibilities of doing lessons in the laboratories	Basic
Code9	Shopping services in school buildings	Basic

Table 12: Kano Categorization (Continued)

Symbols	**Questions**	**Satisfaction Level**
Code10	Student unions and clubs	Basic
Code11	Health services	Basic
Code12	The possibility of good communication with the teaching staff	One-Dimensional
Code13	The possibility of communicating with the administration	Delight
Code14	Transportation facilities on campus	Basic
Code15	How close the bus stations form classrooms	Basic
Code16	How close the car parking	Basic
Code17	Scholarships given by the university body	Basic
Code18	Shopping center on campus	Basic
Code19	Sports and entertainment facilities	Basic
Code20	Organizations of festivals, concerts, and celebrations	Basic
Code21	Advising unit and Tools	Basic
Code22	The Internship Experience	Basic
Code23	Information Technology Services	One-Dimensional
Code24	Online Registration Process	Basic
Code25	The Information in the E-services (Grades, Schedules, Payment Reports, etc.)	One-Dimensional
Code26	Organizing socio-cultural activities	Basic
Code27	Teaching quality	One-Dimensional
Code28	Curve grading system	Basic
Code29	The availability of internet in the campus	Basic
Code30	Organizing some courses with certificate	Basic
Code31	The libraries having got a rich data base	One-Dimensional
Code32	The range of Academic Majors	Basic
Code33	The security system on campus	Basic
Code34	University Policies and Regulations	Basic
Code35	Response to Complaints	One-Dimensional
Code36	Scooter	One-Dimensional
Code37	Online courses	Reverse

Table 13: Summary of Feature Selection Approach

Method	**Feature Selected**	**Common Variables**
Chi-Square	34,37,27,35,18,16,13,7,2	27, 35, 13, 7
Mutual Information	Gender,20,27,23,16,29,21,17, 9	27, 23
Lasso	27,2,37,7,36,34,6,21,35	27, 7, 36, 35
ANOVA	2,7,13,16,21,23,27,32,34	7, 13, 35, 27
Pearson	27,34,2,7,37,32,16,21,35	27, 7, 35

The common features were observed and assessed in a more comprehensive way to enhance the effectiveness of the study. Table 13 (summary of the feature selection approach), shows the detailed features selected by each machine learning approach as well as the Kano model. Moreover, the common features between different ML models as well as the Kano model are presented in the table.

4.3 Prediction Using All Attributes

Results involving different prediction methods have been presented in Figure 23 for the satisfaction datasets. The best results were obtained with XGB Regressor and AdaBoost Regression Model for Cross validation as shown in Table 14. The R squared value for this model was found to be 0.933. The high correlation coefficient value is 0.964. The Root Mean Squared Error of the model is 0.228. The Mean Absolute Error of the model is 0.082. The observation has given rise to the assumption that the contentment of the students can be effectively assessed with the help of this model. Moreover, the coefficient is often found to have a value that is higher than 0.6, which also seems to exhibit the fact that the model is effective in the process of prediction (Li, 2017). Figure 24, shows the line curve for XGB Regressor.

Table 14: Prediction Based On All Features

Prediction Method	**R_Squared**	**RMSE**	**MAE**	**Pearson's Correlation Coefficient**
Multiple Linear Regression	0.344	0.817	0.627	0.602
XGB Regressor	0.933	0.228	0.082	0.964
AdaBoost Regressor	0.933	0.194	0.055	0.965
Random Forest Regressor	0.887	0.304	0.190	0.947
Decision Tree Regressor	0.839	0.385	0.117	0.919
M5P Tree	0.6759	0.596	0.427	0.823
Random Tree	0.8736	0.372	0.101	0.936
Rep Tree Regression	0.6145	0.650	0.410	0.788

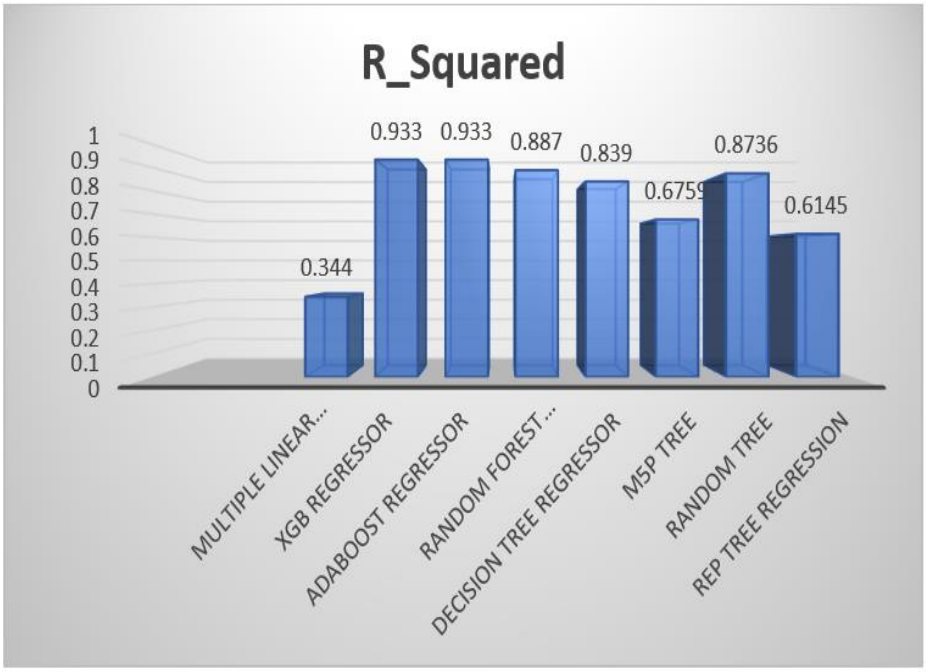


Figure 23: Prediction Based On All Features with Cross Validation

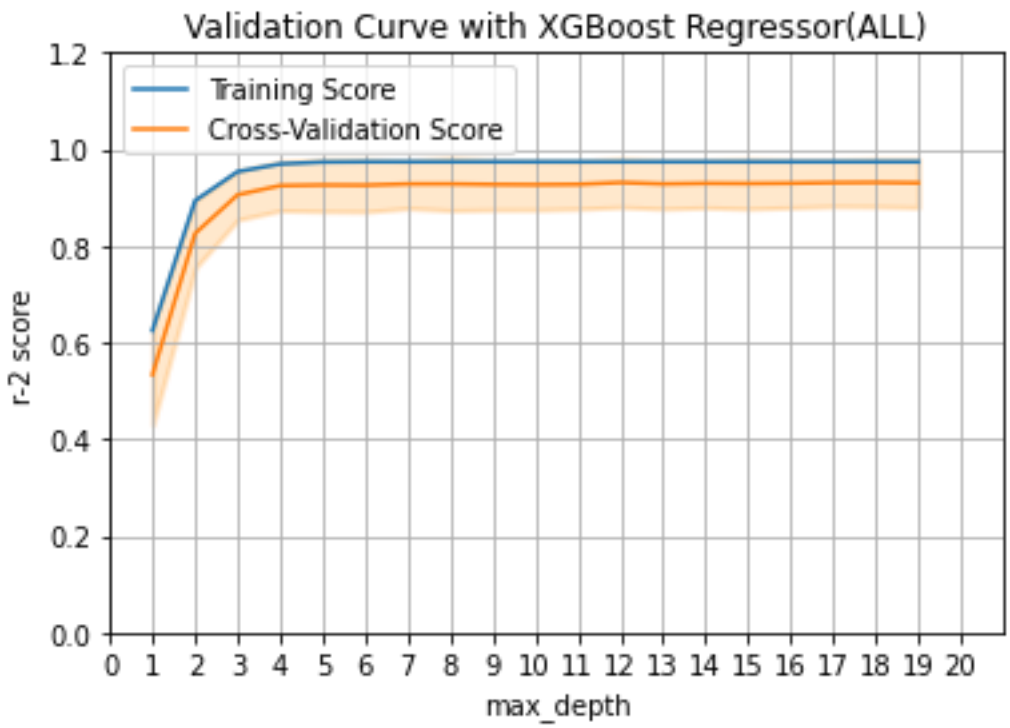


Figure 24: Shows the Line Curve for XGB Regressor

4.4 Prediction Using Selected Attributes

This section of the paper discusses the key results of various ML prediction models on the target variable. Here, different ML model's results for different feature selection methods are given as tablets. In this section, a detailed comparison of different ML models for different feature selection models has been provided (Mostert et al., 2021).

From the conducted feature selection process, four attributes have been selected. Here, these four attributes are common attributes between the ML-based feature selection as well as the Kano model feature selection (one-dimensional and delight features). We also tried different imputations on the feature selection model integration with the Kano model, like taking union attributes etc. However, the "Union Approach" increases the number of attributes. At the same time, taking common attributes for the analysis provides results nearly close to those with all variables.

4.4.1 Multiple Linear Regression (Selected Features)

Table 15 contains the key results of the Multiple Linear regression mode. From the Table 15, the lasso feature selection method integration with the Kano model performed very well with the multiple linear regression model. It has given a higher R-square value as well as a higher Pearson correlation value. In addition, this combination gives a lower RMSE value and MAE value. Here, the R-square value is found to be 0.346, and the Pearson correlation value is found to be 0.607. It means that 34 % of the variables found to be dependent on certain aspects can be examined using the opposite kind of variables. Another main parameter is the RMSE value for different feature selection approaches. Different RMSE values were found. Among them, for the lasso regularization-based feature selection method, the RMSE value and MAE

value are lower and are equal to 0.815 and 0.626 respectively. Here, the linear regression results are very poor as compared to other methods.

Table 15: Multiple Linear Regression

Multiple Linear Regression	**R_Squared**	**RMSE**	**MAE**	**Pearson's Correlation Coefficient**
ALL	0.344	0.817	0.627	0.602
kano	0.286	0.857	0.628	0.548
ANOVA	0.309	0.837	0.637	0.579
Chi-Square	0.331	0.826	0.632	0.591
Lasso	0.346	0.815	0.626	0.607
Mutual information	0.233	0.889	0.666	0.505
Pearson	0.332	0.823	0.631	0.596

4.4.2 XGB Regressor (Selected Feature)

Table 16 contains the key results of the XGB Regressor model with depth=15. From the Table 16, it is clear that the Chi-Square and Lasso based feature selection method integration with the Kano model has performed very well with the XGB Regressor model. It has given a higher R-square value as well as a higher Pearson correlation value. In addition, this combination gives a lower RMSE value and MAE value. Besides this, the R-square value is found to be 0.899, and the Pearson correlation value is found to be 0.949. It means that 89% of the variables found to be dependent on certain aspects can be examined using the opposite kind of variables. Here, the results show that the XGB Regressor model predicts the target variable very well than the multiple linear regression model. Another main parameter is the RMSE value for different feature selection approaches. Different RMSE values were found. Among them, the RMSE value and MAE for the Chi-Square feature selection method is lower, namely 0.297. The MAE value is equal to 0.105. Figure 25 shows outputs of R-squared

based on the XGB Regressor model with different methods for selected features. Figures 26 and 27 shows the line curve for XGB Regressor with Chi-square and Lasso.

Table 16: XGB Regressor

XGB Regressor depth=10	**R_Squared**	**RMSE**	**MAE**	**Pearson's Correlation Coefficient**
ALL	0.933	0.228	0.082	0.964
Kano	0.734	0.505	0.202	0.860
ANOVA	0.887	0.305	0.112	0.941
Chi-Square	0.899	0.296	0.103	0.949
Lasso	0.899	0.291	0.097	0.949
Mutual information	0.764	0.472	0.161	0.876
Pearson	0.895	0.285	0.101	0.945

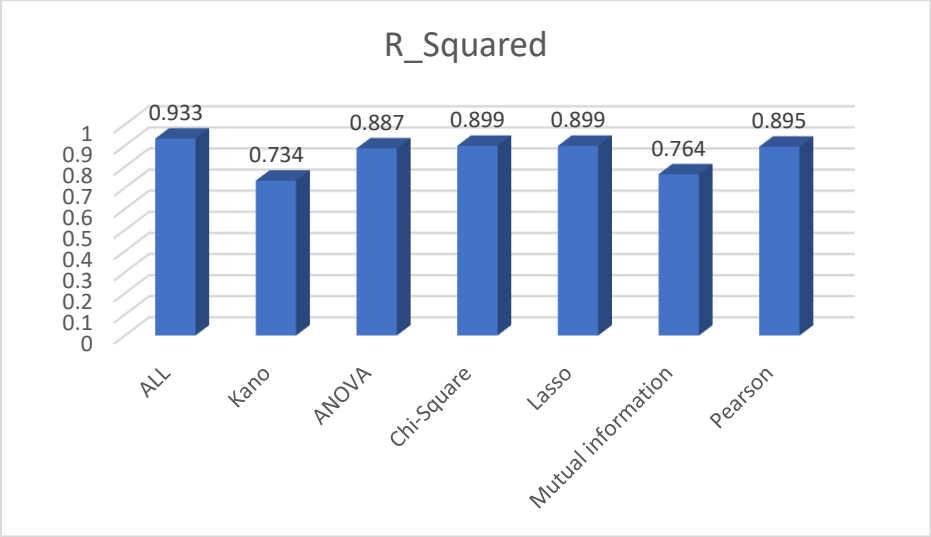


Figure 25: Outputs of R-Squared Based On the XGB Regressor Model with Different Methods for Selected Features

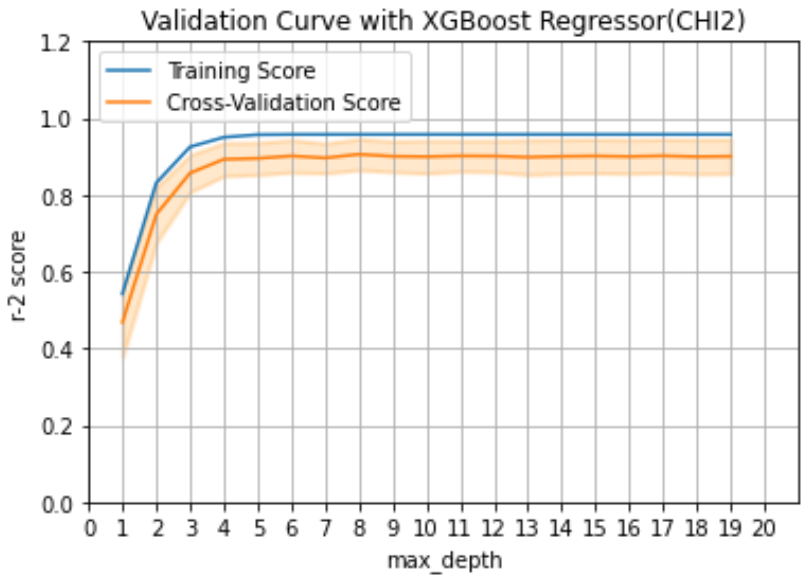


Figure 26: Shows the Line Curve for XGB Regressor with Chi-Square

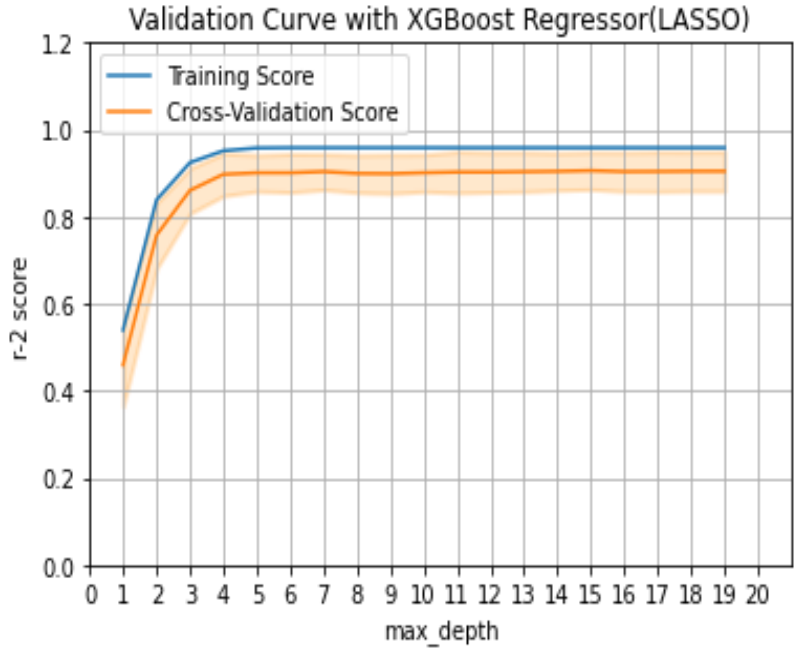


Figure 27: Shows the Line Curve for XGB Regressor with Lasso

4.4.3 AdaBoost Regressor (Selected Feature)

Table 17 presents the key results of the AdaBoost Regressor model. From the Table 17, the Pearson based feature selection method integration with the Kano model has performed very well with the AdaBoost Regressor model with depth = 15. It produced

a higher R-square value as well as a higher Pearson correlation value. Moreover, this combination gives a lower RMSE value and MAE value (Park & Oh, 2018). The R-square value is found to be 0. 908, and the Pearson correlation value is found to be 0. 952. It means that 91% of the variables found to be dependent on certain aspects can be examined using the opposite kind of variables. The outcomes were able to express the fact that the AdaBoost Regressor model with depth = 15 predicts the target variable very well than the multiple linear regressor model. However, the performance of the AdaBoost Regressor model with the Pearson feature selection is more than the XGB Regressor model with Pearson feature selection. The RMSE value for the Lasso feature selection method is lower; 0.269. The MAE value is equal to 0.083.

Table 17: AdaBoost Regressor Depth =15

AdaBoost Regression Depth =15	**R_Squared**	**RMSE**	**MAE**	**Pearson's Correlation Coefficient**
ALL	0.933	0.194	0.0553	0.965
Kano	0.735	0.507	0.190	0.861
ANOVA	0.882	0.316	0.100	0.937
Chi-Square	0.909	0.279	0.087	0.953
Lasso	0.903	0.279	0.089	0.949
Mutual information	0.781	0.460	0.148	0.885
Pearson	0.908	0.272	0.086	0.952

4.4.4 Random Forest Regressor Model (Selected Feature)

Table 18 contains the key results of the Random Forest Regressor model. From the Table 18, it is clear that the Lasso based feature selection method integration with the Kano model has performed very well with the Random Forest Regressor model. It provided a higher R-square value as well as a higher Pearson correlation value. Moreover, this combination produced a lower RMSE value and MAE value. The R-square value was found to be 0.868045, and Pearson correlation value was found to be 0.934255 (Jadhav, 2021). This means that 86% of the variables were found to be

dependent on certain aspects which can be examined using the opposite kind of variables. The outcomes were able to express the fact that the Random Forest Regressor model with a depth of 15 can predict the target variable better than the multiple linear regressor model. However, the performance of this model was slightly less than the performance of the AdaBoost Regressor model with the Lasso feature selection and the XGB Regressor model with the Lasso feature selection. The RMSE value for the Lasso feature selection method was lower; 0.336587. The MAE value is equal 0.190393.

Table 18: Random Forest Regressor Model

Random Regressor depth=15	**Tree Model R_Square**	**RMSE**	**MAE**	**Pearson's Correlation Coefficient**
ALL	0.887	0.304	0.190	0.947
Kano	0.700	0.534	0.288	0.843
ANOVA	0.848	0.369	0.219	0.923
Chi-Square	0.863	0.350	0.208	0.933
Lasso	0.868	0.337	0.193	0.934
Mutual information	0.751	0.490	0.254	0.872
Pearson	0.864	0.340	0.201	0.932

4.4.5 Decision Tree Regressor Model (Selected Feature)

Table 19 contains the key results of the Decision Tree Regressor model. From the Table 19, it is clear that the Pearson feature selection method integration with the Kano model has performed very well with the Decision Tree Regressor model with depth = 15. It has produced a higher R-square value as well as a higher Pearson correlation value. Furthermore, this combination provides a lower RMSE value and MAE value. The R-square value is found to be 0.881, and the Pearson correlation value is found to be 0.938. The decision Tree Regressor model predicts the target variable very well than the multiple linear regressor model. Besides this, the performance of

the Decision Tree Regressor model with the Pearson feature selection is lower than the XGB Regressor model with the Pearson feature selection. The RMSE value for the Pearson feature selection method is lower; 0.303. The MAE value is equal to 0.103.

Table 19: Decision Tree Regressor

Decision Regressor depth=15	**R_Sqpuared**	**RMSE**	**MAE**	**Pearson's Correlation Coefficient**
ALL	0.839	0.385	0.117	0.919
Kano	0.651	0.570	0.236	0.824
ANOVA	0.843	0.353	0.116	0.919
Chi-Square	0.843	0.373	0.118	0.923
Lasso	0.862	0.341	0.110	0.929
Mutual information	0.753	0.486	0.163	0.870
Pearson	0.881	0.304	0.103	0.938

4.5 Common Features

From the conducted feature selection process, four attributes have been selected. Here, these four attributes are common attributes between the ML-based feature selection as well as the Kano model feature selection (one-dimensional and delight features). We also tried different imputations on the feature selection model integration with the Kano model, like taking union attributes etc. However, the “Union Approach” increases the number of attributes. At the same time, taking common attributes for the analysis provides results nearly close to those with all variables.

4.5.1 Multi Linear Regression (Common Features)

Table 20 contains the key results of the Multiple Linear regression mode. From the Table 20, the lasso feature selection method integration with the Kano model performed very well with the multiple linear regression model. It has given a higher R-square value as well as a higher Pearson correlation value. In addition, this

combination gives a lower RMSE value and MAE value. Here, the R-square value is found to be 0.299933, and the Pearson correlation value is found to be 0.556884. It means that 38% of the variables found to be dependent on certain aspects can be examined using the opposite kind of variables. Another main parameter is the RMSE value for different feature selection approaches. Different RMSE values were found. Among them, for the lasso regularization-based feature selection method, the RMSE value is lower and is equal to 0.861678. The MAE value is equal to 0.638478. Here, the linear regression results are very poor as compared to other methods.

Table 20: Multiple Linear Regression

Multiple Linear Regression	**R_Squared**	**RMSE**	**MAE**	**Pearson's Correlation Coefficient**
ANOVA	0.283	0.871	0.641	0.543
Chi-Square	0.283	0.871	0.641	0.543
Lasso	0.300	0.862	0.638	0.557
Mutual information	0.208	0.918	0.675	0.474
Pearson	0.281	0.872	0.642	0.542

4.5.2 XGB Regressor (Common Features)

Table 21 contains the key results of the XGB Regressor model with depth=15. From the Table 21, it is clear that the ANOVA and Chi-Square based feature selection method integration with the Kano model has performed very well with the XGB Regressor model. It has given a higher R-square value as well as a higher Pearson correlation value. In addition, this combination gives a lower RMSE value and MAE value. Besides this, the R-square value is found to be 0.561, and the Pearson correlation value is found to be 0. 756. It means that 57.8% of the variables found to be dependent on certain aspects can be examined using the opposite kind of variables. Here, the results show that the XGB Regressor model predicts the target variable very well than

the multiple linear regression model. Another main parameter is the RMSE value for different feature selection approaches. Different RMSE values were found. Among them, the RMSE value for the ANOVA feature selection method is lower, namely 0.679. The MAE value is equal to 0.397.

Table 21: XGB Regressor

XGB Regressor depth=15	**R_Squared**	**RMSE**	**MAE**	**Pearson's Correlation Coefficient**
ANOVA	0.561	0.679	0.397	0.756
Chi-Square	0.561	0.679	0.397	0.756
Lasso	0.546	0.693	0.410	0.741
Mutual information	0.298	0.863	0.633	0.552
Pearson	0.456	0.754	0.517	0.678

4.5.3 AdaBoost Regressor (Common Features)

Table 22 presents the key results of the AdaBoost Regressor model. From the Table 22, the Chi-Square based feature selection method integration with the Kano model has performed very well with the AdaBoost Regressor model. It produced a higher R-square value as well as a higher Pearson correlation value. Moreover, this combination gives a lower RMSE value and MAE value. The R-square value is found to be 0.542704726, and the Pearson correlation value is found to be 0.747036. It means that 54% of the variables found to be dependent on certain aspects can be examined using the opposite kind of variables. The outcomes were able to express the fact that the AdaBoost Regressor model with depth = 15 predicts the target variable very well than the multiple linear regressor model. However, the performance of the AdaBoost Regressor model with the Chi-Square feature selection is lower than the XGB Regressor model with Chi-Square feature selection. The RMSE value for the Chi-

Square feature selection method is lower; 0.690943. The MAE value is equal to 0.411381.

Table 22: AdaBoost Regressor.

AdaBoost Regression	**R_Squared**	**RMSE**	**MAE**	**Pearson's Correlation Coefficient**
ANOVA	0.543	0.691	0.411	0.747
Chi-Square	0.543	0.691	0.411	0.747
Lasso	0.522	0.709	0.426	0.730
Mutual information	0.272	0.877	0.680	0.540
Pearson	0.423	0.777	0.542	0.660

4.5.4 Random Forest Regressor Model (Common Features)

Table 23 contains the key results of the Random Forest Regressor model. From the Table 23, it is clear that the ANOVA based feature selection method integration with the Kano model has performed very well with the Random Forest Regressor model. It provided a higher R-square value as well as a higher Pearson correlation value. Moreover, this combination produced a lower RMSE value and MAE value. The R-square value was found to be 0.545, and Pearson correlation value was found to be 0.744. This means that 52.7% of the variables were found to be dependent on certain aspects which can be examined using the opposite kind of variables. The outcomes were able to express the fact that the Random Forest Regressor model with a depth of 15 can predict the target variable better than the multiple linear regressor model. However, the performance of this model was slightly lower than the performance of the AdaBoost Regressor model with the ANOVA feature selection and the XGB Regressor model with the ANOVA feature selection. The RMSE value for the ANOVA feature selection method was lower; 0.690. The MAE value is equal to 0.441.

Table 23: Random Forest Regressor

Random Regressor depth=15	**Tree Model R_Square**	**RMSE**	**MAE**	**Pearson's Correlation Coefficient**
ANOVA	0.543	0.692	0.443	0.743
Chi-Square	0.543	0.692	0.443	0.743
Lasso	0.539	0.698	0.448	0.735
Mutual information	0.295	0.865	0.640	0.551
Pearson	0.449	0.760	0.536	0.672

4.5.5 Decision Tree Regressor Model (Common Features)

Table 24 contains the key results of the Decision Tree Regressor model. From the Table 24, it is clear that the ANOVA and Chi-Square feature selection method integration with the Kano model has performed very well with the Decision Tree Regressor model with depth = 15. It has produced a higher R-square value as well as a higher Pearson correlation value. Furthermore, this combination provides a lower RMSE value and MAE value. The R-square value is found to be 0.529404208, and the Pearson correlation value is found to be 0.737. The decision Tree Regressor model predicts the target variable very well than the multiple linear regressor model. Besides this, the performance of the Decision Tree Regressor model with the ANOVA feature selection is similar to the XGB Regressor model with the ANOVA feature selection. The RMSE value for the ANOVA feature selection method is lower; 0.704. The MAE value is equal to 0.407.

Table 24: Decision Tree Regressor

Decision Regressor depth=15	**Tree Model R_Square**	**RMSE**	**MAE**	**Pearson's Correlation Coefficient**
ANOVA	0.529	0.707	0.408	0.736
Chi-Square	0.529	0.705	0.407	0.737
Lasso	0.518	0.714	0.419	0.724
Mutual information	0.298	0.863	0.633	0.552
Pearson	0.435	0.770	0.524	0.665

4.5.6 M5P Tree (Common Features)

Table 25 contains the key results of the M5P tree model. From the Table 25, it is clear that the ANOVA and Chi-Square feature selection method integration with the Kano model has performed very well with the M5P tree model. It has produced a higher R-square value as well as a higher Pearson correlation value. The R-square value is found to be 0.6759, and the Pearson correlation value is found to be 0.5452. Furthermore, this combination provides a lower RMSE value and MAE value. They are found to be 0.8781 and 0.6458 respectively.

Table 25: M5p Tree

M5P tree	**R-square**	**RMSE**	**MAE**	**Pearson correlation coefficient**
ANOVA	0.297	0.878	0.646	0.545
chi	0.297	0.878	0.646	0.545
lasso	0.332	0.856	0.633	0.576
Mutual	0.285	0.886	0.660	0.534
Pearson	0.304	0.874	0.640	0.552

4.5.7 Random Tree (Common Features)

Table 26 contains the key results of the Random Tree model. From the Table 26, it is clear that the Pearson feature selection method integration with the Kano model has performed very well with the Random Tree model. It has produced a higher R-square value as well as a higher Pearson correlation value. The R-square value is found to be 0.5602, and the Pearson correlation value is found to be 0.7514. Furthermore, this combination provides a lower RMSE value and MAE value. They are found to be 0.6946 and 0.4028 respectively.

Table 26: Random Tree

Random Tree	**R-Square**	**RMSE**	**MAE**	**Pearson correlation coefficient**
ANOVA	0.546	0.706	0.411	0.743
chi	0.546	0.706	0.411	0.743
lasso	0.536	0.714	0.420	0.734
Mutual	0.462	0.769	0.545	0.681
Pearson	0.560	0.695	0.403	0.751

4.5.8 Rep Tree Regression (Common Features)

Table 27 contains the key results of the Rep Tree Regression model. From the Table 27, it is clear that the ANOVA and Chi-Square feature selection method integration with the Kano model has performed very well with the Decision Tree Regressor model with depth = 15. It has produced a higher R-square value as well as a higher Pearson correlation value. The R-square value is found to be 0.4209, and the Pearson correlation value is found to be 0.6536. Furthermore, this combination provides a lower RMSE value and MAE value. They are found to be 0.797 and 0.5719 respectively. The results show that XGB Regressor with depth 15 and Decision Tree Regression with depth 15 exhibit the best performance. Only four features have been used to predict the common features between ANOVA and Kano features located under one-dimensional and delight categories. The R-Square value, RMSE, MAE, and Pearson Correlation Coefficient are found to be 0.69, 0.58, 0.32, and 0.83 respectively, which are closer to the model with all attributes.

Table 27: Rep Tree Regression Depth = 15

Rep Tree Regression	**R-Square**	**RMSE**	**MAE**	**Pearson correlation Coefficient**
ANOVA	0.421	0.797	0.572	0.654
chi	0.421	0.797	0.572	0.654
lasso	0.407	0.807	0.567	0.643
Mutual	0.414	0.802	0.599	0.645
Pearson	0.400	0.811	0.589	0.636

The outcomes were able to suggest that the techniques that are used to assess the repetitions were found to be capable of identifying the relationship between the contentment of the clients and different aspects of the services of the institution. Different methods were found to be effective in determining the aspects that are more significant in contributing to the contentment of the clients. In addition, the process of obtaining similar aspects was able to enhance the precision in relation to the assessment of all different characteristics. With the help of this information, the administration team of the institution will be able to significantly improve the contentment of the clients. The outcomes were able to determine the major aspects that were found distinctly in various institutions. So, it will be a wise move to combine various university services for predicting customer satisfaction. Various university services seem to emphasize various aspects; therefore, the unification process will be able to substantially enhance the services of all the involved institutions. Moreover, the study can be equipped in different situations to obtain effective outcomes.

The significance of every aspect in relation to the contentment of the clients has been carefully observed. From this research, it is clear that the maximum values of R-square and Pearson correlation are found to be 0.561 and 0.756 respectively for Decision Tree Regressor as well as XGB Regressor. Moreover, the used feature selection approach is ANOVA Based Feature selection approach as well as Chi-square approach as shown if Figure 28. Here, these results are derived by using four different parameters like R-square value, RMSE, MAE, and Pearson correlation coefficient. The common attributes between the ANOVA features selection method and Kanos one-dimensional and delight features produce the highest Pearson correlation coefficient value; 74%. It is nearest to the results with all the attributes with 96% Pearson correlation coefficient. It was achieved with only four features, which can be

considered a very small number of features as compared to the full model, which has 37 attributes. This shows that the ANOVA technique is effective in the identification of aspects of the students that are found to be effectively contributing to the contentment of the students.

Moreover, the four similar characteristics between Kano and ANOVA feature selection can produce acceptable readings of performance if the information is adequate. The four mentioned characteristics are Food, Dining Hall, Services, and the Possibility of communicating with the administration. They are located under the delight category. The other common services are Information in the E-services (grades, schedules, payment reports, etc.) and teaching quality. They are located under the one-dimensional category. The teaching quality feature has been selected by all feature selection methods, which means that it is the most important attribute. The correlation coefficient between the features and the student satisfaction index is not less than 0.48 for all prediction methods. Figures 29 and 30 show the line curve for XGB Regressor with Chi-square and ANOVA respectively.

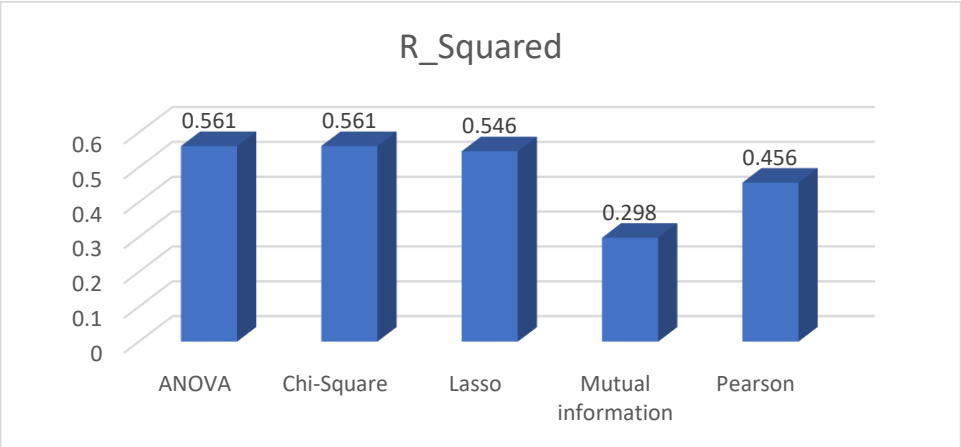


Figure 28: XGB Regressor Model with Depth=15 R-Squared Common Results with ANOVA

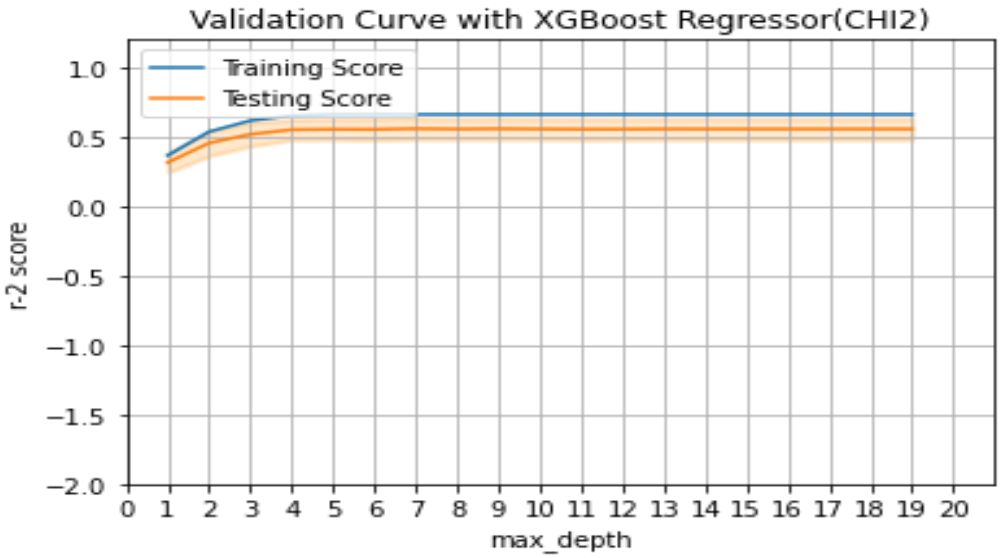


Figure 29: Shows the Line Curve for XGB Regressor with Chi-Square

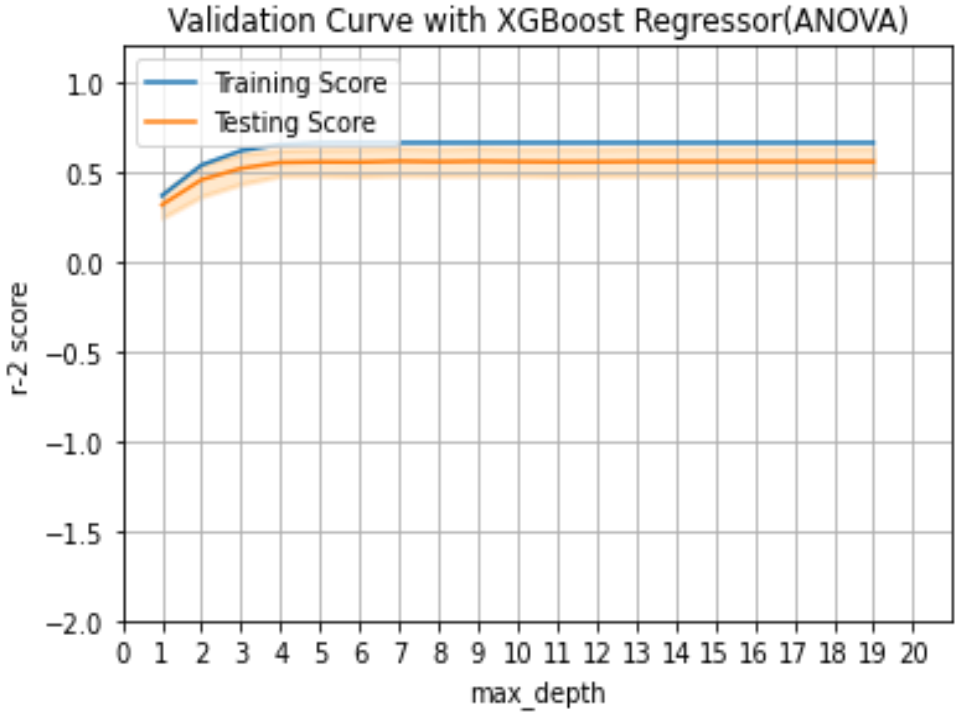


Figure 30: Shows the Line Curve for XGB Regressor with ANOVA.

4.6 Union Attributes between Both Feature Selections and the Kano Model

A union experiment has been conducted for the better clarification of the usefulness of integrating feature selection with Kano model. In this experiment, the prediction was made with the union between the 9 Kano features (7, 13, 25, 27, 35, 12,

23, 31, and 36) and features from each feature selection method. This method includes large number of features as compared to common experiment features. For example, the union between ANOVA and Kano results in 14 features as compared to 4 features in common experiment. Table 28 shows union attributes between both feature selections and the Kano model most important features. In the following sections, the result of the prediction on the union features will be presented. Moreover, these results will be compared with the common experiment results in order to observe the change in accuracy regarding this increase in number of features.

Table 28: Union Attributes Between Both Feature Selections and Kano Model

Method	**Feature Selected**	**Union Variables**
Chi-Square	34,37,27,35,18,16,13,7,2	34,37,18,16,2,7,13,25,27,35,12,23,31,36
Mutual Information	Gender,20,27,23,16,29,21,17,9	G,20,16,29,21,17,9,7,13,25,27,35,12,23,31,36
Lasso	27,2,37,7,36,34,6,21,35	2,37,34,6,21,7,13,25,27,35,12,23,31,36
ANOVA	2,7,13,16,21,23,27,32,34	2,16,21,32,34,7,13,25,27,35,12,23,31,36
Pearson	27,34,2,7,37,32,16,21,35,6	34,2,37,32,16,21,7,13,25,27,35,12,23,31,36

4.6.1 Multiple Linear Regression (Union Features)

Table 29 contains the results of the multiple linear regression models. From the Table 29, it is noticeable that the integration of the Kano method with the Lasso feature selection methods noticeably outperforms other methods in multiple linear regression. Both methods resulted in higher R-square and correlation coefficient values in addition to lower values for RMSE and MAE. For Lasso, R-square and correlation coefficient values were found to be 0.37008435 and 0.615005. For Pearson, R-square and correlation coefficient values were found to be 0.366699769 and 0.61283. This means

that 43% of the variability of the dependent variable was explained by the independent variables in the union group. Lasso presented lower RMSE and MAE values as compared to other methods. For Lasso, RMSE and MAE values are found to be 0.816777 and 0.629896 respectively. By comparing the previous results with the common results, it can be noticed that even with increased number of features, the improvement in accuracy was not that significant. In the common experiment, with 4 features, the R-square value for Lasso was found to be 0.32942 as compared to 0.41105 using 14 features in the union experiment.

Table 29: Multiple Linear Regression

Multiple Linear Regression	**Tree Model R_Square**	**RMSE**	**MAE**	**Pearson's Correlation Coefficient**
ANOVA	0.346	0.831	0.633	0.598
Chi-Square	0.368	0.818	0.632	0.614
Lasso	0.370	0.817	0.630	0.615
Mutual information	0.314	0.853	0.625	0.567
Pearson	0.367	0.818	0.632	0.613

4.6.2 Decision Tree Regression (Union Features)

Table 30 contains the results of Decision Tree Regression Depth = 15 model. From the Table 30, it is noticeable that the integration of the Kano method with ANOVA feature selection methods noticeably outperforms other features selection methods in Decision Tree regression. Both methods resulted in higher R-square and correlation coefficient values in addition to lower values for RMSE and MAE. For ANOVA, R-square and correlation coefficient values were found to 0.856607412 and 0.926664. This means that 86% of the variability of the dependent variable was explained by the independent variables in the union group. ANOVA presented lower RMSE and MAE values as compared to other methods. For ANOVA, RMSE and MAE values are found to 0.856607412 and 0.926664. By comparing the previous

results with the common results, it can be noticed that even with increased number of features, the improvement in accuracy was significant. In the common experiment, with 4 features, the R-square value for ANOVA was found to 0.529404208 as compared to 0.856607412 using 14 features in the union experiment. Number of features should be taken into consideration. Even 0. .856607412 for R-square is a promising result. It required large number of features as compared to the common experiment.

Table 30: Decision Tree Regression Depth = 15

Decision Regressor depth=15	**R_Squared**	**RMSE**	**MAE**	**Pearson's Correlation Coefficient**
ANOVA	0.857	0.383	0.109	0.927
Chi-Square	0.803	0.444	0.123	0.903
Lasso	0.824	0.430	0.120	0.911
Mutual information	0.790	0.468	0.146	0.895
Pearson	0.808	0.429	0.118	0.905

4.6.3 Random Forest Regression (Union Features)

Table 31 contains the results of Random Forest regression Depth = 15 model. From the Table 31, it is noticeable that the integration of the Kano method with Lasso and Pearson feature selection methods noticeably outperforms other features selection methods in Random Forest regression. Both methods resulted in higher R-square and correlation coefficient values in addition to lower values for RMSE and MAE. For Lasso, R-square and correlation coefficient values were found to be 0.875855501 and 0.939275 respectively. This means that 85.5% of the variability of the dependent variable was explained by the independent variables in the union group. Lasso presented lower RMSE and MAE values as compared to other methods. For Lasso, RMSE and MAE values are found to be 0.351066 and 0.201107 respectively.

The results of union between the Kano model most important features and feature selection methods combined with random forest regression model are highly promising. When these results are compared with common features prediction using the same model, the difference in performance metrics can be noticed. In the union experiment, the R-square value for ANOVA is found to be 0.865153. In common, the highest R-square value was 0.527030 using 4 features selected by ANOVA. Large number of features in union experiment decreases the significance of the result.

Table 31: Random Forest Regression Depth = 15

Random Regressor depth=15	**R_Squard**	**RMSE**	**MAE**	**Pearson's Correlation Coefficient**
ANOVA	0.867	0.367	0.206	0.935
Chi-Square	0.871	0.357	0.206	0.937
Lasso	0.876	0.351	0.201	0.939
Mutual information	0.806	0.448	0.238	0.907
Pearson	0.871	0.357	0.206	0.936

4.6.4 AdaBoost Regression (Union Features)

Table 32 contains the results of AdaBoost Regression model. From the Table 32, it is noticeable that the integration of the Kano method with Chi-Square feature selection methods, noticeably, outperforms other features selection methods in AdaBoost Regression. Both methods resulted in higher R-square and correlation coefficient values in addition to lower values for RMSE and MAE. For Chi-Square, R-square and correlation coefficient values were found to be 0.921389092 and 0.959953 respectively. This means that 88.6% of the variability of the dependent variable was explained by the independent variables in the union group. Chi-Square and Pearson presented lower RMSE and MAE values as compared to other methods. For Chi-Square, RMSE and MAE values are found to be 0.275238 and 0.075145 respectively. The results of union between the Kano model most important features

and feature selection methods combined with random forest regression model are highly promising. When these results are compared with common features prediction using the same model, the difference in performance metrics can be noticed. In the union experiment, the R-square value for Chi-Square is found to be 0.921389092. In common, the highest R-square value was found to be 0.542704726 using 4 features selected by Chi-Square. Large number of features in union experiment decreases the significance of the result.

Table 32: AdaBoost Regression

AdaBoost Regression depth 15	**R_Squared**	**RMSE**	**MAE**	**Pearson's Correlation Coefficient**
ANOVA	0.898	0.320	0.091	0.947
Chi-Square	0.921	0.275	0.075	0.960
Lasso	0.913	0.289	0.078	0.956
Mutual information	0.854	0.395	0.116	0.926
Pearson	0.912	0.290	0.083	0.955

4.6.5 XGB Regression (Union Features)

Table 33 contains the results of XGB Regression Depth = 15 model. From the Table 33, it is noticeable that the integration of the Kano method with the Chi-Square feature selection methods noticeably outperforms other features selection methods in XGB Regression Depth = 15. Both methods resulted in higher R-square and correlation coefficient values in addition to lower values for RMSE and MAE. For Chi-Square, R-square and correlation coefficient values were found to be 0.922 and 0.959 respectively. This means that 92.1% of the variability of the dependent variable was explained by the independent variables in the union group. Chi-Square presented lower RMSE and MAE values as compared to other methods. For Chi-Square, RMSE and MAE values are found to be 0.075323099 and 0.2771339 respectively. The results of union between the Kano model most important features and feature selection

methods combined with XGB Regression model are highly promising. When these results are compared with common features prediction using the same model, the difference in performance metrics can be noticed. In the union experiment, the R-square value for Chi-Square is found to be 0.88910. In common, the highest R-square value was found to be 0.578308 using 4 features selected by ANOVA. Large number of features in union experiment decreases the significance of the result. Figure 31 shows the outputs of R-squared based on XGB Regressor depth=10 with the different methods. Figures 32 and 33 show the line curve for XGB Regressor AdaBoost Regressor with Chi-square respectively.

Table 33: XGB Regression Depth = 15

XGB Regressor depth=15	**Tree Model R_Square**	**RMSE**	**MAE**	**Pearson's Correlation Coefficient**
ANOVA	0.899	0.316	0.090	0.948
Chi-Square	0.922	0.272	0.076	0.960
Lasso	0.912	0.288	0.078	0.955
Mutual information	0.854	0.395	0.115	0.926
Pearson	0.918	0.278	0.079	0.958

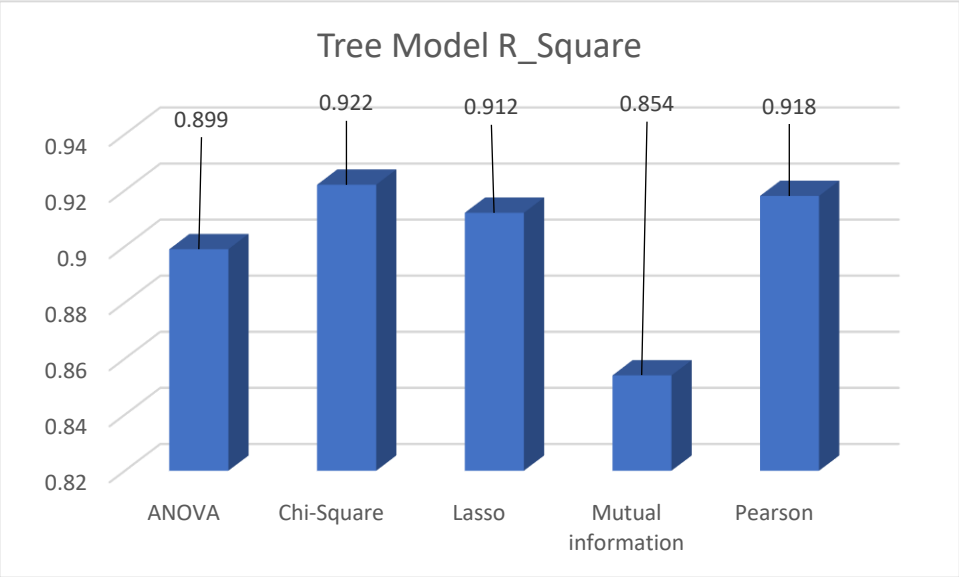


Figure 31: Outputs of R-Squared Based On XGB Regressor Depth=10 with the Different Methods

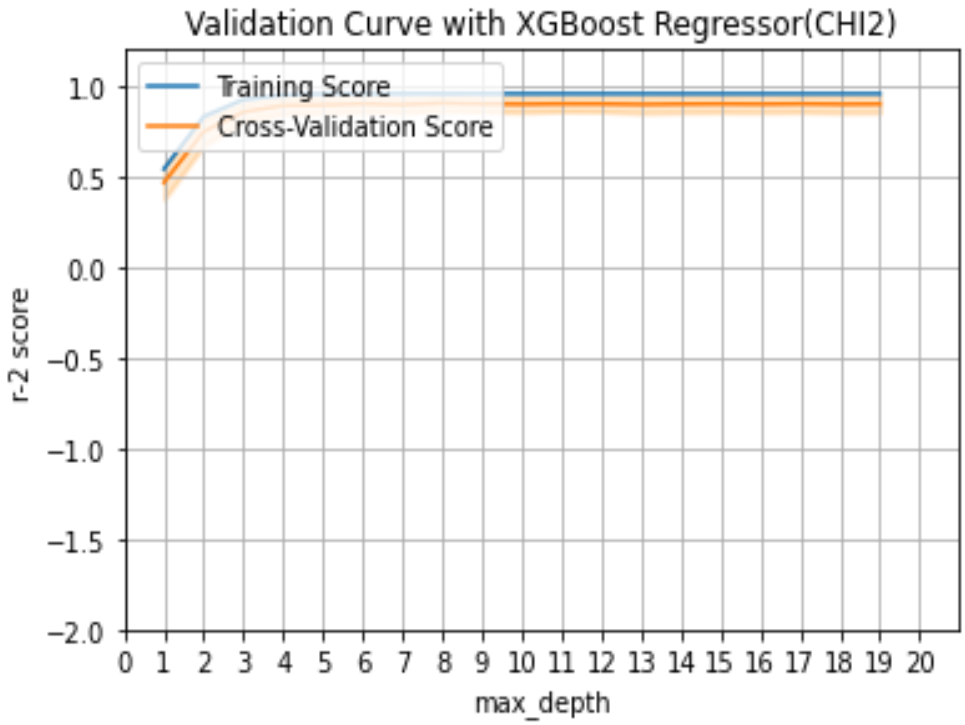


Figure 32: Shows the Line Curve for XGB Regressor with Chi-Square

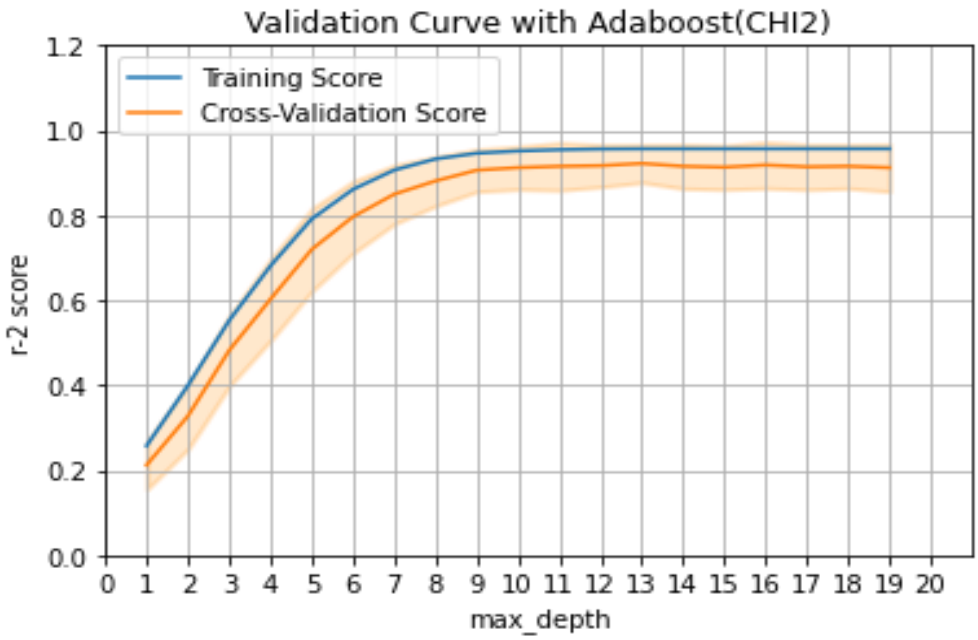


Figure 33: Shows the Line Curve for AdaBoost Regressor with Chi-Square.

In the union experiment, features from the Kano's model and feature-selection methods were taken into consideration. The union result from the two methods was

used to make a prediction using various prediction techniques. The result of the prediction accuracy metric was higher than that of the common experiment. The Chi-Square's feature-selection techniques reached the best accuracy when integrated with the Kano's model in the union experiment. The XGB Regressor and AdaBoost features achieved 0.921 R-square value and 0.959 for Pearson's correlation coefficient. Moreover, this method resulted in small values for RMSE and MAE. Figure 31 shows the outputs of R-squared based on XGB Regressor depth=10 with the different methods.

The flaw in this experiment is the large number of features. The main purpose of this research is to minimize the number of features for business companies' improvers. The union experiment highlights the difference between large and small number of features. It also clarifies the usefulness of common approach, which achieves acceptable performance as compared to the small number of features.

In the union experiment, features from the Kano model and feature selection approaches were taken into account. Using the combined results of the two methods, other prediction approaches were employed to make the predictions. The prediction accuracy metric yielded a greater result than the standard experiment. In a union experiment, the Lasso and Pearson feature selection strategies had the best accuracy when used with the Kano model.

The large number of features in this experiment is a problem. The major goal of this study is to reduce the number of features available to business improvers. The union experiment demonstrates the distinction between large number of features and limited number of features. It also clarifies the utility of a common strategy that achieves acceptable performance despite the limited number of features.

4.7 Clustering

Clustering is another approach in data mining that can be helpful in extracting more useful information from the data. In customer satisfaction, the idea of delivering the needs for each customer is essential. Customers vary significantly in their needs. For this, the model of detecting the factors that affect customer satisfaction has to satisfy the needs of each sector. The idea behind the clustering experiment is to divide the data into 8 clusters based on Elbow method. Each cluster will represent a sector of students, cluster students to different group based on different features, such as transportation, teaching quality and online studying and other features. The features analyzed through clustering can have a big impact on student satisfaction, for example girls or immigrants. Girls, as the most important feature, might be very different from immigrant students. For example, immigrant's first concern is the dorm's cost.4.6.1 Clustering the Kano Dataset.

Eight clusters are created based on Elbow method using Kano dataset. After that, the best features of each cluster are extracted from the same students in satisfaction dataset. Using different feature selection methods, researchers get the best features for each cluster.

4.7.1 Clustering the Kano Dataset

The 8 clusters are created using the Kano dataset. After that, the best features of each cluster are extracted from the same students in the satisfaction dataset. Researchers got the best features for each cluster by using the Kano's and the five-feature-selection methods. After getting the best features, a comparison was made between all clusters to find out the common features. The number of clusters was selected based on the elbow method. We plot the elbow method for clusters k=2 to k=12 and we managed to know that the plot line started decreasing rapidly at points 8-

9 and started increasing a little at k=10. With the elbow, we got good intuition about our best k value which will be 8-9. Figure 34 shows the correlation between CS and number of clusters using elbow method. After this, we run k-means on our data for k=2 to k=12 and we got the best result with k=8 which has the best accuracy score. With this, we can say that the value of k was 8 and it was the best. Following that, the best features of each cluster were extracted from the same students in the satisfaction dataset. Researchers got the best features for each cluster by using different feature-selection methods. After getting the common feature between all of the clusters, it was noticed that Q2 was the most common feature among all feature-selection techniques. Also, it was observed that Q27, Q2, Q7, and Gender were the most common among clusters. The following tables show important features for each cluster in comparison with an important feature for all features.

- Table 34 using Pearson feature selection.
- Table 35 using Lasso feature selection.
- Table 36 using Mutual Information feature selection.
- Table 37 using ANOVA feature selection.
- Table 38 using Chi-Square feature selection.

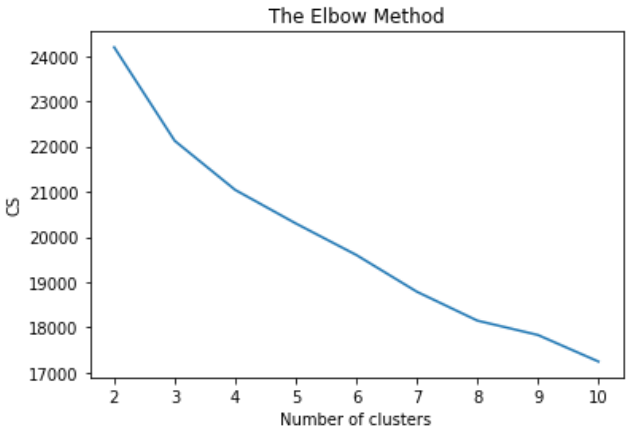


Figure 34: Correlation Between Cs and Number of Clusters

Table 34: Most important features in each cluster based on Pearson feature selection -C means cluster-

C1	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**Common Attributes**
Q34 Q7 Q17 Q35 Q10 Q13 Q27 Q22 Q16	Q21 Q27 Q6 Q2 Q35 Q13 Q16 Q22 Q28 Q32 Q12 Q25 Q34 Q18 Q37 Q26 Q3 Q8 Q7	Q7 Q27 Q2 Q12 Q31 Q8 Q37 Q32 Q6 Q19 Q34 Q29 Q17 Q10 Q4 Q13 Q30 Q14 Q11 Q28 Q5 Q25 Q26 Q9	Q34 Q6 Q7 Q26 Q16 Q37 Q13 Q36 Q35 Q27 Q31 Q22 Q28 Q21 Q32 Q2 Q18 Q25 Q12 Q20 Q10 Q24 Q19	Q27 Q7 Q34 Q2	Q27 Q21 Q25 Q32 Q34 Q16 Q2 Q3 Q37 Q12 Q6 Q7 Q17 Q10 Q28 Q36 Q24	Q37 Q35 Q22 Q34 Q7 Q9 Q24 Q27 Q25 Q17 Q11	Q2 Q27 Q3 Q26 Q35 Q21 Q16 Q18 Q22 Q37 Q20 Q6 Q8 Q9	Q27 Q2

Table 35: Lasso Feature Selection

C1	C2	C3	C4	C5	C6	C7	C8	Common Attributes
Q34	Q2	Q7	Q34	Q27	Q27	Q37	Q2	Q2
Q7	Q21	Q27	Q6	Q2	Q25	Q35	Q27	Q7
Q17	Q6	Q2	Q37	Q7	Q2	Q4	Q26	
Q8	Q27	Q12	Q7	Q31	Q3	Q16	Q21	
Q37	Q13	Q34	Q36		Q37	Q10	Q37	
Q2	Q37	Q37	Q26		Q21	Q34	Q3	
Q35	Q36	Q6	Q31		Q36	Gender	Q36	
Q36	Q35	Q31	Q2		Q34	Q2	Q18	
Q13	Q28	Q8	Q13		Q7	Q3	Gender	
	Gender	Q4	Q20		Q10	Q6	Q4	
	Q3	Q17	Gender		Gender	Q7	Q6	
	Q4	Q36	Q3		Q4		Q7	
	Q5	Gender	Q4		Q6		Q8	
	Q7	Q3	Q5		Q8		Q9	
	Q8	Q5	Q8		Q9			
	Q9	Q9	Q9		Q11			
	Q10	Q10	Q10		Q12			
	Q11	Q11	Q11					
	Q12	Q13	Q12					
		Q14	Q14					
		Q15	Q15					
		Q16	Q16					
		Q18	Q17					
		Q19						

Table 36: Mutual Information Feature Selection

C1	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**Common Attributes**
Gender	Gender	Gender	Gender	GenderQ7	Gender	Q35	Gender	Gender
Q8	Q29	Q15	Q27	Q27	Q36	Q37	Q30	Q27
Q34	Q15	Q21	Q33	Q33	Q33	Gender	Q9	
Q27	Q23	Q29	Q20		Q21	Q20	Q16	
Q4	Q27	Q31	Q19		Q9	Q3	Q32	
Q15	Q20	Q5	Q31		Q7	Q29	Q31	
Q23	Q7	Q9	Q12		Q16	Q31	Q19	
Q16	Q19	Q4	Q5		Q32	Q32	Q20	
Q29	Q4	Q25	Q29		Q29	Q11	Q27	
	Q5	Q10	Q26		Q18	Q25	Q22	
	Q12	Q20	Q11		Q25	Q6	Q8	
	Q28	Q8	Q34		Q2		Q37	
	Q30	Q7	Q21		Q28		Q5	
	Q6	Q30	Q15		Q27		Q18	
	Q16	Q36	Q25		Q17			
	Q31	Q19	Q23		Q26			
	Q33	Q23	Q3		Q10			
	Q35	Q32	Q30					
	Q21	Q6	Q32					
		Q26	Q6					
		Q27	Q22					
		Q35	Q4					
		Q33	Q7					
		Q22						

Table 37: ANOVA Feature Selection

C1	C2	C3	C4	C5	C6	C7	C8	Common Attributes
Q34 Q24 Q27 Q23 Q25 Q8 Q17 Q7 Q31	Q27 Q21 Q25 Q6 Q13 Q35 Q23 Q4 Q2 Q18 Q32 Q16 Q28 Q29 Q12 Q22 Q30 Q34 Q20	Q32 Q19 Q8 Q27 Q7 Q15 Q23 Q2 Q10 Q18 Q31 Q29 Q28 Q26 Q6 Q20 Q11 Q34 Q12 Q22 Q17 Q5 Q9 Q37	Q31 Q20 Q34 Q30 Q6 Q4 Q26 Q32 Q7 Q19 Q16 Q27 Q29 Q18 Q25 Q13 Q23 Q22 Q35 Q36 Q37 Q12 Q10	Q27 Q34 Q11 Q7	Q16 Q27 Q21 Q25 Q6 Q32 Q34 Q12 Q10 Q28 Q2 Q18 Q7 Q13 Q17 Q26 Q3	Q37 Q35 Q26 Q3 Q2 Q27 Q13 Q28 Q4 Q15 Q10	Q2 Q27 Q20 Q3 Q29 Q21 Q18 Q16 Q36 Q37 Q35 Q31 Q4 Q9	Q27

Table 38: Chi-Square Feature Selection

C1	C2	C3	C4	C5	C6	C7	C8	Common Attributes
Q34 Q24 Q13 Q35 Q17 Q25 Q23 Q37 Q10	Q35 Q13 Q21 Q25 Q18 Q16 Q2 Q34 Q28 Q23 Q27 Q6 Q10 Q37 Q22 Q26 Q4 Q32 Q24	Q19 Q32 Q28 Q31 Q8 Q7 Q23 Q2 Q11 Q27 Q18 Q10 Q37 Q34 Q26 Q29 Q20 Q6 Q15 Q14 Q16 Q25 Q22 Q17	Q31 Q20 Q30 Q7 Q18 Q37 Q34 Q6 Q36 Q16 Q26 Q35 Q4 Q19 Q22 Q23 Q10 Q13 Q25 Q29 Q2 Q32 Q28	Q27 Q34 Q11 Q24	Q25 Q16 Q28 Q21 Q27 Q18 Q37 Q34 Q17 Q13 Q6 Q10 Q32 Q24 Q12 Q2 Q23	Q37 Q2 Q28 Q3 Q35 Q26 Q13 Q17 Q10 Q34 Q27	Q20 Q2 Q29 Q3 Q18 Q27 Q37 Q36 Q21 Q23 Q11 Q31 Q4 Q35	No common

Based on the Table 39, we found that Q2 is the most common attributes between the 8 clusters with lasso Feature selection. Consequently, we considered them important, and it is added to kano model 9 features. Then these 10 features are used to predict the R-square value. As shown in Figure 35, there was a high increase in the value of R-square. In contrast, when we added other non-common attributes there was no increase in the R-value. Based on Table 39, R-square for Kano was 73%. After the integration with data, mining Q2 was added with Kano as the most common attributes among 8 clusters, high increase to the value of R-square and reached 88% which is almost the same as R-value for all attributes.

Also, Gender attribute is the second most common features among the 8 clusters Adding Gender to Kano increases the R-square value to 73% and this value was not similar to the value of R-Square for Kano+Q2. Other features which were not frequently common among 8 clusters did not increase the Q-square value and remained around 66%. Also, when we run Kano with Q2, the same increase happened and the value of R-square increased from 73% to 93%.

Clustering in combination with the Kano model has a lot of potentials. The findings of each cluster's prediction utilizing the best four features are highly significant. The clusters yielded significant outcomes.

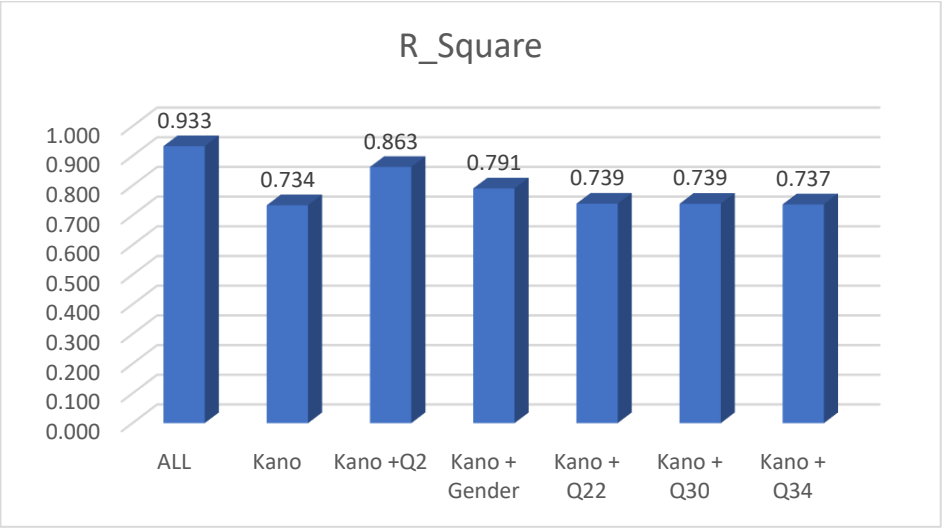


Figure 35: R-Squared Outputs with Different Integration Between Kano and Top Common Features Among 8 Clusters

Table 39: XGB Regressor with Cross Validation

XGB Regressor Depth = 10	**Tree Model R_Square**	**RMSE**	**MAE**	**Pearson's Correlation Coefficient**
ALL	0.933	0.228	0.082	0.964
Kano	0.734	0.505	0.202	0.860
Kano +Q2	0.863	0.333	0.126	0.929
Kano + Gender	0.791	0.448	0.170	0.899
Kano + Q22	0.739	0.500	0.197	0.863
Kano + Q30	0.739	0.500	0.194	0.863
Kano + Q34	0.737	0.502	0.2	0.863

4.7.2 Kano Based Ranking Attributes

Categories in kano model have definite hierarchical rules based on attribute effects on the satisfaction (or dissatisfaction) of customers, with must-be being most influential while indifferent has the least influence. The attributes can be ranked based on Kano responses, in which total satisfaction index is developed; the outcome is better, or worse values are computed using the associations below. Customer satisfaction index (SSI) = (Attractive + One dimensional)/ (Attractive + One

dimensional + Indifference + Must be). Customer dissatisfaction index (DDI) = (-1) x (One dimensional+ Must be)/ (Attractive + One dimensional + Indifference + Must be

$$\text{Better} = (A+O) / (A+O+M+I) \quad (28)$$

$$\text{Worse} = -(O+M)/ (A+O+M+I) \quad (29)$$

SI refers to positive coefficient, and DI refers to negative coefficient (Shokouhyar et al., 2017). A, O, M, I, and R refer to the terms given in Table 40 below.

Table 40: Ranking Factors According to Kano

A	**Attractive**	**Excitement attribute**
M	Must-be	Threshold attribute
O	One-dimensional	Performance attribute
I	Indifferent	Indifferent attribute
R	Reverse	Rejection attribute

According to Mkpojiogu and Hashim (2016) the SI values fall between 0 and 1; the closer the SI to 1, the higher its effect on the satisfaction of the customer, whereas if SI is close to 0, it shows that the specific feature has very little effect on the satisfaction of the customer. The dissatisfaction coefficient (DI) that falls between 0 and -1 must also be considered; if customer satisfaction lies closer to -1, failing to include the feature has a significant effect on the dissatisfaction of the customer (Mkpojiogu & Hashim, 2016). However, if DI is closer to 0, the implication is that the absence of the feature will likely not lead to customer dissatisfaction; the DI always has a negative value (Mkpojiogu & Hashim, 2016).

Based on Table 41, the number of attributes in basic category are 26, and number of attributes in One Dimensional, Attractive, and Reverse are 6,3,1

respectively. Also, it is obvious that all of them are located under one-dimensional category because when they present, they improve customer satisfaction, whereas their absence undermines satisfaction. Based on Table 42, the common attributes between most satisfaction and most dissatisfaction are Code31, code 27, Code25, Code23, and Code12.

Table 41: Kano Category Selection

KANO Category	**Attributes**
Basic	Code2, Code3, Code4, Code5, Code6, Code8, Code9, Code10, Code11, Code14, Code15, Code16, Code17, Code18, Code19, Code20, Code21, Code22, Code24, Code26, Code28, Code29, Code30, Code32, Code33, Code34.
One Dimensional	Code12, Code23, Code25, Code27, Code31, Code35
Attractive	Code7, Code13, Code36
Reverse	Code37

Table 42: Top 10 Features Based On **S**atisfaction and Dissatisfaction

Top 10 Features based on Satisfaction a		**Top 10 Features based on Dissatisfaction**	
Feature	SI	Feature	DI
Code7	0.74	Code24	-0.93
Code13	0.73	Code11	-0.91
Code27	0.71	Code12	-0.90
Code35	0.67	Code29	-0.90
Code25	0.64	Code31	-0.90
Code36	0.59	Code33	-0.90
Code12	0.59	Code23	-0.88
Code23	0.58	Code4	-0.87
Code31	0.57	Code27	-0.87
Code5	0.55	Code25	-0.87

4.7.3 Kano Dataset Clustering Analysis:

As shown in Table 43, in cluster 2, 3, 4, 5 and 7, indifferent attributes are zero, and in cluster 1, and 3, reverse attributes are also zero. Cluster 1 has majority 27 attributes for Basic category, cluster 2 has majority 15 attributes for indifferent, cluster 3 has majority 20 attributes for One Dimensional category, cluster 4 has majority 21 attributes for One Dimensional, and Cluster 5 has majority 31 attributes for Basic category. Cluster 6 has majority 15 attributes for Basic category Cluster 7 has majority 21 attributes for Basic category. Cluster 8 has majority 17 attributes for Basic category So, from the above analysis, it can be concluded that cluster 1, 5, 6, 7, and 8 having majority attributes to Basic Category. But, Cluster 2 has majority attributes to be Attractive and cluster 3 and 4 have majority attributes to be One-Dimensional attributes respectively.

According to the Tables 44, we can notice that all features located under one dimensional and Attractive categories have the highest SI number which supports our research assumption that those features are most related to customer satisfaction if they are present. Figure 36 show the outputs for each feature in related to SSI and DDI value.

No of samples in each cluster:

- Cluster 1 – 59 samples (9.13%)
- Cluster 2 – 145 samples (22.44%)
- Cluster 3 – 65 samples (10.06%)
- Cluster 4 – 101 samples (15.63%)
- Cluster 5 – 83 samples (12.84%)
- Cluster 6 – 113 samples (17.49%)
- Cluster 7 – 15 samples (2.32%)
- Cluster 8 – 65 samples (10.06%)

Table 43: Number of Attributes in Each Category in Each Cluster

Category	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**
Basic	27	0	12	11	31	15	21	17
One Dimensional	6	6	20	21	3	13	6	10
Attractive	3	13	4	2	1	4	5	4
Indifferent	0	15	0	0	0	3	2	0
Reverse	0	2	0	2	1	1	2	5

According to the Table 44, which shows all clusters analysis, we can notice that all features located under one dimensional and Attractive categories have the highest customer satisfaction index (SI) SI number which supports our research

assumption that those features are most related to customer satisfaction if they are present.

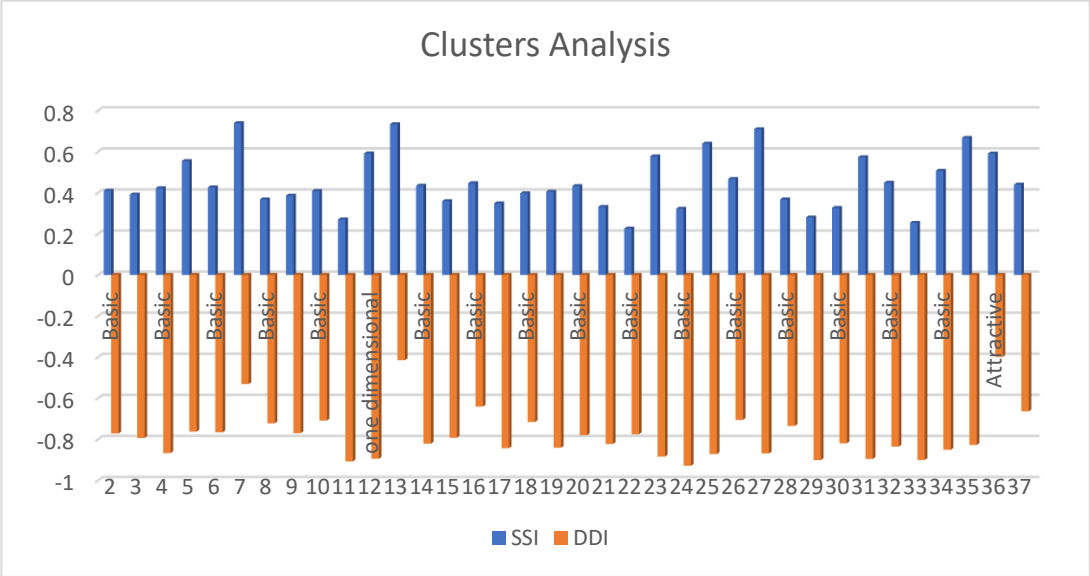


Figure 36: Outputs for Each Feature

Table 44: Data Analysis for Each Feature

Feature	Basic	One dimensional	Excitement	Indifferent	Reverse	Majority	SSI	DDI
2	279	184	62	75	46	Basic	0.410	-0.772
3	297	192	49	78	30	Basic	0.391	-0.794
4	318	220	42	40	26	Basic	0.423	-0.868
5	250	240	116	36	4	Basic	0.555	-0.763
6	284	173	81	58	50	Basic	0.426	-0.767
7	124	194	248	32	48	Attractive	0.739	-0.532
8	301	139	85	84	37	Basic	0.368	-0.723
9	320	147	87	52	40	Basic	0.386	-0.771
10	284	141	104	70	47	Basic	0.409	-0.710
11	425	136	31	26	28	Basic	0.270	-0.908

Table 44: Data Analysis for Each Feature (Continued)

Feature	Basic	One dimensional	Excitement	Indifferent	Reverse	Majority	SSI	DDI
12	241	331	47	20	7	one dimensional	0.592	-0.895
13	136	126	337	32	15	Attractive	0.734	-0.415
14	335	185	90	23	13	Basic	0.434	-0.821
15	352	149	78	53	14	Basic	0.359	-0.793
16	288	108	168	54	28	Basic	0.447	-0.641
17	356	157	55	40	38	Basic	0.349	-0.844
18	255	161	70	95	65	Basic	0.398	-0.716
19	349	176	77	22	22	Basic	0.405	-0.841
20	319	168	102	35	22	Basic	0.433	-0.780
21	373	139	67	43	24	Basic	0.331	-0.823
22	404	91	53	90	8	Basic	0.226	-0.776
23	249	312	54	19	12	one dimensional	0.577	-0.885
24	375	164	23	18	66	Basic	0.322	-0.929
25	192	342	49	29	34	one dimensional	0.639	-0.873
26	268	167	121	60	30	Basic	0.468	-0.706
27	173	381	71	12	9	one dimensional	0.710	-0.870
28	280	120	80	64	102	Basic	0.368	-0.735

Table 44: Data Analysis for Each Feature (Continued)

Feature	Basic	One dimensional	Excitement	Indifferent	Reverse	Majority	SSI	DDI
29	449	128	51	12	6	Basic	0.280	-0.902
30	378	144	64	51	9	Basic	0.327	-0.819
31	247	330	39	28	2	one dimensional	0.573	-0.896
32	319	206	76	27	18	Basic	0.449	-0.836
33	433	132	27	35	19	Basic	0.254	-0.901
34	267	265	52	41	21	Basic	0.507	-0.851
35	168	332	71	33	42	one dimensional	0.667	-0.828
36	162	79	278	85	42	Attractive	0.591	-0.399
37	174	101	81	58	232	reverse	0.440	-0.664

4.8 Deep Learning

Two experiments were performed as deep learning: Multilayer Perception (MLP) and Convolutional Neural Network (CNN). Table 45 to Table 46 have the results of the multilayer perceptron model. It is noticeable that the maximum R-squared value achieved with MLP is 0.549 for all attributes, which is exceptionally low compared to the machine learning prediction techniques with XGB regressor that achieve 0.933. Table 46 shows the results of the convolutional neural network (CNN). The maximum R-squared value is only 0.489, which is also exceptionally low compared to the XGB regressor results. So, this proves that Deep Learning is not suitable for this kind of problem for many reasons. First, the amount of data is small.

Also, it is quite a simple task that requires simple feature engineering and does not require processing unstructured data. Therefore, classical machine learning may be a better choice.

Table 45: Multilayer Perceptron Results

Features	**Train MAE**	**Test MAE**	**Train RMSE**	**Test RMSE**	**Train R2**	**Test R2**
All	0.436	0.517	0.587	0.685	0.685	0.549
Kano	0.598	0.623	0.810	0.839	0.399	0.336
Chi-Square	0.603	0.622	0.791	0.818	0.428	0.367
Mutual Information	0.641	0.662	0.861	0.889	0.321	0.258
Lasso	0.617	0.625	0.803	0.814	0.412	0.374
ANOVA	0.613	0.627	0.803	0.831	0.410	0.345
Pearson	0.620	0.629	0.807	0.819	0.404	0.365
Chi-square common	0.642	0.646	0.867	0.873	0.314	0.278
Mutual Information common	0.672	0.678	0.914	0.919	0.237	0.205
Lasso common	0.639	0.644	0.858	0.866	0.327	0.292
ANOVA common	0.642	0.646	0.867	0.873	0.314	0.278
Pearson common	0.644	0.648	0.869	0.876	0.310	0.273

Table 46: CNN Results

Features	**Train MAE**	**Test MAE**	**Train RMSE**	**Test RMSE**	**Train R2**	**Test R2**
All	0.098	0.112	0.130	0.145	0.615	0.489
Kano	0.141	0.141	0.181	0.183	0.253	0.211
Chi_Square	0.138	0.140	0.176	0.180	0.293	0.234
Mutual_Information	0.140	0.143	0.177	0.180	0.287	0.233
Lasso	0.129	0.131	0.166	0.169	0.374	0.331
ANOVA	0.140	0.141	0.180	0.181	0.263	0.224
Pearson	0.129	0.132	0.164	0.169	0.387	0.325

4.9 Experimental Key Findings

As shown in Figure 37, the cross-validation method at 10 with AdaBoost regressor at depth 15 achieved the highest results with R-square 0.933.

Figure 44 compares all prediction features with extracted features based on feature-selection methods. At cross-validation 10, the XGB regressor at depth 15 with Chi-Square and Lasso feature-selection method through 9 features achieved the closest prediction (0.899) to all features prediction (0.933) using 37 features. Furthermore, the performance of the Kano's model with cross-validation 10 using AdaBoost with depth 15 achieved a value equal to 0.735.

Also, Figure 37 compares all prediction features with extracted features based on common features between higher features of the Kano's model and feature-selection methods. At cross-validation 10, the XGB Regressor with ANOVA and Chi-Square feature-selection method through 4 features achieved the highest prediction

(0.561) in comparison with other feature-selection methods. It was not close to all features prediction value (0.933) using 37 features.

Also, Figure 37 compares all prediction features with extracted features based on union features between higher features of the Kano's model and feature-selection methods. At cross-validation 10, the AdaBoost regressor at depth 10 and XGB regressor depth=10 with the Chi-Square's feature-selection method through 16 features achieved the highest prediction (0.92) in comparison with other feature-selection methods. It was so close to all features prediction value (0.933) using 37 features. Besides, Although the performance of predicting using union features achieved good results, it was not the desired achievement because of the used large number of features.

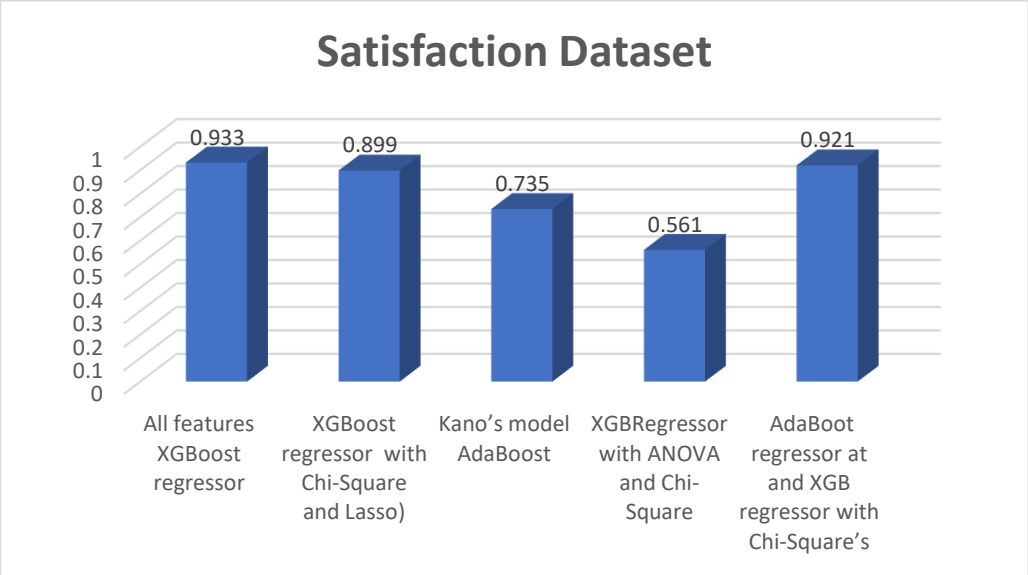


Figure 37: Comparison Between All Features, Selected Features, Kano, Common and Union Features in Terms R2 Based for Satisfaction Dataset

Based on the Table 39, we found that Q2 is the most common attributes between the 8 clusters. Consequently, we considered them important, and it is added to kano model 9 features. Then these 10 features are used to predict the R-square value.

As shown in the Figure 35, there was a high increase in the value of R-square. In contrast, when we added other non-common attributes there was no increase in the R-value. Based on Table 39, R-square for Kano was 73%. After the integration with data, mining Q2 was added with Kano as the most common attributes among 8 clusters, high increase to the value of R-square and reached 86% which is very close to R-value for all attributes which is 93%.

Chapter 5: Conclusion

The main objective of the experiment is the construction of an effective model that can determine student satisfaction at the UAEU university based on 37 services, (i.e., Dormitory, hygiene on the campus, Health services, etc.) provided by the university. Integrating the Kano model with data mining techniques could improve the selection of relevant features that drive customer satisfaction. Different kinds of regression techniques were equipped in the experiment for the purpose of determining the student satisfaction.

At cross validation, the comparison clearly shows that the best model with all attributes is AdaBoost Regression model with depth 15 has R-Square value, RMSE, MAE, and Pearson Correlation Coefficient of 0.93333, 0.19468, 0.0553 and 0.96565 respectively. From this research, it is clear that for common the maximum R-Square value, RMSE, MAE, and Pearson Correlation Coefficient are 0.561, 0.679, 0.397and 0.756 respectively with XGB Regressor with depth 15 using ANOVA or Chi-Square as a features selection method.

According to the results of the integration between the Kano model and the Chi-Square feature selection method (as well as Kano model with ANOVA feature selection method), it was found that food dining hall services, the possibility of communicating with the administration, response to complaints and teaching quality are considered the most important features related to satisfaction. Moreover, the teaching quality feature has been selected through all the selection methods applied, ANOVA, Chi-Square, Lasso, Pearson, Mutual information, which means that it is the most important feature to consider.

It is important to note that the above four features achieved closer results compared to the result gotten when all the attributes for prediction were utilized. This can be considered a very small number of features when compared to the full model with 37 attributes. This reveals that the ANOVA and Chi-Square technique are effective in identifying the aspects of the services that are found to be effectively contributing to the satisfaction of the students.

The clustering approach is highly promising. Using the most common features between the 8 clusters created by k-mean method using the Kano dataset. After that, the best features of each cluster are extracted from the same students in the satisfaction dataset. It is found that Dormitory (Q2) features are the most frequent feature among all clusters when using lasso feature selection method. By integrating Q2 feature to Kano Model 9 features, high increase in the performance of R value. Before the integration, Kano R value was 73% but, after the integration the results achieved 86% which is almost the same as R-value for all attributes.

Also, according to the kano dataset clustering results, we can notice that all features located under one dimensional and Attractive categories have the highest SSI number which supports our research assumption that those features are most related to customer satisfaction if they are present.

In the union experiment, features from the Kano's model and feature-selection methods were taken into consideration. The union result from both methods was used to make a prediction using various prediction techniques. The result of the prediction accuracy metric was higher than that of the common experiment. The Chi-Square's feature-selection techniques reached the best accuracy when integrated with the Kano's model in the union experiment. The AdaBoost Regression and XGB Regressor Depth =10 with Chi-Square through 16 features achieved 0.92 R-square value which

is closer result to all features prediction 0. 933 using 37 features. Moreover, this method resulted in small values for RMSE and MAE.

The flaw in this experiment is a large number of features. The main purpose of this research is to minimize the number of features for business companies improvers. The union experiment highlights the difference between the large and small number of features. It also clarifies the usefulness of the common approach, which achieves acceptable performance as compared to the small number of features also the integration of Kano and the most frequent clusters achieve same result of all attributes with only 10 features.

We have several limitations which are discussed below. First, there are only four common features, which is a small number compared to the total of 37 features used, which were chosen as common features between the kano model and the data mining model. These 4 features' prediction performance is not as good as the performance of all features. But to make the survey easy for the respondents, we expected to have as few features as possible. Therefore, a trade-off exists here between business and data mining. If we want superior data mining accuracy, we need to have more features. However, businesses tend to prefer fewer features so they can concentrate on improving customer pleasure. Second, although deep learning was applied, the results were highly disappointing when compared to traditional machine learning. Also, this kind of research requires datasets that are suitable for customer satisfaction analysis for both approaches: Kano Model and data mining techniques. However, since there is no dataset available from earlier research to satisfy both approaches simultaneously, the intention is to conduct two types of surveys to satisfy both approaches. The drawback of such an approach is that participants may be unwilling to complete the lengthy questioners, resulting in low data collection rates.

Also, this kind of survey is challenging to understand by students. It needs a face-to-face meeting with the students, which was exceedingly difficult because of Covid 19.

Based on the results, the administration team of the institution will be able to effectively make use of the connections to determine the contentment of the students in relation to any changes that are to be made to the characteristics of the institution. There are 646 records that are attributed to small data, and this is one of the very few disadvantages of the experiment. Moreover, the outcomes could be only effective for the institutions that are present within the country.

The key to organizational success relies on the firm's ability to deliver a positive customer experience to retain customers and expand market share. Thus, marketers are encouraged to monitor service delivery and evaluate customer satisfaction levels based on clients reports. Client's feedback based on customer satisfaction level facilitates improving service quality or maintaining the standards of items produced. As documented in several kinds of literature reviewed in this paper, such efforts significantly impact the overall firm's performance and customers loyalty. An institution must significantly invest in software and infrastructure to analyze customer satisfaction levels.

Future work can be directed towards the process of integrating the Kano Model with several other kinds of techniques that can be utilized to improve the process of identifying the attributes. The process of integrating the Kano Model into other methodologies, e.g., data mining and Quality Function Deployment (QFD) which cover the entire development cycle of a product or a process (Hashim & Dawal, 2012). Facilitates comprehensively overcoming each model's limitations, this research study associates the Kano model with evaluating customer satisfaction and contributes significantly to the marketing research theory. Consequently, the results of this study

can play a vital role in streamlining business decision-making in addition to facilitating further scientific research.

References

Abdi, F., & Abolmakarem, S. (2018). Customer Behavior Mining Framework (CBMF) using clustering and classification techniques. Journal of Industrial Engineering International, 15(S1), 1-18. doi: 10.1007/s40092-018-0285-3

Adjimi, A., Hacine-Gharbi, A., Ravier, P., & Mostefai, M. (2019). Mutual Information Based Feature Selection for Fingerprint Identification. Informatica, 43(2). https://doi.org/10.31449/inf.v43i2.1621

Afshar, S. (2015). Sociological Analysis of the Relationship between Social Capital and Cultural Intelligence among the Students of Tabriz University. Journal of Sociological studies, 7(25), 39-57.

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1), 1-24. https://doi.org/10.1186/s40537-019-0191-6

Aktepe, A., Ersöz, S., & Toklu, B. (2015). Customer satisfaction and loyalty analysis with classification algorithms and Structural Equation Modeling. Computers & Industrial Engineering, 86, 95-106. doi: 10.1016/j.cie.2014.09.031

Alnagar, D. K. F. (2020). Using Artificial Neural Network to Predicted Student Satisfaction in E-learning. Am J Appl Math Stat, 8(3), 90-95. doi: 10.12691/ajams-8-3-2

Al Nuaimi, N., & Masud, M. M. (2020). Online streaming feature selection with incremental feature grouping. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(4), e1364. https://doi.org/10.1002/widm.1364

AL-Shatnwai, A. M., & Faris, M. (2020). Predicting Customer Retention using XGB and Balancing Methods. International Journal of Advanced Computer Science and Applications, 11(7), 704-712. doi: 10.14569/IJACSA.2020.0110785

Al Snousy, M. B., El-Deeb, H. M., Badran, K., & Al Khlil, I. A. (2011). Suite of decision tree-based classification algorithms on cancer gene expression data. Egyptian Informatics Journal, 12(2), 73-82. doi:10.1016/j.eij.2011.04.003

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., & Huang, K. (2017). Customer churn prediction in the telecommunication sector using a rough set approach. Neurocomputing, 237, 242-254. doi: 10.1016/j.neucom.2016.12.009

Anderson, E. W., Fornell, C., Lehmann, D. R. (1994). Customer Satisfaction, Market Share, and Profitability: Findings from Sweden. Journal of marketing communications 58:53–66. https://doi.org/10.1177/002224299405800304

Arefi, M., Heidari, M., Morkani, G. S., & Zandi, K. (2012). Application of Kano Model in Higher Education Quality Improvement: Study Master's Degree Program of Educational Psychology in State Universities of Tehran. World Applied Sciences Journal, 17(3), 347-353. doi:10.29015/cerem.464

Arosha Senanayake, S. M. N., & Joshi, S. (2021). A road accident pattern miner (RAP miner). Journal of Information and Telecommunication, 5(4), 484–498. https://doi.org/10.1080/24751839.2021.1955533

Atlason, R. S., Giacalone, D. (2018). Rapid computation and visualization of data from Kano surveys in R. BMC Research Notes 11. Article number: 839 (2018). doi:10.1186/s13104-018-3945-x

Au, S. T., Duan, R., & Jiang, W. (2012). A data mining framework for product and service migration analysis. Annals of Operations Research, 192(1), 105-121. doi:10.1007/s10479-011-0904-5

Avikal, S., Singh, R., & Rashmi, R. (2020). QFD and Fuzzy Kano model-based approach for classification of aesthetic attributes of SUV car profile. Journal of Intelligent Manufacturing, 31(2), 271-284. doi:10.1007/s10845-018-1444-5

Babakus, E., Bienstock, C. C., & Van Scotter, J. R. (2004). Linking Perceived Quality and Customer Satisfaction to Store Traffic and Revenue Growth. Decision Sciences, 35(4), 713-737. https://doi.org/10.1111/j.1540-5915.2004.02671.x

Bailly, A., Blanc, C., Francis, É., Guillotin, T., Jamal, F., Wakim, B., & Roy, P. (2022). Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. Computer Methods and Programs in Biomedicine, 213, 106504. https://doi.org/10.1016/j.cmpb.2021.106504

Baswardono, W., Kurniadi, D., Mulyani, A., & Arifin, D. M. (2019, December). Comparative analysis of decision tree algorithms: Random Forest and C4. 5 for airlines customer satisfaction classification. In Journal of Physics: Conference Series (Vol. 1402, No. 6, p. 066055). 213-222. IOP Publishing. http://doi.org/10.22441/sinergi.2022.2.010

Bell, D., & Mgbemena, C. (2017). Data-driven agent-based exploration of customer behavior. Simulation 94:195–212. https://doi.org/10.1177/0037549717743106

Belyadi, H., & Haghighat, A. (2021). Machine learning guide for oil and gas using python: a step-by-step breakdown with data algorithms codes and applications. Gulf Professional Publishing. Retrieved January 12 2023 from https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=2643527

Bengfort, B. (2020). Recursive Feature Elimination — Yellowbrick v1.3.post1 documentation. Retrieved 7 March 2020, from https://www.scikit-yb.org/en/latest/api/model_selection/rfecv.html#:~:text=Recursive%20feature%20elimination%20(RFE)%20is,number%20of%20features%20is%20reached

Berger, C., Blauth, R., Bolster, C., Burchill, G., DuMouchel, W., Pouliot, F., Richter, R., Rubinoff, A., Shen, D., Timko, M., & Walden, D. (1993). Kano's methods for understanding customer-defined quality. The Center for Quality Management Journal, Vol. 2, No. 4, 3-36

Bisong, E. (2019). Building machine learning and deep learning models on Google cloud platform. 59-64. Berkeley, CA: Apress.

Bjertnaes, O., Skudal, K. E., & Iversen, H. H. (2013). Classification of patients based on their evaluation of hospital outcomes: cluster analysis following a national survey in Norway. BMC Health Services Research 13, Article number: 73 (2013). https://doi.org/10.1186/1472-6963-13-73

Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: a review. Journal of Data Analysis and Information Processing, 8(4), 341-357. doi:10.4236/jdaip.2020.84020

Bohacik, J. (2014). Decision stump for prognosis of mortality rates in heart failure patients. Journal of Information Technologies. 7. 8-14. doi: 10.1007/978-3-030-43192-1_96

Boodhun, N., & Jayabalan, M. (2018). Risk prediction in life insurance industry using supervised learning algorithms. Complex & Intelligent Systems, 4(2), 145-154. https://doi.org/10.1007/s40747-018-0072-1

Botchkarev, A. (2018). Evaluating Performance of Regression Machine Learning Models Using Multiple Error Metrics in Azure Machine Learning Studio. Available at SSRN 3177507. doi:10.2139/ssrn.3177507

Botchkarev, A. (2019). A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms. Interdisciplinary Journal of Information, Knowledge, And Management, 14, 045-076. https://doi.org/10.28945/4184

Busacca, B., & Padula, G. (2005). Understanding the relationship between attribute performance and overall satisfaction: Theory, measurement and implications. Marketing Intelligence & Planning. doi:10.1108/02634500510624110

Buyya, R., Calheiros, R., & Dastjerdi, A. V. (Eds.). (2016). Big data: Principles and Paradigms. Morgan Kaufmann.

Caballero, D. (2017). Feature extraction algorithms from MRI to evaluate quality parameters on meat products by using data mining. ELCVIA: electronic letters on computer vision and image analysis, 16(2), 1-4. doi:10.5565/rev/elcvia.1100.

Chang, C. C., Chen, P. L., Chiu, F. R., & Chen, Y. K. (2009). Application of neural networks and Kano's method to content recommendation in web Pearsonalization. Expert Systems with Applications, 36(3), 5310-5316. doi: 10.1016/j.eswa.2008.06.067

Chapron, K., Plantevin, V., Thullier, F., Bouchard, K., Duchesne, E., & Gaboury, S. (2018). A more efficient transportable and scalable system for real-time activities and exercises recognition. Sensors, 18(1), 268. https://doi.org/10.3390/s18010268

Chen, K. K., Chiu, R. H., & Chang, C. T. (2017). Using beta regression to explore the relationship between service attributes and likelihood of customer retention for the container shipping industry. Transportation Research Part E: Logistics and Transportation Review, 104, 1-16. doi: 10.1016/j.tre.2017.04.015

Chen, L. F. (2012). A novel approach to regression analysis for the classification of quality attributes in the Kano model: an empirical test in the food and beverage industry. Omega, 40(5), 651-659. https://doi.org/10.1016/j.omega.2011.12.004

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. Journal of Big Data, 7(1), 1-26. https://doi.org/10.1186/s40537-020-00327-4

Cheng, Y. S., Kuo, N. T., Chang, K. C., & Hu, S. M. (2019). Integrating the Kano model and IPA to measure quality of museum interpretation service: a comparison of visitors from Taiwan and Mainland China. Asia Pacific Journal of Tourism Research, 24(6), 483-500. https://doi.org/10.1080/10941665.2019.1591474

Choi, J., Seol, H., Lee, S., Cho, H., & Park, Y. (2008). Customer satisfaction factors of mobile commerce in Korea. Internet research. 18(3):313-335 https://doi.org/10.1108/10662240810883335

Chou, J. S., Chiu, C. K., Farfoura, M., & Al-Taharwa, I. (2011). Optimizing the Prediction Accuracy of Concrete Compressive Strength Based on a Comparison of Data-Mining Techniques. Journal of Computing in Civil Engineering, 25(3), 242-253. doi:10.1061/(ASCE)CP.1943-5487.0000088

Christa, S., Suma, V., & Mohan, U. (2022). Regression and decision tree approaches in predicting the effort in resolving incidents. International Journal of Business Information Systems, 39(3), 379-399. https://doi.org/10.1504/IJBIS.2022.122342

Dasu, T., & Johnson, T. (2003). Exploratory data mining and data cleaning. John Wiley & Sons.

De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research, 269(2), 760-772. https://doi.org/10.1016/j.ejor.2018.02.009

Dong, W., Huang, Y., Lehane, B., & Ma, G. (2020). XGB algorithm-based prediction of concrete electrical resistivity for structural health monitoring. Automation in Construction, 114, 103155. doi:10.1016/j.autcon.2020.103155

Doshi, M., & Chaturvedi, S. K. (2014). Correlation based Feature Selection (CFS) Technique to Predict Student Perfromance. International Journal of Computer Networks & Communications, 6(3), 197. doi:10.5121/ijcnc.2014.6315

Drotár, P., Gazda, J., & Smékal, Z. (2015). An experimental comparison of feature selection methods on two-class biomedical datasets. Computers in biology and medicine, 66, 1-10. https://doi.org/10.1016/j.compbiomed.2015.08.010

Du, A. X., Emam, S., & Gniadecki, R. (2020). Review of Machine Learning in Predicting Dermatological Outcomes. Frontiers in Medicine, 7, 266. doi:10.3389/fmed.2020.00266

Eberly, L. E. (2007). Multiple Linear Regression. Topics in Biostatistics, 165-187. https://doi.org/10.1007/978-1-59745-530-5_9

El Aissaoui, O., El Alami El Madani, Y., Oughdir, L., & El Allioui, Y. (2018). A fuzzy classification approach for learning style prediction based on web mining technique in e-learning environments. Education And Information Technologies, 24(3), 1943-1959. https://doi.org/10.1007/s10639-018-9820-5

Eid, R. (Ed.). (2013). Managing customer trust, satisfaction, and loyalty through information communication technologies. IGI Global. DOI:10.4018/978-1-4666-3631-6

Elomaa, T., & Kaariainen, M. (2001). An Analysis of Reduced Error Pruning. Journal of Artificial Intelligence Research, 15, 163-187. https://doi.org/10.1613/jair.816

Farhadloo, M., Patterson, RA., Rolland, E. (2016). Modeling customer satisfaction from unstructured data using a Bayesian approach. Decision Support Systems, Elsevier B.V. 90:1–11. https://doi.org/10.1016/j.dss.2016.06.010

Ferreira, J. C., Monteiro, V. D. F., & Afonso, J. L. (2012). Data mining approach for range prediction of electric vehicle.

Fida, M. A. F. A., Ahmad, T., & Ntahobari, M. (2021, October). Variance Threshold as Early Screening to Boruta Feature Selection for Intrusion Detection System. In 2021 13th International Conference on Information & Communication Technology and System (ICTS). 46-50. IEEE.

Fonti, V., & Belitser, E. (2017). Feature selection using lasso. VU Amsterdam research paper in business analytics, 30, 1-25.

Gacto, M., Soto-Hidalgo, J. M., Alcala-Fdez, J., & Alcala, R. (2019). Experimental Study on 164 Algorithms Available in Software Tools for Solving Standard Non-Linear Regression Problems. IEEE Access, 7, 108916-108939. doi: 10.1109/access.2019.2933261

Galimberti, G., & Soffritti, G. (2011). Tree-Based Methods and Decision Trees. Modern Analysis of Customer Surveys: With Applications Using R, 283-307. https://doi.org/10.1002/9781119961154.ch15

Gao, Y., Rasouli, S., Timmermans, H., & Wang, Y. (2018). Trip stage satisfaction of public transport users: A reference-based model incorporating trip attributes, perceived service quality, psychological disposition and difference tolerance. Transportation Research Part A: Policy and Practice, 118, 759-775. https://doi.org/10.1016/j.tra.2018.10.029

Geler, Z., Savić, M., Bratić, B., Kurbalija, V., Ivanović, M., & Dai, W. (2021). Sentiment prediction based on analysis of customers assessments in food serving businesses. Connection Science, 33(3), 674-692. https://doi.org/10.1080/09540091.2020.1870436

Gokhan, A. K. S. U., & Keceoglu, C. R. (2019). Comparison of results obtained from logistic regression, CHAID analysis and decision tree methods. Eurasian Journal of Educational Research, 19(84), 115-134.

Gómez Fernández, D., Salas López, R., Rojas Briceño, N. B., Silva López, J. O., & Oliva, M. (2022). Dynamics of the Burlan and Pomacochas Lakes Using SAR Data in GEE, Machine Learning Classifiers, and Regression Methods. ISPRS International Journal of Geo-Information, 11(11), 534. https://doi.org/10.3390/ijgi11110534

Grömping, U. (2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest. The American Statistician, 63(4), 308-319. https://doi.org/10.1198/tast.2009.08199

Gu, Q., Li, Z., & Han, J. (2012). Generalized fisher score for feature selection. arXiv preprint arXiv:1202.3725. https://doi.org/10.48550/arXiv.1202.3725

Gupta, M., & Shri, C. (2018). Understanding customer requirements of corrugated industry using Kano model. International Journal of Quality & Reliability Management, 35(8), 1653-1670. doi: 10.1108/ijqrm-04-2017-0074

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier. https://doi.org/10.1016/C2009-0-61819-5

Hand, D., Manilla, H., & Smyth, P. (2007). Principles of Data Mining. Drug safety, 30(7), 621-622.

Hashim, A. M., & Dawal, S. Z. M. (2012). Kano Model and QFD integration approach for Ergonomic Design Improvement. Procedia Social and Behavioral Sciences, Elsevier B.V. 57:22–32. https://doi.org/10.1016/j.sbspro.2012.09.1153

Hassan, M. (2018). Customer Profiling and Segmentation in Retail Banks Using Data Mining Techniques. International Journal of Advanced Research in Computer Science, 9(4), 24-29. doi: 10.26483/ijarcs.v9i4.6172

Hassan, M. M. T. M., & Tabasum, M. (2018). Customer profiling and segmentation in retail banks using data mining techniques. International journal of advanced research in computer science, 9(4), 24-29. https://doi.org/10.26483/ijarcs.v9i4.6172

Hazra, T. K., Sarkar, R., & Kumar, A. (2016). Handwritten English Character Recognition Using Logistic Regression and Neural Network. International Journal of Science and Research, 5, 750-754.

He, X., Ai, X., Jing, Y., & Liu, Y. (2016). Partner selection of agricultural products supply chain based on data mining. Concurrency and Computation: Practice and Experience, 28(4), 1246-1256. https://doi.org/10.1002/cpe.3625

Hira, Z. M., & Gillies, D. F. (2015). A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. Advances in bioinformatics, 2015. DOI:10.1155/2015/198363

Hota, L., & Dash, P. K. (2021). Prediction of Customer Churn in Telecom Industry: A Machine Learning Perspective. doi:10.36647/ciml/02.02.a001

Hu, W. H., Tang, S. H., Chen, Y. C., Yu, C. H., & Hsu, W. C. (2018, July). Promotion recommendation method and system based on random forest. In Proceedings of the 5th Multidisciplinary International Social Networks Conference. 1-5. https://doi.org/10.1145/3227696.3227708

Huiskonen, J., & Pirttilä, T. (1998). Sharpening logistics customer service strategy planning by applying Kano's quality element classification. International journal of production economics, 56, 253-260. doi:10.1016/S0925-5273(97)00065-0

Hung, H., Jou, Z. Y., & Huang, S. Y. (2018). Robust mislabel logistic regression without modeling mislabel probabilities. Biometrics, 74(1), 145-154. doi: 10.1111/biom.12726

Huang, S. H. (2015). Supervised feature selection: A tutorial. Artif. Intell. Res., 4(2), 22-37.

Iannace, G., Ciaburro, G., & Trematerra, A. (2019). Wind Turbine Noise Prediction Using Random Forest Regression. Machines, 7(4), 69. https://doi.org/10.3390/machines7040069

Idris, A., & Khan, A. (2017). Churn Prediction System for Telecom using Filter–Wrapper and Ensemble Classification. The Computer Journal, bxv123. doi: 10.1093/comjnl/bxv123

Ingaldi, M., & Ulewicz, R. (2019). How to Make E-Commerce More Successful by Use of Kano's Model to Assess Customer Satisfaction in Terms of Sustainable Development. Sustainability, 11(18), 4830. doi: 10.3390/su11184830

Ittner, C. D., & Larcker, D. F. (1998). Are Nonfinancial Measures Leading Indicators of Financial Performance? An analysis of Customer Satisfaction. Journal of Accounting Research 36:1-35. https://doi.org/10.2307/2491304

Jadhav, D. A. (2021). An enhanced and secured predictive model of Ada-Boost and Random-Forest techniques in HCV detections. Materials Today: Proceedings. https://doi.org/10.1016/j.matpr.2021.05.071

Jahanshahi, A. A., Gashti, M. A. H., Mirdamadi, S. A., Nawaser, K., & Khaksar, S. M. S. (2011). Study the effects of customer service and product quality on customer satisfaction and loyalty. International Journal of Humanities and Social Science, 1(7), 253-260. https://doi.org/10.20431/2349-0349.0902004

Jangaraj, A., Sunitha, G., Madhavi, R. K., & Kora, P. (2021). Prediction of house price using XGB regression algorithm. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(2), 2151-2155. doi**:** https://doi.org/10.17762/turcomat.v12i2.1870

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. Electronic Markets, 31(3), 685-695. doi: 10.1007/s12525-021-00475-2

Johnson, M. D., Herrmann, A., & Gustafsson, A. (2002). Comparing customer satisfaction across industries and countries. Journal of Economic Psychology, 23(6), 749-769. doi: 10.1016/s0167-4870(02)00137-x.

Joshi, N., Katyayan, P., & Ahmed, S. A. (2021, April). Comparing Classical ML Models with Quantum ML Models with Parametrized Circuits dor Sentiment Analysis Task. In Journal of Physics: Conference Series (Vol. 1854, No. 1), 012032. Doi: 10.1088/1742-6596/1854/1/012032

Jović, A., Brkić, K., & Bogunović, N. (2015, May). A review of feature selection methods with applications. In 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO) 1200-1205. IEEE.

Južnik Rotar, L., & Kozar, M. (2017). The Use of the Kano Model to Enhance Customer Satisfaction. Organizacija, 50(4), 339-351. doi: 10.1515/orga-2017-0025

Kano, N., Seraku, N., Takahahsi, F., & Tsuji, S. (1984). Attractive Quality and Must-Be Quality. Hinshitsu (Quality, The Journal of Japanese Society for Quality Control), 14, 39-48.

Kaya, E., Dong, X., Suhara, Y., Balcisoy, S., Bozkaya, B., & Pentland, A. (2018). Behavioral attributes and financial churn prediction. EPJ Data Science, 7(1). doi: 10.1140/epjds/s13688-018-0165-5

Kazemi, M., Estemdad, N., & Poorebrahimi, A. (2015). Customer Behavior in Banking Industry: Comparison of Data Mining Techniques. Journal of Advanced Management Science, 13-16. doi: 10.12720/joams.3.1.13-16

Koduri, S. B., Gunisetti, L., Ramesh, C. R., Mutyalu, K. V., & Ganesh, D. (2019, May). Prediction of crop production using AdaBoost regression method. In Journal of Physics: Conference Series (Vol. 1228, No. 1), 012005. IOP Publishing.

Kratsios, A., & Hyndman, C. (2021). Neu: A Meta-Algorithm for Universal UAP-Invariant Feature Representation. Journal of Machine Learning Research, 22, 92. https://doi.org/10.48550/arXiv.1809.00082

Krishnamurthy, A. (2011). High-dimensional clustering with sparse gaussian mixture models. Unpublished paper, 191-192.

Kwak, C., & Clayton-Matthews, A. (2002). Multinomial logistic regression. Nursing research, 51(6), 404-410.

Lameski, P., Zdravevski, E., Mingov, R., & Kulakov, A. (2015). SVM Parameter Tuning with Grid Search and Its Impact on Reduction of Model Over-Fitting. In Rough sets, fuzzy sets, data mining, and granular computing. 464-474. Springer, Cham.

Lee, Y. C., & Huang, S. Y. (2009). A new fuzzy concept approach for Kano's model. Expert Systems with Applications, Elsevier Ltd 36:4479–4484. https://doi.org/10.1016/j.eswa.2008.05.034

Leonard, L. C. (2017). Web-based Behavioral Modeling for Continuous User Authentication (CUA). In Advances in Computers (Vol. 105). 1-44. Elsevier. https://doi.org/10.1016/bs.adcom.2016.12.001

Li, S., & Zhang, X. (2019). Research on orthopedic auxiliary classification and prediction model based on XGB algorithm. Neural Computing and Applications, 32(7), 1971-1979. doi:10.1007/s00521-019-04378-4

Li, Y. H. (2017). Text feature selection algorithm based on Chi-Square rank correlation factorization. Journal of Interdisciplinary Mathematics, 20(1), 153-160. https://doi.org/10.1080/09720502.2016.1259769

Li, Y. F., Xie, M., & Goh, T. N. (2009). A study of mutual information-based feature selection for case-based reasoning in software cost estimation. Expert Systems with Applications, 36(3), 5921-5931. https://doi.org/10.1016/j.eswa.2008.07.062

Lin, Z., & Vlachos, I. (2018). An advanced analytical framework for improving customer satisfaction: A case of air passengers. Transportation Research Part E: Logistics and Transportation Review, 114, 185-195. doi: 10.1016/j.tre.2018.04.003

Liu, H., & Motoda, H. (2012). Feature selection for knowledge discovery and data mining (Vol. 454). Springer Science & Business Media.

Liu, X., Xu, Y., Montes, R., Ding, R., & Herrera, F. (2019). Alternative Ranking-Based Clustering and Reliability Index-Based Consensus Reaching Process for Hesitant Fuzzy Large Scale Group Decision Making. IEEE Transactions on Fuzzy Systems, 27(1), 159-171. https://doi.org/10.1109/tfuzz.2018.2876655

Luiz de Freitas Vieira, J., & Almeida Có, M. (1997). https://sobraep.org.br/artigo/sistema-retificador-inversor-com-corrente-pulsada-no-barramento-cc-para-acionamento-de-motores-de-inducao/. Eletrônica De Potência, 2(1), 35-42. doi: 10.18618/rep.1997.1.035042

Luo, H., Cheng, F., Yu, H., & Yi, Y. (2021). SDTR: Soft Decision Tree Regressor for Tabular Data. IEEE Access, 9, 55999-56011. Digital Object Identifier 10.1109/ACCESS.2021.3070575

Lu, H., & Yuan, J. (2018). Student Performance Prediction Model Based on Discriminative Feature Selection. International Journal of Emerging Technologies in Learning (iJET), 13(10):55. doi: 10.3991/ijet.v13i10.9451

Madzík, P., Budaj, P., Mikuláš, D., & Zimon, D. (2019). Application of the Kano Model for a Better Understanding of Customer Requirements in Higher Education—A Pilot Study. Administrative Sciences, 9(1), 11. doi: 10.3390/admsci9010011

Magnini, V. P., Honeycutt Jr. E. D., & Hodge, S. K. (2003). Data mining for hotel firms: use and limitations. Cornell Hotel Restaur Admin Q 44:94-105. https://doi.org/10.1016/S0010-8804(03)90022-X

Martilla, J. A., & James, J. C. (1977). Importance-Performance Analysis. Journal of marketing, 41(1), 77-79. https://doi.org/10.2307/1250495

Matzler, K., Hinterhuber, H. H., Bailom, F., & Sauerwein, E. (1996). How to delight your customers. Journal of Product & Brand Management 5:6-18. https://doi.org/10.1108/10610429610119469.

Matzler, K., & Hinterhuber, H. (1998). How to make product development projects more successful by integrating Kano's model of customer satisfaction into quality function deployment. Technovation, 18(1), 25-38. doi: 10.1016/s0166-4972(97)00072-2

McHugh, M. L. (2013). The chi-square test of independence. Biochemia medica, 23(2), 143-149. doi: 10.11613/BM.2013.018

Meng, Q., Wei, X., & Meng, W. (2016). A Decision Method to Maximize Service Quality under Budget Constraints: The Kano Study of a Chinese Machinery Manufacturer. Sci Program 2016:1-12. https://doi.org/10.1155/2016/7291582

Meyer, C., & Schwager, A. (2007). Understanding customer experience. Harvard business review, 85(2), 116.

Mikulić, J., & Prebežac, D. (2011). A critical review of techniques for classifying quality attributes in the Kano model. Managing Service Quality: An International Journal 21:46–66. https://doi.org/10.1108/09604521111100243

Mittal, M., Goyal, L. M., Hemanth, D. J., & Sethi, J. K. (2019). Clustering approaches for high-dimensional databases: A review. Wires Data Mining and Knowledge Discovery, 9(3). https://doi.org/10.1002/widm.1300

Mkpojiogu, E. O. C., & Hashim, N. L. (2016). Understanding the relationship between Kano model's customer satisfaction scores and self-stated requirements importance. Springerplus, 5(1). https://doi.org/10.1186/s40064-016-1860-y

Mostert, W., Malan, K. M., & Engelbrecht, A. P. (2021). A Feature Selection Algorithm Performance Metric for Comparative Analysis. Algorithms, 14(3), 100. doi: 10.3390/a14030100

Nguyen, L. T. K., Chung, H. H., Tuliao, K. V., & Lin, T. M. Y. (2020). Using XGBoost and skip-gram model to predict online review popularity. SAGE Open, 10(4), 2158244020983316. https://doi.org/10.1177/2158244020983316

Novikov, A. V. (2019). PyClustering: Data Mining Library. Journal of Open-Source Software, 4(36), 1230. https://doi.org/10.21105/joss.01230

Okazaki, S., Díaz-Martín, A. M., Rozano, M., Menéndez-Benito, H. D. (2015). Using Twitter to engage with customers: a data mining approach. Internet Research 25:416–434. https://doi.org/10.1108/IntR-11-2013-0249

Olsen, L. L., Witell, L., & Gustafsson, A. (2014). Turning customer satisfaction measurements into action. Journal of Service Management. DOI:10.1108/JOSM-01-2014-0025

O'Shea, K., & Nash, R. (2015). An Introduction to Convolutional Neural Networks. arXiv preprint arXiv:1511.08458.https://doi.org/10.48550/arXiv.1511.08458

Othman, R., Belkaroui, R., & Faiz, R. (2017). Extracting Product Features for Opinion Mining Using Public Conversations in Twitter. Procedia Computer Science, 112, 927-935. doi: 10.1016/j.procs.2017.08.122

Pandey, N., Patnaik, P. K., & Gupta, S. (2020). Data pre-processing for machine learning models using python libraries. International Journal of Engineering and Advanced Technology, 9, 1995-1999.

Park, B. U., Mammen, E., Lee, Y. K., & Lee, E. R. (2013). Varying Coefficient Regression Models: A Review and New Developments. International Statistical Review, 83(1), 36-64. https://doi.org/10.1111/insr.12029

Park, J., & Oh, J. S. (2018). Treatment Effects Analysis on Credit Guarantee using Survey Questionnaire. Journal Of Industrial Economics and Business, 31(3), 1079-1112. https://doi.org/10.22558/jieb.2018.06.31.3.1079

Pascal Ezenkwu, C., Ozuomba, S., & kalu, C. (2015). Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services. International Journal of Advanced Research in Artificial Intelligence, 4(10). doi: 10.14569/ijarai.2015.041007

Pavya, K., & Srinivasan, B. (2018). Review of Literature on Filter and Wrapper Methods for Feature Selection. International Journal of Engineering Sciences & Research Technology, 7(1). doi: 10.5281/zenodo.1135878

Pekel, E. (2020). Estimation of soil moisture using decision tree regression. Theoretical and Applied Climatology, 139(3), 1111-1119. doi:10.1007/s00704-019-03048-8

Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning 1:1, 81-106. http://dx.doi.org/10.1007/BF00116251

Ramaswami, M., & Bhaskaran, R. (2009). A study on feature selection techniques in educational data mining. arXiv preprint arXiv:0912.3924. https://doi.org/10.48550/arXiv.0912.3924

Raschka, S., Patterson, J., & Nolet, C. (2020). Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence. Information, 11(4), 193. https://doi.org/10.3390/info11040193

Redfern, R., & Davey, C. L. (2003). Supply chain market orientation in new product development in the UK: A pilot case study. Journal of Fashion Marketing and Management: An International Journal.

Rust, R. T., & Zahorik, A. J. (1993). Customer satisfaction, customer retention, and market share. Journal of retailing, 69(2), 193-215. https://doi.org/10.1016/0022-4359(93)90003-2

Sabbeh, S. F. (2018). Machine-Learning Techniques for Customer Retention: A Comparative Study. International Journal of Advanced Computer Science and Applications, 9(2). doi: 10.14569/IJACSA.2018.090238

Sadewo, H., Satria, Y., & Burhan, H. (2021). Application of Mean Shift Clustering to optimize matching problems in ridesharing for maximize the total number of matches. Journal of Physics: Conference Series, 1821(1), 012019. https://doi.org/10.1088/1742-6596/1821/1/012019

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1249. https://doi.org/10.1002/widm.1249

Sahin, E. K. (2020). Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGB, gradient boosting machine, and random forest. SN Applied Sciences, 2(7). https://doi.org/10.1007/s42452-020-3060-1

Sauerwein, E., Bailom, F., Matzler, K., & Hinterhuber, H. H. (1996). The Kano model: how to delight your customers. In: International working seminar on production economics (Vol. 1, No. 4). 313-327

Schüller, D., & Pekárek, J. (2018). Customer Satisfaction Measurement - Clustering Approach. Acta Universitatis Agriculturae Et Silviculturae Mendelianae Brunensis, 66(2), 561-569. https://doi.org/10.11118/actaun201866020561

Seufert, E. B. (2013). Freemium economics: Leveraging analytics and user segmentation to drive revenue. Elsevier.

Shahin, A., & Akasheh, S. (2017). Classifying customer requirements using Kano model and Kano map: the case of hospital services. International Journal of Productivity and Quality Management, 21(4), 500. doi:10.1504/ijpqm.2017.10005846

Shahraki, A., Abbasi, M., & Haugen, Ø. (2020). Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. Engineering Applications of Artificial Intelligence, 94, 103770. https://doi.org/10.1016/j.engappai.2020.103770

Sheena, Kumar, K., & Kumar, G. (2016). Analysis of Feature Selection Techniques: A Data Mining Approach. IJCA Proceedings on International Conference on Advances in Emerging Technology ICAET.

Shen, X. X., Tan, K. C. and Xie, M. 2000. An integrative approach to innovative product development using Kano's model and QFD. European Journal of Innovation Management, 3(2): 91–99.

Shokouhyar, S., Tavallaee, R., Mohsenian, F., & Nia, M. S. (2017). Assessing customer satisfaction in food industry with fuzzy Kano approach: a case study in food company. International Journal of Business and Management 9:189–200. doi:10.7537/marsrsj071215.03.

Shrestha, D. L., & Solomatine, D. P. (2006). Experiments with AdaBoost.RT, an Improved Boosting Scheme for Regression. Neural Computation, 18(7), 1678-1710. https://doi.org/10.1162/neco.2006.18.7.1678

Sukarsa, I. M., Pinata, N. N. P., Rusjayanthi, N. K. D., & Wisswani, N. W., (2021). Estimation of gourami supplies using gradient boosting decision tree method of XGB. TEM Journal, 10(1), 144-151. https://doi.org/10.18421/tem101-17

Sulistiani, H., Muludi, K., & Syarif, A. (2019, October). Implementation of Dynamic Mutual Information and Support Vector Machine for Customer Loyalty Classification. In Journal of Physics: Conference Series. IOP Publishing.Conference Series, vol. 1338, no. 1, 012050. doi:10.1088/1742-6596/1338/1/012050

Suppers, A., Van Gool, A. J., & Wessels, H. J. C. T. (2018). Integrated Chemometrics and Statistics to Drive Successful Proteomics Biomarker Discovery. Proteomes. 6. 20. 10.3390/proteomes6020020. doi:10.3390/proteomes6020020

Tama, B. A. (2015). Data mining for predicting customer satisfaction in Fast-food restaurant. Journal of Theoretical and Applied Information Technology 75

Tang, M., & Liao, H. (2021). From conventional group decision making to large-scale group decision making: What are the challenges and how to meet them in big data era? A state-of-the-art survey. Omega, 100, 102141. https://doi.org/10.1016/j.omega.2019.102141

Tariq, M. U., Babar, M., Poulin, M., & Khattak, A. S. (2021). Distributed model for customer churn prediction using convolutional neural network. Journal of Modelling in Management, (ahead-of-print). https://doi.org/10.1108/JM2-01-2021-0032

Taylor, S. A., & Baker, T. L. (1994). An assessment of the relationship between service quality and customer satisfaction in the formation of consumers' purchase intentions. Journal of retailing, 70(2), 163-178. https://doi.org/10.1016/0022-4359(94)90013-2

Tsami, M., Adamos, G., Nathanail, E., Budiloviča, E. B., Jackiva, I. Y., & Magginas, V. (2018). A decision tree approach for achieving high customer satisfaction at urban interchanges. Transport and Telecommunication, 19(3), 194. doi: https://doi.org/10.2478/ttj-2018-0016

Tirenni, G., Kaiser, C., & Herrmann, A. (2007). Applying decision trees for value-based customer relations management: Predicting airline customers' future values. Journal of Database Marketing & Customer Strategy Management, 14(2), 130-142. https://doi.org/10.1057/palgrave.dbm.3250044

Tontini, G. (2007). Integrating the Kano Model and QFD for Designing New Products. Total Quality Management & Business Excellence, 18(6), 599-612. doi: 10.1080/14783360701349351

Turisová, R. (2015). A generalization of traditional Kano model for customer requirements analysis. Quality Innovation Prosperity, 19(1), 59-73. doi:10.12776/qip.v19i1.407

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. IEEE access, 7, 60134-60149. doi: 10.1109/ACCESS.2019.2914999

Uyanık, G. K., & Güler, N. (2013). A Study on Multiple Linear Regression Analysis. Procedia-Social and Behavioral Sciences, 106, 234-240. https://doi.org/10.1016/j.sbspro.2013.12.027

Violante, M. G., & Vezzetti, E. (2017). Kano qualitative vs quantitative approaches: an assessment framework for products attributes analysis. Computers in Industry, Elsevier B.V. 86:15–25. https://doi.org/10.1016/j.compind.2016.12.007

Wang, Y., Fan, X., & Cai, Y. (2014). A comparative study of improvements Pre-filter methods bring on feature selection using microarray data. Health Information Science and Systems, 2(1), 1-9. doi: 10.1186/2047-2501-2-7

Wang, H., & Li, Z. S. (2021). An AdaBoost-based tree augmented naive Bayesian classifier for transient stability assessment of power systems. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, Vol 236(3), https://doi.org/10.1177/1748006X211047308

Wardy, W., Mena, B., Nongtaodum, S., No, H. K., & Prinyawiwatkul, W. (2014). Exploring the Drivers of Purchase Intent and Consumer Satisfaction of Chicken Eggs Using Principal Component Analysis and the Kano Model. Journal Of Sensory Studies, 29(6), 463-473. doi: 10.1111/joss.12127

Witell, L., Löfgren, M., & Dahlgaard, J. J. (2013). Theory of attractive quality and the Kano methodology–the past, the present, and the future. Total Quality Management & Business Excellence, 24(11-12), 1241-1252. doi:10.1080/14783363.2013.791117

Wu, Z., Jing, L., Wu, B., & Jin, L. (2022). A PCA-AdaBoost model for E-commerce customer churn prediction. Annals of Operations Research, 1-18. doi:10.1007/s10479-022-04526-5

Xiao, J., Xiao, Y., Huang, A., Liu, D., & Wang, S. (2015). Feature-selection-based dynamic transfer ensemble model for customer churn prediction. Knowledge and information systems, 43(1), 29-51. doi:10.1007/s10115-013-0722-y

Xiong, Y., Li, Q., Wang, P., & Ye, M. (2021). Informative Gene Selection Based on Cost-Sensitive Fast Correlation-Based Feature Selection. Current Bioinformatics, 16. https://doi.org/10.2174/1574893616666210601111850

Xu, Q., Jiao, R. J., Yang, X., Helander, M., Khalid, H. M., & Opperud, A. (2009). An analytical Kano model for customer need analysis. Design Studies, Elsevier Ltd 30:87-110. https://doi.org/10.1016/j.destud.2008.07.001

Yanan, W., Jiekang, Wu., & Xioaming, M. (2018). Intelligent Scheduling Optimization of Seasonal CCHP System Using Rolling Horizon Hybrid Optimization Algorithm and Matrix Model Framework. IEEE Access, Vol 6, 75132 – 75142. doi: 10.1109/ACCESS.2018.2878044

Yang, J., Liu, K., Zhang, M., Xu, Q., & Jin, P. (2021). Research and Application of Business Ability Evaluation Based on DBSCAN Algorithm and Entropy Method. Journal of Physics: Conference Series, 1881(3), 032064. https://doi.org/10.1088/1742-6596/1881/3/032064

Yi, H. S., Lee, B., Park, S., Kwak, K. C., & An, K. G. (2019). Prediction of short-term algal bloom using the M5P model tree and extreme learning machine. Environmental Engineering Research, 24(3), 404–411. https://doi.org/10.4491/eer.2018.245

Yu, S. H. (2007). An Empirical Investigation on the Economic Consequences of Customer Satisfaction. Total Quality Management and Business Excellence 18:555–569. https://doi.org/10.1080/14783360701240493

Zeinalizadeh, N., Shojaie, A. A., & Shariatmadari, M. (2015). Modeling and analysis of bank customer satisfaction using neural networks approach. International Journal of Bank Marketing, Vol. 33 No. 6, pp. 717-732. https://doi.org/10.1108/IJBM-06-2014-0070

Zhang, M. (2019). Use Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Algorithm to Identify Galaxy Cluster Members. IOP Conference Series: Earth and Environmental Science, 252, 042033. https://doi.org/10.1088/1755-1315/252/4/042033

Zhang, T. (2004). Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization. The Annals of Statistics, 32(1), 56-85. https://www.jstor.org/stable/3448494

Zhao, Y., Xu, X., & Wang, M. (2019). Predicting overall customer satisfaction: big data evidence from hotel online textual reviews. International Journal of Hospitality Management, 76, 111-121. doi: 10.1016/j.ijhm.2018.03.017

Zheng, Z., Wu, X., & Srihari, R. (2004). Feature selection for text categorization on imbalanced data. ACM SIGKDD Explorations Newsletter, 6(1), 80-89. doi: 10.1145/1007730.1007741

Zhu, X., Zhang, P., & Xie, M. (2021). A Joint Long Short-Term Memory and AdaBoost regression approach with application to remaining useful life estimation. Measurement, 170, 108707. https://doi.org/10.1016/j.measurement.2020.108707

List of Publications

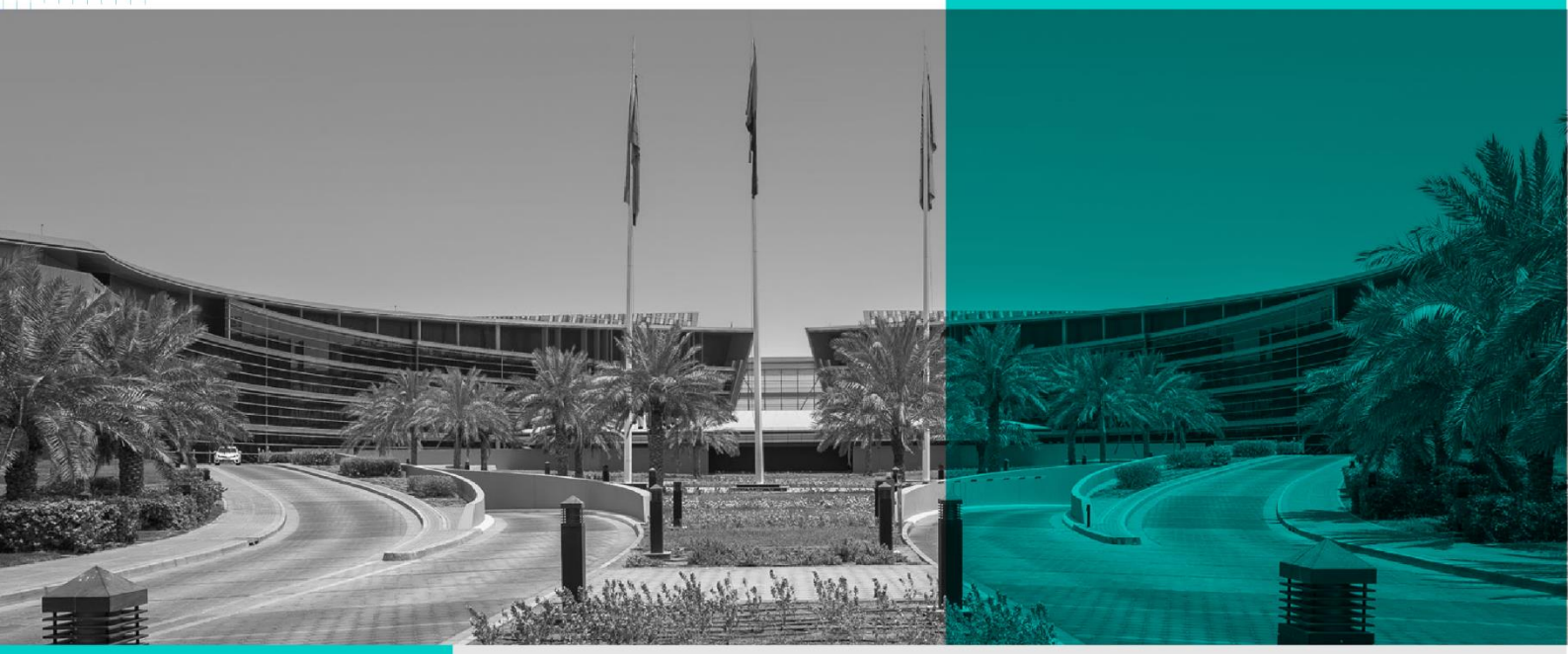
Al Rabaiei, K., Alnajjar, F., & Ahmad, A. (2021). Kano Model Integration with Data Mining to Predict Customer Satisfaction. Big Data and Cognitive Computing, 5(4), 66. 1-18. doi:10.3390/bdcc5040066

Appendix

List of questions

1. Gender	1. جنس
2. Dormitory	2. السكن الجامعي
3. residence services and cleaning in the housing	3. خدمات الإقامة والنظافة في السكن
4. Cleaning and hygiene in the campus	4. التنظيف والنظافة في الحرم الجامعي
5. Modern equipment and decoration in the classrooms: (projection machine, data machine, etc.)	5. التجهيزات الحديثة والديكور في الفصول الدراسية: (آلة العرض ، آلة البيانات ، وما إلى ذلك) الفصول
6. Uncrowded classroom	6. الدراسية غير المكتظة
7. Food dining hall services	7. خدمات قاعة الطعام
8. The possibilities of doing lessons in the laboratories	8. إمكانيات عمل الدروس في المختبرات
9. Shopping services in school buildings	9. خدمات التسوق في المباني المدرسية
10. Student unions and clubs	10. اتحادات ونوادي الطلاب
11. Health services	11. خدمات طبية
12. The possibility of good communication with the teaching staff	12. إمكانية التواصل الجيد مع أعضاء هيئة التدريس
13. The possibility of communicating with the administration	13. إمكانية التواصل مع الإدارة
14. Transportation facilities on campus	14. مرافق النقل في الحرم الجامعي
15. How close the bus stations form classrooms?	15. قرب محطات توقف الباصات من الفصول الدراسية
16. How close the car parking?	16. قرب مواقف سيارات
17. Scholarships given by the university body	17. ا لمنح الدراسية المقدمة من الهيئة الجامعية
18. Shopping center on campus	18. مركز تسوق في الحرم الجامعي
19. Sports and entertainment facilities	19. المرافق الرياضية والترفيهية
20. organizations of festivals, concerts and celebrations	20. المهرجانات والحفلات والاحتفالات المنظمة
21. Advising unit and Tools	21. وحدة الارشاد وأدواتها
22. The Internship Experience	22. تجربة التدريب
23. Information Technology Services	23. خدمات تكنولوجيا المعلومات
24. Online Registration Process	24. عملية التسجيل عبر الإنترنت
25. The Information in the E-services (Grades, Schedules, Payment Reports, etc.)"	25. المعلومات في الخدمات الإلكترونية (الدرجات ، الجداول ، تقارير الدفع ، إلخ) "
26. Organizing socio-cultural activities	26. تنظيم الأنشطة الاجتماعية والثقافية
27. Teaching quality	27. جودة التدريس
28. curve grading system	28. نظام التقييم بالمنحنى للدرجات
29. Organizing some courses with certificate	29. توفر خدمات الإنترنت في الحرم الجامعي
30. The availability of internet in the campus	30. تنظيم بعض الدورات مع الشهادة
31. The library’s having got a rich data base	31. المكتبة لديها قاعدة بيانات غنية
32. The range of Academic Majors	32. مجموعة التخصصات الأكاديمية
33. The security system on campus	33. نظام الأمن في الحرم الجامعي
34. University Policies and Regulations	34. سياسات ولوائح الجامعة
35. Response to Complaints	35. الرد على الشكاوى
36. Do you like to use scooter inside the university?	36. هل تحب قيادة دراجة (سكوتر) في الجامعه
37. Online courses	37. الدراسة عبر الانترنت
38. Overall satisfaction about the study at UAEU	38. الرضاء العام عن الدراسة في الجامعةَ

UAEU
جامعة الإمارات العربية المتحدة
United Arab Emirates University



UAE UNIVERSITY DOCTORATE DISSERTATION NO. 2022:38

This dissertation develops a method to integrate the Kano model and data mining approaches to select relevant attributes that drive customer satisfaction, with a specific focus on higher education. The significant contribution of it is to improve the quality of UAEU academic support and development services provided to their students by solving the problem of selecting features that are not methodically correlated to customer satisfaction, which could reduce the risk of investing in features that could ultimately be irrelevant to enhancing customer satisfaction.

Khaled Al Rabaiei received his PhD in Informatics and Computing from the Department of Computer Science and Software Engineering, College of Information Technology at UAE University, UAE. He received his MSc in computer system and networking from the College of Electrical Engineering, London South Bank University. He received his Bachelor degree in Electrical Engineering from the College of Electrical Engineering, United Arab Emirates University.

www.uaeu.ac.ae

Online publication of dissertation:
https://scholarworks.uaeu.ac.ae/etds/

UAEU عمادة المكتبات Libraries Deanship
جامعة الإمارات العربية المتحدة
United Arab Emirates University
Digital Library Services Section - قسم الخدمات المكتبية الرقمية