

# Predicting Likelihood of Depression using Machine Learning

---

Tristan Dos Remendos

*Supervisor(s):*  
Seun Olukanmi



A research proposal submitted in partial fulfillment of the requirements for the  
degree of Bsc in Computer Science Honours

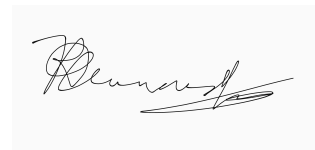
in the

School of Computer Science and Applied Mathematics  
University of the Witwatersrand, Johannesburg

31 May 2024

# Declaration

I, Tristan Dos Remendos, declare that this proposal is my own, unaided work. It is being submitted for the degree of Bsc in Computer Science Honours at the University of the Witwatersrand, Johannesburg. It has not been submitted for any degree or examination at any other university.

A handwritten signature in black ink, appearing to read 'Tristan Dos Remendos', is displayed within a light gray rectangular box.

Tristan Dos Remendos

31 May 2024

## *Abstract*

Depression stands as a significant public health challenge, contributing substantially to disability rates globally. Traditional approaches to monitoring and surveillance of depression often face delays and resource constraints, hindering timely interventions. In South Africa, research on the prevalence and determinants of depression remains limited, posing challenges for designing targeted interventions and allocating resources effectively. This research aims to address these gaps by leveraging machine learning models and socioeconomic data from the National Income Dynamics Study to predict the likelihood of depression among individuals in South Africa. The proposed methodology involves preprocessing and preparing the National Income Dynamics Study dataset, including labeling participants for depression using the Center for Epidemiologic Studies Depression scale. Various machine learning algorithms, such as logistic regression, random forests, deep neural networks, and support vector machines, will be explored and evaluated. The top-performing models will be rigorously validated and tuned to optimize their predictive capabilities. Additionally, the relative importance of different socioeconomic and demographic features in predicting depression likelihood will be analyzed. By providing a novel and potentially complementary approach to depression surveillance, this research aims to enhance our understanding of the socioeconomic factors associated with depression in the South African context. The findings may inform evidence-based strategies for mitigating the burden of depression and guide resource allocation for mental health interventions in the region.

# Acknowledgements

I would like to thank the University of the Witwatersrand for their continuous involvement in my academic and personal development as well as my supervisor Dr. Seun Olukanmi for her wisdom, guidance and unwavering support.

# Contents

<b>Declaration</b>	<b>i</b>
<b>Abstract</b>	<b>ii</b>
<b>Acknowledgements</b>	<b>iii</b>
<b>List of Figures</b>	<b>vi</b>
<b>List of Tables</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Related Work . . . . .	2
1.1.1 A history of Depression . . . . .	3
Historical Statistics for Depression . . . . .	3
Depression in South Africa . . . . .	4
1.1.2 Prediction and Diagnosis of Depression . . . . .	5
Traditional Approaches . . . . .	6
Novel Approaches . . . . .	7
Machine Learning for Depression Measures . . . . .	10
1.2 Problem Statement . . . . .	12
1.3 Research Question . . . . .	13
1.4 Research Aims and Objectives . . . . .	13
1.4.1 Research Aims . . . . .	13
1.4.2 Objectives . . . . .	13
1.5 Limitations . . . . .	14
1.6 Overview . . . . .	14
<b>2 Research Methodology</b>	<b>16</b>
2.1 Research design . . . . .	16

2.2	Data . . . . .	17
2.2.1	Data Collection . . . . .	17
	Data from the National Income Dynamics Study . . . . .	17
2.2.2	Data Preprocessing . . . . .	18
	Data labelling for Depression . . . . .	18
	Data Cleaning . . . . .	19
	Data Transformation and Feature Engineering . . . . .	19
2.2.3	Data Analysis . . . . .	20
2.3	Machine Learning Models . . . . .	20
2.3.1	Logistic Regression Model . . . . .	21
2.3.2	Random Forest model . . . . .	22
2.3.3	Deep Neural Network . . . . .	23
2.3.4	Support Vector Machine . . . . .	25
2.4	Analysis . . . . .	26
2.5	Limitations . . . . .	28
2.6	Ethical Considerations . . . . .	29
<b>3</b>	<b>Schedule of Work</b>	<b>30</b>
3.1	Schedule of Work . . . . .	30
3.2	Potential Difficulties . . . . .	31
<b>4</b>	<b>Conclusion</b>	<b>32</b>
	<b>Bibliography</b>	<b>34</b>

# List of Figures

1.1	Depression cases, by SDI regions, from 1990 to 2017. [38]	4
2.1	Simplification of random forest majority voting process	23
2.2	Visualisation of an arbitrarily large deep neural network	24

# List of Tables

1.1	Achievements of Different Model Types . . . . .	11
3.1	Schedule of work for blocks 3 and 4 . . . . .	31



# Chapter 1

## Introduction

Depression stands as a significant public health challenge globally, contributing substantially to disability rates [36]. With its increasing prevalence, effective surveillance strategies become imperative for monitoring trends, allocating resources, and informing interventions [4]. Traditional surveillance methods often encounter delays and resource constraints, necessitating innovative approaches for timely responses [29].

In recent years, the utilization of machine learning models in conjunction with large-scale datasets has emerged as a promising avenue for mental health surveillance [22]. Leveraging socio-economic and demographic features from comprehensive datasets provides a newer understanding of depression prevalence and its determinants. Despite the potential of such methodologies, research in this domain, particularly in regions like South Africa, remains scarce.

This research aims to address this gap by employing machine learning models to predict the likelihood of depression among individuals in South Africa. Utilizing the National Income Dynamics Study dataset, which longitudinally surveys around 30,000 South African individuals every 2-3 years [48], offers a strong foundation for analysis. By integrating socio-economic, demographic, and health-related variables, this study seeks to develop predictive models capable of identifying individuals at risk of depression.

The motivation for this research therefore stems from the scarcity of studies focusing on depression prediction in South Africa, despite its pressing public health implications [49]. By utilizing advanced analytical techniques and a rich dataset,

this research endeavors to advance our understanding of depression epidemiology in the region and contribute to the development of targeted interventions.

Through a comprehensive analysis of socio-economic and demographic factors associated with depression, this study aims to shed light on the unique determinants of mental health in South Africa. By improving our understanding of these factors, we can facilitate the design of tailored interventions and resource allocation strategies to mitigate the burden of depression in the population [22].

In the subsequent chapters, we will delve into the methodology, potential results, and implications of employing machine learning models for depression prediction in South Africa, offering insights into both the methodological advancements and the public health impact of this research study.

## **1.1 Related Work**

Depression, a pervasive mental health disorder, has garnered increasing attention in both historical discourse and contemporary medical research. Since ancient times, perceptions of depression have undergone significant transformations, reflecting evolving societal attitudes and advancements in medical understanding [7]. In recent decades, the recognition of depression as a leading cause of disability worldwide has increased the urgency of effective surveillance and intervention strategies [36].

The structure of the review will proceed as follows: firstly, an exploration of the historical evolution of depression, followed by an examination of the prevalence of depressive disorders globally, and in South Africa. Subsequently, the focus will shift to methodologies employed in depression surveillance, comparing traditional methods to novel approaches that have arisen due to the development of machine learning technologies. Finally, the review will conclude with a discussion of various statistical research methodologies commonly associated with public health surveillance and their efficacy when combined with internet-based sources.

### 1.1.1 A history of Depression

Depression has been recognised since ancient times, though the concepts and understanding of it have evolved considerably over history. Across various eras, depression saw changes in its perception ranging from a trait associated with genius and creativity, to imbalances of chemicals in the body. It was only after the 20th century, that the idea of depression became medicalised and treated as a serious health issue [7].

Around this time, the modern clinical concepts of depressive disorders began to take shape, with classifications developing in the Diagnostic and Statistical Manual (DSM) [2] and criteria for major depressive disorder, dysthymia, bipolar disorder and others emerging [7]. Biological treatments like antidepressant medications were also introduced during this period.

#### Historical Statistics for Depression

Depression is a leading cause of disability worldwide, affecting people of all ages, backgrounds, and socioeconomic statuses. Its incidence continues to rise, as the World Health Organization (WHO) calculated approximately 280 million people suffered from depressive disorder in 2019 [36].

Apart from the significant impact of depression alone, mental health disorders collectively affect populations worldwide. According to the WHO, around 970 million people globally live with various mental health conditions [36]. This highlights the widespread prevalence and urgency for scientific study and targeted interventions to address the challenges posed by mental illness on a global level.

An article in the Journal of Psychiatric Research assessed the age-standardized incidence rate (ASR) of depression in over 190 countries [38]. The data used in the study was sourced from the Global Health Data Exchange, a website that provides a comprehensive database for all health-related data.

SDI refers to the sociodemographic index, which uses information on the economy, education, and fertility rate of different countries to represent their social/economic

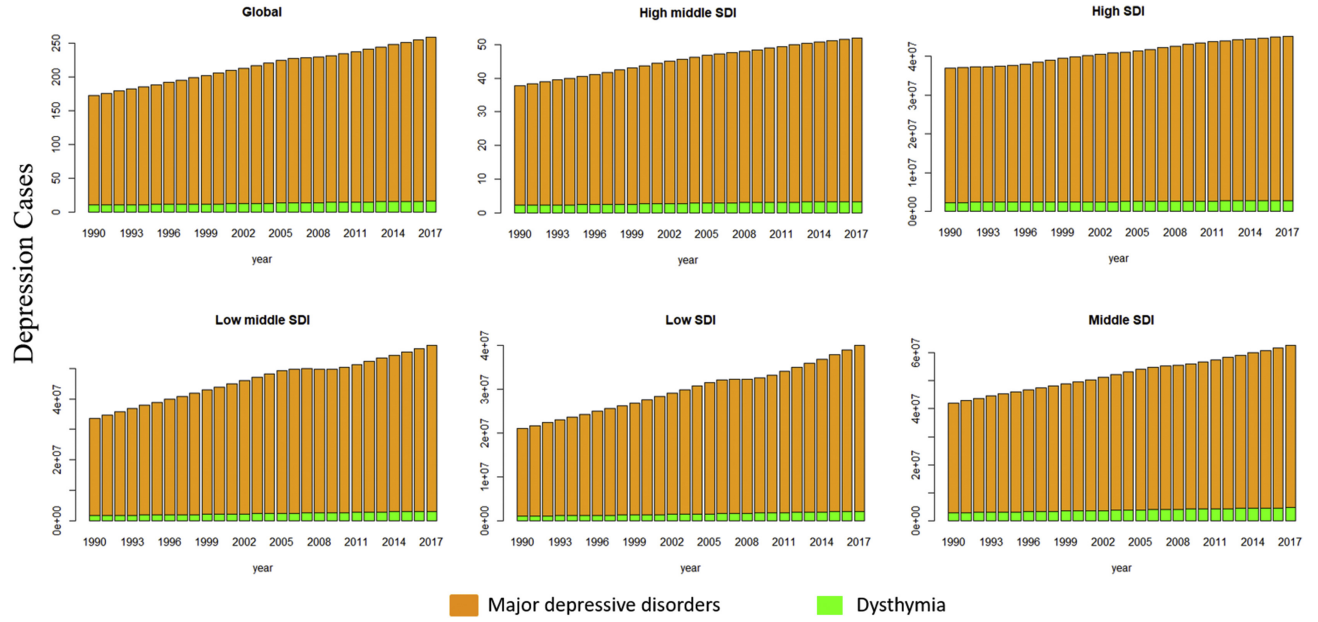


FIGURE 1.1: Depression cases, by SDI regions, from 1990 to 2017. [38]

development. According to this study, it was observed that the incidence for depression cases in countries across different SDI regions is steadily increasing over time (see figure 1.1). This further states the importance of addressing depressive disorders as a health issue across the world.

### Depression in South Africa

The history of depression and mental health in South Africa has been significantly shaped by the country's apartheid past and its socio-economic disparities. During the apartheid era, mental health services were largely segregated and inaccessible to the majority of the population, particularly in rural areas and for non-white communities [39].

After the end of apartheid in 1994, efforts were made to reform the mental health system and provide more equitable access to services. However, progress has been slow, and mental health remains a neglected area within the broader healthcare system [39, 17].

In terms of research on depression specifically, a study found that the proportion of South Africans who at one point were diagnosed with major depressive disorder was almost 10% [49]. More recently, the South African National Health and Nutrition Examination Survey (SANHANES-1) conducted in 2011-2012 reported a 4.5% prevalence of depression among South African adults [51].

However, it must be noted that the available research on depression in South Africa is still limited and can be relatively outdated. There is certainly a need for more comprehensive and ongoing surveillance to better understand the epidemiology and burden of depression across different regions, population groups, and socio-economic strata [39, 17].

### **1.1.2 Prediction and Diagnosis of Depression**

According to Centres for Disease Control (CDC), public health surveillance plays a vital role in detecting health issues early, monitoring trends, assessing intervention effectiveness, allocating resources efficiently, and informing policy development [19]. By systematically collecting, analyzing, and disseminating data on specific health events within a population, surveillance and monitoring enables timely responses to emerging health epidemics, evaluation of control measures, and evidence-based decision-making. The CDC believes this continuous monitoring helps address public health issues by identifying and addressing health challenges effectively.

Over time, there have been different approaches to depression monitoring, and advancements in technology have enhanced our ability to statistically monitor and analyze data not only for mental health disorders but also for other health incidents [30]. Each new approach yields different benefits and limitations when being compared to one another.

## Traditional Approaches

Traditional surveillance of depression typically involves monitoring individuals' mental health through diagnostic assessments rather than laboratory testing. Public health departments analyze data from clinical evaluations and self-report measures to identify trends and patterns in the occurrence and prevalence of depression within populations [35].

Current measures to predict or diagnose depression primarily involve clinical assessments and standardized screening tools. Clinicians often rely on structured or semi-structured interviews, such as the Structured Clinical Interview for DSM-5 (SCID), to diagnose depression based on established criteria [3]. Commonly used questionnaires include the Patient Health Questionnaire-9 (PHQ-9), which is a self-administered tool assessing the severity of depression symptoms, and the Beck Depression Inventory (BDI), a 21-item self-report inventory that measures the presence and severity of depressive symptoms [33, 5]. Additionally, the Hamilton Depression Rating Scale (HAM-D), a clinician-administered questionnaire, is frequently used to evaluate depression severity [24].

However, this approach is not without its limitations. One overarching drawback of the traditional method is the significant delay between diagnosing a disorder and reporting it. This delay, often referred to as a "time lag" can vary from days to weeks, hindering the timely implementation of interventions [43]. Moreover, studies have shown substantial diagnosis delays for certain conditions. For example, research conducted in Spain found that the average delay in diagnosing major depressive disorder was nearly 10 weeks [29]. Consequently, the combined time required for physical diagnosis of depressive disorders and the subsequent reporting delay pose significant challenges in promptly treating patients with depressive disorders.

Additionally, the resource demands associated with traditional surveillance methods can be burdensome, particularly for lower-income countries. Moreover, these methods may result in inaccuracies in reported information due to delays in data

collection and reporting [30]. The use of standardized tools like the Beck Depression Inventory (BDI) or the Patient Health Questionnaire (PHQ-9) for patient self-reporting may also lead to inaccuracies in the diagnosis of depression, since this source relies on the accuracy of the diagnosis scales used to classify patients with depression [33]. Early intervention and effective diagnosis are crucial in alleviating the burden of mental disorders, underscoring the paramount importance of interpretable and timely monitoring techniques [29].

### **Novel Approaches**

It is important to outline the way in which avenues for potential sources of depression screening data have opened up over time. As mentioned in [1.1.1](#), depression was not treated as a serious health issue until the 20th century [7].

In the late 20th century, the rise of cognitive-behavioral therapy (CBT) provided an effective psychotherapeutic approach for depression, addressing negative thought patterns and behaviors. Increased public awareness and destigmatization efforts also gained momentum [16]. The emergence and evolution of structured diagnostic criteria (e.g., DSM-III) and standardized assessment tools (e.g., Beck Depression Inventory) [26] would then further pave the way for effective depression diagnosis methods.

Eventually, the advent of the internet and digital technologies opened new avenues for monitoring and treating depression:

- Online screening tools and self-help resources became widely available, increasing accessibility and promoting early intervention [28].
- Telepsychiatry and online therapy platforms emerged, offering remote mental health services [40].
- Social media data analysis and natural language processing (NLP) techniques enabled the detection of depression signals from online behaviors and language patterns [23].

The use of internet platforms would also evolve the way in which traditional approaches were executed. For example, surveys and questionnaires paired with researched depression marking scales could now be sent out over the internet. This improved the amount of information researchers could gather about the general mental health status of the public [15].

Current research into novel approaches focuses on leveraging advanced technologies such as machine learning, wearable devices, and multimodal data analysis for early detection, personalized treatment, and continuous monitoring of depression [22]. There are a number of depression prediction methods that have emerged as a result of newer technologies and computational power:

**Digital Phenotyping:** The use of data from smartphones and wearable devices to monitor behavior and physiological signals indicative of depression [37, 41]. This type of approach evolved mainly due to the proliferation of smartphones and wearable devices and advances in sensor technology.

Pros:

1. Allows for continuous, real-time monitoring.
2. Can capture objective data unobtainable through self-report.
3. Scalable and potentially cost-effective.

Cons:

1. Privacy and ethical concerns regarding data collection.
2. Requires user compliance and technology access.
3. Data interpretation can be complex.

**Comparison to Traditional Methods:** Provides more continuous and objective data but lacks the depth of clinical interviews.

**Natural Language Processing (NLP):** Analyzing text from social media, electronic health records, or speech to detect depressive symptoms [31, 50]. Technologies that



enabled this kind of approach include increases in the availability of large text corpora, and computational power.

Pros:

1. Can analyze large volumes of data efficiently.
2. Captures naturalistic expressions of mood and thought patterns.
3. Non-intrusive data collection.

Cons:

1. Quality of data can vary widely.
2. Potential privacy issues.
3. Requires sophisticated algorithms to interpret nuances in language.

Comparison to Traditional Methods: Offers insights from naturally occurring language but may lack the structure and reliability of standardized tools.

**Machine Learning Models:** Utilizing various ML algorithms to predict depression based on diverse datasets [32, 12]. Machine learning models have been a researched field for some decades now, however, growth in data availability and computational power has led to their massive use across the world for various research tasks.

Pros:

1. Can handle large, complex datasets.
2. Capable of identifying subtle patterns and interactions in data.
3. Adaptive and continuously improving with more data.

Cons:

1. Requires large, annotated datasets for training.
2. Model interpretability can be challenging.

### 3. Potential biases in training data can affect outcomes.

Comparison to Traditional Methods: Provides scalability and efficiency but may lack the personalized touch of clinician-administered assessments.

It's important to note that while technological advancements have enabled novel approaches, researchers claim they should be considered complementary to, rather than replacements for, traditional methods involving human expertise and personalized care [22].

### **Machine Learning for Depression Measures**

Machine learning techniques have been increasingly applied to the task of predicting depression, leveraging large datasets and computational power to identify patterns and develop predictive models.

Compared to traditional methods, machine learning approaches offer the potential for more objective, scalable, and data-driven predictions. They can leverage diverse data sources and identify complex patterns that may not be easily discernible through human observation or manual analysis [32].

Furthermore, if compared to some novel approaches, machine learning models can potentially provide more robust and generalizable predictions by learning from larger datasets and accounting for multiple variables simultaneously [10].

In the South African context, there is a scarcity of studies specifically applying machine learning techniques for predicting depression. However, given the potential benefits of these approaches and the increasing availability of digital data sources, this area warrants further exploration, considering cultural and linguistic factors specific to South African populations.

I have compiled a list of machine learning methods and associated studies that demonstrate their use, as well as, what information the models were able to predict (see table [1.1](#)).

Model Type	Achievements	References
LASSO/Logistic Regression	Able to screen community dwellers in Korea for depression with an accuracy of 82%	[12]
Support Vector Machines	Were able to determine specific biomarker indicators for depression including genitality	[52]
Random Forests	Detected depression and PTSD in sexually abused children with accuracies of 88% and 76% respectively	[20]
Neural Networks	Screened survey participants for depression with a Recurrent Neural Network and found predictive factors causing depression	[25]
Bayesian Networks	Used a Bayesian Classifier to predict depression in senior citizens	[6]

TABLE 1.1: Achievements of Different Model Types

One common application of machine learning is the use of supervised learning algorithms to predict depression based on clinical and demographic data. For instance, Random Forests and Support Vector Machines (SVMs) have been employed to classify individuals as depressed or non-depressed using features such as age, gender, and medical history [42]. Additionally, logistic regression models have been utilized for similar classification tasks due to their interpretability and ease of implementation [46].

Deep learning methods, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have also shown promise in analyzing complex data types such as text and speech. CNNs have been used to detect depression from social media posts by capturing linguistic patterns indicative of depressive symptoms [45], but this falls more so into the field of natural language processing. Similarly, RNNs, especially Long Short-Term Memory (LSTM) networks, have been applied to time-series data to predict depression trajectories based on longitudinal health records [11].

Unsupervised learning techniques, including clustering algorithms and Principal Component Analysis (PCA), have been used to identify subgroups of patients with similar symptom profiles, aiding in the personalization of treatment plans [18]. Moreover, Natural Language Processing (NLP) methods, integrated with machine learning, have been utilized to analyze textual data from electronic health records and patient interviews to identify depressive symptoms and trends over time [34].

Ultimately, it is evident that the use of machine learning for predicting depression or depression symptoms has promising results. Research into these models, as well as further improvements in the space can be incredibly valuable for the future of depression diagnosis/intervention. This is especially true in the context of South Africa, where the amount of research conducted in the area of newer machine learning techniques for public health surveillance is sparse (as mentioned before).

## **1.2 Problem Statement**

Depression is a significant public health issue globally, contributing substantially to disability rates. Traditional approaches to monitoring and surveillance of depression often face delays and resource constraints, hindering timely interventions. In South Africa, research on the prevalence and determinants of depression remains limited, despite its pressing implications for public health. This lack of comprehensive and up-to-date data on depression epidemiology in the region poses challenges for designing targeted interventions and allocating resources effectively. The proposed research aims to address this gap by constructing machine learning models trained on socio-economic data from the National Income Dynamics Study to predict the likelihood of depression among individuals in South Africa. By providing a novel and potentially supplementary approach to depression surveillance and enhancing our understanding of the socio-economic factors associated with depression in the region, this research aims to possibly inform evidence-based machine learning strategies for mitigating the burden of depression in the South African population.

## 1.3 Research Question

Can machine learning models effectively predict the likelihood of depression among individuals in South Africa using socio-economic and demographic data from the National Income Dynamics Study?

## 1.4 Research Aims and Objectives

### 1.4.1 Research Aims

The aim of this research is to develop machine learning models that can accurately predict the likelihood of depression among individuals in South Africa by utilizing socio-economic, demographic, and health-related data from the National Income Dynamics Study (NIDS) dataset.

### 1.4.2 Objectives

The objectives of the research are:

1. To preprocess and prepare the NIDS dataset for machine learning analysis, including labeling participants for depression using the Center for Epidemiologic Studies Depression (CES-D) scale.
2. To explore and evaluate the performance of various machine learning algorithms in predicting depression risk from the NIDS data.
3. To rigorously validate and tune the top-performing machine learning models to optimize their predictive capabilities for depression.
4. To analyze the relative importance of different socio-economic and demographic features in the models for predicting depression likelihood.
5. To assess the potential real-world applicability and impact of using the developed machine learning models for depression surveillance and risk estimation in South Africa.

## 1.5 Limitations

There are a number of limitations that can be described from the current approach of the research, relating to its scope. These can be split into different areas:

### Practical Limitations:

- Implementing these machine learning models in real-world settings for depression surveillance and risk assessment may face technical, logistical, and resource challenges.
- Privacy and ethical concerns around using personal socio-economic data for mental health predictions could limit the adoption and scalability of such models.

### Interpretation Limitations:

- While the models may identify important predictive features, establishing causal relationships between socio-economic factors and depression risk may require further study.
- The interpretability of complex machine learning models can be limited, making it challenging to fully understand how predictions are being made. For medical professionals in the industry, this might be an unappealing aspect of this approach, since it is quite likely that they would want to understand exactly why the model was making its decisions.

Limitations of the methodology will be described in section [2.5](#).

## 1.6 Overview

The remaining sections of the proposal are structured as follows: Chapter [2](#) describes the methodology that will be employed to carry out the research. This includes discussions of the data handling methods that will be used, types of machine learning models that will be created and how they will be compared, as well as potential ethical considerations and limitations of the methodology. Chapter [3](#)

describes a general outline for the schedule of work and a description of potential difficulties that may be faced. Chapter 4 includes a summary of the proposal, concluding the report.

## Chapter 2

# Research Methodology

The methodology aims to investigate the benefits of using machine learning models to predict the likelihood of depression in survey participants. For this to be a success, various objectives of the research (outlined in section [1.4.2](#)) will need to be completed. This study ultimately hopes to provide valuable contributions to the field of infodemiology and depression monitoring.

### 2.1 Research design

The study is structured and split into several phases to achieve its objectives systematically:

1. Collection of Depression data from the South African National Income Dynamics Study (NIDS)
2. Depression labelling of survey participants with the use of the CESD-10 scale
3. Cleaning and preprocessing the data, as well as feature engineering and selection, to improve its quality and effectiveness for Machine Learning Classification
4. Creation of various Machine Learning Models to determine the most effective type of model for the problem
5. Validation and Hyper-parameter tuning of the various machine learning models to ensure the best performance for each model
6. Evaluation of machine learning models against particular metrics that are descriptive for this type of problem



7. Interpretation and reporting of findings, as well as discussion of results and possible limitations or caveats

## 2.2 Data

For this research study, the main source of data will be from the National Income Dynamics Study (N.I.D.S) [48].

### 2.2.1 Data Collection

#### Data from the National Income Dynamics Study

There are a number of ways to obtain health data for machine learning models, namely medical surveys, epidemiological/infodemiological studies and government health agencies such as the National Department of Health. A review that performed an infodemiological study on depression in South Africa use information from the South African Stress and Health study to investigate the statistics of depression in South Africa [47]. Another study used data from the National Income Dynamics Study (NIDS), to map clusters of depression cases across South African provinces [14].

The NIDS is a longitudinal study that surveys around 30000 South African individuals every 2-3 years. The NIDS datasets include various biographical, demographical and health information over multiple waves of surveys conducted within the past decade. [48]

Importantly, the NIDS datasets do not contain exclusive case information per participant for depression. However, there are a number of questions in the survey that relate to the participant's emotional state. The answers to these questions will be used to label participants for depression during data preprocessing.

## 2.2.2 Data Preprocessing

Data preprocessing is a crucial step in many machine learning and health studies, as it ensures the quality and integrity of the data before any analysis is performed. As mentioned in 2.2.1, the NIDS does not contain case information for depression, therefore more than just standard machine learning practices for data pre-processing will be necessary to convert the data into a useful format for the models. The following methods will be used:

### Data labelling for Depression

The study that used the NIDS datasets to create spatial clusters of depression dwellers in South Africa was able to create an incidence cohort of participants for depression [14]. This study will be referred to as the “NIDS study” from this point on in the proposal. Their method for depression labelling will be followed closely.

At the moment of making this proposal, there are 5 waves of surveys available on the NIDS website (<http://www.nids.uct.ac.za/>). The NIDS study used 3 waves. Each wave contains 10 questions that relate to the participant’s emotional state. These questions can also be found on a 20-item medical self-report scale for depression known as the Center for Epidemiologic Studies Depression Scale (CES-D) [48].

Importantly, the CES-D scale contains questions which can be answered with one of four options that indicate different frequencies of having experienced certain sentiments or symptoms. Questions with negative connotations such as “I felt fearful” would be scored according to the participant’s answer from 1-4. Questions with positive connotations such as “I was happy” would be reverse scored, and so an answer of “Most or all of the time (5-7 days)” would score 0 since it indicates the participant is not likely to be depressed. The scores for each answer are then summed and the cutoff score for a person to be labelled as “likely depressed” would be 20. (A person who scored 20 or above when their scores are all summed is therefore likely to be depressed)

Associated research has indicated that the 10 item abridged version with a cutoff score of 10 is still very accurate for depression classification [1, 14]. The NIDS study

used a cutoff score of 10 and therefore this study will be doing the same.

This presents the method of labelling each participant for depression that will be used for this research. The target feature/variable for all the models is then created from this labelling approach.

### **Data Cleaning**

Any missing data will be handled according to common practice. This will involve techniques like imputation (e.g., mean/median imputation, multiple imputation), or simply removing observations with missing values, depending on the extent and patterns of missingness [44].

Outliers can distort machine analyses and certain models make assumptions that outliers are not present [44], so they may need to be removed, winsorized, or treated separately, depending on the specific context and assumptions.

### **Data Transformation and Feature Engineering**

This would include any processes related to the augmentation/transformation of the data in order to improve its usability with machine learning models.

Normalization or standardization is a common pre-processing step. Some statistical models assume that the data is normally distributed or has a similar scale [44]. Techniques like z-score normalization or min-max scaling will be applied to transform the data to meet these assumptions, for any models that require these assumptions to be true.

Feature selection will be an important step for this research, due to the fact that depression can have many indicators or predictive factors [25]. Features relating to socio-demographic status will be included, such as age, marital status, income etc. Features relating to patient health will also be included: history of illnesses, smokes tobacco etc.

I intend to split the data into various subsets for training, validation and testing tasks exclusively. Since the dataset presents multiple waves, I will be using the first three waves as training data, the fourth wave as validation data and the fifth wave as testing data. The use of prior waves as training data will also give us an idea of how well models trained on past data can predict future data. This could present another strength of machine learning models for this type of classification problem.

### **2.2.3 Data Analysis**

Exploratory data analysis (EDA) follows data preprocessing and helps researchers understand data characteristics, identify patterns, and guide subsequent analyses [9].

Specifically, for this research, visualisation of the dataset will be useful for understanding the way in which the data is distributed. This would include histograms, pie charts and q-q plots for various features in the dataset.

Descriptive statistics like means, medians, and frequencies provides initial summaries which will also be useful for this research. Looking at various features in the data such as age, gender, etc. could provide interesting insights before the methodology is conducted.

## **2.3 Machine Learning Models**

Depending on the type of machine learning model being applied to a dataset, varying assumptions are made about the nature of the dataset. Based on research into current machine learning methods being used to predict depression, I have decided to compare the performance of 4 of the most popular types of models:

1. Logistic Regression
2. Random Forests
3. Deep Neural Networks
4. Support Vector Machines

### 2.3.1 Logistic Regression Model

Logistic regression is a foundational algorithm in machine learning, largely used for classification problems containing two target classes [27]. This is achieved through the logistic (sigmoid) function, which maps real numbers into the range between 0 and 1. The sigmoid function is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

where  $z$ , the input to the function, is typically a weighted sum of input features and model parameters.

The core of logistic regression lies in its ability to convert a sum of weighted input features into a probability. Specifically, the model computes a weighted sum of the input features, adds a bias term, and then applies the sigmoid function to this linear combination. The resultant probability is then used to classify the input by comparing it to a predefined threshold, typically 0.5 [22].

The creation of a logistic regression model involves several critical steps. Initially, the dataset must be prepared, including handling missing values, encoding categorical variables, and normalizing or standardizing the features. This will mean that the dataset used for training this model will be different to the datasets used to train other models, so as to account for the assumptions made by logistic regression models [27].

Mathematically, the training process involves maximizing the log-likelihood function, which sums the contributions of all data points, representing how probable the observed outcomes are given the model's parameters. The log-likelihood function is expressed as:

$$\begin{aligned} \ell(\beta) &= \sum_{i=1}^n [y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)] \\ &= \sum_{i=1}^n [y_i \mathbf{X}_i \beta - \log(1 + \exp(\mathbf{X}_i \beta))] \end{aligned}$$

The logistic regression model is then initialized, and the training process begins, where the model's parameters (weights and bias) are estimated. This estimation is done by maximizing the likelihood of observing the given data, which is often achieved through optimization techniques such as gradient descent. Maximising this log-likelihood therefore fits the model to the training data [27].

For predicting from the NIDS dataset, logistic regression will be a relatively low-cost algorithm and is less "black-box" in nature when compared to other models [32]. This justifies its use for the objectives of the research.

### 2.3.2 Random Forest model

Random forests are an ensemble learning technique that builds upon the foundational principles of decision trees to enhance predictive performance and robustness [8]. In essence, a random forest constructs a multitude of decision trees during training, each based on different bootstrap samples of the original dataset and considering random subsets of features at each split. This approach, known as *bootstrap aggregation* or *bagging*, helps to mitigate the risk of overfitting that individual decision trees often face by averaging out their predictions [8].

For classification tasks, a random forest aggregates the results of its constituent trees through majority voting. This ensemble method leverages the diversity among the trees to achieve greater generalization accuracy [20].

Mathematically this method can be described as, for a random forest classifier, the prediction  $\hat{y}$  for an input  $x$  is given by the mode of the predictions from  $n$  decision trees in the forest:

$$\hat{y} = \text{mode}(\hat{y}_1(x), \hat{y}_2(x), \dots, \hat{y}_n(x))$$

Although random forests are computationally more intensive and less interpretable than single decision trees, their ability to handle large datasets with high dimensionality, coupled with their resistance to overfitting, makes them a powerful tool

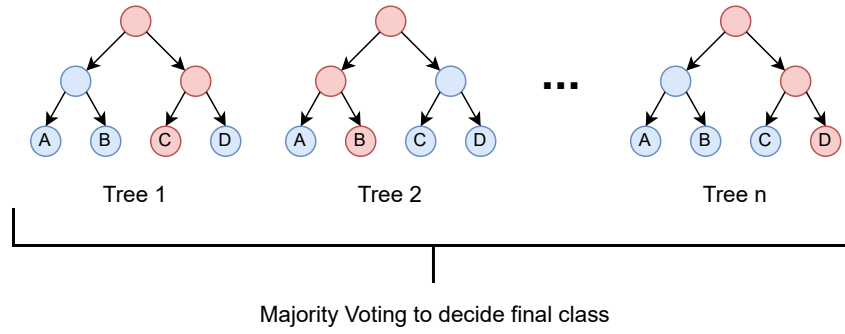


FIGURE 2.1: Simplification of random forest majority voting process

in various machine learning applications [20]. This strength will be particularly useful for handling the high number of features in the processed NIDS datasets.

### 2.3.3 Deep Neural Network

Neural networks are a class of machine learning models inspired by the biological neurons in the human brain. These models consist of interconnected layers of artificial neurons that process input data to make predictions or decisions [21]. These networks are called “deep” if they have more than one hidden layer. Deep Neural Networks (DNN) are capable of learning complex patterns and representations from data, making them powerful tools for tasks such as medical classification or natural language processing.

The fundamental building block of a neural network is the neuron, which receives input signals, performs a weighted sum of these inputs, applies an activation function to produce an output, and passes it to the next layer.

During training, neural networks learn to adjust the weights of connections between neurons to minimize a loss function, thus improving their ability to make accurate predictions on new data. This training process typically involves a technique known as *backpropagation* [21], where the network iteratively updates its weights by computing gradients of the loss function with respect to the model parameters.

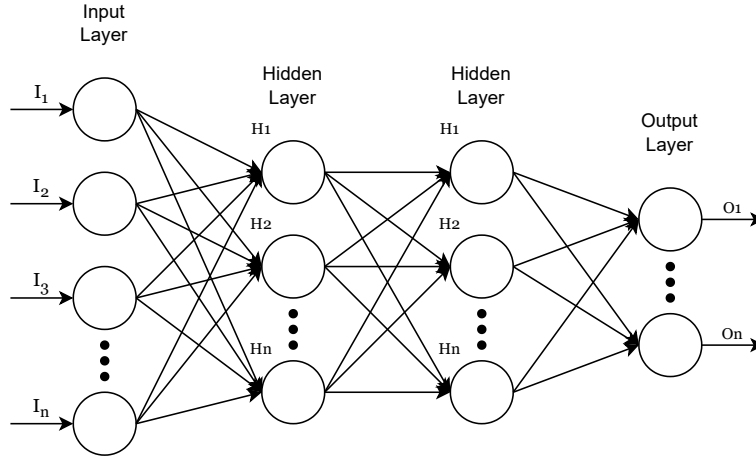


FIGURE 2.2: Visualisation of an arbitrarily large deep neural network

The loss function typically used for classification tasks is known as the Cross-Entropy loss function, and is defined as:

$$L = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

where:

- $L$  represents the cross-entropy loss.
- $N$  is the total number of samples.
- $y_i$  is the true label (ground truth) of the  $i$ -th sample.
- $\hat{y}_i$  is the predicted probability of the  $i$ -th sample belonging to the positive class.

For the purpose of this research, the activation functions will be sigmoid since they bound values between 0 and 1 which is convenient for a binary classification task. The structure of the network (such as number of layers, weights and layer types) will be adjusted and tuned according to the validation data.



### 2.3.4 Support Vector Machine

A Support Vector Machine (SVM) is a supervised learning algorithm that can be used a number of machine learning tasks [13]. In classification, an SVM finds the optimal hyperplane that separates the data into different classes, maximizing the margin between the classes. This hyperplane is chosen such that it has the maximum distance to the nearest data points of any class, hence maximizing adaptability to unseen data.

The data from the NIDS datasets is likely not linearly separable. To account for this kind of data, SVMs can use a kernel function which maps the input features into a higher-dimensional space where the classes become separable by a hyperplane [13]. The most commonly used kernels are:

#### Polynomial kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j + c)^d$$

#### Radial Basis Function (RBF) kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp \left( -\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2} \right)$$

#### Sigmoid kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\alpha \mathbf{x}_i^T \mathbf{x}_j + c)$$

A function defined as the decision function is a key component that determines how the SVM classifies new data points as it decides. The decision function takes an input data point  $x$  and computes a score or output based on its relationship with the trained SVM model [13]. This score indicates the confidence of the SVM in assigning the input data point to a particular class.

Mathematically, for a non-linear SVM, the decision function is represented as:

$$f(\mathbf{x}) = \sum_{i=1}^N \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b$$

where:

- $\alpha_i$  are Lagrange Multipliers
- $N$  is the number of support vectors
- $y_i$  are class labels
- $K(\mathbf{x}_i, \mathbf{x})$  is the kernel function
- $b$  is the bias term

When building the model, hyper-parameters will be randomly initialised, before being tuned on validation data. Another study which created SVMs for depressive disorder detection decided to validate their models on all the different kernel types rather than selecting one [52]. This approach is more likely to yield an accurate model and will therefore be used in this research method.

## 2.4 Analysis

Evaluation of machine learning performance is a critical process, especially for classification tasks like predicting depression risk. Since we will be classifying depression labels, a confusion matrix can be constructed to assess different aspects of the models' performance [22].

- *True Positive* (TP) - Predicted depression was likely and participant was actually self-reported to be depressed
- *False Positive* (FP) - Participant was self-reported as not depressed but model predicted they were likely to be depressed
- *True Negative* (TN) - Predicted depression was not likely and participant was actually self-reported to not be depressed
- *False Negative* (FN) - Participant was self-reported as depressed but model predicted they were not likely to be depressed

I have compiled a list of evaluation metrics used in papers involving the models that will be created as part of this research methodology [52, 20, 32, 6]. These metrics will be used to evaluate the models.

### **Accuracy**

Accuracy represents the proportion of correct predictions out of the total number of instances [25, 22]:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}}$$

Importantly, in the presence of data imbalances, accuracy can become less informative [44].

### **Precision**

Precision measures the proportion of true positive instances among all instances predicted as positive, quantifying the ability to avoid false positives [25, 22]:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

### **Recall**

Recall measures the proportion of true positive instances correctly identified by the model, quantifying the ability to find all positive instances [25, 22]:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

### **F1-Score**

The F1-score is an average (harmonic mean) between recall and precision [22, 10]:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

### **Specificity**

Specificity measures the proportion of true negative instances correctly identified by the model [22, 10]:

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

### **Area Under the Receiver Operating Characteristic (ROC) Curve (AUC-ROC)**

The ROC curve plots the true positive rate (recall) against the false positive rate (1 - specificity) at different classification thresholds. The AUC-ROC summarizes the model's ability to distinguish between positive and negative instances, with a value of 1 representing a perfect classifier and 0.5 representing a random classifier [22].

### **Area Under the Precision-Recall Curve (AUC-PR)**

The precision-recall curve plots precision against recall at different thresholds. It is particularly useful for imbalanced data, emphasizing the performance on the positive class [22].

## **2.5 Limitations**

The NIDS dataset, while comprehensive, may not capture all relevant socio-economic, demographic, and health factors that could influence depression risk. There may be important variables missing from the dataset. Moreover, the dataset is limited to South Africa, which could affect the generalizability of the findings and models to other countries or regions with different socio-economic and cultural contexts.

Furthermore, Machine learning models are ultimately based on the data used to train them. If there are biases or inaccuracies in the NIDS data, these could be reflected in the model predictions [44].

Importantly, the process of labeling participants for depression using the CES-D scale may not capture the full complexity and nuances of depressive disorders. The CES-D scale itself is not a clinical or medical diagnosis for depression, it is a self-reporting measure and should be treated as such for the purpose of this research [1].

## 2.6 Ethical Considerations

**Data Privacy and Consent:** The use of personal socio-economic and health data for machine learning models raises concerns about data privacy and the need for informed consent from participants. Participants' data in the NIDS surveys is anonymized [14, 48], abiding by these privacy rules. However, ethical approval/clearance for the use of the NIDS datasets will be required.

**Ethical Use of Predictive Models:** The development of predictive models for depression risk raises ethical questions about how these models will be used, particularly in terms of potential discrimination or stigmatization. Clear guidelines and governance frameworks should be established to ensure the ethical and responsible use of these models, avoiding unintended consequences or harm [22].

## Chapter 3

# Schedule of Work

### 3.1 Schedule of Work

For the aim of the research to be achieved, to build machine learning models to predict the likelihood of depression in NIDS survey participants (see methodology in [2](#)), the compilation of all 5 waves will need to be adjusted/processed differently depending on the model. Ideally, all of the data preparation for the models can be completed in the beginning phase of block 3. Thereafter, each model will need to be created, trained, validated and tested on their datasets. In block 4, these models should be complete and conclusions/interpretations can be drawn from all of the different models' results. Finally, the report will be drawn up before creating the presentation of the research for the School of Computer Science and Applied Mathematics Innovation Day. (see table [3.1](#))

Week	Scheduled task
July 08-15	Data preprocessing and labelling
July 15-22	Data splitting and further cleaning/preparation
July 22-29	Logistic Model Creation and Training
July 29 - August 05	Logistic Model Validation and Evaluation
August 05-12	Random Forest Model Creation and Training
August 12-19	Random Forest Model Validation and Evaluation
August 19-26	Deep Neural Net Creation and Training
August 26 - September 02	Deep Neural Net Validation and Evaluation
September 02-09	Support Vector Machine Creation and Training
September 09-16	Support Vector Machine Validation and Evaluation
September 16-23	Extra time for completion of any prior tasks
September 23-30	Interpretation of results and summary analysis
September 30 - October 01	Writing Report
October 01-08	Writing Report
October 08-15	Writing Report
October 15-22	Creating presentation for Innovation Day
October 22-29	Extra time for completion of report/presentation
October 29 - November 03	Examination period begins

TABLE 3.1: Schedule of work for blocks 3 and 4

## 3.2 Potential Difficulties

It is difficult to give an informed estimate of how long each model will take. It is possible that some models might take much longer than others, due to implementation difficulties etc. I have assigned extra time after the model creation and evaluation phases are over to account for any model creation/implementation difficulties.

Furthermore, there is a large amount of data pre-processing that must occur, this also might take longer than expected but 2 weeks seems comfortable based on what I have already seen from the actual data.

## Chapter 4

# Conclusion

This research proposal has identified significant gaps in the current literature regarding the prediction and surveillance of depression in South Africa using machine learning models. While traditional approaches to depression monitoring face challenges of delays, resource constraints, and limited scalability, the application of advanced analytical techniques that utilise large-scale socioeconomic datasets presents an opportunity to address these limitations.

The primary aim of this research is to develop machine learning models that can accurately predict the likelihood of depression among individuals in South Africa by utilising socioeconomic, demographic, and health-related data from the National Income Dynamics Study (NIDS) dataset. By harnessing the power of algorithms such as logistic regression, random forests, deep neural networks, and support vector machines, this study seeks to enhance our understanding of the complex interplay between socioeconomic factors and depression risk.

The proposed research has the potential to make valuable contributions to the field of public health surveillance and depression epidemiology in South Africa. Firstly, by training on longitudinal data from the NIDS surveys, the developed models may uncover novel insights into the socioeconomic determinants of depression within the South African context. Secondly, the application of machine learning techniques to this domain could pave the way for more efficient and scalable approaches to depression surveillance, complementing traditional methods and informing targeted interventions. Finally, the findings of this study may contribute to the broader discourse on the responsible and ethical integration of advanced analytical tools in mental health research and practice.



By addressing the pressing need for effective and timely surveillance strategies, this research holds the potential to guide resource allocation, policy development, and targeted interventions aimed at mitigating the burden of depression in South Africa.

# Bibliography

- [1] Elena M Andresen et al. "Screening for depression in well older adults: evaluation of". In: *Prev Med* 10 (1994), pp. 77–84.
- [2] American Psychiatric Association. *The Diagnostic and Statistical Manual of Mental Disorders*. Originally published in 1952, latest revision in 2022, p. 947. ISBN: 978-0-89042-554-1.
- [3] R Michael Bagby and Thomas A Widiger. "Five Factor Model personality disorder scales: An introduction to a special section on assessment of maladaptive variants of the five factor model." In: *Psychological Assessment* 30.1 (2018), p. 1.
- [4] Joana M Barros, Jim Duggan, and Dietrich Rebholz-Schuhmann. "The application of internet-based sources for public health surveillance (infoveillance): systematic review". In: *Journal of medical internet research* 22.3 (2020), e13680.
- [5] Aaron T Beck et al. "Comparison of Beck Depression Inventories-IA and-II in psychiatric outpatients". In: *Journal of personality assessment* 67.3 (1996), pp. 588–597.
- [6] Ishita Bhakta and Arkaprabha Sau. "Prediction of depression among senior citizens using machine learning classifiers". In: *International Journal of Computer Applications* 144.7 (2016), pp. 11–16.
- [7] Michel Bourin. "History of depression through the ages". In: *Archives of Depression and Anxiety* (2020).
- [8] Leo Breiman. "Random forests". In: *Machine learning* 45 (2001), pp. 5–32.
- [9] D. Bzdok and N. Altman. *Machine Learning: A Guide to Current Research*. Vol. 12. 3. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2018, pp. 1–207. DOI: [10.2200/s00868ed1v01y201810aim040](https://doi.org/10.2200/s00868ed1v01y201810aim040).
- [10] Danilo Bzdok, Martin Krzywinski, and Naomi Altman. "Machine learning: supervised methods". In: *Nature methods* 15.1 (2018), p. 5.

- [11] Anna Cascarano et al. "Machine and deep learning for longitudinal biomedical data: a review of methods and applications". In: *Artificial Intelligence Review* 56.Suppl 2 (2023), pp. 1711–1771.
- [12] Seo-Eun Cho, Zong Woo Geem, and Kyoung-Sae Na. "Predicting depression in community dwellers using a machine learning algorithm". In: *Diagnostics* 11.8 (2021), p. 1429.
- [13] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine learning* 20 (1995), pp. 273–297.
- [14] Diego F Cuadros et al. "Spatial structure of depression in South Africa: A longitudinal panel survey of a nationally representative sample of households". In: *Scientific reports* 9.1 (2019), p. 979.
- [15] Reza Che Daniels, Kim P Ingle, and Timothy SL Brophy. "Determinants of attrition between Waves 1 and 2 of South Africa's National Income Dynamics Study–Coronavirus Rapid Mobile Survey (NIDS-CRAM)". In: *South African Journal of Economics* 90.4 (2022), pp. 535–552.
- [16] Keith S Dobson. "A meta-analysis of the efficacy of cognitive therapy for depression." In: *Journal of consulting and clinical psychology* 57.3 (1989), p. 414.
- [17] S. Docrat, C. Lund, and D. Chisholm. "Sustainable financing options for mental health care in South Africa: findings from a situational analysis and key informant interviews". In: *International Journal of Mental Health Systems* 13.1 (2019), pp. 1–14.
- [18] Andrew T Drysdale et al. "Resting-state connectivity biomarkers define neurophysiological subtypes of depression". In: *Nature medicine* 23.1 (2017), pp. 28–38.
- [19] Andrew Friede, Joseph A Reid, and Howard W Ory. "CDC WONDER: a comprehensive on-line public health information system of the Centers for Disease Control and Prevention." In: *American Journal of Public Health* 83.9 (1993), pp. 1289–1294.
- [20] Emel Sari Gokten and Caglar Uyulan. "Prediction of the development of depression and post-traumatic stress disorder in sexually abused children using a random forest classifier". In: *Journal of Affective Disorders* 279 (2021), pp. 256–265.

- [21] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. <http://www.deeplearningbook.org>. MIT Press, 2016.
- [22] Sarah Graham et al. “Artificial intelligence for mental health and mental illnesses: an overview”. In: *Current psychiatry reports* 21 (2019), pp. 1–18.
- [23] Sharath Chandra Guntuku et al. “Detecting depression and mental illness on social media: an integrative review”. In: *Current Opinion in Behavioral Sciences* 18 (2017), pp. 43–49.
- [24] Max Hamilton. “The Hamilton Depression Scale—accelerator or break on antidepressant drug discovery”. In: *Psychiatry* 23.1 (1960), pp. 56–62.
- [25] Jiatong Han et al. “Depression prediction based on LassoNet-RNN model: A longitudinal study”. In: *Heliyon* 9.10 (2023).
- [26] Robert MA Hirschfeld. “The epidemiology of depression and the evolution of treatment”. In: *The Primary Care Companion for CNS Disorders* 14.Suppl 1: Editor Choice (2012), p. 26328.
- [27] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.
- [28] Thomas K Houston, Lisa A Cooper, and Daniel E Ford. “Internet support groups for depression: a 1-year prospective cohort study”. In: *American Journal of Psychiatry* 159.12 (2002), pp. 2062–2068.
- [29] Raúl Huerta-Ramírez et al. “Diagnosis delay in first episodes of major depression: a study of primary care patients in Spain”. In: *Journal of affective disorders* 150.3 (2013), pp. 1247–1250.
- [30] Amy Kristen Johnson et al. “Nowcasting Sexually Transmitted Infections in Chicago: Predictive Modeling and Evaluation Study Using Google Trends”. In: *JMIR PUBLIC HEALTH AND SURVEILLANCE* (2020).
- [31] Christian Karmen, Robert C Hsiung, and Thomas Wetter. “Screening internet forum participants for depression symptoms by assembling and enhancing multiple NLP methods”. In: *Computer methods and programs in biomedicine* 120.1 (2015), pp. 27–36.
- [32] Ronald C Kessler et al. “Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports”. In: *Molecular psychiatry* 21.10 (2016), pp. 1366–1371.

- [33] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. “The PHQ-9: validity of a brief depression severity measure”. In: *Journal of general internal medicine* 16.9 (2001), pp. 606–613.
- [34] Mrinal Kumar et al. “Detecting changes in suicide content manifested in social media following celebrity suicides”. In: *Proceedings of the 26th ACM conference on Hypertext & Social Media*. 2015, pp. 85–94.
- [35] Claudia Lang. “Inspecting mental health: depression, surveillance and Care in Kerala, South India”. In: *Culture, Medicine, and Psychiatry* 43.4 (2019), pp. 596–612.
- [36] Sian Lewis et al. *World mental health report*. Tech. rep. World Health Organisation, 2022.
- [37] Yunji Liang, Xiaolong Zheng, and Daniel D Zeng. “A survey on big data-driven digital phenotyping of mental health”. In: *Information Fusion* 52 (2019), pp. 290–307.
- [38] Qingqing Liu et al. “Changes in the global burden of depression from 1990 to 2017: Findings from the Global Burden of Disease study”. In: *Journal of Psychiatric Research* 126 (2020), pp. 134–140. URL: <https://www.sciencedirect.com/science/article/pii/S0022395619307381>.
- [39] C. Lund et al. “Public sector mental health systems in South Africa: inter-provincial comparisons and policy implications”. In: *Social Psychiatry and Psychiatric Epidemiology* 45.3 (2010), pp. 393–404.
- [40] Hilty Donald M et al. “The effectiveness of telemental health: a 2013 review”. In: *Telemedicine and e-Health* 19.6 (2013), pp. 444–454.
- [41] Jukka-Pekka Onnela and Scott L Rauch. “Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health”. In: *Neuropsychopharmacology* 41.7 (2016), pp. 1691–1696.
- [42] Sohrab Saeb et al. “The relationship between clinical, momentary, and sensor-based assessment of depression”. In: *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. IEEE. 2015, pp. 229–232.
- [43] Luana Gorgueira Santos. “Surveillance of Tuberculosis by analysing google trends”. In: *Católica Lisbon School of Economics and Business* (2023).

- [44] Abhishek Sheetal, Zhou Jiang, and Lee Di Milia. "Using machine learning to analyze longitudinal data: A tutorial guide and best-practice recommendations for social science researchers". In: *Applied Psychology* 72.3 (2023), pp. 1339–1364.
- [45] Guangyao Shen et al. "Depression detection via harvesting social media: A multimodal dictionary learning solution." In: *IJCAI*. 2017, pp. 3838–3844.
- [46] Gregory E Simon et al. "Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records". In: *American Journal of Psychiatry* 175.10 (2018), pp. 951–960.
- [47] Mark Tomlinson et al. "The epidemiology of major depression in South Africa: results from the South African Stress and Health study: mental health". In: *South African Medical Journal* 99.5 (2009), pp. 368–373.
- [48] UCT. *National Income Dynamics Study*. Accessed: 2024-05-20. 2024. URL: <http://www.nids.uct.ac.za/>.
- [49] D. R. Williams et al. "Twelve-month mental disorders in South Africa: prevalence, service use and demographic correlates in the population-based South African Stress and Health Study". In: *Psychological Medicine* 38.2 (2008), pp. 211–220.
- [50] JT Wolohan et al. "Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP". In: *Proceedings of the first international workshop on language cognition and computational models*. 2018, pp. 11–21.
- [51] V. Pillay-van Wyk et al. "Mortality trends and differentials in South Africa from 1997 to 2012: second National Burden of Disease Study". In: *The Lancet Global Health* 4.9 (2016), e642–e653.
- [52] JS Yu et al. "A support vector machine model provides an accurate transcript-level-based diagnostic for major depressive disorder". In: *Translational psychiatry* 6.10 (2016), e931–e931.