



MULTIMEDIA UNIVERSITY OF KENYA

FACULTY OF COMPUTING & INFORMATION TECHNOLOGY

**GAMBLE-AWARE: USING BIG DATA AND
DECISION TREES TO DETECT ONLINE SPORTS
GAMBLING DISORDER (GD)**

BY

NAME: GEOFREY OBARA ONGETA

REG. No: CIT-227-050/2018

SUPERVISOR:

APRIL, 2023

Project Proposal submitted in partial fulfilment of the requirements of Bachelor of Science in
Software Engineering/Computer Science

DECLARATION

I hereby declare that this Project Proposal is my own work and has, to the best of my knowledge, not been submitted to any other institution of higher learning.

Student: Geoffrey Obara Ongeta

Registration Number: CIT-227-050/2018

Signature:

Date.....

This Project Proposal has been submitted as a partial fulfilment of requirements for the Bachelor of Science in Software Engineering of Multimedia University of Kenya with my approval as the University supervisor.

Supervisor: _____

Signature:

Date:

ACKNOWLEDGEMENTS

First, I take this opportunity to thank the Almighty God for being with me all throughout this journey till the completion of my project. He has given me enough strength, health, wisdom and has protected me. Moreover, I extend my gratitude to my supervisor, Mr. Adunya. His valuable guidance and feedback have given me brilliant ideas on how to improve the project. He also was a great assistance in come up with this research paper that is up to standard. I also want to thank all the lecturers who have ever taught me. The foundation they laid in for me has greatly helped me formulate this research project. Also, I would like to appreciate my classmates for the teamwork and support that has helped me towards improving this project. Lastly, special gratitude go to my parents who have been a source of encouragement and for being with me every step of the way

ABSTRACT

Gambling addiction, also known as pathological gambling, is a long-standing impulse control condition with several negative consequences. Online gambling is diagnosed in the same way that conventional gambling problems are, but it has its own set of issues that are not evident or easily detected in regular gambling. People who have an Internet connection can participate in online gambling. Also, because there is no control over the setting in which internet gambling occurs, people can gamble while under the influence of drugs or alcohol and end up spending even more money. With these specific issues, it is critical to identify users who have a proclivity to become problem gamblers later on. This study attempts to answer the issue of whether it is feasible to identify probable gambling addicts on a high traffic gambling website using predictive analysis and machine learning technologies. This paper's work includes recording user data and developing a data pipeline for storing and aggregating user data. Machine learning algorithms were then utilised to categorise users using the Problem Gambling Severity Index (PGS1) based on their site usage history over time.

TABLE OF CONTENTS

DECLARATION	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
TABLE OF CONTENTS	v
LIST OF ABBREVIATIONS	viii
TABLE OF FIGURES	ix
LIST OF TABLES	x
CHAPTER 1: INTRODUCTION	1
1.1 Background of Study	1
1.2 Problem Statement	2
1.3 Aim of Study	2
1.3.1 Research Objectives	2
1.4 Significance of Study	2
1.5 Scope	2
1.6 Assumptions	3
1.7 Limitations	3
CHAPTER 2: LITERATURE REVIEW	4
2.1 Introduction	4
2.2 Related Systems	4
2.2.1 Ludens: A Gambling Addiction Prevention.	4
2.2.3 GamblingLess: Curb Your Urge	5
2.2 Limitations of systems	5
2.3 Solutions of proposed system	5
CHAPTER 3: METHODOLOGY	6
3.1 Introduction	6
3.2 Methodology	6
3.3 Data Sets	8
3.4 Project Resources	9
3.4.1 Human Resource	9
3.4.2 Reusable Components	10
3.4.3 Hardware and Software tools	10
3.5 Project Schedule	10

CHAPTER 4: SYSTEM ANALYSIS	12
4.1 Introduction	12
4.2 Analysis of current system	12
4.2.1 Problem Gambling Severity Index	12
4.3 System Requirements	14
4.3.1 User Requirements	14
4.3.2 Functional requirements	14
4.3.3 Non- functional Requirements	15
CHAPTER 5: SYSTEM DESIGN	16
5.1 Introduction	16
5.2 Architectural Design	16
5.3 Data Set design	18
5.3.1 Data set and feature list	18
5.3.2 Feature engineering	20
5.4 User Interface Design	22
5.4.1 Streamlit	22
CHAPTER 6: IMPLEMENTATION AND TESTING	24
6.1 Introduction	24
6.2 Development Environment	24
6.3 System Components	24
6.3.1 Data Collection and Preparation:	24
6.3.2 Model Selection and Optimization:	24
6.3.3 Model Deployment:	24
6.3.4 Monitoring and Maintenance:	25
6.3.5 Visualization and Interpretability:	25
6.4 Test Data	25
6.4.1 Invariance Test	25
6.4.2 Directional Expectation Test	25
6.4.3 Minimum Functionality Test	26
6.5 Test Results	26
CHAPTER 7: CONCLUSION	27
7.1 Achievements and Lessons learnt	27
7.2 Conclusions	27
7.3 Recommendations	28

BIBLIOGRAPHY	29
APPENDIX	31
APPENDIX1: Project Gantt chart	31

LIST OF ABBREVIATIONS

DSE	Data Science Edge
SAIC	Science Applications International Corporation
BDA	Big Data Analytics
CRISP-DM	Cross-Industry Standard Process Model-Data Mining
CBT	Cognitive Behavioural Therapy
EMI	Ecological Momentary Intervention
ERP	Exposure Response Prevention
BCLB	Betting and Licencing Control Board
PGSI	Problem Gambling Severity Index

TABLE OF FIGURES

Figure 1: CRISP-DM Methodology	6
Figure 2: Data Analytics Ladder.....	7
Figure 3: Data Science Edge.....	7
Figure 4: Resource Pyramid.....	9
Figure 5: Machine Learning Workflow	11
Figure 6: PGSI Flow Chart	13
Figure 7: High Level System Architecture	16
Figure 8: Data Logging Architecture.....	17
Figure 9: Python Log Analyser Architecture.....	17
Figure 10: Process Automation Architecture.....	18
Figure 11: Data Types in Dataset.....	18
Figure 12: Labels against Frequency	19
Figure 13: Time Spent Gambling against Frequency	19
Figure 14: Feature Importance.....	20
Figure 15: Random Forest Epoch Training	21
Figure 16: Confusion Matrix	21
Figure 17: Model Performance Metrics.....	22
Figure 18: Streamlit User Interface Design	23
Figure 19: System Prediction Output.....	23
Figure 20: System Prediction Output.....	23
Figure 21: Project Gantt chart.....	31

LIST OF TABLES

Table 1: Collected Data Definition	8
Table 2: Project Schedule	11

CHAPTER 1: INTRODUCTION

1.1 Background of Study

At its most basic level, "sports betting" entails putting a financial wager on the outcome of a sporting event, as well as incidents that occur inside the bigger match or fixture. The popularity of sports betting and gambling during sporting events is a relatively new phenomenon. Previously, sports betting was limited to an individual physically placing a wager on the outcome of a horse race, but in the mid-1990s, two significant shifts happened. First, some bookies ventured beyond horse and greyhound racing and began accepting wagers on team sports. Second, some bookies started taking bets over the phone and later on the internet.

Increased Internet accessibility has resulted in significant advancements in our daily lives, as well as changes in how individuals gamble. For most people, gambling is merely a sort of enjoyment; but, for others, the large choice of betting and gaming activities available through Internet-enabled devices can quickly turn into a disorder with major social and psychological ramifications (Deans, Thomas, Daube, & Derevensky, 2016). In Spain, for example, the majority of gambling activity is still conducted on land, despite the fact that in 2013, online gambling rose from 20.15 percent of gambling winnings to 26.48 percent in 2015 (Dirección General de Ordenación del Juego, 2015).

Since the launch of online sports betting in Kenya back in 2013, a total of 28 companies have been licensed. As of 2017, the companies had 5 million active consumers in Kenya making bets on their websites, generating a total revenue of 20 billion Kenyan shillings (Omondi, 2018). The emergence of online betting platforms, combined with the simplicity of sending money via mobile wallets, has increased young participation in sports betting in Kenya (Koross, 2016). Kenyans bet the most, at 5000 shillings per individual bet, compared to other gamblers in Sub-Saharan Africa, primarily on football games in the English Premier League (Ssewanyana & Bitanihirwe, 2018). In Kenya, the youth market makes up the majority of the new demographic target market for sports betting operators, who engage them through social media platforms (Kiragu, 2016). Because of the high rate of unemployment among Kenyan youngsters, sports betting has become popular among them as a way to make money while having fun (Wangari, 2017).

1.2 Problem Statement

Internet gambling eliminates the need for social connection and delivers constant, real-time feedback (Bonnaire, 2012; Gainsbury, 2015). Because Internet gambling is often done in private, Internet gamblers are more likely to continue their problematic activity until it reaches a crisis point (Gainsbury, Russell, Hing, Wood, & Blaszczynski, 2013). These considerations raise the possibility that Internet gambling may play a role in the development of gambling disorders (GD). According to research, online gamblers are more likely to acquire gambling-related disorders and engage in risky activities (Kairouz et al., 2012; Wood & Williams, 2007). One particularly disturbing aspect of online gambling is that online gamblers are less likely than offline gamblers to recognize their gambling problems (Petry, 2006). As a result, it is critical that an online solution for predicting GD prevalence among internet gamblers be developed.

1.3 Aim of Study

1.3.1 Research Objectives

1. Collect, analyse and interpret data on online gambling tendencies among gamblers.
2. Employ big data analytics and machine learning model to detect Gambling Disorder (GD)
3. Utilise user gambling tendencies to provide healthy gambling notes to users.

1.4 Significance of Study

While gambling is marketed as a get-rich-quick scheme, frequent gambling establishes a tendency that most consumers are unaware of. The study's goal is to raise awareness of Gambling Disorder as a likely illness that can be avoided by monitoring gambling tendencies.

1.5 Scope

The proposed system boundaries are constrained to online sports gambling participants. The system utilises big data analytics and a machine learning model to classify gambler into six categories depending on their likelihood to develop GD. Moreover, the system would not make medical appointments to those affected, nor will the system bar the user from their website usage. The system is to be built over two academic semesters and submitted for the fulfilment of a bachelor's degree.

1.6 Assumptions

1. The system runs on a gambling website as a module that offers additional services to user as a self - assessment tool
2. The users adhere to the system's recommendations.
3. The project scope remains the same throughout development of the system.

1.7 Limitations

1. The system does not offer adequate treatment mechanisms for chronic GD patients.
2. The system cannot deter a gambling addict from using a gambling service.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Recently developed Internet-delivered approaches have shown promise as a viable treatment option for problem gamblers who are reluctant to seek face-to-face treatment (Canale et al., 2016; Myrseth, Brunborg, Eidem, & Pallesen, 2013). A recent pilot randomised controlled trial in Sweden tested the feasibility of an Internet-based treatment for problem gamblers and concerned significant others, and found that this novel intervention successfully lowered the symptoms of problem gambling and measures of depression and anxiety for gamblers (Nilsson, Magnusson, Carlbring, Andersson, & Gumpert, 2017). Lowering barriers to treatment via Internet-delivered approaches is especially relevant in Sweden since more than half (55%) of the gamblers in this country report playing online (Swedish Gambling Authority, 2015).

2.2 Related Systems

2.2.1 Ludens: A Gambling Addiction Prevention.

Ludens is a gambling addiction prevention program that has four goals: inform participants about gambling and gambling addiction; sensitise participants to the risk of gambling for health, especially addiction; promote a change in attitudes toward gambling; and alert participants to risky behaviours that can lead to addiction. The prevention program was implemented during 2017 to 2019. Fourteen psychologists presented it to 2372 adolescents (48.8% females, 51.2% males) aged 14–19 years, none of whom were university students, recruited from 42 Spanish high schools in 132 groups taking different courses. The main dependent variables analysed were the monthly frequencies of gambling, at-risk gambling, and gambling addiction (as measured by the National Opinion Research Centre DSM-IV Screen for Gambling Problems, adapted to diagnose gambling disorder according to DSM-5, in which pathological gambling is considered an addictive disorder). Given that all of the gamblers were adolescents, fulfilment of 1–3 the DSM-5 diagnostic criteria were considered to indicate a risk of problem gambling. After the administration of Ludens, statistically significant reductions were observed in the three variables of interest: monthly frequency of gambling, percentage of adolescents with risky gambling, and percentage of adolescents with gambling disorder. The results were analysed according to sex and age (minors vs. adolescents between 18 and 19 years old). The results obtained after applying the prevention program indicate that Ludens is effective as a universal prevention program for gambling addiction.

2.2.3 GamblingLess: Curb Your Urge

Cognitive Behaviour Therapy is the treatment of choice for Gambling Disorder with stimulus control (SC) with Exposure Response Prevention (ERP). Hawker, Markouris, Youssef & Dowling conducted a single-arm study that supports the acceptability, feasibility and preliminary efficacy of an app-delivered Ecological Momentary Intervention (EMI) for craving management in people with gambling problems. The app's EMI feature recommends using 12 urge-curb tips or exercises that take 1 to 5 min to complete. The content is related to psychoeducation, relaxation techniques, and mindfulness (e.g., about my urge, delay and distraction, and urge surfing). Smartphone apps have also been demonstrated to be feasible and acceptable as CBT (Cognitive Behaviour Therapy) adjunctive components to enhance homework completion in people suffering from a gambling disorder (e.g., decisional balance exercise, functional analysis of gambling behaviour, development of healthy alternatives to gambling, problem-solving, and relapse prevention exercises). Moreover, a randomised controlled trial in which a self-help CBT program was combined with a messaging app and showed promising results for overcoming the high dropout rate of unguided internet-based interventions for gambling disorders. Every day at 9 pm, participants in the intervention group received monitoring, personalised feedback, and messages based on CBT. Only 6.7% of the participants dropped out at follow-up and 77% continued participating during the trial period. Recent RCT protocol studies include apps for assessing and delivering interventions for gambling problems. It is important to improve the quality of psychological programs considering smartphone apps.

2.2 Limitations of systems

1. Despite the wide range of effective treatment options, only a few persons experiencing gambling problems seek treatment.

2.3 Solutions of proposed system

The proposed internet based solution aim to counter limitation of existing system by:

1. Making it possible to assess symptoms and to diagnose via the internet.
2. Reducing diagnosis costs.
3. Collecting data for the opportunity to integrate such treatments into regular clinical settings.

CHAPTER 3: METHODOLOGY

3.1 Introduction

This chapter delves into the approaches used to develop the proposed systems. Descriptions of the data sets utilised to actualize the system, project resources, and project scheduling approaches are also included in this chapter.

3.2 Methodology

A lifecycle of data collection, iteration, analysis, and action to achieve a mission goal is the general paradigm for data science. Several areas are outlined in the theoretical framework:

- Iteration
- Intermediate output
- Prototyping
- Listening to what the data are telling us
- A data-value pyramid for structuring the process
- Pursuing the critical path to a product
- Documenting the analytics process as it unfolds.

With agile ways to cut development time, the SDLC has altered considerably. A partnership established the Cross-Industry Standard Process Model-Data Mining (CRISP-DM) to deliver probabilistic accurate answers rather than deterministic ones.

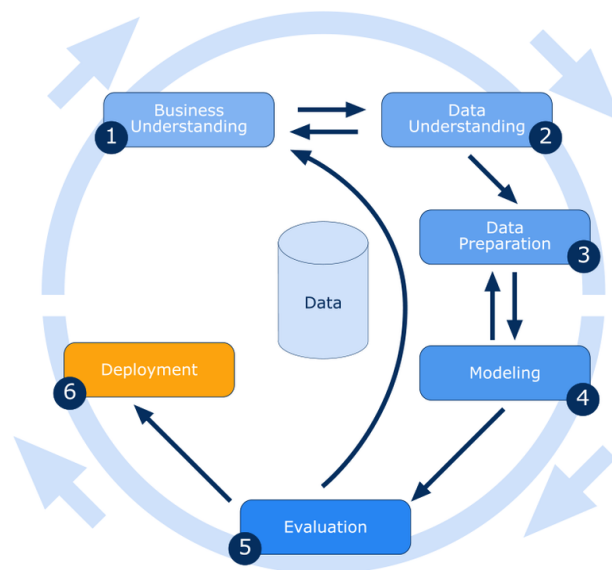


Figure 1: CRISP-DM Methodology

The CRISP-DM is motivated by the intended result. As indicated in the diagram below, analytics can be defined as a ladder of increasing complexity.

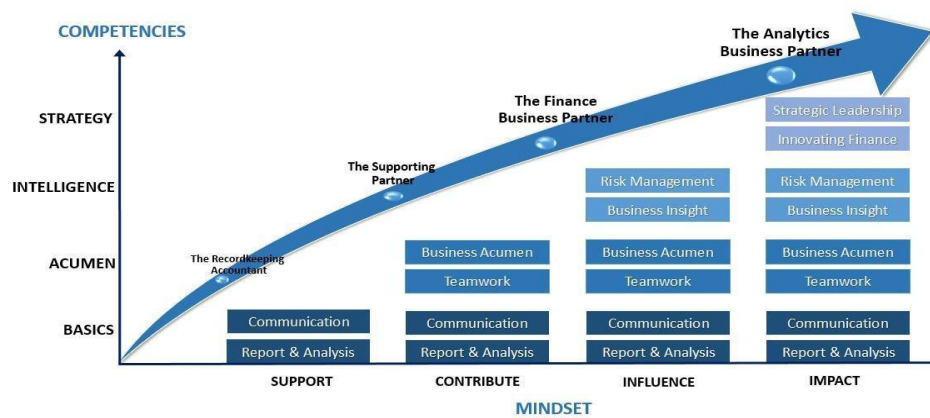


Figure 2: Data Analytics Ladder

Advanced analytics require experimentation and an iterative exploration process of testing and evaluation. SAIC has developed a Big Data Analytics model extending CRISP-DM to incorporate new technologies for big data and cloud. This BDA process model is called Data science Edge (DSE)

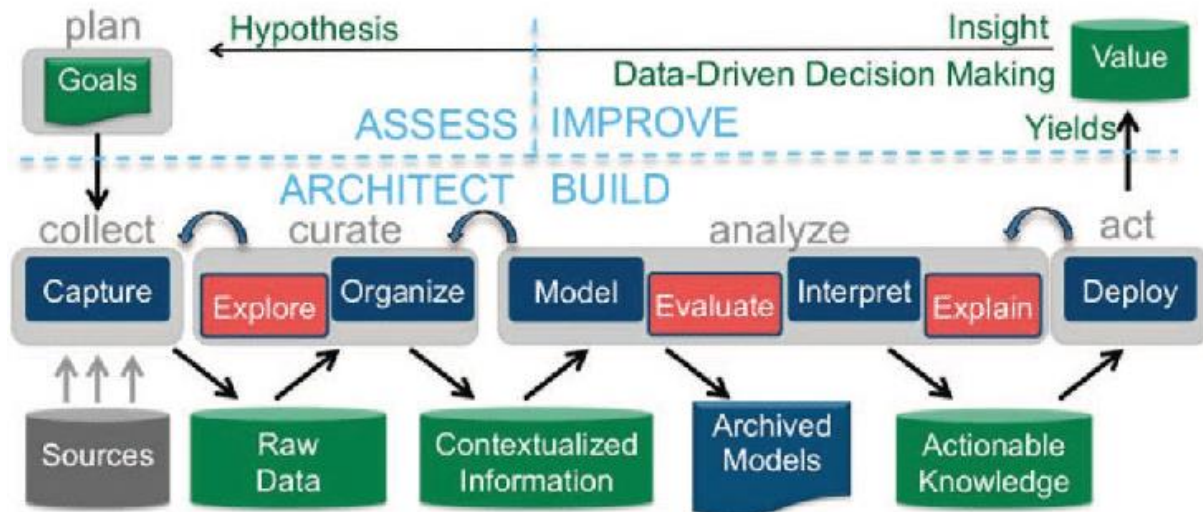


Figure 3: Data Science Edge

The DSE is a five-step model of planning, data collection, curation, analysing and acting. These steps align with CRISP-DM but consider collecting data from external sources and data storage. Appropriate NoSQL technologies are considered when working with DSE since the data sources are tethered together. The steps explicitly determine the data storage types.

3.3 Data Sets

A gambling site X has a definition about who is a betting addict by definition based on a user's history on the site. If a user satisfies any of the following criteria, they are considered a gambling addict.

- Users who spend huge sums of money on the gambling site.
- Users who spend a lot of hours on the gambling site are likely addicts
- Users who have self-excluded themselves on the website.
- Users who deposit money more often are more likely to be gambling addicts.
- Users who play more bets are likely gambling addicts.
- Users who log in more frequently are gambling addicts.

Data was collected from the gambling sites to form the data set used in the proposed system. The table below specifies the type of data collected.

Table 1: Collected Data Definition

Data	Description
Basic User Details	Basic user details such as the currency, country, age and login ID is tracked for each user.
Time Spent Per Page	how much time user is spending on each page
Usage Summary History	Summary (aggregation) of how much money user is spending per each section of the site
Deposit/Withdrawal History	Each deposit and withdrawals the user has made during the specified time.
Site Usage	This is a different way to track time on the site since time spent per page can be unreliable in cases when the user does not stay on the same page and closes the tab/page after staying inactive on it.

3.4 Project Resources

Project resources simply mean resources that are required for successful development and completion of a project. In the project planning phase, identification of resources that are required for completion of a project and how they will be allocated is a key element and a very important task to do. There are three types of resources that are considered. They are very essential for execution and completion of projects on time and on budget. These resources can be denoted by pyramids which is also known as Resource Pyramid. This is shown in following diagram:

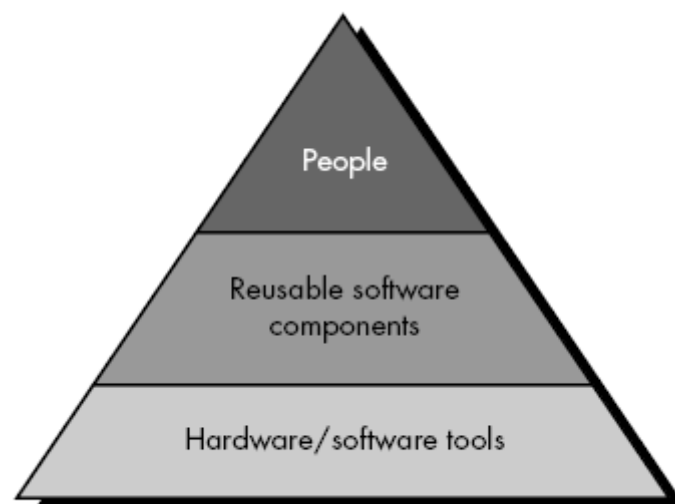


Figure 4: Resource Pyramid

Software planning for resources entails consideration of the following characteristics.

- Description of resource
- Resource availability
- Time of resource when it will be available
- Duration of resource availability

3.4.1 Human Resource

Due to the small project size, a single individual would act as the project manager and perform all software engineering tasks while consulting specialist as required. External resources required are potential customers and project supervisors.

3.4.2 Reusable Components

Reusable components are categorised into off the shelf components, full experience components, partial experience components and new components. Off the shelf components would be the data sets, partial component would be pre-trained AI models and new component would a customized machine learning model.

3.4.3 Hardware and Software tools

The tools required for the completion of the project are personal computers for research and coding, notebooks for note taking, internet services for online research, VS Code coding environment and Git version control system.

3.5 Project Schedule

The project schedule is a network of software engineering jobs that connects scope, effort estimates, and deadlines. Parallelism and task dependency are managed via the software project schedule. The 90-90 rule would be used to complete the job on schedule. According to the rule, 90 percent of the project must be done in 90 percent of the time allotted, and the remaining 10 must be completed in 90 percent of the time allotted. However, there are a number of factors that may prevent the project from reaching the deadline. These include, but are not limited to:

- Changing user requirements
- Honest underestimation of work load
- Inconsistent risk analysis
- Technical difficulties
- Unforeseen human difficulties.

The tool used in the scheduling process are Gantt charts and work flow diagrams. Gantt charts provide graphical illustrations of the schedule for planning, coordinating and tracking tasks

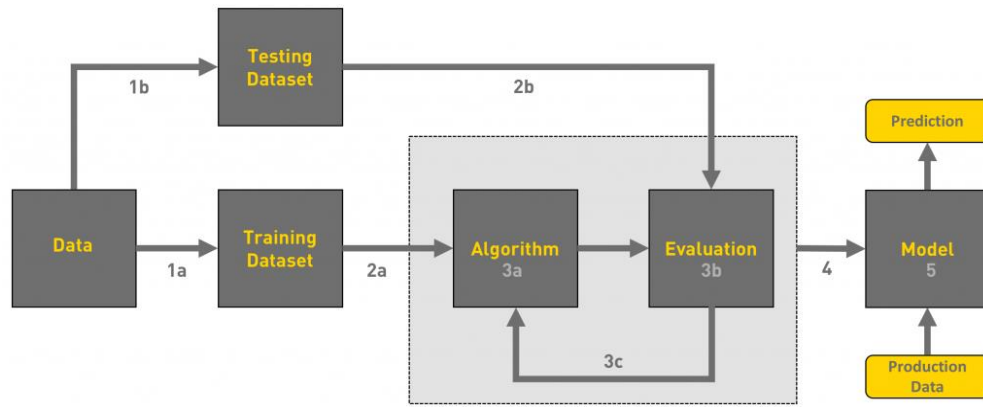


Figure 5: Machine Learning Workflow

while the task activity diagrams show dependencies of tasks. Below is a tabular representation of project task, milestones and deliverables. See Appendix 1 for the comprehensive Gantt chart

Table 2: Project Schedule

Task ID	Tasks and Milestones	Start Date	End Date
1	Draft project proposal		
1.1	Conduct background research	01/12/2022	12/12/2022
1.2	Conduct a literature review	15/12/2022	26/12/2022
1.3	Evaluate and choose a research and SDLC methodology	29/12/2022	09/01/2023
2	Trained model with an API		
2.1	Source data set and perform data cleaning	12/01/2023	14/01/2023
2.2	Split data set into training and testing data	15/01/2023	16/01/2023
2.3	Develop Decision tree algorithm to train the model	19/01/2023	30/01/2023
2.4	Evaluate model accuracy	05/02/2023	14/02/2023
2.5	Retrain model tweaking features	17/02/2023	04/03/2023
2.6	Package model and build an API	04/03/2023	15/03/2023
2.7	Type the project documentation.	17/01/2023	17/03/2023
3	Create presentation slides		
3.1	Design presentation plan	22/03/2023	25/03/2023
3.2	Create presentation slides	25/03/2023	27/03/2023
3.3	Proofread presentation slides	27/03/2023	29/03/2023

CHAPTER 4: SYSTEM ANALYSIS

4.1 Introduction

System analysis is the process of examining a system in order to identify its components, interrelationships, and properties. It involves breaking down a system into its constituent parts, studying how these parts work together, and identifying opportunities for improvement. The goal of system analysis is to design or modify a system to better meet the needs of its users or stakeholders.

4.2 Analysis of current system

4.2.1 Problem Gambling Severity Index

Used in the Health Survey for England, Scottish Health Survey and the Welsh Problem gambling Survey. The PGSI was specifically developed for use among the general population rather than within a clinical context by Ferris and Wynne (2001)

The PGSI consists of nine items and each item is assessed on a four-point scale: never, sometimes, most of the time, almost always. Responses to each item are given the following scores:

- never = zero
- sometimes = one
- most of the time = two
- almost always = three

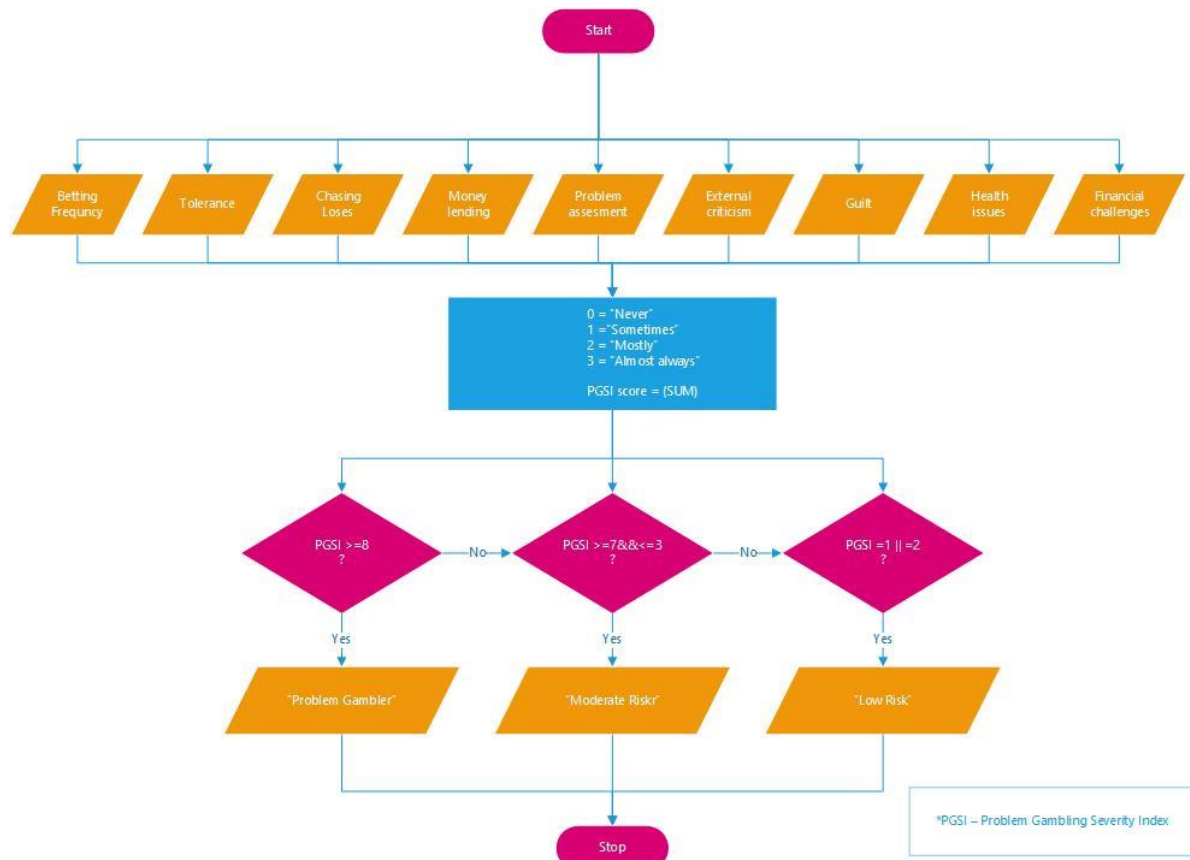


Figure 6: PGSI Flow Chart

A PGSI score of eight or more represents a problem gambler. This is the threshold recommended by the developers of the PGSI and the threshold. The PGSI was also developed to give further information on sub-threshold problem gamblers. Scores between three and seven represent ‘moderate risk’ gambling (gamblers who experience a moderate level of problems leading to some negative consequences) and a score of one or two represents ‘low risk’ gambling (Gamblers who experience a low level of problems with few or no identified negative consequences).

4.3 System Requirements

4.3.1 User Requirements

User requirements describe the features and functionality that the user or stakeholders expect from the system. Some user requirements are:

- a) User input: The system should accept user input related to online gambling behavior, such as account balance, number of bets placed, and funds withdrawn.
- b) Prediction accuracy: The system should be able to accurately predict the risk of online gambling disorder based on the user's inputs.
- c) User interface: The system should have a user-friendly interface that is easy to navigate and understand.
- d) Speed and responsiveness: The system should be fast and responsive, providing real-time predictions to the user.

4.3.2 Functional requirements

Functional requirements are specific actions or capabilities that a system must be able to perform in order to satisfy the needs of the user or stakeholders.

- a) Input validation: The system should validate user inputs to ensure they are within acceptable ranges and formats.
- b) Decision tree logic: The system should implement the decision tree algorithm to accurately predict online gambling disorder based on the user inputs.
- c) Model accuracy: The system should be tested to ensure it provides accurate predictions of online gambling disorder.
- d) User feedback: The system should provide clear and concise feedback to the user regarding their level of risk for online gambling disorder.
- e) Predictive capabilities: The system should be able to predict online gambling disorder based on the user inputs with a high degree of confidence.
- f) Real-time prediction: The system should be able to provide real-time predictions to the user based on their inputs.
- g) Compliance: The system should comply with all relevant laws and regulations, including data privacy and online gambling regulations.

4.3.3 Non- functional Requirements

Non-functional requirements are the attributes that describe how well the system performs its functions

- a) Performance: The system should be able to provide accurate predictions within a reasonable amount of time.
- b) Scalability: The system should be able to handle large datasets without compromising on the quality of predictions.
- c) Reliability: The system should be able to function correctly and consistently even under adverse conditions, such as unexpected inputs or data anomalies.
- d) Security: The system should be able to protect the data it processes and ensure that only authorized personnel have access to it.
- e) Maintainability: The system should be designed in a way that makes it easy to maintain and update as needed.
- f) Usability: The system should be user-friendly and easy to understand, with clear instructions and feedback.
- g) Accessibility: The system should be accessible to users with disabilities, such as visual or hearing impairments.
- h) Compatibility: The system should be compatible with the hardware and software environments it will be deployed in.
- i) Portability: The system should be easily transferable to different environments or platforms without significant modification.
- j) Legal and ethical considerations: The system should comply with all relevant laws and ethical standards, including data privacy laws and regulations.

CHAPTER 5: SYSTEM DESIGN

5.1 Introduction

System design is the process of defining the architecture, modules, interfaces and data for a system to satisfy given requirements.

5.2 Architectural Design

The overall structure and design of a computer system or software application is referred to as system architecture. It includes the system's various components and modules, how they interact with one another, and how they work together to achieve the desired results.

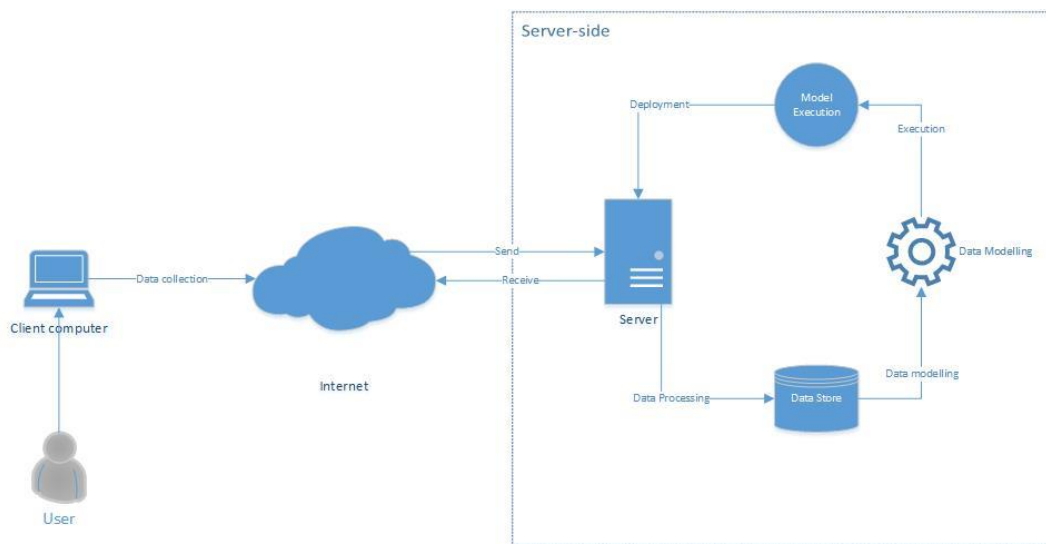


Figure 7: High Level System Architecture

The diagram above summarises the overall stem architecture. The components are the user, software such as the data, machine learning model and hardware. Hardware components are client nodes, laptops, mobile phone or PC's. Basically, data flows from the client to the server and back. The user provides input variables which are sent as HTTP request to the server which connects to the data warehouse. Data is pre-processed before being passed on to the data modelling engine which evaluated the inputs against the model to generate an accurate prediction which is send back to the users via HTTP requests.

The data collected is sent to Linux servers and stored as log files. Ngix is responsible for storing the data as log files. The figure below shows how the data logging model works.

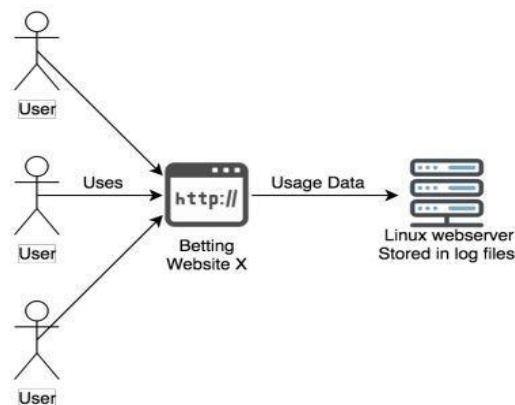


Figure 8: Data Logging Architecture

A python log analyser is run through cron to read the daily log, clean the data and store it in the database. The figure below shows the basic architecture of how data pre-processing is done.

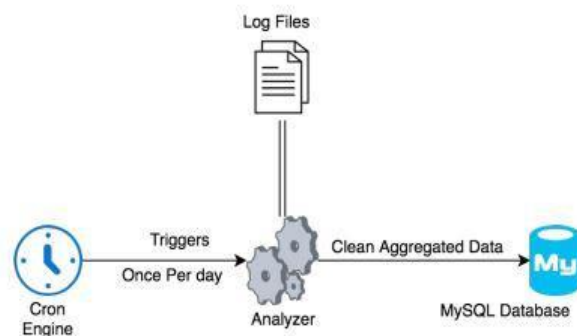


Figure 9: Python Log Analyser Architecture

The aggregated daily usage data of each user is used to build a predictor for classifying users. Gamblers who are identified by the PGSI were used to create a machine learning model of betting addiction. The features were first scaled, oversampled the minority class, used grid search to find the best hyper parameters for the algorithm and then trained the model for potential betting addicts.

Once the model is trained, it can predict users based on their behaviour history for potential betting addiction. This whole process is automated and repeated every day through cron scripts.

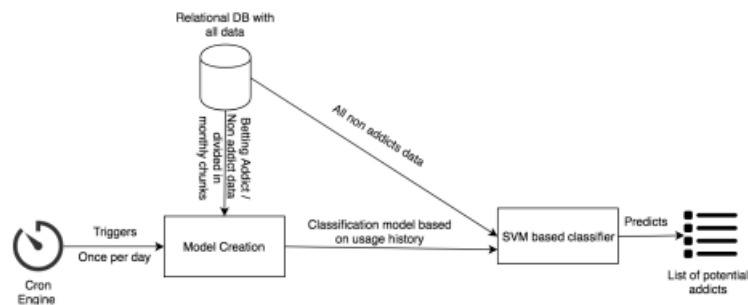


Figure 10: Process Automation Architecture

5.3 Data Set design

5.3.1 Data set and feature list

The tracked, aggregated and filtered data has 5000 records of daily user activity and an independent variable label with the PGSI scale for that day. Figure shows the list of features used for the classification task, data types and the number of records.

```

data.info()
✓ 0.4s

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 5 columns):
#   Column                Non-Null Count  Dtype
---  -
0   User Id                5000 non-null  int64
1   Time                  5000 non-null  int64
2   Win/Loss Ratio         5000 non-null  float64
3   Number of Deposits     5000 non-null  int64
4   label                  5000 non-null  int64
dtypes: float64(1), int64(4)
memory usage: 195.4 KB
  
```

Figure 11: Data Types in Dataset

The relevant features used were User Id for identification purposes, time spent gambling, number of deposits made and win/loss ratio. Which is calculated by:

$$wlr = \text{Account Balance} + \text{Total Funds Withdrawn} \div \text{Total Funds Deposited}$$

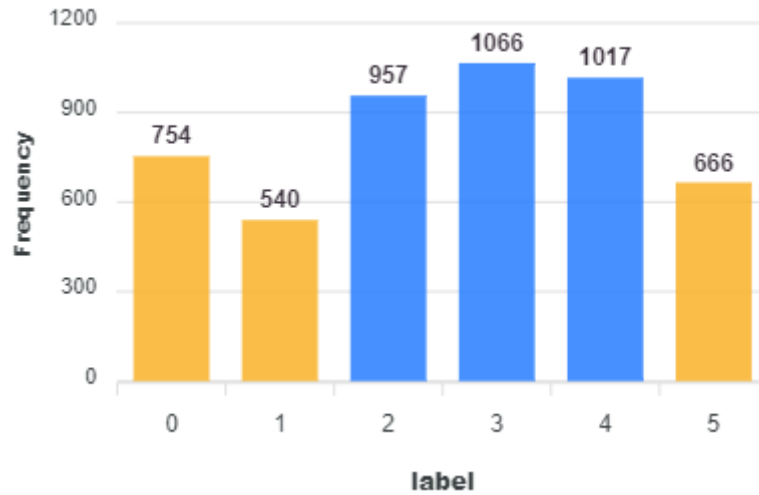


Figure 12: Labels against Frequency

Figure 12 shows the frequency of each label. Users labelled 0 which is the least problematic gambling were 754 while users with label 5 which is the highest scale in the PGSI were 666

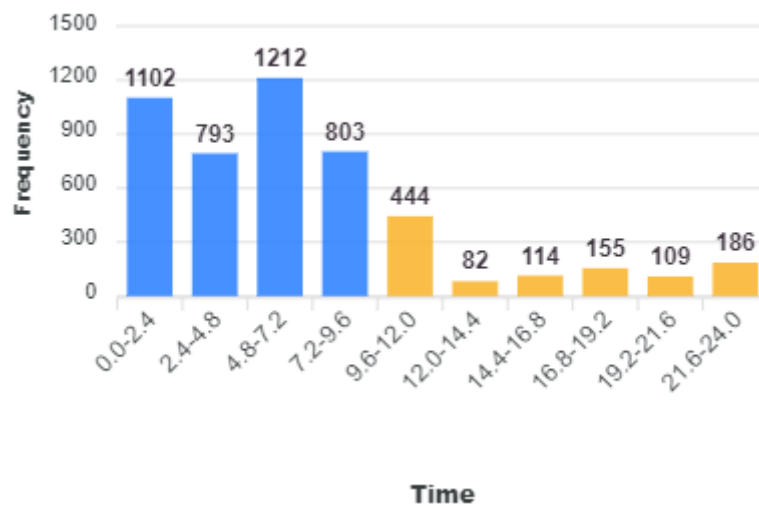


Figure 13: Time Spent Gambling against Frequency

Figure 13 shows the distribution of time spent gambling against the frequency in the dataset.

Some features such as gender, location, device information were dropped due to their low importance to the model. More feature engineering is covered in the next section.

5.3.2 Feature engineering

Feature engineering is the process of selecting, transforming, and creating new variables or features from raw data to improve machine learning model performance. It involves domain knowledge, intuition, and creativity to extract meaningful information from data that can help in building accurate models. The importance of feature engineering lies in its ability to improve the predictive power of a model by reducing noise, extracting relevant information, and identifying important variables. Good feature engineering can also reduce the amount of data required to train a model, increase interpretability, and reduce overfitting.

Data Normalization

The dataset features are rescaled so that they have the properties of a standard normal distribution with 0 mean (μ) and 1 standard deviation (σ). Data normalization resulted in features which had a different unit but were all in a similar scale.

Oversampling

The classification training data had an imbalance class distribution. More than 60% of the users were in the lower risk of GD and the remaining were on a higher scale. Creating a predictive model from such a data yields a highly biased classifier which predicts the majority class most number of times. For this reason the minority data was oversampled to match the number of majority class elements. Python's imbalanced-learn class was used to oversample the minority class using SMOTE method. It resulted in a dataset with 5000 records for each class.

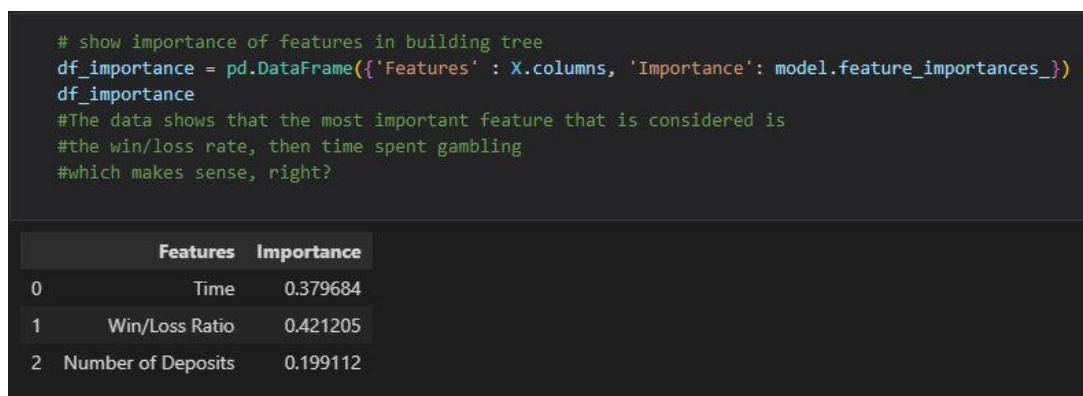


Figure 14: Feature Importance

Figure 14 shows feature importance with win/loss ratio as the most important feature followed by time spent gambling and number of deposits as the least important feature.

After the feature engineering the data was split into 80% train data and 20% test data. Two different algorithms Decision Tree Classifier and Random Forest were used for classification. Random Forest however proved insufficient with an accuracy of less than 70%.

185	02:36:16	Training epoch complete. - epoch 84, accuracy 0.6655, loss 1.4698, val_accuracy 0, val_loss 0
186	02:36:25	Training epoch complete. - epoch 85, accuracy 0.666, loss 1.4677, val_accuracy 0, val_loss 0
187	02:36:35	Training epoch complete. - epoch 86, accuracy 0.6658, loss 1.4702, val_accuracy 0, val_loss 0
188	02:36:45	Training epoch complete. - epoch 87, accuracy 0.6661, loss 1.4655, val_accuracy 0, val_loss 0
189	02:36:54	Training epoch complete. - epoch 88, accuracy 0.6672, loss 1.46, val_accuracy 0, val_loss 0,
190	02:37:04	Training epoch complete. - epoch 89, accuracy 0.6679, loss 1.4604, val_accuracy 0, val_loss 0
191	02:37:14	Training epoch complete. - epoch 90, accuracy 0.6674, loss 1.4584, val_accuracy 0, val_loss 0
192	02:37:23	Training epoch complete. - epoch 91, accuracy 0.6685, loss 1.4579, val_accuracy 0, val_loss 0
193	02:37:33	Training epoch complete. - epoch 92, accuracy 0.6686, loss 1.4507, val_accuracy 0, val_loss 0
194	02:37:43	Training epoch complete. - epoch 93, accuracy 0.6695, loss 1.4506, val_accuracy 0, val_loss 0
195	02:37:52	Training epoch complete. - epoch 94, accuracy 0.6698, loss 1.4463, val_accuracy 0, val_loss 0
196	02:38:02	Training epoch complete. - epoch 95, accuracy 0.6698, loss 1.4461, val_accuracy 0, val_loss 0
197	02:38:12	Training epoch complete. - epoch 96, accuracy 0.6702, loss 1.447, val_accuracy 0, val_loss 0
198	02:38:21	Training epoch complete. - epoch 97, accuracy 0.6708, loss 1.4432, val_accuracy 0, val_loss 0

Figure 15: Random Forest Epoch Training

The models were further evaluated using K-Fold method using 10 folds. The mean accuracy, recall, precision accuracy, F-measure and confusion matrix were used as an evaluation criteria for model performance.

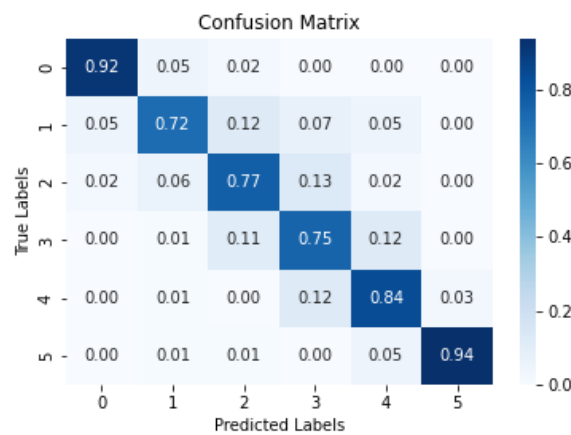
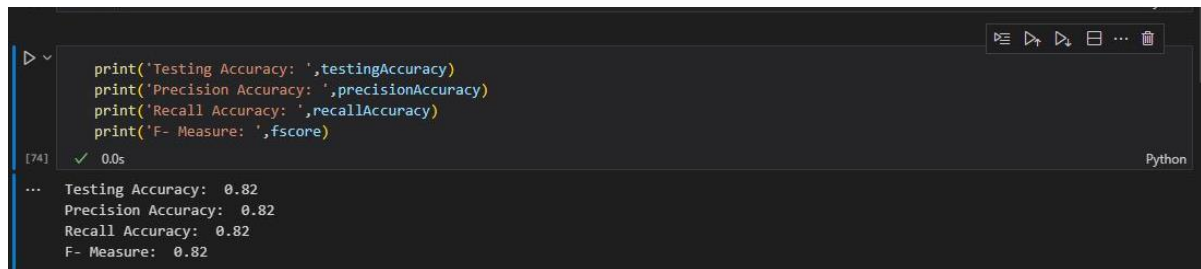


Figure 16: Confusion Matrix

Confusion matrices are important when evaluating model prediction. The darker of blue in figure 16 show frequency of prediction in percentage. The x-axis represents the predicted labels while the y-axis are the true labels from the training data. The model performed very well predicting users in label 0 and label 5.

A screenshot of a Jupyter Notebook interface. The top part shows a code cell with four lines of Python code: `print('Testing Accuracy: ',testingAccuracy)`, `print('Precision Accuracy: ',precisionAccuracy)`, `print('Recall Accuracy: ',recallAccuracy)`, and `print('F- Measure: ',fscore)`. Below the code, the output is displayed: `Testing Accuracy: 0.82`, `Precision Accuracy: 0.82`, `Recall Accuracy: 0.82`, and `F- Measure: 0.82`. The notebook interface includes a toolbar with icons for running, saving, and other actions, and a status bar at the bottom indicating the language is Python.

```
print('Testing Accuracy: ',testingAccuracy)
print('Precision Accuracy: ',precisionAccuracy)
print('Recall Accuracy: ',recallAccuracy)
print('F- Measure: ',fscore)
```

```
Testing Accuracy: 0.82
Precision Accuracy: 0.82
Recall Accuracy: 0.82
F- Measure: 0.82
```

Figure 17: Model Performance Metrics

The model displayed a testing accuracy of 82%, a precision accuracy of 82%, recall accuracy of 82% and F-1 score of 82%. The model passed the expected threshold of 80%

5.4 User Interface Design

5.4.1 Streamlit

Streamlit is a Python open-source library that makes it easy to create web-based user interfaces for machine learning and data science applications. The user interface of Streamlit is intended to be both simple and flexible, allowing developers to quickly prototype and deploy their applications with minimal effort. Streamlit includes a number of built-in widgets, such as sliders, buttons, and text inputs that can be used to create interactive user interface elements. Third-party widgets and visualizations can also be used by developers to enhance the functionality of their applications. The UI of Streamlit is designed to be reactive, which means that it updates in real-time as the user interacts with the application. This allows for a more fluid and intuitive user experience, as well as the creation of applications that respond to user input in real time.

Streamlit offers features for data visualization, data manipulation, and machine learning in addition to its user interface capabilities. As a result, it is a versatile tool for developing a wide range of applications, ranging from simple data visualizations to complex machine learning models.

Overall, the Streamlit user interface is a powerful and versatile tool for developing web-based applications, and its ease of use and real-time reactivity make it an excellent choice for rapid prototyping and deployment.

Gambling Disorder Predictor

Current Account Balance

100000.00 - +

Total funds deposited

30000.00 - +

Total funds withdrawn

0.00 - +

Hours Played

4.00 - +

Number of Deposits Made

1.00 - +

Predict

You were placed in Category 0: Your behaviour does not seem to be destructive.

Made with Streamlit

Figure 18: Streamlit User Interface Design

The user interface for Gamble Aware is quite simple and intuitive. Simplicity and ease of use was a key priority for a playground environment for crowdsourcing. The User interface consists of form inputs where a user keys in numerical data corresponding to their gambling site usage. The user is then obligated to click on the button labelled “predict” to get a category that corresponds to their site usage.

Predict

You were placed in Category 4: You have been placed on the high risk scale of developing GD. We suggest you stop gambling immediatly and seek professional help.

Figure 19: System Prediction Output

CHAPTER 6: IMPLEMENTATION AND TESTING

6.1 Introduction

The implementation and testing chapter outlines the development environment that the system was built in, the system components and testing plan before deploying the system.

6.2 Development Environment

A development environment is a collection of procedures and tools for developing, testing and debugging an application or program (Techopedia, 2016). The development environment for Gamble Aware was Microsoft's Visual Studio Code, a Visual Studio subsidiary tool. This environment appealed to me because of its debugging capabilities, source control capabilities, powerful extensions such as code time, which tells you how much time you spend coding, and superior code wrapping and formatting features.

I used pip package manager to keep track of dependencies. Pip enables a developer to create a list of dependencies for a project. When transferring the project to a new environment, the same list can be used by invoking the package manager to install everything on the list..

6.3 System Components

6.3.1 Data Collection and Preparation:

This involves collecting and cleaning the relevant data sets, performing feature engineering, and creating a clean and organized data pipeline for machine learning algorithms to train on.

6.3.2 Model Selection and Optimization:

This involves selecting the appropriate machine learning algorithm for the problem at hand and fine-tuning its hyper parameters for optimal performance. This is usually done through cross-validation and grid search.

6.3.3 Model Deployment:

Once a model has been trained and optimized, it needs to be deployed into a production environment where it can be accessed by end-users. Gamble Aware was deployed through a web app that users can use as a playground.

6.3.4 Monitoring and Maintenance:

Machine learning models need to be continuously monitored and maintained to ensure they are performing as expected. This involves tracking metrics such as accuracy and precision, and updating the model periodically with new data. The model's web app will collect metrics for performance analysis.

6.3.5 Visualization and Interpretability:

It's important to be able to understand and explain the results of machine learning models. This can involve visualizing the data and model outputs in a way that is easy to understand for non-technical stakeholders. The visualizations were made in python libraries and some are displayed in this documentation.

6.4 Test Data

Due to the complexity of testing machine learning models, a workflow had to be prepared for unexpected events while working with black-box model and shifting output/input relationship. Best practices observed were training for robustness, interpretability and reproducibility. Post-train tests were employed to check on the internal behaviour of the model. The tests done were invariance tests, directional expectation test and minimum functionality tests.

6.4.1 Invariance Test

The invariance test defines input changes that are expected to leave model outputs unaffected. The common method for testing invariance is related to data augmentation. You pair up modified and unmodified input examples and see how much this affects the model output. I employed mechanisms such as checking whether a change in a feature would greatly affect the models output.

6.4.2 Directional Expectation Test

Directional expectation tests can be used to define whether input distribution changes expected effects on the output. We made logical assumptions such as, the more hours a user spends on gambling, the more likely they are to be problem gamblers. Another assumption would be the more times a user makes deposits, they are more likely to be placed higher on the index.

6.4.3 Minimum Functionality Test

The minimum functionality test helps decide whether individual model components behave as you expect. The reasoning behind these tests is that overall, output-based performance can conceal critical upcoming issues in your model. Minimum functionality tests were performed on the model in three different ways.

- Samples were created that were easy for the model to predict to see whether the model consistently delivered expected results.
- Splitting the data sets into test data and training data. The segments were made smaller to see whether the model would produce expected results
- Testing for failures that were previously identified during manual error analysis.

6.5 Test Results

The model achieved an accuracy of 0.827 on the test set, indicating that it correctly classified 83% of the samples. The precision and recall scores were 0.83 and 0.82, respectively, indicating that the model correctly identified 83% of the positive samples and 82% of the negative samples.

The F1-score was 0.82, which indicates a good balance between precision and recall. Overall, the model performed well on the test set, indicating its ability to accurately predict the target variable.

CHAPTER 7: CONCLUSION

7.1 Achievements and Lessons learnt

Throughout the course of this project, I've had the opportunity to learn a wide range of skills and techniques that have been extremely beneficial to me throughout the course of this project. One of the most useful skills I've acquired is how to use Python for data science. This has included learning about various data structures, libraries, and data analysis and visualization techniques. The results of this project have demonstrated the effectiveness of machine learning in predicting the target variable and have highlighted the importance of feature selection and hyper parameter tuning for improving model performance. In addition, I learned a lot about the CRISP-DM methodology, which has helped me understand how to approach data-driven projects in a structured and organized way. Another important area of study has been project management, which has taught me the value of communication, goal-setting, and progress tracking. I've also learned about decision trees, which are an effective tool for making complex decisions with multiple variables. Finally, I've gained an understanding of medical problem gambling diagnosis, which is particularly relevant to the project's focus. Overall, I believe that this project has provided me with a wealth of valuable knowledge and experience that will serve me well in my future endeavours.

7.2 Conclusions

In conclusion, to address the growing prevalence of online gambling addiction, the development of an internet-based gambling disorder diagnosis tool is critical. With the rise of online gambling platforms, it is now easier than ever for people to engage in excessive gambling behaviours, which can have a negative impact on their personal and financial well-being. An internet-based diagnosis tool can provide a cost-effective and easily accessible method of identifying gambling disorder symptoms, allowing people to seek treatment before their condition worsens. Furthermore, such a tool could be used by healthcare professionals to screen their patients for gambling disorder, resulting in earlier diagnosis and intervention. As internet gambling grows in popularity, the need for effective and easily accessible diagnostic tools becomes more pressing.

7.3 Recommendations

We must take proactive measures to address the problem of problem gambling. To begin, public education campaigns should be launched to raise awareness about the dangers of excessive gambling and the resources available to those in need. This can include promoting responsible gambling practices and making support services such as counselling and self-exclusion programs available. Second, governments and gambling operators should collaborate to put stricter regulations and guidelines in place to prevent and reduce gambling harm. Measures such as mandatory pre-commitment limits, exclusion zones, and stricter advertising regulations may be included. Healthcare professionals should be educated on the signs and symptoms of gambling disorder, as well as have the resources necessary to provide effective treatment and support to their patients. We can help prevent and mitigate the harms associated with problem gambling by implementing these recommendations, as well as improve the well-being of individuals and communities affected by this issue. Finally, I would recommend further research in the field of online prevalence of PG, by developing tools that avail measures to users directly on their devices.

BIBLIOGRAPHY

- Deans, Emily & Thomas, Samantha & Daube, Mike & Derevensky, Jeffrey & Gordon, Ross. (2016). Creating symbolic cultures of consumption: An analysis of the content of sports wagering advertisements in Australia. *BMC Public Health*. 16. 10.1186/s12889-016-2849-8.
- Hawker, C.O.; Merkouris, S.S.; Youssef, G.J.; Dowling, N.A. A smartphone-delivered ecological momentary intervention for problem gambling (GamblingLess: Curb Your Urge): Single-arm acceptability and feasibility trial. *J. Med. Internet Res.* 2021, 23, e25786. [CrossRef]
- Chóliz, M., Marcos, M. & Bueno, F. Ludens: A Gambling Addiction Prevention Program Based on the Principles of Ethical Gambling. *J Gambl Stud* (2021). <https://doi.org/10.1007/s10899-021-10066-7>
- Sagoe, Dominic & Griffiths, Mark & Erevik, Eilin & Høyland, Turid & Leino, Tony & Lande, Ida & Sigurdsson, Mie & Pallesen, Ståle. (2021). Internet-based treatment of gambling problems: A systematic review and meta-analysis of randomized controlled trials. *Journal of Behavioral Addictions*. 10. 10.1556/2006.2021.00062.
- Bonnaire, C. (2011). Internet gambling: what are the risks?. *L'encephale*, 38(1), 42-49.
- Gainsbury, S., & Wood, R. (2011). Internet gambling policy in critical comparative perspective: The effectiveness of existing regulatory frameworks. *International Gambling Studies*, 11(3), 309-323.
- Bell, R. & Boldero, J. (2011), Factors affecting youth gambling: A comprehensive model of the antecedents and consequences of gambling in young people. Department of Justice, Victoria. Retrieved from, http://responsiblegambling.vic.gov.au/sites/default/files/Factors_affecting_youth_gambling.pdf. accessed 14 November, 2013)
- Omondi, T. (2018, April 08). Question and Answer session with Chairman of BCLB. Daily Nation.
- Ssewanyana, D., & Bitanirwe, B. (2018). Problem gambling among young people in sub-Saharan Africa. *Frontiers in public health*, 6, 23.

- Kiragu, M. N. (2016). Market Penetration Strategies by Football Betting Firms in Kenya. Nairobi, Kenya: University of Nairobi; Unpublished Thesis.
- Koross, R. (2016, November). University Students Gambling: Examining the Effects of Betting on Kenyan University Students' Behavior. *International Journal of Liberal Arts and Social Science*, 57-66.
- Ladouceur, Robert, Caroline Sylvain, Hélène Letarte, Isabelle Giroux, and Christian Jacques. "Cognitive treatment of pathological gamblers." *Behaviour research and therapy* 36, no. 12 (1998): 1111-1119.
- Gainsbury, S. M. (2015). Online gambling addiction: the relationship between internet gambling and disordered gambling. *Current addiction reports*, 2(2), 185-193.
- Chóliz, M., & Saiz-Ruiz, J. (2016). Regular el juego para prevenir la adicción: hoy más necesario que nunca. *Adicciones*, 28(3), 174-181.
- Hawker, C. O., Merkouris, S. S., Youssef, G. J., & Dowling, N. A. (2021). A Smartphone-Delivered Ecological Momentary Intervention for Problem Gambling (GamblingLess: Curb Your Urge): Single-Arm Acceptability and Feasibility Trial. *Journal of medical Internet research*, 23(3), e25786. <https://doi.org/10.2196/25786>
- Nilsson, A., Magnusson, K., Carlbring, P. et al. The Development of an Internet-Based Treatment for Problem Gamblers and Concerned Significant Others: A Pilot Randomized Controlled Trial. *J Gambl Stud* 34, 539–559 (2018). <https://doi.org/10.1007/s10899-017-9704-4>
- Petry, N. M., Ammerman, Y., Bohl, J., Doersch, A., Gay, H., Kadden, R., ... & Steinberg, K. (2006). Cognitive-behavioral therapy for pathological gamblers. *Journal of consulting and clinical psychology*, 74(3), 555.
- Techopedia. (2016, November 11). Development Environment. Retrieved from Techopedia.Com: <https://www.techopedia.com/definition/16376/developmentenvironment>

APPENDIX

APPENDIX1: Project Gantt chart

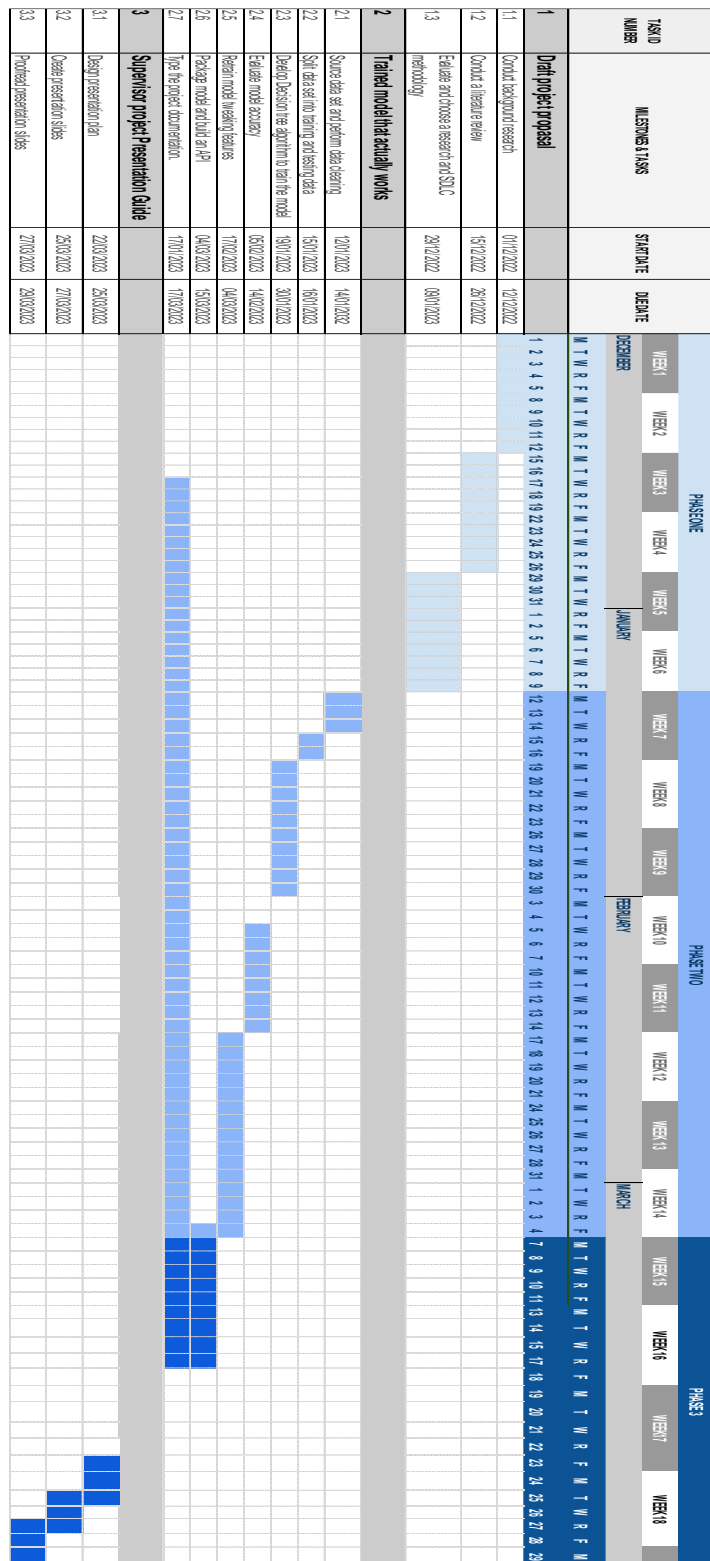


Figure 21: Project Gantt chart