Bachelor's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Economics and Management
NTNU Business School

Sindre Lavrans Nygaard Engen

Tackling the Game: Exploring the Factors Influencing Sports Piracy Streaming

Bachelor's thesis in Business Administration, Business Analytics
Supervisor: Denis Becker
April 2023

NTNU
Norwegian University of
Science and Technology

Sindre Lavrans Nygaard Engen

Tackling the Game: Exploring the Factors Influencing Sports Piracy Streaming

Bachelor's thesis in Business Administration, Business Analytics
Supervisor: Denis Becker
April 2023

Norwegian University of Science and Technology
Faculty of Economics and Management
NTNU Business School

NTNU
Norwegian University of
Science and Technology

PREFACE

I have always been interested in the intersection of sports, business, and analytics. This bachelor thesis is a testament to my interest in these fields, and over the course of my education, I have become increasingly cognizant of the escalating issue of piracy streaming, particularly within the realm of sports.

By employing different analytical techniques, I aim to offer insights into the factors and variables that influence piracy streaming. This research could have practical implications for industry stakeholders who are looking to reduce piracy and promote the legal streaming of sports events.

I am grateful for the support and guidance provided by my supervisor, Denis Becker, whose expertise, and mentorship have been instrumental in providing the foundation needed to write this thesis. The course Essentials of Business Analytics (BBAN3001) has been a highlight of my education, sparking my interest in the field and motivating me to pursue further studies that involve analytics.

Finally, I want to express my gratitude to Jon Martin Denstadli and Harry Arne Solberg, who provided me with the necessary data for this thesis. Their contributions and guidance have been critical in enabling this study.

ABSTRACT

This study investigates factors influencing sports piracy streaming behavior and develops accurate predictive models. Utilizing a dataset comprised of 330 respondents, this study implemented logistic regression, Naïve Bayes, and decision tree classifier models to analyze the relationships between variables and piracy streaming.

Results reveal that future intentions to engage in piracy streaming, interest in foreign football leagues, and social influence are significant predictors of piracy streaming behavior. The logistic regression and Naïve Bayes model achieved high accuracies.

The findings support the Theory of Planned Behavior and the Issue-Risk-Judgment model, emphasizing the importance of understanding consumer behavior and social influences in anti-piracy measures. Future research should address study limitations, explore alternative algorithms, and evaluate intervention strategies.

SAMMENDRAG

Denne oppgaven undersøker faktorer som påvirker piratstrømming av sport. Ved hjelp av et datasett med 330 respondenter ble logistisk regresjon, Naïve Bayes og decision tree classifier benyttet for å analysere forholdet mellom forskjellige variabler og piratstrømming.

Resultatene viser at fremtidige intensjoner om å piratstrømme, interesse for utenlandske fotballigaer og sosial påvirkning er betydelige faktorer som påvirker piratstrømming. De prediktive analysene som ble gjennomført oppnådde svært høy nøyaktighet.

Funnene i denne oppgaven støtter teorien om Planned Behavior og Issue-Risk-Judgment, samt understreker betydningen av å forstå forbrukeratferd og sosiale påvirkninger i strategier for å redusere piratstrømming. Fremtidig forskning bør ta for seg oppgavens begrensninger, utforske alternative algoritmer, samt implementere og evaluere mulige strategier som hindrer piratstrømming.

LIST OF FIGURES AND TABLES

TABLE OF CONTENTS

1. INTRODUCTION

The rapid evolution of digital technology has led to a revolution in the way people access and consume content, offering increased accessibility to media and entertainment worldwide. Amidst these advancements, piracy streaming has emerged as a critical challenge for content creators, distributors, and rights holders, particularly in the realm of sports broadcasting. The exclusive nature and high demand for live sports events have made piracy streaming a significant concern for stakeholders in the sports industry, exacerbated by soaring consumer costs and increasingly exclusive rights holders.

This bachelor thesis aims to explore the multifaceted issue of piracy streaming in sports, delving into the factors that contribute to its prevalence which can identify potential strategies to address piracy streaming and safeguard the interests of stakeholders in the industry.

The thesis will begin by explaining the sports piracy ecosystem and will move further by going through relevant theories and literature. Furthermore, it will focus on the data and the methods used in the thesis, as well as present the results of the analyses. Lastly, the results will be discussed, and the thesis will conclude by answering which factors that contribute to piracy streaming, and potential strategies to address piracy streaming.

2. THE SPORTS PIRACY ECOSYSTEM

Piracy streaming refers to the unauthorized distribution and consumption of movies, TV series, and other media content by a distributor who lacks the consent of the original licensee. In the context of sports, piracy streams are typically consumed through various methods (Wong, 2015). These methods include live streaming via the internet through television services or web servers, which bypasses authorized distribution channels; recorded telecasts made available on file-sharing networks, enabling users to access and download content without proper authorization; highlights generated and uploaded to various websites, infringing on the original rights holder's exclusive distribution rights; and illegal set-top boxes and "signal boxes" that unscramble encrypted signals of pay-TV content, providing unauthorized access to viewers.

The process of piracy streaming commences with a pirate, an individual who either steals cable and satellite feeds for redistribution or rips online streams from official sites using advanced coding skills or screen-recording techniques (Bushnell, 2019). Once the pirate obtains the stream, it is typically redistributed automatically through various websites. The stolen video or stream is often hosted on a separate site but embedded on a destination site, which may claim no legal responsibility for the content. These destination sites, where consumers access piracy streams, frequently exist in legal gray areas. Some sites disguise themselves as informational websites, with the owners arguing that redistribution of copyrighted content is not the site's primary purpose. This tactic makes it challenging for licensees and police to lawfully dismantle such operations.

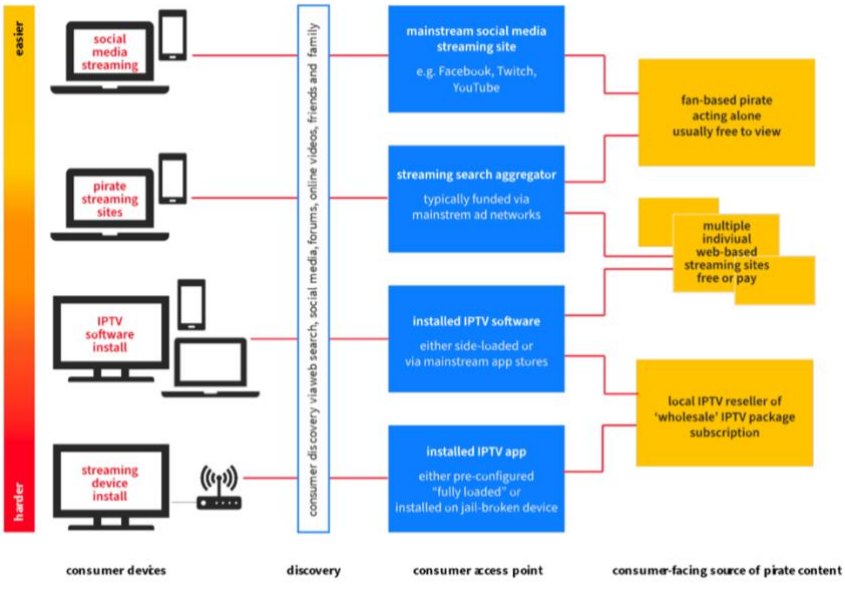


Figure 1: How fans gain access to pirated material (Ampere Analysis, 2020)

Piracy streaming poses significant financial challenges for the sports media industry. A recent report estimates that the sports media industry, which is valued at $50 billion, loses an estimated $28 billion annually to piracy (Ampere Analysis, 2021). In an environment often characterized by the "Winner's Curse", such substantial losses could potentially result in bankruptcy and layoffs for licensees. A more immediate concern arising from piracy streaming is its unregulated black market nature. It is reasonable to assume that criminal actors play a significant role in piracy streaming operations. If the estimated losses translate directly into net income for pirates, even a fraction of the total amount could be used to cause considerable harm by criminal entities. Furthermore, piracy streaming can result in higher costs for legitimate consumers. For instance, Viaplay was forced to raise the monthly fee for a Premier League subscription by $10 due to an insufficient number of subscribers (Jerijervi, 2023).

3. THEORY AND LITERATURE REVIEW

In this chapter, we explore the underlying factors and examine various theories that can help explain the complex issue of piracy streaming. By analyzing the findings from Synamedia and LEADERS reports (Synamedia, 2020; Synamedia, 2021), as well as drawing from the Theory of Planned Behavior and Benjamin Tan's Issue-Risk-Judgment Model (Ajzen, 1991; Tan, 2022), we seek to gain a better understanding of what drives consumers to engage in piracy streaming.

3.1 Synamedia and LEADERS Reports

In the Synamedia report on global sports piracy, it is revealed that there are eight key factors driving individuals to engage in piracy (Ampere Analysis, 2020). These factors include the desire to watch content not included in their subscriptions, the lack of legitimate providers for an event in a given country, the need for multi-device or non-TV viewing options, the preference for simple user experiences without contracts or installations, the unwillingness to commit to a legal subscription for certain sports or events, limited funds leading to cheaper alternatives, resentment towards paying for sports content, and the need for access to major sports events or complete games in a league where rights are split among multiple providers.

The report also identifies three different segments of consumers: Casual spectators, loyal stalwarts, and fickle superfans. Casual spectators, making up 43% of respondents, are occasional sports fans who prefer tournaments over leagues and mainly have basic or free TV. They are typically over 55 years old. Loyal stalwarts, comprising 26% of respondents, have traditional viewing habits, watching pay TV sports channels on their main TV screen at home. They are often older middle-aged males with families. Lastly, fickle superfans, accounting for 31% of respondents, enjoy a diverse range of sports and engage in multiscreen viewing on various devices, both at home and outside. They are usually in their 20s and early 30s.

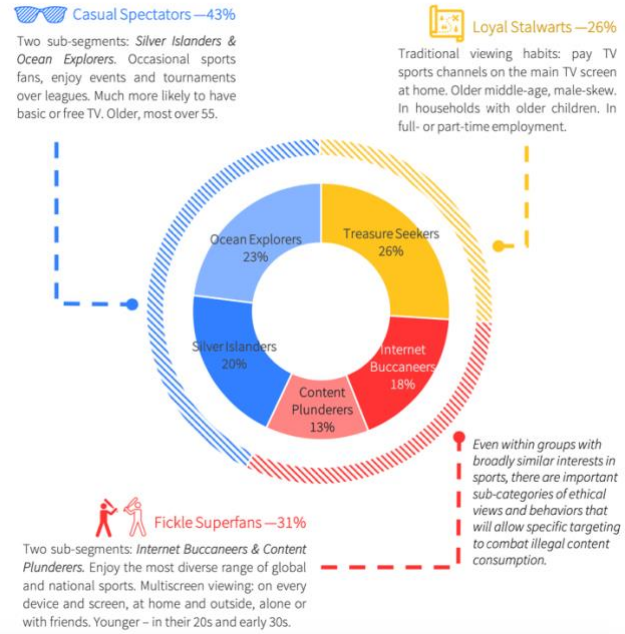


Figure 2: Three different segments of consumers (Ampere Analysis, 2020)

A LEADERS report echoes many of the same triggers that lead consumers to seek out piracy streaming, such as flexibility, ease of use, availability, content choice, and price (Synamedia, 2021). The report also highlights three main challenges for corporations losing to piracy streams: A lack of awareness, failure to prioritize, and increased vulnerability on multiple fronts.

3.2 Theory of Planned Behavior

The Theory of Planned Behavior (TPB), developed by Icek Ajzen, provides another possible explanation for piracy streaming (Ajzen, 1991). A central factor in TPB is an individual's intention to perform a specific behavior. Intensions represent all the motivational factors that influence a behavior, reflecting how much effort people are willing to exert to perform that behavior. The stronger the intention to engage in a behavior, the more likely the behavior is to be performed. However, the behavior must be under voluntary control.

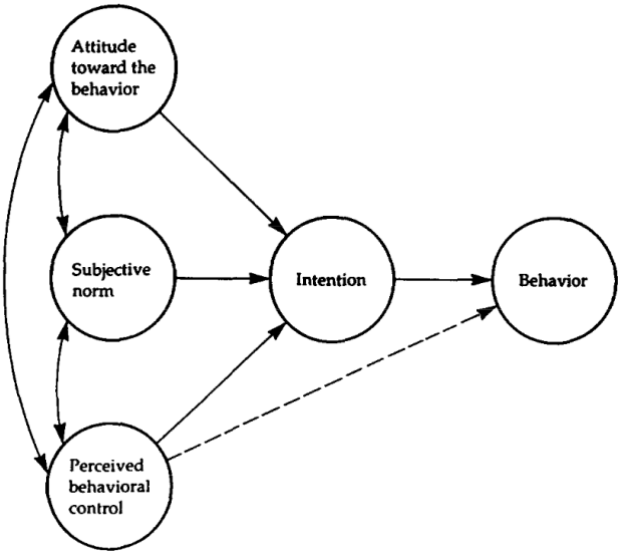


Figure 3: Theory of planned behavior (Ajzen, 1991)

While intentions are crucial, most behaviors also depend on non-motivational factors, such as opportunity and resources. The combination of motivational and non-motivational factors represents an individual's actual control over a behavior. If the required resources and opportunities are available and the intention to perform the behavior is present, the person should be able to carry out the behavior.

Actual behavioral control influences the likelihood of achieving a behavior but perceived behavioral control, or the individual's perception of their ability to perform the behavior is even more critical. This concept, known as perceived self-efficacy, suggests that people's behavior is strongly influenced by their confidence in their ability to perform it. In TPB, the construct of self-efficacy or perceived behavioral control is placed within a more general framework that outlines the relationships among beliefs, attitudes, intentions, and behavior.

TPB posits that perceived behavioral control, along with behavioral intention, can be used to directly predict behavioral achievement. This hypothesis is based on two rationales. First, when holding intention constant, the effort to successfully perform the intended behavior is likely to increase with perceived behavioral control. For example, if two individuals have equally strong intentions to learn something and both attempt to do so, the one with greater confidence in their ability to succeed is more likely to achieve it. Second, perceived behavioral control can often serve as a substitute for measuring actual control, provided that the individual's perceptions are accurate. The accuracy of these perceptions depends on the person's information about the behavior, changing requirements or resources, and the presence of new or unfamiliar elements in the situation.

3.3 Benjamin Tan's Issue-Risk-Judgment Model

Benjamin Tan's theory, which aims to understand consumer ethical decision-making about the purchase of pirated software, may also apply to piracy streaming (Tan, 2002). Tan's issue-risk-judgment model examines the impact of moral intensity, perceived risks, and moral judgment on the consumer's ethical decision-making process. The model also considers situational variables such as price, gender, age, education, income, and past purchase experience.

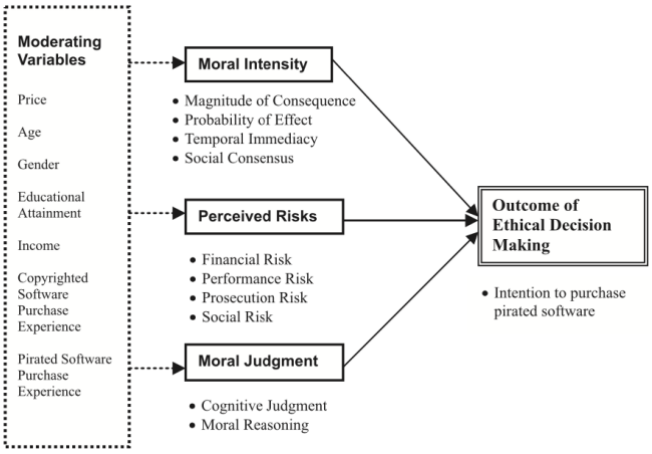


Figure 4: Issue-Risk-Judgment Model (Tan, 2002)

Several studies have explored the role of moral intensity in ethical decision-making, supporting the issue-contingency nature of ethical decisions. Four components of Jones' issue-contingent model are relevant in the context of Tan's theory: Magnitude of consequences, probability of effect, temporal immediacy, and social consensus. The magnitude of consequences assesses the extent to which consumers are aware that their purchase of pirated software negatively affects the original software producers (copyright holders). The probability of effect asks consumers to consider the likelihood that these adverse effects will occur. Temporal immediacy concerns

the immediacy of the time lapse between the consumer's action and the consequence of the action. Social consensus encourages consumers to contemplate their behavior's acceptance or rejection within a social environment, such as among family members and friends.

Numerous studies have also examined how perceived risks influence consumer decisions and behavior. Three of the six common aspects of risk are relevant in the context of software piracy: Performance, financial, and social risk. Performance risk applies because pirated products lack a warranty and may not function as expected. Financial risk is relevant due to potential time loss and incidental expenses incurred if the pirated products fail to perform as anticipated, necessitating the reinstatement of computers and data systems. Social risk is applicable as consumers may be conscious of the image they project to their peers. Additionally, Tan argues that prosecution risk should be included, as installing pirated software infringes copyright law but is unlikely to result in physical risk.

Research in cognitive moral development has consistently demonstrated a direct relationship between higher stages of moral judgment and increased ethical behavior. Ethical decision-making, reasoning, and intended ethical behavior generally rise as individuals utilize higher stages of moral reasoning. The connection between moral thinking and moral behavior is widely supported as significant.

Furthermore, the model considers moderating variables, such as the influence of situational factors affecting consumers' decision-making process and the contingent effect. Studies have shown that ethical behavior varies according to consumer demographics, including gender, age, and education. Additionally, consumers' past buying behavior and available economic incentives may have a contingent effect on ethical decisions. Consumers who have previously purchased copyrighted or pirated software are likely to repeat the buying pattern when acquiring new software. A larger price difference also presents a greater economic incentive to act unethically.

3.4 Methods and Data in Theory and Literature

The Synamedia report and the LEADERS report are based on an online quantitative study involving more than 6,000 sports fans aged 18-64 (Ampere Analysis, 2020). Synamedia commissioned the study, which was conducted by Ampere Analysis (Ampere Analysis, 2021). The study took place in March 2020, just before the coronavirus-related lockdown and the suspension of multiple sports. Ten markets/countries were included in the study: Brazil, Egypt, Germany, India, Italy, Jordan, Malaysia, Saudi Arabia, the UK, and the US.

Respondents were selected based on their experience watching sports on TV. A k-means cluster analysis was used to create segments by grouping similar consumers into distinct groups. The analysis resulted in five distinct groups of consumers, each exhibiting varied attitudes and behaviors regarding accessing illegal sports content streams. Six different clustering criteria were used: Pay TV and OTT services in the home, passion for watching sports, frequency of accessing illegal content, types of services and devices used to watch illegal content, drivers of using services and devices to watch illegal content, and opinions on the ethics of consuming illegal content. Additionally, industry interviews were conducted throughout the process to aid in survey design, report formulation, and ensuring stakeholder concerns were adequately reflected.

TPB is based on previous research in social psychology, including studies on attitudes, norms, and behavior (Ajzen, 1991). The data used to develop the theory came from several sources, including studies on people's attitudes and beliefs toward specific behaviors, their perceptions of social norms, and their perceived control over their actions. The data was collected using self-reported measures such as surveys, questionnaires, and interviews. These measures were designed to assess people's intentions, attitudes, subjective norms, and perceived behavioral control. The data was then analyzed using statistical techniques such as regression analysis to determine the relationship between these variables and behavior.

Benjamin Tan's Issue-Risk Judgment model focuses on three factors: Moral intensity, perceived risk, and moral judgment (Tan, 2002). Tan used a self-administered questionnaire featuring scenarios in which the respondent had to imagine themselves needing word-processing software but do not possess it. Respondents then had to choose between purchasing original (copyrighted) software or pirated software.

The copyrighted software was priced at $250 and $600, while the pirated software cost $50. Respondents were randomly assigned to the three different price levels, and participation was voluntary. The survey was pre-tested in three focus groups to evaluate the validity and ensure respondents understood the scenarios. A stratified random sampling approach was adopted to obtain a sample of respondents with a distribution that approximated population statistics. Out of 950 identified respondents, 400 responses were received, and 23 were rejected due to incomplete responses. A two-step hierarchical regression analysis was employed as the appropriate method of statistical estimation.

3.5 Results and Implications

The Synamedia report concludes with several solutions and industry implications (Ampere Analysis, 2020). It suggests that rights holders, buyers, and distributors collaborate with industry technology providers to address the needs of each consumer group individually. This involves creating focused product offerings, ensuring content is fully protected across all platforms, networks, and devices, and concentrating on security solutions that enable flexible access and payment models, protect multiscreen access, and combat illegal access technologies. Importantly, the report emphasizes the need to implement these solutions without increasing access complexity or frustrating paying sports fans to maximize disruption of the piracy ecosystem.

The LEADERS report identifies deterrents to reduce consumption of pirate sports services, such as imposing fines, emphasizing team financial impact, banning supporters, delivering poor video quality, and stopping streams midgame (Ampere Analysis, 2021). The report also presents strategies for fighting piracy, including monitoring, pre-breach security, post-breach security, disruption, contractual obligations, dynamic IP blocking, administrative blocking, notice and takedown, awareness, and strategic prioritization, by shifting anti-piracy measures from the cost column, and establishing a unified, funded body specifically targeting piracy.

TPB proposes that perceived behavioral intention, along with perceived behavioral intention, can directly predict behavioral achievement. There are two rationales for the relationship between perceived behavioral control and behavioral achievement. First, when holding intention constant, the more control a person perceives they have over behavior, the more likely they are to make an effort to perform it successfully. Second, perceived behavioral

control can serve as a substitute for actual control in predicting behavior, as long as people's perceptions of control are accurate. These perceptions can be influenced by a person's knowledge about the behavior, the resources available to them, and changes in the situation.

Benjamin Tan's Issue-Risk Judgment model demonstrates that, when testing the effect of moral intensity, only social consensus and the magnitude of consequences are significantly correlated with consumers' purchase intention (Tan, 2002). The significance of magnitude of consequences suggests that if consumers perceive a higher moral imperative regarding pirated software purchases, they are more likely to avoid making an unethical decision. Social consensus negatively influences consumers' purchase intentions, indicating that social acceptability reduces an individual's probability to act unethically, and a high degree of social consensus diminishes the likelihood of ambiguity. All components of perceived risk significantly influence consumers' purchase intention. Cognitive judgment and moral reasoning are significant predictors of consumers' purchase intention, supporting the connection between consumers' moral judgment and ethical decision-making. Additionally, several variables are significantly related to consumers' purchase intentions, such as price, age, gender, purchase experience of copyrighted software, and purchase experience of copyrighted software. Consumers who have previously purchased copyrighted software are less likely to purchase pirated software, while those with a history of pirated software purchases are more likely to continue buying pirated software.

4. DATA

4.1 Dataset

The dataset used in this thesis was acquired via convenience sampling in collaboration with several professors from different universities and colleges. The dataset originates from a digital survey distributed by these professors to their students, rendering the calculation of a response rate infeasible. Although a total of 330 responses were received, the number of responses is relatively low considering the number of people it reached out to, which is primarily due to the extensive time required for completion, leading to some students not completing the survey at all. The dataset contains 60 variables, and the dependent variable is *'pirat_selv'*, which indicates whether the respondent has engaged in piracy streaming or not.

4.2 Data Cleaning

The initial dataset required cleaning before analysis due to the presence of incorrect entries, such as strings, zero values, and other errors. We visualized all unique values in each variable, manually checking for inconsistencies. The dataset contained numerous blank values, primarily attributable to the survey's conditional questions. However, some blank and incorrect values unrelated to the conditional questions required addressing first. Chapter 9.1 shows all the variables included in the dataset, in addition to their descriptions.

The three *'intensjon'* variables about intent to piracy stream (*'intensjon_1'*, *'intensjon_2'*, and *'intensjon_3'*) have some blank values that are unrelated to conditional questions. Since respondents were to answer these questions on a scale of 1 to 7, with 4 being neutral, we set all blank values to 4. As there is only one blank value in each of the three variables, this change to the average is unlikely to significantly affect the results. Additionally, we converted the variable values to integers.

The *'income'* and *'wtp_PL'* variables also contained blank values unrelated to the conditional questions. The variables have respectively one and three blank values, which we replaced with the average of the non-blank values in the respective variables, ensuring the substitution would not significantly impact the results. Furthermore, we converted all values in these two variables to integers.

We amended a value in *'tv_kamper_utl_sett'* to correspond with the intended range of 1-6, as the dataset contained values of 1-5 and 9. We replaced all instances of 9 with 6 and converted all values to integers. Additionally, as value 6 is representing "I don't know", we assigned all 84 blank values to 6 as that is the value that corresponds the best with the blank values. Furthermore, the variable *'svartid'* was removed due to its irrelevance to the research question and difficulty in converting to integers. The variable contained text strings representing the time spent answering the survey.

The conditional questions provide a challenge, as we needed to either include or exclude all blank values resulting from these questions. To perform a logistic regression analysis, we removed 16 variables that contained numerous blank values due to conditional questions and were non-essential to our research objective. We also removed *'wtp_utl'* and *'wtp_no'* as they contained respectively 82 and 133 blank values. We could have converted them to the average non-blank values, but that significantly impact the results. Furthermore, we removed *'piratstrøm_antall'* as the variable is correlating too much with the dependent variable. We then proceeded to clean three critical variables *('persgevinst'*, *'persrisiko_1'*, and *'persrisiko_2'*) with numerous blank values due to the third and final conditional question. Each variable contained 144 blank values, and the blank values were converted to 4 ("Verken eller"). Lastly, we converted the dependent variable, which originally had three different values ("Yes", "No", and "Idk"), to a dummy variable. This conversion allowed for the use of logistic models, as "Idk" essentially equated to "No". Lastly, we converted all values to integers, resulting in a dataset with 330 entries and 40 variables in our dataset.

4.3 Limitations

The sample size of 330 respondents is probably relatively small compared to the number of people who were reached out to. This could affect the statistical power of the analyses and the reliability of the results, making it difficult to detect small effects or identify rare patterns. Furthermore, the dataset contained several instances of missing values, either due to conditional questions or other issues. While some blank values were addressed by replacing them with the average or neutral value, this approach might introduce bias into the dataset and significantly impact the results.

5. METHODS

5.1 Logistic Regression Model

We will use a logistic regression model to analyze the data as the dependent variable is a dummy variable. Logistic regression is a generalized linear model that utilizes the logistic function to model the probability of an outcome variable (Friedman and Hastie and Tibshirani, 2008, p. 119). The logistic function is formulated as:

$$\log \frac{\Pr\,(G = K - 1|X = x)}{\Pr(G = K|X = x)} = \beta_{(K-1)0} + \beta_{K-1}^{T}\, x$$

The logistic regression model was constructed and trained with statsmodels and the scikit-learn API library.

5.1.1 Lasso

Lasso is utilized in the analysis because of issues with quasi separation and multicollinearity, which potentially could lead to inflated or unreliable coefficient estimates. Lasso, or Least Absolute Shrinkage and Selection Operator, is a regularization technique that incorporates an L1 penalty term into the objective function (Friedman and Hastie and Tibshirani, 2008, p. 125). In the context of logistic regression, lasso is formulated as:

$$\max_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} \left[y_i(\beta_0 + \beta^T x_i) - \log(1 + e^{\beta_0 + \beta^T x_i}) \right] - \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

5.1.2 Pseduo R^2

Pseudo R^2 serves as the goodness of fit metric for logistic regressions, and the metric gives an approximate accuracy of the logistic regression model. with McFadden's Pseudo R^2 being the default approach (Shafrin, 2016). It is expressed as (Wikipedia, 2023):

$$R^2_{McF} = 1 - \frac{\ln L_M}{\ln L_0}$$

5.1.3 Coefficients and p-values

Coefficients represent the relationship between predictor and outcome variables on a logistic scale, while p-values are utilized to determine the significance of these relationships.

5.1.4 Assumptions to Logistic Regression Model

To get accurate, reliable, and interpretable results, several assumptions underlie logistic regression models (Statistics Solutions, No date):

1. *Binary outcome*

The dependent variable must be binary (0 or 1) as logistic regression is designed to model success or failure probabilities based on predictor variables.

2. *Independence of observations*

The observations should be independent of each other. This means that there is no correlation or clustering among the observations that might cause issues in the model estimation.

3. *No multicollinearity*

The predictor variables should not be highly correlated. Multicollinearity can lead to unstable estimates and inflated standard errors.

4. *Linearity of log-odds*

There must be a linear relationship between the log-odds of the outcome and the predictor variables. The log-odds of the outcome must change linearly with the predictors.

5. *Large sample size*

Large sample size is required to ensure reliable parameter estimates. A small sample size might lead to overfitting or biased estimates.

5.2 Naïve Bayes

Naïve Bayes is a predictive model and a probabilistic classification method based on Bayes' theorem (Friedman and Hastie and Tibshirani, 2008, p. 210). It assumes that all the independent variables are independent of each other, given the class variable. Bayes theorem can be expressed as follows:

$$P(C|X) = \frac{P(X|C) * P(C)}{P(X)}$$

5.3 Decision Tree Classifier

A decision tree classifier is a hierarchical mode used for classification and regression tasks. It recursively splits the data into subsets based on the values of the input features, making a decision at each node of the tree (Friedman and Hastie and Tibshirani, 2008, p. 308). The goal of the decision tree is to create a model that predicts the value of the target variable based on the input features.

5.3.1 Gini Impurity

Gini impurity is a measurement used to build decision trees to determine how the features of the dataset should split nodes to form the tree (Karabiber, Unknown). Gini impurity is expressed as follows:

$$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2$$

5.3.2 K-Folds Cross-Validation

K-fold cross-validation is a technique that partitions the dataset into equal-sized subsets, or "folds" (Friedman and Hastie and Tibshirani, 2008, p. 241). The model is then trained and tested k times, each time using a different fold as the test set and the remaining k-1 folds as the training set. The model's overall performance is estimated by averaging the test accuracies obtained in each of the k iterations.

5.4 Limitations

The dataset was collected via convenience sampling in collaboration with professors from different universities and colleges, which might not be representative of the broader population engaged in piracy streaming. This could limit the external validity of the survey's findings and introduce potential biases related to demographics or other factors. Moreover, the dataset is a snapshot of a specific point in time, and the piracy streaming landscape is constantly evolving due to advancements in technology, changes in legislation, and shifts in consumer behavior. The survey's findings might not apply to different periods or contexts, limiting the generalizability of the results.

The models used in this thesis all rely on certain assumptions. Three of the five assumptions for logistic regression are met, but some issues with the independence of observations (quasi-separation) and multicollinearity necessitated the use of lasso regularization to reduce the model to only six variables. Although the value of quasi-separation is much lower with only six variables, there is still a grade of quasi-separation present, which could compromise the performance of the logistic regression model. The other models' performances might also be compromised if their assumptions are violated.

There might be external factors or unmeasured variables that could have affected piracy streaming behavior but were not included in the analyses. These factors might confound the relationships between the observed variables and piracy streaming behavior. Additionally, hindsight bias must be kept at a minimum to stop significant implications for decision-making, learning, and interpersonal relationships.

6. RESULTS AND DISCUSSION

6.1 Correlation

Pearson Correlation Matrix

	pirat_selv	perskontroll_1	intensjon_1	intensjon_3	interesse_4	perskontroll_2	intensjon_2	kjønn	pirat_venner	interesse_5	holdning_piracy_1
pirat_selv	1.000000	0.713463	0.679683	-0.659906	0.654987	0.651104	0.624263	0.571324	-0.570700	0.560002	0.541370
perskontroll_1		1.000000	0.646875	-0.597463	0.579222	0.850893	0.647937	0.598588	-0.603412	0.525375	0.602050
intensjon_1			1.000000	-0.639544	0.496722	0.588178	0.871513	0.450735	-0.419776	0.433708	0.626547
intensjon_3				1.000000	-0.457588	-0.506476	-0.673001	-0.475252	0.496797	-0.402155	-0.580763
interesse_4					1.000000	0.513662	0.471079	0.551103	-0.472546	0.821479	0.378655
perskontroll_2						1.000000	0.557291	0.529113	-0.565048	0.476256	0.536010
intensjon_2							1.000000	0.445861	-0.430548	0.426585	0.654597
kjønn								1.000000	-0.390731	0.518155	0.429943
pirat_venner									1.000000	-0.435169	-0.385787
interesse_5										1.000000	0.359175
holdning_piracy_1											1.000000

Table 1: The independent variables with the 10 highest absolute correlation values

The results from the correlation matrix (Table 1) show a high correlation between the dependent variable, and the *'perskontroll'* and *'intensjon'* variables, which are about the respondent's knowledge and ability to find piracy streams, and their intentions to engage in piracy streaming in the future if given the opportunity. This suggests that those who have piracy streamed tend to be more positive about piracy streaming than those who haven't. It also suggests that those whom piracy stream exhibit a higher degree of comfort in continuing this behavior in the future, indicating that there are minimal barriers or potential penalties that would discourage them from accessing and viewing pirated content.

Additionally, there is a high correlation between the dependent variable and the 'interesse' variables, specifically *'interesse_4'* and *'interesse_5'*. These variables are about the respondent's interest in foreign football leagues and their interest in the Norwegian men's national team. The results imply that as the interest in foreign leagues and the Norwegian men's national team increases, so does the likelihood of engaging in piracy streaming. This may suggest that the growing interest in foreign leagues could be a significant factor contributing to the increase in piracy streaming.

Lastly, there is a substantial correlation between the dependent variable and the variables *'kjønn'* and *'pirat_venner'*. These results suggest that men are more likely to piracy stream than women and that the likelihood of an individual engaging in piracy streaming increases if their friends or acquaintances do so. This finding suggests that social influence and peer behavior might be significant factors in driving piracy streaming. In addition, it is a

substantial correlation between the dependent variable and *'holdning_piracy1'*. This is a natural correlation as a more positive attitude toward piracy streaming would naturally increase the likelihood of an individual engaging in piracy streaming.

6.2 Logistic Regression Model

Logit Regression Results

Dep. Variable:	pirat_selv	No. Observations:	330
Model:	Logit	Df Residuals:	324
Method:	MLE	Df Model:	5
Date:	Sun, 16 Apr 2023	Pseudo R-squ.:	0.7245
Time:	10:01:02	Log-Likelihood:	-62.286
converged:	True	LL-Null:	-226.06
Covariance Type:	nonrobust	LLR p-value:	1.191e-68

	coef	std err	z	P>|z|	[0.025	0.975]
perskontroll_1	0.7974	0.459	1.738	0.082	-0.102	1.696
intensjon_1	1.0397	0.317	3.279	0.001	0.418	1.661
perskontroll_2	0.5297	0.437	1.213	0.225	-0.326	1.386
intensjon_3	-1.3848	0.334	-4.144	0.000	-2.040	-0.730
interesse_4	1.6836	0.313	5.378	0.000	1.070	2.297
pirat_venner	-0.8192	0.334	-2.452	0.014	-1.474	-0.165

Table 2: Logistic Regression Model Results

The Pseudo R^2 value in the logistic regression model is 0.7245, which indicates that the model accounts for 72.45% of the variance in the dependent variable. The relatively high value suggests that the model is reasonably effective in capturing the relationship between the dependent and the selected independent variables.

The LLR p-value is extremely small and effectively zero. This suggests that the independent variables in the model significantly contribute to the prediction of the outcome, thereby lending support to the overall validity of the logistic regression model.

Interpretation:

- *'perskontoll_1'*: The coefficient is 0.7974, and the p-value is > 0.05. This indicates that the respondent's knowledge of how to piracy stream is not significantly related to the dependent variable.
- *'intensjon_1'*: The coefficient is 1.0397, and the p-value is < 0.05. This indicates a significant positive relationship between the respondent's plans to piracy stream and the dependent variable. As the variable increases by one unit, the log-odds of the dependent variable increase by 1.0397, holding other variables constant.
- *'perskontroll_2'*: The coefficient is 0.5297, and the p-value is > 0.05. This indicates that the respondent's ability to easily find piracy streams is not significantly associated with the dependent variable.
- *'intensjon_3'*: The coefficient is -1.3848, and the p-value is < 0.05. This indicates a significant negative relationship between the respondent's intentions of never engaging in piracy streaming and the dependent variable. As the variable increases by one unit, the log-odds of the dependent variable decreases by 1.3848, holding other variables constant.
- *'interesse_4'*: The coefficient is 1.6836, and the p-value is < 0.05. This indicates a significant positive relationship between the respondent's interest in foreign football leagues and the dependent variable. As the variable increases by one unit, the log-odds of the dependent variable increase by 1.6836, holding other variables constant.
- *'pirat_venner'*: The coefficient is -0.8192, and the p-value is < 0.05. This indicates a significant negative relationship between the respondent's friends and acquaintances engaging in piracy streaming and the dependent variable. As the variable increases by one unit, the log-odds of the dependent variable decreases by 0.8192, holding other variables constant.

Furthermore, Figure 5 presents the confusion matrix for the logistic regression model, illustrating the model's classification performance. The model has an accuracy of 0.89, meaning it can predict the respondents' engagement in piracy streaming with 89% accuracy. From a total of 66 inputs, there are 59 accurate predictions and 7 wrong predictions. Of these 7, there are 2 false positives and 5 false negatives (Wikipedia, 2023).

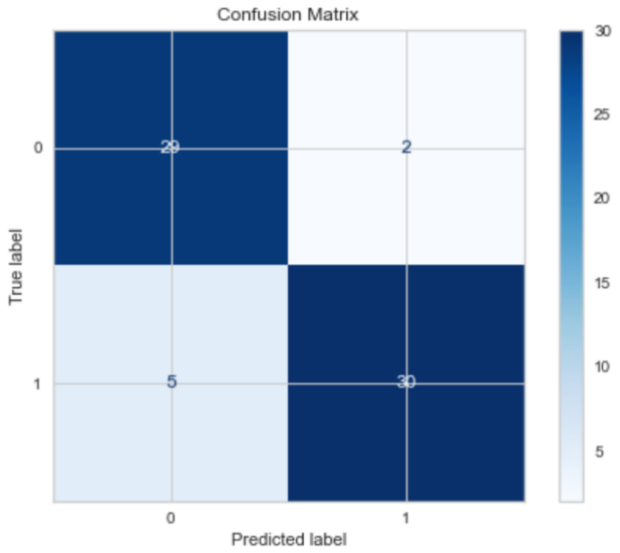


Figure 5: Logistic Regression Model Results

The following classification report offers a comprehensive evaluation of the model's performance, detailing precision, recall, and F1-score metrics for both classes – piracy streamers and non-piracy streamers - along with the overall accuracy (Table 4 in Appendix). The report shows that the model is performing well in classifying both classes, with an overall accuracy of approximately 89%. The high values for precision, recall, and F1-score across both classes indicate that the model has managed to minimize misclassification by finding a balance between false positives and false negatives. In summary, the classification report corroborates the findings illustrated by the confusion matrix.

6.3 Naïve Bayes

The confusion matrix for the Naïve Bayes model reveals that out of the 63 inputs, there are 62 correct predictions and 1 false negative (Figure 8 in Appendix). Consequently, the model exhibits an accuracy rate of 98%.

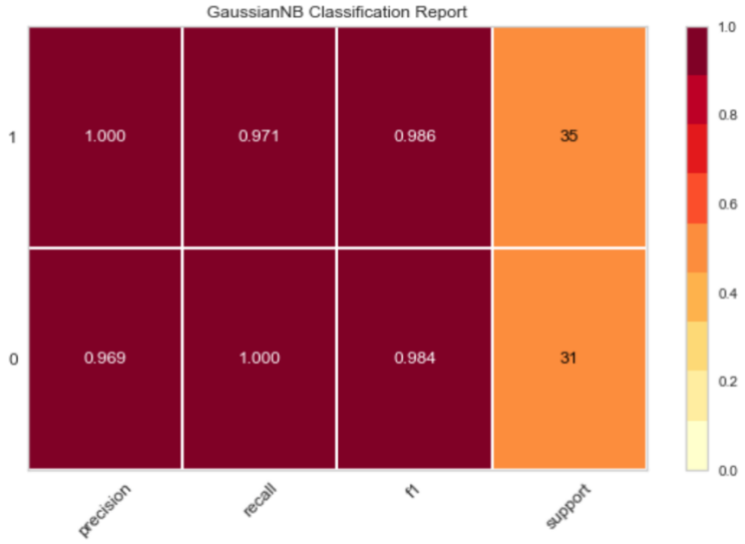


Figure 6: Naïve Bayes Classification Report

As illustrated by the classification report, the Naïve Bayes model achieves perfect precision for class 1 and perfect recall for class 0, while having values over 0.969 for all other metrics. To verify the absence of overfitting, cross-validation scores, mean accuracy, and standard deviation were assessed. Two of the five folds attained a perfect accuracy of 100%, with the lowest accuracy score being 96.96% (Figure 9 in Appendix). Furthermore, the mean accuracy is 98.79%, accompanied by a standard deviation of 1.13% (Figure 10 in Appendix). These findings suggest that the Naïve Bayes model is performing consistently well across various data subsets, indicating successful generalizing.

6.4 Decision Tree

The decision tree has an accuracy of 85.85%, with the root node splitting the data based on the variable 'intensjon_1' using a condition of <= 1.5. The tree has 13 internal nodes and 15 leaf nodes, which means that the tree utilizes 13 unique variables.

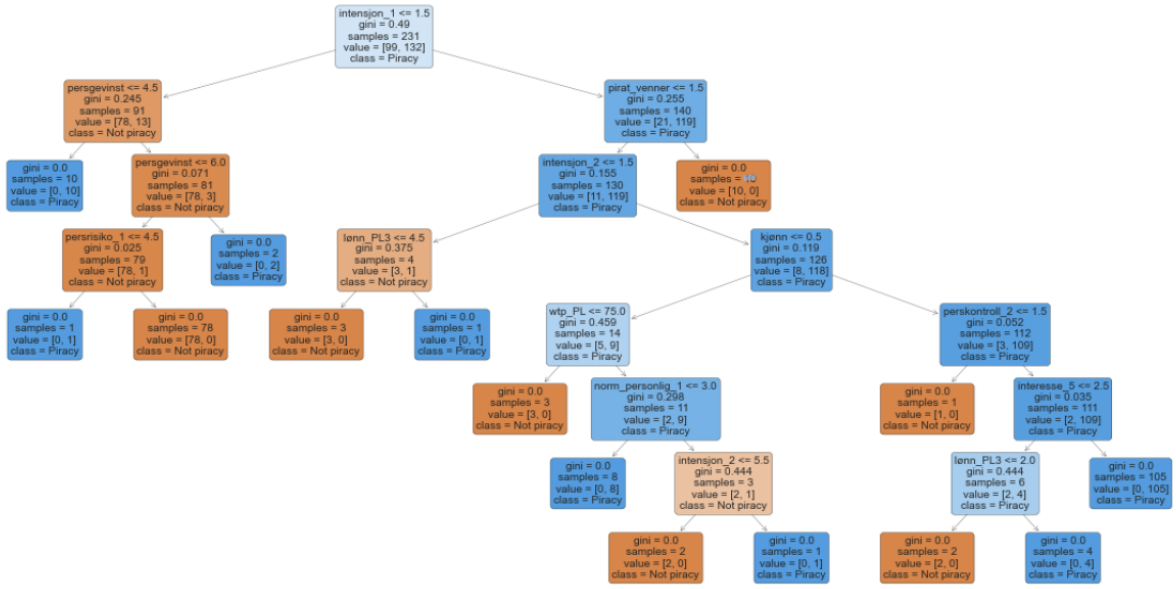


Figure 7: Decision Tree

According to the feature importance in the decision tree, *'intensjon_1'* emerges as the most crucial variable for making decisions when the tree is classifying the data points. The variable has a value of 0.4875 which significantly surpasses the subsequent highest value (Figure 11 in Appendix). The next most influential features are *'persgevinst'* and *'pirat_venner'*, with values of respectively 0.1795 and 0.1375. Additional, albeit less influential, feature importances with values below 0.1 include '*intensjon_2'*, *'perskontroll_2'*, *'wtp_PL'*, *'kjønn'*, *'lønn_PL3'*, *'persrisiko_1'*, *'norm_personlig_1'*, and *'interesse_5'*.

The k-folds cross-validation has accuracy levels ranging from 90% to 100% across the five distinct folds, culminating in a mean test accuracy of 95.15% (Figure 12 in Appendix). While the mean feature importances bear some resemblance to the original decision tree's feature importances, notable differences exist. The mean feature importance displaying the highest value is *'persgevinst'* at 0.41, trailed by *'intensjon_1'* at 0.21, and *'persrisiko_2'* at 0.16. The remaining mean feature importances have values below 0.1 and include *'perskontroll_1'*, *'pirat_venner'*, *'persrisiko_1'*, and *'interesse_3'*. Although *'intensjon_1'* was the most vital feature in the original decision tree, it ranks second in importance with k-folds cross-

validation. Meanwhile, *'persgevinst',* which initially was the second most significant feature in the original decision tree, assumes the position of the most important variable with k-folds cross-validation. Apart from these two variables, only *'persrisiko1'* and *'pirat_venner'* are involved as essential variables in both the original decision tree and with k-folds cross-validation.

6.5 General Discussion

The logistic regression model accounted for 72.45% of the variance in the dependent variable, suggesting a reasonably effective model in capturing the relationships between the variables. Significant relationships were found between piracy streaming and future intentions to piracy stream, intentions of never engaging in piracy streaming, interest in foreign football leagues, and the influence of friends and acquaintances engaging in piracy streaming. This information underscores the relevance of intentions and social influences in understanding piracy streaming behavior. The logistic regression model demonstrated an accuracy of 89% in predicting piracy streaming behavior. The model's high precision, recall, and F1-score across both classes indicate its effectiveness in minimizing misclassification, further supporting the findings.

The Naïve Bayes model achieved an impressive 98% accuracy in predicting piracy streaming behavior, with consistent performance across various data subsets. This consistency suggests that the model is generalizing well and offers valuable insights into factors contributing to piracy streaming. The decision tree classifier had an accuracy of 85.85% and highlighted the importance of variables such as intentions to piracy stream, perceived benefits, and social influence in predicting piracy streaming behavior. The k-folds cross-validation analysis further confirmed the importance of these variables and the overall consistency of the decision tree classifier.

The results are consistent with the theories and models discussed in the literature. The Synamedia report and the LEADERS report both emphasize the importance of addressing the needs of different consumer groups and implementing targeted solutions to reduce sports piracy. Our results show that significant relationships exist between piracy streaming and future intentions to engage in piracy streaming, interests in foreign football leagues, and the influence of friends and acquaintances who engage in piracy streaming. These findings

highlight the importance of understanding consumer behavior and social influences when designing targeted anti-piracy measures.

The Theory of Planned Behavior (TPB) posits that perceived behavioral intention and perceived behavioral control directly predict behavior achievement. Our logistic regression model supports this assertion, as it accounts for 72.45% of the variance in the dependent variable, piracy streaming behavior. The logistic regression model further corroborates TPB, demonstrating an accuracy of 89% in predicting piracy streaming behavior.

Benjamin Tan's Issue-Risk Judgment model also finds support in our results. The models suggest that factors such as social consensus, the magnitude of consequences, and cognitive judgment significantly influence consumers' ethical decision-making. In our survey, variables such as intentions to piracy stream, perceived benefits, and social influence were critical in predicting piracy streaming behavior. This highlights the importance of addressing consumers' moral judgment and ethical-decision making when combating sports piracy.

7. CONCLUSION

This thesis aimed to understand the factors influencing piracy streaming, focusing on sports events, and to develop accurate predictive models for identifying such behavior. By utilizing various statistical models, we have found that intentions to engage in piracy streaming, interest in foreign football leagues, and social influence significantly predict piracy streaming behavior. These findings are in line with the theoretical frameworks discussed, such as the Theory of Planned Behavior and the Issue-Risk-Judgment model.

The thesis also demonstrates the effectiveness of the logistic regression and Naïve Bayes models in predicting piracy streaming behavior with high accuracy. These models can provide valuable insights for decision-makers and content owners in developing targeted anti-piracy measures and understanding the dynamics of piracy streaming.

7.1 Future Research

Several avenues for future research are not covered in this thesis. A larger and more diverse sample could provide better insights into the nuances of piracy streaming behavior across different populations. Unmeasured variables and other external factors that are not covered in the survey could also significantly affect piracy streaming, such as internet speed, access to legal streaming options, and anti-piracy laws. Furthermore, longitudinal studies could examine the changes in piracy streaming behavior over time, providing a better understanding of how piracy streaming evolves due to technological advancements, legislation changes, and shifts in consumer preferences. Lastly, it would be interesting to develop intervention strategies based on the findings of this thesis and evaluate whether these strategies could be successful to reduce piracy streaming.

8. REFERENCES

Ajzen, I. (1991) *The Theory of Planned Behavior.* Academic Press: University of Massachusetts at Amherst.

Ampere Analysis. (2020) *Charting global sports piracy*. Synamedia. Available at: https://www.synamedia.com/whitepapers-reports/charting-global-sports-piracy/ (17.04.23)

Ampere Analysis. (2021) *The super aggregators: Sports IP pirates have gone OTT*. Synamedia/LEADERS. Available at: https://leadersinsport.com/sport-business/reports/leaders-special-report-the-super-aggregators-sports-ip-pirates-have-gone-ott/ (17.04.23)

Bushnell, H. (2019). *Inside the complex world of illegal sports streaming*. Available at: https://sports.yahoo.com/inside-the-complex-world-of-illegal-sports-streaming-040816430.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuYmluZy5jb20v&guce_referrer_sig=AQAAAI6zf_Jw4DR6-IECUe-ClbV2CxY1vt75EJE6-e1HVI0-0N3bopUEA99mEWGABTgM3ToiScIjlQ4iRw_CRXK2gCw_q7uts8r8fZ54w1c2ogBWlSE7TzJI5cfQuZSCbCuT-X-SEA4j-D2MQ7TEhtlz95pghNkKM7_8Q8sFH7k0E5P7 (17.04.23)

Friedman, J. and Hastie, T. and Tibshirani, R. (2008) *The Elements of Statistical Learning.* 2nd edition. Stanford, California: Springer. Available at: https://hastie.su.domains/Papers/ESLII.pdf (17.04.23)

Jerijervi, D. R. (2023) *Viaplay-underskudd på nesten 300 millioner – taper store penger på global utrulling*. Available at: https://kampanje.com/medier/2023/02/viaplay-legger-til-890-000-abonnenter/ (17.04.23)

Karabiber, F. (Unknown) *Gini Impurity*. Available at: https://www.learndatasci.com/glossary/gini-impurity/ (17.04.23)

Shafrin, J. (2016) *What is Pseudo R-squared?* Available at: https://www.healthcare-economist.com/2016/12/28/what-is-a-pseudo-r-squared/ (17.04.23)

Statistics Solutions. (No date) *Assumptions of Logistic Regression*. Available at: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/) (17.04.23)

Tan, B. (2002) *Understanding consumer ethical decision making with respect to purchase of pirated software*. Nanyang Technological University, Nanyang Business School.

Wikipedia (2023) *Confusion matrix*. Available at: https://en.wikipedia.org/wiki/Confusion_matrix (17.04.23)

Wikipedia (2023) *Pseudo-R*-squared. Available at: https://en.wikipedia.org/wiki/Pseudo-R-squared (17.04.23

Wong, D. (2015) *The EPL drama – paving the way for more illegal streaming? Digital piracy of live sports broadcasts in Singapore*. Leisure studies. Centre for Business in Society, Coventry University.

9. APPENDIX

9.1 Dataset

Table 3: Included variables in dataset

Variable	Description
interesse_1	How interested are you in sports?
interesse_2	How interested are you in Norwegian women's club football?
interesse_3	How interested are you in Norwegian men's club football?
interesse_4	How interested are you in foreign football leagues?
interesse_5	How interested are you in Norwegian men's national team?
interesse_6	How interested are you in Norwegian women's national team?
favorittlag_utl	Do you have a favorite team in a foreign league?
tv_kamper_utl_sett	What proportion of these matches did you see?
favorittlag_no	Do you have a favorite team in the Norwegian Eliteserien?
pirat_venner	Do friends/acquaintances of you engage in piracy streaming football matches?
pirat_selv	Have you piracy streamed football matches?
holdning_piracy_1	Do you think piracy streaming football matches is a bad or good idea?
holdning_piracy_2	Do you think piracy streaming football matches is dumb or smart?
holdning_piracy_3	Do you think piracy streaming football matches is not beneficial or beneficial?
norm_sosial_1	Will people close to you dislike it if you piracy stream football matches?
norm_sosial_2	Will people close to you look down on you if you piracy stream football matches?
norm_sosial_3	Does any person close to you think it is okay to piracy stream football matches?
norm_personlig_1	Will you feel guilty if you piracy stream football matches?
norm_personlig_2	Does piracy streaming go against your principles?
norm_personlig_3	Do you experience piracy streaming football matches as illegal?
perskontroll_1	Do you have knowledge of how to piracy stream football matches?
perskontroll_2	Can you easily find football matches to piracy stream?
piratstrøm_antall	How often have you piracy streamed football matches during the last year?
persgevinst	Do you piracy stream football matches to save money?
persrisiko_1	Would you consider abstaining from piracy streaming if the risk of getting penalized is too big?

persrisiko_2	Would an increased risk for viruses/hacking make you abstain from piracy streaming football matches?
intensjon_1	Do you plan to piracy stream football matches in the near future?
intensjon_2	Would you piracy stream football matches if you get the opportunity?
intensjon_3	Would you never piracy stream football matches?
abonnement_PL1	Is 649 NOK monthly to watch Premier League unfair or fair?
abonnement_PL2	Is 649 NOK monthly to watch Premier League unreasonable or reasonable?
abonnement_PL3	Is 649 NOK monthly to watch Premier League wrong or right?
abonnerer_PL_kanal	Do you subscribe to TV channels that broadcasts matches from Premier League?
wtp_PL	What would you maximum be willing to pay monthly to watch matches from Premier League?
lønn_PL1	Average salary for players in Premier League is 37 million NOK, do you think that is unfair or fair?
lønn_PL2	Average salary for players in Premier League is 37 million NOK, do you think that is unreasonable or reasonable?
lønn_PL3	Average salary for players in Premier League is 37 million NOK, do you think that is wrong or right?
rettferdiggjør_1	Does the high price of watching TV matches justify piracy streaming?
rettferdiggjør_2	Do the extremely high salaries for football players justify piracy streaming?
kjønn	What is your gender?
inntekt	What was your gross taxable income in 2021?

9.2 Reports, Scores, and Matrixes

Table 4: Classification Report Logistic Regression Model

	precision	recall	f1-score	support
0	0.852941	0.935484	0.892308	31
1	0.9375	0.857143	0.895522	35
accuracy	0.893939	0.893939	0.893939	0.893939
macro avg	0.895221	0.896313	0.893915	66
weighted avg	0.897783	0.893939	0.894012	66

Figure 8: Confusion Matrix Naïve Bayes

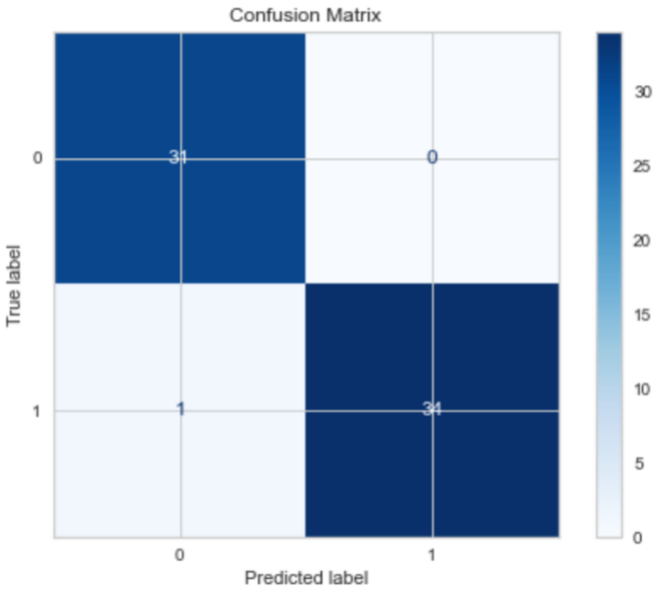


Figure 9: Cross-Validation Accuracy Scores

```
[0.98484848 1.         0.96969697 0.98484848 1.        ]
```

Figure 10: Cross-Validation Mean Accuracy and Standard Deviation

```
Mean accuracy: 98.79%
Standard deviation: 1.13%
```

Figure 11: Feature Importances Decision Tree Classifier

```
intensjon_1: 0.4875
persgevinst: 0.17951668584579983
pirat_venner: 0.1375388500388501
intensjon_2: 0.04408292324958985
lønn_PL3: 0.036826599326599326
wtp_PL: 0.027892561983471068
kjønn: 0.024007435465768823
persrisiko_1: 0.0174530111238972
norm_personlig_1: 0.017141108050198965
perskontroll_2: 0.01689331376831377
interesse_5: 0.011147511147511172
```

Figure 12: K-Folds Cross-Validation Accuracy

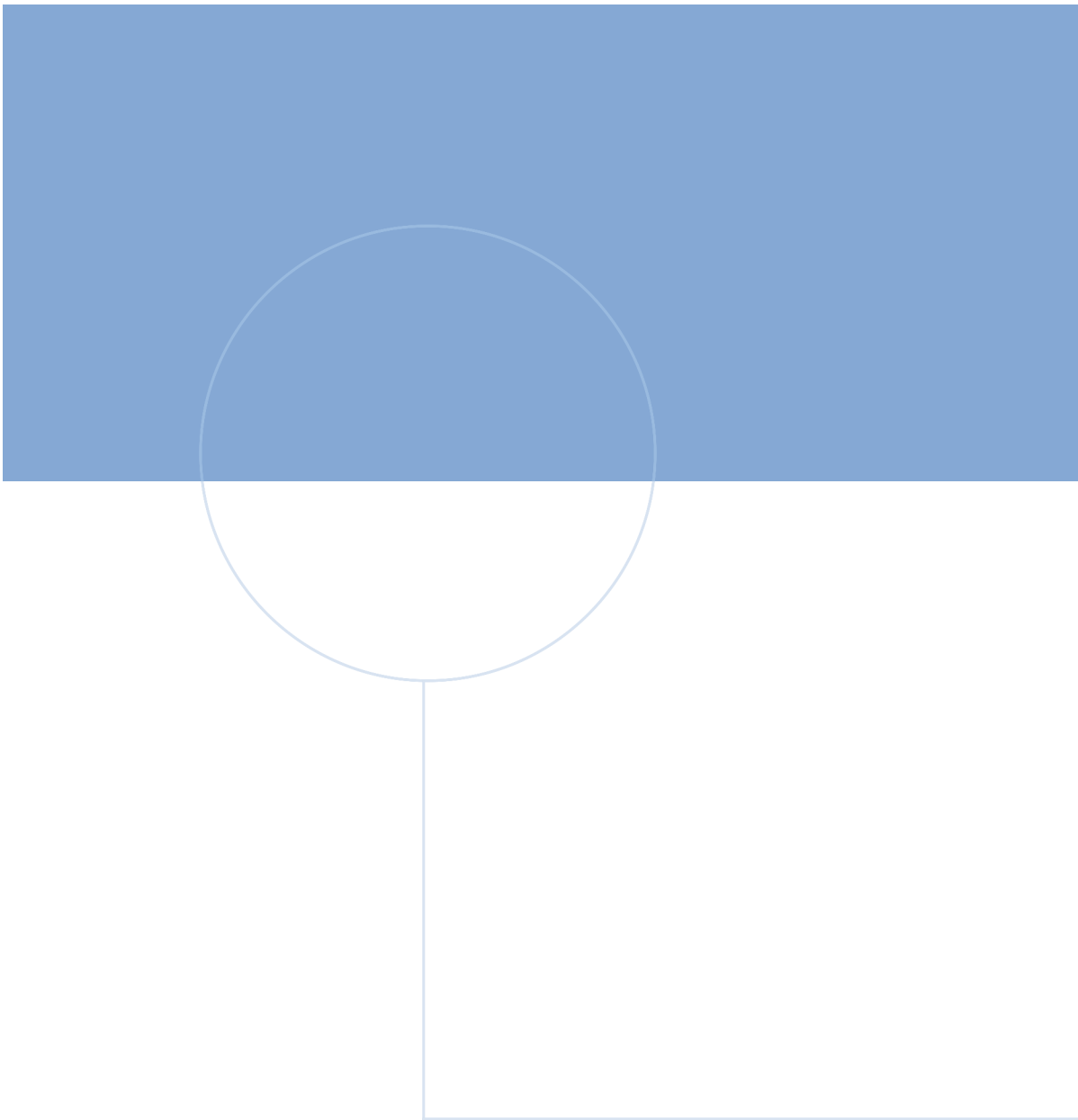
```
Test accuracies: [0.9393939393939394, 0.9242424242424242, 0.90909090909
09091, 0.9848484848484849, 1.0]
Mean test accuracy: 0.9515151515151515

Mean feature importances:
               Feature  Importance
21         persgevinst        0.41
24         intensjon_1        0.21
23        persrisiko_2        0.16
19      perskontroll_1        0.09
9         pirat_venner        0.05
22        persrisiko_1        0.05
2          interesse_3        0.01
```

9.3 Python Code

The code is provided in a zip-file. The dataset is not included in the zip-file because of privacy concerns. Reach out to Jon Martin Denstadli to get access to the data.

Zip-file is named “Bachelor_SLNE” and contains:

- “Bachelor.ipynb”. This file is a Jupyter Notebook file that contains the Python code used in this thesis.
- “Bachelor.py”. This file is a Python file that contains the Python code used in this thesis.
- “Spørreskjema_studenter.docx”. This file is a Microsoft Word file that contains the survey used in this thesis.



NTNU

Norwegian University of
Science and Technology