REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Information Technologies

CLASSIFICATION OF USER CONSUMPTION DATA-BASED CONSUMER PROFILES USING AI

Ibrahim Ramadan Jaboua JABOUA

Master's Thesis

Supervisor

Assoc. Prof. Dr. Sefer KURNAZ

İstanbul, 2025

CLASSIFICATION OF USER CONSUMPTION DATA-BASED CONSUMER PROFILES USING AI

Ibrahim Ramadan Jaboua JABOUA

Information Technologies

Master’s Thesis

ALTINBAŞ UNIVERSITY

2025

The Thesis titled CLASSIFICATION OF USER CONSUMPTION DATA-BASED CONSUMER PROFILES USING AI prepared by IBRAHIM RAMADAN JABOUA JABOUA and submitted on 13/06/2025 has been accepted unanimously for the degree of Master of Science in Information Technologies.

_____

Assoc. Prof. Dr. Sefer KURNAZ

Supervisor

Thesis Defense Committee Members:

Assoc. Prof. Dr. Sefer KURNAZ	Department of Computer Engineering, Altınbaş University	_____
Asst. Prof. Dr. Hameed Mutlag Farhan FARHAN	Department of Computer Engineering, Altınbaş University	_____
Asst. Prof. Dr. Serdar KARGIN	Department of Biomedical Engineering, İstanbul Arel University	_____

I hereby declare that this thesis meets all format and submission requirements of a Master's thesis.

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare that all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa, as required by the abovementioned rules and conduct.

Ibrahim Ramadan Jaboua JABOUA

Signature

ABSTRACT

CLASSIFICATION OF USER CONSUMPTION DATA-BASED CONSUMER PROFILES USING AI

JABOUA, Ibrahim Ramadan Jaboua

M.Sc., Information Technologies, Altınbaş University,

Supervisor: Assoc. Prof. Dr. Sefer KURNAZ

Date: June/2025

Pages: 62

Over-The-Top (OTT) platforms have grown dramatically in popularity in recent years, giving consumers access to a variety of multimedia material. Understanding consumer behavior has therefore become essential for platform providers and advertising. In order to categorize users based on their consumption rate which may be classed as low, medium, or high this research suggests the construction of a user consumption categorization system based on machine learning (ML) as Decision Tree (DT) , Random Forest (RF) , Regression Algorithm , k-Nearest Neighbour (KNN), Bayesian Algorithm (BA), Gradient Boosting (GB) and XGBoost Classifier . In order to tailor the data and get it ready for clustering, the research will employ data preprocessing techniques. Next, it will evaluate several ML classifiers to see which one is the most accurate at predicting the user's consumption rate. Each algorithm's limits will be investigated, and the system will be combined with data analytics and mining software. By offering a useful application for categorizing user consumption habits, this research will advance the fields ML and may be useful to OTT platform providers, advertisers, and data analytics experts.

Keywords: OTT, ML, Classification, Behaviour, High, Low, Medium.

TABLE OF CONTENTS

Pages

LIST OF TABLES

LIST OF FIGURES

Pages

ABBREVIATIONS

AI	:	Artificial Intelligence
ML	:	Machine Learning
ACC	:	Accuarcy
PREC	:	Precision
REC	:	Recall
F1-S	:	F1-Score

1. INTRODUCTION

1.1 INTRODUCTION

The way we consume multimedia has significantly altered as a result of technological improvements. The day when our only sources of pleasure and information were cable TV networks has long since passed. The popularity of OTT services like YouTube and Netflix serves as evidence that in the current day, our major source of multimedia consumption has switched to internet-connected smartphones [1].

Multimedia may now be seen without a TV from anywhere in the globe because to the simplicity and accessibility of OTT content. According to a PricewaterhouseCoopers estimate, the worldwide OTT industry is anticipated to develop rapidly, with an average annual growth rate of 14%, to reach USD 156.9 billion by 2024 [2]. The number of OTT subscription services has increased as a result of this expansion and is expected to reach 650 million by 2021 [3].

Due to this expansion, the OTT industry has intensified competition, with major firms like YouTube (Yub), Netflix (Nflix), Amazon Prime, and Hulu actively extending their global footprint. In the meanwhile, businesses like Disney, Apple, Warner Media (WM), and HBO are breaking into the industry by utilizing their current content.

Network broadcast media and TV over IP businesses must contend with fierce rivalry as a result [4, 5]. Internet service providers (ISPs) are battling OTT media providers over bandwidth and payment difficulties in the meanwhile. ISPs desire profit growth since OTT services use quite a lot of network resources [6]. ISPs must increase their throughput capacity in order to meet consumer demands for a greater quality of service (QoS) [7]. ISPs frequently restrict throughput (throttling) based on the popular OTT service in order to solve bandwidth overloads, which causes customer complaints. ISPs must stronger and more timely consumer OTT use analysis tools to manage this balancing act, reduce network performance concerns, and satisfy customer requests. Additionally, ISPs would have more negotiating power with OTT providers if they had a trustworthy tool for this purpose [9].

As a result of the COVID-19 epidemic, internet access speeds and bandwidth have improved dramatically all around the world. Customers may now buy the media material they want at

reasonable costs and get it whenever and wherever they want. Traditional media industries are being disrupted by businesses. OTT services, which are described as "video content offered over internet or IP-based channels" [10], are expanding quickly as a result of the ability of customers to select individualized platforms and content in accordance with their own schedules. With an average annual growth rate of 29.4% from 2020 to 2027, the worldwide OTT market is anticipated to reach USD 1039.03 billion [11]. As of 2019 [12], there were 182 million OTT customers in the US, the country with the largest OTT market. 84.2% of users of digital video watched YouTube, while 67.6% watched Netflix [13]. Global service providers and telecommunications carriers are in fierce rivalry with one another as a result of the expansion of the OTT market.

The suggested machine learning technique aims to enhance our comprehension of consumer consumption trends on OTT platforms. The approach entails gathering open-source data and modifying it to satisfy the needs of the machine learning algorithms. The data will be preprocessed in order to get it ready for the classification task. The most efficient method for categorizing OTT platform usage patterns will be identified by a comparison examination of several machine learning classification algorithms. This analysis will provide us a complete understanding of the algorithms' performance.

This study has the potential to advance the fields of machine learning and data analysis by demonstrating how the algorithms for categorizing OTT platform usage patterns may be used in real-world settings. The study's findings can also give OTT platform providers and marketers useful information on how to better understand and target their viewers. This may therefore result in better user experiences and raised interest in OTT platforms.

Figure 1.1: OTT Platform**.**

1.2 OBJECTIVE OF THE RESEARCH

This research proposal aims to develop a user consumption classification system based on the contribution of Artificial Intelligence, specifically Machine Learning. The goal of the research is to classify user consumption patterns according to their usage rate, which can be classified as low, medium, or high.

To achieve this goal, the research will follow a multi-step process. Firstly, data preprocessing techniques will be applied to the data in order to customize it and prepare it for the machine learning algorithms. The data will then be divided into different groups, based on the consumption rate, which will serve as the basis for the clustering.

Once the data is prepared, the research will investigate which machine learning classifier is the most accurate in determining the user's consumption rate. A variety of classifiers will be considered, including Decision Trees, Random Forests, Support Vector Machines, and others. Each classifier will be evaluated using a set of evaluation metrics, such as accuracy, precision, recall, and F1-score, to determine its performance.

The classification report for each algorithm will be explained in detail, including the performance of each classifier and the results of the evaluation metrics. The research will also explore the limitations of each algorithm and the factors that may affect their performance.

Finally, the developed system will be integrated with data analytics and mining applications, to provide a comprehensive solution for user consumption classification. The system will allow data analysts to gain valuable insights into user consumption patterns, which can be used to improve the user experience and optimize the use of OTT platforms.

This research will contribute to the field of Artificial Intelligence and Machine Learning by providing a practical application of these technologies for classifying user consumption patterns. The results of the research will also have the potential to benefit OTT platform providers and advertisers, as well as data analytics professionals, by providing valuable insights into user behavior and consumption patterns.

1.3 RESEARSH SINIFICANCE

Finally, the developed system will be integrated with data analytics and mining applications, to provide a comprehensive solution for user consumption classification. The system will allow data analysts to gain valuable insights into user consumption patterns, which can be used to improve the user experience and optimize the use of OTT platforms.

This research will contribute to the field of Artificial Intelligence and Machine Learning by providing a practical application of these technologies for classifying user consumption patterns. The results of the research will also have the potential to benefit OTT platform providers and advertisers, as well as data analytics professionals, by providing valuable insights into user behavior and consumption patterns.

This research aims to address these limitations by applying machine learning classifiers and data preprocessing techniques to enhance the accuracy of the user consumption classification. The proposed research builds upon previous studies in this field, such as [4], and seeks to improve upon their results by utilizing more advanced preprocessing techniques and a more thorough evaluation of the performance of different machine learning algorithms.

By conducting this research, we aim to make a significant contribution to the field of AI and Machine Learning and provide valuable insights for OTT platform providers, advertisers, and data analytics professionals. The results of the research have the potential to improve our understanding of user behavior and consumption patterns, which can lead to improved user experiences and increased engagement with OTT platforms.

1.4 CONTRIBUTION AND MOTIVATION

1.4.1 Contribution

This research aims to contribute to the field of AI and Machine Learning by improving the accuracy and effectiveness of user consumption classification. By applying machine learning algorithms and data preprocessing techniques, the research aims to provide a more comprehensive solution for user profiling. This research seeks to improve upon the existing approaches to user profiling and classification by utilizing advanced preprocessing techniques and a more thorough evaluation of the performance of different machine learning algorithms.

In addition, this research will contribute to the understanding of user behavior and consumption patterns. The results of the research have the potential to provide valuable insights into user behavior and preferences, which can inform OTT platform providers, advertisers, and data analytics professionals on how to better engage with users and optimize their interactions with OTT platforms.

1.4.2 Motivation

The motivation for this research is driven by the growing need for a better understanding of user behavior and consumption patterns. As OTT platforms continue to grow in popularity, it is becoming increasingly important to understand how users interact with these platforms and the resources they consume. This information can be used to improve the user experience, optimize the use of OTT platforms, and better target advertisements.

However, current approaches to user profiling and classification are often limited in their accuracy and effectiveness. This is why this research seeks to address these limitations and make a significant contribution to the field of AI and Machine Learning. The results of the

research have the potential to provide valuable insights into user behavior and consumption patterns, which can inform OTT platform providers, advertisers, and data analytics professionals on how to better engage with users and optimize their interactions with OTT platforms.

1.5 BACKGROUND

The use of OTT platforms has grown significantly in recent years, and with this growth has come an increased need for a better understanding of user behavior and consumption patterns. These platforms, such as Netflix, Hulu, and Amazon Prime Video, offer users access to a wide range of content, from movies and TV shows to music and video games. As such, they are becoming an increasingly important part of people's daily lives and entertainment habits.

As the popularity of OTT platforms continues to grow, it is becoming increasingly important for OTT platform providers, advertisers, and data analytics professionals to gain a deeper understanding of how users interact with these platforms and the resources they consume. This information can be used to improve the user experience, optimize the use of OTT platforms, and better target advertisements.

However, current approaches to user profiling and classification are often limited in their accuracy and effectiveness. These approaches are usually based on simple metrics, such as the number of hours spent on a platform or the number of resources consumed, and do not take into account other important factors, such as user preferences and engagement.
This is why this research seeks to address these limitations and make a significant contribution to the field of AI and Machine Learning by improving the accuracy and effectiveness of user consumption classification. The results of the research have the potential to provide valuable insights into user behavior and consumption patterns, which can inform OTT platform providers, advertisers, and data analytics professionals on how to better engage with users and optimize their interactions with OTT platforms.

1.6 CONCLUSION

The increasing popularity of OTT platforms has led to a growing need for a better understanding of user behavior and consumption patterns. This information can be used to

improve the user experience, optimize the use of OTT platforms, and better target advertisements. However, current approaches to user profiling and classification are often limited in their accuracy and effectiveness, which is why this research proposes a machine learning approach to enhance the accuracy and effectiveness of user consumption classification.

The proposed research will utilize machine learning algorithms and data preprocessing techniques to provide a more comprehensive solution for user profiling. The results of the research have the potential to provide valuable insights into user behavior and consumption patterns, which can inform OTT platform providers, advertisers, and data analytics professionals on how to better engage with users and optimize their interactions with OTT platforms. This research aims to contribute to the field of AI and Machine Learning by improving the accuracy and effectiveness of user consumption classification.

2. LITERATURE REVIEW

2.1 INTRODUCTION

An summary of earlier studies on the subject is provided in the Literature Review section of a research paper on OTT employing intelligent strategies. It covers OTT-related themes, including its influence on the broadcasting business, its development, difficulties, and prospects, as well as the clever methods employed to enhance the distribution and consumption of OTT material. Key findings should be highlighted, opportunities for further study should be identified, and a critical critique of the available literature should be presented in the chapter. It provides the backdrop and context for the research investigation.

2.2 RELATED WORK

Previous studies on OTT services tried to improve our comprehension of user behavior in relation to the spread of new media. Based on the primary research topics for each study, three general categories may be made. The first category focuses on how user perceptions and use intentions regarding OTT services are formed. Scholars were intrigued by the possibility of effective dissemination of OTT services over traditional media when they initially entered the media industry. As OTT was seen as a novelty in the media market, they investigated consumers' behavioral intentions using an innovation diffusion model [14, 15]. The second category assesses how well OTT services compete with traditional pay-TV options including cable TV, IPTV, and satellite TV. In order to forecast the eventual replacement of pay-TV by OTT services, these studies assessed the amount of rivalry between pay-TV and OTT services. OTT was viewed as the product of innovation in information and communication technology (ICT) in the studies pertaining to the first study subject. These studies were built on the premise that, because to technological advancement, everyone with an internet connection may watch whatever movie they want, whenever they want, on any device [16].

Two information systems theories were used to analyze user behavior in the OTT market: the Technology Acceptance Model (TAM) and the Theory of Planned Behavior (TPB). Perceived usability and perceived usefulness, according to TAM, are two important criteria that affect behavioral intent about technology usage. [17] A theoretical framework for

examining user behavior in the context of OTT services was made available by the application of TAM and TPB.

According to the TAM, writers in [18] investigated the factors that influence people's intentions to use Netflix . 251 people participated in the survey, and the results showed that perceived usefulness had a favorable influence on respondents' intentions to use OTT services, while reported ease of use had no appreciable impact.

Similar to [19], researchers looked into OTT usage's motivations using a modified version of the UTAUT2 (Unified Theory of TAM) model. They discovered that the UTAUT2 model was a useful resource for describing OTT usage.

In [20], the researchers looked at how participant inventiveness, subjective norms, and reputation goals affected participants' happiness and intention to continue using in the setting of uses and gratifications. They used the TAM and the expectation-confirmation model to analyze the findings of an online poll they performed in Korea with 315 participants. This research contributed to the body of literature by offering an empirically tested theoretical model that illustrated the factors that influence customers' intentions to use over-the-top (OTT) services at a time when OTT was still new on the market and few people were familiar with it. But in recent years, the media landscape has undergone a substantial transformation, with OTT services in many industrialized nations—including Korea—becoming more common and outpacing traditional broadcasting services.

Dimmick's niche theory was frequently employed in research that examined the rivalry between pay-TV and OTT providers. According to this argument, consumers will gravitate toward new media with greater features due to their limited time and resources, which will lead to a decline in the usage of older media [21,22]. The competitive relationships between three television mediums (OTT, IPTV, and digital cable) were examined using this hypothesis, and the results showed that OTT was the most competitive, with digital cable and IPTV being roughly equally competitive.

According to scholars in [23], who used niche theory to describe the competitive dynamics in video platforms in Korea, there was little to no competition in Korea between traditional pay-TV and OTT services.

According to Chen's study [24] in Taiwan, which applied niche theory, OTT services and traditional TV had a high degree of similarity in terms of entertainment and usability, with OTT services having a competitive advantage in all areas. When OTT services initially entered the market, these studies assisted in identifying the beneficial characteristics of those services in comparison to traditional media and their direct rivalry with traditional TV providers. The shortcoming of the niche analysis is that it only considers the replacement of old media while ignoring the complimentary nature of new and old media. As a result, it is no longer appropriate to see OTT services and pay-TV as competitors between old and new media.

Authors [25] looked at South Korean consumers' preferences for OTT service characteristics such as pricing, content quality, convenience, service dependability, and user interface. They gathered survey information from 1,100 Korean respondents and discovered that customers valued content quality over convenience and cost. These studies emphasize the need of comprehending consumer preferences for various OTT service characteristics, as this may assist providers in better tailoring their offers to satisfy consumer wants and preferences.
The mixed logit analysis was carried out on the conjoint survey data gathered from Korean OTT users, according to research by authors in [26]. The findings showed that Korean consumers wanted pay-TV services to be less expensive when combined with OTT services, and that as the number of bundled services rose, so did their sensitivity to the discount rate. The study also showed that in Korea, OTT services and TV broadcasting services work well together.

In different research by authors [27], the authors used conjoint analysis to analyze the marginal willingness to pay for certain OTT service features between China and Korea. The findings indicated that resolution, followed by the recommendation system and viewing possibilities, were the most significant characteristics for Chinese customers. The suggestion system, however, was the feature that Korean customers loved the most, followed by the watching choices.

Several large corporations, including Netflix, YouTube, Amazon PrimeVideo, and Hulu, now enjoy a considerable market share in the US OTT industry and hence dominate the OTT business. These businesses control around 75%, 55%, 44%, and 32%, respectively, of the

US OTT market, according to the figures mentioned in [28], and they collectively account for 79% of the US market share.

New competitors are, however, entering the market and attempting to stand out. For instance, according to [29], Disney launched Disney+ in November 2019 and 50 million US members had signed up by July 2020. Similar to HBO, AT&T introduced HBO Max after purchasing Warner Media. This service delivers both recent and old HBO programming, and as of June 2020, according to [30], it has over 34 million US members.

Traditional TV networks have both a difficulty and an opportunity in the rapidly changing OTT industry. On the one hand, it can threaten their current business plans, but on the other, it gives them an opportunity to enter new markets by providing OTT services utilizing their current content and service infrastructure. As an illustration, Hulu, which was introduced in 2008 and features material from FOX, NBC, and ABC [31], has become the market leader. With almost 3 million monthly users in Korea, Wavve provides material from KBS, MBC, and SBS [32].

ISPs' market positioning and business strategies are more intricate than those of conventional TV networks. IP-TV and internet service providers must make investments in and keep up a high-quality broadband infrastructure for both services. In addition, bundling tactics must be used to stop users from dropping IP-TV in favor of OTT services just [34].
KT, which controls the largest portion of the IP-TV market in Korea, is considering collaborating with Netflix in order to achieve synergies. With the use of this agreement, Netflix would be able to significantly increase its market share in Korea, and KT, which has a sizable number of wired network customers, could increase its subscriber base and increase network usage fees by providing Netflix content [35]. It will be more important for ISPs to ensure steady QoS in order to prevent customer churn as 5G technology is widely adopted and the number of users utilizing wireless OTT services rises. The continued adoption of real-time video-streaming services like Twitch and Discord will provide a challenge for consistent service delivery in particular [36,37].

Network traffic analysis was used by researchers to identify users in [34]. To differentiate and identify users, they created the idea of flow-bundle-level characteristics. They used a

dataset of flow records with 65 users to compare four different ML models. Using the random forest technique, the authors were able to identify people with an accuracy of 83%. A profiling model that describes user behavior and temporal dynamics was put out by the authors in [35]. According to their research, 80% of users visit six cells everyday, with 22% never returning to a site and 30% doing so every 6–10 days. Additionally, users' upload/download rates were greater throughout the holidays. This knowledge of recurring mobility patterns can help forecast heavy traffic loads and enhance network administration. The authors of [37] looked into the relationship between user behavior and phone usage. To find distinctive representations of user behavior, they employed a SOM and a multilayer perceptron set up as an autoencoder. To create user behavior profiles and isolate a single person based on their activity for a particular device, the authors employed application, cell tower, and website usage data.

According to research by authors in [38], user identification based on behavior patterns is essential since it's simple to alter IP address allocation to get around security measures. The authors described a novel method for identifying users that makes use of NetFlow traffic patterns. A number of variables are taken from the network flows and used to a random forest model to categorize individuals. The model proved effective in identifying users with a high degree of precision (94%), making it a good option for forensic science. With this approach, identifying and describing users no longer require analysis of the full stream (header and payload).

The authors of [39] looked at the prospect of identifying a user's gender based on the mobile apps they had downloaded. The goal was to make smartphone systems more user-friendly and offer more specialized services and goods. The study discovered that gender may be deduced from usage patterns and via empirical analysis discovered substantial disparities in application categories, functionality, and icon designs. An accuracy of 76.62% was shown by the study's prediction model.

The writers of [40] carried out research to describe the Twitter network of verified users. The study examined 79,213,811 relationships among verified Twitter users and 231,246 user accounts. The findings revealed that the verified users' sub-graph shares some characteristics with the full Twitter user graph, such as a small diameter, but also differs significantly in a number of important ways, including a power-law out-degree distribution, a slight degree of

dissortativity, and a higher reciprocity rate. The scientists also noted that, in contrast to the actions of regular Twitter users, the activity levels of verified users demonstrated a stagnant pattern over time.

The Millennial generation (born 1983–1996) and Generation Z (born 1997–2006) have grown up with modern technology and cultural changes, including OTT services. Baby Boomers had an average of 8 [34] subscriptions, compared to 17 [34] and 14 [34] for Millennials and Gen-Z, respectively. According to research, 70% [41] of millennials (Generation Z) and just 39% [41] of baby boomers (Generation X) currently subscribe to Netflix. The qualities of customers differ depending on the nation. Consumers in India use video streaming services for more than 45 billion hours [42], more than twice as long as those in the US. With about 2000 hours [42] of video material seen per person, South Korea has the greatest consumption rate.

As a result, several groups of customers in the OTT market may be created, each with a different number of members. Researchers [31].'s analysis of mobile user data traffic (including OTT) revealed that heavy users, who account for just 2–4% of all users, consume the majority of data. They gathered mobile traffic information from Finland, Germany, the UK, Japan, and Brazil and used a clustering approach to categorize users. The percentages of heavy, regular, and light users were 3.5%, 41.9%, and 54.6% in Finland, the country with the greatest sample size. The amounts of data uploaded and downloaded by each category were 328.7 MB, 64.3 MB, and 9.4 MB, respectively. Only around 3% of people were heavy users, yet they consumed more data than nearly 97% of other users. While the proportions varied, a comparable tendency was discovered in other nations [31]. Additionally, research that categorized users by scrutinizing actual OTT traffic discovered same tendencies. The high, medium, and low consumption groups, respectively, had 84, 50, and 582 people, according to authors in [29].

Similar to this, real-world settings frequently show extremely tiny or extremely large numbers of samples for particular classes [43]. As a result, the most populous classes may be overrepresented in analysis results [44]. In machine learning, this is referred to as the "class-imbalance problem" and is one of the ten most difficult problems to solve [45]. Despite the fact that there have been a few studies on OTT traffic, few have taken into account the class disparity issue, which is evident in OTT demographics. If this issue is not

resolved, computations for real-world grouping and pricing strategies will suffer from performance degradation issues. We offer a solution to solve these issues in ML approaches when used in real OTT situations to address this issue.

Other than OTT, the issue of class-imbalance in categorization occurs frequently. This problem has been ignored in some research on Twitter spam detection, but it has been addressed in others using artificial resampling techniques [46,47]. However, the minor classes could not have enough data characteristics if resampling is carried out on a little amount of data. Cost-based approaches have been used in binary classification to address this issue [33,48], reducing the cost of misclassifications for the smaller classes. A cost-based strategy was used in this study, utilizing a technique that can be applied to a dataset with several classes in line with the traits of OTT customers.

2.3 MACHINE LEARNING

Machine learning (ML) is a method for increasing data handling effectiveness by training computers to learn from the data. ML is used to extract pertinent information when interpreting data gets challenging. Because there are so many datasets accessible, demand for ML has skyrocketed, and many sectors employ ML to evaluate data. The aim of machine learning (ML) is to allow computers to learn from data, and research has been done to provide techniques enabling computers to learn without explicit programming. To deal with this issue, particularly with big datasets, a variety of strategies have been put forth by mathematicians and programmers.

There are serval ML methods used in OTT classification, we will present in the next subsection.

2.3.1 Decision Tree

The DT methodology is a well-liked data mining technique that builds classification schemes or forecasting formulas for a target variable based on a number of variables. Using an inverted tree structure with root, internal, and leaf nodes, this non-parametric technique divides a population into segments that resemble branches. Large, complex datasets are handled well without the need for intricate parametric frameworks. Study data can be divided into training and validation datasets with a sizable enough sample size. A decision tree model

is created using the training dataset, and the validation dataset is utilized to identify the best tree size for the final model.

DT [49] is a supervised classification technique that makes use of a standard tree structure, which has a root, nodes (the places where branches split), branches, and leaves. The nodes of the decision tree are depicted as circles, while the branches are represented as connecting segments. Typically, the tree is drawn from left to right, starting at the base and moving downward. The tree begins at the root node, and the terminus node is referred to as the "leaf" node. Two or more branches, which each represent a specific attribute and a range of values, can be extended from an internal node (non-leaf node). The values of the specified feature are divided into these value ranges as partition points. The diagram in Figure 1 shows the tree structure.

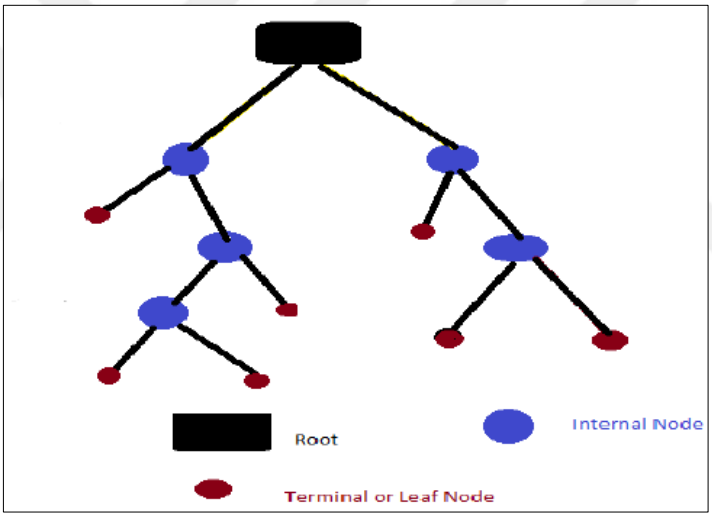

Figure 2.1: Tree Structure**.**

The grouping of data in a DT is dependent on the values of the characteristics included in the supplied data. The pre-classified data used to build the tree is divided into classes according to the attributes that effectively segment the data. Following this, the data items are divided into subgroups based on the values of these attributes, and this procedure is repeated for each subset. Until every piece of data in the current subset belongs to the same class, the process keeps going until it comes to an end.

2.3.2 Random Forest

A versatile, user-friendly ML technique called RF often gives excellent results even without hyper-parameter adjustment. Due to its ease of use and the fact that it can be applied to both

classification and regression problems, it is also among the most popular algorithms [50]. A supervised learning system called Random Forest generates a forest and somewhat randomizes it. The ensemble of Decision Trees it constructs was typically trained using the bagging approach. The bagging method's general tenet is that combining learning models improves the end outcome [50]. RF has the key benefit of being applicable to both classification and regression problems, which make up the majority of modern machine learning systems.

Given that random forest is frequently seen as a key part of ML, we will concentrate on describing its use in categorization. The structure of a RF for categorization is shown in the next image.

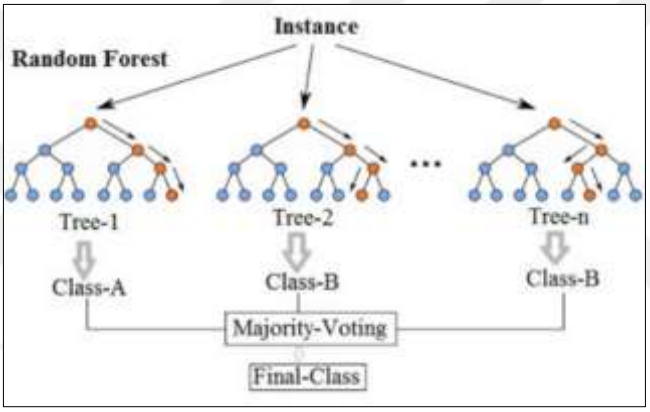

Figure 2.2: RF Illustration**.**

A well-liked and simple ML method called RF is frequently employed for classification jobs. It is renowned for generating effective outcomes with little to no hyperparameter adjustment. With a random forest classifier, overfitting is less likely to happen if there are enough trees in the forest. For real-time predictions, the method may become sluggish and ineffective if there are many trees. Although the algorithm may be trained quickly, making predictions might take some time, especially if a more precise forecast is needed, which would need adding additional trees. Understanding relationships in data may be better accomplished through alternative methods as RF is a predictive modeling tool and not a descriptive tool.

2.3.3 Regression Algorithm

A predictive analytics method called regression analysis [51] makes advantage of the correlation between dependent (target) and independent variables. There are several popular

regression models, including linear regression, logistic regression, stepwise regression, ordinary least squares regression (OLSR), multivariate adaptive regression splines (MARS), and locally estimated scatterplot smoothing (LOESS).

2.3.3.1 Linear regression model

A statistical method called simple linear regression is used to look at the connection between two continuous variables. It presupposes that the input variable(s) and the output variable, which may be computed as a linear combination of the input variable, have a linear connection (s). It is referred to as simple linear regression if there is just one input variable. It is referred to as multiple linear regression if there are several input variables. The estimation of traffic arrival times and the forecasting of salaries and salaries are two typical applications of linear regression. An illustration of a linear regression model is shown in Figure 2.3.

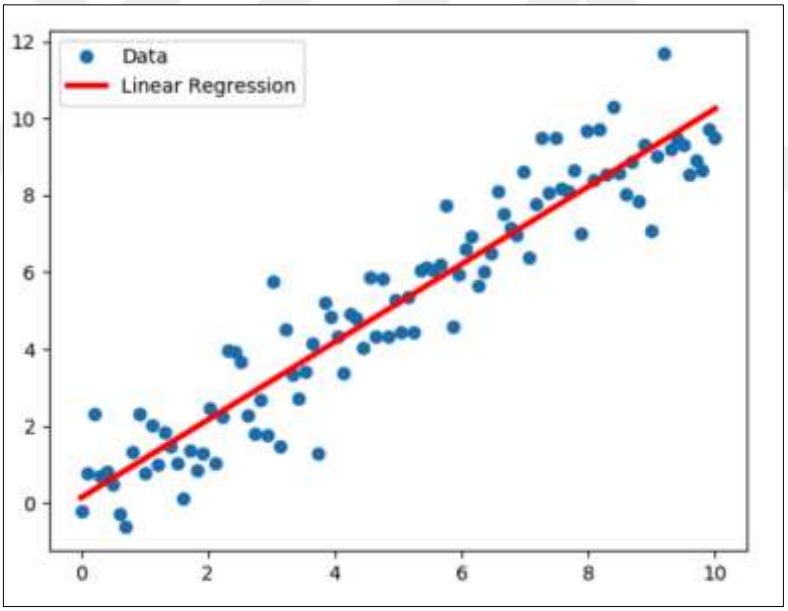

Figure 2.3: Sample Example for A Linear Regression Model.

2.3.3.2 Logistic regression

When the response is binary, particularly in fraud detection, credit card scoring, and clinical trials, logistic regression is a commonly used regression approach. This supervised machine learning algorithm's ability to handle many dependent variables, which may be continuous or dichotomous, is one of its key features. The fact that it offers a quantitative number to

gauge the degree of correlation between variables is another benefit. However, there are several drawbacks to logistic regression, such as a lack of resilience and a heavy reliance on the model. Despite this, logistic regression is utilized in a number of different areas, including the real estate industry for house value prediction and the insurance industry for client lifetime value estimation. It may also be used to anticipate if a consumer is likely to purchase a certain scenario or to provide a continuous outcome.

2.3.3.3 Multivariate regression algorithm

When there are several predictor variables in a regression model, a method known as multivariate multiple regression is used to predict a response variable based on a set of explanatory factors. It is a straightforward supervised machine learning technique that benefits greatly from matrix operations for implementation. Multivariate regression may be carried out using the definitions and operations for matrix objects included in the Python "numpy" package. In the retail industry, multivariate regression is frequently used to examine consumer decisions based on a variety of variables, including brand, price, and product, and to identify the ideal mix of these variables to boost shop traffic.

2.3.4 K-Nearest Neighbour (KNN)

The KNN [52] decision rule is a widely used method for solving pattern recognition issues, however it has a flaw in that it treats all labeled samples equally when selecting which class to place the pattern in, regardless of how typical they are. In order to solve this problem, three approaches for determining the fuzzy memberships of the labeled samples have been developed. The experimental findings and comparisons to the crisp version of each approach are described. One of the easiest machine learning techniques and an excellent approach to learn about classification and ML is the kNN method. The most comparable data points in the training set must be located, and a classification must be made based on those data points' classifications. This approach, however straightforward to comprehend and use, has several applications in a variety of fields, such as recommendation systems, semantic searching, and anomaly detection.

The KNN algorithm is a ML method that gauges how closely fresh data points' attributes match those in the training set. This closeness decides whether the algorithm classifies the new data point as a continuous value (regression) or a discrete class (classification).
As illustrated in Figure 4, the algorithm classifies the new data point in the classification scenario according to which class its K nearest neighbors are most likely to belong to. To forecast the value of the new data point in regression, the algorithm computes the average or median of the values of its KNN.

KNN is a straightforward yet effective method that has been used in a number of fields, including anomaly detection, semantic search, and recommendation systems.

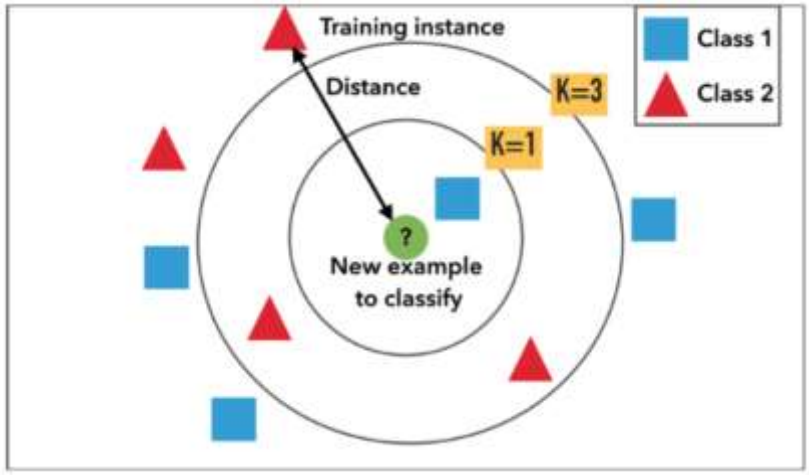

Figure 2.4: Example of k-NN Classification.

2.3.5 Support Vector Machine

A supervised learning technique called the SVM may be applied to classification, regression, and density estimation. Using the structural risk minimization (SRM) [56] approach, a reasonable upper limit or boundary on the generalization error is established. Then, using predetermined training parameters, the best model structure is sought for. SVM uses a nonlinear transformation with a predetermined kernel, such as a linear, polynomial, sigmoidal, radial basis, or hybrid, to map the input space into a higher dimensional feature space.

Due to the fact that both SVM and Artificial Neural Network (ANN) may be seen as two-layered networks with nonlinear weights in the first layer and linear weights in the output layer, they are comparable to one another [34]. However, unlike ANN, SVM does not choose

its parameters for the first layer by adaptive learning, but rather through training input vectors. The fact that SVM needs fewer training samples and variables is one of its benefits. In a regression-based SVM, the training data sets are defined as [x_i, y_i], where xi∈Rn is the input vector, n is the dimension of input vector, and $y_i \in [-1,1]$ is the output vector. SVM uses quadratic programming techniques to find optimal hyperplanes that separate an input class from a target class. This can be expressed mathematically as shown in (1) (2). The references are kept in their original positions.

$$min\frac{1}{2}w^t w + c\sum_{i=1}^{n}\xi_i \tag{2.1}$$

$$y_i(w\varphi(x_i) + b) + \xi_i - 1 \geq 0 \tag{2.2}$$

The SVM model involves mapping the training sample in a high dimensional feature space using the function $\varphi(x_i)$, with w is the weight vector, b is the bias term, C is the penalty for the error term, and $\xi_i \geq 0$ being the slack variable used to prevent influence of noisy data. The slack parameter ξ_i and parameter C are used to avoid overfitting. Figure 2.5 provides an illustration of how SVM selects the optimal hyperplane that maximizes the margin.

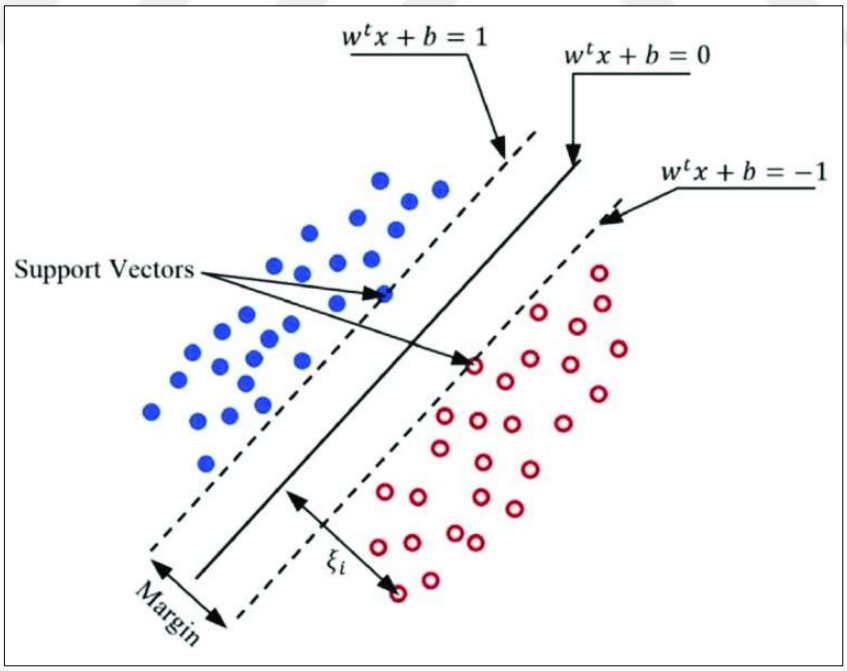

Figure 2.5: SVM Operating Concepts Illustrated Graphically**.**

2.3.6 Bayesian Algorithms

Some ML algorithms use the Bayes [53] Theorem to solve classification and regression issues, including Naive Bayes (NB), Gaussian Naive Bayes (GNB), Multinomial Naive Bayes(MNB), Averaged One-Dependence Estimators (AODE), Bayesian Belief Network (BBN), and Bayesian Network (BN).

2.3.6.1 Naive bayes

By assuming conditional independence across the training dataset, the complexity of the Bayesian classifier may be decreased and it can become more useful. Using this premise, the NB method drastically decreases the complexity of the issue to just 2n.

If the probability distribution governing X is unrelated to the value of Y given Z, conditional independence is the assumption that X and Y are conditionally independent given Z. In other words, if Z happens, knowing whether X happens doesn't tell you anything about the chance of Y happening, and vice versa. The assumption of conditional independence states that the probability distribution of A is independent of the value of B given the response variable A in the case of a response variable A and input attribute B.

2.3.6.2 Multinomial naive

Multinomial Naive Natural language processing and text analysis frequently employ the Bayes classification technique, which is based on Bayes' theorem. It is predicated on the ideas that feature counts have a multinomial distribution and that features (words in a text) are conditionally independent given the document's category and class label.
The formula for Multinomial Naive Bayes is as follows:

$$P(y|x) = P(x|y) * \frac{P(y)}{p(x)} \tag{2.3}$$

where:

a. the class identifier is y. (category of the document)
b. The vector of features is x. (counts of words in the document)
c. The probability of y in the posterior given x is $P(y|x)$ (i.e., the probability that the document belongs to category y, given its features x)

d. The probability of x given y is $P(x|y)$ (i.e., the probability of observing the feature vector x, given that the document belongs to category y)
e. The prior likelihood of y is $P(y)$ (i.e., the probability of a document being in category y before considering its features)
f. The proof is $P(x)$ (i.e., the probability of observing the feature vector x, regardless of the category).

We determine the posterior probability for each potential class label and select the one with the highest likelihood to categorize a new document:

$$y_{hat} = argmax_y P(y|x) \tag{2.4}$$

where:

The anticipated class label for the new document is y hat.

By considering each feature count as independent and applying the product rule of probability, we may simplify the computation of the likelihood term under the NB assumption of conditional independence:

$$P(x|y) = \prod P(x_i|y)^{\{count(x_i)\}} \tag{2.5}$$

where:

x_i is the count of the i_{th} feature (word) in the document.

$count(x_i)$ is how many times the document uses the i_{th} feature.

$P(x_i|y)$ is the probability of observing the i-th feature given that the document belongs to category y

To estimate the probabilities $P(x_i|y)$ and $P(y)$, In order to prevent zero probability, we employ maximum likelihood estimation (MLE) or a version like Laplace smoothing. These methods add a little pseudo-count to each feature count.

2.3.7 Gradient Boosting

A effective ML method called GB [54] combines a number of straightforward models with boosting and gradient-based optimization to enhance prediction performance. Weak learners are frequently employed with boosted trees, which are generally binary split decision trees.

A boosted classifier (or regressor) produces an additive form that yields a more accurate composite hypothesis by establishing a number of hypotheses and merging them. Due to their superior accuracy and quick operation, boosted trees are at the heart of many cutting-edge solutions in a variety of learning disciplines.

Having the additive form of with an input feature vector of x:

$$H(x) = \sum_{t} \alpha_t \ h_t(x) \tag{2.6}$$

In GB, the ensemble of depth-3 trees is built in a greedy manner to minimize a regularized objective on the training loss. The weight assigned to each hypothesis, $h_t(x)$, is represented by α_t. A broad schematic representation of the ensemble is shown in Fig. 5. Gradient-boosting is used to build the trees, which are an adaptable combination of numerous basic models that increase prediction performance.

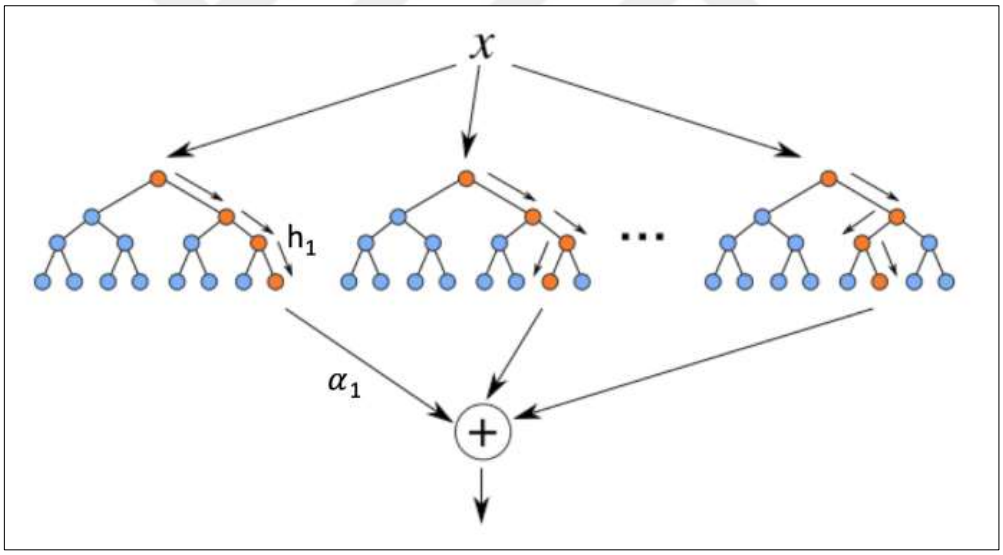

Figure 2.6: An Enhanced Ensemble of DT Is Shown Schematically.

2.3.8 XGBOOsT

GBDT architecture is used to create ML algorithms in XGBoost [55], a distributed library of gradient boosting techniques. Figure 2 depicts the XGBoost's fundamental structure. In order to lower the residual, the first tree's residual is passed into the second tree, and so on. Using this method, the residuals in the forecasts are successively corrected, which enhances the model's performance. Due to its excellent efficiency and efficacy, XGBoost has grown

to be a well-liked tool for many machine learning applications. It is tailored for distributed computing.

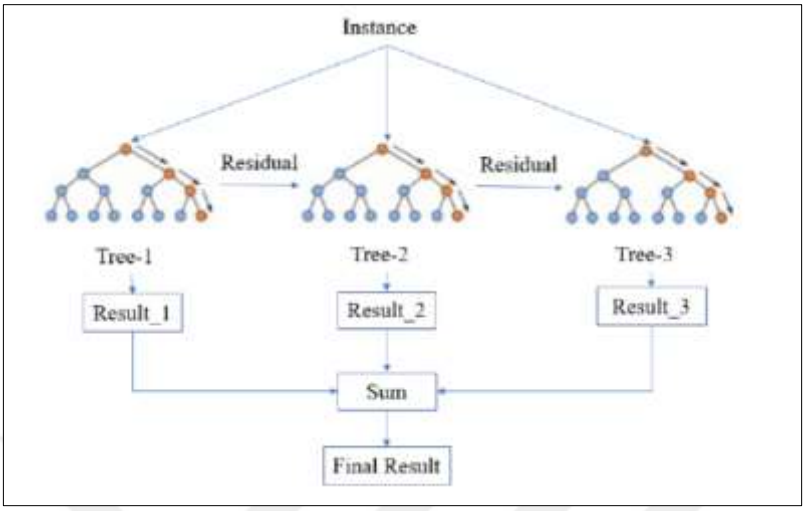

Figure 2.7: XBOOT Example**.**

Under the GBDT framework, XGBoost is a distributed library of GB algorithms that implements machine learning algorithms [55]. Figure 2.7 depicts the XGBoost's fundamental structure. In order to lower the residual, the first tree's residual is passed into the second tree, and so on. Using this method, the residuals in the forecasts are successively corrected, which enhances the model's performance. Due to its great efficiency and efficacy, XGBoost—which is tailored for distributed computing—has grown to be a prominent tool for many ML applications.

2.4 CONCLUSION

In this chapter, we looked at the relevant research on OTT services and the ML techniques applied to this field. We discovered that OTT services have used a number of machine learning approaches, including recommendation systems, content analysis, and predictive modeling, to enhance user engagement, content discovery, and personalized experiences. These techniques have shown encouraging results and the potential to grow the OTT market. However, there are still certain issues that need to be resolved in the implementation of ML in OTT services, such as data privacy and bias.

3. MEDHODOLOGY

3.1 INTRODUCTION

This chapter's goal is to outline a system for categorizing OTT streaming service consumers according to how they utilize the services. The rapid expansion of OTT services has fueled intense competition among streaming service providers for customers' attention. For the creation of successful marketing strategies and enhancing user engagement, it is essential to comprehend how people consume information. To overcome this difficulty, we created a system that categorizes OTT consumers as low, medium, or high consumption users using a variety of machine learning algorithms. We will go through the procedures for data collecting and preprocessing, the ML algorithms that were employed, and the model training and validation process in this chapter. We'll also go through the methodology's findings and their consequences.

3.2 RESEARCH APPROACH

To classify OTT consumers as low, medium, or high consumption users, our suggested technique makes use of a number of machine learning algorithms. To create focused marketing tactics and boost user engagement, it is challenging to precisely detect and categorize user consuming behavior. This technique solves this difficulty. To obtain and prepare user activity data for analysis, our technique involves a data collection and preparation stage. We next evaluate the data and categorize consumers based on their consumption patterns using machine learning methods like DT, logistic regression, and KNN. Futhermore, our technique offers a methodical and useful way to categorize OTT consumers according to their consumption habits, which may be helpful for streaming service providers trying to optimize their offerings and user experience. Figure 3.1 shows the components of our framework, which we will go over in more depth in the sections below.

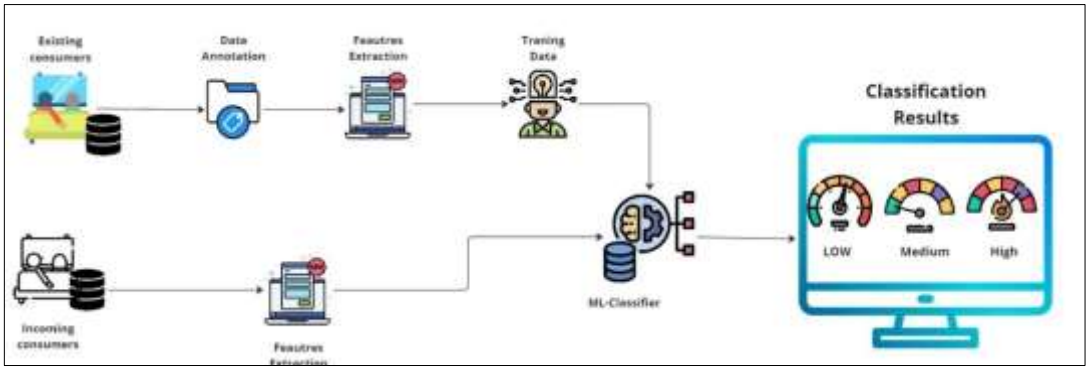

Figure 3.1: Researsh Approash

3.2.1 Dataset

The statistics offered details how consumers consumed content on the Universidad del Cauca Network throughout April, May, and June of 2019. The dataset is a single CSV file with 1249 instances and 114 characteristics. Each instance represents a user's consumption profile, which contains a summary of the user's actions in relation to 56 OTT applications found in the various IP flows that were recorded to construct the dataset.
Two categories of attributes—time occupation (in seconds) and data occupation (in bytes) per application—summarize user behavior. 56 OTT applications, including Amazon, Apple, Dropbox, Facebook, Google, Instagram, Netflix, Skype, TikTok, WhatsApp, YouTube, etc. are covered by the dataset. Users are divided into three categories: Low, Medium, and High Consumption Users.

The dataset is essential for network operators to gather crucial data on consumption patterns in order to develop new data plans targeted at certain consumers and have a proper understanding of the network. Additionally, by analyzing and comprehending user behavior, this data might enhance network performance.

3.2.2 Data Annotation

After gathering the data consumption patterns of current OTT users, it is necessary to first annotate the data in order to categorize the consumers depending on their OTT usage. In order to ascertain the users' OTT usage habits, the data gathered from the users must be analyzed.

The OTT consumer traffic is then divided into three categories based on this analysis: high consumption, medium consumption, and low consumption. Consumers that use OTT heavily are referred to as "high consumption" consumers, while those who use it sparingly are referred to as "low consumption" users and those who use it on average are referred to as "medium consumption" users.

Users are categorized depending on how they consume data, which varies greatly between the three categories. High consumption consumers frequently consume a lot of data and invest a lot of time in OTT services. Low consumption customers, on the other hand, use extremely little data and spend very little time using OTT services. Users who consume in the middle fall between these two categories.

Network operators must categorize customers based on their OTT consumption habits in order to provide new data plans targeted at certain users and get an accurate picture of the network. The performance of the network may be enhanced by using this data to evaluate and comprehend user behavior.

3.2.3 Features Extraction

The following phase is feature extraction after gathering OTT user traffic and annotating it. Finding specific, quantifiable data characteristics that may be utilized for ML entails feature extraction. The objective is to identify the most important elements that will improve the classification's accuracy.

The factors relating to user behavior that may be included in the features derived from OTT user traffic include total data used, number of sessions, duration of sessions, frequency of usage, time of day, and kind of OTT application. The average data rate, the volume of data transferred, the number of transactions, and the OTT application's reaction time are further features that may be derived.

The issue description and the characteristics of the OTT traffic are used to determine which features to extract. The classification accuracy increases as more relevant characteristics are extracted. But it's crucial to strike a balance between the quantity of data collected and the computing expense of the machine learning algorithms being utilized for categorization.

In conclusion, feature extraction is a crucial stage in the analysis of OTT user traffic data since it chooses the characteristics that will be utilized to train the classification ML models. The standard and applicability of the derived characteristics determine how accurate the categorization will be.

3.2.4 ML-Classifier

In order to create ML models that can categorize incoming customers according to their OTT usage habits, characteristics from the acquired consumer data are first extracted from them and utilized as training data. The objective is to create a model that can correctly predict a new user's categorization based on their OTT traffic statistics.
Extraction of the pertinent characteristics from traffic data is the first stage when fresh data from arriving customers has to be reviewed. These characteristics are then fed into the machine learning model, which categorizes the new user as having high, medium, or low usage.

In conclusion, important elements from traffic data are extracted in order to categorize arriving customers based on their OTT consumption habits. An ML model that has been trained on the gathered data is then used to forecast categorization. It is possible to use a variety of strategies to solve the issue of class imbalance, which can result in decreased classification accuracy.

3.2.4.1 DT

Recursively dividing the training data into subsets depending on the most important characteristic, the method constructs a decision tree until either a stopping requirement is satisfied or the subsets are pure (i.e., they include only one class label).
The fact that DT is simple to perceive and comprehend is one of its key benefits. The decision-making principles for DT may be simply retrieved from the tree structure and displayed. Because of this, DT is a suitable option for applications where interpretability and explainability are crucial, like OTT consumption analysis, where the decision-making process must be open and simple to understand.

DT also have the ability to handle numerical and categorical variables, as well as missing values, which is another benefit. By evaluating the significance of each feature throughout

the tree development process, decision trees may also be utilized for feature selection. This can help to reduce the number of dimensions in the data and boost the effectiveness of the classification model.

In the supplied code snippet, a DecisionTreeClassifier object is created and trained using the matching class labels (y train) and the training data that has been converted (x train transformed). The class labels of fresh data may then be predicted using the trained model.
With the extra benefit of being simple to perceive and comprehend, decision trees are a strong and versatile tool for classification and regression problems. Their suitability for OTT consumption analysis, where transparency and interpretability are crucial, results from this.

3.2.4.2 RF

Popular ML method RF is particularly suitable for OTT data analysis. This is because it can handle noisy data as well as big datasets with plenty of characteristics.
Being a very precise algorithm, RF has the capacity to reach high levels of accuracy even while working with complicated data, which is one of its key benefits. Additionally, because of its robustness, this technique is less prone to overfitting than others. This is due to the fact that the model's variance is decreased by the employment of several decision trees, each of which is trained using a different random subset of the data.
Another benefit of RF is that it is a relatively quick method, allowing for the usage of real-time applications and handling of big datasets. Additionally, because RF is an ensemble approach, it can handle both numerical and categorical data and can capture a larger variety of feature interactions.

3.2.4.3 LG

ML algorithms like logistic regression (LG) are frequently employed for categorization issues. Logistic Regression may be used in OTT analysis to determine whether a user belongs to a particular consumption group (e.g., high, medium, or low consumption).
LG has the benefit of being computationally efficient and having a short learning curve, even for huge datasets. Additionally, it works effectively when the input characteristics and the output labels have a linear connection. Furthermore, LG may offer insights into the connection between the input characteristics and the output labels, which can be helpful for

comprehending the elements that lead to excessive OTT usage. However, when the connection between the input data and the output labels is nonlinear, Logistic Regression may not be as accurate as more sophisticated ML techniques like RF and NN.

3.2.4.4 KNN

A simplistic ML technique used for classification problems is called KNN. Based on how closely a user's usage habits resemble those in the training data, KNN may be used to categorize a user's consumption behavior as high, medium, or low in the context of OTT analysis.

The KNN algorithm has the benefit of being easy to comprehend and use. It is also a non-parametric method, which means it makes no assumptions about the distribution of the data at its core. When the distinction between classes is not well defined, KNN can still be useful. KNN may be computationally costly, particularly on big datasets, as it necessitates computing the distances between each data point in the training set and the data point to be categorized. The accuracy of the method may also be sensitive to the size of the input characteristics and the value of k.

3.2.4.5 SVM

The fitting of two alternative SVM models to the training data is demonstrated in the Python code. While the second model, svmlr, utilizes a linear kernel to separate the data, the first model, SVMRBF, uses a RBF kernel.

SVMs are a well-liked kind of supervised learning model that may be applied to regression or classification applications. They operate by identifying the best hyperplane for classifying the data. To help the hyperplane better distinguish between the classes, the data are transformed into a higher-dimensional space using the kernel function.
The capacity of SVMs to effectively handle high-dimensional data is one of its advantages. They may be used with both linear and non-linear kernels and are useful when dealing with non-linear data. Due to its ability to handle a sizable number of data and non-linear connections between those features, SVMs might be helpful in the context of OTT for categorizing consumers based on their consumption behavior.

3.2.4.1 Multinomial naive

Based on the Bayes theorem, the MNB method is a straightforward yet powerful probabilistic classifier that is frequently employed for text classification tasks including spam filtering, sentiment analysis, and topic categorization. Given the attributes retrieved from the dataset for OTT data analysis, MNB may be used to categorize customers according to their consumption patterns.

MNB is advantageous for OTT classification because it is quick and effective and performs well with high-dimensional datasets that have sparse features, which are typical of OTT data. In addition, it can handle many classes and needs less training data than other machine learning algorithms. Additionally, MNB makes the premise that characteristics are independent, which may not totally hold true when applied to OTT data but can still produce accurate classification results.

3.2.4.2 GB

An example of an ensemble learning technique is gradient boosting classifier, which combines many weak models (in this case,DT) to produce a stronger, more precise model. It functions by training DT consecutively, with each succeeding tree aiming to fix the mistakes caused by the one before it. The weighted average of all the trees' forecasts makes up the final projection.

GradientBoostingClassifier has the benefit of being resilient to outliers and noisy data, as well as being able to handle high-dimensional data with many features. It can also do well in classification tasks, particularly when the weak models are diversified and well-tuned. In contrast to more straightforward models like logistic regression or DT, it could take longer to train and forecast.

3.2.4.3 XGBoost

Extreme Gradient Boosting, often known as XGBoost, is a strong and effective machine learning algorithm that has gained popularity in the OTT sector for usage in a variety of tasks including content recommendation, churn prediction, and user behavior analysis.

One of XGBoost's key benefits is its capacity to manage complicated, sizable, and high-dimensional datasets, which has helped it become a popular option in the OTT sector where enormous volumes of user data are created daily. The capacity to model non-linear correlations and interactions between variables, which is critical for properly anticipating user preferences and behaviors, contributes to its high prediction accuracy and performance. Furthermore, XGBoost has the capacity to deal with missing data and lessen overfitting using regularization methods like L1 and L2 regularization, which is particularly helpful in OTT where data might be noisy or partial. Parallel processing is used in its implementation, allowing for quicker training and prediction times even with big datasets.

Nevertheless, XGBoost's great performance, accuracy, and capacity to handle sizable and complicated datasets with ease have helped it establish itself as a standard method in the OTT space.

3.2.5 Evaluation Metrics

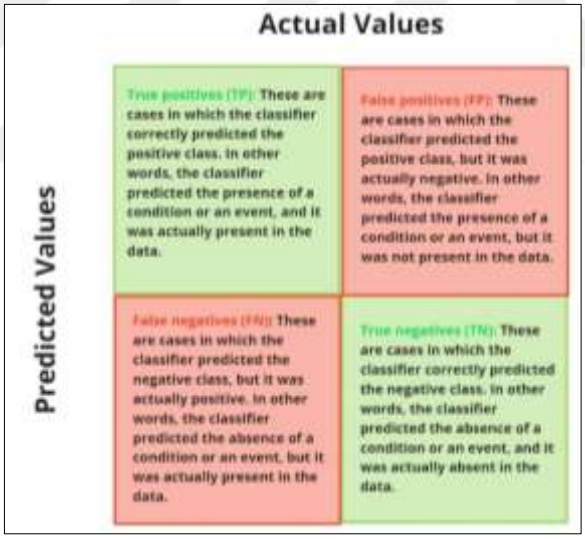

Figure 3.2: Confusion Matrix (CM) Definition.

a. The percentage of cases that are properly categorized out of all the examples is known as accuracy. It is an uncomplicated measure that is simple to comprehend. However, if the data are unbalanced, which means that one class is far more frequent than the other, accuracy might be deceiving. A classifier that consistently predicts the majority class in such circumstances may have great accuracy but not be practical.

b. Occurrences that are anticipated as positive are subtracted from all other instances to determine the precision. It is a gauge of a classifier's capacity to steer clear of occurrences that are projected to be positive but are really negative, known as false positives. Precision is crucial in instances where false positives can be expensive or harmful, such in medical diagnosis.

c. Recall quantifies the percentage of real positive incidents among all positive cases. It evaluates the classifier's capacity to find every successful instance, even those that are overlooked (FN). Recall is crucial for spotting fraud or spam emails, for example, because FN can be expensive or harmful.

d. A harmonic mean of recall and accuracy is the F1-score. It gives a single measure of the classifier's performance while balancing both measures. When there is a trade-off between accuracy and recall, which are both crucial, the F1-score is especially helpful.

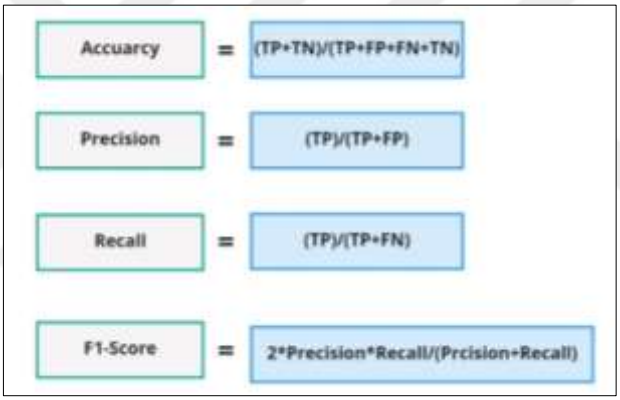

Figure 3.3: Evaluation Metrics**.**

In conclusion, measures used to assess the effectiveness of classification models include accuracy, precision, recall, and F1-score. Depending on the particular needs of the assignment, they are useful in various circumstances and gauge various elements of the model's performance.

3.3 CONCLUSION

In conclusion, this chapter offers a thorough framework for classifying users of OTT streaming services according to how they use such services. To identify consumers with low, medium, and high consumption, the system combines data gathering, preprocessing, and ML techniques. The results of this study offer insightful information that OTT service providers may utilize to create targeted marketing campaigns and raise user engagement.

Understanding how people consume information is crucial for the success of these firms given the fast proliferation of OTT services and growing competition. This system offers a helpful framework for tackling this issue and can be improved upon in the future. The outcomes of each ML technique employed in the system for classifying users of OTT streaming services will be shown in the next chapter. This will allow for a thorough assessment of each method's efficacy and a comparison of how well it performs. We can better understand which approach for classifying OTT customers by comparing the accuracy of each model for low, medium, and high consumption users.

4. EVALUATION OF PERFORMANCE

4.1 INTRODUCTION

In the chapter previously, we described a method for classifying users of OTT streaming services based on how they use such services. We used data collection, preprocessing, and machine learning techniques to achieve this. When we categorize users into low, medium, and high consumption groups in this chapter, we will explain the outcomes of each machine learning technique that was used in the system. By comparing the effectiveness of each approach, we can better determine which ML algorithm OTT service providers should employ to target particular customer demographics and increase user engagement. The outcomes of the ML models, including their confusion matrices and accuracy scores, will be presented first. The performance of each approach will then be compared, and the potential causes of their shortcomings will be examined. We will next go through the ramifications of our results for the OTT market, including how they may be applied to create powerful marketing plans and raise user engagement. The findings and analysis given in this chapter make a significant contribution to the marketing of OTT streaming services and shed light on how machine learning methods may be used to classify user behavior.

4.2 RESULTS

The results of each ML algorithm employed in the system for classifying users of OTT streaming services are shown in the "Results" section. We'll highlight the outcomes of the KNN, DT, RF, LR, XGBoost, SVM, GB, and MNB models in particular. We will give an overview of the model's performance in classifying low, medium, and high consumption consumers by presenting their respective confusion matrices and accuracy ratings. Insights into the most successful strategy for OTT service providers to target particular consumer groups and increase user engagement will be gained through the study of these data, which will enable us to compare the efficacy of each ML algorithm.

4.2.1 DT Result

The findings (Fig 4.1) are for a DT classifier that was evaluated on 250 samples after being trained on a dataset. The model's accuracy is 0.944, meaning that it can accurately predict the class label for 94.4% of the test items.

The CM demonstrates that the model accurately identified 93 samples as belonging to class 0, 71 samples as belonging to class 1, and 72 samples as belonging to class 2. While class 0 should have been used, the model incorrectly categorized 6 samples as class 1 and 2 samples as class 2. Moreover, it incorrectly assigned 4 samples to class 0, when class 2 should have been assigned, and 1 sample to class 0, when class 1 should have been.

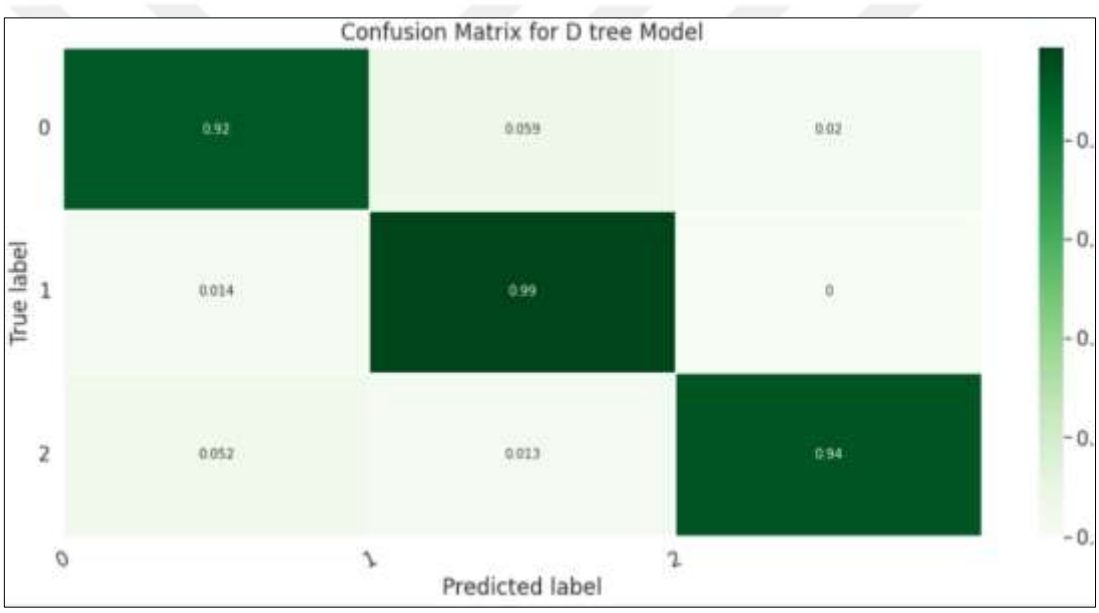

Figure 4.1: DT-CM**.**

The class 0 PREC score is 0.95, which indicates that 95% of the samples that the model predicted to be in class 0 really were. The class 0 REC score is 0.92, which indicates that 92% of the samples that truly belonged to the class 0 were properly identified by the model. The harmonic mean of PREC and REC, which serves as the F1-S for class 0, is 0.93.

The PREC, REC, and F1-S for classes 1 and 2 are computed similarly. The model is doing well in terms of both PREC and REC, as evidenced by the accuracy and recall for each of the three classes being over 0.90. PREC, REC, and F1-S have a weighted average of 0.95, which shows (Fig 4.2) that the model is operating effectively as a whole.

	precision	recall	f1-score	support
0	0.95	0.92	0.93	101
1	0.91	0.99	0.95	72
2	0.97	0.94	0.95	77
accuracy			0.94	250
macro avg	0.94	0.95	0.94	250
weighted avg	0.95	0.94	0.94	250

Figure 4.2: DT-Metrics.

Generally, the findings imply that the DT classifier is operating efficiently and successfully assigning the samples to the appropriate classes with high ACC. Nonetheless, there is definitely potential for improvement, particularly for the samples that were incorrectly identified.

4.2.2 RF Results

The results (Fig 4.3) are for a classifier called a RF that was trained on a dataset and evaluated on a batch of 250 samples. The model's ACC is 0.976, which means that it can accurately predict the class label for 97.6% of the test items.

According to the CM, the model correctly identified 98 samples as belonging to class 0, 72 samples as belonging to class 1, and 74 samples as belonging to class 2. Three samples were incorrectly assigned to class 1 by the model when they should have been in class 0. Moreover, it incorrectly categorized two samples as class 0, when they should have been class 2.

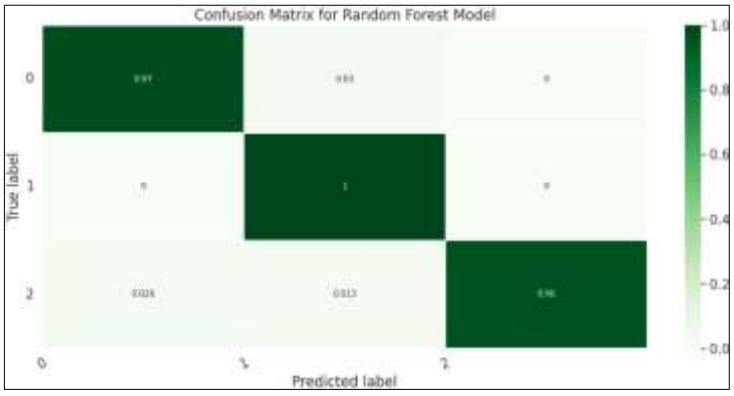

Figure 4.3: RF-CM.

According to (Fig 4.4) the PREC score for class 0, which is 0.98, 98% of all the samples that the model predicted to be in class 0 really were. The class 0 REC score is 0.97, which indicates that 97% of the samples that truly belonged to this class were properly identified by the model. The harmonic mean of PREC and REC is 0.98, and this value represents the F1-S for class 0.

The accuracy, REC, and F1-S for classes 1 and 2 are computed in a similar manner. The PREC and REC for all three classes are more than 0.95, demonstrating the model's excellent performance in terms of PREC and REC.

Essentially, the findings imply that the random forest classifier outperforms the DT classifier and is better able to accurately categorize the samples into the various classes. The model is performing extremely well in terms of REC and ACC, which suggests that it accurately identifies the majority of samples. Nonetheless, there is definitely potential for improvement, particularly for the samples that were incorrectly identified.

	precision	recall	f1-score	support
0	0.98	0.97	0.98	101
1	0.95	1.00	0.97	72
2	1.00	0.96	0.98	77
accuracy			0.98	250
macro avg	0.98	0.98	0.98	250
weighted avg	0.98	0.98	0.98	250

Figure 4.4: RF-Metrics.

4.2.3 LG Results

The data shown (Fig 4.5) relates to an LG classifier that was tested on a batch of 250 samples after being trained on a dataset. The model's ACC is 0.972, which indicates that for 97.2% of the test items, it can correctly predict the class label.

The model accurately recognized 96 samples as belonging to class 0, 72 samples as belonging to class 1, and 75 samples as belonging to class 2, as shown by the confusion matrix. The model wrongly classified three samples as belonging to class 1, when they really

belonged in class 0. Moreover, it misclassified two samples as class 0, when in fact they were in class 2.

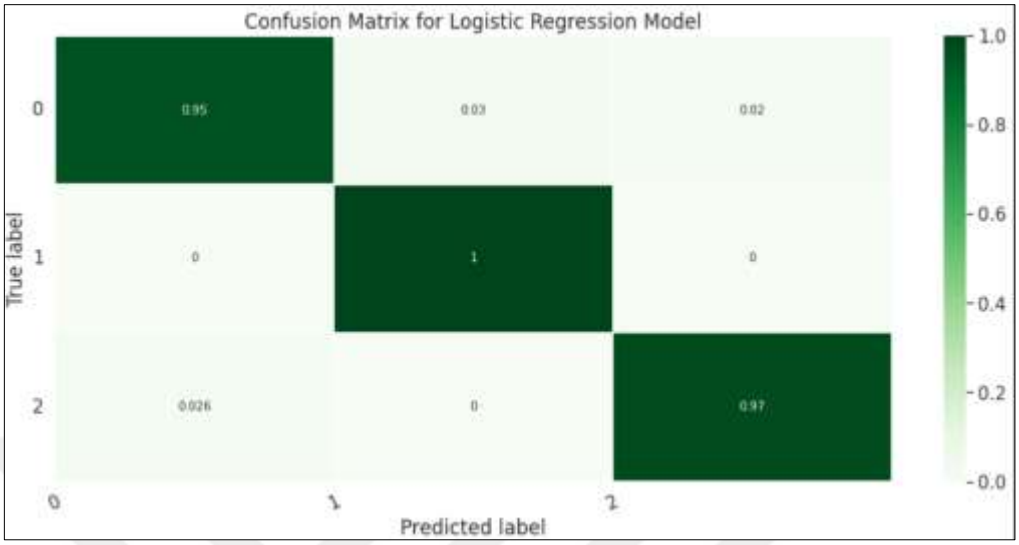

Figure 4.5: LG-CM.

As the model correctly identified 98% of the samples as belonging to class 0, class 0 got an ACC score of 0.98. Recall for class 0 is 0.95, which means that the model correctly recognized 95% of the data that really belonged to that class. The F1-S for class 0 is 0.96. ACC, REC, and F1-S for classes 1 and 2 are also calculated. The model performs well in terms of accuracy and recall, as seen by the fact that all three classes have precision and recall values > 0.95. The model is doing well, as seen by the 0.97 weighted average of ACC, REC, and F1-S.

Basically, the findings (Fig 4.6) indicate that the logistic regression classifier is functioning quite well and is able to accurately categorize the data into the appropriate groups. The model has good ACC and REC scores, which means that most of the samples are accurately identified by the model. Nonetheless, there is definitely potential for improvement, particularly for the samples that were incorrectly identified.

	precision	recall	f1-score	support
0	0.98	0.95	0.96	101
1	0.96	1.00	0.98	72
2	0.97	0.97	0.97	77
accuracy			0.97	250
macro avg	0.97	0.97	0.97	250
weighted avg	0.97	0.97	0.97	250

Figure 4.6: LG-Metrics.

4.2.4 KNN Results

The results (Fig 4.7) pertain to a KNN classifier that was tested on a batch of 250 samples following training on a dataset. The class label for 94.4% of the test items can be correctly predicted by the model, which has an ACC of 0.944. The CM shows that 95 samples were correctly classified as belonging to class 0, 69 samples as belonging to class 1, and 72 samples as belonging to class 2 by the model. The model wrongly classified two samples that belonged in class 0 as class 1. Moreover, it misclassified 4 samples as class 2 when they should have been class 0, and 5 samples that should have been class 0 as class 0.

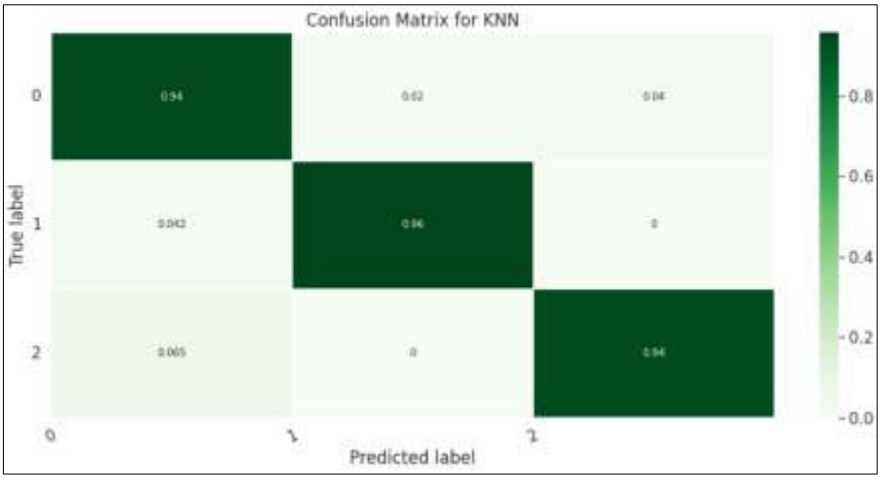

Figure 4.7: KNN-CM.

The model's predictions (Fig 4.8) regarding a sample's class were accurate 92% of the time, as shown by the PREC score for class 0. The class 0 REC score shows that the model accurately identified 94% of the genuine class 0 samples. Similar conclusions may be drawn from the PREC and REC findings for other classes. The harmonic mean of the PREC and

REC scores is the F1-S, which provides a thorough assessment of the model's performance. The F1 scores for Classes 0, 1, and 2 are respectively 0.93, 0.97, and 0.94.

The PREC, REC, and F1-S scores for the KNN classifier are 0.95, 0.94, and 0.95, respectively, according to the macro average, which takes the mean of the PREC, REC, and F1 scores over all classes. The weighted average accounts for the amount of samples in each class and gives classes with more samples greater weight. With the KNN classifier, the weighted average PREC, REC, and F1-S are 0.94, 0.94, and 0.94, respectively.

	precision	recall	f1-score	support
0	0.92	0.94	0.93	101
1	0.97	0.96	0.97	72
2	0.95	0.94	0.94	77
accuracy			0.94	250
macro avg	0.95	0.94	0.95	250
weighted avg	0.94	0.94	0.94	250

Figure 4.8: KNN-Metrics.

4.2.5 SVM Results

4.2.5.1 Linear SVM

For the test data, the linear SVM model had an ACC of 0.976, demonstrating the model's strong predictive ability. The CM (Fig 4.9) demonstrates that most samples were accurately predicted, with just a small number of misclassifications generated by the model. The model accurately predicted 97 out of 101 samples for class 0, 72 out of 72 samples for class 1, and 75 out of 77 samples for class 2 in particular.

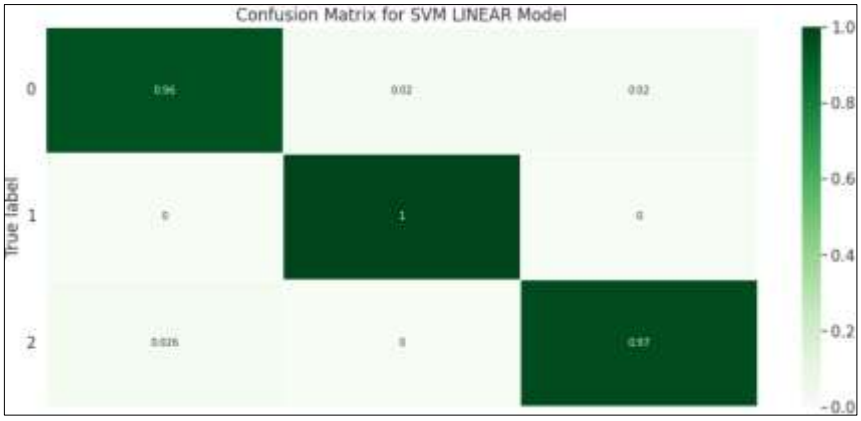

Figure 4.9: Linear SVM-CM.

The model performs (Fig 4.10) well for each class, as seen by PREC, REC, and F1-S measures that all have scores over 0.97. This shows that the model can predict each class correctly without favoring any one class. The model did exceptionally well overall on the test data, obtaining pinpoint efficiency and favorable performance indicators.

	precision	recall	f1-score	support
0	0.98	0.96	0.97	101
1	0.97	1.00	0.99	72
2	0.97	0.97	0.97	77
accuracy			0.98	250
macro avg	0.98	0.98	0.98	250
weighted avg	0.98	0.98	0.98	250

Figure 4.10: Linear SVM-Metrics.

4.2.5.2 RBF SVM

For the test set, the RBF SVM model had an ACC of 0.948, accurately classifying 94.8% of the samples. The CM (Fig 4.11) demonstrates that the model incorrectly categorized 90 samples as class 0 when they were genuinely class 0 and 9 samples as class 1 and 2 samples as class 2, respectively. Similar to this, it accurately identified class 1 for all 72 samples, but class 0 for two class 2 samples. 75 class 2 samples were accurately categorized, whereas two class 2 samples were incorrectly classified as class 0.

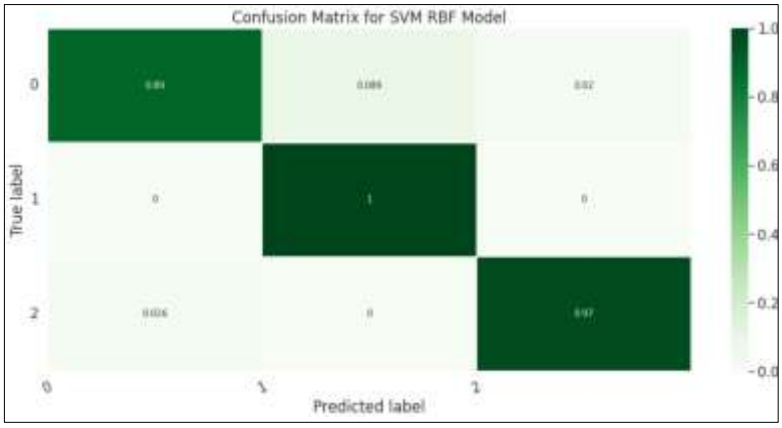

Figure 4.11: RBF SVM-CM.

According to the criteria of REC and F1-S , the model (Fig 4.12) performed well on the majority class 1 with accuracy and recall of 0.89 and 1.00, respectively. For the other groups, precision is high but recall is frequently low. For example, class 0 has a PREC of 0.98 but a recall of 0.89, indicating that the model is very good at identifying genuine negatives but may miss some TP. The model's weighted average F1-S, or total harmonic mean PREC and REC score, was 0.95, demonstrating the model's balance of ACC and REC across all classes.

	precision	recall	f1-score	support
0	0.98	0.89	0.93	101
1	0.89	1.00	0.94	72
2	0.97	0.97	0.97	77
accuracy			0.95	250
macro avg	0.95	0.96	0.95	250
weighted avg	0.95	0.95	0.95	250

Figure 4.12: RBF SVM-Metrics.

4.2.6 Multinomial Naive Results

Given the supplied dataset, the Multinomial Naive Bayes model achieved an ACC of 0.912. According to the CM (Fig 4.13), the model correctly predicted 65 instances of class 2 and 72 instances of class 1 out of 101 cases in class 0. Regrettably, it was incorrect in its predictions for 12 cases of class 3 and 7 cases each of classes 0, 2, and 3.

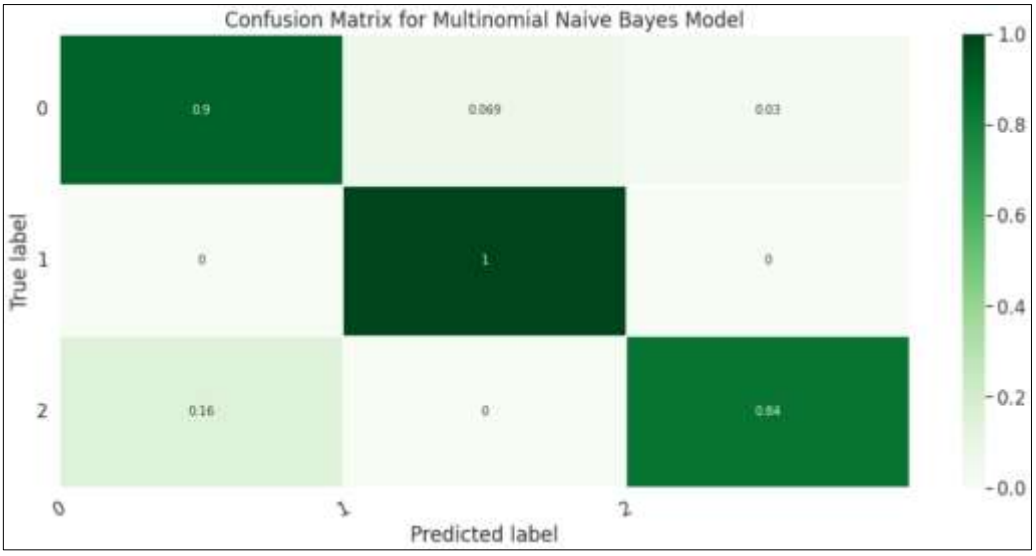

Figure 4.13: Multinomial Naive-CM**.**

The metrics table (Fig 4.14) also displays the Prec, REC, and F1-S for each class. Class 0 had a PREC of 0.88, Class 1 had a PREC of 0.91, and Class 2 had a PREC of 0.96. The REC was 0.90 for class 0, 1.00 for class 1, and 0.84 for class 2. Class 0 had an F1-S of 0.89, class 1 of 0.95, and class 2 of 0.90. The PREC, REC, and F1-S all had weighted averages of 0.91. PREC, REC, and F1-S had a macro average of 0.92.

	precision	recall	f1-score	support
0	0.88	0.90	0.89	101
1	0.91	1.00	0.95	72
2	0.96	0.84	0.90	77
accuracy			0.91	250
macro avg	0.92	0.92	0.91	250
weighted avg	0.91	0.91	0.91	250

Figure 4.14: Multinomial Naive-Metrics.

4.2.7 GB Results

One of the greatest accuracies among the classifiers you've assessed, GB's accuracy on the test data was 0.98. The CM (Fig 4.14) reveals that there were just a few incorrect classifications: two samples from class 1 were incorrectly categorized as class 0, two

samples from class 2, and one sample from class 0 was incorrectly classified as class 1. The classifier successfully identified samples from each class in general.

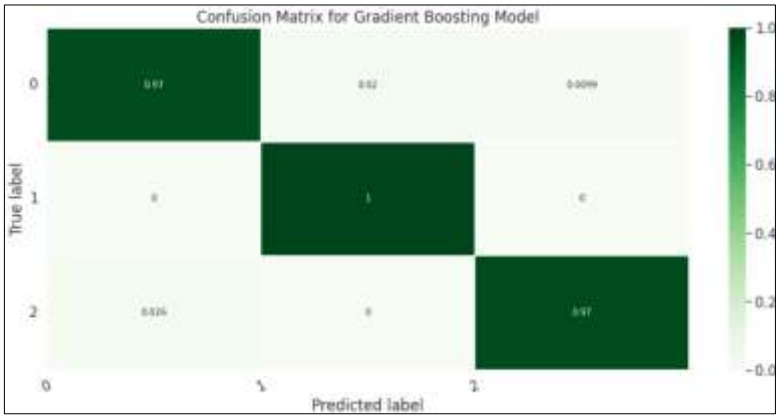

Figure 4.15: GB-CM.

When compared using the criteria (Fig 4.15), the PREC, REC, and f1-score for each class were all pretty high. For instance, class 2 had a PREC of 0.99, meaning that almost all of the samples assigned to that class genuinely belonged there. Although almost all of the data from class 0 being properly identified by the classifier, the REC for that class was just 0.97. To obtain the F1-S, which provides a single number that summarizes the performance of the classifier for each class, the PREC and REC are weighted averaged. For GB, the F1-S for each class was higher than 0.98, highlighting the classifier's excellent performance.

	precision	recall	f1-score	support
0	0.98	0.97	0.98	101
1	0.97	1.00	0.99	72
2	0.99	0.97	0.98	77
accuracy			0.98	250
macro avg	0.98	0.98	0.98	250
weighted avg	0.98	0.98	0.98	250

Figure 4.16: GB-Metrics.

4.2.8 XGBoost Results

For the dataset, the XGBoost classifier achieved an ACC of 0.976. The CM (Fig 4.17) demonstrates that, out the 101 samples in class 0, 97 were correctly categorized as class 0, 3 as class 1, and 1 as class 2. All 72 of the class 1 samples were correctly categorized. 75 of

the 77 samples in class 2 were correctly identified as belonging to that category, while two were mistakenly placed in class 0.

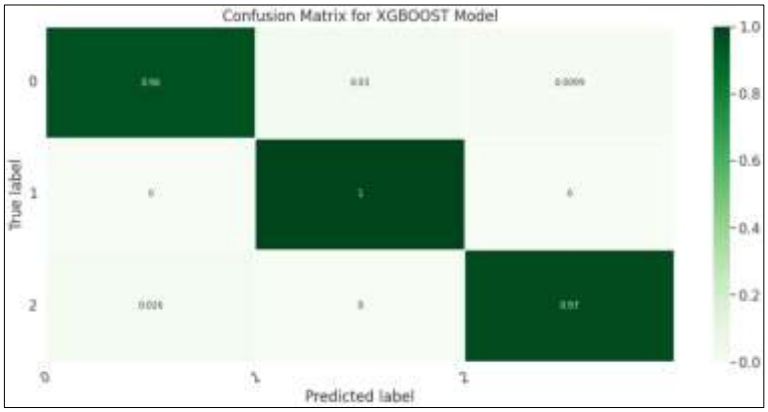

Figure 4.17: XGBoost-CM.

Each class's PREC, REC, and F1-S are also reported. The PREC, REC, and F1-S values for class 0 were all 0.98. The PREC, REC, and F1-score for class 1 were 0.96, 1.00, and 0.98 respectively. The PREC, REC, and F1-Se values for class 2 were all 0.99. For the dataset, the XGBoost classifier performed well overall with high ACC and F1-S for each class. The classifier classified samples in class 1 very well, earning a flawless PREC and REC score. The Fig 4.18 illustrates the results.

	precision	recall	f1-score	support
0	0.98	0.96	0.97	101
1	0.96	1.00	0.98	72
2	0.99	0.97	0.98	77
accuracy			0.98	250
macro avg	0.98	0.98	0.98	250
weighted avg	0.98	0.98	0.98	250

Figure 4.18: XGBoost Results.

4.3 COMPARAISION

The models with the greatest ACC, PREC, RECC, and F1-S values are GB and XGBoost. While LG and DT have lower performance measures but still provide reasonable results, Linear SVM and RF trail closely behind. The models with the lowest scores are KNN and RBF SVM.

We must take into account a number of aspects, including model performance, computational complexity, and interpretability, in order to select the appropriate model for the OTT problem. GB and XGBoost both earn the greatest ACC and F1-S in this instance, demonstrating their ability to offer more accurate predictions for the OTT problem. Yet they are also computationally demanding models that can take a lot of time and resources to train. Because they offer a decent mix between accuracy and complexity, Linear SVM or RF can be a better option if computational economy is our top priority.

In general, the optimal model for the OTT problem will rely on the particular needs and restrictions of the problem, and before reaching a final choice, it is necessary to do a comprehensive study of the data and models.

Table 4.1: Comparaison Methods.

Method	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.944	0.90	1.00	0.95
Linear SVM	0.976	0.98	0.97	0.97
RBF SVM	0.948	0.97	0.89	0.93
Multinomial Naïve Bayes	0.912	0.88	0.90	0.89
Gradient Boosting	0.980	0.98	0.97	0.98
XGBoost	0.976	0.98	0.96	0.97
K-Nearest Neighbors	0.968	0.96	0.97	0.96
Random Forest	0.984	0.98	0.98	0.98
Decision Tree	0.944	0.95	0.94	0.94

4.4 CONCLUSION

This chapter concluded with a mechanism for categorizing OTT streaming service consumers according to their usage trends. To do this, we employed methods for data gathering, preprocessing, and machine learning. We identified which ML algorithm OTT service providers could use to target certain consumer demographics and boost user engagement by grouping users into low, medium, and high consumption categories and analyzing the efficacy of each strategy. We evaluated the effectiveness of each strategy while

looking into the possible reasons for their weaknesses. We showed the results of the ML models, including their confusion matrices and accuracy ratings. We also talked about the implications of our findings for the OTT industry, such as how they may be used to develop effective marketing strategies and boost user engagement. Our research has a big impact on OTT streaming service marketing and clarifies how machine learning techniques may be applied to categorize user behavior.

5. CONCLUSION AND FUTURE WORK

In order to categorize consumers depending on their rate of consumption, this research has suggested creating a user consumption categorization system based on machine learning. The study has shown that several ML algorithms, such as DT, RF, and SVM, are capable of categorizing consumer consumption patterns with high accuracy. Insights into user behavior and consumption patterns may be gathered by combining the built system with data analytics and mining programs. These insights can be utilized to enhance the user experience and maximize the usage of OTT platforms.

Future research in this field might concentrate on enhancing the performance of the created system by taking into account more sophisticated ML techniques, including deep learning. In order to help platform owners and marketers improve user experience, the research may also look at the possibility of forecasting consumer behavior. The research can also examine the moral and privacy issues raised by the collecting and application of user data in such systems. OTT platform providers, advertisers, and data analytics experts might all gain significantly from the creation of a user consumption classification system, making it an intriguing subject for future study.

REFERENCES

[1] FCC. Annual Assessment of the Status of Competition in the Market for the Delivery of Video Programming. MB Docket No. 14–16. FCC 15-41. (accessed on 28 October 2020).

[2] Marketsand Markets. Over-The-Top Services Market by Type (Online Gaming, Music Streaming, VoD and Communication), Monetization Model (Subscription-based, Advertising-based, and Transaction-based), Streaming Device, Vertical, and Region–Global Forecast to 2024. Available (accessed on 28 October 2020).

[3] Statistica. Number of Over-the-top (OTT) Subscription Video Service Subscribers Worldwidefrom2012to2021.Availableonline: https://www.statista.com/statistics/821883/number-ottsubscribers/#statisticContainer (accessed on 28 October 2020).

[4] Kim, J.; Kim, S.; Nam, C. Competitive dynamics in the Korean video platform market: Traditional pay TV platforms vs. OTT platforms. *Telemat. Inform.* 2016, *33*, 711–721.

[5] Park, E.A. Business strategies of Korean TV players in the age of over-the-top (OTT) video service. *Int. J. Commun.* 2018, *12*, 4646–4667.

[6] Sujata, J.; Sohag, S.; Tanu, D.; Chintan, D.; Shubham, P.; Sumit, G. Impact of Over the Top (OTT) Services on Telecom Service Providers. *Indian J. Sci. Technol.* 2015, *8*, 145.

[7] Dai, W.; Baek, J.W.; Jordan, S. Feature article: Network Neutrality [Neutrality between a vertically integrated cable provider and an over-the-top video provider]. *J. Commun. Netw.* 2016, *18*, 962–974.

[8] Hu, M.; Zhang, M.; Wang, Y. Why do audiences choose to keep watching on live video streaming platforms? An explanation of dual identification framework. *Comput. Hum. Behav.* 2017, *75*, 594–606.

[9] Rojas, J.S.; Rendon, A.; Corrales, J.C. Consumption Behavior Analysis of Over the Top Services: Incremental Learning or Traditional Methods? *IEEE Access* 2019, *7*, 136581–136591.

[10] Federal Communications Commission. Annual Assessment of the Status of Competition in the Market for the Delivery of Video Programming. MB Docket No. 14-16. FCC 15-41. 2015. (accessed on 23 May 2021).

[11] Rake, R.; Gaikwad, V.; Over-the-top (OTT) Market Outlook—2027. Allied Market Research. 2020. Available online: https://www.alliedmarketresearch.com/over-the-topservices-market (accessed on 19 April 2021).

[12] von Abrams, K. The Global Media Intelligence Report. *eMarketer*. 2018. Available online: https://www.emarketer.com/content/global-media-intelligence-2018,(accessed on 19 April 2021).

[13] Benes, R. US Digital Video. *eMarketer*. 2019. Available online: https://www.emarketer.com/content/us-digital-video-2019, (accessed on 19 April 2021).

[14] KISDI. A Study on Internet Video Content Distribution and Consumption; Korea Information Society Development Institute: Jincheon, Republic of Korea, 2019.

[15] Koreaherald. Netflix Removes 30-Day Free Trial. Available online: http://www.koreaherald.com/view.php?ud=20210407000710 (accessed on 24 August 2022).

[16] Yousaf, A.; Mishra, A.; Taheri, B.; Kesgin, M. A cross-country analysis of the determinants of customer recommendation intentions for over-the-top (OTT) platforms. Inf. Manag. 2021, 58, 103543

[17] Leung, L.; Chen, C. Extending the theory of planned behavior: A study of lifestyles, contextual factors, mobile viewing habits, TV content interest, and intention to adopt mobile TV. Telemat. Inform. 2017, 34, 1638–1649.

[18] Malewar, S.; Bajaj, S. Acceptance of OTT video streaming platforms in India during covid-19: Extending UTAUT2 with content availability. J. Content Community Commun. 2020, 12, 89–106.

[19] Cebeci, U.; Ince, O.; Turkcan, H. Understanding the Intention to Use Netflix: An Extended Technology Acceptance Model Approach. Int. Rev. Manag. Mark. 2019, 9, 152.

[20] Bhattacharyya, S.S.; Goswami, S.; Mehta, R.; Nayak, B. Examining the factors influencing adoption of over the top (OTT) services among Indian consumers. J. Sci. Technol. Policy Manag. 2022, 13, 652–682.

[21] Kim, D.; Park, N. Effects of OTT Service Users Use Motivations on Satisfaction and Intention of Continued Use. J. Broadcast. Telecommun. Res. 2016, 93, 77–110. (In Korean).

[22] Dimmick, J.W. Media Competition and Coexistence: The Theory of the Niche; Routledge: London, UK, 2002

[23] Gaskins, B.; Jerit, J. Internet news: Is it a replacement for traditional media outlets? Int. J. Press/Politics 2012, 17, 190–213

[24] Kim, J.; Kim, S.; Nam, C. Competitive dynamics in the Korean video platform market: Traditional pay TV platforms vs. OTT platforms. Telemat. Inform. 2016, 33, 711–721.

[25] Chen, Y.N.K. Competitions between OTT TV platforms and traditional television in Taiwan: A Niche analysis. Telecommun. Policy 2019, 43, 101793

[26] Kim, S.; Lee, C.; Lee, J.; Kim, J. Over-the-top bundled services in the Korean broadcasting and telecommunications market: Consumer preference analysis using a mixed logit model. Telemat. Inform. 2021, 61, 101599.

[27] Kim, M.S.; Kim, E.; Hwang, S.; Kim, J.; Kim, S. Willingness to pay for over-the-top services in China and Korea. Telecommun. Policy 2017, 41, 197–207.

[28] Marvin, R. Netflix, YouTube, Prime Video, and Hulu Dominate Streaming, for Now. Available online: https://www.pcmag.com/news/netflix-youtube-prime-video-andhulu-dominate-streaming-for-now (accessed on 28 October 2020).

[29] Webb, K. Disney Plus Can't Compete with Netflix when it Comes to Original Content, but its Affordable Price and Iconic Franchises Make it a Great Value for Families. Available online: https://www.businessinsider.com/disney-plusreview (accessed on 28 October 2020).

[30] Spangler, T. HBO Max and HBO Have 36.3 Million Subscribers, Up 5% From End of 2019, AT&T Says. Available online: https://variety.com/2020/digital/news/hbo-max-subscribers-subscribers-q2-att1234714316/ (accessed on 28 October 2020).

[31] Kim, J.; Kim, S.; Nam, C. Competitive dynamics in the Korean video platform market: Traditional pay TV platforms vs. OTT platforms. *Telemat. Inform.* 2016, *33*, 711–721.

[32] YonhapNews. Local OTT Giant Wavve Sees Drop in Active Users, Netflix Soars: Report. Available online: https://en.yna.co.kr/view/AEN20200617003700320?input2106m (accessed on 28 October 2020).

[33] Kim, J.; Nam, C.; Ryu, M.H. IPTV vs. emerging video services: Dilemma of telcos to upgrade the broadband. *Telecommun. Policy* 2020, *44*, 101889.

[34] Yoo-chul, K. Netflix May Pay for KT's Network. Available online: http://www.koreatimes.co.kr/www/tech/2020/07/133293720.html (accessed on 28 October 2020).

[35] Hu, M.; Zhang, M.; Wang, Y. Why do audiences choose to keep watching on live video streaming platforms? An explanation of dual identification framework. *Comput. Hum. Behav.* 2017, *75*, 594–606.

[36] Johnson, M.R.; Woodcock, J. And Today's Top Donator is: How Live Streamers on Twitch.tv Monetize and Gamify Their Broadcasts. *Soc. Media Soc.* 2019, *5*.

[37] Nielsen. The Nielsen Total Audience Report: August 2020. Available online: https://www.nielsen.com/us/en/insights/report/2020/the-nielsen-total-audience-report-august-2020/ (accessed on 19 April 2021).

[38] Webb, K. Disney Plus can't Compete with Netflix when it Comes to Original Content, but its Affordable Price and Iconic Franchises Make it a Great Value for Families. *Business Insider*. 2020. Available online: https://www.businessinsider.com/disney-plus-review (accessed on 19 April 2021).

[39] Spangler, T.; Littleton, C. HBO Max and HBO Have 36.3 Million Subscribers, Up 5% From End of 2019, AT&T Says. *VARIETY*. 2020. Available online: https://variety.com/2020/digital/news/hbo-max-subscribers-subscribers-q2-att-1234714316/ (accessed on 19 April 2021).

[40] Kim, J.; Kim, S.; Nam, C. Competitive dynamics in the Korean video platform market: Traditional pay TV platforms vs. OTT platforms. *Telemat. Informat.* 2016, *33*, 711–721. [Google Scholar] [CrossRef]

[41] Stoll, J. Netflix Subscriptions in the U.S. 2020, by Generation. *Statista*. 2021. Available online: https://www.statista.com/statistics/720723/netflix-members-usa-by-age-group/#statisticContainer (accessed on 19 April 2021).

[42] AppAnnie. The State of Mobile 2020 Report. 2019. Available online: https://www.appannie.com/en/go/state-of-mobile-2020/ (accessed on 19 April 2021).

[43] Li, C.; Liu, S. A comparative study of the class imbalance problem in Twitter spam detection. *Concurr. Comput.* 2018, *30*, e4281.

[44] Liu, S.; Wang, Y.; Zhang, J.; Chen, C.; Xiang, Y. Addressing the class imbalance problem in twitter spam detection using ensemble learning. *Comput. Sec.* 2017, *69*, 35–49.

[45] Yang, Q.; Wu, X. 10 challenging problems in data mining research. *Int. J. Inf. Technol. Decis. Mak.* 2006, *5*, 597–604

[46] Inuwa-Dutse, I.; Liptrott, M.; Korkontzelos, I. Detection of spam-posting accounts on Twitter. *Neurocomputing* 2018, *315*, 496–511.

[47] Kudugunta, S.; Ferrara, E. Deep neural networks for bot detection. *Inform. Sci.* 2018, *467*, 312–322.

[48] Sze-To, A.; Wong, A.K. A weight-selection strategy on training deep neural networks for imbalanced classification. In Proceedings of the International Conference Image Analysis and Recognition, Montreal, QC, Canada, 5–7 July 2017; Karray, F., Campilho, A., Cheriet, F., Eds.; Springer: Cham, Switzerland, 2017; pp. 3–10.

[49] Han X, Zhu X, Pedrycz W, Li Z. A three-way classification with fuzzy decision trees. Applied Soft Computing. 2023 Jan 1;132:109788.

[50] Hu J, Szymczak S. A review on longitudinal data analysis with random forest. Briefings in Bioinformatics. 2023 Jan 18.

[51] Wang X, Wang X, Ma B, Li Q, Wang C, Shi Y. High-performance reversible data hiding based on ridge regression prediction algorithm. Signal Processing. 2023 Mar 1;204:108818.

[52] Helmina A, Lailani FK, Fatihaturahmi F, Jalinus N, Waskito W. Use of the K-Nearest Neighbor (KNN) Algorithm in New Categories Based on Books. Jurnal Pendidikan dan Konseling (JPDK). 2023 Jan 11;5(1):2570-9.

[53] Kitson, Neville Kenneth, Anthony C. Constantinou, Zhigao Guo, Yang Liu, and Kiattikun Chobtham. "A survey of Bayesian Network structure learning." *Artificial Intelligence Review* (2023): 1-94.

[54] Nasiboglu R, Nasibov E. WABL method as a universal defuzzifier in the fuzzy gradient boosting regression model. Expert Systems with Applications. 2023 Feb 1;212:118771.

[55] Han Y, Kim J, Enke D. A machine learning trading system for the stock market based on N-period Min-Max labeling using XGBoost. Expert Systems with Applications. 2023 Jan 1;211:118581.

[56] Kurani A, Doshi P, Vakharia A, Shah M. A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting. Annals of Data Science. 2023 Feb;10(1):183-208.