Jashore University of Science and Technology

Jashore-7408, Bangladesh

Behavioral Cluster Analysis of Online Gambling and Its Association with Big Five Personality Profiles

Submitted By:

Sovo Hasan

Student ID: 200127

Registration Year: 2020-21

Supervisor

Dr. Mohammad Nowsin Amin Sheikh

Assistant Professor

Dept. of Computer Science and Engineering

Dept. of Computer Science and Engineering

Jashore University of Science and Technology

May, 2026

Jashore University of Science and Technology

Jashore-7408, Bangladesh

This thesis is approved as a partial fulfillment of the requirements for the degree.

Signature of the Student

Signature of the Supervisor

Signature of the Chairman

Dedication

This thesis is dedicated to my parents who have provided unrelaxing support, encouragement and sacrifices upon which my academic experience has been based. I dedicate also this work to my teachers and mentors and those who have been so helpful to me with wisdom and patience. Lastly, I would like to say that all those who are dealing with gambling-related harm and whose lives might one day be healthier, safer, and more responsible because of some evidence-based approaches are the ones to whom I would like to dedicate this research.

Acknowledgments

I would like to express my sincere appreciation to my supervisor, Dr. Mohammad Nowsin Amin Sheikh, Assistant Professor, Department of Computer Science and Engineering, Jashore University of Science and Technology, for his supervision, feedback and encouragement, contributing to the success of this research. I also would like to thank the academic support and feedback of the professors of Department of Computer Science and Engineering. I would also like to thank my friends who have been the source of motivation and knowledge for me at the time of work. Lastly, I acknowledge the existing research in the fields of behavioral data science, machine learning, psychology of personality and gambling research where I built this thesis, where I founded my research.

Table of Contents

Dedication i

Acknowledgments ii

List of Abbreviations ix

Abstract xi

Chapter 1: Introduction 2

1.1 Overview 2

1.2 Background 2

1.3 Problem Statement 4

1.4 Research Questions 5

1.5 Objectives 5

1.6 Contributions of the Study 6

1.7 Significance of the Study 6

1.8 Chapter Organization 7

Chapter 2: Literature Review 9

2.1 Overview 9

2.2 Online Gambling and Behavioral Analytics Studies 9

2.3 Personality and Broader Gambling Risk Studies 10

2.4 Research Gap and Comparison of Proposed Framework 11

Chapter 3: Methodology 14

3.1 Overview 14

3.2 Proposed Methodology 14
3.3 Dataset Description 15
3.4 Data Preprocessing 18
3.4.1 Bustabit Preprocessing 18
3.5 Feature Engineering 19
3.6 Clustering 20
3.6.1 Behavioral Clustering 21
3.6.2 Personality Clustering 22
3.7 Cross-Domain Centroid Mapping 23
3.8 Supervised Validation Models 23
3.9 Evaluation Metrics 24
3.10 Mathematical Formulation of Evaluation Measures 25

Chapter 4: Results **28**

4.1 Overview 28
4.2 Bustabit Behavioral Evaluation 29
4.3 Big Five Personality Evaluation 35
4.4 Cross-Domain Hungarian Mapping 39
4.5 Supervised Risk Classification Results 42
4.6 Generalization Analysis of Supervised Models 45
4.6.1 Training and Validation Performance 46
4.6.2 Validation and Test Performance 47
4.7 Discussion 49
4.8 Comparison with Existing Studies 49

Chapter 5: Conclusion **54**

5.1 Overview 54
5.2 Conclusion 54

5.3 Limitations 55
5.4 Future Work 55

Chapter 6: References **58**

List of Figures

3.1 Proposed methodology of the study. 15

3.2 Visual representation of the clustering framework used in the study. 21

4.1 Distribution of selected Bustabit gameplay variables including bet amount, profit, and non-negative cash-out values. 30

4.2 Elbow curve for Bustabit KMeans clustering showing the selected best value of $k = 5$. 32

4.3 PCA projection of the Bustabit five-cluster KMeans solution. 33

4.4 Elbow curve for Big Five KMeans clustering showing the selected best value of $k = 5$. 37

4.5 PCA projection of the Big Five five-cluster KMeans solution. 38

4.6 Pairwise distance heatmap for cross-domain alignment between Bustabit behavioral clusters and Big Five personality clusters. 40

4.7 Train accuracy VS validation accuracy across supervised models. 47

4.8 Train accuracy VS validation accuracy across supervised models. 48

List of Tables

2.1 Summary of Online Gambling and Behavioral Analytics Studies 10

2.2 Summary of Personality and Broader Gambling Risk Studies 11

2.3 Comparison of Proposed Framework with Existing Works 12

3.1 Number of samples used at different stages of analysis 16

3.2 Summary of the Datasets Used in This Study 17

4.1 Sample characteristics and analytical files used in the study 29

4.2 Descriptive summary of engineered Bustabit behavioral constructs 29

4.3 Bustabit KMeans diagnostic metrics across candidate numbers of clusters . . . 31

4.4 Comparison of alternative clustering methods on the Bustabit feature space . . 32

4.5 Behavioral interpretation of Bustabit clusters 34

4.6 Standardized Bustabit behavioral cluster centroids 34

4.7 Raw-scale Bustabit cluster profile summary (behavioral constructs) 35

4.8 Raw-scale Bustabit cluster profile summary (monetary and incentive variables) 35

4.9 First five scored Big Five trait profiles after preprocessing 36

4.10 Big Five KMeans diagnostic metrics across candidate numbers of clusters . . . 36

4.11 Standardized Big Five cluster centroids 37

4.12 Hungarian cluster-matching table 39

4.13 Pairwise distance matrix for Bustabit–Big Five alignment 39

4.14 Comparison of alternative cluster-mapping methods 40

4.15 Sensitivity of Hungarian mapping across distance metrics 41

4.16 Integrated behavioral and personality interpretation by cluster 41

4.17 Classification performance across supervised models 43

4.18 Classification report for the best-performing model 44

4.19 Confusion matrix for the Logistic Regression model 44
4.20 Training and validation accuracy across supervised models 46
4.21 Validation and test accuracy across supervised models 48
4.22 Comparison of the present study with selected existing studies 50

List of Abbreviations

No.	Abbreviation	Full Form / Meaning
1	JUST	Jashore University of Science and Technology
2	CSE	Computer Science and Engineering
3	AI	Artificial Intelligence
4	ML	Machine Learning
5	OCEAN	Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism
6	BFI	Big Five Inventory
7	IPIP	International Personality Item Pool
8	PGSI	Problem Gambling Severity Index
9	SOGS	South Oaks Gambling Screen
10	DSM	Diagnostic and Statistical Manual of Mental Disorders
11	ICD	International Classification of Diseases
12	COM-B	Capability, Opportunity, Motivation, and Behaviour model
13	PCA	Principal Component Analysis
14	KMeans	K-means clustering algorithm

No.	Abbreviation	Full Form / Meaning
15	DBSCAN	Density-Based Spatial Clustering of Applications with Noise
16	GMM	Gaussian Mixture Model
17	KNN	K-Nearest Neighbors
18	SVM	Support Vector Machine
19	RF	Random Forest
20	DT	Decision Tree
21	GB	Gradient Boosting
22	LR	Logistic Regression
23	XGBoost	Extreme Gradient Boosting
24	TP	True Positive
25	FP	False Positive
26	TN	True Negative
27	FN	False Negative

Abstract

The swift development of online gambling has posed new issues in determining risky behaviours of players and psychological factors that relate to gambling behaviour. In contrast to the traditional gambling environment, the online gambling environment is always able to record behaviour which means that behavioural patterns can be analysed using data analytics. This thesis explores how online gambling behaviours are aggregated in groups of players, and how the behaviours are associated with the big five personality traits. The paper examines big data of online gambling through machine learning methods to uncover latent behavioural groups, where the frequency of gambling, size of bets, degree of involvement, reward pursuit, losses and control over gambling behaviour are considered. Players were clustered into segments representing various gambling behaviours, including regulated and low-risk players, to very active, loss-oriented and impulsive gamblers, using clustering algorithms. There were five primary clusters of behaviour, including regulated low-risk players, highly engaged regular players, high-loss players and low-control unstable players. Simultaneously, the data about personality (according to the Big Five model (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism)) were analysed to create five personality clusters. The behavioural clusters were compared to personality clusters using matching based on similarities. The results indicated that there is very high structural similarity between the two. The group of individuals exhibiting erratic, risky and excessive gambling behaviours were usually linked to the high levels of neuroticism and low levels of conscientiousness and moderate and controlled gambling behaviours were more linked to the emotionally stable, conscientious and responsible individuals. Results suggest that the type of gambling behaviours are not by chance, but they can be an expression of more general psychological dispositions. The study also looks at prediction models to identify more high-risk gaming behaviours by use of simulated behavioural character-

istics. Machine learning models, such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), Extra Trees, Gradient Boosting (GB), Logistic Regression (LR), Adaptive Boosting (AdaBoost), and Decision Tree (DT) were applied to predict high-risk and low-risk gambling behavior. Among the tested models, Logistic Regression achieved the best overall performance, with a test accuracy of 99.04% and an F1 score of 98.81% This indicates the excellent predictive power of behavioural attributes and effectiveness of behavioural analytics in automated monitoring. This paper demonstrates the feasibility of reinterpreting online gambling behaviours into useful behavioural clues, and associating them with personality-based behavioural patterns in a structural manner. The proposed paradigm can help to improve the emergence of gambling technologies to support responsible gambling, early detection systems and future interdisciplinary studies including machine learning, psychology and online behavioural analytics.

Keywords: Online gambling, Bustabit crash game, Behavioral profiling, Big Five (OCEAN), K-means clustering, Hungarian assignment algorithm, Risk classification, Machine learning.

Chapter 1

Introduction

Chapter 1: Introduction

1.1 Overview

The rise of online technology has led to a growth in online gambling and in turn, new forms of gambling behavior that are rapid, recurrent and interactive. Of particular interest is online crash gambling because it's characterised by continuous betting, unpredictable outcomes, and potentially high-risk behaviours. This gambling behavior may also be linked to individual personality characteristics, such as those represented in the Big Five personality model.

This thesis investigates online crash gambling behavior and its potential association with the Big Five personality model. This chapter provides background for the study, outlines the research problem, research aim and questions, the importance of the study, and thesis structure.

1.2 Background

Gambling has a long history as a social and economic activity, but the digitalisation of gambling changed the scale and dimension of gambling risk. Land-based gambling is limited by the operating hours, travel, promotion and architecture of gambling houses. The online environment might change the activity into continuous, private and algorithmically supported. This is important because gambling harm is not just about the loss of money but can also be emotional distress, social isolation, poor work or academic performance, family disintegration and constant attempts to recover losses [1–3]. That gambling disorder has been added to major diagnostic systems is another sign that problematic gambling is now recognised as a serious behavioural addiction and not simply a leisure preference [4, 5].

The importance of online gambling is that the digital platform keeps a record of one's gambling. The stake, cash-out, win, lose, session, bonus and reaction to previous wins and losses can be logged. Using these logs, it is possible to consider gambling in terms of decisions, not outcomes. A person who plays one big stake and then cashes out might not be similar to someone who continues to play small stakes, although the net profit or loss for both of them might be identical.

This is the crux of the behavioral analytics approach as risk can be expressed as fast play, volatile staking, regular exposure to losses or erratic cash-out [6–8].

This type of study is very applicable to crash gambling in which the game has a quick structure, is highly repetitive and the decision to cash-out is continuous. In a crash game, a multiplier is increased until it crashes, the player must cash-out before the crash to win money. The game design itself results in cues being produced related to risk, timing, volatility and loss. A player who always quits early could be different to a player who waits until a very high multiplier. Similarly, an escalation in gambling after a loss might be different to the risk process of a player who doesn't.

At the same time, there have been several studies that indicated the importance of individual differences in gambling. Big Five personality system describes the general personality profiles on openness, conscientiousness, extraversion, agreeableness and neuroticism [9–11]. It has been found in previous gambling studies and meta-analyses that neuroticism and low conscientiousness are typically associated with problematic gambling, and impulsivity and poor self-control are also important in determining the vulnerability of gambling [12–14]. The findings suggest that personality-profile structures may be conceptually related to the behavioral profiles, which are generated based on data collected from these platforms.

Machine learning provides the tools for exploring such associations. Unsupervised clustering can be applied to identify latent behavioural sub-groups that are not specified using a diagnostic class and supervised models can be used to test whether engineered behavioural features can distinguish between high and low risk profiles of a disease or condition [15–17]. This paper adopts a hybrid approach: it first creates interpretable features for players based on Bustabit gambling records, then clusters the behavioral and the personality data and finally matches the centres of clusters using the Hungarian assignment algorithm [18]. It is an exploratory, but not diagnostic objective. The objective of the study is not to claim that by examining a player's gambling history you can determine their true personality type; rather it seeks to investigate whether there is significant correlation between independently attained behavioural

and personality structures, in the form of clusters.

Although previous studies have examined online gambling behaviours and personality variables, there is a lack of research into their correlation. A wealth of research investigates them separately and it is unclear if online crash gambling behaviour patterns can be significantly related to higher level personality. This gap highlights the need for a user-friendly model to better explain risky gambling behaviour.

1.3 Problem Statement

Previous research has produced reliable research results in three interrelated areas. First, online, even mixed-mode, gamblers are more engaged and problematic than land-based-only gamblers [19, 20]. Second, risk of gambling is associated with personality traits such as neuroticism, conscientiousness and impulsivity [12, 13]. Third, account-based or behavioural information can be used to detect risky play, or predict self-reported problem gambling or help with responsible gambling interventions by researchers working in machine learning to detect it before it occurs using machine learning methods [21–23]. However, these areas of research are studied separately.

The link between observable gambling behavior and more comprehensive personality profiling is still missing. The questionnaire-based studies often directly assess the traits, but may not assess the high-frequency decisions. Platform-data studies are more behavioral studies that may not be psychometrically measured. The data that have been used in this study (Bustabit and Big Five) are independent, so one cannot make inferences about individuals. The challenge however is to build a framework that can compare the structure of the two worlds at the cluster level without overemphasizing the psychological meaning of the results.

The problem addressed in this thesis is a methodological as well as interpretative. The methodological question that the study raises is how the crash-game logs be translated into player-level constructs. Interpretatively, the question is whether or not the constructs can be clustered into

clusters that in an exploratory sense can be linked to independently constructed Big Five personality clusters. This is important because a responsible gambling system should have explainable and not just black-box signals. A system that can reveal why one type of behavioural pattern is more volatile than others might be more helpful in early warning and monitoring and development of future interventions.

1.4 Research Questions

The research questions of this study are:

- Is it possible to convert online crash gambling records to meaningful individual-level behavioral profiles?
- Which distinct behavioural clusters can be uncovered from Bustabit crash gaming logs?
- Is it possible to match Big Five personality clusters (independently constructed) with behavioral gambling clusters at an aggregate level?
- How well can engineered behavioral features classify higher-risk/ lower-risk gambling types in a supervised learning environment?

1.5 Objectives

The main purpose of the study is to develop a machine-learning model of the analysis of online crash gambling behavior, and explore its relationship with Big Five personality-profile structures.

The study's objectives are:

- To clean and aggregate Bustabit crash-game data to a player-level.

- To identify clusters of behavior using unsupervised clustering.
- To process and score an open Big Five data set into the OCEAN personality scales and form personality clusters.
- To match behavioral clusters with personality clusters with Hungarian algorithm.
- To evaluate the distinguishability of the behavioral feature space with the supervised models.

1.6 Contributions of the Study

1. Produced player-specific behavioural features from online gambling data.
2. Clustered different groups of gamblers. Examined the links between behavioural clusters and Big Five personality traits.
3. Assessed the behavioural features using machine learning techniques.
4. Provided insights for future responsible gambling strategies.
5. Highlighted the risk-related importance of behavioural analysis in identifying vulnerable gambling behaviours.

1.7 Significance of the Study

The present study produces several contributions to the field of gambling analytics, behavioural modeling and computational psychology. First, it demonstrates that crash-game logs could be used to offer structured gambling-behavioral representations (instead of using single win-loss events to offer such representations). Second, it offers an exploratory approach of the matching of the gambling behavior profiles to the personality-profile structure even in situations where the two datasets are not interrelated at the subject level. Third, it provides a machine-learning

perspective on risk which could potentially be used in the future in responsible gambling systems, early warning systems, and behavioral surveillance systems. Finally, the research paper is relevant to the broader field of data-driven behavioral analytics in that it presents the methods of combining the unsupervised clustering, centroid matching, and supervised learning in a single system.

1.8 Chapter Organization

Chapter 1: In Chapter 1, the research topic is introduced, the background on online crash gambling and Big Five personality is discussed, the research problem is identified, the aims, research questions, the significance of the study and the structure of the thesis is presented.

Chapter 2: Chapter 2 also presents the related literature. It describes the problem gambling, online gambling harms, adolescent gambling, social influences, personality correlates, machine learning techniques, clustering, supervised classification models and the research gap that motivates the proposed approach.

Chapter 3: Chapter 3 discusses the methodology. It covers the Bustabit and Big Five data, data cleaning, feature engineering plan, unsupervised clustering model, Hungarian mapping to the clusters, supervised validation models, model evaluation metrics and the ethical implications of the research design.

Chapter 4: Chapter 4 report the results It shows the sample characteristics, summary statistics of the user behavior, clustering diagnostic statistics, principal component analysis (PCA) statistics, cross-brand mapping statistics, and the supervised validation model statistics. Each table and figure will be accompanied by an interpretation to make sure the numbers will be linked with the goals of the study.

Chapter 5: Chapter 5 concludes the thesis with a summary of the findings, limitations and suggestions on future research.

Chapter 2

Literature Review

Chapter 2: Literature Review

2.1 Overview

This chapter presents literature review of related work to the current study. Earlier research has covered gambling harm, online gambling, personality traits, young adults and adolescents, and machine learning in gambling analytics [2, 3, 13, 21]. These papers demonstrate that online gambling provides valuable behavioural data, and psychological and survey-based research shows why some people may be more susceptible to gambling-related harms [6, 7, 12, 13]. But most prior studies have been focused on a single aspect. So, there is a need for an interpretable platform to jointly examine online gambling behavior and personality types [13, 22, 24].

2.2 Online Gambling and Behavioral Analytics Studies

Previous research has included studies on gambling-related harm and risk in online context. Blaszczynski and Nower noted that problem gamblers are not a monolithic group, while Braverman and Shaffer suggested markers of high-risk online gambling [2, 6]. Dragicevic et al. also examined online casino play for player protection [7]. In machine learning based gambling analytics, Philander compared data-mining procedures to identify high-risk online gamblers, Percy et al. used supervised machine learning to predict self-exclusion, and Hopfgartner et al. and Auer and Griffiths used account level data of online players to predict self-reported problem gambling [21–23, 25]. These studies demonstrate that gambling behavior data can be applied to risk detection, but most do not consider personality interpretation [21–23].

Table 2.1 lists key studies that employed gambling behavior logs, markers and machine learning approaches to detect online gambling risk. These findings are consistent with the notion that online gambling traces can show important patterns of behavior, but they primarily involve risk detection, prediction, or responsible gambling interventions but not interpretation based on personality.

Table 2.1: Summary of Online Gambling and Behavioral Analytics Studies

Authors	**Approach / Models**	**Datasets**	**Performance**	**Criticism**
Blaszczynski and Nower	Pathways model	Gambling literature	N/A	Not based on online behavioral data
Braverman and Shaffer	Behavioral marker analysis	Internet gambling records	Markers identified	No personality analysis
Dragicevic et al.	Behavioral risk analysis	Online casino data	Risk patterns observed	Limited psychological interpretation
Philander	Classification / regression; ANN	Sports betting gamblers	ANN most reliable	Poor generalization on new samples
Percy et al.	LR, BN, NN, RF	845 online gamblers	RF best performer	Harder to identify self-excluders
Hopfgartner et al.	Logistic + 5 ML models	1,743 casino gamblers	ROC-AUC: 0.654–0.717	Country-wise variation
Auer and Griffiths	LR, RF, GB, DT, SVM	1,611 online gamblers	LR AUC 0.789; RF 0.776	Moderate class imbalance

2.3 Personality and Broader Gambling Risk Studies

The second group of studies looked at personality and general vulnerability. Dudfield et al. reported in a meta-analysis that the best predictors of problem gambling are higher neuroticism and lower conscientiousness [13]. Similarly, MacLaren et al. showed similar findings with university students [12]. Low emotional stability and external locus of control were also significant predictors of problem gambling in a large Australian sample by von der Heiden and Egloff [14]. Further, Slutske et al. examined the prospective associations between personality and gambling, and Canale et al and Potenza et al found that internet or online gambling among adolescents is associated with more severe gambling [26–28]. Botella-Guijarro et al. also applied the COM-B model to adolescent gambling behaviour [29]. These studies confirm the role of personality and risk context among adolescents, but are mostly questionnaire- or survey-based, and do not use

detailed platform data of gambling behavior [12, 13, 27, 29].

Table 2.2: Summary of Personality and Broader Gambling Risk Studies

Authors	**Approach / Models**	**Datasets**	**Performance**	**Criticism**
Dudfield et al.	Meta-analysis	20 samples, 32,222 participants	$r = .31$ (N), $r = -.28$ (C)	Not based on platform data
MacLaren et al.	Big Five correlation study	University students	High N, low C found	Questionnaire-based only
von der Heiden and Egloff	Big Five + locus of control	$>$12,500 adults	Low stability predictive	Not online-only
Slutske et al.	Prospective cohort study	Young adult birth cohort	Risk traits identified	Not online gambling specific
Canale et al.	Survey analysis	14,778 adolescents	21.9% vs 4.0% problem gambling	Youth-focused only
Potenza et al.	Chi-square + logistic regression	2,006 student gamblers	Higher at-risk/problem rate	Adolescent sample only
Botella-Guijarro et al.	COM-B + PLS-PM	354 adolescents, 3 waves	Risk factors predicted behavior	Not platform telemetry

Table 2.2 highlights studies on personality, adolescent gambling risk and general vulnerability. These studies suggest that personality traits like neuroticism and conscientiousness and adolescent gambling risk and vulnerability are associated with gambling problems. But most are questionnaire- or survey-based studies, and do not directly link these psychological factors to online gambling logs.

2.4 Research Gap and Comparison of Proposed Framework

In general, the literature reveals that gambling harms, online behavioural tracking, personality vulnerabilities and machine learning are all relevant aspects of risky gambling behaviour

[2, 6, 13, 21]. But a gap remains as most research studies are limited to either online gambling behavior data or psychological traits [12, 22, 24]. Few attempts are made to link online gambling behavior to Big Five personality profiles within an interpretable framework [13, 24]. Thus, the current study aims to fill this gap by creating behavioral features from online crash gambling, identifying groups of gamblers through clustering, and examining their structural association with the Big Five personality profiles.

Table 2.3: Comparison of Proposed Framework with Existing Works

Study	**Online data**	**Big Five**	**Crash focus**	**Clustering**	**Supervised**	**B–P mapping**	**RG view**
Braverman	✓	✗	✗	✗	✗	✗	✓
Dudfield	✗	✓	✗	✗	✗	✗	Indirect
Philander	✓	✗	✗	✗	✓	✗	✓
Hopfgartner	✓	✗	✗	✗	✓	✗	✓
Lannes	✓	✗	✗	✓	✗	✗	Indirect
MacLaren	✗	✓	✗	✗	✗	✓	Indirect
Auer	✓	✗	✗	✗	✓	✗	✓
Our Work	✓	✓	✓	✓	✓	✓	✓

Table 2.3 compares the present framework with some previous works. It highlights existing approaches that often focus on only one or two aspects, for example, online gambling risk categorization, association with psychological traits, or machine learning prediction. On the other hand, the current work integrates online gambling behavior, Big Five personality profiling, crash gambling, clustering, supervised learning, and comparison of gambling structure with Big Five personality profile in one framework.

Chapter 3

Methodology

Chapter 3: Methodology

3.1 Overview

This chapter outlines the approach taken to study online crash gambling behavior and its exploratory link to Big Five personality type. The research involves two data sets: Bustabit crash game and Big Five (OCEAN) personality profiles. The Bustabit gameplay data are converted into individual-level features that represent risk, chasing, engagement, volatility, control, and incentives. The features are then used to cluster players into groups using unsupervised learning. Big Five questionnaire data are independently clustered to form personality clusters. Lastly, the cluster spaces are compared via Hungarian matching of cluster centroids and supervised learning models are applied to assess whether the engineered features can classify high-risk and low-risk personality profiles of gamblers [6, 7, 13, 21].

3.2 Proposed Methodology

The focus of this study is to explore online gambling behaviour and its potential relationship with personality. To this end, two major sources of data were considered, Bustinat gameplay data and Big Five (OCEAN) personality traits. Bustabit gaming data were analysed to find features describing risk seeking, chasing losses, the degree of involvement and consistency of cash-out activities. Then, these features were clustered to discover common gambling behavior patterns using unsupervised machine learning, primarily K-means clustering, and compared with DBSCAN, Agglomerative Clustering and Gaussian Mixture Models (GMM). In order to explore the association between gambling behavior and personality, the Hungarian assignment algorithm was applied to optimally pair behavioral clusters with personality clusters, based on the Big Five personality profile. Finally, several supervised learning algorithms, such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), Extra Trees, Gradient Boosting (GB), Logistic Regression (LR), Adaptive Boosting (AdaBoost), and Decision Tree (DT) were applied to predict high-risk and low-risk gambling behavior from the

engineered features [16, 18, 30–41].

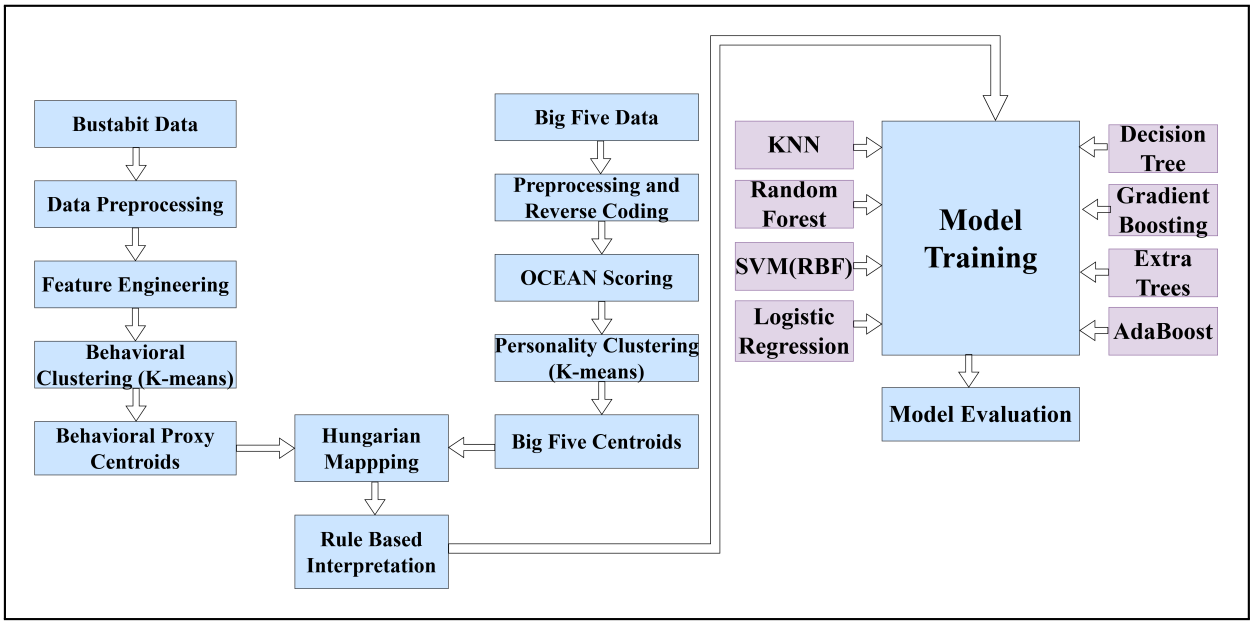

Figure 3.1: Proposed methodology of the study.

Figure 3.1 shows the analytical pipeline for the study. The left part of the figure illustrates how Bustabit crash-game log data are converted from round-level data into behavioral features, and into behavioral clusters. The right side illustrates how the Big Five questionnaire items are reverse scored to produce OCEAN personality dimensions, and then clustered as personality profiles. The middle of the figure represents the Hungarian matching stage, which involves the matching of behavioral centroids and personality centroids. The bottom of the workflow represents the supervised validation stage, in which various machine learning models are trained to assess whether the engineered feature space of behavior features can be used to discriminate among higher-risk and lower-risk gambling behavior.

3.3 Dataset Description

We used two publicly-available datasets. The first is the Bustabit crash-game dataset, comprising of round-level online gambling data. The submitted file includes 50,000 rounds of gameplay

observations comprised of 9 variables: player ID, bet size, profit, cash out (yes or no), bust multiplier and bonus information. As the intention of the research is to make inferences about consistent patterns rather than individual rounds, later the data was aggregated at the player level.

The second data set is the Big Five personality data set. When tab-delimited, the uploaded file has 19,719 questionnaire data rows and 57 columns. These are 50 personality items covering the Big Five (OCEAN) factors and 7 demographic items (race, age, gender, handedness, source, country and English native). These data were used to create clusters of personality profiles for later comparison (structurally) with the clusters of behavior.

Table 3.1: Number of samples used at different stages of analysis

Dataset	**Classification**		**Elbow Method**	**Clustering**	**Behavioural / Trait Analysis**	**Mapping**
	Train	**Test**				
Bustabit	40000(80%)	10000(20%)	30000	4150	4150	5 clusters
Big Five	–	–	10000	19719	19719	5 clusters

Table 3.1 shows the distribution of samples in the various stages of the study. The Bustabit data set was used for supervised classification and unsupervised clustering. For instance, 40,000 samples were used for training and 10,000 samples for testing while 30,000 samples were also used in the elbow method for choosing the number of clusters. Following data cleaning and aggregation, 4,150 player profiles were used for clustering and player behaviour analysis. The Big Five dataset was not used for supervised classification, but was used for personality trait analysis and clustering. A sample of 10,000 responses was used for the elbow method and the entire data set of 19,719 responses was used for clustering and trait analysis. Given the lack of a link between the two datasets at the individual level, the final mapping was done at the cluster level with five clusters from each dataset.

Table 3.2 lists the datasets and the derived structures used in this paper. The Bustabit data consisted of raw game-level gambling behaviour data, which was then aggregated to obtain player-level behavioural summaries for clustering and behavioural profiling. Meanwhile, the Big Five

Table 3.2: Summary of the Datasets Used in This Study

Data Type	**File**	**Features**
Original Gambling Behaviour Data	bustabit.csv	Game-level gambling data featuring player ID, bet, cashout, winnings, and other transaction features used to calculate player-level risk features.
Behavioural Summary Profiles	Derived from bustabit.csv	Player-level summary features like RiskIntensity, LossChasing, VolatilitySeeking, Control, Engagement and IncentiveSensitivity which were used for clustering and behavioural analysis.
Raw Personality Data	data.csv	Item-level Big Five personality responses and background information such as age, gender, race, education, handedness, source and country.
Personality Profiles at Trait-level	Derived from data.csv	Standardised Big Five trait scores (Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism) after reversing and pre-processing of the questionnaire items.
Cluster-level Mapping Data	Derived from both datasets	Five behavioural clusters (from Bustabit dataset) and five personality clusters (from Big Five dataset), used for Hungarian mapping (matching of clusters).
Temporal Classification Dataset	Derived from bustabit.csv	Leakage-free temporal samples from previous gambling behaviour to determine future risk, used for supervised classification - training and testing data.

dataset were item-level responses to a personality questionnaire, which were used to derive the standardized OCEAN personality traits. The two datasets were not merged at the person level, so a mapping analysis was used to compare the two on a cluster level. Further, the Bustabit data was also used to create a temporal classification dataset for leakage-free supervised risk prediction.

3.4 Data Preprocessing

This part explains the preprocessing done on the two datasets prior to clustering and classification. The two datasets were not processed together as they had a different structure and purpose.

3.4.1 Bustabit Preprocessing

The preprocessing of the Bustabit data set starts with the choice of the variables that are the most pertinent to the behavioral analysis. These are player identifier, bet amount, profit, cash-out value, bust multiplier and information relating to bonuses. All variables are transformed into a uniform numeric representation, and invalid or missing data are treated with care to ensure that missing data do not skew the subsequent feature engineering step.

Once cleaned, the round level records are player-aggregated. Summary statistics are then generated with respect to each player which include average bet, total bet, total profit, total losses, number of games, mean cash-out, variability of cash-out, variability of bust multiplier, and bonus use. This change is needed since the analysis is based on stable behavioral patterns, as opposed to individual gambling outcomes. It can also be used to construct interpretable behavioral constructs like engagement, control and loss chasing since player level aggregation is possible.

slug subsection: Big Five Preprocessing and Reverse Coding.

The processing of the Big Five data is on questionnaire-item level. To begin with, the responses

of all the items are transformed into numbers. The invalid responses and incomplete records are eliminated where it is necessary. Reverse coded items are then coded back again in such a way that larger scores will always mean larger levels of the desired trait. The reason behind this step is that personality questionnaires tend to have both positively and negatively phrased questions to minimize bias when responding to them.

Once the response is reverse coded, the item responses are averaged on a trait level to get openness, conscientiousness, extraversion, agreeableness and neuroticism scores. These trait scores are then standardized and thereafter clustered such that no one trait has the monopoly of the distance calculation due to differences in scale. It is a characteristic description, in line with the existing practice of scoring on the Big Five [42–44].

3.5 Feature Engineering

The summarized Bustedabit clean data are converted into six behavioral constructs that can be interpreted. These scales are meant to capture the intensity of gambling, persistence in losses, variability of behavior, self-regulation, level of activity and reward sensitivity. They are characterized as follows:

$$RiskIntensity_i = AvgBet_i \times Games_i \tag{3.1}$$

$$LossChasing_i = |TotalLosses_i| \tag{3.2}$$

$$VolatilitySeeking_i = std(BustMultiplier_i) \tag{3.3}$$

$$Control_i = \frac{1}{1 + std(CashedOut_i)} \tag{3.4}$$

$$Engagement_i = Games_i \tag{3.5}$$

$$IncentiveSensitivity_i = TotalBonus_i \tag{3.6}$$

These gambling features make the data more meaningful than round-level features. For instance, it is more informative to refer to repeated loss exposure as loss chasing rather than as a raw cumulative gambling loss value, and it is more informative to refer to stable cash-out as control rather than as a raw standard deviation value. This construct-based representation is also in line with past research on behavioral gambling analytics [6–8].

3.6 Clustering

In this study, clustering was applied to detect latent profile structures in the gambling data and Big Five personality data. Given that the study was exploratory, rather than diagnostic, unsupervised learning was preferred over pre-labeling. Clustering allowed us to cluster similar observations and check whether there was any structure in these two domains before we compared them. In both situations, K-means was the primary clustering algorithm used due to its relatively simple interpretation of centroid-based clusters [16, 30]. We also explored other methods like DBSCAN, Agglomerative Clustering and Gaussian Mixture Models to make sure the identified structure was still reasonable if different assumptions were made for clustering [31–33, 45].

For a set of standardized observations $x_1, x_2, \ldots, x_n$, K-means partitions the data into K clusters by minimizing the within-cluster sum of squared distances:

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \tag{3.7}$$

where C_k denotes the k-th cluster and μ_k denotes the centroid of that cluster. Each observation

is assigned to the nearest centroid based on Euclidean distance.

Figure 3.2 shows the clustering framework used in this study. The behavioral and personality datasets are clustered separately using K-means, evaluated using internal validation metrics, and then compared at the centroid level for later mapping and interpretation.

3.6.1 Behavioral Clustering

Clustering of the standardized Bustabit feature matrix was done post-player aggregation and feature engineering. This was an attempt to discover behavioral clusters based on gambling activity rather than to build a predictive model with labels of clinical interest. The behavioral clustering used behavioral constructs of risk intensity, loss chasing, volatility seeking, control, engagement and incentive sensitivity. These variables were chosen to represent various characteristics of gambling behavior and offer a meaningful portrayal of player behaviour.

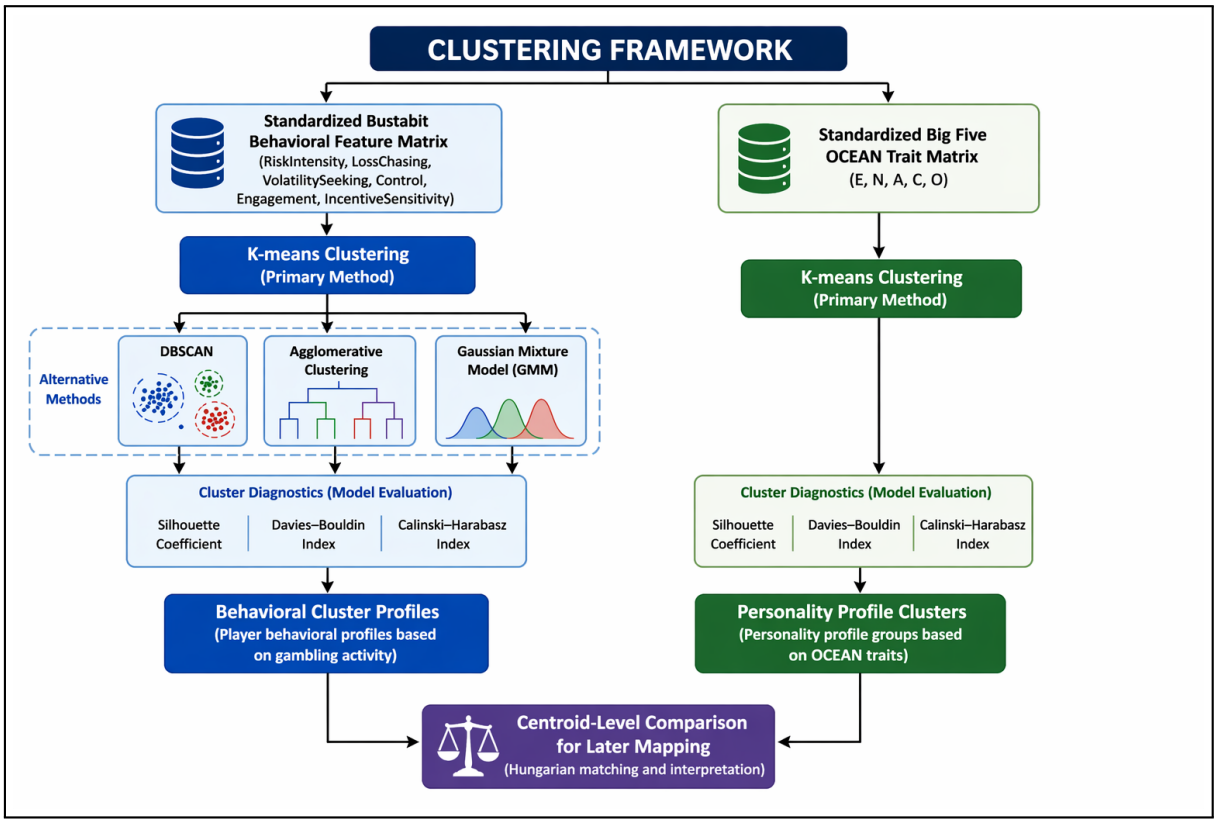

Figure 3.2: Visual representation of the clustering framework used in the study.

Let the behavioral feature vector for player i be defined as:

$$x_i = [RiskIntensity_i, LossChasing_i, VolatilitySeeking_i, Control_i, Engagement_i, IncentiveSensitivity_i] \tag{3.8}$$

These vectors for each player were scaled and then used in the clustering algorithms. The main method used for grouping the player profiles was K-means clustering. We tested several different values of k and chose the final cluster solution based on internal criteria and interpretability. To ensure the behavioral solution was not specific to K-means, additional clustering algorithms were applied, including DBSCAN, Agglomerative Clustering and Gaussian Mixture Models. This allowed us to ensure that the final grouping structure was not influenced by a particular clustering algorithm.

3.6.2 Personality Clustering

Personality clustering was conducted independently from the standardized OCEAN trait matrix of the Big Five data. The goal of this step was to define large-scale personality-profile groupings which could then be compared with clusters of gambling behavior. Given that personality traits are dimensional and overlapping, the identified clusters were considered as approximate personality-profile groups rather than discrete personality categories.

Let the personality vector for respondent j be defined as:

$$p_j = [E_j, N_j, A_j, C_j, O_j] \tag{3.9}$$

where E, N, A, C, and O represent Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness respectively. These standardized trait vectors were clustered using K-means in order to create personality-profile groups. A five-cluster solution was selected for the Big Five dataset in order to maintain conceptual comparability with the behavioral clustering stage. This

made the later cluster-matching process more interpretable. The resulting personality clusters represented different combinations of OCEAN traits, which then served as the psychological comparison space for the later centroid-mapping analysis.

3.7 Cross-Domain Centroid Mapping

The Bustabit and Big Five data are not paired, so they can't be compared on a row-by-row basis. Thus, the matching is done at the centroid level. First, the behavioural clusters are represented their centroids. Then the behavioral centroids are interpreted in a proxy OCEAN space using conceptual links. For instance, control is associated with conscientiousness, loss sensitivity is associated with neuroticism, engagement is associated with extraversion, and volatility seeking is associated with openness. Interpretation of agreeableness is tempered because it is not as well represented in gambling data.

Having computed the same types of centroid representations for both the gambling and personality clusters, the Hungarian assignment algorithm is applied to compute the optimal one-to-one mapping between clusters [18, 34]. This is suitable because it considers the total cost of assignments (across all clusters) as opposed to local optimisations. The mapping is purely exploratory and should only be interpreted as structural correspondence.

3.8 Supervised Validation Models

Following the unsupervised learning stage, we use supervised learning models to evaluate the discriminative power of the engineered behavioral feature space to discriminate between high-risk and low-risk behavioral profiles. This supervised stage does not replace the clustering process, it is an internal validation step.

We employ the following classifiers: K-Nearest Neighbors, Support Vector Machine, Logistic Regression, Decision Tree, Random Forest, Extra Trees, Gradient Boosting, AdaBoost and

XGBoost. These classifiers were chosen since they cover different families of machine learning algorithms: distance-based, linear, margin-based, tree-based, ensemble, and boosting [35–41, 54–61]. Comparing these models provides insight on whether the behavioral features are linearly separable or if nonlinear/ensemble methods are required.

3.9 Evaluation Metrics

The clustering solutions are evaluated using internal validation metrics. The K-means clustering objective is defined as:

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \tag{3.10}$$

where C_k denotes the k-th cluster and μ_k denotes its centroid.

The Silhouette coefficient for an observation i is calculated as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{3.11}$$

where $a(i)$ is the average distance between observation i and all other points in its own cluster, and $b(i)$ is the minimum average distance between observation i and points in another cluster.

The Davies–Bouldin index is given by:

$$DB = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right) \tag{3.12}$$

where S_i and S_j are within-cluster scatter values, and M_{ij} is the distance between cluster centroids i and j.

The Calinski–Harabasz index is calculated as:

$$CH = \frac{B_K/(K-1)}{W_K/(N-K)} \tag{3.13}$$

where B_K is the between-cluster dispersion matrix, W_K is the within-cluster dispersion matrix, K is the number of clusters, and N is the total number of observations.

Inertia is calculated as:

$$Inertia = \sum_{i=1}^{N} \|x_i - \mu_{c_i}\|^2 \tag{3.14}$$

where μ_{c_i} is the centroid of the cluster assigned to observation x_i.

For centroid mapping, the Euclidean distance between a behavioral centroid b_i and a personality centroid p_j is:

$$D_{ij} = \|b_i - p_j\|_2 \tag{3.15}$$

and the Hungarian assignment seeks the optimal mapping:

$$\min_{\pi} \sum_{i=1}^{K} D_{i,\pi(i)} \tag{3.16}$$

where $\pi(i)$ denotes the assigned personality cluster for behavioral cluster i.

The main metrics for supervised learning are accuracy, precision, recall and F1-score. Accuracy measures the fraction of objects classified correctly and F1-score is a composite measure of precision and recall and is particularly useful for imbalanced datasets [46–48, 64, 65, 67–72].

3.10 Mathematical Formulation of Evaluation Measures

The supervised evaluation measures are calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{3.17}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.18}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{3.19}$$

$$F1\text{-score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3.20}$$

Where:

- TP = True Positive
- FP = False Positive
- TN = True Negative
- FN = False Negative

These measures are reported because accuracy can be misleading in the case of unbalanced two-class problems. In these situations, the measures of precision, recall, and F1-score can be used to better understand classification accuracy. In this study, these measures are used to compare the performance of different supervised models for classification of high-risk versus low-risk gambling.

Chapter 4

Results

Chapter 4: Results

4.1 Overview

This chapter details the empirical results of the study in terms of the research goals. The findings are divided into six key steps. First, the Bustabit gambling data are assessed through descriptive summaries, diagnostics, interpretation, and visualisation of player behavioural patterns. Second, the Big Five personality data are evaluated through trait scoring, clustering, and centroid interpretation. Third, the structural relationship between the behavioural and personality spaces is examined using Hungarian mapping. Fourth, supervised machine learning is used to evaluate the discriminative power of the derived behavioural feature space for risk classification. Fifth, the supervised results are compared with existing research. Finally, the key findings are interpreted in terms of model behaviour, interpretability, and practical importance.

The study used two separate datasets with different original structures. The Bustabit raw dataset contained 50,000 round-level gambling records with 9 original variables. After preprocessing, aggregation, and feature engineering, these records were converted into 4,149 unique player-level profiles with 6 engineered behavioural features: *RiskIntensity*, *LossChasing*, *Volatility-Seeking*, *Control*, *Engagement*, and *IncentiveSensitivity*. These 6 features were used for clustering and behavioural interpretation. However, for supervised risk classification, *RiskIntensity* was removed from the input feature set because the risk label was derived from it. Therefore, the final supervised classification stage used 5 leakage-safe behavioural features.

The Big Five raw dataset contained 19,719 questionnaire records with 57 original variables, including personality items and demographic/background variables. After preprocessing, reverse coding, and trait-score calculation, the dataset was transformed into 19,718 scored personality profiles with 5 final standardized personality features. These features were *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*. Thus, while the original Big Five dataset contained 57 variables, the preprocessed analytical version used 5 interpretable trait-level features.

Table 4.1: Sample characteristics and analytical files used in the study

Dataset	Unit of analysis	N
Bustabit raw file	Round-level play records	50,000
Bustabit analytical file	Unique players	4,149
Big Five raw file	Questionnaire records	19,719
Big Five scored file	Scored personality profiles	19,718

Table 4.1 show that there are two distinct data sets. The Bustabit raw data file is larger in terms of events, but it is only 4,149 player profiles. This is crucial because a large number of rounds does not necessarily translate to a large number of players. The Big Five data contains 19,718 scored Big Five personality profiles, which offers a large space for comparison of personalities. The table also explains why the study is performed at the cluster level, rather than the individual.

4.2 Bustabit Behavioral Evaluation

In this section, we perform statistical, cluster diagnosis, interpretation and visualization on the Bustabit gambling data. This involves examining whether the behavioral feature space engineer can be used to cluster gambling behavior.

Table 4.2: Descriptive summary of engineered Bustabit behavioral constructs

Statistic	RiskIntensity	LossChasing	VolatilitySeeking	Control	Engagement	IncentiveSensitivity
count	4,149	4,149	4,149	4,149	4,149	4,149
mean	35,372.993	10,111.877	25.254	0.878	12.051	9.559
std	372,777.234	92,632.147	388.704	0.212	24.817	22.31
min	1	0	0	0.018	1	0
25%	100	0	0	0.832	1	0
50%	880	64	1.574	1	3	2.71
75%	5,426	1,000	6.457	1	10	8.4
max	16,262,350	3,886,342	19,140.036	1	291	319.16

Table 4.2 gives an overview of the Bustabit variables. This right skew of some gambling variables is represented by the median-maximum gaps. The majority of players are low or medium risk, but some players are very high risk intensity and loss chasing. This is a common feature of gambling telemetry, and suggests that standardisation is needed before clustering.

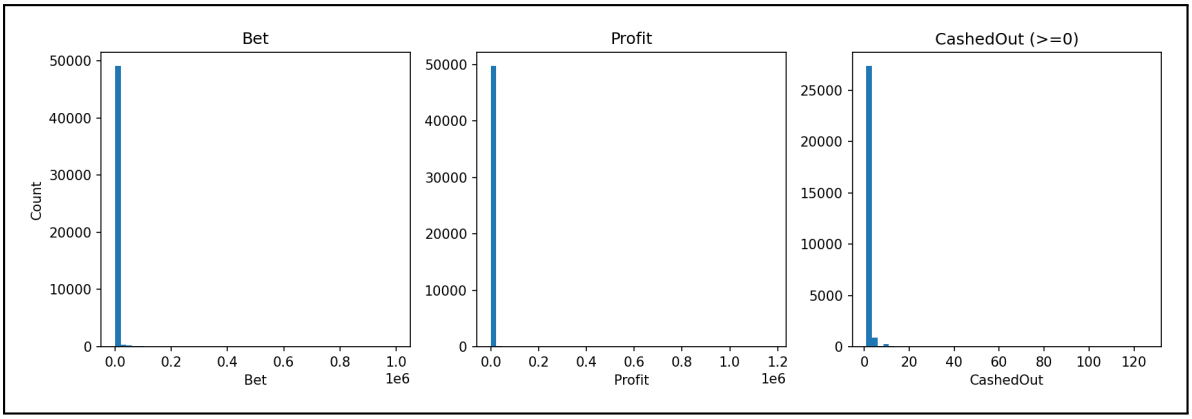

Figure 4.1: Distribution of selected Bustabit gameplay variables including bet amount, profit, and non-negative cash-out values.

Figure 4.1 shows visually the skewness in the descriptive statistics. Bets and profits are concentrated around the minimum value with a few extreme values extending the scale. The distribution of cash-outs is also negatively skewed with a long tail. Aggregation and conservative interpretation are justified for these reasons, as a few players can have a large influence on the distributions.

Table 4.3: Bustabit KMeans diagnostic metrics across candidate numbers of clusters

k	**Silhouette(↑)**	**Davies–Bouldin(↓)**	**Calinski–Harabasz(↑)**	**Inertia(↓)**
2	0.688	1.059	449.42	3,090.456
3	0.689	0.834	512.045	2,211.037
4	0.646	0.745	670.353	1,484.581
5	0.665	0.762	636.584	1,259.339
6	0.619	0.833	604.853	1,108.752
7	0.623	0.742	669.492	888.387
8	0.590	0.606	759.178	705.062
9	0.616	0.626	808.898	595.243
10	0.610	0.719	796.963	543.613

Table 4.3 presents internal K-means diagnostics for candidate values of k. Although $k = 2$ and $k = 3$ show slightly higher Silhouette values, the five-cluster solution provides a better balance between interpretability and structural comparison with the Big Five clustering. The selected $k = 5$ solution avoids oversimplifying the population while still preserving meaningful behavioral profiles.

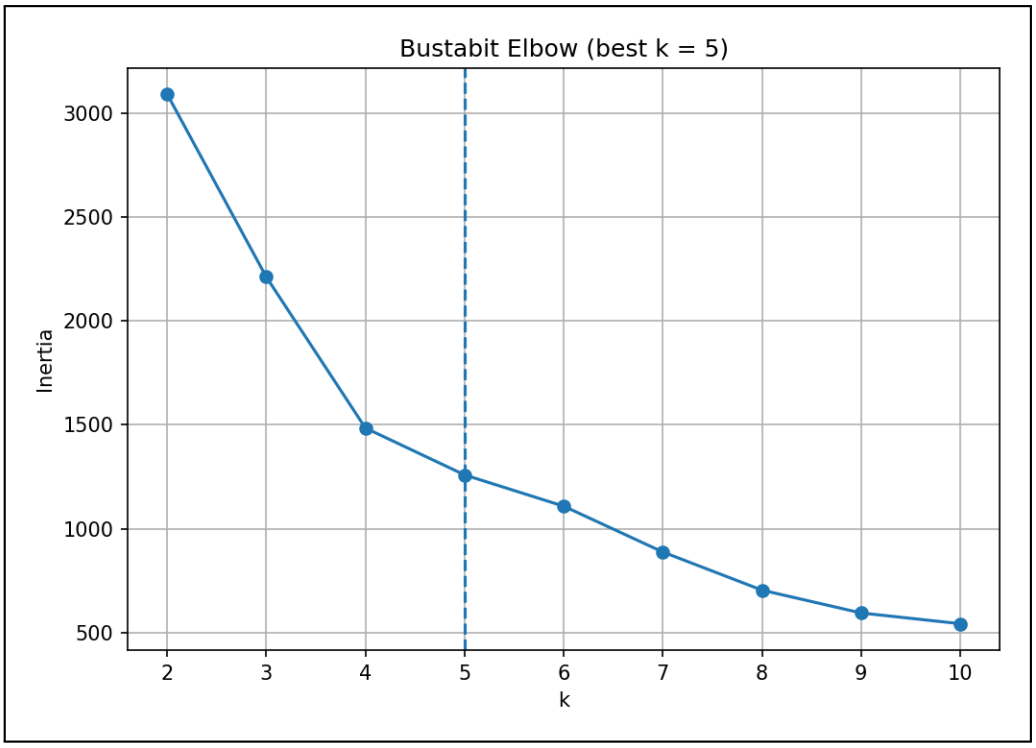

Figure 4.2: Elbow curve for Bustabit KMeans clustering showing the selected best value of $k = 5$.

Figure 4.2 illustrates the elbow pattern for Bustabit clustering. The curve drops sharply from $k = 2$ to $k = 4$ and then begins to flatten, which suggests diminishing returns from adding more clusters. The selected solution at $k = 5$ provides enough detail to distinguish low-engagement, high-engagement, high-stake, volatile, and low-control behavioral profiles without over-fragmenting the sample.

Table 4.4: Comparison of alternative clustering methods on the Bustabit feature space

Method	**Silhouette(↑)**	**Davies–Bouldin(↓)**	**Calinski–Harabasz(↑)**	**Noise%**
KMeans	0.659	0.619	1,192.357	0
Agglomerative	0.586	0.670	1,015.389	0
GMM	0.179	1.949	388.210	0
DBSCAN	0.865	0.176	182.859	2.82

Table 4.4 compares K-means with other clustering algorithms. DBSCAN has a high Silhouette value, as well as a low Davies–Bouldin score, but it also classifies some of the data points as noise, and has a lower Calinski-Harabasz value. K-means and agglomerative clustering yield more robust full-sample clusterings, while GMM does not perform as well in this feature space. This exercise supports the use of K-means as the preferred method as it provides a good balance between interpretability, coverage, and diagnostic performance.

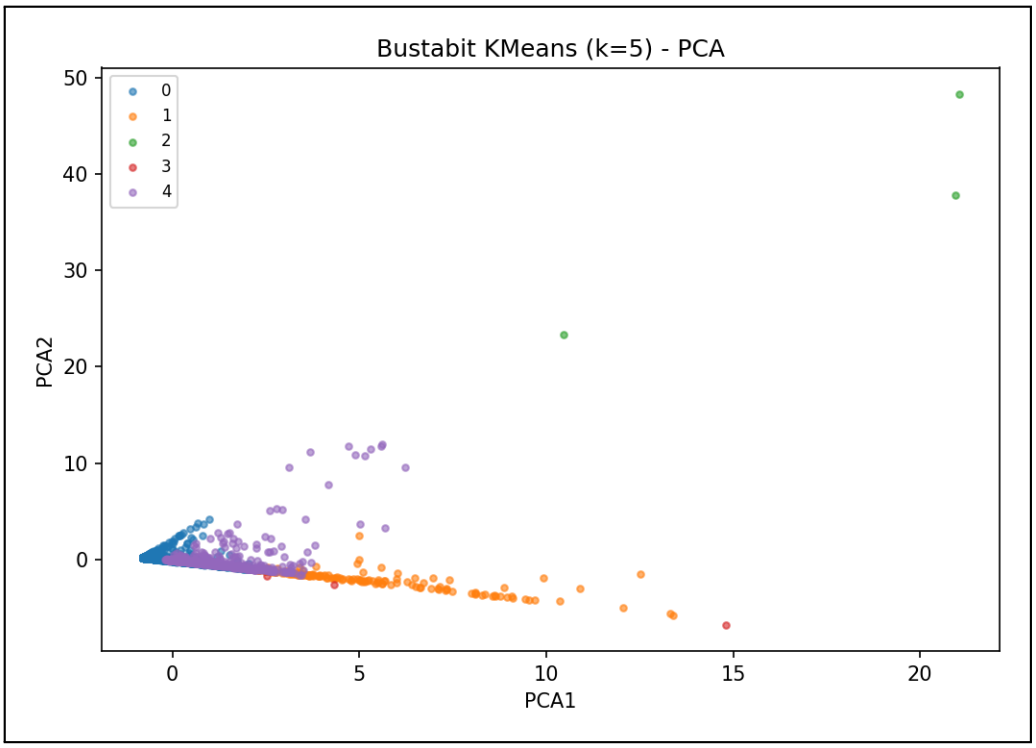

Figure 4.3: PCA projection of the Bustabit five-cluster KMeans solution.

Figure 4.3 shows the PCA pattern of the five-cluster Bustabit solution. The clusters are fairly distinct, but not completely isolated (as is common in behavioral data). There are some remote observations away from the central dense region, particularly on the first principal component, which shows that the model includes both typical and extreme high-intensity players.

Table 4.5: Behavioral interpretation of Bustabit clusters

Cluster	Cluster Name	Players	Risk Rate	Avg Bet	Total Profit	Loss Chasing	Control	Engagement
0	HighBet / LowEng / LowLoss	3,208	0.334	4,238.35	2,528.359	3,757.343	0.973	5.48
1	LowBet / HighEng / HighLoss	143	0.741	567.502	20,753.137	14,252.322	0.699	114.972
2	HighBet / HighEng / HighLoss	3	1.000	218,716.716	3,672,828	2,585,625	0.690	55.667
3	LowBet / HighEng / LowLoss	3	0.333	401.433	12,328.467	9,557	0.691	69.667
4	HighBet / LowEng / HighLoss	792	0.602	7,282.508	27,734.843	25,349.736	0.528	19.701

Table 4.5 summarises the five clusters of Bustabit players. The largest cluster (Cluster 0) is characterized by low engagement and high control. Cluster 1 has higher engagement and higher rate of high-risk behavior. Cluster 2 is very small but extreme with very high average stake, total profit and loss chasing. Cluster 4 is a dense cluster of high bets and high losses with less control. These findings suggest there is no single pattern of gambling risk; it may manifest as engagement, bet size, potential loss, or poor control.

Table 4.6: Standardized Bustabit behavioral cluster centroids

BustaCluster	Z_RiskIntensity	Z_LossChasing	Z_VolatilitySeeking	Z_Control	Z_Engagement	Z_IncentiveSensitivity
0	-0.064	-0.069	-0.042	0.447	-0.265	-0.254
1	0.066	0.045	0.050	-0.843	4.148	4.114
2	29.370	27.807	-0.009	-0.886	1.758	1.338
3	0.083	-0.006	33.283	-0.881	2.322	2.069
4	0.135	0.165	0.035	-1.652	0.308	0.275

Table 4.6 displays the standardized centroids and allows the comparison of the clusters with the entire population. Clusters 2 and 3 are, for example, very distinct in risk seeking and loss chasing, and Cluster 3 is very distinct in volatility seeking. Cluster 1 is distinctive in engagement and incentive sensitivity and Cluster 4 is distinctive in very low control. The standardized

centroids enable the separation of relative meaning from absolute values.

Table 4.7: Raw-scale Bustabit cluster profile summary (behavioral constructs)

Cluster	RiskIntensity	LossChasing	VolatilitySeeking	Control	Engagement
0	11,571.776	3,757.343	8.899	0.973	5.48
1	60,148.720	14,252.322	44.563	0.699	114.972
2	10,982,628	2,585,625	21.930	0.690	55.667
3	66,203.667	9,557	12,961.037	0.691	69.667
4	85,722.893	25,349.736	39.029	0.528	19.701

Table 4.8: Raw-scale Bustabit cluster profile summary (monetary and incentive variables)

Cluster	IncentiveSensitivity	AvgBet	TotalProfit	TotalLosses	Games
0	3.883	4,238.35	2,528.359	-3,757.343	5.48
1	101.331	567.502	20,753.137	-14,252.322	114.972
2	39.417	218,716.716	3,672,828	-2,585,625	55.667
3	55.723	401.433	12,328.467	-9,557	69.667
4	15.695	7,282.508	27,734.843	-25,349.736	19.701

Tables 4.7 and 4.8 present the raw-scale centroid values of the Bustabit clusters. Table 4.7 summarizes the main behavioral constructs, while Table 4.8 reports the monetary and incentive-related variables. Together, these tables show the actual magnitude of differences among clusters and make the behavioral interpretation more practical.

4.3 Big Five Personality Evaluation

stage assesses the Big Five using trait scoring, cluster diagnostics, centroid profiling and plotting. The aim of this step is to identify large-scale personality-profile clusters for comparison to the gambling clusters.

Table 4.9: First five scored Big Five trait profiles after preprocessing

E	N	A	C	O
4.4	1.1	4.6	4.7	4.3
2.2	3.1	3.5	4.2	2.6
3.5	4.6	3.8	4.9	4.5
2.2	4.3	3.7	2.6	4.1
3.4	3.0	4.4	3.4	3.4

Table 4.9 displays a subset of the scored Big Five traits. The rows are the five OCEAN traits after preprocessing and scoring. The table illustrates that the sample varies in profile shape, rather than being high or low on all traits, which will be useful for clustering later.

Table 4.10: Big Five KMeans diagnostic metrics across candidate numbers of clusters

k	Silhouette(↑)	Davies–Bouldin(↓)	Calinski–Harabasz(↑)	Inertia(↓)
2	0.202	1.763	425.971	5,865.720
3	0.176	1.837	341.115	5,175.194
4	0.167	1.681	318.430	4,597.746
5	0.162	1.618	289.967	4,242.369
6	0.157	1.545	272.535	3,940.011
7	0.155	1.510	254.924	3,721.303
8	0.152	1.567	239.289	3,549.159
9	0.154	1.507	229.216	3,378.536
10	0.154	1.496	217.804	3,253.452

Table 4.10 shows that Silhouette values are less in Big Five than in Bustabit clusters. This is unsurprising given that personality traits are not naturally discrete. We chose the five clusters as they offer meaningful patterns of profiles and are comparable to the five behavioral clusters.

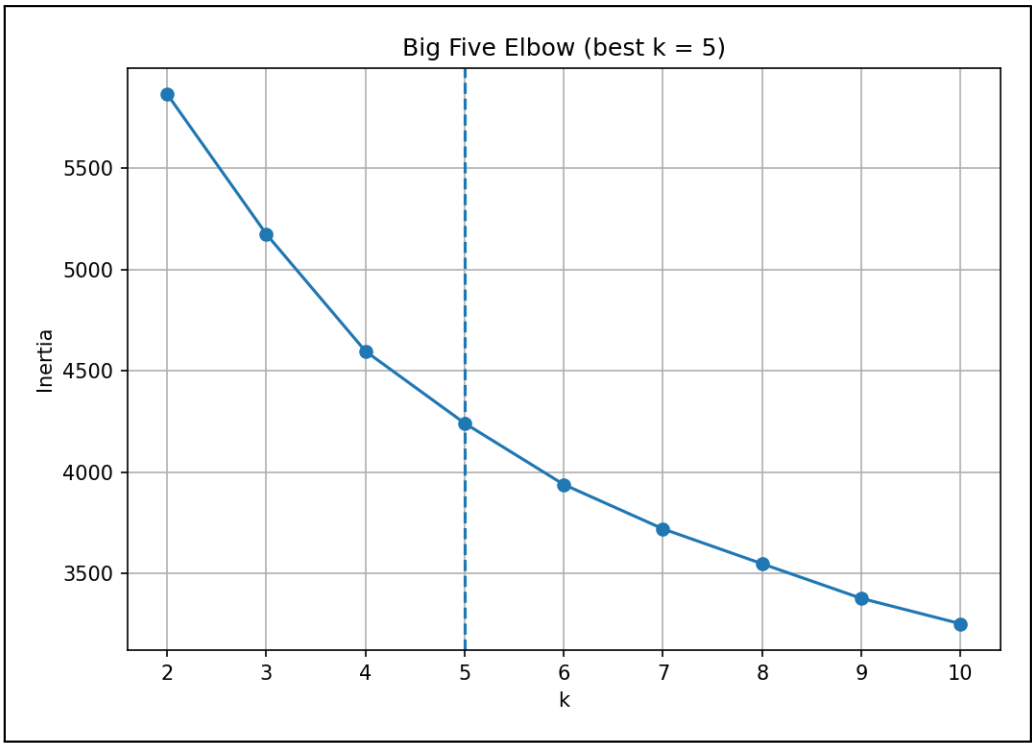

Figure 4.4: Elbow curve for Big Five KMeans clustering showing the selected best value of $k = 5$.

Figure 4.4 shows the Big Five elbow curve. The inertia curve shows a smooth decrease with the increasing k; the chosen $k = 5$ offers a reasonable compromise between simplicity and descriptiveness. The less clear elbow than seen with Bustabit is due to the fact that the data are continuous.

Table 4.11: Standardized Big Five cluster centroids

Big5Cluster	**Z_E**	**Z_N**	**Z_A**	**Z_C**	**Z_O**
0	-0.715	0.846	0.297	-0.031	0.456
1	0.604	-0.877	0.605	0.906	0.300
2	0.886	0.071	0.335	-0.765	0.380
3	-0.731	-0.062	-1.597	-0.180	0.333
4	-0.405	0.278	-0.249	-0.262	-1.337

Table 4.11 summarises the standardised centroids of the Big Five clusters. Cluster 1 is high

on conscientiousness and low on neuroticism, Cluster 2 is high on extraversion and low on conscientiousness, Cluster 0 is high on neuroticism and low on extraversion, Cluster 3 is low on agreeableness and has relatively high openness, and Cluster 4 has higher neuroticism and lower openness. These cluster centroid patterns serve as the psychological comparison space for mapping later.

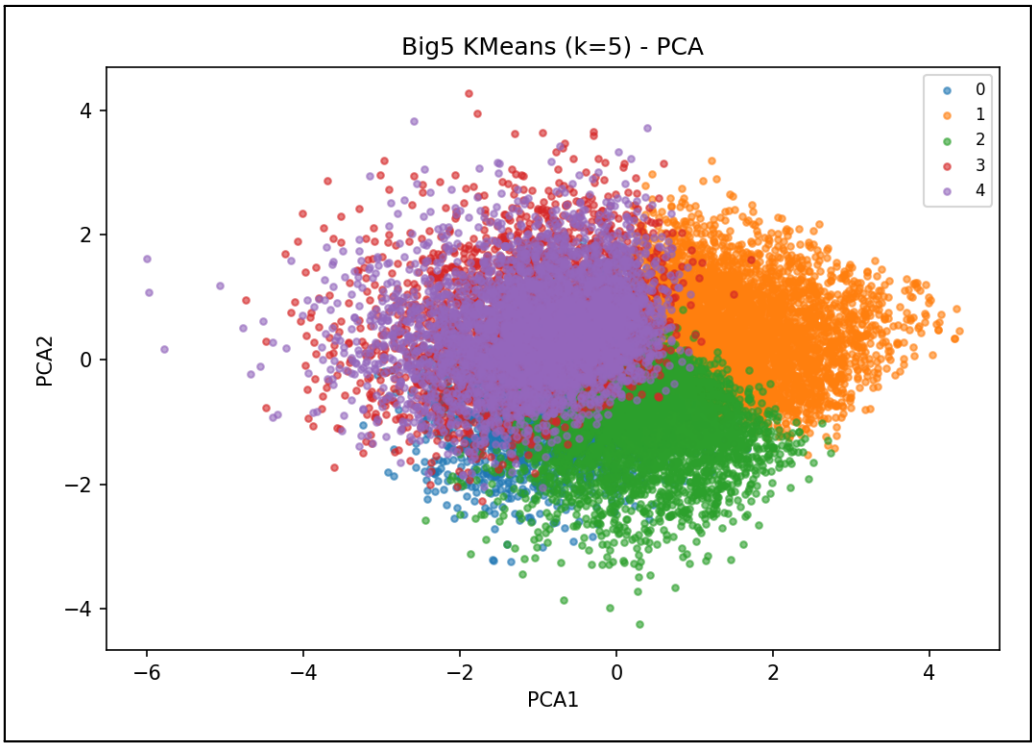

Figure 4.5: PCA projection of the Big Five five-cluster KMeans solution.

Figure 4.5 shows the PCA map of the Big Five five-clusters solution. The clusters are visually more overlapping than in the Bustabit space (as is typical of personality data). Still, the plot exhibits general concentration areas of distinct personality-profile patterns. This justifies the conservative interpretation of personality clusters as profile structure rather than classification.

4.4 Cross-Domain Hungarian Mapping

This section discusses the centroid-level comparison of the behavioural gambling clusters and Big Five personality clusters. The data sets are independent so that we compare at the centroid level.

Table 4.12: Hungarian cluster-matching table

BustaCluster	Big5Cluster	Big Five Profile
0	1	High C, Low N
1	2	High E, Low C
2	0	High N, Low E
3	3	High O, Low A
4	4	High N, Low O

Table 4.12 presents the one-to-one Hungarian pairing of Bustabit clusters and Big Five clusters. The matched profiles indicate that the more regulated behavioral clusters are structurally more similar to the high-conscientiousness/low-neuroticism profiles and the less stable clusters are structurally more similar to the higher-neuroticism or lower-control profiles. These findings are descriptive and exploratory.

Table 4.13: Pairwise distance matrix for Bustabit–Big Five alignment

Cluster	P0	P1	P2	P3	P4
B0	2.841	2.660	4.206	2.249	2.867
B1	3.384	2.480	1.644	2.749	2.732
B2	1.639	4.353	2.809	2.610	2.260
B3	2.744	2.838	2.159	2.107	4.201
B4	2.613	3.900	2.537	1.486	2.022

Table 4.13 shows the distance matrix used for the cluster alignment. Lower values mean closer

in the normalized proxy space. The Hungarian algorithm finds the set of pairings that optimises the total cost of all the clusters rather than local minima.

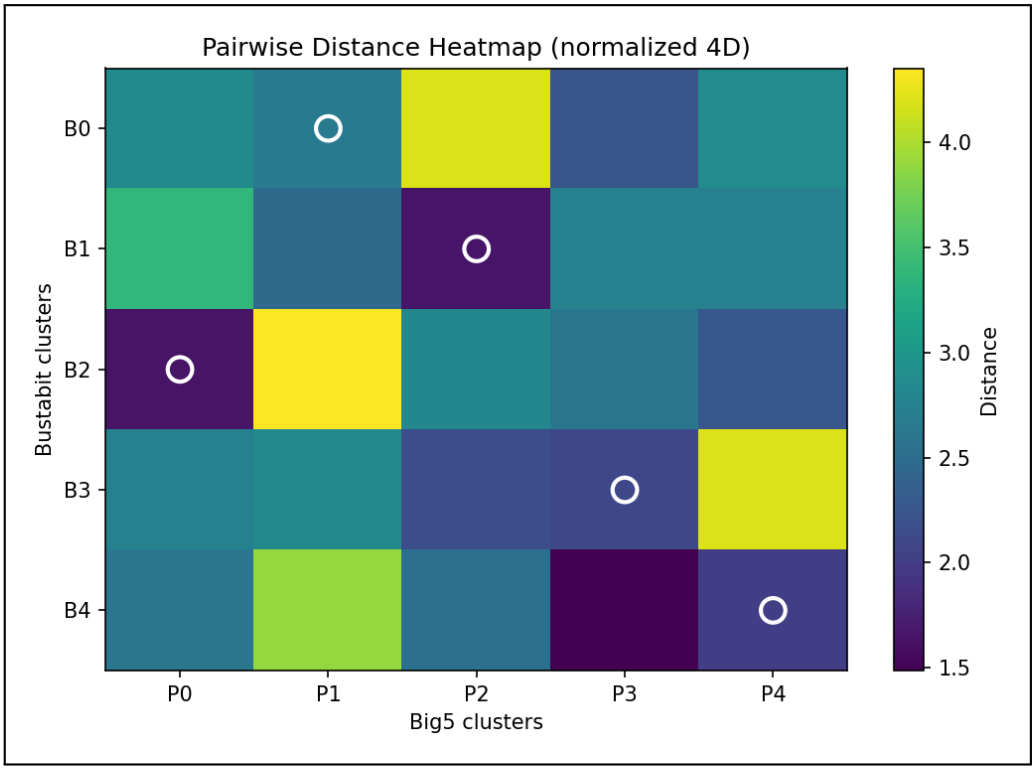

Figure 4.6: Pairwise distance heatmap for cross-domain alignment between Bustabit behavioral clusters and Big Five personality clusters.

Figure 4.6 displays the distance matrix. The heatmap is darker for smaller distances, and the selected cells are highlighted. The heatmap makes the structural correspondence more understandable because it enables the reader to visually compare the alternatives.

Table 4.14: Comparison of alternative cluster-mapping methods

Method	**Total Cost**	**Mapping (Busta→Big5)**
Hungarian (optimal 1-to-1)	10.072	{0: 1, 1: 2, 2: 0, 3: 3, 4: 4}
Greedy (local choices)	11.630	{4: 3, 2: 0, 1: 2, 0: 1, 3: 4}

Table 4.14 Hungarian matching versus greedy matching. The Hungarian algorithm has a smaller

cost than that of the greedy matching, thus the chosen one-to-one alignment is optimal for the current distance matrix.

Table 4.15: Sensitivity of Hungarian mapping across distance metrics

Distance	Total Cost	Mapping
Euclidean	10.072	{0: 1, 1: 2, 2: 0, 3: 3, 4: 4}
Cosine	2.574	{0: 1, 1: 2, 2: 0, 3: 3, 4: 4}

Table 4.15 shows mapping sensitivity under Euclidean and cosine distance. The mapping is the same under both metrics because it is not merely a function of one distance metric. While the costs differ due to different definitions, the structure is the same.

The integrated summary in Table 4.16 includes the number of subjects in each behavioral cluster, the rate of high risk, the main behavioral means, the labels used by the Hungarians, and the rule-based interpretations. It is the most comprehensive table in the chapter because it connects the statistical clustering to the interpretation. The table also reveals that high risk can be linked to various combinations of engagement, loss chasing, control and risk intensity.

Table 4.16: Integrated behavioral and personality interpretation by cluster

BustaCluster	Players	HighRiskRate	MeanRiskIntensity	MeanLossChasing	MeanControl	HungarianTrait	RuleTrait
0	3,208	0.334	11,571.776	3,757.343	0.973	High C, Low N (Big5 profile)	Self-controlled / regulated
1	143	0.741	60,148.720	14,252.322	0.699	High E, Low C (Big5 profile)	Self-controlled / regulated
2	3	1.000	10,982,628	2,585,625	0.690	High N, Low E (Big5 profile)	Impulsive / emotionally volatile
3	3	0.333	66,203.667	9,557	0.691	High O, Low A (Big5 profile)	Highly engaged / persistent
4	792	0.602	85,722.893	25,349.736	0.528	High N, Low O (Big5 profile)	Impulsive / emotionally volatile

4.5 Supervised Risk Classification Results

This section evaluates whether the engineered behavioural feature space can distinguish between low-risk and high-risk gambling profiles using supervised machine learning models. The aim of this experiment is not to claim clinical diagnosis or confirmed gambling disorder prediction. Instead, the purpose is to examine whether the behavioural indicators constructed from the Bustabit data contain enough structured information to support internally defined risk classification.

In the earlier version of the experiment, *RiskIntensity* was included as one of the supervised learning features. However, since the target variable *RiskLevel* was derived from *RiskIntensity*, including this feature in the model input could create direct target leakage. To make the supervised evaluation more reliable and defensible, *RiskIntensity* was removed from the classification feature set. Therefore, the final supervised models were trained using only the remaining behavioural indicators: *LossChasing*, *VolatilitySeeking*, *Control*, *Engagement*, and *IncentiveSensitivity*.

The classification task was performed on the player-level behavioural dataset, where each row represents an aggregated player profile. The dataset was divided into training and testing sets using an 80–20 split. Several supervised machine learning algorithms were evaluated, including Logistic Regression, K-Nearest Neighbours, Support Vector Machine, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, and Extra Trees. Table 4.17 shows the classification performance of these models.

Table 4.17: Classification performance across supervised models

Model	Test Accuracy	F1 Score
LogisticRegression	**0.9904**	**0.9881**
GradientBoosting	0.9807	0.9758
KNN	0.9795	0.9744
DecisionTree	0.9783	0.9730
SVM(RBF)	0.9771	0.9714
RandomForest	0.9759	0.9701
ExtraTrees	0.9651	0.9569
AdaBoost	0.9566	0.9443

Among the tested models, Logistic Regression achieved the best overall performance, with a test accuracy of 0.9904 and an F1 score of 0.9881. This is an important result because Logistic Regression is a relatively simple and interpretable linear model. Its strong performance suggests that, even after removing the direct leakage feature, the remaining behavioural indicators provide a clear separation between internally defined low-risk and high-risk players.

The strong performance of Logistic Regression also indicates that the decision boundary between the two risk groups may be largely explainable using a weighted combination of behavioural variables. In other words, features such as *LossChasing*, *VolatilitySeeking*, *Control*, *Engagement*, and *IncentiveSensitivity* appear to contain sufficient discriminatory information for classifying players according to the internally constructed risk labels. This supports the usefulness of the proposed behavioural feature engineering approach.

However, the high performance should be interpreted carefully. Although *RiskIntensity* was removed to prevent direct leakage, the target labels were still internally derived from behavioural patterns rather than external clinical or harm-based outcomes. Therefore, the model should be understood as performing behavioural risk classification, not clinically validated gambling-harm prediction. External validation using measures such as PGSI scores, self-exclusion

records, deposit-limit changes, or clinical screening results would be needed before making stronger real-world claims.

Table 4.18 presents the detailed classification report for the best-performing model, Logistic Regression.

Table 4.18: Classification report for the best-performing model

Class/Summary	Precision	Recall	F1-score	Support
Low-risk	0.99	0.98	0.99	498
High-risk	0.98	0.99	0.99	332
Accuracy	–	–	0.99	830
Macro avg	0.99	0.99	0.99	830
Weighted avg	0.99	0.99	0.99	830

The classification report shows that the model performs strongly for both classes. For the low-risk class, the model achieved a precision of 0.99, recall of 0.98, and F1-score of 0.99. For the high-risk class, the model achieved a precision of 0.98, recall of 0.99, and F1-score of 0.99. The high recall for the high-risk class is especially important in the context of responsible gambling, because failing to detect high-risk players may have more serious consequences than incorrectly flagging a small number of low-risk players.

The confusion matrix further supports this interpretation. Out of 498 low-risk players in the test set, 490 were correctly classified as low-risk, while only 8 were misclassified as high-risk. All 332 high-risk players were correctly classified as high-risk. This means that the model did not miss any high-risk players in the test set.

Table 4.19: Confusion matrix for the Logistic Regression model

Actual Class	Predicted Low-risk	Predicted High-risk
Low-risk	490	8
High-risk	0	332

From a responsible gambling perspective, this result is meaningful because the model demonstrates strong sensitivity toward high-risk cases. In practical harm-reduction systems, identifying potentially risky players early is more important than achieving only overall accuracy. A model that misses high-risk players may fail to support timely intervention, whereas a model that flags some additional players can still be reviewed through further responsible gambling checks.

Overall, the supervised classification results show that the engineered behavioural features form a highly structured representation of gambling behaviour. The updated experiment is more methodologically sound than the earlier version because it removes the direct leakage feature from the model input. Nevertheless, the results remain dependent on internally constructed labels. Therefore, the findings should be interpreted as evidence that the proposed behavioural feature space is useful for risk-oriented classification, while future work should validate these predictions against external harm indicators.

4.6 Generalization Analysis of Supervised Models

Here, we examine the generalization ability of the supervised learning models by comparing training, validation and test accuracy. This comparison aims to verify whether the models have a stable learning pattern and to detect overfitting phenomena. We show two rounds of comparisons. First, we compare training accuracy with validation accuracy. Then, validation accuracy is compared with test accuracy. The tables and figures shown here give a comprehensive picture of model stability prior to the final results of the study.

4.6.1 Training and Validation Performance

Table 4.20: Training and validation accuracy across supervised models

Model	Train Accuracy	Validation Accuracy	Train–Validation Gap
LogisticRegression	**0.9917**	**0.9910**	**0.0008**
SVM(RBF)	0.9858	0.9840	0.0017
GradientBoosting	0.9989	0.9825	0.0163
KNN	0.9871	0.9819	0.0052
RandomForest	0.9865	0.9777	0.0088
ExtraTrees	0.9745	0.9717	0.0028
DecisionTree	0.9858	0.9717	0.0141
AdaBoost	0.9605	0.9524	0.0081

Table 4.20 shows that all supervised models had high training and validation accuracy, which suggests that the extracted behavioral features are very useful for classification. The train–validation gaps of Logistic Regression and SVM(RBF) are very small, which indicates good generalization. Gradient Boosting and Decision Tree display slightly larger gaps, which could be indicative of slightly higher overfitting to the training data than the simpler models. But overall, the gaps are still rather small, implying that overfitting is not a big issue in this phase.

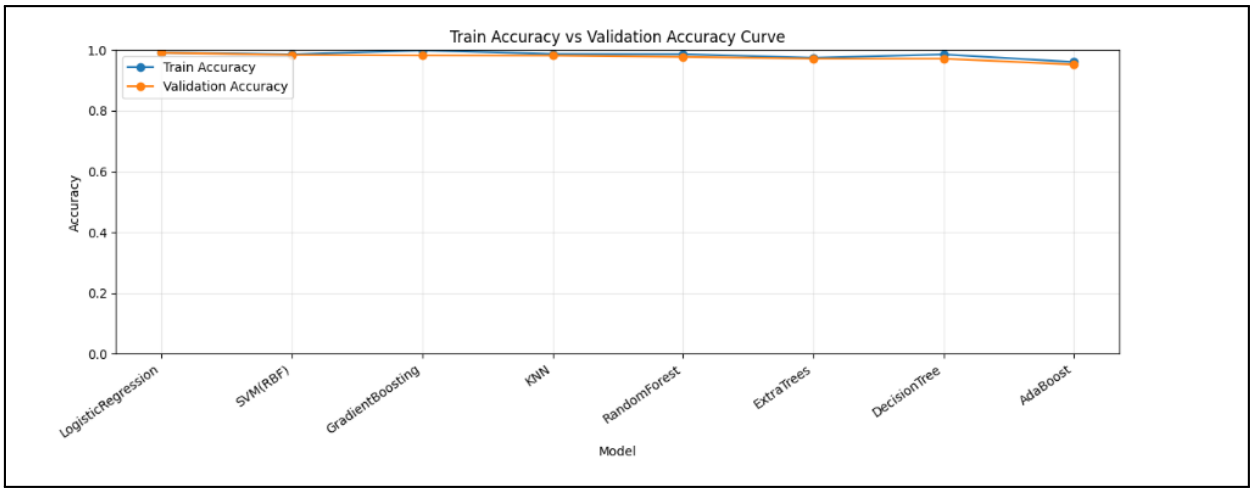

Figure 4.7: Train accuracy VS validation accuracy across supervised models.

Figure 4.7 visually confirms the numerical comparison reported in Table 4.20. The training and validation lines are pretty close for most models, suggesting that their predictive performance is similar for both datasets. The figure also facilitates the observation of the slight performance decreases from training to validation, particularly for Gradient Boosting, Random Forest and Decision Tree. In general, the visual evidence supports the claim that the models have good generalisation performance from training to validation.

4.6.2 Validation and Test Performance

Table 4.21 shows the differences between accuracy on validation and test sets, to assess the final predictive performance. The gaps between the validation and test performances are small for most models, meaning that their predictions are consistent for new data. Logistic Regression, Gradient Boosting, KNN and Random Forest have the smallest differences, which indicate high stability across the validation and test sets. Decision Tree and AdaBoost have slightly higher performance on the test set than on the validation set, which could be due to randomness. Overall, the table suggests that the supervised models perform well after the validation stage.

Table 4.21: Validation and test accuracy across supervised models

Model	Validation Accuracy	Test Accuracy	Validation–Test Gap
LogisticRegression	**0.9910**	**0.9904**	**0.0006**
SVM(RBF)	0.9840	0.9771	0.0069
GradientBoosting	0.9825	0.9807	0.0018
KNN	0.9819	0.9795	0.0024
RandomForest	0.9777	0.9759	0.0018
ExtraTrees	0.9717	0.9651	0.0066
DecisionTree	0.9717	0.9783	-0.0066
AdaBoost	0.9524	0.9566	-0.0042

Figure 4.8 shows the validation vs test accuracy for all supervised models. These two lines are almost identical, which supports the observations in Table 4.21. The figure also shows the ranking of models is consistent. This reinforces the conclusion that the feature space captures meaningful aspects of the behavior and the classifiers are well-generalized to new data.

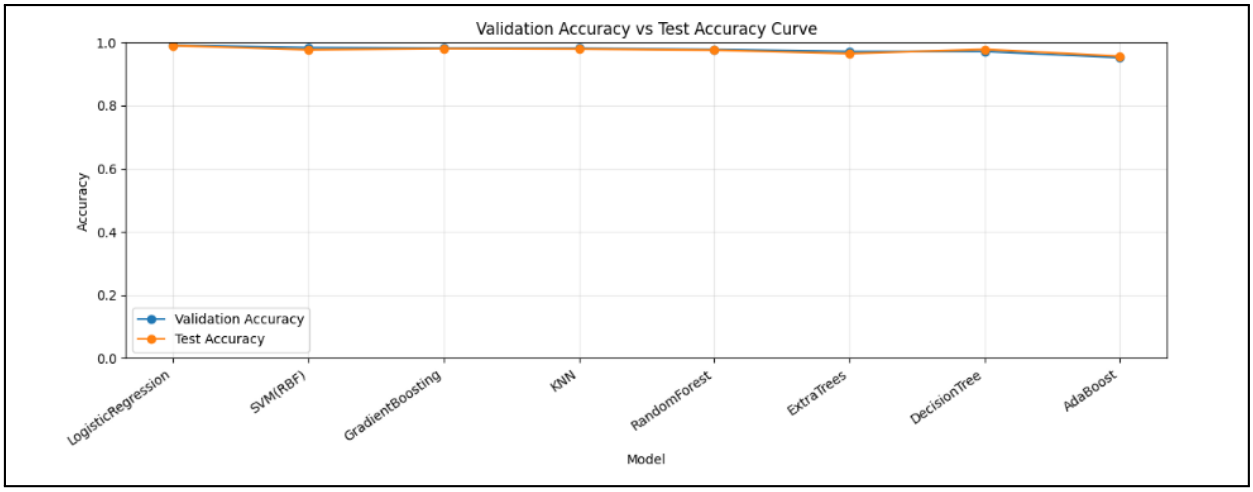

Figure 4.8: Train accuracy VS validation accuracy across supervised models.

Taken together, the two tables and two figures indicate a consistent performance of the supervised models across training, validation, and test. The associated performance differences

suggest that the engineered behavioral constructs are a robust feature space that is highly discriminative. These results offer further evidence of the validity of the supervised stage and enhance the empirical basis of the study prior to the presentation of the final conclusion.

4.7 Discussion

The supervised results show that different models performed differently on the same behavioural feature space. Logistic Regression achieved the highest result because the final leakage-safe features appear to separate low-risk and high-risk players in a clear and consistent way. Since Logistic Regression works well when classes can be separated through a weighted combination of features, it was able to classify the two risk groups effectively.

Tree-based models such as Decision Tree, Random Forest, and Gradient Boosting also performed well, but slightly lower than Logistic Regression. These models are useful for capturing nonlinear and rule-based relationships. However, in this case, the risk pattern may not require highly complex decision rules. Therefore, the simpler Logistic Regression model performed better than more complex models.

AdaBoost and Extra Trees showed comparatively lower performance. AdaBoost may overfocus on difficult or borderline samples, while Extra Trees uses highly randomised splits that may lose some structured information in a small feature space. Overall, the result suggests that the engineered behavioural features form a clear risk structure, but the classification should still be interpreted as internally derived behavioural risk classification, not clinically validated gambling-harm prediction.

4.8 Comparison with Existing Studies

This section compares the current study with several past studies as listed in the same order as references in the paper. Earlier studies sometimes do not report classification accuracy, and so

the comparison is based on the quantitative performance measure used in each study, such as accuracy, ROC–AUC, correlation, prevalence contrast, or best-model summary [13, 21–23, 25, 27–29].

Table 4.22: Comparison of the present study with selected existing studies

Study	**Method / Context**	**Reported Accuracy / Quantitative Output**
Dudfield et al. [13]	Meta-analysis of Big Five and problem gambling	Neuroticism: $r = .31$; Conscientiousness: $r = -.28$
Philander [21]	Classification / regression; ANN	Artificial Neural Network reported as the most reliable hold-out classifier
Hopfgartner et al. [22]	Logistic regression and five machine learning models	ROC–AUC = 0.654–0.717
Auer and Griffiths [23]	LR, RF, GB, DT, SVM	Logistic Regression AUC = 0.789; Random Forest AUC = 0.776
Percy et al. [25]	LR, BN, NN, RF for self-exclusion prediction	Random Forest reported as the best-performing supervised model
Canale et al. [27]	Survey analysis of adolescent internet gambling	Problem gambling rate = 21.9% among online gamblers vs 4.0% among non-online gamblers

Study	Method / Context	Reported Accuracy / Quantitative Output
Potenza et al. [28]	Chi-square and logistic regression on adolescent internet gambling	Higher at-risk/problem gambling rate reported among internet gamblers
Botella-Guijarro et al. [29]	COM-B + PLS-PM behavioral prediction	Risk factors significantly predicted gambling behavior
Our Study	Behavioral feature engineering + supervised risk classification	**Logistic Regression accuracy = 0.9904; F1 = 0.9881** Gradient Boosting accuracy = 0.9807; F1 = 0.9758 KNN accuracy = 0.9795; F1 = 0.9744

Table 4.22 shows that quantitative analysis is supported by earlier studies that demonstrate the usefulness of statistical and machine learning approaches for studying risky gambling behaviour. Philander, Hopfgartner, Auer and Griffiths, and Percy show the value of machine learning and account-based behavioural variables in identifying or predicting risky gambling patterns, while Dudfield, Canale, Potenza, and Botella-Guijarro provide quantitative evidence from personality, adolescent gambling, and behavioural prediction research. In comparison, the current study achieved very strong internal supervised classification performance after removing the direct leakage feature *RiskIntensity* from the model input. Logistic Regression produced the best result, with an accuracy of 0.9904 and an F1 score of 0.9881, showing that the remaining engineered behavioural indicators provide a highly structured and interpretable basis for distinguishing internally defined low-risk and high-risk gambling profiles.

However, this comparison should be interpreted carefully because the studies use different datasets, target definitions, modelling approaches, and evaluation metrics. Some previous stud-

ies used clinical, survey-based, or harm-related outcomes, whereas the current study uses internally constructed behavioural risk labels. Therefore, the high performance of the current study should not be treated as evidence of clinically validated gambling-harm prediction.

Chapter 5

Conclusion

Chapter 5: Conclusion

5.1 Overview

This chapter provides a summary of the thesis' findings, it discusses the limitations of the study and proposes areas for future research. This research sought to apply machine learning to model online crash gambling behavior, and to determine how it structurally relates to Big Five personality profiles. To achieve this, Bustabit gameplay data were transformed into player-level behavioural constructs, and a Big Five data set was used to obtain clusters of personality profiles. We then compared cluster spaces at the centroid level, using a Hungarian algorithm, and used supervised learning models to test the discriminative power of engineered behavioural features.

5.2 Conclusion

This thesis shows that gambling logs of online crash-gambling can be used to engineer informative player-level profiles. By reducing Bustabit round-level gambling data to meaningful constructs such as risk intensity, loss chasing, volatility seeking, control, engagement and incentive sensitivity, distinct sub-groups of gambling behavior were identified using clustering techniques. The results indicate that it's not a single pattern of gambling, but a set of different patterns from relatively controlled gambling to more volatile and risky gambling.

The study also demonstrates that these behavioral patterns can be compared (exploratively) with independently derived Big Five personality structures. While the two data sets were not connected at the individual level, the comparison at the centroid level showed that there were similarities between the structures of the behavioral clusters and the personality clusters. More specifically, higher-risk or more unstable patterns of gambling were more closely linked with profiles that are more neurotic and less conscientious, while more stable or lower-risk patterns of gambling were more linked with less neurotic and more conscientious profiles.

Finally, the supervised validation results demonstrated the engineered behavioral features con-

stitute a structured and discriminative feature space. Tree-based and ensemble-based classification showed a particularly high discriminative ability for higher- and lower-risk patterns. In conclusion, the thesis proposes an interpretable approach for online gambling behavior analytics, which is based on feature engineering, unsupervised clustering, centroid-level Hungarian matching, and supervised validation.

5.3 Limitations

- The Bustabit and Big Five data sets were not matched, so the study was not able to look at the link between personality and gambling on an individual level. So, the centroid alignment can be regarded as structural and exploratory, not direct measurement of personality.
- The mapping of gambling behavior to personality was based on conceptual assumptions, such as control and conscientiousness, and loss sensitivity and neuroticism. While these assumptions are theoretically valid, such links are not the same as psychometric validations.
- The supervised approach was evaluated using the internally generated labels, rather than external clinical risk or harm data.
- Publicly available data sources were employed, which might not cover all gambling platforms and players.

5.4 Future Work

- We should work with linked data on both gambling behavior and personality so that we can directly study the association between individuals.

- It would also be helpful to include external harm data (PGSI, self-exclusion, changes of deposit limits or clinical screening) to better validate the approach.
- Another important direction is temporal modeling, where sequence-based methods could be used to study how gambling risk develops over time.
- Research should be conducted on best practice for ethical use of behavioral risk detection systems to ensure that such systems are transparent and auditable, and used to prevent harm rather than exploit the consumer.

Chapter 6

References

References

[1] H. R. Lesieur and S. B. Blume, "The South Oaks Gambling Screen (SOGS): A new instrument for the identification of pathological gamblers," *American Journal of Psychiatry*, vol. 144, no. 9, pp. 1184–1188, 1987.

[2] A. Blaszczynski and L. Nower, "A pathways model of problem and pathological gambling," *Addiction*, vol. 97, no. 5, pp. 487–499, 2002.

[3] F. Calado and M. D. Griffiths, "Problem gambling worldwide: An update and systematic review of empirical research (2000–2015)," *Journal of Behavioral Addictions*, vol. 5, no. 4, pp. 592–613, 2016.

[4] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. Arlington, VA, USA: American Psychiatric Publishing, 2013.

[5] E. Dow, S. Kearney, and M. Day, "Absolute risks and decision tools for communicating the risks of visual impairment from myopia-related diseases," *Investigative Ophthalmology and Visual Science*, vol. 66, no. 4, p. 82, 2025.

[6] J. Braverman and H. J. Shaffer, "How do gamblers start gambling: Identifying behavioural markers for high-risk internet gambling," *European Journal of Public Health*, vol. 22, no. 2, pp. 273–278, 2012.

[7] S. Dragicevic, G. Tsogas, and A. Kudic, "Analysis of casino online gambling data in relation to behavioural risk markers for high-risk gambling and player protection," *International Gambling Studies*, vol. 11, no. 3, pp. 377–391, 2011.

[8] M. M. Auer and M. D. Griffiths, "The use of personalized behavioral feedback for online gamblers: An empirical study," *Frontiers in Psychology*, vol. 6, p. 153256, 2015.

[9] L. R. Goldberg *et al.*, "The international personality item pool and the future of public-domain personality measures," *Journal of Research in Personality*, vol. 40, no. 1, pp. 84–96, 2006.

[10] O. P. John, E. M. Donahue, and R. L. Kentle, "Big Five Inventory," *Journal of Personality and Social Psychology*, 1991.

[11] J. M. Digman, "Personality structure: Emergence of the five-factor model," *Annual Review of Psychology*, vol. 41, pp. 417–440, 1990.

[12] V. V. MacLaren, L. A. Best, M. J. Dixon, and K. A. Harrigan, "Problem gambling and the five factor model in university students," *Personality and Individual Differences*, vol. 50, no. 3, pp. 335–338, 2011.

[13] F. W. Dudfield, J. M. Malouff, and J. Meynadier, "The association between the five-factor model of personality and problem gambling: A meta-analysis," *Journal of Gambling Studies*, vol. 39, no. 2, pp. 669–687, 2023.

[14] J. M. von der Heiden and B. Egloff, "Associations of the Big Five and locus of control with problem gambling in a large Australian sample," *PLOS ONE*, vol. 16, no. 6, p. e0253046, 2021.

[15] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[16] J. MacQueen, "Multivariate observations," in *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, Oakland, CA, USA: University of California Press, 1967, vol. 1, pp. 281–297.

[17] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[18] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, nos. 1–2, pp. 83–97, 1955.

[19] N. Hing *et al.*, "Gambling prevalence and gambling problems amongst land-based-only, online-only and mixed-mode gamblers in Australia: A national study," *Computers in Human Behavior*, vol. 132, p. 107269, 2022.

[20] S. M. Gainsbury, "Online gambling addiction: The relationship between Internet gambling and disordered gambling," *Current Addiction Reports*, vol. 2, pp. 185–193, 2015.

[21] K. S. Philander, "Identifying high-risk online gamblers: A comparison of data mining procedures," *International Gambling Studies*, vol. 14, no. 1, pp. 53–63, 2014.

[22] N. Hopfgartner, M. Auer, D. Helic, and M. D. Griffiths, "Using artificial intelligence algorithms to predict self-reported problem gambling among online casino gamblers from different countries using account-based player data," *International Journal of Mental Health and Addiction*, pp. 1–23, 2024.

[23] M. Auer and M. D. Griffiths, "Using machine-learning algorithms to predict self-reported problem gambling among a sample of online gamblers," *International Journal of Mental Health and Addiction*, pp. 1–29, 2026.

[24] L. M. P. Lannes, "Unsupervised learning applied to the segmentation of users of online gambling platforms in Portugal: The effects of the Covid-19 pandemic on user behavior and segmentation," Master's thesis, Universidade NOVA de Lisboa, Portugal, 2021.

[25] C. Percy, M. Franca, S. Dragičević, and A. d'Avila Garcez, "Predicting online gambling self-exclusion: An analysis of the performance of supervised machine learning models," *International Gambling Studies*, vol. 16, no. 2, pp. 193–210, 2016.

[26] W. S. Slutske, A. Caspi, T. E. Moffitt, and R. Poulton, "Personality and problem gambling:

A prospective study of a birth cohort of young adults," *Archives of General Psychiatry*, vol. 62, no. 7, pp. 769–775, 2005.

[27] N. Canale, M. D. Griffiths, A. Vieno, V. Siciliano, and S. Molinaro, "Impact of Internet gambling on problem gambling among adolescents in Italy: Findings from a large-scale nationally representative survey," *Computers in Human Behavior*, vol. 57, pp. 99–106, 2016.

[28] M. N. Potenza *et al.*, "Correlates of at-risk/problem internet gambling in adolescents," *Journal of the American Academy of Child and Adolescent Psychiatry*, vol. 50, no. 2, pp. 150–159.e3, 2011.

[29] A. Botella-Guijarro, D. Lloret-Irles, J. V. Segura-Heras, and J. A. Moriano-Leon, "Characterization and prediction of gambling behavior in adolescents using the COM-B model," *PLOS ONE*, vol. 17, no. 11, p. e0277520, 2022.

[30] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[31] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, 1996, vol. 96, no. 34, pp. 226–231.

[32] J. H. Ward Jr., "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[33] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[34] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the Society for Industrial and Applied Mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

[35] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[36] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[37] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[38] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[39] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[40] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[41] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. Hoboken, NJ, USA: John Wiley and Sons, 2013.

[42] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Journal of Personality*, vol. 60, no. 2, pp. 175–215, 1992.

[43] O. John, "The Big-Five trait taxonomy: History, measurement, and theoretical perspectives," 1999.

[44] C. J. Soto and O. P. John, "Short and extra-short forms of the Big Five Inventory-2: The BFI-2-S and BFI-2-XS," *Journal of Research in Personality*, vol. 68, pp. 69–81, 2017.

[45] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density

estimation," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 611–631, 2002.

[46] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[47] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, p. e0118432, 2015.

[48] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, 2020.

[49] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proc. SODA*, 2007, vol. 7, pp. 1027–1035.

[50] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, vol. 4, no. 4. New York, NY, USA: Springer, 2006.

[51] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[52] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[53] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[54] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer-Verlag, 2002.

[55] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 11, pp. 559–572, 1901.

[56] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: John Wiley and Sons, 2009.

[57] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.

[58] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: An overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86–97, 2012.

[59] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. 2009.

[60] G. James, *An Introduction to Statistical Learning with Applications in R*. 2013.

[61] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[62] P. T. Costa Jr. and R. R. McCrae, *NEO Personality Inventory*. Washington, DC, USA: American Psychological Association, 2000.

[63] L. R. Goldberg, "An alternative description of personality: The Big-Five factor structure," *Journal of Personality and Social Psychology*, vol. 59, no. 6, pp. 1216–1229, 1990.

[64] S. P. Whiteside and D. R. Lynam, "The five factor model and impulsivity: Using a structural model of personality to understand impulsivity," *Personality and Individual Differences*, vol. 30, no. 4, pp. 669–689, 2001.

[65] L. Clark, "Decision-making during gambling: An integration of cognitive and psychobiological approaches," *Philosophical Transactions of the Royal Society B*, vol. 365, no. 1538, pp. 319–330, 2010.

[66] J. Haefeli, S. Lischer, and J. Schwarz, "Early detection items and responsible gambling features for online gambling," *International Gambling Studies*, vol. 11, no. 3, pp. 273–288, 2011.

[67] M. Griffiths, "Internet gambling: Issues, concerns, and recommendations," *CyberPsychology and Behavior*, vol. 6, no. 6, pp. 557–568, 2003.

[68] A. Broda, D. A. LaPlante, S. E. Nelson, R. A. LaBrie, A. L. Bosworth, and H. J. Shaffer, "Virtual harm reduction efforts for Internet gambling: Effects of deposit limits on actual Internet sports gambling behavior," *Harm Reduction Journal*, vol. 5, no. 1, p. 27, 2008.

[69] S. E. Nelson, D. A. LaPlante, A. J. Peller, A. Schumann, R. A. LaBrie, and H. J. Shaffer, "Real limits in the virtual world: Self-limiting behavior of Internet gamblers," *Journal of Gambling Studies*, vol. 24, no. 4, pp. 463–477, 2008.

[70] R. A. LaBrie, D. A. LaPlante, S. E. Nelson, A. Schumann, and H. J. Shaffer, "Assessing the playing field: A prospective longitudinal study of Internet sports gambling behavior," *Journal of Gambling Studies*, vol. 23, no. 3, pp. 347–362, 2007.

[71] D. A. LaPlante, J. H. Kleschinsky, R. A. LaBrie, S. E. Nelson, and H. J. Shaffer, "Sitting at the virtual poker table: A prospective epidemiological study of actual Internet poker gambling behavior," *Computers in Human Behavior*, vol. 25, no. 3, pp. 711–717, 2009.

[72] J. MacKillop, M. T. Amlung, L. R. Few, L. A. Ray, L. H. Sweet, and M. R. Munafò, "Delayed reward discounting and addictive behavior: A meta-analysis," *Psychopharmacology*, vol. 216, no. 3, pp. 305–321, 2011.