A Machine Learning Model to Predict Substance Abuse Relapse in Nairobi County, Kenya

By

Emmy Jepchirchir

171221

Submitted in the Partial Fulfilment of the Requirements for the Degree of Master of Science in Information Technology at Strathmore University

School of Computing and Engineering Sciences

Strathmore University

Nairobi, Kenya.

June 2025

Declaration and Approval

Declaration

I confirm that this research is entirely my own work and has not been previously submitted to any institution for academic certification. Any content referenced from existing published sources has been properly cited, with due recognition given to the respective authors.

Name: Emmy Jepchirchir

Signature ……Date …29th May 2025……

Approval

The thesis of Name of Emmy Jepchirchir was reviewed and approved for examination by the following:

Dr Esther Khakata

School of Computing and Engineering Sciences,

Strathmore University.

Dr. Julius Butime,

Dean, School of Computing & Engineering Sciences,

Strathmore University

Prof. Bernard Shibwabo,

Director of Graduate Studies,

Strathmore University

Abstract

Substance use disorder (SUD) remains a critical public health issue in Nairobi County, Kenya, where high relapse rates undermine efforts to achieve long-term recovery for individuals affected by addiction. Despite the availability of treatment programs and rehabilitation centers, relapse rates continue to exceed 50% within the first year of recovery, indicating the limitations of current intervention strategies. The complexity of relapse, driven by psychological, social, and biological factors, requires a more sophisticated approach to prediction and prevention. This research explored how machine learning could be applied as a predictive tool to identify individuals at risk of relapses. Specifically, the study aimed to enhance intervention approaches and the recovery results in Nairobi County. The data used in the study was collected from a secondary data source (Kaggle) with features that include several variables, such as demographic information, substance education, behavioral factors, psychological evaluations, and socioeconomic indicators. The data collected was then used to train and validate the machine learning model, which utilized predictive algorithms such as decision trees, support vector machines, random forest, Artificial Neural Network and the K-nearest neighbors to determine the probability of relapse. Eventually, the study evaluated the model's performance by utilizing the standard performance metrics. By developing this predictive model, the study anticipated handling numerous major challenges in the management of substance abuse disorder in Nairobi County. First, by identifying individuals at high risk of relapse, the model will enable healthcare providers to implement targeted interventions, thereby reducing relapse rates and improving the chances of sustained recovery. Second, the model's predictions can inform resource allocation, ensuring limited healthcare resources are directed toward those most in need. Third, the research will contribute to public health policy by providing data-driven insights that can be used to design more effective prevention and treatment programs. The developed models in this research performed well with the adopted and implemented model, Artificial Neural Network achieving an accuracy of ~0.9774.

Keywords: Substance Use Disorder (SUD), Machine Learning Model, Relapse.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI	-	Artificial Intelligence
ANN	-	Artificial Neural Networks
CBT	-	Cognitive Behavioral Therapy
KNN	-	K-Nearest Neighbors
LoR	-	Logistic Regression
MCBT	-	Mindfulness Cognitive Behavioral Therapy
ML	-	Machine Learning
MLE	-	Maximum Likelihood Estimation
NB	-	Naïve Bayes
ROC	-	Receiver Operating Characteristics
RF	-	Random Forest
RMSE	-	Root Mean Squared Error
SUD	-	Substance Use Disorder
SVM	-	Support Vector Machines
WHO	-	World Health Organization

Definition of Terms

Machine Learning	A branch of artificial intelligence (AI) that focuses on the development of algorithms that enable computers to learn from and make predictions based on data (Sarker, 2021).
Maximum Likelihood Estimation	A statistical method used to estimate the parameters of a probability distribution based on observed data. The core idea is to find the parameter values that maximize the likelihood function, which represents the probability of observing the given data under a specific statistical model (Sarker, 2021).
Mindfulness Cognitive Behavioral Therapy	A therapeutic approach that combines principles from cognitive behavioral therapy (CBT) with mindfulness practices (Anundo et al, 2022).
Relapse	Refers to the return of a condition or behavior after a period of improvement, commonly used in medical and psychological contexts. In substance use, it signifies a significant setback in the recovery process, where an individual who has previously achieved abstinence begins to engage in substance use again (American Editorial, 2024).
Substance Use Disorder	Refers to the harmful or hazardous use of psychoactive substances, including alcohol and illicit drugs (World Health Organization,2024).

Acknowledgements

The progress made thus far of this thesis marks a significant milestone in my academic journey, and I am deeply grateful to all those who contributed to its success.

First and foremost, I would like to express my sincere appreciation to my principal supervisor, Dr Esther Khakata and lead supervisor for MSc IT 2024-2025 Prof. Ismael Ateya, for their invaluable support, expert guidance, and constant encouragement throughout this research. Their insightful feedback, patience, and commitment to excellence played a crucial role in shaping this study. I am truly grateful for the time and effort they dedicated to mentoring me.

I extend my gratitude to Strathmore University, School of Computing and Engineering Sciences whose resources and academic environment provided a strong foundation for this research. I also acknowledge the contributions of my professors and academic mentors, whose teachings and guidance have significantly influenced my approach to research and critical thinking.

I am profoundly thankful to my family for their unwavering love, patience, and encouragement. Their constant support and belief in my abilities have been a source of strength, especially during challenging times. I also extend my appreciation to my friends and colleagues for their moral support, intellectual discussions, and words of motivation that kept me going.

Lastly, to everyone who has played a role, whether big or small, in the working of this thesis, thank you from the bottom of my heart.

Chapter 1: Introduction

1.1 Background of the Study

Substance use disorder (SUD) presents a notable global public health challenge (Kabisa et al., 2021), with relapse being a critical obstacle to successful long-term recovery. In Nairobi County, Kenya, the problem of substance abuse is compounded by socioeconomic inequalities, limited access to mental health services, and a rapidly urbanizing environment that exacerbates stressors contributing to substance use. While various intervention programs and rehabilitation centers are in place, high relapse rates indicate a need for more effective strategies.

Nairobi County, as the capital of Kenya, faces unique challenges in managing substance use disorder. The county's diverse population, rapid urbanization, and socioeconomic disparities contribute to the prevalence of substance abuse, with a population of at least 19.1% of individuals in the county being reported as abusing at least one Substance or drug (NACADA, 2022). Alcohol, tobacco, khat, cannabis, heroin, and prescription drugs are among the most commonly abused substances in the region (NACADA, 2022). According to recent studies, the socioeconomic environment in Nairobi, characterized by unemployment, poverty, and inadequate mental health services, have created conditions conducive to both the initiation and persistence of substance abuse. Compounding this issue is the stigma linked to substance dependency, which is cited to be the main demotivating factor when it comes to people looking for support and interfering with effective intervention (NACADA, 2022).

The healthcare system in Nairobi, while relatively more developed than in other parts of Kenya, is often overstretched. Rehabilitation centers are limited in number and capacity, and there is a significant stigma associated with seeking help with substance use disorders. This stigma, combined with socioeconomic challenges, often leads to individuals delaying or avoiding treatment altogether. Moreover, even when treatment is sought, the high relapse rate, which is estimated to be over 50% within the first year of recovery, points to the inadequacy of current intervention methods (Mwove, 2019).

Relapse is a complex phenomenon influenced by several factors, including psychological, socioeconomic, and biological elements (Mukora & Mbugua, 2023). The process of relapse often follows a predictable pattern where there is an initial trigger, followed by a period of intense

craving and eventually a return to substance use. However, the precise contributors to relapse can vary significantly from one individual to another. These contributors can be stress, being exposed to substance-related cues, social pressures, underlying mental health complications in the form of anxiety or depression, and even genetic predispositions.

In Nairobi County, additional challenges such as poverty, unemployment, and the lack of social support structures further exacerbate the risk of relapse. The traditional approaches to relapse prevention, which often rely on counselling and behavioral therapies, may not be sufficient to address the complex and various nature of relapse. There is a clear need for more personalized and proactive approaches to predicting and preventing relapse, and therefore, a gap that Machine learning can play an important role in filling.

Machine learning, a branch of artificial intelligence, has shown great promise in predicting outcomes in various fields, such as healthcare (Nyabuti & Mohammed, 2024). Unlike conventional statistical approaches that depend on predefined models and assumptions, machine learning algorithms can access massive and complicated information and identify patterns and links that cannot be immediately apparent to human analysts. This capability makes machine learning particularly well-suited for predicting relapse, as it can consider a wide range of variables simultaneously and adapt to the precise nature of the dataset.

Several machine learning techniques have been applied to determine patient outcomes in healthcare settings. These algorithms have been used to predict everything from disease onset to patient readmission rates, often with high accuracy. In the case of substance use disorders, machine learning models can be programmed on data that includes demographic information, treatment history, psychological assessments, and socioeconomic indicators to predict the likelihood of relapse.

1.2 Statement of the Problem

Despite existing treatment programs and rehabilitation efforts, relapse remains a significant hurdle in the recovery process, with many individuals returning to Substance even after refraining for a while (Kuyeya,2021). The unpredictability of relapse poses a critical challenge for healthcare providers, who often lack the tools to identify individuals at high risk and intervene effectively.

Current intervention strategies in Nairobi County are predominantly reactive rather than proactive, with most efforts focusing on treating relapse after it occurs rather than preventing it. This approach not only strains the already limited healthcare resources but also reduces the chances of long-term recovery for individuals struggling with addiction. The complexity of relapse, influenced by psychological, social, and biological determinants, further complicates the development of effective prevention strategies.

Given the high stakes associated with relapse, there is an urgent need for innovative approaches to predict and prevent it. Machine learning, with its capacity to assess large and complex datasets, offers a promising solution (Nyabuti & Mohammed, 2024). However, there is a lack of research on applying machine learning techniques to predict substance relapse in the case of Nairobi County. Existing models from other regions may not be directly applicable due to the unique socioeconomic and cultural factors determining substance use in this area (Nairobi County, Kenya).

This research will fill this gap by developing a machine learning model that can accurately identify individuals at high risk of relapse and provide healthcare providers with a valuable tool to enhance intervention strategies. The goal is to reduce relapse rates, improve recovery outcomes, and contribute to the broader effort of combating substance use disorder in Nairobi County.

1.3 Main Objective of the Study

This research's primary objective was to develop a robust and accurate machine learning model that can predict substance abuse relapse in Nairobi County, Kenya, tailored specifically for rehabilitation centres in Kenya.

1.4 Specific Objectives

i. To identify the factors influencing relapse on the use of substances.
ii. To review the existing approaches in the prediction of substance abuse relapse in Nairobi County.
iii. To review existing machine learning models and algorithms that have been used in the prediction of relapse in substance abuse.
iv. To develop a machine learning model that will assist with the prediction of relapse into substance abuse.
v. To validate the developed model.

1.5 Research Questions.

i. What are the determinants of relapse on the use of a substance?
ii. What are the existing approaches in the prediction of substance abuse relapse in Nairobi County?
iii. What are existing machine learning models and algorithms that have been used in the prediction of relapse in substance abuse?
iv. How can a machine learning model that will assist with predicting relapse into substance abuse be developed?
v. How can the performance of the developed model be validated?

1.6 Study Justification

In the case of substance use disorders, machine learning can offer a data-driven approach to identify individuals at high risk of relapse, enabling healthcare providers to intervene proactively. Nonetheless, a considerable literature gap focuses on applying machine learning techniques to predict substance relapse in Nairobi County. Most existing models are developed in different socio-cultural contexts, which may not directly apply to Nairobi's unique challenges.

This research was justified by its potential to address a pressing public health issue in Nairobi County, provide a locally relevant solution, and contribute to the broader substance use disorder treatment and prevention field. By leveraging machine learning, the study aims to offer a novel approach to predicting and preventing relapse, ultimately improving the lives of those affected by substance use disorder in Nairobi County.

The scalability and cost-effectiveness of machine learning further justify its application in Nairobi County. The healthcare system faces constraints in resources and personnel, limiting its capacity to monitor patients after treatment. Once trained, a machine learning model will provide consistent and scalable relapse risk assessments, allowing healthcare providers to focus on delivering targeted interventions. Additionally, the insights generated from the model will guide policymakers in allocating resources effectively, identifying high-risk groups, and developing data-driven public health strategies.

Adopting this technology aligns Nairobi County with global trends in leveraging artificial intelligence to solve complex public health issues. Machine learning enables the identification of people at risk on time and enhances the understanding of relapse patterns across the population.

By integrating such a model, Nairobi County can significantly reduce relapse rates, improve treatment goals and promote sustainable recovery for those battling substance abuse. This initiative represents a transformative step toward a more effective, efficient, and personalized approach to addressing substance abuse challenges in the region.

1.7 Study Scope and Limitations

1.7.1 Scope

The primary focus of this study was to develop a machine learning model capable of predicting substance relapse among individuals undergoing treatment for substance use disorder in Nairobi County, Kenya. This model would have the potential to provide stakeholders with a data-driven approach to predicting and preventing relapse, ultimately improving recovery outcomes and public health in the region.

1.7.2 Limitations

While this research has the potential to provide valuable insights and tools for predicting and preventing substance relapse in Nairobi County, it is vital to recognize and address the inherent limitations. Data availability and quality, complexity of relapse, generalizability and ethical considerations are some of the limitations that underscore the need for ongoing research and collaboration between researchers, healthcare providers, and policymakers to fully realize the benefits of machine learning.

Data availability and quality remain among the main challenges when it comes to research or studies in healthcare. Access to reliable and comprehensive data is essential for building effective models. In many regions, including Nairobi County or Kenya in general, data on substance use and relapse may be sparse, incomplete, or biased, making it difficult to train reliable models.

The complexity of relapse in substance use disorders poses another challenge for ML applications. Substance relapse is determined by a variety of factors that can range from psychological, social, environmental, and biological. Capturing the full complexity of these factors in a machine-learning model is challenging. Generalizability is another challenge where studies focus on specific demographics or settings and may not reflect the diversity of SUD patients across different regions or backgrounds. Models trained on a specific population, such as Nairobi County,

may not generalize to other regions or populations due to cultural, socioeconomic, and healthcare system differences.

Using personal data in machine learning models, especially for predicting sensitive outcomes like relapse, raises privacy concerns. Adhering to ethical considerations by ensuring informed consent, data anonymization, and equitable use of AI tools is critical to prevent harm or bias against vulnerable populations.

Chapter 2: Literature Review

2.1 Introduction

Use of machine learning (ML) models increases for predicting multiple severe consequences including substance use and relapse decisions in the area of healthcare. Substance use forecasting entails the ability to extrapolate alterations from high volumes of data merged with multiple parameters such as ancestry, psychological histories, effaced treatment, and sociological prescribes. Historical standard predicting techniques have major problems when it comes to nonlinear manner of these factors. However, with the help of machine learning algorithms, such problems can be solved with the assistance of artificial intelligence, which, passing through tens of thousands of web pages, selects the most important patterns that a person would not be able to determine. Din research by Al-Sharan et al. (2022), machine learning models remains promising for identifying relapse with decent accuracy, allowing healthcare providers the best approaches for suitable intercession and care. First, the use of DL makes prediction possible in using other data source since it is more flexible than DL.

Different techniques of machine learning have been discussed in this regard including support vector machine, decision tree, neural network, and random forest of akin structure that allow for quality sorting of a complicated data set. As observed by Hosseinzadeh et al. (2021), these models are promising in comparison with traditional statistical methods in terms of laying out various trends and interactions that contribute to the use of substances and show better prediction capability. Machine learning types also have shown high potential because they are able to integrate unstructured data from EHRs or social media. Despite these tremendous opportunities, the issue of data privacy, model explanation, and algorithm value fairness, among others, remain major challenges to be solved (Zhang & Sun, 2023).

2.2 Determinants Substance Abuse Relapse

2.2.1 Individual-Specific Determinants

Individual factors significantly influence the likelihood of substance abuse relapse, as they encompass a range of personal characteristics and circumstances that can either promote recovery or lead to a return to substance use. The most critical individual factors are personal willingness or desire to abstain from substances. Research by Afkar et al. (2017) indicated that a strong commitment to recovery correlates with lower relapse rates; for instance, personal willingness has

been associated with an odds ratio of 8.186, highlighting its importance in the recovery process (Amir et al., 2021). Besides, when there are co-occurring mental health complications, there can be challenges in achieving recovery, as individuals may turn to substances to cope with underlying psychological issues. This interplay between mental health and substance use underscores the importance of integrated treatment strategies that handle the two issues at the same time.

Another essential individual factor is the availability of substances. The ease with which an individual can access drugs significantly impacts their recovery journey (Malik et al., 2023). According to Amir and others (2021), people who have easier access to substances are also the ones who are most likely to relapse, with an odds ratio of 3.392, indicating a strong correlation between drug availability and relapse risk (Amir et al., 2021). Furthermore, personal circumstances such as relationship status also play a role; for instance, being single has been linked to higher relapse rates compared to being married, suggesting that social support systems can be protective against relapse (Amir et al., 2021). The emotional state of an individual, including feelings of loneliness or rejection from peers and family, can further exacerbate the risk of returning to substance use.

The role played by social influences, especially from family and friends, cannot be overlooked when shaping an individual's recovery trajectory. The presence of addicted friends or family members increases the likelihood of relapsing, as these relationships can create environments where substance use is normalized or encouraged (Sohrabpour et al., 2024; Kabisa et al., 2021). Additionally, experiences of rejection or loss, such as the death of a significant other or divorce, can lead individuals back to substance use as a means of coping with emotional pain (Afkar et al., 2017). Collectively, these individual factors illustrate a nuanced relationship between personal choices, social relationships and environmental influences in determining the risk of substance abuse relapse.

2.2.2 Familial/Genetic Determinants

Familial factors contribute significantly to the risk of having substance abuse disorders (SUDs). A significant aspect of this influence is the genetic predisposition to addiction, which can account for approximately 50% of an individual's susceptibility to substance use disorders (Grant & Chamberlain, 2020). Genetic factors can increase the likelihood of substance use by affecting neurotransmitter systems, particularly those involving dopamine, associated with the brain's

reward pathways. People who come from families with a record of addiction are predisposed to inherit such genetic traits, which makes them vulnerable to developing similar issues themselves. A study by Mousali et al. (2021) indicates that the presence of a first-degree family member with a substance use disorder significantly elevates the risk, with some research suggesting an eight-fold increase in risk among relatives of individuals with drug disorders across various substances, including opioids and alcohol (NACADA, 2024).

In addition to genetic factors, environmental influences within the family context are critical in shaping an individual's relationship with substances. Children raised in environments where drug or alcohol use is normalized are more likely to adopt similar behaviors. Factors such as parental substance abuse, ineffective parenting styles, and exposure to trauma or domestic violence can create a chaotic environment that increases the risk of developing SUDs (Mwove, 2019). These negative experiences can hinder healthy emotional development and coping mechanisms, leading individuals to seek substances as a form of escape or relief from stressors (NACADA, 2024). Furthermore, familial relationships characterized by conflict or having no support can perpetuate a state of loneliness and isolation and increase susceptibility to substance use as a coping strategy.

Communication about addiction within families has also been noted to be essential for prevention and awareness. Families that openly discuss their history of addiction can help younger members understand their risks and encourage healthy coping strategies (DrugFree, 2024). For instance, educating children about the implications of their family history can empower them to make informed choices regarding substance use. Additionally, fostering strong familial bonds and supportive relationships can serve as protective factors against addiction. By emphasizing resilience and providing resources for mental health support, families can address the perils that come with their genetic and environmental backgrounds (NACADA, 2024; Price, 2024).

2.2.3 Social and Environmental Determinants

Social and environmental determinants greatly influence the likelihood of substance abuse relapse, influencing individuals both directly and indirectly (Price, 2024). One of the most critical social factors is peer pressure (Wainaina, 2020). Individuals who associate with friends or family members who use drugs or alcohol are at a heightened risk of relapse. Research indicates that being surrounded by peers who engage in substance use can evoke strong urges to use again, with an

odds ratio of 2.4 for those living with peers compared to those in more supportive environments (Kabisa et al., 2021). This social dynamic creates an environment where substance use is normalized, therefore hindering individuals from sustaining their sobriety.

Environmental cues contribute significantly to triggering relapse. Exposure to specific places smells, or objects linked to substance abuse from the past can lead to intense cravings. For instance, encountering a familiar drug dealer or visiting locations where one previously used Substance can evoke memories and feelings that trigger a relapse. Studies have shown that these environmental triggers are significant contributors to relapse, emphasizing the importance of avoiding high-risk situations following treatment (Malik et al., 2023). Moreover, the overall availability of drugs in one's environment further exacerbates this risk; individuals living in areas with high drug accessibility are more likely to relapse due to the constant temptation and ease of obtaining substances (Kuyeya, 2021).

Additionally, lack of social or community support can significantly hinder recovery efforts (NACADA, 2022). Individuals who do not have a robust support system, whether from family, friends, or community resources, may struggle more with coping mechanisms when faced with stressors or triggers (Sohrabpour et al., 2024). A limited support network can lead to feelings of isolation and helplessness, increasing the likelihood of returning to Substance use as a coping strategy. Furthermore, interpersonal conflicts with family or friends can create negative emotions that contribute to relapse; studies indicate that conflicts are involved in over 50% of all relapses (Afkar et al., 2017). Addressing these social and environmental determinants is important when developing functional prevention approaches and supporting individuals in their recovery journeys.

2.2.4 Psychological and Behavioral Determinants

Psychological and behavioral factors are critical in understanding the dynamics of substance abuse relapse. One of the most significant psychological factors is stress. High levels of stress, particularly from work, relationships, or financial pressures, can lead individuals to seek relief through substance use (American Editorial, 2024). When coping skills are inadequate, people can resort to abusing drugs and alcohol to manage their stress, which can create a cycle of dependency and increase the risk of relapse. Negative emotions such as anxiety, depression, and anger also contribute to this risk; studies indicate that interpersonal conflicts and unresolved

emotional issues can lead to cravings for substances as a form of escape from these feelings (Wainaina, 2020). Interpersonal problems are involved in more than 50% of all relapses, highlighting the need to address emotional health in recovery programs (Sohrabpour, 2024).

Self-efficacy is another crucial psychological factor, which refers to an individual's belief in their capacity to refrain from abusing drugs in challenging situations. Higher self-efficacy is associated with lower relapse rates, as individuals who believe they can cope with triggers and stressors without resorting to substances are more likely to succeed in their recovery efforts (Malik et al., 2023). Conversely, low self-efficacy can lead to feelings of helplessness and increased vulnerability to relapse when faced with high-risk situations. Additionally, personality traits such as neuroticism are characterized by emotional instability and low conscientiousness have been linked to higher relapse rates. Individuals with these traits may struggle more with managing their emotions and maintaining consistent recovery efforts, making them more susceptible to returning to substance use (Malik et al., 2023; Sohrabpour, 2024).

Behavioral factors also contribute significantly to relapse dynamics. A typical scenario is when intense cravings are triggered as a result of being exposed to people, situations or places that are linked with past substance use. For example, encountering friends who use drugs or visiting locations where one previously used Substance can trigger memories and desires that are difficult to resist (American Editorial, 2024). Furthermore, peer pressure remains a powerful influence; being around individuals who engage in substance use can create an environment where relapse becomes more likely. The presence of addicted friends or family members has been identified as a strong predictor of relapse risk (Kabisa et al., 2021).

2.2.5 Economic Determinants

Unemployment is one of the critical economic factors associated with increased relapse rates. Unemployed individuals often experience heightened stress and financial instability, which can lead to a return to Substance use as a coping mechanism. Research indicates that unemployment not only exacerbates feelings of hopelessness but also limits access to treatment resources, thereby increasing the risk of relapse (Afkar et al., 2017). According to Afkar and others (2017), a large portion of individuals who relapse cite job loss or employment instability as contributing factors to their return to substance use.

Income level also plays a major role in recovery outcomes. Higher income is often linked to better access to healthcare services, including addiction treatment programs and mental health support (Kabisa et al., 2021). On the other hand, lower income levels can lead to increased stressors, such as housing instability and food insecurity, which may drive individuals back to substance use. Research has demonstrated that as family income increases, the significance of various relapse-related factors decreases, suggesting that economic stability can avoid some of the risks linked to addiction relapse. This relationship underscores the importance of addressing economic barriers in recovery programs to enhance long-term success (Sohrabpour, 2024).

Economic problems are often intertwined with other social issues, such as marital discord and lack of familial supervision. These interconnected factors can create a challenging environment for individuals in recovery (Gant & Chamberlain, 2020). For example, financial stress may lead to conflicts within families or relationships, further complicating an individual's ability to maintain sobriety. Studies indicate that economic hardship significantly correlates with higher relapse rates, sensitizing on the importance of all-rounded strategies that address economic stability and emotional support in recovery strategies (Afkar et al., 2021).

2.3 Existing Approaches in Prediction of Substance Abuse Relapse in Nairobi County.

2.3.1 Mindfulness Cognitive Behavioral Therapy (MCBT)

A comparative study by Anundo, Muaka & Ongaro (2022) evaluated the effectiveness of MCBT against the traditional 12-Steps model for relapse prevention in rehabilitation centres across Nairobi. The findings indicated that MCBT may offer better outcomes in preventing relapse by equipping individuals with coping mechanisms to handle triggers associated with substance use. This approach highlights integrating mindfulness practices into cognitive behavioral strategies to enhance treatment efficacy.

MCBT focuses on enhancing self-awareness and promoting acceptance of thoughts and feelings without judgment, which can significantly help individuals manage cravings and triggers associated with substance use.

2.3.2 Biopsychosocial Determinants

According to the research conducted by Mwove (2019), the study identified that biopsychosocial determinants have contributed to the substance abuse relapse prediction in Nairobi County. By understanding and addressing the biopsychosocial determinants like accessibility to

drugs or substances, age, family conflicts or financial instability through community interventions and policy changes, relapse rates can be reduced.

The biopsychosocial model also emphasizes that the interconnectedness of biological, psychological, and social factors can help predict relapse. This model provides a holistic approach that considers the underlying psychological issues and social circumstances surrounding the individual (Mwove, 2019).

2.3.3 Community-Based Interventions

Community-based detoxification and rehabilitation programs have been implemented to address alcohol dependence in many regions, not only in Nairobi County. These programs involve outpatient detoxification and regular follow-ups by health workers, focusing on individualized care based on factors like intensity of alcohol intake and craving levels. This method provides ongoing support and monitoring to reduce relapse rates (Kuria,2013).

By focusing on this approach, the intervention aims to create supportive environments that facilitate recovery and reduce the stigma associated with substance use disorders. This can involve establishing self-help groups and peer support networks that empower individuals in recovery to share experiences and strategies for maintaining sobriety.

2.3.4 Holistic Approaches

Holistic approaches to addiction treatment focus on addressing the whole person rather than just the symptoms of substance use disorders (Kuria, 2013). The approach employs a comprehensive strategy that recognizes that addiction Influences multiple dimensions of a person's life, including emotional, mental, physical, and spiritual well-being. The treatment creates a balanced recovery experience by integrating these various therapeutic ways.

Additionally, the holistic treatment or approach not only focuses on evidence-based practices but also implements traditional medical interventions. While CBT may address cognitive distortions related to substance use, Anundo, Muaka & Ongaro (2022) elaborates that self-awareness and emotional regulation skills can be achieved by having individuals engage in mindfulness and meditation.

2.4 Empirical Review

According to the results of the study by Katardjiev et al. (2019), the random forest algorithm showed the best performance, dropping only a relatively small amount in predictive power when moving onto the test set, being able to explain 84.1% of the variance in the data, and with the lowest RMSE recorded on the test set. However, this study failed to attain an acceptable generalizability given the methodological approaches applied were based on one company's construct for assessing sobriety.

Lai et al. (2021) developed a machine learning framework for forecasting smoking cessation results to develop a forecasting framework that uses machine learning algorithms to predict the outcome of smoking cessation. The model developed was easily applicable, and it showed potential in achieving customized and precision medicine for addressing smoking cessation.

The study by Lai et al. (2021) utilized the ANN framework, which showed improved functionality with a specificity of 0.567, sensitivity of 0.704, an ROC value of 0.660, and accuracy of 0.640 for prediction in smoking cessation results. However, the study's major limitation was some bias and misclassification with the patients who self-reported on some factors which influenced the adoption of the combined outcome.

In 2023, Xin et al. (2023) assessed the prediction of relapse risk among Chinese drug abusers based on interpretable machine learning technology. The research presented an innovative approach to predicting relapse risk among Chinese drug abusers using interpretable machine learning techniques (Extreme gradient boost, Logistic regression, random forest, and Classification and regression tree). The developed model could integrate diverse data sources (medical records, social factors, genetic information) to improve prediction accuracy. However, the model lacked generalizability as it was modelled only with one specific population dataset (Chinese drug abusers), and it may not generalize well to other groups without extensive retraining and validation.

2.5 Theoretical Review

In 2019, Katardjiev et al. (2019) conducted research on a Machine Learning Based Approach to predicting Alcoholic Relapses. The assessment explored the utilization of various machine learning algorithms to forecast relapses in patients undergoing alcohol addiction

treatment. Utilizing clinical trial data from Contigo Care, the authors model four algorithms: support vector machines, random forests, decision trees, and K-nearest neighbors. The study reveals that the random forest model performs better than baseline predictors, suggesting its potential for effectively forecasting relapse trends. The research highlights a notable gap in the application of machine learning within addiction care, proposing that future work could enhance prediction accuracy through larger datasets and advanced deep learning techniques. This proof of concept demonstrates that relapse patterns can be predicted several days in advance, facilitating preventive measures in treatment plans. The research, however, lacked a certain level of generalizability, as the methodology used was built on a single company's own construct for measuring sobriety.

Islam et al. (2021) designed a machine-learning algorithm to predict dependence on drugs and alcohol with a case study of the Bangladeshi population. The study aimed to develop a predictive framework that identifies individuals at risk of addiction based on various socio-demographic and psychological factors. The study results indicated that logistic regression performed better than other models, attaining the highest accuracy. It also demonstrated the feasibility of using machine learning for addiction prediction and emphasized the importance of understanding the various natures of addiction. The research, however, had some limitations as it lacked generalization due to a limited data sample size whose characteristics could not reflect or represent diverse demographics. The study also depended on self-reporting data that could lead to inaccuracies as the individuals may underreport or misrepresent their drug use and related factors.

According to a study by Ebrahimi et al. (2021) on the prediction of the risk of alcohol addiction using machine learning, the research identified multiple factors influencing the risk of developing alcohol use disorder. Genetic predisposition, particularly a family history of alcoholism, significantly heightens the risk, while psychological factors such as stress, anxiety, and personality disorders further complicate the landscape. Behavioral patterns, including binge drinking and substance abuse, alongside social influences like peer pressure, contribute to the development of alcohol use disorder. However, the research highlighted in its review that many studies utilized small or single-site datasets, showing the study results' limited generalizability and applicability to broader populations.

2.6 Machine Learning Algorithms

Machine Learning Algorithms in substance abuse contexts are increasingly being utilized to enhance prediction, prevention, and treatment strategies. Their application spans various aspects, including identifying risk factors, predicting treatment outcomes, and analyzing patterns of substance use. Some main algorithms used and adopted include random forest, logistic regression, CART, KNN decision tree, and ANN, among others.

2.6.1 Random Forest

Random Forest is a robust supervised machine learning Framework that performs well in regression and classification tasks by constructing many decision trees during training and aggregating their predictions, as demonstrated in Figure 2.1 (Bharathidason & Sujdha, 2023). Utilizing ensemble learning, it employs bootstrapping to create random subsets of the training data and randomly selects features for each split in the trees, which enhances diversity and reduces overfitting. This method leads to high accuracy and resilience against noise, making Random Forest particularly effective for complex datasets (Sarker, 2021).

A study by Katardjiev et al. (2019) adopted this algorithm due to its robust outfitting ability, as it can handle complex datasets, and the model achieved 78.1% accuracy compared to other algorithms. Despite this advantage, random forests are prone to some demerits, such as computational complexity, as the trees keep increasing, which may make the model complex and lack interpretability.

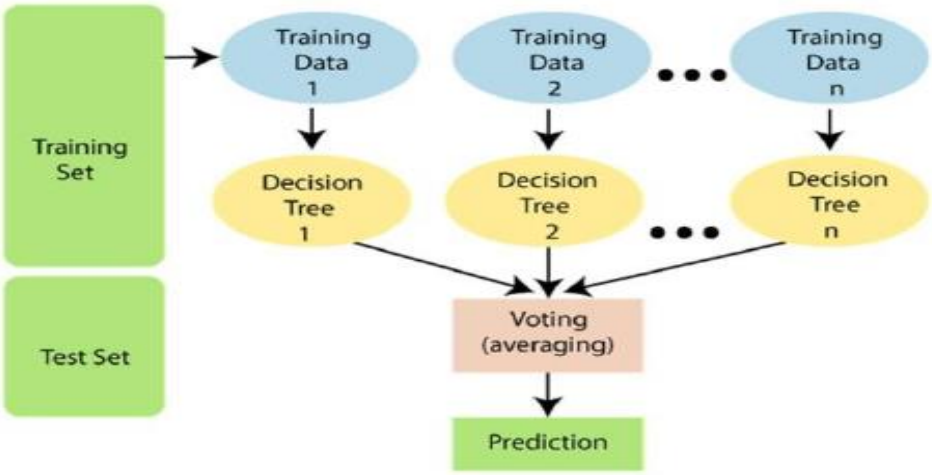


Figure 2.1: Showing Random Forest Algorithm (Bharathidason & Sujdha, 2023)

2.6.2 Logistic Regression

Logistic regression is an algorithm mainly utilized for binary classification tasks, aiming to forecast the likelihood that an instance belongs to one of two classes. The algorithm estimates parameters using maximum likelihood estimation (MLE), making it a simple yet powerful baseline model in predictive analytics (Sarker, 2021).

Islam et (2021) adopted logistic regression in their study to predict addiction to drugs and alcohol. The algorithm achieved an accuracy of 97.71 % compared to other algorithms such as KNN, SVM, Naïve Bayes and CART due to its ease of interpretability. However, logistic regression has challenges, such as sensitivity to outliers where the model may not perform well in the presence of outliers.

2.6.3 Classification and Regression Tree (CART)

A decision tree is a supervised learning algorithm for classification and regression tasks, characterized by its flowchart-like structure (Praba et al., 2021). It begins at a root node, representing the entire dataset, and splits into branches based on decision rules derived from feature values, ultimately leading to leaf nodes that indicate outcomes. The algorithm employs measures like information gain and the Gini index to determine the best attributes for splitting at each node, aiming to create homogeneous subsets of data (Bharathidason & Sujdha, 2023). Generally, these decision trees are simple and can be easily interpreted and visualized; therefore, they are more reliable for understanding complex decision-making processes. However, they tend to overfit when used with noisy or imbalanced datasets, necessitating techniques like pruning to optimize performance and enhance generalization. Figure 2.2 below shows the structure of a Decision tree algorithm (Bharathidason & Sujdha, 2023).

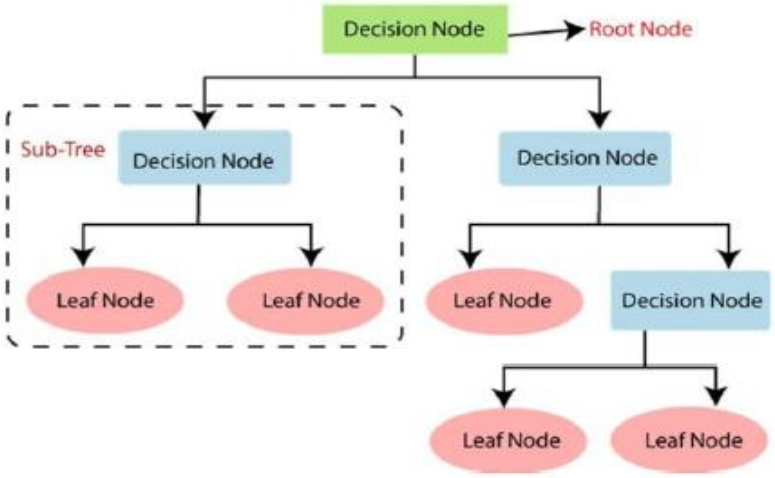


Figure 2.2: Showing Decision Tree Algorithm (Bharathidason & Sujdha, 2023)

The Classification and Regression Trees (CART) algorithm has been applied in various studies to predict substance abuse relapse. Xin et al. (2023) trained the CART model to predict drug abusers who were at risk of relapse in the Chinese setup, and the model was able to achieve an accuracy of 73.61 %, which was not a bad performance compared to the XGboost model, which was adopted with 76.01 & accuracy. Its simplicity and interpretability make it easier to adopt and implement; however, decision trees become overly complex if not pruned properly, which could lead to overfitting on training data, and this model may not be adopted.

2.6.4 Support Vector Machines (SVM)

Figure 2.3 (Praba, 2021) shows the structure of SVM, a widely used machine learning algorithm that analyzes data for classification and regression analysis (Mahesh, 2020). For this algorithm, the data used for training and testing can be split into training and testing sets. The model is imported from the support vector classifier. The model is built with the parameter kernel='poly', resulting in the model having a higher accuracy rate for prediction, fit the model in the training data, and predict with testing data, that gives the model a higher accuracy rate and compute the accuracy percentage (Bharathidason & Sujdha, 2023).

In 2021, Park et al. conducted research on a machine learning prediction of dropping out of outpatients with alcohol use disorders. Based on the results of the trained models, SVM was among the best-performed models, with an accuracy of 70.23 % and a sensitivity of 64.70 %. With its effectiveness in high dimensionality handling, SVM can handle complex datasets such as data

from substance abuse research. The algorithm, however, has its limitations, whereby with large datasets, the computational functionality becomes complex compared to other simpler models.

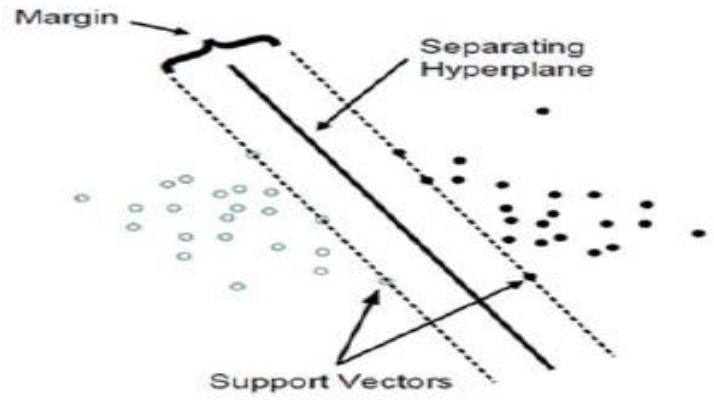


Figure 2.3: Showing Support Vector Machine (Praba, 2021)

2.6.5 Artificial Neural Networks (ANN)

According to Bharathidason & Sujdha (2023), Artificial Neural Networks (ANN) are derived from biological neural networks in the human brain. Like the brain's neurons, ANNS have interlinked nodes organized into layers. The nodes use the network to communicate, similar to the human brain structure. ANN structure is demonstrated in Figure 2.4 below.

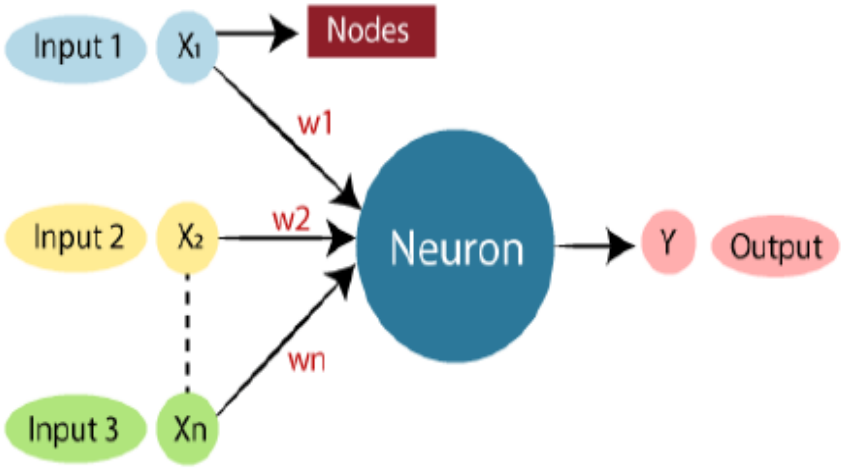


Figure 2.4: Showing Artificial Neural Network (Mahesh, 2020)

In 2021, Lai et al. developed a model to predict smoke cessation outcomes where models like support vector machine (SVM), regression (LoR), k-nearest neighbors (KNN), artificial neural network (ANN), random forest (RF), naïve Bayes (NB), and logistic classification and regression

tree (CART), were trained to predict the final smoking status of the patients in six months. ANN was adopted in this study after achieving the best performance compared to other models, with an accuracy of 64 % and sensitivity of 70.4%. The main advantage of ANN is that it can handle complex datasets and has fault tolerance in case of any damaged neurons or connections in the network. However, the algorithm has or is prone to computational complexity as the network keeps growing, which may lead to a lack of interpretability.

2.6.6 K-Nearest Neighbor (KNN)

The "k" in KNN stands for the number of nearest neighbors to consider, which needs to be chosen before using the algorithm (Bharathidason & Sujdha, 2023). KNN uses data and classifies new data points based on similar measures, such as the Euclidean distance function. One of the biggest challenges of KNN is choosing the optimal number of neighbors to be considered (Sarker, 2021). Figure 2.5 shows the structure of the KNN algorithm.

According to research by Islam et al. (2022), KNN was among the best-performing models just after random forest after achieving an accuracy of 90.98%. The research was aimed at developing a machine-learning model for predicting individual substance abuse with associated risk factors. One of the main benefits of KNN is that it is simple to implement and interpret and is flexible. The algorithm, however, has some disadvantages in that the computational cost increases as the size of the dataset grows, and also, in the presence of irrelevant features, the model may become inaccurate as it is sensitive to noise.

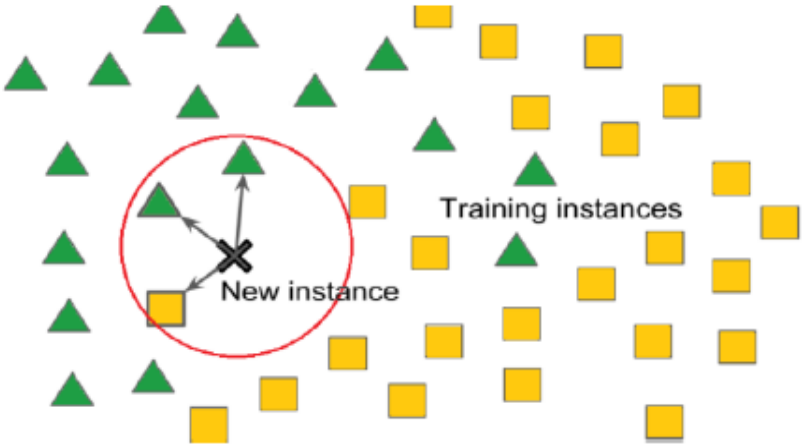


Figure 2.5: Showing K-nearest Neighbor (Bharathidason & Sujdha, 2023)

2.7 Literature Review Summary

The review highlights several factors influencing substance abuse relapse, categorized into individual, familial, social, psychological, behavioral, and economic dimensions. Individual factors such as personal commitment to recovery and the presence of co-occurring mental health disorders play critical roles in relapse likelihood. Social dynamics, particularly peer pressure and environmental triggers, significantly impact recovery outcomes. Psychological stress and low self-efficacy are crucial in understanding relapse dynamics, while economic factors such as unemployment and income levels further complicate recovery efforts.

The application of machine learning (ML) models in predicting substance abuse and relapse emphasizes their ability to analyze complex datasets that incorporate various influencing factors. Traditional predictive models often struggle with non-linear relationships among demographic, psychological, treatment history, and social influences. In contrast, ML algorithms can process large volumes of data to identify patterns that conventional methods may overlook. Research indicates that ML models, including decision trees, support vector machines, and neural networks, outperform traditional statistical techniques in capturing intricate patterns related to substance abuse.

2.8 Conceptual Framework

A conceptual framework provides a visual or theoretical framework that showcases the main components of the research and how they interact, and it serves as a blueprint (Luft et al., 2022). The conceptual framework guides the study and helps explain the logical structure behind the model for predicting substance relapse.

According to Figure 2.6, the raw data was collected from a secondary data source (Kaggle). The collected data was analyzed and processed into a more presentable standard. Eventually, the data was used to train the ML algorithms to be able to predict the possibilities of an individual relapsing from substance use. Lastly, the model was assessed to establish its precision and accuracy.

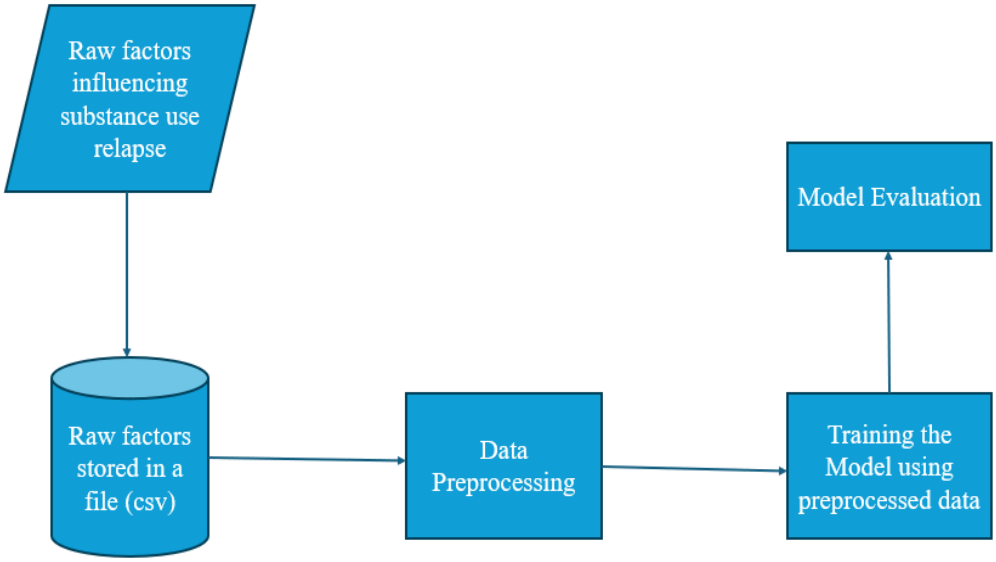


Figure 2.6: Showing Conceptual Framework

Chapter 3: Research Methodology

3.1 Introduction

Creswell & Creswell (2018) describe research methodology as a structural approach the researcher employs to conduct the study. The research methodology encompasses the specific procedure, such as the tools and techniques used to collect, process and interpret the data. This section demonstrates the process through which the research outcome will be achieved. Specifically, the section discusses the methodological approaches that will be used to attain the aims and objectives of the study. These methodological approaches include the research design, data collection, the target population and sampling, the data analysis approach, the research validity and reliability, and ethical considerations in the study process.

3.2 Research Design

According to Creswell & Creswell (2018), the research design is the blueprint or structure that guides the investigation of a research problem. It outlines how a research study will be conducted, including the methods and procedures used to gather and analyze data, as well as the structure and organization of the study. An applied research approach was used on this study as this research sought to solve a practical problem.

3.3 Study Population and Sampling Strategies.

3.3.1 Target Population

The target population for the study on machine learning model to predict substance abuse in Nairobi County, Kenya, was consistent of adults aged 18 to 65, encompassing diverse demographic backgrounds such as gender, socioeconomic status, and various substance use patterns in rehabilitation centres within Nairobi County. This population included individuals who have experienced substance dependence on various substances, including alcohol, tobacco, khat, cannabis, heroin, and prescription drugs, among others. It was also crucial to focus on those with a documented history of relapse, as their experiences would contribute important insights into the factors that influence recurring substance use. Additionally, incorporating individuals with co-

occurring mental health disorders would help capture the complexities of their recovery journeys, as these conditions often exacerbate substance use issues.

3.3.2 Sampling size

Sampling size refers to the number of subjects or units selected from a population for inclusion in a research study. It is a critical aspect of research design as it determines the extent to which the findings from the sample can be generalized to the larger population. The sampling size is big enough to give the needed trust in the data and the researcher (Saunders et al., 2019). The sample size of the secondary dataset obtained from Kaggle was 10000 samples, however after data cleaning 4654 samples were utilized in the development of the model.

3.4 Data Collection Method and Analysis

The data collection includes the approaches the researcher uses to collect raw data from primary and secondary sources for analysis. This step is crucial to the entire research, and it encompasses collecting relevant data to help meet the research objectives and fulfil the research aim. The data collection processes vary depending on the nature of the study and the data collection type.

3.4.1 Data Collection

Due to limited timelines, the research employed secondary data collected from the open public data site Kaggle. The dataset comprised of several features such as age, gender, socioeconomic status, peer influence, family background, mental health and more, which are the factors that influence substance use relapse. The data quality assessment was done on the dataset to ascertain that the data was complete, consistent and had integrity. This included data cleaning to ensure that the data was in a state that could be processed with ease. Once data cleaning had been done, data transformation was carried out. This was for feature engineering to create variables close to the features needed from the dataset to develop the model.

After data preprocessing, data splitting was done, where the dataset would be partitioned to train_test split. This study used 20% of the data for testing and 80% for training. Though it is not a strict rule that the sample size for testing and training machine learning should be 20% and 80%, respectively. An 80/20 split gives the model a sizable enough sample for training while leaving enough data for testing and performance evaluation. Doing this can prevent the model

from overfitting to the training set, which could result in subpar performance on fresh, untested data.

I7 7

	A	B	C	E	F	G	H	I	J	K	L	M	N	O
1	Year	Age_Group	Gender	Drug_Experimentation	Socioeconomic_Status	Peer_Influence	School_Programs	Family_Background	Mental_Health	Access_to_Counseling	Parental_Supervision	Substance_Education	Community_Support	Media_Influence
2	2024	15-19	Both	32.4	High	5	Yes	1	5	No	4	No	3	1
3	2024	14-Oct	Female	41.57	High	6	Yes	10	5	No	9	Yes	9	3
4	2023	14-Oct	Both	56.8	High	6	Yes	2	7	Yes	2	No	5	1
5	2024	40-49	Both	42.9	Middle	10	No	9	7	Yes	2	No	10	9
6	2023	15-19	Male	39.62	High	1	No	2	4	Yes	4	No	10	3
7	2022	70-79	Male	47.29	High	10	Yes	7	4	No	4	No	4	2
8	2021	30-39	Female	26.6	Low	4	No	7	1	Yes	2	No	4	10
9	2022	14-Oct	Female	54.67	Middle	5	Yes	9	2	No	4	No	4	2
10	2021	70-79	Both	63.65	Low	9	Yes	1	6	Yes	5	Yes	2	6
11	2020	60-69	Female	17.62	Low	2	No	3	6	Yes	5	Yes	9	9
12	2020	40-49	Male	66.66	High	5	No	5	3	No	9	No	2	10
13	2022	50-59	Both	11.41	High	2	No	8	3	Yes	10	No	10	4
14	2024	25-29	Male	19.46	Low	5	Yes	2	4	Yes	1	No	10	8
15	2022	60-69	Female	42.47	High	8	No	4	1	No	5	Yes	3	2
16	2020	15-19	Male	65.31	Low	9	No	1	10	No	2	Yes	9	1
17	2020	60-69	Male	62.21	Low	5	Yes	9	9	Yes	2	No	1	4
18	2023	30-39	Both	57.28	High	2	Yes	9	3	Yes	4	No	8	1
19	2024	15-19	Female	39.62	Middle	10	No	8	4	Yes	8	No	6	8
20	2021	30-39	Female	22.28	High	1	Yes	6	8	No	7	Yes	2	6
21	2021	15-19	Male	30.66	Low	3	Yes	7	5	No	5	Yes	9	7
22	2023	30-39	Both	45.36	Low	6	Yes	4	3	No	10	Yes	4	6
23	2022	70-79	Male	50.57	Middle	6	Yes	9	2	No	2	Yes	7	8
24	2021	25-29	Female	67.8	High	3	Yes	4	7	Yes	6	No	5	3
25	2020	60-69	Male	14.25	High	2	Yes	3	6	No	10	No	3	5

Figure 3.1: Showing Sample Data

3.4.2 Data Analysis

In this research, data analysis included cleaning data to remove any missing values and inspecting and transforming the data into a format to gain insight into the research's objectives. The study combined both quantitative and qualitative data analysis techniques. The intensity of the research approach was determined depending on time sufficiency and data resources.

3.5 Model Development

3.5.1 Algorithms to be used

This research developed the model using decision trees, random forests, K-Nearest Neighbor, Artificial Neural Network and support vector machine (SVM) algorithms. These models are particularly well-suited for diverse datasets and provide unique advantages depending on the problem's complexity and the nature of the data. Their flexibility and capability to handle simple and complex relationships between features and target variables make them widely adopted in machine learning applications.

A decision tree is a simple yet powerful classification algorithm that models' decisions based on a tree-like conditions structure. It splits data into subsets using feature thresholds that maximize information gain or reduce impurity. Each internal node represents a decision based on a feature, branches denote possible outcomes, and leaves indicate the final prediction or class label. Decision trees are easy to interpret and visualize, making them valuable for understanding relationships between predictors and outcomes.

Random Forest was utilized to ensure diversity among trees. This randomness makes Random Forest robust to noise and overfitting, while its ensemble nature enhances predictive performance. Random Forest is highly versatile, can handle numerical and categorical data, and provides feature importance scores.

Support Vector Machines, on the other hand, aim to find an optimal hyperplane to separate data into distinct classes. By maximizing the margin between the hyperplane and the nearest data points (support vectors), SVMs ensure a strong generalization capacity. SVMs are particularly effective for high-dimensional data and cases where classes are not linearly separable, using kernel tricks to map data to higher dimensions.

3.5.2 Model Testing & Validation

Model testing involved evaluating the predictive model's performance using a separate test dataset that is not be used during training. This ensured the model could generalize to unseen data and not overfit the training set. Standard testing techniques include cross-validation, where the data is split into multiple subsets to assess performance across different segments.

On the other hand, model validation ensured that the predictive model performs well on the test data and new, unseen data from real-world scenarios. This research validated the models through model performance metrics validation, which evaluated how well a predictive model performs based on specific metrics that assess various aspects of the model's accuracy, reliability, and generalizability. These metrics, which include the Precision, Recall, Accuracy and Confusion matrix, provided insights into how well the model made predictions, identified patterns, and handled different types of errors.

3.5.3 Model Deployment

The model was deployed via a web application that integrates the trained model into a backend API, such as Flask, which would receive user data and run. The front-end interface, built with technologies like HTML or CSS, which allowed the users to input data and view the results. The backend API processes the data, sends it to the model and returns a response while the application is deployed on a cloud platform for scalability and accessibility.

3.6 Reliability and Validity of the Study

The validity of this study was ensured by having data sets with no missing values and taking the data through intense cleaning and transformation to ensure consistency in the data set. Additionally, validity and reliability could be achieved using already successful and established algorithms, such as the Random Forest Support Vector Machine, to help optimize the model's functionality. Further reliability would come from adhering to ethical consideration frameworks outlined by the Strathmore Ethical Review Committee and ensuring that the study's methodological approaches align with what is considered acceptable.

3.7 Results Utilization and Dissemination

3.7.1 Results Utilization

The results of this research will be used to inform local health authorities and policymakers about the critical factors influencing relapse rates. This information will be vital for designing targeted interventions that address specific demographic groups or risk factors prevalent in Nairobi County. Also, the model will help allocate resources more effectively by identifying high-risk populations or areas within Nairobi that require additional support or intervention efforts.

3.7.2 Results Dissemination

This will be achieved through collaboration with rehabilitation centres within Nairobi, which will help facilitate the integration of research findings into existing health programs. This partnership will help in tailoring interventions that address specific local needs based on the research outcomes.

3.8 Ethical Considerations.

In the process of data collection, analysis and presentation, this study ensured that the privacy and confidentiality of study was adhered to. This was ascertained by refraining from collecting data with a person and identifying information like names, addresses, or contacts. Also,

the researcher sought an ethical review from SU-ISERC to ensure that all the ethical and consideration aspects of the research will be made. The ethical approval was issued by the Strathmore Ethical Review Committee for this research providing a go ahead to implement the model development. No further data collection clearance was needed as the researcher utilized the publicly available dataset.

Chapter 4: System Design and Architecture

4.1 Introduction

The system design and architecture focuses on the development of web-based substance abuse relapse prediction system that can be used by the various stakeholders involved in the management of the substance abuse recovery based on the conceptual model presented in figure 2.6. System design and architecture helps in clarifying the functionality of the developed system and the iteration between different components.

4.2 Requirement Analysis

Requirement Analysis stage focuses on evaluating users' expectations to ensure the model aligns with all stakeholders' needs, in line with the research study's objectives. Similarly, and in accordance with these objectives, the following discussion primarily addresses the functional and non-functional requirements of the developed model.

4.2.1 Functional Requirements

Functional requirements describe what the model has to do by pinpointing the tasks, actions and/or activities that must be accomplished.

The functional requirements of the model include:

i. The system should allow users to self-register on first visit.

ii. The system should allow user to login.

iii. The system should allow users to input parameters for prediction.

iv. The system should allow users to make prediction.

v. The system should be able to display prediction results.

4.2.2 Non-Functional Requirements

Non-functional requirements are the general requirement attributes and qualities that the system must have for optimal performance.

The non-functional requirements include:

i. Performance & Scalability

The system will be able to handle multiple concurrent requests efficiently.

ii. **User friendly**

The system will have a user-friendly user interface making it easy to learn, operate, prepare inputs and interpret outputs as users interact with it.

iii. **Availability & Reliability**

The system will be dependable and will be able to function around the clock.

iv. **Usability & User Experience**

Simple, intuitive web interface for non-technical users.

v. **Efficiency**

The system will be able to handle capacity and throughput within the specified maximum time.

vi. **Integrity**

The system data will be maintained accurately, authentically and without corruption.

vii. **Maintainability**

The system will allow for ease of finding faults and errors, and fix them to meet the user requirements

4.3 System Architecture

Figure 4.1 below represents the system architecture for the substance abuse relapse prediction system. It will illustrate the general interaction of various components to achieve system functionality. The raw data will be pre-processed to create training and test datasets. Thereafter, the model will be trained and validated then converted to a format that can be embedded into a web application. The web interface then will display the predictions as per users' requests, at the same time it will also allow users to input their queries.

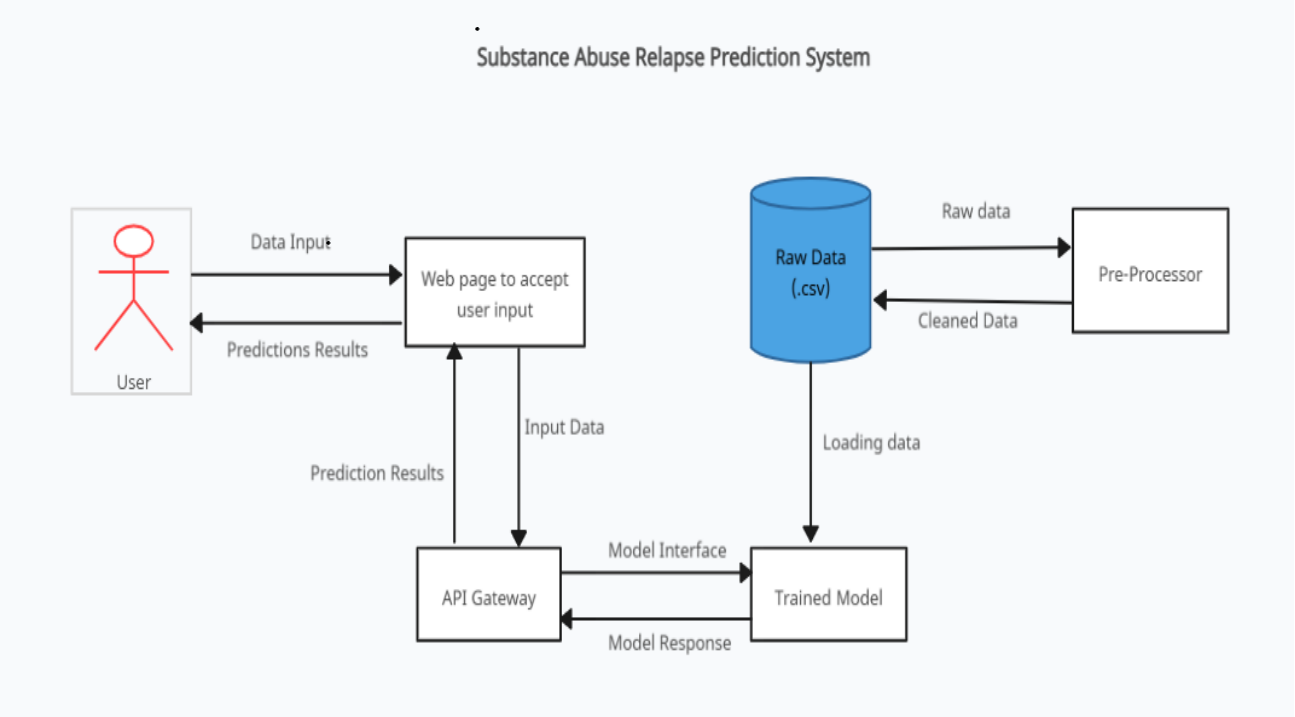


Figure 4.1 System Architecture

4.4 Use-Case Diagram

A use-case diagram illustrates the interactions and sequence of actions between actors and the system, showcasing how users engage with the system. Figure 4.2 outlines the functionalities that the proposed model is expected to deliver. The primary actors in this system are the users and the system administrator. The system administrator is responsible for collecting and cleaning the data, training the algorithm, and testing the developed model. The insights and knowledge derived from this process are then presented to the users via a web interface.

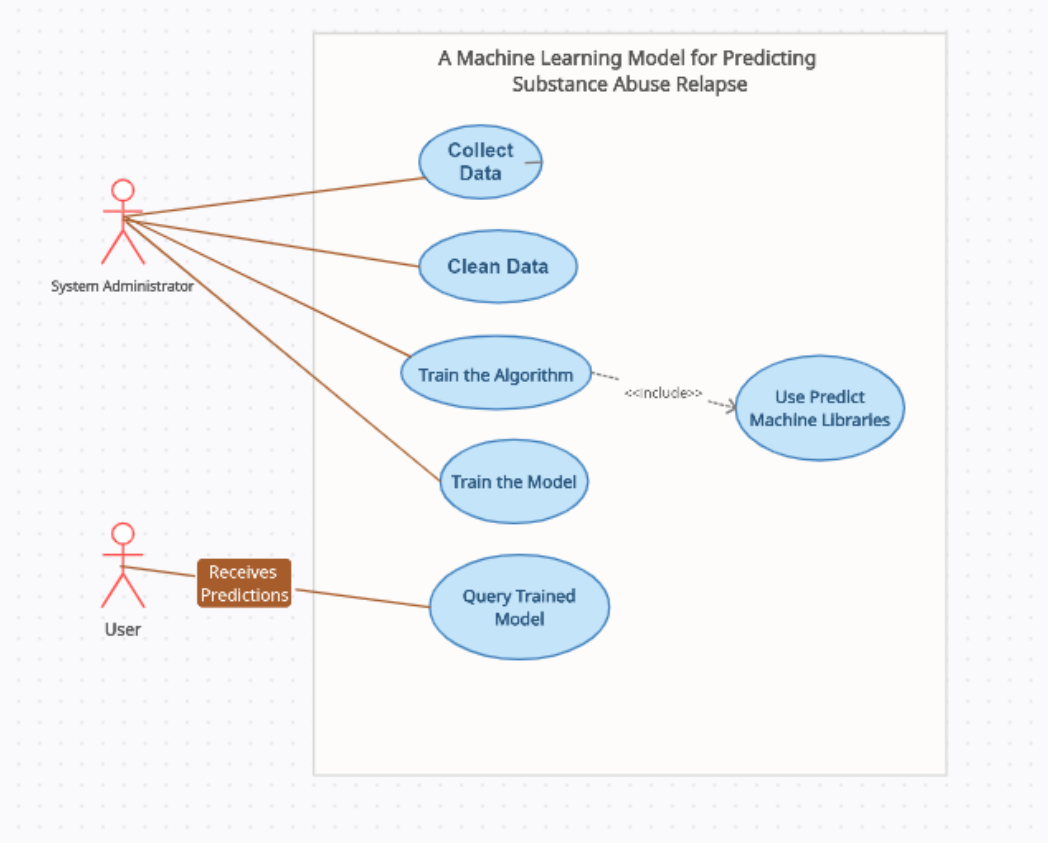


Figure 4.2: Use-Case Diagram

Use Case: Data Collection
Primary Actor/s: System Administrator
Brief Description: This use case describes how the System Administrator will collect the data
Pre-condition: Data is relevant to substance abuse relapse
Post-condition: The System Administrator identifies the data sources and retrieves the data
Main Success Scenarios

Actor Responsibility	**System Responsibility**

1. The system administrator identifies the relevant data sources.	
2. The System administrator collects the data	
	3.The collected data is stored and saved in the system.

Table 4.1: Description of Data Collection

Use Case: Data Cleaning	
Primary Actor/s: System Administrator	
Brief Description: This use case describes how the System Administrator will clean the collected raw dataset/s.	
Pre-condition: Adequate data available	
Post-condition: Pre-processed dataset	
Main Success Scenarios	
Actor Responsibility	**System Responsibility**
1. The system Administrator imports the libraries	
2. The System Administrator imports the dataset	
3. The system Administrator carries out data cleaning to check out for missing values and handle them.	
4. The system administrator splits the dataset into training and testing sets	
	5. The system performs feature extraction and saves the features in the system.

Table 4.2: Description of Data Cleaning

Use case: Model Training	
Primary Actor/s: The System Administrator	
Brief Description: This use-case describes how the System Administrator will train the model.	
Pre-condition: Pre-processed dataset, training system available	
Post-condition: Trained model that can predict likelihood of substance abuse relapse	
Main Success Scenarios	
Actor Responsibility	**System Responsibility**
1. The Administrator selects pre-processed data	
2. The Administrator selects training and testing set percentage	
3. The Administrator selects the output format and executes the training command.	
	4. Splits dataset into training sets as executed by the administrator
	5. Outputs the trained model.

Table 4.3: Description of Model Training

4.5 Sequence Diagram

A sequence diagram visually represents the interactions between objects or components within a system, illustrating the order in which these interactions occur. It is commonly used to model a system's dynamic behavior, aiding in understanding and analyzing system requirements. Figure 4.5 represents the sequence diagram for the proposed system.

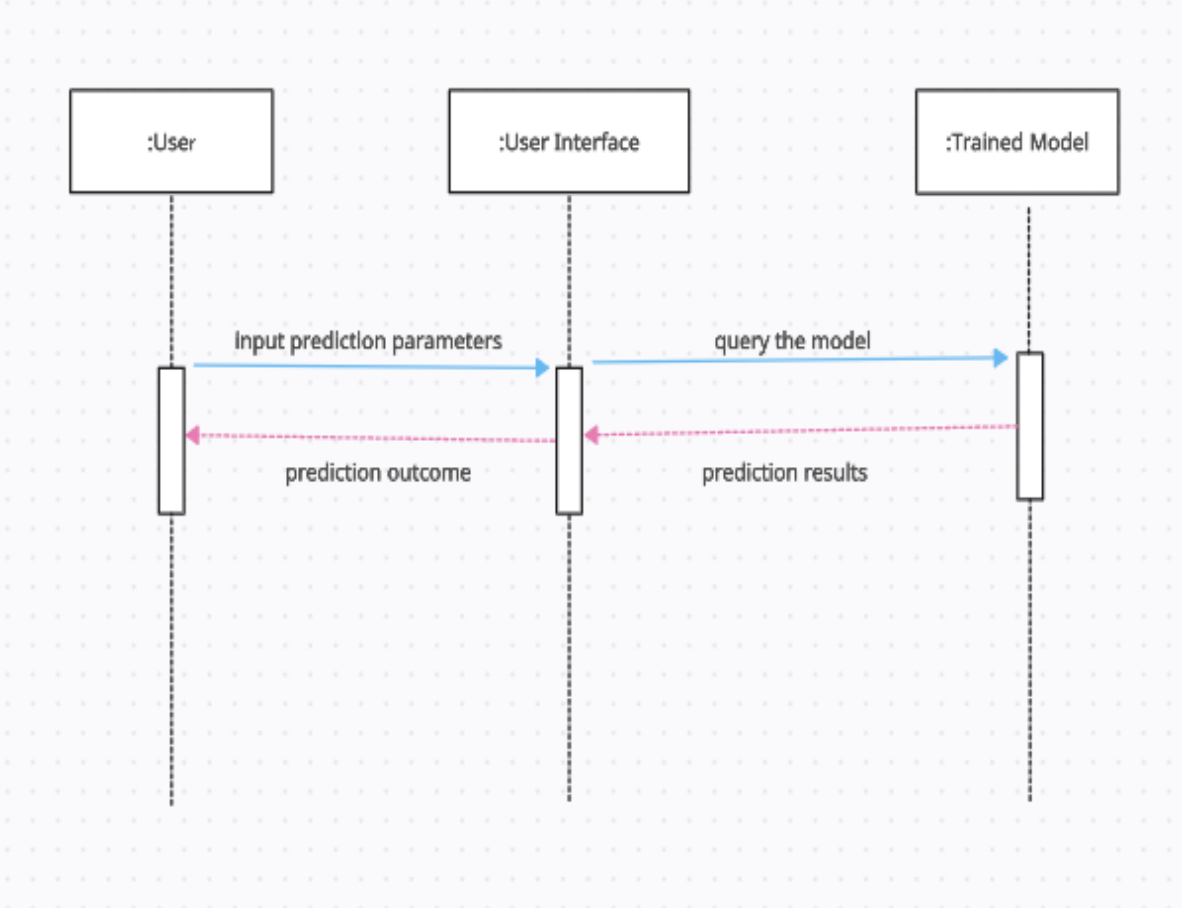


Figure 4.3: Sequence Diagram

4.6 Context Diagram

4.6.1 Low-level Context Diagram

A context diagram is a visual representation of a system and its interactions with external entities, such as users, other systems, or organizations. It provides an overview of the system's boundaries, inputs, and outputs without delving into internal processes or data flow details. Context diagrams help stakeholders understand how a system fits into its broader environment and clarify system requirements. The user and the trained model are the main entities interacting in this proposed system. Figure 4.4 shows how external entities interact with the model.

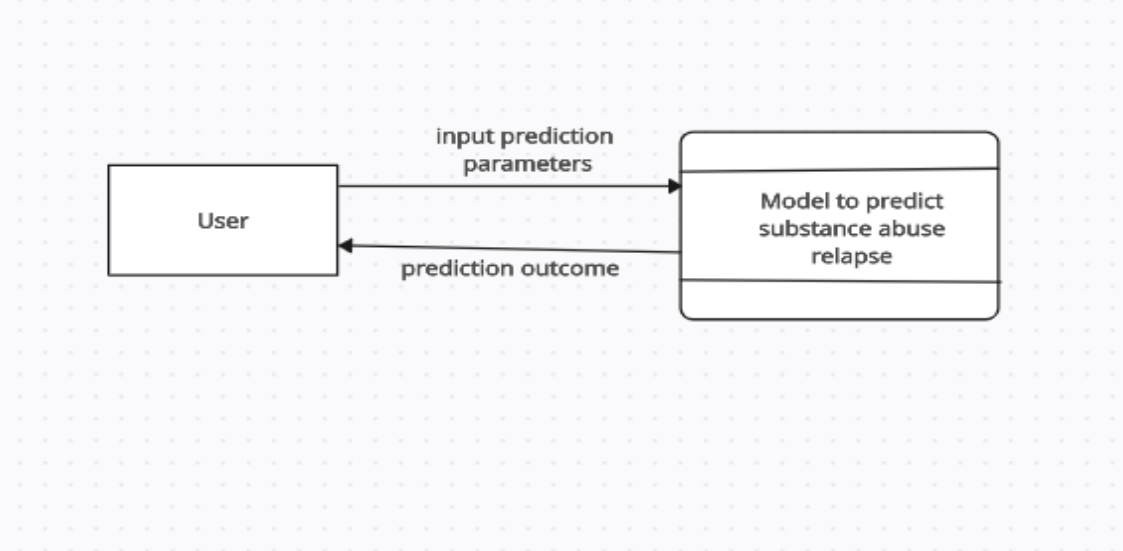


Figure 4.4: Low-Level Context Diagram

4.6.2 High-level Context Diagram

The high-level context diagram as per Figure 4.5 below represents the entities, data stores and processes in the proposed system, which provides a clear understanding of the data flow process in the system.

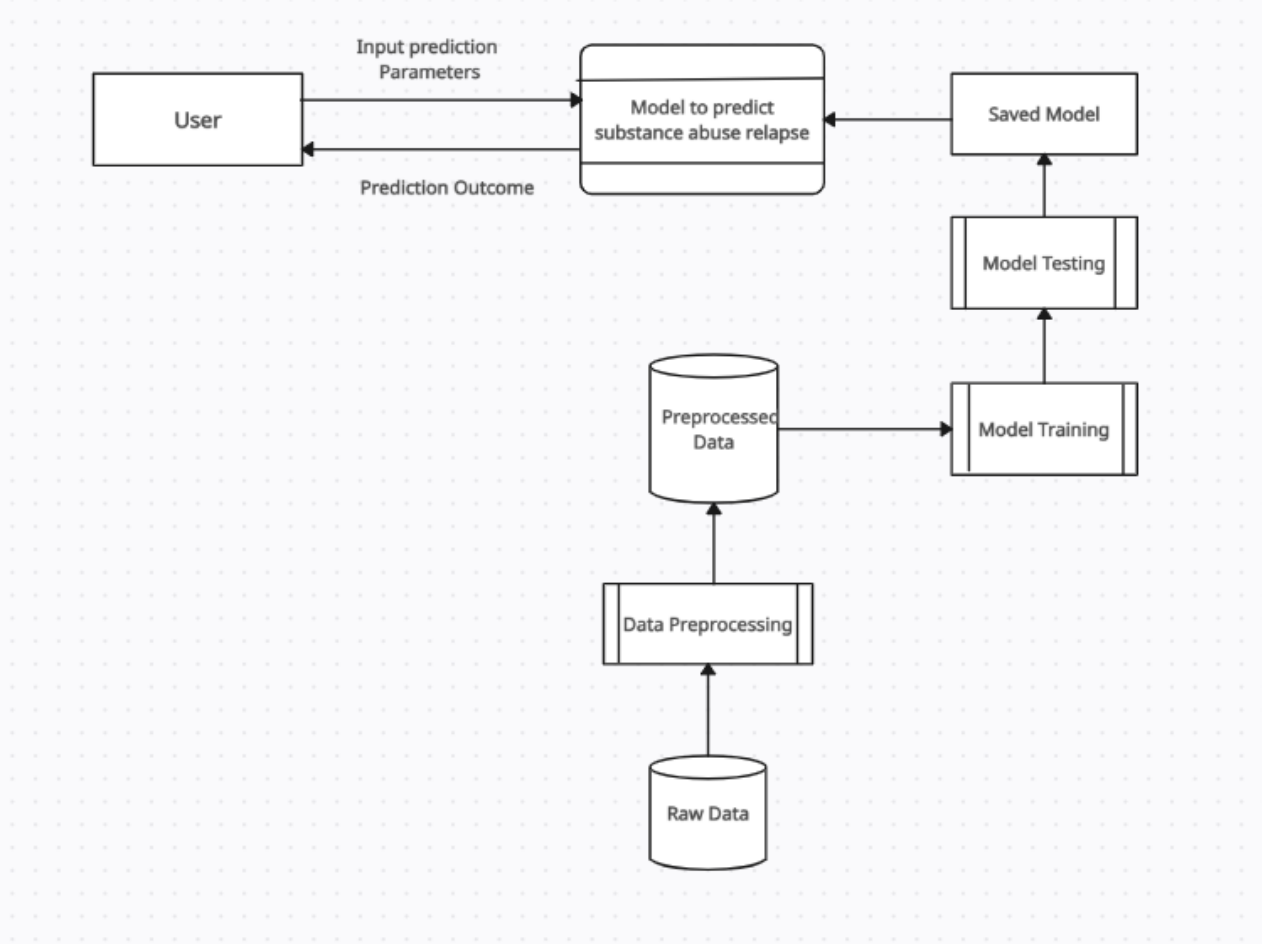


Figure 4.5: High-Level Context Diagram

Chapter 5: System Implementation and Testing.

5.1 Introduction

The system implementation and testing phase describes how the substance abuse relapse model was developed. It covers how the developed model was tested and validated. The process began by obtaining the dataset used from Kaggle, then followed by building the model using python and its' libraries, and the supervised learning algorithm after the preprocessing of the obtained dataset had been done. Finally, it will show or describe the use of the developed model in predicting substance abuse relapse through a web-based user interface.

5.2 System Implementation

The implementation of the substance abuse relapse model entailed several processes, data collection and preprocessing, model selection, model training and validation, and model deployment.

5.2.1 Dataset Extraction

The raw research data used was collected through a secondary source, Kaggle (https://www.kaggle.com/datasets/waqi786/youth-smoking-and-drug-dataset/data). The main reason this dataset was used is because it contained crucial features which have high contribution in substance use relapse. The data was exported from Kaggle in the format of a CSV file. The dataset provides an In-depth Look at Trends in Substance Use (2020-2024).

5.2.2 Data Preprocessing

After data collection was done, preprocessing of the data was necessary as the data was in an unstructured format which is not suitable for use in machine learning techniques. Thus, required pre-processing before the data could be used in the model training. Figure 5.1 below shows sample data extracted from Kaggle.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Year	Age_Group	Gender	Smoking_Prevalence	Drug_Experimentation	Socioeconomic_Status	Peer_Influence	School_Programs	Family_Background	Mental_Health	_Counseling	Parental_Supervision	Substance_Education	Community_Support	Media_Influence
2	2024	15-19	Both	18.85	32.4	High	5	Yes	1	5	No	4	No	3	1
3	2024	14-Oct	Female	34.88	41.57	High	6	Yes	10	5	No	9	Yes	9	3
4	2023	14-Oct	Both	42	56.8	High	6	Yes	2	7	Yes	2	No	5	1
5	2024	40-49	Both	33.75	42.9	Middle	10	No	9	7	Yes	2	No	10	9
6	2023	15-19	Male	47.9	39.62	High	1	No	2	4	Yes	4	No	10	3
7	2022	70-79	Male	20.14	47.29	High	10	Yes	7	4	No	4	No	4	2
8	2021	30-39	Female	38.38	26.6	Low	4	No	7	1	Yes	2	No	4	10
9	2022	14-Oct	Female	7.87	54.67	Middle	5	Yes	9	2	No	4	No	4	2
10	2021	70-79	Both	11.61	63.65	Low	9	Yes	1	6	Yes	5	Yes	2	6
11	2020	60-69	Female	23.98	17.62	Low	2	No	3	6	Yes	5	Yes	9	9
12	2020	40-49	Male	42.85	66.66	High	5	No	5	3	No	9	No	2	10
13	2022	50-59	Both	48.72	11.41	High	2	No	8	3	Yes	10	No	10	4
14	2024	25-29	Male	37.71	19.46	Low	5	Yes	2	4	Yes	1	No	10	8
15	2022	60-69	Female	44.99	42.47	High	8	No	4	1	No	5	Yes	3	2
16	2020	15-19	Male	11.78	65.31	Low	9	No	1	10	No	2	Yes	9	1
17	2020	60-69	Male	41.94	62.21	Low	5	Yes	9	9	Yes	2	No	1	4
18	2023	30-39	Both	49.99	57.28	High	2	Yes	9	3	Yes	4	No	8	1
19	2024	15-19	Female	24.61	39.62	Middle	a	No	8	4	Yes	8	No	6	8
20	2021	30-39	Female	15.08	22.28	High	1	Yes	6	8	No	7	Yes	2	6
21	2021	15-19	Male	15.4	30.66	Low	3	Yes	7	5	No	5	Yes	9	7
22	2023	30-39	Both	15.17	45.36	Low	6	Yes	4	3	No	10	Yes	4	6
23	2022	70-79	Male	10.49	50.57	Middle	6	Yes	9	2	No	2	Yes	7	8
24	2021	25-29	Female	9.95	67.8	High	3	Yes	4	7	Yes	6	No	5	3
25	2020	60-69	Male	26.59	14.25	High	2	Yes	3	6	No	10	No	3	5
26	2022	15-19	Male	18.41	15.59	Middle	7	No	3	9	Yes	6	No	8	4
27	2023	15-19	Female	46	32.2	High	1	No	7	1	Yes	3	No	5	10
28	2020	14-Oct	Both	14.8	69.46	Middle	3	No	6	10	Yes	8	No	8	8

Figure 5.1: Sample of Collected Data

To start with preprocessing, data normalization was applied to rescale numerical features to a standardized range, usually between 0 and 1 or -1 and 1. This process prevented features with higher magnitudes from overshadowing others during training, ensuring balanced contributions from all features. Subsequently, categorical encoding was employed to convert categorical variables such as Gender, socioeconomic_status, Access_to_counselling, and Substance_Education into numerical representations suitable for analysis. Additionally, using the available features the target variable that is “Substance_Relapse” was generated using a weighted formula based on the risk factors. For example, on socioeconomic_status low status

Data Analysis was also done to ensure that there are no missing values on the dataset and no outliers. Through this analysis, no outliers or missing values were identified within the dataset. Figure 5.2 shows a sample of the cleaned data.

	Age_Group	Gender	Drug_Experimentation	Socioeconomic_Status	Peer_Influence	Family_Background	Mental_Health	Access_to_Counseling	Substance_Education	Community_Support	Media_Influence	Substance_Relapse
2	0	2	32.4	2	5	1	5	0	0	3	1	0
3	4	2	42.9	1	10	9	7	1	0	10	9	1
4	0	0	39.62	2	1	2	4	1	0	10	3	0
5	7	0	47.29	2	10	7	4	0	0	4	2	1
6	3	1	26.6	0	4	7	1	1	0	4	10	0
7	7	2	63.65	0	9	1	6	1	1	2	6	1
8	6	1	17.62	0	2	3	6	1	1	9	9	0
9	4	0	66.66	2	5	5	3	0	0	2	10	1
10	5	2	11.41	2	2	8	3	1	0	10	4	0
11	2	0	19.46	0	5	2	4	1	0	10	8	0
12	6	1	42.47	2	8	4	1	0	1	3	2	0
13	0	0	65.31	0	9	1	10	0	1	9	1	1
14	6	0	62.21	0	5	9	9	1	0	1	4	1
15	3	2	57.28	2	2	9	3	1	0	8	1	1
16	0	1	39.62	1	10	8	4	1	0	6	8	1
17	3	1	22.28	2	1	6	8	0	1	2	6	0
18	0	0	30.66	0	3	7	5	0	1	9	7	0
19	3	2	45.36	0	6	4	3	0	1	4	6	1
20	7	0	50.57	1	6	9	2	0	1	7	8	1
21	2	1	67.8	2	3	4	7	1	0	5	3	1
22	6	0	14.25	2	2	3	6	0	0	3	5	0
23	0	0	15.59	1	7	3	9	1	0	8	4	0
24	0	1	32.2	2	1	7	1	1	0	5	10	0
25	6	0	18.92	1	9	7	10	1	1	1	10	0
26	5	2	11.52	0	6	3	10	0	1	3	8	0

Figure 5.2: Sample of Cleaned Data

5.2.3 Feature Selection

For the substance abuse relapse to be predicted, the predictive model was developed with various features which provided invaluable insights into substance abuse relapse factors. The features used in the substance abuse relapse model include:

i.) Age_Group
ii.) Gender
iii.) Drug_Experimentation
iv.) Socioeconomic_Status
v.) Peer_Influence
vi.) Family_Background
vii.) Mental_Health
viii.) Access_to_Counseling
ix.) Substance_Education
x.) Community_Support

xi.) Media_Influence

xii.) Substance_Relapse (Target Variable)

Table 5.1 below illustrates further the description of these features drawn from the dataset which is used in the prediction.

Variable	Description	Possible Values
Age_Group	This is classification of individuals into different are ranges.	Age is between 15-69 years
Gender	This attribute shows whether the person is male or female or Both	Female (1), Male (0)
Drug_Experimentation	Shows the levels of drug experimentation	Ranges from 0.0 to 100.00
Socioeconomic_Status	A combined measure of an individual or family's social and economic standing, encompassing factors like income, education, occupation, and wealth.	Low (0), Middle (1), High (2)
Peer_Influence	Attributes on how peers impact a person's behavior	Ranges from 1 to 10 (Low to high Influence)
Family_Background	Shows the family dynamics and support systems	Ranges from 1 to 10 (1- Absent, 2-Hihgly Toxic, 3-Dysfunctional, 4-At-risk, 5-Unstable, 6-Mixed Environment, 7-Mostly Stable, 8-Positive, 9-Very Positive, 10-Highly Supportive)
Mental_Health	This is attributed to individuals' emotional, psychological, and social well-being.	Ranges from 1 to 10 (1- Crisis Level, 2-Severe Condition, 3-Significant Impairment, 4-Struggling, 5-Moderate Concerns, 6-Mild Concerns, 7-Stable, 8-Good, 9-Very Good, 10-Excellent)
Access_to_Counseling	Attributes to the availability and ease with which individuals can receive professional mental health support and guidance to address substance abuse and related issues	Yes (1), No (0)
Substance_Education	This attributes to the level of access to information on the impact of substances	Yes (1), No (0)

Community_Support	The social support individuals receive from their community ties and organizations, encompassing resources and assistance from neighbors, local groups, and community services	Ranges from 1 to 10 (1-No support, 2-Minimal support, 3-Underserved, 4-Weak support, 5-Neutral, 6-limited support, 7-Moderately supportive, 8-Strong support, 9-Very strong support, 10-Exceptional support)
Media_Influence	The attribute shows the impact of media on an individuals' behaviors and influencing their social norms and values	Ranges from 1 to 10 (1-Extremely Harmful Influence, 2-Very Harmful Influence, 3-Strong Negative Influence, 4-Moderate Negative Influence, 5-Mild Negative Influence, 6-Neutral, 7-Moderate Positive, 8-Generally Positive, 9-Very Positive Influence, 10-Strong Positive Influence)
Substance_Relapse	This is attributed to the outcome (target variable) whether the individual is at high risk of relapse or not based on the other variables.	Yes (1), No (0)

Table 5.1: Description of the features.

5.2.4 Model Training

For this research, supervised learning techniques/algorithms were used to train the classification model because these algorithms can learn from labeled data allowing them to generalize well on new, unseen dataset. Additionally, due to the use of labeled data the classification models can be evaluated using metrics such as accuracy, precision, recall and F1-score making performance assessment straightforward. The model was trained with K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest, and Decision Tree algorithms and Artificial Intelligence Neural Networks (ANN) algorithms.

To perform model training, the cleaned dataset which was saved in csv file was imported into Pythons' data frame using the panda's library. Thereafter, features were split into two sets, training and testing sets. Using the train_test_split method on scikit-learn python library, the

features were split into 80% training set and 20% testing set in order to test the model's performance. The features were also optimized using the fit_transform method in scikit-learn library. Figure 5.3 shows the python code to optimize and split the features.

```
# Separate features and target
X = Dataset.drop(columns=["Substance_Relapse"])
y = Dataset["Substance_Relapse"]

# Split dataset into training and testing sets (80% train, 20% test for balance)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)
Y_labels = y_test

# Normalize continuous feature
scaler = MinMaxScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Figure 5.3 Data splitting for training and testing

After data splitting into training and testing sets, five different algorithms were trained to implement the models: K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest, and Decision Tree algorithms and Artificial Intelligence Neural Networks (ANN) algorithms.

5.2.5 Performance Evaluation

Performance evaluation is a procedure to evaluate the trained machine learning models used in the classification. This will explore further the performance of the five trained models: K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest, and Decision Tree algorithms and Artificial Intelligence Neural Networks (ANN). To ensure the accuracy of these models, both cross-validation and hyperparameter tuning techniques were performed to make sure the accuracy was valid and consistent, and that the models were not overfitting or memorizing training data.

5.2.5.1 Decision Tree Model

A Decision Tree is a supervised learning algorithm that splits data into branches based on feature values, forming a tree-like structure. It makes decisions by following a sequence of conditions from root to leaf nodes. Decision trees are easy to interpret but prone to overfitting, which can be mitigated using pruning or ensemble methods like Random Forest.

After the dataset was loaded and split into 2 sets, training and testing as per figure 5.3, Decision Tree performance was done using the performance metrics: Accuracy, Precision, Recall, F1-score and Confusion Matrix. Figure 5.4 below shows the performance of Decision Tree.

Accuracy	Precision	Recall	FI-Score
0.96	0.96	0.95	0.96

Table 5.2: Decision Tree Algorithm Performance

```
# Train Decision Tree model
dt = DecisionTreeClassifier(random_state=42)
dt.fit(X_train, y_train)

y_pred = dt.predict(X_test)

# Get training accuracy for DT
dt_train_acc = dt.score(X_train, y_train)
dt_test_acc = dt.score(X_test, y_test)

# Perform 5-fold cross-validation
dt_cv_scores = cross_val_score(dt, X_train, y_train, cv=5)

# Evaluate Decision Tree model
accuracy = accuracy_score(y_test, y_pred)
print(f"DT Model Accuracy: {accuracy:.4f}")
print("Classification Report:\n", classification_report(y_test, y_pred))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred))
```

```
DT Model Accuracy: 0.9603
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.95      0.96       461
           1       0.96      0.97      0.96       470

    accuracy                           0.96       931
```

Figure 5.4: Initialization and Prediction Using Decision Tree Model

Decision Tree Model achieved an overall accuracy of ~0.9603, showing that the model performed quite well with the presented dataset. Using python's seaborn and matplotlib libraries, below figure 5.5 illustrates confusion matrix for the predictions of Decision Tree model

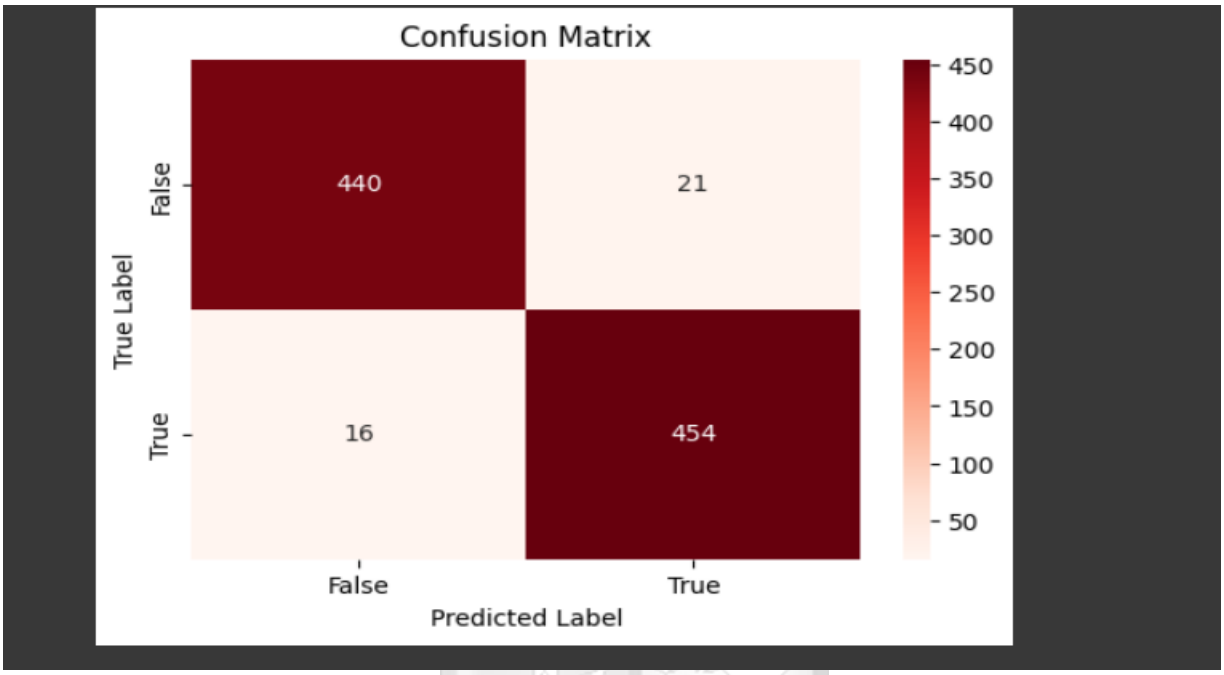


Figure 5.5: Decision Tree Model Confusion Matrix

5.2.5.2 Random Forest Model

Random Forest is an ensemble learning algorithm that builds multiple decision trees and combines their outputs to improve accuracy and reduce overfitting. Each tree is trained on a random subset of data, and the final prediction is based on majority voting (classification). It is robust to noise and works well for high-dimensional datasets.

Using the sklearn.ensemble import RandomForestClassifier, the model was trained with the defined 80% training dataset and tested with 20% of the dataset. This model proved to achieve high performance of ~0.9807 which is higher compared to the Decision Tree model.

Accuracy	Precision	Recall	FI-Score
0.98	0.98	0.98	0.98

Table 5.3: Random Forest Algorithm Performance

```python
# Train Random Forest model
rf = RandomForestClassifier(n_estimators=100, random_state=42)
rf.fit(X_train, y_train)

rf_preds = rf.predict(X_test)

# Get training accuracy for RF
rf_train_acc = rf.score(X_train, y_train)
rf_test_acc = rf.score(X_test, y_test)

# Perform 5-fold cross-validation
rf_cv_scores = cross_val_score(rf, X_train, y_train, cv=5)

# Evaluate Decision Tree model
accuracy = accuracy_score(y_test, rf_preds)
print(f"Random Forest Model Accuracy: {accuracy:.4f}")
print("Classification Report:\n", classification_report(y_test, rf_preds))
print("Confusion Matrix:\n", confusion_matrix(y_test, rf_preds))
```

```
Random Forest Model Accuracy: 0.9807
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.98      0.98       461
           1       0.98      0.99      0.98       470

    accuracy                           0.98       931
   macro avg       0.98      0.98      0.98       931
weighted avg       0.98      0.98      0.98       931
```

Figure 5.6: Initialization and Prediction Using Random Forest Model

Figure 5.6 represents the initialization and prediction of the Random Forest model showing its accuracy, precision, recall, and F1-score. Additionally, below figure 5.7 illustrates the confusion matrix of the model.

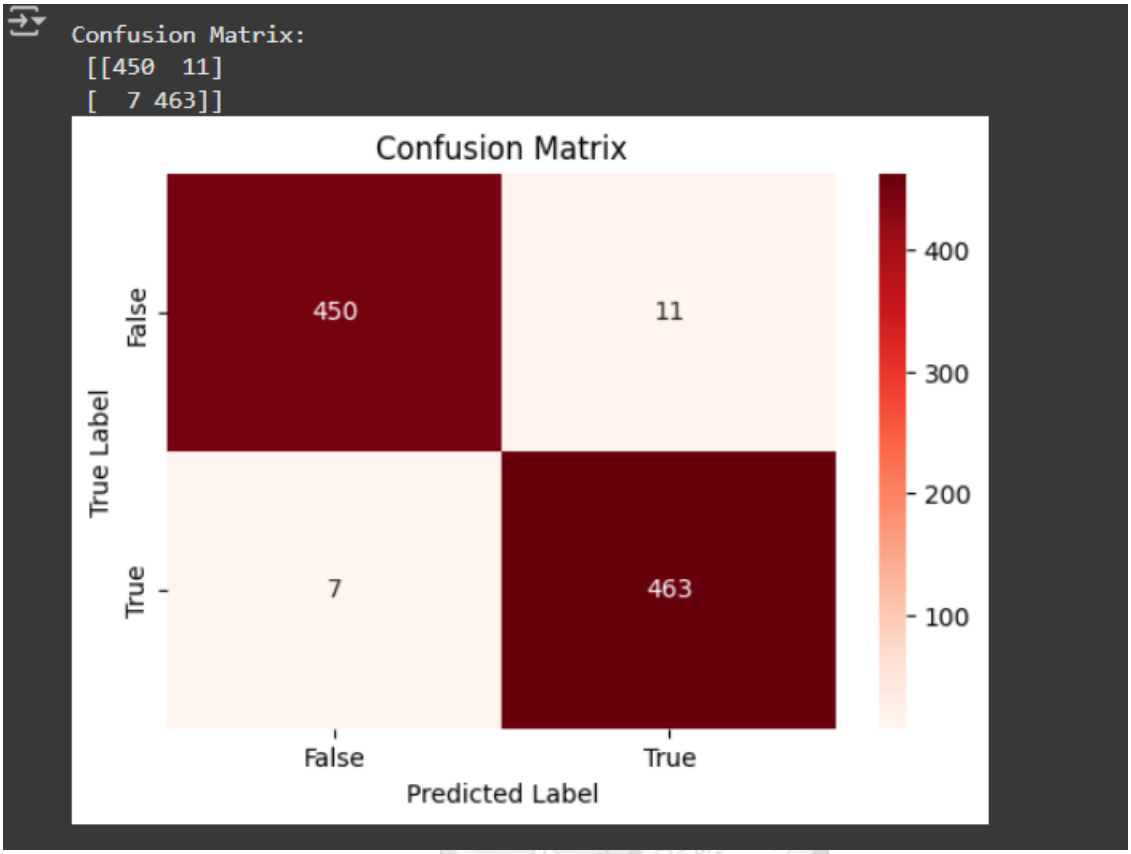


Figure 5.7: Random Forest Model Confusion Matrix

5.2.5.3 Support Vector Machine (SVM) Model

SVM is a supervised learning algorithm used for classification and regression. It finds the optimal hyperplane that maximizes the margin between different classes. For non-linearly separable data, SVM uses kernel functions to project data into a higher-dimensional space, making it effective for complex classification tasks.

Support Vector Machine algorithm was noted to have outperformed all the other algorithms achieving and accuracy of ~0.9968. Cross-validation of both 5-fold and 10-fold including hyperparameter tuning were carried out on this model to ensure that the model was not in any way outfitting or memorizing on the training data, and the results were consistent with the overall accuracy. Figures 5.8, and 5.9 below represent the prediction of the SVM model and Confusion Matrix respectively.

```python
# Train SVM model
svm = SVC(kernel="linear")
svm.fit(X_train, y_train)

svm_preds = svm.predict(X_test)

# Get training accuracy for SVM
svm_train_acc = svm.score(X_train, y_train)
svm_test_acc = svm.score(X_test, y_test)

# Perform 10-fold cross-validation
svm_cv_scores = cross_val_score(svm, X_train, y_train, cv=10)

# Evaluate Support Vestor Machine model
accuracy = accuracy_score(y_test, svm_preds)
print(f"SVM Model Accuracy: {accuracy:.4f}")
print("Classification Report:\n", classification_report(y_test, svm_preds))
print("Confusion Matrix:\n", confusion_matrix(y_test, svm_preds))
```

```
SVM Model Accuracy: 0.9968
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       461
           1       1.00      1.00      1.00       470

    accuracy                           1.00       931
   macro avg       1.00      1.00      1.00       931
weighted avg       1.00      1.00      1.00       931
```

Figure 5.8: Initialization and Prediction Using SVM Model

```
Confusion Matrix:
 [[460   1]
 [  2 468]]
```

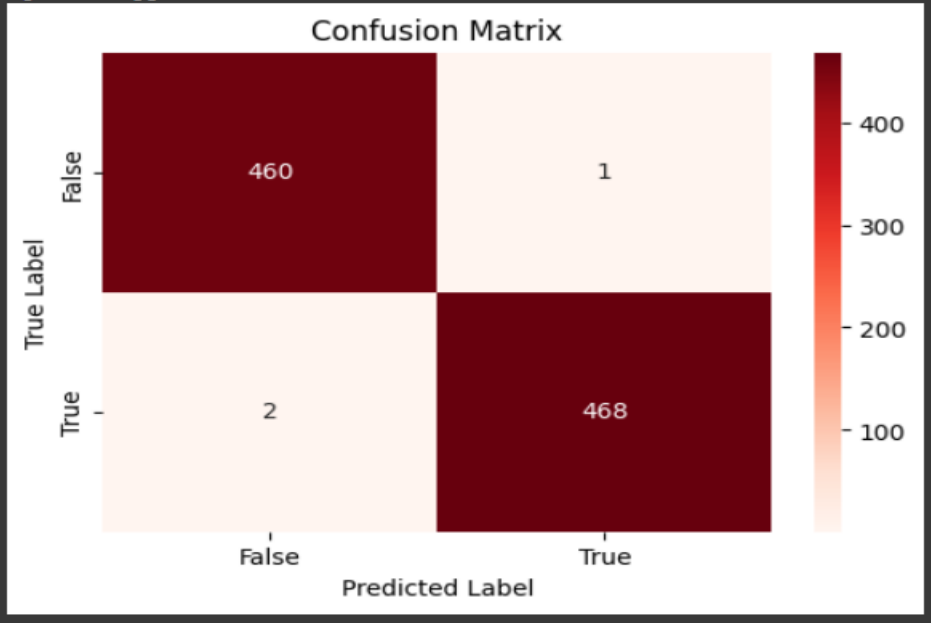


Figure 5.9: SVM Confusion Matrix

5.2.5.4 K-Nearest Neighbors (KNN) Model

KNN is a simple, instance-based learning algorithm that classifies data points based on the majority class of their *k* nearest neighbors. It calculates the distance (e.g., Euclidean distance) between data points and assigns the most common class among the nearest *k* neighbors. KNN is useful for classification and regression but can be computationally expensive for large datasets.

As defined while using the SVM, DT, RF algorithms, KNN was also trained and tested using 80/20 dataset split. The model's performance was also evaluated using accuracy, precision, recall and F1-score metrics and Confusion Matrix as well. The KNN model performed well as well, achieving a high accuracy of ~0.9871, with ~0.99 (precision), ~0.99 (recall), and ~0.99 (F1-score). Figures 5.10, and 5.11 below represent prediction performance scores and Confusion Matrix of KNN model respectively.

```
# Train Decision Tree model
knn = KNeighborsClassifier()
knn.fit(X_train, y_train)

knn_preds = knn.predict(X_test)

# Get training accuracy for KNN
knn_train_acc = knn.score(X_train, y_train)
knn_test_acc = knn.score(X_test, y_test)

# Perform 5-fold cross-validation
knn_cv_scores = cross_val_score(knn, X_train, y_train, cv=5)

# Evaluate KNN model
accuracy = accuracy_score(y_test, knn_preds)
print(f"KNN Model Accuracy: {accuracy:.4f}")
print("Classification Report:\n", classification_report(y_test, knn_preds))
print("Confusion Matrix:\n", confusion_matrix(y_test, knn_preds))
```

```
KNN Model Accuracy: 0.9871
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.98      0.99       461
           1       0.98      1.00      0.99       470

    accuracy                           0.99       931
   macro avg       0.99      0.99      0.99       931
weighted avg       0.99      0.99      0.99       931
```

Figure 5.10: Initialization and Prediction Using KNN Model

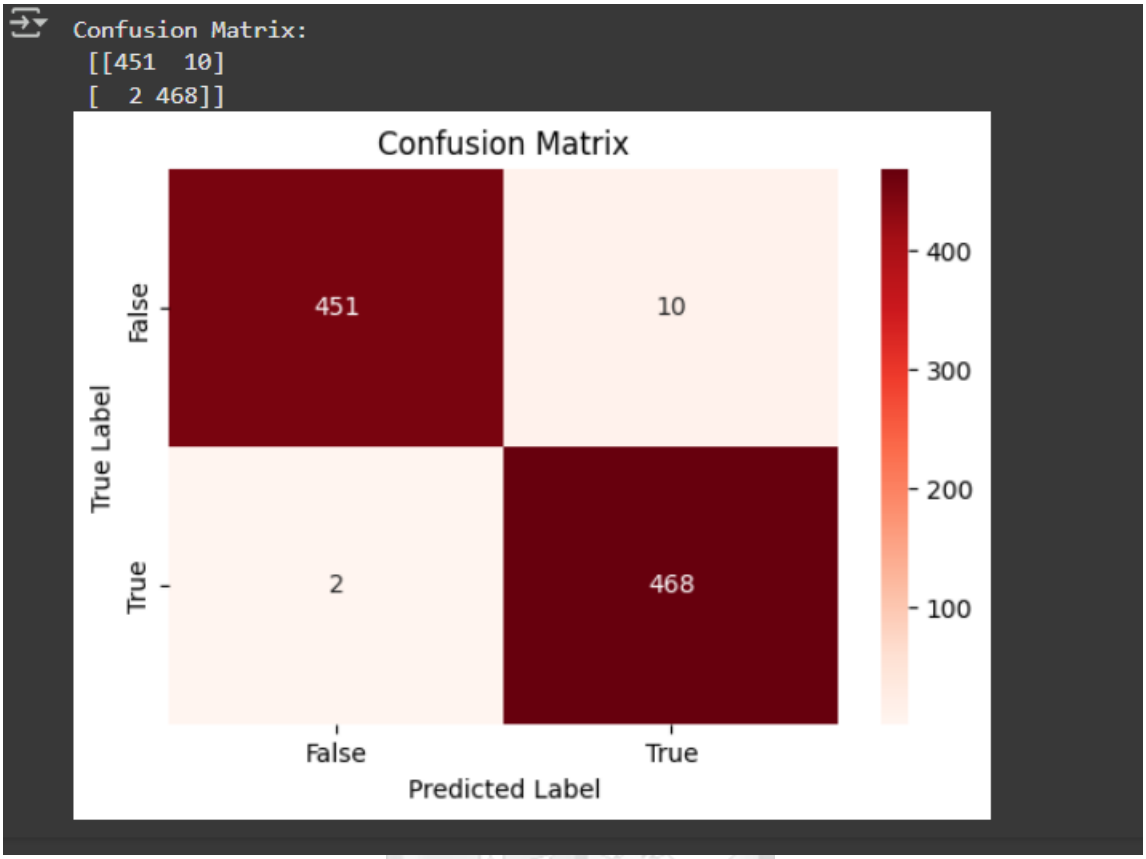


Figure 5.11: KNN Model Confusion Matrix

5.2.5.5 Artificial Neural Network (ANN) Model

ANN is a computational model inspired by the human brain, consisting of layers of interconnected neurons. It includes input, hidden, and output layers, with each neuron applying activation functions to transform inputs. ANNs are used for complex pattern recognition, classification, and regression tasks, making them fundamental to deep learning applications.

The ANN model was applied in this case to predict the classifications of the individuals who are at high risk of substance abuse relapse. The classification model was trained using the Multilayer Perception (MLP) algorithm, the MLPClassifier is used to build the model. The model was built with two hidden layers consisting of 64 and 32 neurons, with the max_iter set at 500 meaning the number of maximum iterations (epochs) the model will undergo during training. The figures below 5.12, and 5.13 depict initialization and training of ANN model and the Confusion Matrix respectively. The model achieved an accuracy of ~0.9774.

```python
# Train ANN model
ann = MLPClassifier(hidden_layer_sizes=(64, 32), activation='relu', solver='adam', max_iter=500, random_state=42)
ann.fit(X_train, y_train)

ann_preds = ann.predict(X_test)

# Get training accuracy for ANN
ann_train_acc = ann.score(X_train, y_train)
ann_test_acc = ann.score(X_test, y_test)

# Performing 10-fold cross-validation
ann_cv_scores = cross_val_score(ann, X_train, y_train, cv=10)

# Evaluate ANN model
accuracy = accuracy_score(y_test, ann_preds)
print(f"ANN Model Accuracy: {accuracy:.4f}")
print("Classification Report:\n", classification_report(y_test, ann_preds))
print("Confusion Matrix:\n", confusion_matrix(y_test, ann_preds))
```

```
ANN Model Accuracy: 0.9774
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.96      0.98       461
           1       0.96      0.99      0.98       470

    accuracy                           0.98       931
   macro avg       0.98      0.98      0.98       931
weighted avg       0.98      0.98      0.98       931
```

Figure 5.12: Initialization and Prediction Using KNN Model

```
Confusion Matrix:
 [[443  18]
 [  3 467]]
```

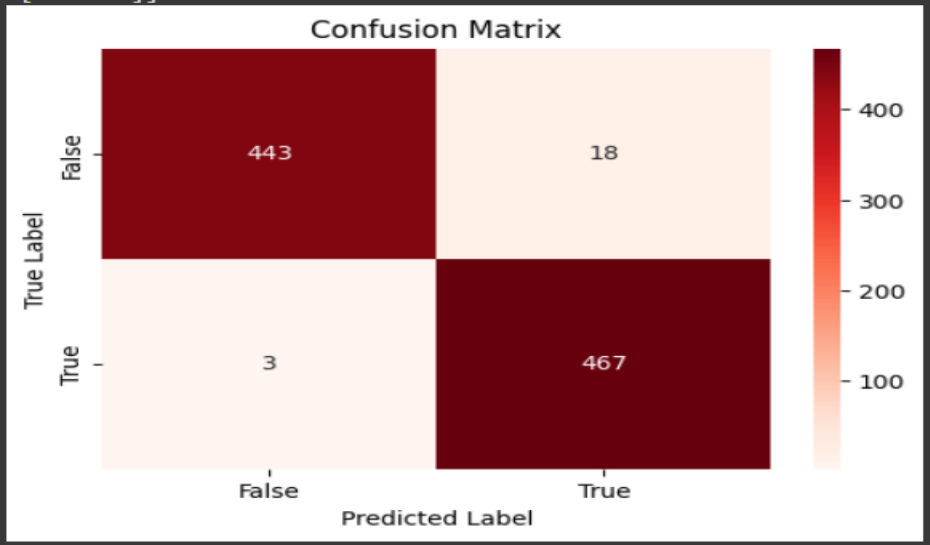


Figure 5.13: ANN Model Confusion Matrix

5.3 Models' Findings and Results.

As has been illustrated above, five different models namely Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Artificial Neural Network (ANN) were evaluated for the classification task. The evaluation results below on table 5.4 gives a comparison of the performance of the models. Based on these results, all the models achieved quite high accuracy, or performance shows that the dataset used was balanced, and noise free and applied normalization helped the models perform better.

Model	Accuracy	Precision	Recall	F1-score
Decision Tree	0.96	0.96	0.96	0.96
Random Forest	0.98	0.98	0.98	0.98
Support Vector Machine	0.99	1	1	1
K-Nearest Neighbors	0.98	0.99	0.99	0.99
Artificial Neural Network	0.97	0.98	0.98	0.98

Table 5.7: Trained Models Performance

5.4 Adopted Model for System Deployment

From the shared performance of the models as illustrated on table 5.7, SVM model showed to have the highest performance compared to other models followed by KNN, Random Forest, ANN and Decision Tree respectively. However, the overall results of all the models show that they all performed well as they depict slightly lower accuracy difference from each other.

For the immense advantages that ANN model present compared to the other models, despite achieving slightly lower accuracy compared to Random Forest, SVM and KNN achieving accuracy of ~0.9774, this model was adopted for the implementation and deployment of this research. The main reasons ANN was chosen is because of its robustness, adaptability and scalability in the real-world compared to the other models, ANNs can handle real-time, continuous learning when deployed, adapting to new trends over time.

5.5 Substance Abuse Relapse Prediction User Interface

For the deployment of the developed model, a web-based user interface was developed to allow end-users to make predictions through the website. To access the prediction system, users have to register or login to the portal. The users can create the new account by signing up using their google account or creating the account by filling in relevant information like username and password. Figure 5.14 below shows the registration and login interface.

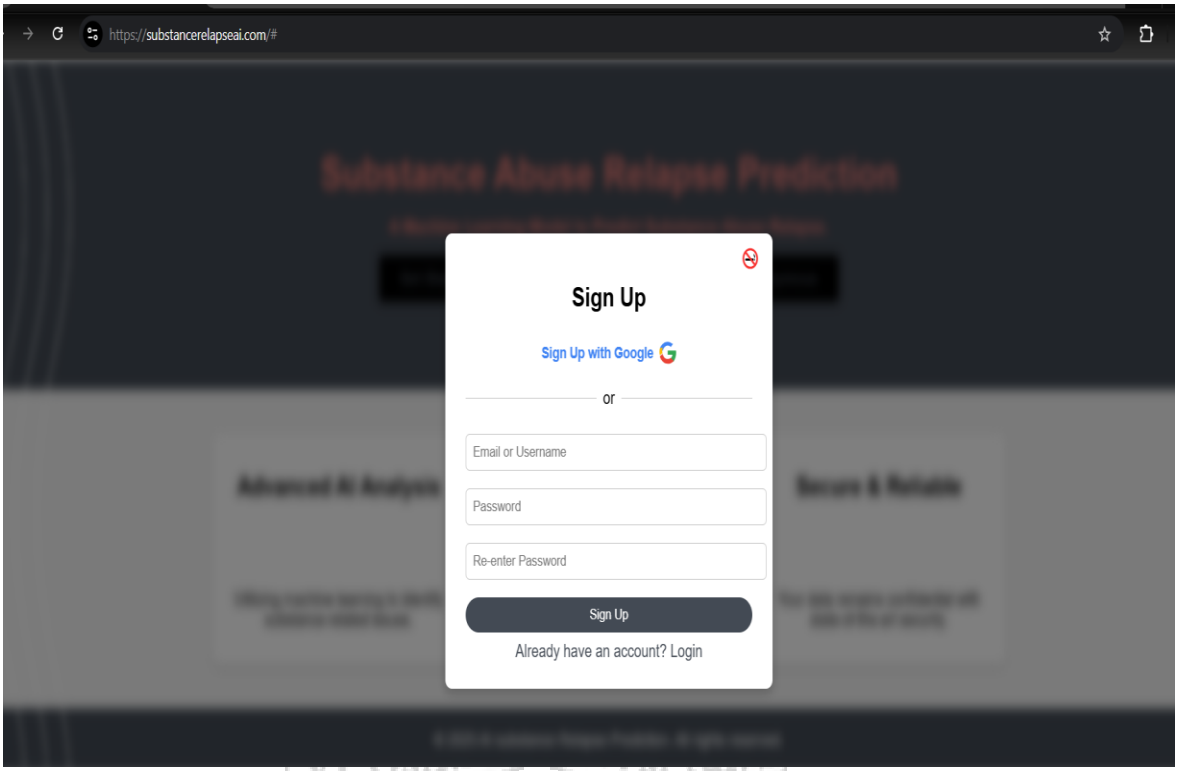


Figure 5.14: Login/Sign Up Interface

Once the user is logged in, they can access the user dashboard screen. From here, the user/s are able to make new predictions or have a view of the history of previously made predictions. Also on this dashboard screen, the user/s can view their profile and make any changes to the account/profile if needed.

Under the prediction interface (New prediction), the user/s is required to input or select the appropriate data in order to for the prediction model to output or generate the probability of relapse. Figure 5.16 below depicts the prediction interface.

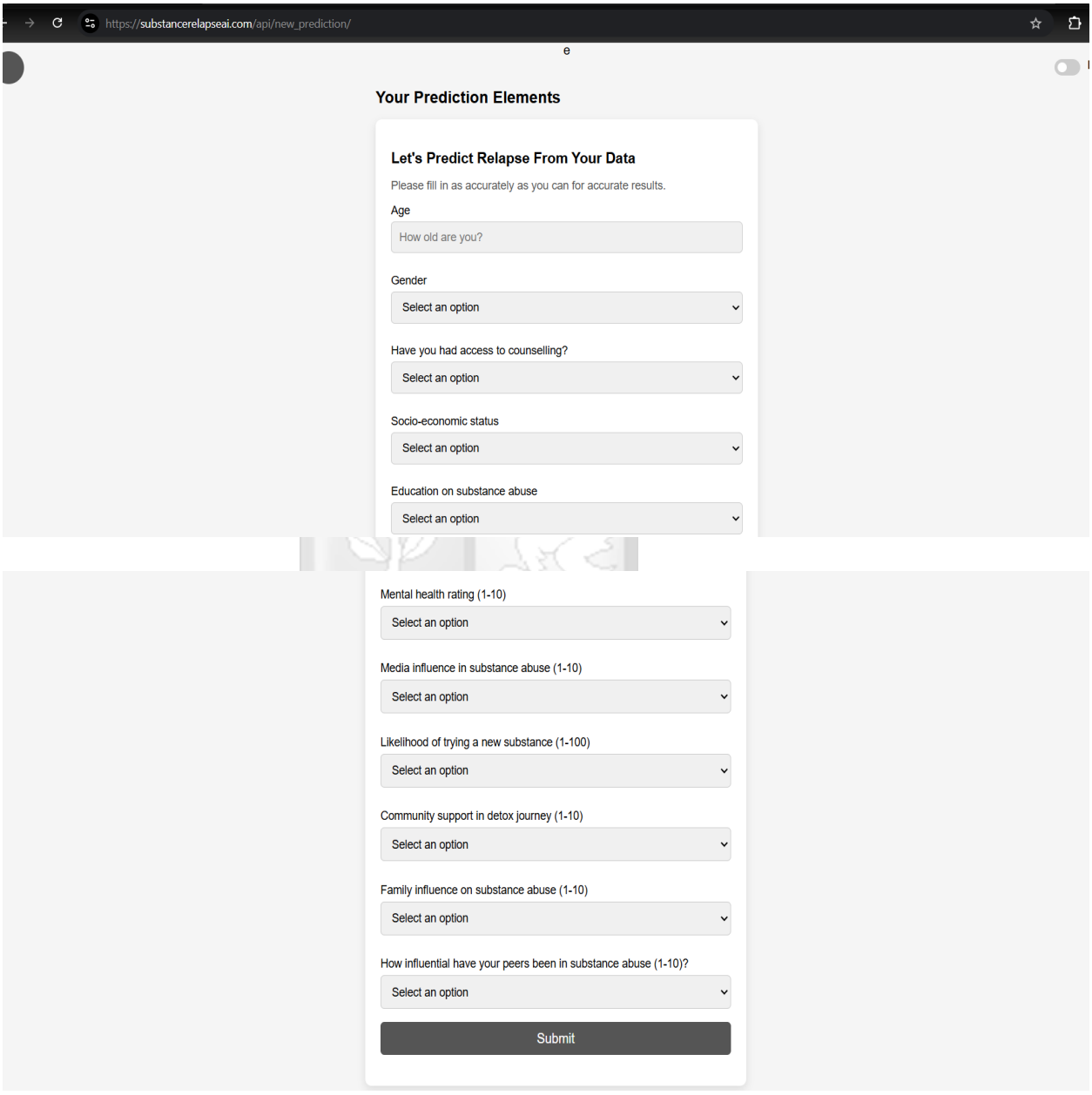


Figure 5.15: Input User Interface

For the prediction history capability, the interface shows list of previously made predictions showing whether the individual/patient is likely to relapse from substance abuse after recovery or while undergoing recovery. Figure 5.17 shows prediction history interface.

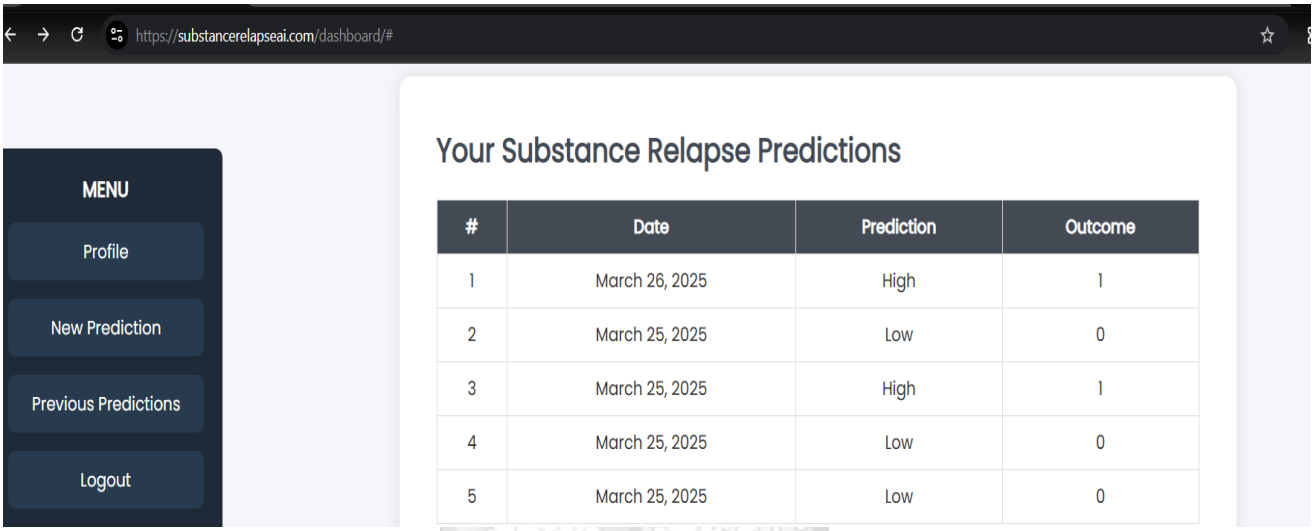


Figure 5.16: Prediction History Interface

Chapter 6: Discussions

6.1 Introduction

Discussions provide an overview of the significant findings and outcomes identified throughout this research. From literature review all the way through to design and implementation of this research, going through the set specific objectives of this research to ensure that they were met.

6.2 Identification of factors influencing relapse on use of substances

Through literature review, several determinants were identified for the influence of substance use relapse. From individual-specific determinants, familial determinants(background), social determinants, behavioral determinants, and economic determinants**.**

As per research by Kabisa et al. (2021), Income level plays a major role in recovery outcomes. Higher family income is often linked to better access to healthcare services, including addiction treatment programs and mental health support. On the other hand, lower income levels can lead to increased stressors, such as housing instability and food insecurity, which may drive individuals back to substance use. It was also confirmed by NACADA (2024) that lack of social or community support can significantly hinder recovery efforts, moreover Sohrabpour et al., (2024) established that Individuals who do not have a robust support system, whether from family, friends, or community resources, may struggle more with coping mechanisms when faced with stressors or triggers.

Additionally, it was confirmed from the literature review that family background has significant impact on substance use relapse. According to a study by Mousali et al. (2021) indicates that the presence of a first-degree family member with a substance use disorder significantly elevates the risk, with some research suggesting an eight-fold increase in risk among relatives of individuals with drug disorders across various substances.

These findings from this determinant's literature review contributed to overall research as this helped in determining the best dataset featuring during the data collection process these features for precise predictions.

6.3 Reviewing of existing approaches in prediction of substance abuse relapse on Nairobi County

This objective aimed at identifying traditional methods used to predict substance abuse relapse in Nairobi County to be specific. It was identified during the literature review that Mindfulness Cognitive Behavioral Therapy, biopsychosocial determinants, community support and holistic approaches are used to determine if the individual is at risk of relapse. For example, according to the research conducted by Mwove (2019), the study identified that biopsychosocial determinants have contributed to the substance abuse relapse prediction in Nairobi County.

These approaches have proved yes, traditional approaches can be used to predict substance abuse relapse, but these approaches are not effective enough as they are more reactive than proactive, whereby it's likely that an individual is termed at risk of relapse when they have already relapsed. Therefore, the need to use modern technologies like AI that leverage on historic data to give proactive predictions and intervention.

6.4 Reviewing existing machine learning models and algorithms that have been used in the prediction of relapse in substance abuse

Several machine learning models and algorithms have been developed and used in the fields related and close to substance abuse relapse but not specifically like the case in this research. For instance, Park et al. (2021) used the Support Vector Machine (SVM) algorithm to develop a machine learning model to predict the dropping out of outpatients with alcohol use disorders, achieved an accuracy of 70.23%. On the other hand, Bharathidason & Sujdha, (2023) used and adopted random forest algorithm in the development of a model to predict drug addiction. Random forest was adopted for this study because it had achieved the highest accuracy of 90.99% compared to other algorithms.

For this research, KNN, SVM, Decision Trees, Random Forest and ANN algorithms were all trained for the classification task. All these models achieved high accuracies compared to the previous research done. As indicated on the above explanation under section 5.4, the ANN model was adopted because of the advantages it has on complex and unstructured data like medical data and also looking at the future utilization of the model.

6.5 Developing a machine learning model that will assist with the prediction of relapse into substance abuse

The main objective of this research was to develop a machine learning model for the prediction of substance use relapse. This objective was achieved by identifying the factors influencing substance abuse relapse, which helped in drawing the features necessary or crucial to predict relapse. A dataset was obtained containing the identified features: Family background, Substance education, peer pressure, socioeconomic status, mental health, access to counselling, community support, and model training was done using different algorithms. Through this process, the ANN model adopted for the implementation of this research/solution.

6.6 Validation of the developed Model

This objective was achieved using test data which was set for testing. The main reason for the evaluation of the model is to ensure that it performs well in the real world, on unseen data. Using the performance metrics precision, recall, and f1-score the ANN model was noted to perform well with precision rate of 98%, recall rate of 98% and f1-score rate of 98% as well. This showed that the ANN model can accurately and effectively identify individuals at high risk of relapse.

6.7 Contributions of the Developed Model

The developed substance abuse relapse prediction system during this research will make a significant contribution to the management of substance abuse recovery to rehabilitation centers. The system will provide early identification of individuals at high risk of relapse, allowing for early interventions by rehabilitation centers, counselors and healthcare providers to prevent eventual relapse. Additionally, the developed systems can be leveraged to better utilize rehabilitation resources whereby the system will help the rehabilitation centers to optimize their efforts by focusing on individuals most likely to relapse hence reducing strain on the facilities leading to better service delivery.

Through the system, data-driven policy making will be promoted as insights gained from the predictive model can be used to craft evidence-based interventions for long-term substance abuse relapse prevention. Data trends can help to monitor relapse patterns over time, allowing authorities and organizations to adjust policies dynamically to address emerging challenges. Moreover, integrating predictive analytics into policy frameworks enables proactive intervention strategies instead of reacting to relapse cases after they have already occurred.

6.8 Shortfalls of Research/Model

Despite being able to achieve high performance on the developed model, the research did not exhaust on all the possible features which contribute substance abuse relapse given that secondary data was utilized. The researcher had anticipated using more features on the model to broaden on the factors influencing substance abuse relapse, however there were limitations on the access of data since most medical related data is rarely publicly accessible, and the primary data collection option was not viable.

Chapter 7: Conclusions, Recommendations and Future Work

7.1 Conclusions

The main objective of this research was focused on developing a machine learning model that can predict substance abuse relapse. This objective was further divided into five specific objectives to adequately achieve the main objective. These specific objectives were as follows:

To identify the factors influencing relapse on the use of substances, to review the existing approaches in the prediction of substance abuse relapse in Nairobi County and to review existing machine learning models and algorithms that have been used in the prediction of relapse in substance abuse: These three specific objectives were achieved through conducting literature review from existing researches/studies. Identifying factors influencing relapse on the use of substance, reviewing existing approaches in the prediction of substance abuse relapse in Nairobi County and also reviewing the existing machine learning models and algorithms helped in understanding the concept of substance use relapse.

To develop a machine learning model that will assist with the prediction of relapse into substance abuse: To achieve this objective, secondary data was collected from an open-source publicly available platform, Kaggle. The dataset contained features which crucially contribute to the relapse of the use of substance. The dataset was preprocessed and split into both training and testing sets. The training set was used in training the model, and the testing set was used in model testing. Chapter 4 and 5 fully explored how this objective was met.

To validate the developed model: The last objective of this research was to validate the developed model. This objective was achieved by using evaluation metrics such accuracy, precision, recall and F1 Score. These metrics are crucial in assessing the model's performance and also to compare different models or algorithms in order to choose the best one for the specific purpose.

7.2 Recommendations

This research demonstrates that machine learning can be used to predict substance abuse relapse among recovering or recovered individuals instead of using traditional methods. It will be helpful in detecting patients at risk of relapse on the use of substance and provide the rehabilitation centers with opportunities to provide prompt interventions. Stakeholders from the rehabilitation centers within Nairobi County can utilize this tool/system to provide proactive and effective response to patients at risk of relapse.

As a result of the outcome obtained from implementation and testing of this research, below are some of the recommendations made:

The research used a secondary dataset which could have limited features compared to if the research had utilized primary datasets. The research recommends implementation of the model using primary datasets as this would exhaust the features relating to substance abuse relapse.

7.3 Suggestions for Future Work

For future work based on the findings drawn by the researcher, the suggestions below are made:

Expanding the data sources: Future research should focus on collecting more comprehensive data from diverse sources such as hospitals, rehabilitation centers, and community support groups. Partnering with these institutions will help address data gaps and biases while improving the model's generalizability.

Personalized and Adaptive Intervention Systems – Machine learning models should not only predict relapse but also recommend personalized interventions. Future research could focus on dynamic treatment plans, where individuals receive customized counseling, therapy, or community support based on real-time risk levels.

Longitudinal Studies and Continuous Model Improvement – To improve accuracy and long-term effectiveness, future work should involve long-term tracking of individuals in recovery and periodic model retraining based on updated data.

References

Afkar, A., Rezvani, S. M & Sigaroudi, E.A (2017). Measurement of Factors Influencing the Relapse of Addiction: A Factor Analysis. International Journal High Risk Behavior Addict, 6(3).

Al-Shagran, H., Al-Afeef, M., & Al-Hasanat, M. (2022). Predicting substance abuse relapse using machine learning algorithms. Journal of Addiction Research and Therapy, 13(2), 1-10.

American Editorial (2024). Addiction Relapse: Risk Factors, Coping & Treatment Options. https://americanaddictioncenters.org/treat-drug-relapse.

Anundo, J.A., Muaka, C.A & Ongaro, K (2022). A Comparative Study on Effectiveness of Mindfulness Cognitive Behavior Therapy and 12-Steps Model on Relapse Prevention Among Persons with Substance Use Disorder in Selected Rehabilitation Centers in Nairobi and Kajiado Counties in Kenya.4,3.

Bharathidason, S & Sujdha, C (2023). A Comparative Study of Machine Learning Algorithms for Drug Addiction Prediction. Journal of Harbin Engineering University. 44(12), 1006-7043.

Creswell, J. W & Creswell, J. D. (2018). Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.). SAGE Publications.

DrugFree (2024). When Addiction is in Your Family Tree. https://drugfree.org/article/when-addiction-is-in-your-family-tree/.

Ebrahimi, A., Wiil, U. K., Schmidt, T., Naemi, A., Nielsen, A. S., Shaikh, G. M., & Mansourvar, M. (2021). Predicting the Risk of Alcohol Use Disorder Using Machine Learning: A Systematic Literature Review. IEEE Access, 9, 151697-151712

Grant, J. E & Chamberlain, S. R (2020). Family history of substance use disorders: Significance for mental health in young adults who gamble. J Behav Addict. 9(2), 287-289.

Hosseinzadeh, A., Mansoori, P., & Heidari, M. (2021). Machine learning techniques for drug addiction relapse prediction: A systematic review. Journal of Substance Use, 26(6), 628-635.

Ibitoye, O.J., A & Nwosu, O.C. (2021). A MACHINE LEARNING MODEL FOR SOBRIETY AND RELAPSE ANALYSIS IN DRUG REHABILITATION. International Journal of Information System and Computer Science.

Islam, A., Rahman, S.A., Habib, T.M & Islam, A.A (2021). Prediction of addiction to drugs and alcohol using machine learning: A case study on Bangladeshi population. International Journal of Electrical and Computer Engineering. 11(5),4471-4480

Islam, U.I., Haque, E., Alsalman, D., Islam, M.N., Moni, M.A & Sarker, I.H(2022). A machine learning model for predicting individual substance abuse with associated risk-factors. In: Annals of Data Science. 10(6),1607-1634.

Kabisa, E., Biracyaza, E., Habagusenga, J., & Umubyeyi, A. (2021). Determinants and prevalence of relapse among patients with substance use disorders: case of Icyizere Psychotherapeutic Centre, 16,13.

Kuria, M.W (2013). Factor associated with relapse and remission of alcohol dependent persons after community-based treatment. Open Journal of Psychiatry. 3,2.

Kuyeya Fatuma (2021). Effectiveness of Treatment and Rehabilitation Programs for Drug and Substance Dependence in Mombasa County, Kenya. African Journal of Alcohol and Drug Abuse, 6.

Lai, C.C., Huang, W.H., Chang, B.C.-C.& Hwang, L.C (2021). Development of Machine Learning Models for Prediction of Smoking Cessation Outcome. Int. J. Environ.Res. Public Health 2021, 18, 2584.

Luft, A.J., Jeong,Sophia., Idsardi,R & Gardner,Grant (2022). Literature Reviews, Theoretical Frameworks, and Conceptual Frameworks: An Introduction for New Biology Education Researchers. 32,3.

Magill, M., Ray, L., Kiluk, B., Hoadley, A., Bernstein, M., Tonigan, S.J & Carroll, K (2019). A meta-Analysis of Cognitive-Behavioral Therapy for Alcohol or Other Drug Use Disorders: Treatment Efficacy by Contrast Condition. 87()12, 1093-1105.

Mahesh, B (2020). Machine Learning Algorithms – A Review. International Journal of Science and Research (IJSR). 2319-7064.

Malik, N.I., Saleem, S., Ullah, I., Rehan, S.T., De Berardis, D., & Atta, M (2023). Psychosocial Factors Affecting Drug Relapse among Youth in Punjab, Pakistan. J. Clin. Med. 12, 2686.

Ministry of Health- MOH (2017). The National Protocol for Treatment of Substance Use Disorders in Kenya 2017. https://www.afro.who.int/sites/default/files/2017-09/The%20National%20Protocol%20for%20treatments%2014%2007%202017.pdf.

Mousali, A.A., Bashirian, S., Barati, M., Mohammadi, Y., Moeini, B., et al (2021). Factors affecting substance use relapse among Iranian addicts, 10(129).

Mukora, W., M & Mbugua, M., R. (2023). Determinants of relapse among alcohol and drug abusers in selected rehabilitation centers in Nairobi County, Kenya. International Journal of Research in Social Sciences, 13,8 (2023)

Mwove, M., P (2019). Biopsychosocial Determinants of Relapse among Patients diagnosed with Substance related disorders at Mathari National Teaching and Referral Hospital, Kenya.

NACADA (2024). Does Alcohol and Drug Addiction Run in Your Family? https://nacada.go.ke/does-alcohol-and-drug-addiction-run-your-family.

National Authority for the Campaign Against Alcohol and Drug Abuse (NACADA). (2022). National Survey on Alcohol and Drug Abuse. https://nacada.go.ke/sites/default/files/2023-05/National%20Survey%20on%20the%20Status%20of%20Drugs%2

Nyabuti, S., & Mohammed, M. (2024). AI's Potential in The Fight Against Alcohol and Drug Abuse. https://nacada.go.ke/ais-potential-fight-against-alcohol-and-drug-abuse.

Praba. R., Darshan. G., Roshanraj, K. T & Surya, P. B (2021). Study On Machine Learning Algorithms. International Journal of Scientific Research in Computer Science, Engineering and Information Technology 7(4), 67-72.

Price, G. (2024). Family Footprints: The Invisible Risk Factors in Substance Addiction. https://www.virtuerecoverykilleen.com/rehab-blog/family-history-of-substance-addiction-risk.

Sarker, I.H (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science (2021) 2:160

Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., Burroughs, H., & Jinks, C. (2017). Saturation in qualitative research: exploring its conceptualization and operationalization. Quality & Quantity, 52(4), 1893–1907.

Sohrabpour, M., Kamyab, A., Yari, A., Harsini, P.A & Jeihooni, A.K (2024). The factors affecting substance abuse relapse based on theory of planned behavior in male addicts covered by addiction treatment centers in Southern Iran. BMC Public Health, 24(1265).

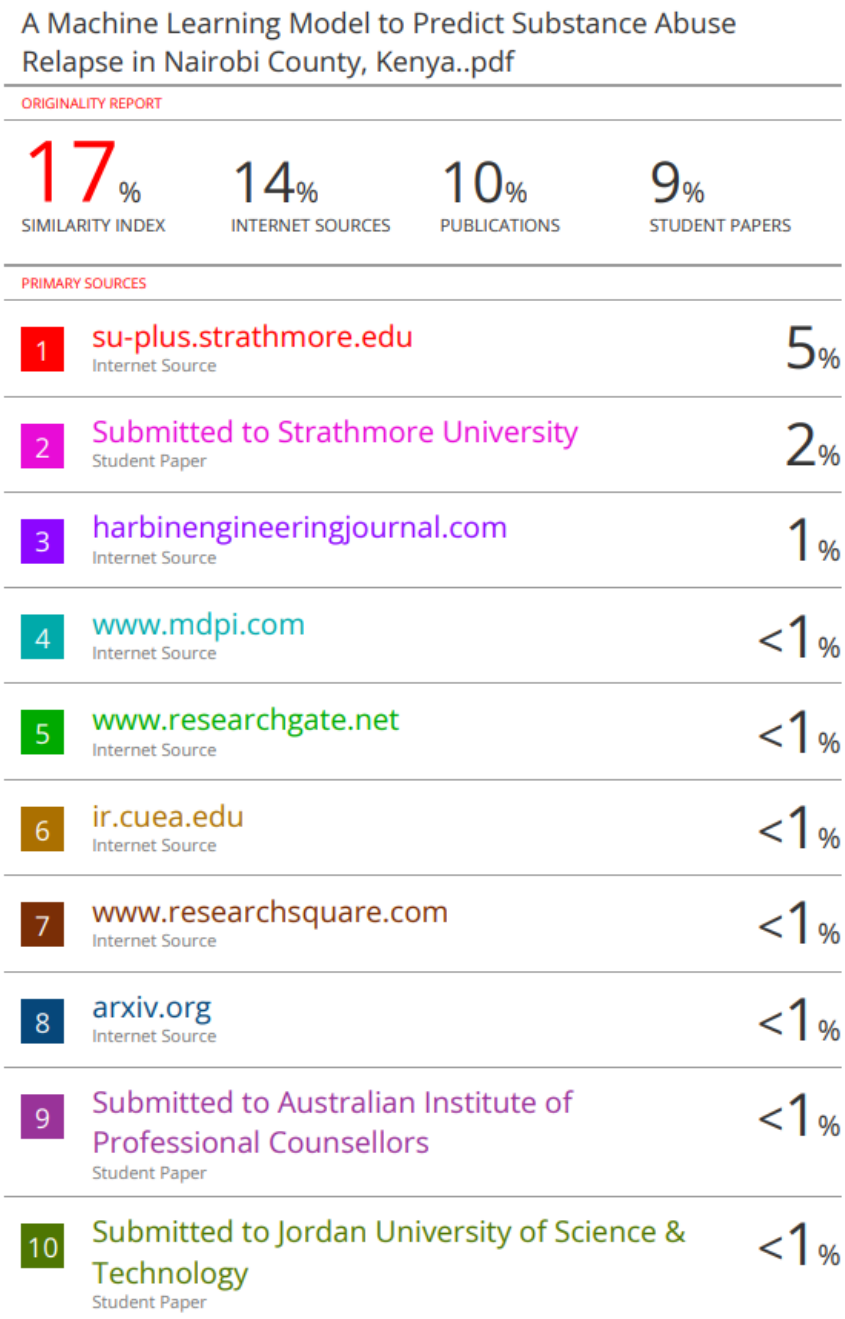
Wainaina, N.V(2020). Factors Influencing Alcohol Relapse Among Patients in Alcohol and Substance Abuse Treatment and Rehabilitation Programme (ASATREP) in Kiambu County, Kenya.

Xin, L., Jisheng, X., Zheng, T.Z., Qiuyue, H., Zhicheng, Z., Hui, W. & Xue, L. (2023) Prediction of Relapse Risk Among Chinese Drug Abusers Based on Interpretable Machine Learning Technology, Journal of Substance Abuse Treatment, 145, 108928.

Zhang, Y., & Sun, X. (2023). Ethical challenges in applying machine learning to substance abuse prediction. Healthcare Informatics Journal, 30(1), 87-98.

Appendices

Appendix A: Originality Report



A Machine Learning Model to Predict Substance Abuse Relapse in Nairobi County, Kenya..pdf

ORIGINALITY REPORT

17%	14%	10%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	su-plus.strathmore.edu Internet Source	5%
2	Submitted to Strathmore University Student Paper	2%
3	harbinengineeringjournal.com Internet Source	1%
4	www.mdpi.com Internet Source	<1%
5	www.researchgate.net Internet Source	<1%
6	ir.cuea.edu Internet Source	<1%
7	www.researchsquare.com Internet Source	<1%
8	arxiv.org Internet Source	<1%
9	Submitted to Australian Institute of Professional Counsellors Student Paper	<1%
10	Submitted to Jordan University of Science & Technology Student Paper	<1%

Appendix B: Ethical Clearance Confirmation

Strathmore UNIVERSITY

3rd February 2025

Ms Jepchirchir Emmy,
emmy.jepchirchir@strathmore.edu

Dear Ms Jepchirchir,

RE: A Machine Learning Model to Predict Substance Abuse Relapse in Nairobi County, Kenya

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2608/25.** The approval period is from **3rd February 2025 to 2nd February 2026.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu