CLASSIFYING MENTAL HEALTH OUTCOMES USING THE MHI-5 SCREENING TOOL

A COMPARATIVE STUDY OF RANDOM FOREST, XGBOOST, SUPPORT VECTOR MACHINE AND LOGISTIC REGRESSION

NOHÉMI R.C.COSSTER

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

u1246538

COMMITTEE

dr. Marijn van Wingerden

Msc. Ratislav Hronský

LOCATION

Tilburg University

School of Humanities and Digital Sciences

Department of Cognitive Science &

Artificial Intelligence

Tilburg, The Netherlands

DATE

May 19th, 2025

WORD COUNT

8799

ACKNOWLEDGMENTS

To say that I went through an evolution during this process would not be an overstatement. One need only look at the progression of titles this thesis has carried: from the initial "Wear It Green: Predicting Mental Health with the MHI-5 Screening Tool" to the equally ambitious "Green Screens: Predicting Mental Health Outcomes", and finally to my more precise current title. None of this evolution—awkward names and all—would have been possible without the support of the following people.

I would like to express my sincere gratitude to my thesis supervisor, Dr. Marijn van Wingerden, for his guidance and support throughout this project. I am also thankful to my fellow students, especially Ilona and Tuur. A special thanks goes to my dear mom, Ludwina Gijsbertha, and my sister Egelique—no matter the project, they always have my back. Happy reading!

CLASSIFYING MENTAL HEALTH OUTCOMES USING THE MHI-5 SCREENING TOOL

A COMPARATIVE STUDY OF RANDOM FOREST, XGBOOST, SUPPORT VECTOR MACHINE AND LOGISTIC REGRESSION

NOHÉMI R.C.COSSTER

Abstract

Given the growing global burden of mental health issues, screening tools are widely used for early identification, yet their predictive potential remains underexplored. This thesis evaluates the predictive power of four machine learning models—Logistic Regression, Random Forest, Support Vector Machine, and XGBoost—in classifying mental health outcomes using the MHI-5 screening tool. Using longitudinal data from the Dutch LISS panel, these models were compared across three tasks: contemporaneous prediction (2022), year-ahead prediction (2023), and prediction of mental health change between 2022 and 2023 (improved, stable, deteriorated). The study also compared balancing techniques (oversampling versus undersampling) and feature selection approaches (theory-driven versus data-driven). Logistic Regression with Random Forest feature importance feature selection method achieved strong performance while using a minimal set of predictors (F1 = 0.61, recall = 0.76), making it a robust and efficient choice for identifying individuals at risk. The model's performance remained stable for year-ahead prediction (F1 = 0.56; p = 0.55). However, predicting change in mental health status proved more challenging, yielding a macro F1 of 0.35 across balanced classes. These findings support the development of scalable, data-driven tools for early detection of mental health risks in the general population.

0.1 *Source/Code/Ethics/Technology Statement*

This research utilizes data from the LISS panel (Longitudinal Internet Studies for the Social Sciences), operated by a non-profit institute Centerdata (Tilburg University in the Netherlands). The collected data is anonymous. This thesis did not involve data collection from humans or animals. All figures were created by the author. The data and code used in this thesis belong to the original owner, both during and after completion. The author generated all of the figures shown in this work. Software utilized in this study is listed with their respective references in Appendix C. The code and data used in this project are available in the GitHub repository here. Regarding language assistance, ChatGPT (OpenAI, 2021) was used to enhance the clarity of the author's original content. Additionally, Grammarly (Grammarly Inc., 2023) contributed to grammatical improvements and spell-checking. No other typesetting tools or services were employed.

1 INTRODUCTION

1.1 *Motivation*

Ursula von der Leyen, president of the European Commission, recently highlighted the growing mental health crisis among young people, and pledged to launch a Europe-wide inquiry into its effects (European Commission, 2024). This reflects the increasing global concern about mental health, with the World Health Organization (WHO) reporting that one in eight people globally experience mental disorders, such as anxiety and depression, affecting nearly 970 million individuals in 2019 (WHO, 2022). Early identification and intervention are critical in addressing this widespread problem, and mental health screening tools have emerged as essential in this process. While formal diagnosis of mental health disorders requires comprehensive assessments by trained professionals—who use diagnostic systems like the DSM-V or ICD-11, along with clinical interviews and detailed analysis of the patient's history—screening tools offer an efficient and accessible first step for identifying individuals who may be at risk.

A prominent example of such screening tools is the Mental Health Inventory 5 or MHI-5. The MHI-5 is a five-item scale derived from the longer Mental Health Inventory, assessing general mental health with a focus on symptoms of anxiety and depression (Berwick et al., 1991). The scale includes five questions measuring psychological distress and well-being over the past month, such as feelings of nervousness, depression, and overall mental health status. Many studies have validated the MHi-5,

confirming its reliability and ability to accurately measure mental health, and its effectiveness (Rumpf et al., 2001; Means-Christensen et al., 2005). Also, its brevity and ease of use make it an effective tool in large-scale mental health research and clinical practice (Hoeymans et al., 2004). Though the MHI-5 is widely used to assess mental health, research has primarily centered on descriptive and explanatory methods, with little emphasis on its use for predictive modeling. Given the urgency of addressing mental health, this research aims to comprehensively evaluate machine learning approaches for MHI-5 prediction, comparing model performance across different prediction tasks and temporal contexts.

1.2 *Problem Statement*

While the MHI-5 is widely used for mental health assessment, its potential for predictive modeling remains underexplored. This represents a significant data science challenge because mental health outcomes involve complex, non-linear relationships between diverse predictors that traditional statistical methods struggle to capture.

Existing research primarily employs traditional statistical approaches with limited predictor sets. For example, Van der Velden et al. (2019) examined how social media use, loneliness, and demographics influence MHI-5 outcomes using logistic regression but did not explore advanced machine learning techniques or broader predictor categories. Comparative analyses evaluating multiple machine learning models for MHI-5 prediction, and studies focusing on MHI-5 as a primary target variable, are virtually absent.

Furthermore, research investigating predictor-outcome relationships over time is scarce. While establishing causality requires confounding analyses, experimental designs, or advanced causal inference methods, longitudinal studies that incorporate temporal precedence—where predictors precede the target variable—offer a critical first step in exploring causal pathways. However, such longitudinal applications of machine learning to MHI-5 remain underexplored. Studies leveraging rich longitudinal data from the LISS panel to predict MHI-5 are also limited. This thesis addresses these gaps by systematically applying and comparing machine learning models to predict MHI-5 outcomes both contemporaneously and longitudinally, utilizing LISS panel data to explore predictive relationships and potential causal pathways.

1.3 *Scientific Relevance*

The current gap in the literature limits our understanding of how screening tools like the MHI-5 can predict mental health outcomes. Current MHI-5 research relies primarily on traditional statistical methods that cannot capture complex, non-linear relationships between predictors and mental health outcomes, while machine learning models excel at identifying intricate patterns and revealing the most important predictors.

No studies have systematically incorporated comprehensive predictor categories or examined what happens to MHI-5 outcomes one year later using longitudinal approaches or explored mental health trajectory changes over time. Research predicting whether individuals improve, remain stable, or decline is virtually absent. Additionally, the field lacks rigorous comparative evaluations of multiple ML algorithms, feature selection methods, and data balancing techniques specifically for MHI-5 prediction.

This study addresses these gaps through a systematic comparison of multiple ML algorithms for MHI-5 prediction, a comprehensive evaluation of feature selection and data balancing approaches, and novel prediction of mental health trajectory changes using extensive longitudinal LISS panel data. Hence, this thesis makes significant methodological contributions to the scientific community by expanding the predictive modeling of MHI-5 scores through advanced ML techniques.

1.4 *Social Relevance*

As shown in the motivation section of this thesis, mental health is a growing global concern, with one in eight people worldwide experiencing disorders such as anxiety and depression (WHO, 2022). This amounted to nearly 970 million individuals in 2019. It is neither feasible nor scalable to send everyone to a psychologist for diagnosis, given the limited resources in mental healthcare systems. Tools like the MHI-5 play a critical role as screeners, helping to flag individuals who may be at risk of developing mental health problems.The goal, therefore, is not to "reduce someone's MHI-5 score" directly, but to identify and address the underlying factors that contribute to poor scores. By catching at-risk individuals early, we can help prevent deterioration and connect them with the most appropriate form of support or treatment.

This research contributes by identifying key predictors associated with high-risk MHI-5 profiles. Many of these predictors—such as emotional instability, or social disconnection—are modifiable. If we can reliably predict who is at risk and understand what factors contribute to that risk, then interventions can be targeted at those drivers. This reduces the need

to rely exclusively on formal diagnoses or clinical treatments, which may only come into play after problems have become severe.

In this way, predicting MHI-5 scores is not just a technical task, but a means to support preventive public health strategies. It enables earlier identification, better resource allocation, and the design of personalized or population-level interventions. Ultimately, by addressing the upstream drivers of mental distress, this approach helps reduce the overall burden on individuals, communities, and healthcare systems.

1.5 *Research Strategy & Research Questions*

This thesis investigates the performance of various machine learning models to predict MHI-5 in 2022 and 2023 using LISS panel data. Additionally, it assesses the predictive value of a range of variables from 2022 in relation to mental health outcomes across two years. This study builds upon Van der Velden (2019), which used logistic regression to examine how social media use, loneliness, and demographics influence MHI-5. However, Van der Velden's study lacked predictive evaluation metrics, as it only fit the training data. In contrast, this thesis emphasizes predictive performance (e.g., accuracy, F1-score), thereby offering a more robust and generalizable contribution.

A combination of cross-sectional and longitudinal designs is employed to investigate mental health status using three main approaches:

1. A contemporaneous study examining how 2022 predictors relate to MHI-5 outcomes in 2022 (binary classification)
2. A year-ahead study using 2022 predictors to forecast MHI-5 outcomes in 2023 (A stability test)
3. A change-focused study predicting whether mental health improved, remained stable, or declined between 2022 and 2023 (multiclass classification)

Best research practices are applied to build the most robust model by investigating optimal feature selection methods and balancing techniques. Feature selection is essential for identifying the key determinants of mental health, while balancing techniques are used to address class imbalance—common in mental health research. In addition, model optimization is performed through hyperparameter tuning and comprehensive error analysis.
The main research questions are:

RQ1 *Which machine learning models are most effective in predicting binary classifications of mental health status based on MHI-5 scores in 2022 using LISS panel data?*

RQ2 *How well do machine learning models trained on 2022 data generalize to predicting mental health status in 2023 based on MHI-5 scores using LISS panel data, and how stable is their predictive performance over time?*

RQ3 *Which machine learning models are most effective in predicting changes in mental health status (improved, stable, or deteriorated) based on MHI-5 score differences between 2022 and 2023?*

The sub-research questions are:

SQ1 *How does the performance of logistic regression compare to that of random forest, gradient boosting, and support vector machine?*

SQ2 *What are the most important predictors of mental health status in 2022, 2023, and of changes between these years?*

SQ3 *Which feature selection method (theory-driven or data-driven) yields optimal prediction performance for MHI-5 outcomes in 2022, and for predicting changes?*

SQ4 *Which balancing technique (oversampling or undersampling) produces the best predictive performance for MHI-5 outcomes in 2022, and for predicting changes?*

SQ5 *Does model performance vary across different demographic groups?*

2 RELATED WORK

Recent years have seen growing interest in applying machine learning to mental health prediction, yet limited attention has been given to the MHI-5 as a primary outcome. Despite its widespread use as a screening tool, MHI-5 has mostly been explored descriptively rather than predictively. This chapter reviews existing literature on machine learning applications in mental health, relevant predictors, and the use of longitudinal data. It also highlights the research gap concerning predictive modeling of MHI-5 in cross-sectional, longitudinal, and change-focused settings.

While MHI-5 is a widely used mental health measure, research specifically focused on predicting MHI-5 is limited, possibly due to its role as a brief screening tool rather than a diagnostic instrument (Strand et al., 2003). Machine learning approaches have shown promise in predicting various mental health outcomes, with models like Random Forest (RF), Support Vector Machines (SVM), Logistic Regression, and K-Nearest Neighbors (KNN) being commonly applied (Spyrou et al., 2016). The literature shows that Random Forest models consistently perform well in predicting mental health problems, emerging as the best-performing classifier in several studies (Abdul Rahimapandi et al., 2022; Y. Li, 2023).

Random Forest models have demonstrated strong performance in predicting mental health outcomes, with reported accuracies ranging broadly across studies depending on the dataset, outcome type, and modeling approach. For instance, (Spyrou et al., 2016) achieved 95.45% accuracy using Random Forest to classify geriatric depression in elderly participants, using EEG recordings from 34 participants with cognitive impairment and depression and 32 control subjects. On the lower end, Hornstein et al. (2021) achieved 71% accuracy when predicting binary treatment responses, defined as either a 5-point reduction on the PHQ-9 depression scale or a 4-point reduction on the GAD-7 anxiety scale.

It is important to note that evaluation metrics vary significantly depending on several factors: the specific mental health outcome being predicted (binary vs. continuous), class imbalance in the dataset, sample size, feature selection methods, the characteristics of the dataset used for training and testing (including train-test split ratios), data quality, hyperparameter tuning, and the choice of evaluation metrics themselves (accuracy vs. precision vs. recall vs. F1-score).

However, studies specifically predicting MHI-5 scores are scarce. While Shaikh Mohammad and Siddiqui (2021) achieved 95.83% accuracy using Random Forest Regressor and Bayesian Optimization to predict mental health status using questionnaire and screening tools (MHI-5, PHQ-9, and BDI), they did not focus specifically on MHI-5 as the primary outcome

variable. Similarly, Elovainio et al. (2020) used Random Forest algorithms but included MHI-5 only as a predictor variable for mental health service use rather than as an outcome.

2.1 *LISS Panel Data*

The LISS (Longitudinal Internet Studies for the Social Sciences) panel provides a rich source of longitudinal data on a representative sample of Dutch households, including mental health measures like the MHI-5. Despite the wealth of data available, peer-reviewed studies specifically utilizing LISS panel data for mental health research remain relatively scarce.

Van der Velden and colleagues have conducted several influential studies using the LISS panel data with MHI-5 as a key mental health indicator, though primarily using traditional statistical approaches rather than machine learning methods. In their 2019 study, van der Velden, Setti, et al. (2019b) examined the relationship between social networking sites (SNS) use and mental health problems using logistic regression analyses. They found that when controlling for prior mental health problems, loneliness, and demographics, SNS use did not significantly predict mental health outcomes as measured by MHI-5. Further extending this research, van der Velden, Das, and Muffels (2019) investigated the stability of mental health problems among Dutch young adults, again as measured by MHI-5, across three cohorts (2007, 2012, and 2017) using the LISS panel. Using latent profile analysis, they identified four distinct mental health profiles: "healthy" (82.2%), "at risk" (9.6%), "clinical" (4.2%), and "treatment" (3.9%).They found that these mental health profiles showed considerable stability over time, with most individuals maintaining their profile membership across the study period. Furthermore, in their 2022 longitudinal cohort study, van der Velden et al. (2022) examined the impact of COVID-19 on adolescent mental health, again using the MHI-5 as a key outcome measure. These studies collectively demonstrate the potential of LISS panel data for investigating factors affecting mental well-being in the Dutch population over time, while simultaneously revealing a significant gap: despite the use of MHI-5 as an outcome measure in traditional statistical analyses, the predictive potential of this screening tool remains largely unexplored through machine learning approaches.

2.2 *Predictors Of Mental Health*

Previous research has extensively examined various predictors and determinants across different mental health outcomes. Key factors that have

been found to predict mental health conditions include personality traits and social support networks (Burešová et al., 2020). Educational attainment has also been identified as a significant predictor Gloster et al. (2020), along with personal life values (Cohen & Cohen, 1995). Demographic characteristics play a crucial role, particularly age, gender, and income levels (Kohn et al., 2022; Shields-Zeeman & Smit, 2022; Wilhelm, 2014).

More novel predictors have emerged from recent research, such as the analysis of everyday language patterns on social media platforms to forecast mental health conditions (Thorstad & Wolff, 2019). van der Velden, Setti, et al. (2019b) similarly found associations between social networking site use and later mental health outcomes in a prospective population-based study.Additional determinants have been highlighted in prior research, including eating and snacking behaviors as discussed by Smith (2011), the role of religious involvement noted by Levin (2010), psychedelic use examined by Krebs and Johansen (2013), and political influences explored by Bhugra and Ventriglio (2023). While these studies, primarily cross-sectional in nature, have established these variables as predictors for various mental health outcomes, their specific predictive power for MHI-5 scores remains unexplored. This gap presents an opportunity to evaluate these established determinants using machine learning approaches to predict MHI-5 scores specifically.

2.3 *Longitudinal Setting*

Longitudinal studies play a critical role in identifying causal relationships in mental health research. By analyzing predictors and their influence on MHI-5 scores over time, this study moves beyond mere cross-sectional correlations to explore potential causal pathways. As highlighted by Kraemer et al. (2001), temporal precedence—where changes in predictors occur before changes in outcomes—is a fundamental criterion for establishing causality.

While randomized controlled trials are the gold standard for establishing causality, they are often impractical or unfeasible in large-scale population studies. Instead, longitudinal observational data, such as the LISS panel, provide a valuable alternative for examining these relationships over time, despite inherent limitations.

Pollar and Harigovind (2018) applied machine learning models to classify and predict outcomes in longitudinal data. They trained classifiers on data from previous visits to predict the binary colonization status (colonized or not) at a subsequent visit. For their imbalanced dataset, Gradient Boosting (GB) and Support Vector Machines (SVM) with a radial basis function (RBF) kernel outperformed other classifiers, with GB achieving an

accuracy of 71.72% and F1-score of 70.21, and SVM achieving an accuracy of 67.58% and F1-score of 66.15. For MHI-5, it is similarly important to investigate how predictors in one year influence subsequent MHI-5 scores a year later, aligning with the criterion of temporal precedence. Additionally, this study evaluates the predictive power of machine learning models in forecasting future mental health outcomes at a one-year interval—an area that remains under-researched with respect to MHI-5.

2.4 *Predicting Change*

While most studies in mental health prediction focus on contemporaneous or prospective outcomes at a single time point, fewer have attempted to predict change in mental health over time, especially in a categorical format (e.g., improved, stable, deteriorated). However, understanding how and why mental health changes—rather than simply assessing current status—is critical for understanding mental health trajectories or developmental trends. This approach complements cross-sectional as well as longitudinal analyses by explicitly modeling the direction of change, offering insight into factors associated with improvement, stability, or deterioration over time. Despite its importance, studies specifically focusing on predicting change in MHI-5 scores using machine learning are virtually nonexistent. In adjacent fields, however, some efforts have been made to model change over time. For instance, Zacher and Rudolph (2021) examined predictors of mental health change during COVID-19 using repeated survey data, though without employing classification algorithms.Similarly, studies in other health domains have used longitudinal modeling to capture individual change trajectories—for example, in cognitive decline (Petersen et al., 2018) and BMI development (Harring et al., 2006).These studies illustrate how repeated measurements can be used to track and predict meaningful changes in health status over time. Although these applications have traditionally relied on statistical methods such as linear mixed models or trajectory analysis, the same principles can be adapted within supervised machine learning frameworks to classify individuals based on whether their mental health improves, worsens, or remains stable. The application of machine learning techniques to predict categorical changes in mental health status over time represents a significant gap in the current literature.

3 METHOD

This section outlined the comprehensive methodology employed in this thesis with a flow-chart (See figure 1) and elaboration on its components within the preprocessing and modeling stages of this research. Each step is crucial for ensuring reliable MHI-5 outcome predictions.

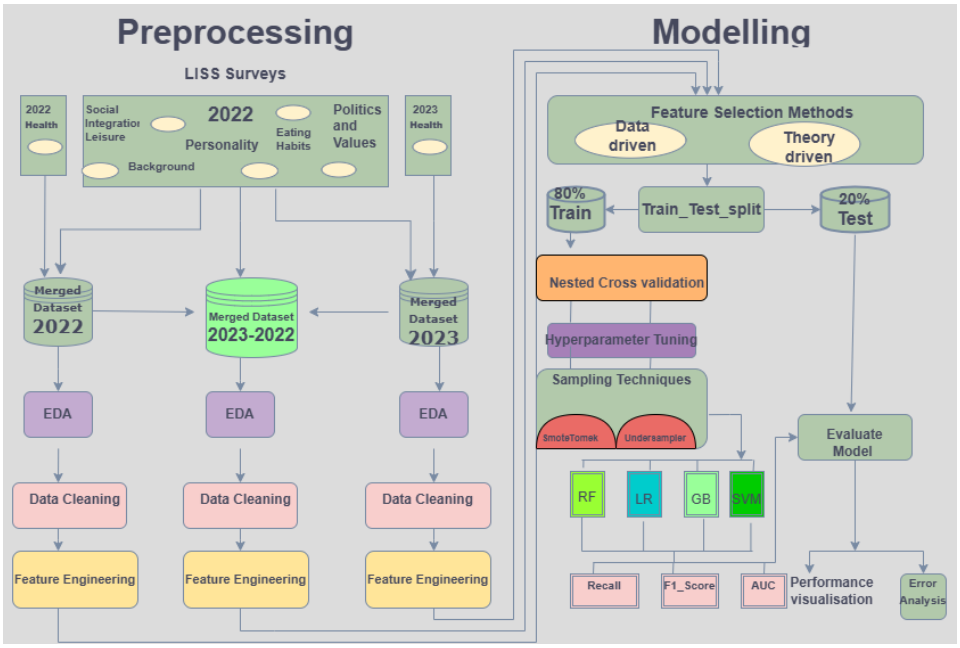

Figure 1: Overview of Methodology (Flow-Chart)

3.1 *Data Description*

3.1.1 *Source*

The data for this study comes from the Longitudinal Internet Studies for the Social Sciences panel (LISS) (Scherpenzeel, 2018). Since late 2007, this internet research platform has provided academics with publicly available data upon request. The LISS panel includes 5,500 to 6,000 participants in every wave from randomly chosen Dutch homes aged 16 and above. The participants in the LISS panel are typical of the Dutch population (De Vos, 2010), allowing for confident generalization of findings to the larger population.

3.1.2 *Structure*

Each annual wave of the LISS panel comprises responses from the same participants to the 'core studies', which are surveys covering various themes. Background characteristics, including age, gender, and income, are measured monthly. Additional surveys can be distributed to the panel, and all data collected can be linked to respondents through their unique encrypted identifiers. This is the `nomem_encr` and it is the number of the household member encrypted as an integer. This approach facilitates the creation of datasets containing a wide range of variables for the same individuals tracked across multiple years within the LISS panel data archive.

In addition to background variables, this research utilizes datasets that combines information from several 'core studies' within the LISS panel from year 2022 and 2023. Relevant predictive variables from surveys including, 'Social Integration and Leisure', 'Health', 'Personality', 'Politics and Values', 'Eating Habits' and 'Religion and Ethnicity' are merged. Details on these variables and their sources are listed in Table Appendix A. Additionally, the target variable—respondents' MHI-5 questionnaire responses—was taken from the 2022 Health Core Study for the contemporaneous analysis, and from the 2023 wave for the Year Ahead study. This process resulted in three merged datasets. The first, Dataset 2022, includes all relevant variables along with the target variable (MHI-5) from 2022. The second, Dataset 2023, is identical, except that the target variable is taken from 2023. The third, Dataset "The Difference", is also identical, except that the target variable represents the difference score between MHI-5 values in 2022 and 2023. Only subjects with complete data across all variables were retained.

3.2 *Preprocessing and EDA*

3.2.1 *Exploratory Data Analysis (EDA)*

Exploratory Data Analysis (EDA) was conducted to gain preliminary insights into the structure and relationships within the dataset. Distributions and box plots were reviewed to assess skewness and identify anomalies; no extreme outliers were detected. Likert-scale items were treated as continuous variables. Pearson correlation analysis revealed limited multicollinearity overall, but in cases of high correlation between variables, one variable was removed to reduce redundancy and improve model performance. Cramér's V was computed to assess associations among categorical variables (see Appendix). Scatter plots were used to explore linear and nonlinear patterns between predictors and the target variable. Importantly, EDA also revealed the need to reverse-code two MHI-5 items, which con-

sist of 5 items, before summing the scale, ensuring accurate calculation of the total mental health score based on consistent item directionality. These analyses informed subsequent feature engineering and modeling strategies.

3.2.2 *Data Cleaning*

Data cleaning is an essential step to ensure data quality by addressing missing values, errors, and outliers, which could negatively impact model performance. This cleaning process enhances the reliability of the input data, leading to more accurate and robust classifier predictions.

The datasets were filtered to include only subjects with data for all relevant columns. Subsequently, duplicates were removed. The decision to remove duplicates presents a trade-off: while it may result in the loss of potentially valuable data points, retaining duplicates introduces data leakage and unreliable model accuracy scores (Tamilselvi & Gifta, 2011).

After data filtering and removing duplicates, the final dataset sizes were as follows:

- 2022 dataset: 995 entries
- 2023 dataset: 895 entries
- Change dataset (2022–2023): 895 entries

3.2.3 *Missing Data Imputation*

Multiple imputation is a powerful technique for handling missing data in datasets. When data is missing not at random, simply excluding cases with missing values can introduce significant biases and reduce statistical power, leading to increased Type I errors (Little & Rubin, 2002; Rubin, 1976). Multiple imputation addresses this issue by creating multiple plausible versions of the incomplete dataset, each with the missing values replaced by imputed values. These imputed values are estimated based on the observed data, taking into account the uncertainty associated with the missing data. By analyzing each of the imputed datasets separately and then combining the results using specialized rules, multiple imputation produces unbiased estimates and valid statistical inferences (Van Buuren, 2018). The MICE (Multivariate Imputation by Chained Equations) algorithm is an implementation of multiple imputation (Van Buuren & Groothuis-Oudshoorn, 2011). The `mice` package in R was used to impute missing values in both datasets, specifically for the income column, which contained a substantial amount of missing data. For other features with less critical missingness, simple imputation using scikit-learn's `SimpleImputer` was applied.

3.2.4 *Feature Selection*

Feature engineering techniques, including feature selection, were applied to address the curse of dimensionality, reduce overfitting, and improve classifier performance and computational efficiency. As an initial step, a correlation matrix (excluding the target variable) was used to identify and remove highly correlated features, minimizing redundancy.

Two distinct feature selection strategies were employed: a theory-driven approach and a data-driven method. For the literature-based approach, key predictors of mental health outcomes were selected from the longitudinal LISS panel dataset based on a review of relevant literature, particularly the peer-reviewed study by van der Velden (van der Velden, Setti, et al., 2019a). Variables were grouped into four conceptual blocks: demographic (e.g., age, gender, education, employment), health (e.g., sleep problems, snacking, psychedelic use), social factors (e.g., social media use, loneliness), and psychology (e.g., personality traits, self-esteem, values). Psychological constructs were represented by the Big Five personality traits—emotional stability, extraversion, agreeableness, conscientiousness, and intellect/imagination (Goldberg et al., 2006)—as well as by self-esteem (Rosenberg, 2015) and a composite value orientation score.The latter, referred to in the dataset as `composite_values`, was constructed from standardized scores of both instrumental and terminal values, following Rokeach's theory of human values. These two value types were combined into a single metric to represent personal life values (Rokeach, 1973). For multi-item features, scale directions were inverted as necessary. A series of models were trained to evaluate the predictive value of each block: a baseline model included demographic and social features, while additional models assessed the impact of adding the health or psychology blocks, as well as a combined model with all four blocks.

For the data-driven approach, two techniques were utilized. Lasso, an L1 regularization method (Tibshirani, 1996), assumes linear relationships between features and the target variable. It uses L1 regularization to penalize large coefficients, encouraging the shrinking of less important features' coefficients towards zero. This helps prevent overfitting, improves model generalizability, and enhancing interpretability.In our analysis, the Lasso-based feature set consisted of 13 features with non-zero coefficients. Complementarily, the `feature_importances_` attribute of Random Forest provides a robust alternative for mixed data types and nonlinear data structures, evaluating feature significance by calculating each feature's contribution to reducing impurity metrics such as Gini impurity or mean squared error (Breiman, 2001).For this approach, the top 10 features based on importance ranking were selected for model training.

Table 1 summarizes the features selected through both theory-driven and data-driven approaches. The Lasso-selected features were identified using five-fold cross-validation, while Random Forest feature importance was based on a single train-test split.

Table 1: Overview of Feature Sets Used in Model Development

Feature Set	**Description**	**Features**
Demographic	Theory-driven	`age`, `sex`, `education`, `employment_status`, `Personal_Net_Income_Category`, `living_arrangement`, `marital_status`
Social	Theory-driven	`hours_on_social_media`, `loneliness_score`
Health	Theory-driven	`SmallSnacks_Daily`, `LargeSnacks_Weekly`, `used_hallucinogens`, `Self_Rated_Health`, `Health_Hindrance_to_Daily_Functioning`, `Sleeping_Problems`
Psychological	Theory-driven	`emotional_stability`, `intellect_imagination`, `conscientiousness`, `agreeableness`, `extraversion`, `self_esteem`, `composite_values`
Lasso-selected	Data-driven	`social_media_frequency`, `age`, `marital_status`, `living_arrangement`, `emotional_stability`, `intellect_imagination`, `self_esteem`, `Self_Rated_Health`, `Health_Hindrance_to_Daily_Functioning`, `Sleeping_Problems`, `Personal_Net_Income_Category`, `Employment_status`, `hours_on_social_media`
Random Forest Feature Importance	Data-driven	`emotional_stability`, `self_esteem`, `Self_Rated_Health`, `age`, `Health_Hindrance_to_Daily_Functioning`, `Sleeping_Problems`, `conscientiousness`, `composite_values`, `intellect_imagination`, `extraversion`

3.2.5 *Target Variables and Class Imbalance*

The MHI-5 score was calculated by summing responses to five Likert-scale items measuring mental health symptoms (see Table 2), transformed to a 0–100 scale. For binary classification, a threshold of 60 was applied to distinguish between individuals with "no mental health issues" (≥ 60) and "mental health issues" (< 60). Kelly et al. (2008) validated this cutoff using the misclassification rate minimization method in their methodological study comparing five statistical approaches for defining MHI-5 thresholds. This procedure was applied consistently to both the 2022 and 2023 datasets.

Table 2: MHI-5 items and response scale. Items 2 and 4 are reverse-scored (higher values indicate better mental health), while items 1, 3, and 5 are directly scored (higher values indicate poorer mental health).

Item	**Question**	**Response Scale**
1	How much of the time during the past month have you been a very nervous person?	All of the time (1) – None of the time (6)
2	During the past month, how much of the time have you felt calm and peaceful?	All of the time (6) – None of the time (1)
3	How much of the time during the past month have you felt downhearted and blue?	All of the time (1) – None of the time (6)
4	During the past month, how much of the time have you been a happy person?	All of the time (6) – None of the time (1)
5	How much of the time during the past month have you felt so down in the dumps that nothing could cheer you up?	All of the time (1) – None of the time (6)

Note. Items reproduced from the MHI-5 scale (Kelly et al., 2008).

Figure 2 shows the histogram of the target variable distribution for both years, illustrating a clear class imbalance.

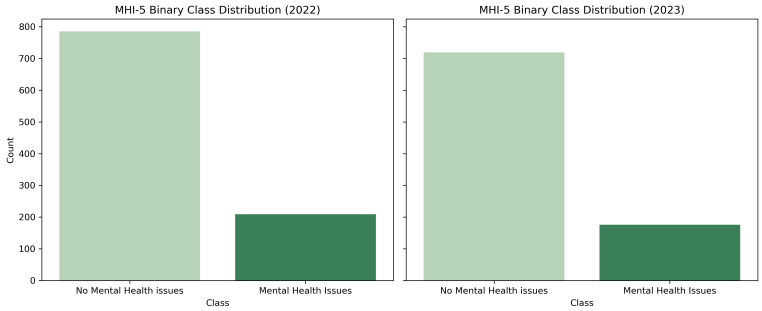

Figure 2: Distribution of MHI-5 Binary Class for 2022 and 2023

Class imbalance can lead to biased classifiers that favor the majority class. As noted by Chawla et al. (Chawla et al., 2002), a classifier may consistently predict the majority class, resulting in deceptively high accuracy scores that do not reflect true model performance. One approach to addressing this issue was to compare three balancing strategies: no balancing (baseline), random undersampling, and SMOTE-Tomek.

Using no balancing served as a baseline to assess the effect of sampling techniques on model performance. Random undersampling mitigates class imbalance by reducing the number of majority class instances, though it may result in information loss (He & Garcia, 2009). SMOTE-Tomek was selected as a hybrid approach combining SMOTE's synthetic minority oversampling with Tomek Links' data cleaning. While SMOTE alone can introduce noise and class overlap, the Tomek component helps eliminate ambiguous instances (Chawla et al., 2002; Sasada et al., 2020). Prior research (Le et al., 2018) has shown SMOTE-Tomek to outperform alternatives like Borderline-SMOTE and ADASYN. In this study, SMOTE-Tomek improved the performance of the Random Forest classifier by 1.7%.

For the third research question examining changes in mental health, a new target variable was constructed by calculating the difference between standardized MHI-5 scores from 2022 and 2023. These difference scores were then categorized into three classes: (1) improved mental health (positive difference scores), (2) stable mental health (difference scores near zero), and (3) declined mental health (negative difference scores). To ensure balanced representation across categories, thresholds were carefully selected to create three approximately equal-sized groups, transforming the analysis into a balanced multiclass classification problem with a three-class target variable, as illustrated in Figure 3.

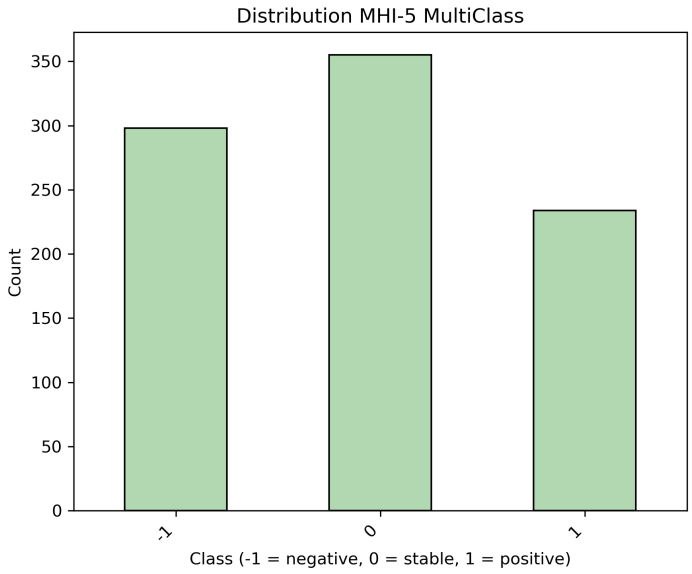

Figure 3: Distribution of the 3-Class Target Variable Representing Change in MHI-5 (2022–2023)

3.2.6 *Data Partition*

To partition the data, I employed a stratified 80-20 split using `train_test_split` from the `scikit-learn` library (see Appendix 5.5), where 80% was allocated for training and 20% for testing. The stratified sampling was crucial due to the significant class imbalance in the target variable and to ensure representative datasets. This stratification ensured that both training and test sets maintained the same proportion of samples for each class as the original dataset, preventing sampling bias and enabling more reliable model evaluation.

3.3 *Modelling*

3.3.1 *Scaling*

The `StandardScaler` was used to standardize continuous features by transforming them to have a mean of zero and a standard deviation of one. This process helps ensure that the features are on a comparable scale, which can improve the performance of certain algorithms, particularly those sensitive to feature magnitudes, such as Support Vector Machines and Logistic Regression (Géron, 2017). Note that scaling was applied only to continuous variables; categorical variables, which are typically handled via one-hot encoding, were not present in this dataset, as the LISS Panel data did not include categorical features requiring encoding. Feature scaling was

performed within the inner loop of the nested cross-validation to avoid data leakage. This process is described in more detail in a later section.

3.3.2 *Models*

Four machine learning models were used to predict MHI-5 scores: Logistic Regression, Random Forest, Support Vector Machine (SVM), and Gradient Boosting. Model selection was informed by prior cross-sectional and longitudinal studies reviewed in this thesis.

3.3.3 *Logistic Regression*

Logistic Regression is a linear model used for binary classification tasks. It predicts the probability of a binary outcome by modeling the relationship between predictors and the log-odds of the target variable (Cramer, 2002). This model was selected because it has been previously used by van der Velden et al. (van der Velden, Setti, et al., 2019a) to analyze MHI-5 using LISS panel data. Logistic Regression also serves as a baseline model due to its simplicity and interpretability, offering insights into the linear relationships between predictors and MHI-5.

Table 3: Hyperparameter tuning – Logistic Regression

Hyperparameter	**Values tested**
C (inverse regularization strength)	0.01, 0.1, 1, 10

3.3.4 *Random Forest*

Random Forest is an ensemble learning method introduced by Breiman (Breiman, 2001) that constructs multiple decision trees using bootstrap sampling and random feature selection at each node. The model's prediction is determined by majority voting across all trees (Parmar et al., 2019). Particularly effective for non-linear relationships and feature interactions, Random Forest has demonstrated robust performance in mental health outcome prediction (G. Li & Jiang, 2023; Rahimapandi et al., 2022).

Hyperparameters tuned include `n_estimators`, `max_features`, `max_depth`, `min_samples_split`, and `min_samples_leaf`, with candidate values displayed in Table 4.

Table 4: Hyperparameter tuning – Random Forest

Hyperparameter	Values tested
`n_estimators`	50, 100, 150, 200, 250
`max_depth`	None, 5, 10, 15, 20
`max_features`	'sqrt', 'log2'
`min_samples_split`	2 to 10
`min_samples_leaf`	1 to 4

3.3.5 *Support Vector Machine (SVM)*

Support Vector Machine (SVM) is a supervised learning algorithm that finds the optimal hyperplane separating data classes by maximizing the margin between support vectors (the closest data points to the hyperplane), thereby improving the model's generalization ability (Cortes & Vapnik, 1995). For non-linearly separable data, SVM employs kernel functions such as the Radial Basis Function (RBF) to map data into higher-dimensional spaces. The regularization parameter `C` controls the trade-off between maximizing the margin and minimizing misclassification errors. SVM handles high-dimensional data effectively, especially when combined with class balancing techniques. Hyperparameters tuned for SVM are summarized in Table 5.

Table 5: Hyperparameter tuning – Support Vector Machine

Hyperparameter	Values tested
`kernel`	'linear', 'rbf', 'poly'
`C`	0.1 to 100 (log scale)
`gamma`	'scale', 'auto', 0.1, 0.01, 0.001
`degree`	2 to 5
`class_weight`	None, 'balanced'

3.3.6 *Extreme Gradient Boosting (XGBoost)*

Extreme Gradient Boosting (XGBoost) is an advanced ensemble learning algorithm based on gradient boosting, designed to enhance predictive accuracy and computational efficiency (Chen & Guestrin, 2016). Unlike Random Forest, which constructs trees independently in parallel, XGBoost

builds decision trees sequentially, with each new tree correcting the errors of its predecessors. This iterative approach allows XGBoost to capture more complex patterns, including higher-order feature interactions, leading to improved predictive performance.

A key advantage of XGBoost over Random Forest is its regularization techniques (L1 and L2 penalties), which reduce overfitting—particularly relevant when handling high-dimensional data (Ke et al., 2017). Additionally, XGBoost efficiently processes missing values and handles imbalanced datasets through weighted loss functions and sampling techniques. Its computational efficiency is further enhanced by parallel processing, tree pruning, and early stopping, making it well-suited for large-scale data analysis (Natekin & Knoll, 2013). The hyperparameters considered for tuning and their candidate values are shown in Table 6.

Table 6: Hyperparameter tuning – XGBoost

Hyperparameter	**Values tested**
`n_estimators`	50, 100, 200, 400, 500
`learning_rate (eta)`	0.01 to 0.1 (log-uniform)
`max_depth`	2, 4, 6, 10, 15
`min_child_weight`	1, 3, 5
`subsample`	0.6, 0.8, 1.0
`colsample_bytree`	0.6, 0.8, 1.0
`gamma`	0 to 1 (step 0.1)
`reg_alpha (L1)`	0.001 to 5 (log-uniform)
`reg_lambda (L2)`	0.001 to 5 (log-uniform)
`early_stopping_rounds`	10

3.3.7 *Hyperparameter Tuning*

Effective machine learning models require careful calibration of two distinct types of parameters: model parameters that are automatically adjusted during training, and hyperparameters that demand manual configuration and can profoundly impact model performance. For this study, a dual approach to hyperparameter optimization was implemented: GridSearchCV for Logistic Regression and Optuna for SVM, Random Forest, and XGBoost models. GridSearchCV exhaustively evaluates all hyperparameter combinations within a defined search space—appropriate for Logistic Regression's relatively small parameter space. For the more complex models,

Optuna's Bayesian optimization approach was employed, which intelligently explores the hyperparameter landscape by learning from previous trials to focus on promising configurations, reducing computational cost while improving results.

3.3.8 *Performance Evaluation*

To assess the predictive performance of the models, multiple evaluation metrics were applied across the three research questions. Given the significant class imbalance in the dataset—with mental health issues representing the minority class—metrics specifically designed to handle skewed distributions were prioritized. Additionally, systematic temporal comparisons were conducted to quantify both the models' generalization capability across different time periods and their predictive stability when facing evolving mental health patterns.

For RQ1 (Predicting MHI-5 in 2022), classification performance was evaluated using, binary F1-Score and Recall to ensure robust assessment. Binary F1-score was specifically selected over weighted F1-score as it focuses on the minority class (those experiencing mental health issues), which aligns with our primary interest, which is having a model that predicts individuals with mental health issues correctly. While weighted F1-score would give disproportionate importance to the majority class due to the substantial class imbalance, binary F1-score provides a balanced measure of precision and recall specifically for individuals with mental health concerns. This choice is particularly important in mental health interventions, where the cost of false positives can be significant. Additionally, Recall was included as a critical metric since it directly measures how many individuals with mental health issues are correctly identified—a crucial consideration when the goal is to provide appropriate support and intervention to those who need it, rather than optimizing for overall accuracy.—-Besides, We evaluated classification performance using both AUC-ROC and Precision-Recall (PR) curves. While AUC-ROC is a standard binary classification metric, it can be misleading in imbalanced datasets. The PR curve provides a more nuanced view by focusing on the positive class (individuals with mental health issues), revealing the trade-off between precision and recall. This approach ensures a comprehensive assessment of model performance, particularly when identifying the minority class is critical. — Confusion matrix was also used to help identify misclassifications. Permutation Feature Importance (PFI) was used to analyze the contribution of each predictor to the classification decision.

For RQ2 (testing temporal generalization and predictive stability), the same classification metrics were applied to evaluate the model's ability to generalize to future data. In addition to hold-out test performance, we

assessed predictive stability through a paired t-test comparing fold-level F1-scores from a nested cross-validation on a matched sample of participants present in both years (N = 894). To further understand shifts in model behavior and predictor relevance, confusion matrices and Permutation Feature Importance were examined for both 2022 and 2023 hold-out test sets.

For RQ3 (predicting the change in MHI-5 scores between 2022 and 2023), evaluation metrics were adapted to reflect the multiclass nature of the target variable. Macro F1 score, recall, AUC-ROC, and precision recall (PR) curves were used. Macro F1-score was selected because it calculates metrics independently for each class and then takes an unweighted mean, treating all classes equally regardless of their frequency in the dataset. This was particularly appropriate as we intentionally divided the target variable into three balanced categories: improvement, stability, and deterioration in mental health. By using macro averaging, we ensured fair performance evaluation across all change categories without bias toward any particular class.Confusion matrices for multiclass predictions helped identify specific misclassifications. PFI was applied to determine the most influential features driving mental health changes over time.as we intentionally divided the target variable into three balanced classes.

3.3.9 *Experimental Setup*

This study followed a structured experimental workflow that builds upon and extends the research design of van der Velden et al. (van der Velden, Setti, et al., 2019a).The process included preprocessing and exploratory data analysis, model training using nested cross-validation, hyperparameter tuning, evaluation on a hold-out test set, and statistical significance testing to compare model performance. While van der Velden et al. measured MHI-5 scores at two time points (short-term and long-term), their approach was primarily descriptive, examining which predictors maintained significance when additional variables were introduced. This study replicates their temporal design by analyzing data from 2022 and 2023, but employs a machine learning approach to develop predictive models rather than focusing solely on predictor significance. This distinction is critical—a variable like social media usage might lose statistical significance in a descriptive model yet still contribute to prediction accuracy in a machine learning context.

This research also expands the feature set beyond van der Velden, Setti, et al. focus on social and demographic factors by incorporating psychological and health-related variables. To systematically assess the impact of these additional feature categories, six different feature combinations were evaluated for each time period: (1) demographic and social factors only (*So-*

cio+Demo), (2) demographic, social, and health factors (*Socio+Demo+Health*), (3) demographic, social, and psychological factors (*Socio+Demo+Psy*), (4) all factors combined (*Full*), (5) all factors with Random Forest feature importance selection (*RF FeatureImportance*), and (6) all factors with Lasso technique selection (*Lasso*). This comprehensive approach, resulting in 18 distinct analytical notebooks (6 feature sets across 3 datasets), allowed for direct comparison of how different factor combinations influence mental health predictions in short-term (2022), long-term (2023), and longitudinal change settings.

The `imbPipeline` from `scikit-learn` was implemented to address the class imbalance in the target variable. Feature scaling was incorporated into the pipeline to ensure uniform treatment of variables with different distributions and ranges.

A 5×3 fold nested cross-validation approach was employed to ensure reliable model evaluation. The inner loop performed hyperparameter tuning and compared three different balancing/sampling techniques. The outer loop was used for final model selection, where different models were compared based on their mean binary F1 score, weighted F1 score, standard deviation, and recall to assess class-specific performance.

Models demonstrating both higher average performance and lower standard deviation during the training phase were prioritized. Once the best model configurations were selected, they were evaluated on a 20% hold-out test set to assess their final generalization performance.

Finally, to assess whether the overall best-performing model—defined as the model with the highest average F1 score across all feature sets—significantly outperformed the baseline model (replicating van der Velden's demographic and social feature approach), a paired sample t-test was conducted. This statistical test was applied separately for predictions on the 2022 and mental health change datasets. To evaluate temporal stability of the predictive models, an additional analysis was performed using participants present in both 2022 and 2023 datasets (n=894). The best-performing model was trained and evaluated on both years' matched data using identical procedures, with paired t-tests comparing cross-validation performance scores to assess stability over time. To further interpret model behavior and identify the most influential predictors contributing to MHI-5 classification, Permutation Feature Importance (PFI) was applied to the final models, with feature importance rankings compared across time periods to evaluate predictor consistency.

As a final exploratory step for RQ3, we tested whether explicitly modeling the ordinal structure of mental health change (*deteriorated* < *stable* < *improved*) improved classification. Prior research suggests that ordinal models may reduce misclassification errors by respecting the inherent order

in outcome categories (Cardoso & da Costa, 2007; Rennie & Srebro, 2005). Using `mord.LogisticIT()`, an ordinal logistic regression model, we trained on the same RF-based feature set and stratified train-test split as in the nominal 3-class task. Macro F1-score and confusion matrices were used for evaluation. Although not part of the primary model comparison, this analysis explored the potential added value of ordinal modeling.

4 RESULTS

This section presents the results of the 5×3 nested cross-validation of the model comparison of the four machine learning models, as well as their performance on the 2022 and 2023 scores of the hold out test set participants, as well as the mental health change classification. Additionally, the outcomes of the paired t-tests and the error analyses are reported.

4.1 *Results research question 1*

4.1.1 *Model Comparison Based on F1 binary and Recall (2022)*

Figure 4 shows the distribution of F1 binary scores across the 5×3 nested cross-validation for each model trained on the 2022 dataset. Figure 5 presents the corresponding recall score distributions. `LR-Full` achieved the highest mean F1 score ($\sim$0.67), while `LR-Lasso` demonstrated strong recall performance (mean $\sim$0.83). However, `LR-RF_FeatureImportance` was selected as the optimal model.

Statistical validation using the Friedman test confirmed significant differences among models overall ($p < .05$), indicating that model selection is statistically meaningful. However, subsequent Nemenyi post-hoc analysis revealed that *LR-RF_FeatureImportance* and *LR-Full* are not significantly different (rank difference: 1.4 < critical difference: 3.37), demonstrating statistical equivalence in performance. This finding validates that both models achieve comparable predictive capability.

Given this statistical equivalence, model selection was based on practical considerations favoring parsimony. *LR-RF_FeatureImportance* was selected as the optimal model despite not achieving the absolute highest mean scores, maintaining strong performance (F1 $\sim$ 0.65, recall $\sim$ 0.75). Most importantly, *LR-RF_FeatureImportance* utilizes only the top 10 features selected through random forest importance ranking, compared to *LR-Full* (all features/blocks) and *LR-Lasso* (13 features). This more parsimonious and data-efficient feature set significantly reduces model complexity and potential overfitting risk while maintaining equivalent predictive power.

The principle of parsimony (Occam's Razor) supports selecting the simpler model when performance is statistically equivalent (McFadden, 2023). Fewer features mean reduced computational requirements, improved interpretability, and lower risk of overfitting to training data peculiarities. Additionally, a more compact feature set enhances model robustness and facilitates real-world implementation where feature collection may be resource-constrained.

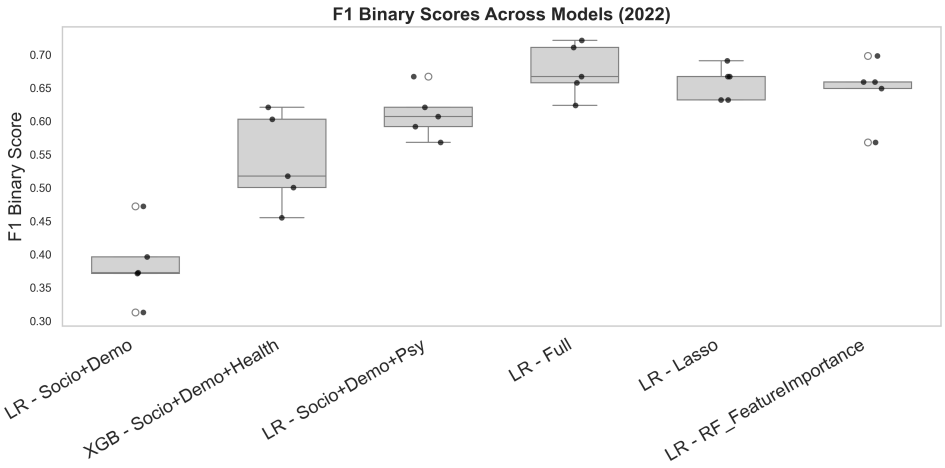

Figure 4: Distribution of F1 binary scores for each model trained on the 2022 dataset, based on 5 outer test folds from nested cross-validation. Each box shows the interquartile range (IQR) of the F1 scores, with the line indicating the mean. The black dots represent individual fold test scores.

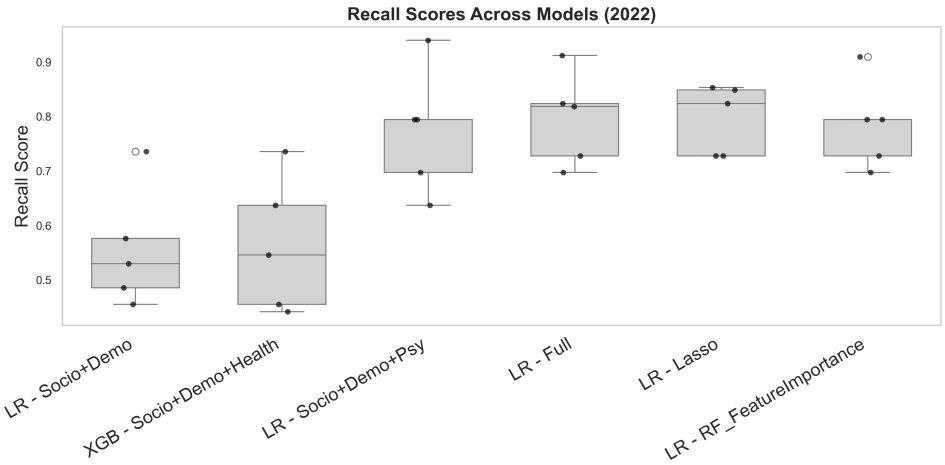

Figure 5: Distribution of recall scores for each model trained on the 2022 dataset, based on 5 outer test folds from nested cross-validation. Each box shows the interquartile range (IQR) of the recall scores, with the line indicating the mean. The black dots represent individual fold test scores.

4.1.2 *Test Set Results*

Table 7 shows the best-performing Logistic Regression (LR) model's hyperparameters along with the sampling strategy used to address class imbalance. After hyperparameter optimization, the model was subsequently evaluated on the held-out test set. The results, shown in Table 8,

indicate a slight decrease in F1 score and recall for class 1 compared to the nested cross-validation performance. The gap between training and test performance is called the *generalization gap*. A small, acceptable gap is normal — it's a sign of good generalization. A large gap suggests overfitting. In this case, the LR model maintained strong test performance with minimal generalization gap, supporting its robustness.

The selected model used `RandomUnderSampler`, which reduces the dominance of the majority class by balancing the class distribution through undersampling. This often enhances the model's ability to detect minority class instances—in this context, individuals at risk of poor mental health (class 1). The relatively balanced recall scores across both classes suggest that this sampling strategy was particularly well-suited to the LR model and the underlying data, leading to reliable and generalizable predictions.

Table 7: Best Logistic Regression Model and Sampling Strategy (RF Feature Importance, 2022)

Component	**Value**
`classifier__C`	0.01
`sampling strategy`	RandomUnderSampler

Table 8: Test Set Performance of the Best Model (Logistic Regression with RF Feature Importance, 2022)

Metric	**Value**
F1 Binary (class 1)	0.6095
Recall (class 0)	0.8025
Recall (class 1)	0.7619
AUC-ROC	0.8615
PR AUC	0.6368

Figure 6 shows the confusion matrix for the best-performing model (Logistic Regression) on the 2022 test set. The model demonstrates a relatively balanced error distribution, with a slight bias toward identifying individuals as at risk. In the context of mental health screening, such a bias is advantageous: prioritizing sensitivity helps ensure that individuals with potential mental health problems are more likely to be flagged, aligning with the goal of early identification and intervention.

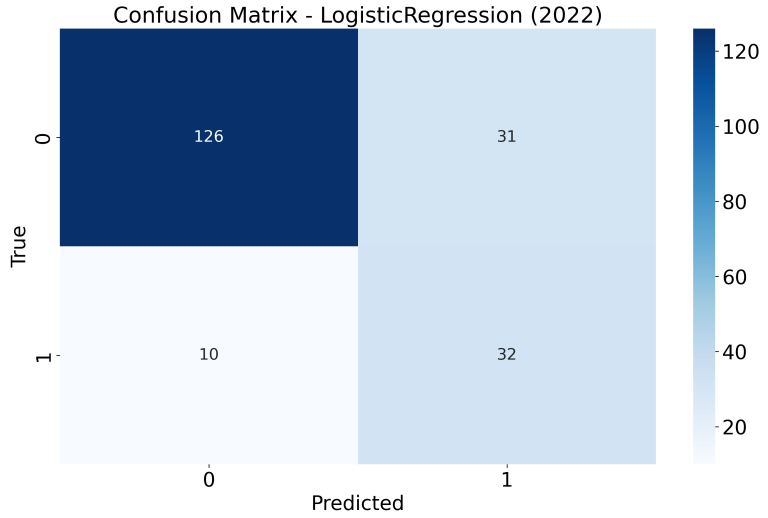

Figure 6: Confusion matrix of the best-performing model (Logistic Regression with RF FeatureImportance, test set 2022).

Although the model was trained on features selected using Random Forest importance, we additionally applied permutation feature importance to assess the true predictive value of each feature. It does this by measuring the drop in performance when a feature's values are randomly shuffled. This makes permutation importance particularly valuable for verifying whether top-ranked features do, in fact, influence the final predictions. The results show that features like *emotional_stability*, *Health_Hindrance_to_Daily_Functioning*, and *age* had the strongest actual impact, reinforcing the relevance of the RF-based selection (see Figure 7). Notably, features such as *conscientiousness*, *self_esteem*, and *composite_values* showed negative or near-zero importance in the permutation analysis, indicating they do not contribute meaningfully to the Logistic Regression model's predictive performance. This suggests their presence may add noise rather than value in this context.

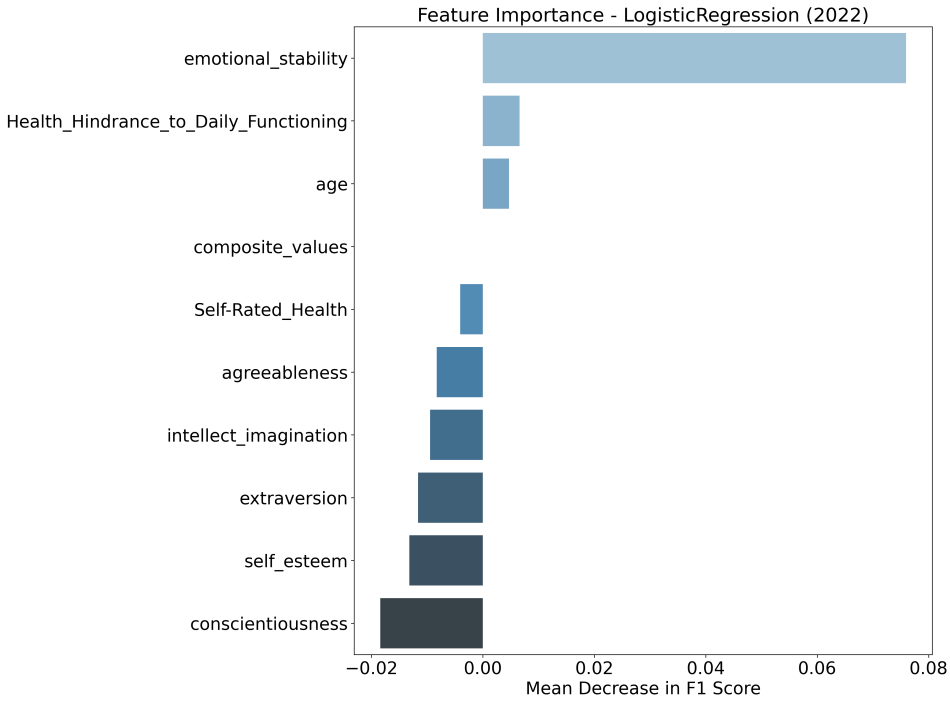

Figure 7: Permutation feature importance for the best-performing model (Logistic Regression with RF FeatureImportance, test set 2022).

4.1.3 *Error Analysis*

Figure 8 shows F1 binary scores across demographic groups for the best-performing model. There were notable differences across education levels, with higher scores among respondents with higher educational attainment. However, the model exhibited minimal variation by gender, generation, or ethnic background, suggesting only limited bias toward these groups. Lastly, respondents identifying as Christian had higher F1 scores than those in the "missing religion" group. In the LISS panel, only participants who indicated they are religious were asked to specify their religion (e.g., Christian, Islam, Buddhism), so the "missing" group likely includes both non-religious individuals and those who chose not to respond.

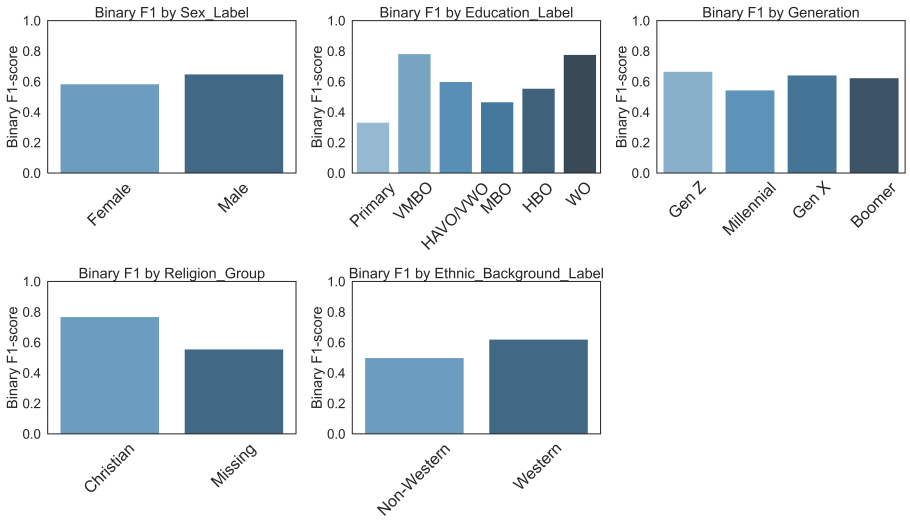

Figure 8: F1 binary scores for the best model (Logistic Regression with RF Feature Importance feature set and Random Undersampler) across demographic groups. The figure shows performance variation by sex, education level, generation, religion, and ethnic background.

4.1.4 *Basic model vs best model: Statistical significance*

Table 9 summarizes the mean and standard deviation of F1 binary and recall scores across the five-fold nested cross-validation for the baseline model (Logistic Regression with socio-demographic features) and the best-performing model (Logistic Regression with RF Feature Importance-based features).

To determine whether the performance improvement was statistically significant, a paired sample t-test was conducted. The result ($t = 8.78$, $p = 0.0009$) indicates a significant improvement in performance, confirming that the RF Feature Importance-based Logistic Regression model significantly outperformed the baseline. Figure 9 illustrates this improvement across individual folds.

Table 9: Cross-validated F1 performance of the baseline and best model (2022)

Model	**F1 Binary**
Baseline (Socio+Demo, Logistic Regression)	0.385 ± 0.051
Best (RF-based Features, Logistic Regression)	0.646 ± 0.043

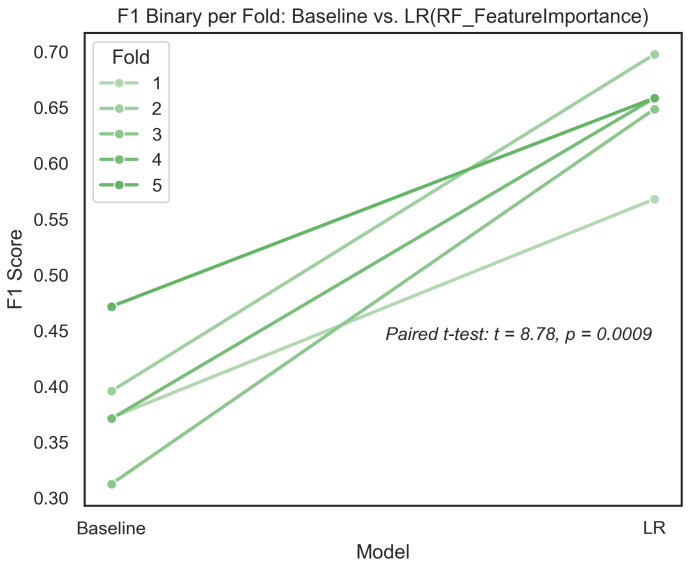

Figure 9: F1 binary scores per fold for the baseline and best model (Logistic Regression with RF-based features). The figure visualizes the performance improvement across all cross-validation folds.

4.1.5 *Samplers*

To address class imbalance in the 2022 dataset, three sampling strategies were systematically evaluated: no sampling, random undersampling, and SMOTE-Tomek. These techniques were integrated into the modeling pipelines and compared during model optimization, as part of the hyperparameter tuning process within the inner loop of the nested cross-validation framework. For each feature set, the combination of model and sampler yielding the highest F1 binary score across folds was identified. Table 15 summarizes the best-performing sampling strategy per feature set. As shown, SMOTE-Tomek was the most frequently selected sampler, demonstrating superior performance in four out of six feature sets. These findings suggest that, although sampling needs may vary depending on the input features, SMOTE-Tomek offered the most consistent performance gains in this study.

Table 10: Best-performing sampling strategy per feature set (Mean F1 Binary ± STD, 2022 dataset)

Feature Set	No Sampling	Random Undersampling	SMOTE-Tomek
Socio + Demo		0.385 ± 0.051	
Socio + Demo + Health			0.539 ± 0.063
Socio + Demo + Psychology			0.611 ± 0.033
Socio + Demo + Health + Psy			**0.676 ± 0.036**
Lasso Selected Features			0.657 ± 0.023
RF Feature Importance		0.646 ± 0.043	

4.2 *Results research question 2*

Generalization Performance: 2022 to 2023

To evaluate the generalizability of the best-performing model, the Logistic Regression model trained on RF Feature Importance features was applied to the 2023 score of the hold-out test set participants. It achieved an F1 binary score of 0.5610, with recall scores of 0.8333 for class 0 and 0.6571 for class 1 (see Table 11). Compared to its 2022 performance, the model was less effective in identifying individuals at risk (class 1), suggesting possible overfitting to patterns in the 2022 data and sensitivity to data shift—that is, changes in the distribution of features or class labels between 2022 and 2023.Nonetheless, this indicates only a modest decline compared to its 2022 test set performance (F_1 = 0.6095), suggesting reasonable generalization to future data.

Table 11: Test Set Performance of the Best Model (Logistic Regression with RF Feature Importance, 2023)

Metric	Value
F1 Binary (class 1)	0.5610
Recall (class 0)	0.8333
Recall (class 1)	0.6571
AUC-ROC	0.8615
PR AUC	0.6368

The 2023 confusion matrix (Figure 10) shows the model became slightly worse at identifying class 1 cases—fewer true positives and more false

negatives than in 2022. While it still handled class 0 well, its sensitivity to the minority class declined.

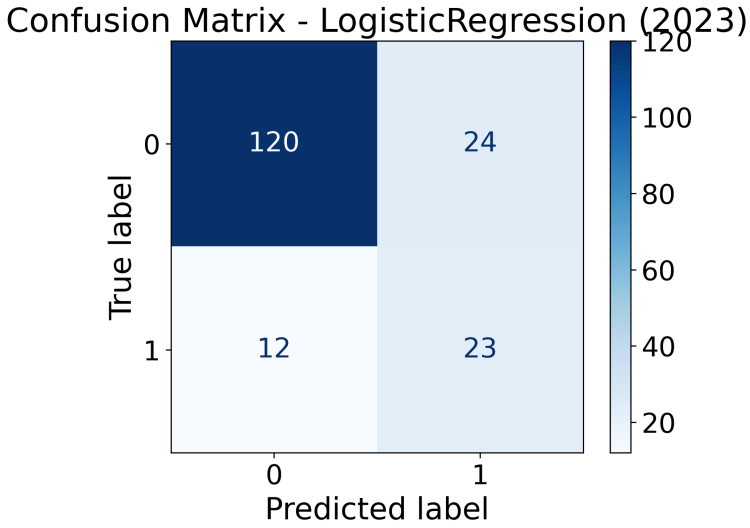

Figure 10: Confusion matrix of the best-performing model (Logistic Regression with RF FeatureImportance, test set 2023).

The 2023 permutation feature importance analysis (Figure 11) shows that psychological variable `emotional_stability` remained the dominant predictor, with its importance increasing compared to 2022. However, other features such as `self_esteem`, `health_hindrance_to_daily_functioning`, and `age` gained relevance, suggesting the model relied more on a broader set of features to make predictions. In contrast, features like `composite_values` and `intellect_imagination` remained largely uninformative.

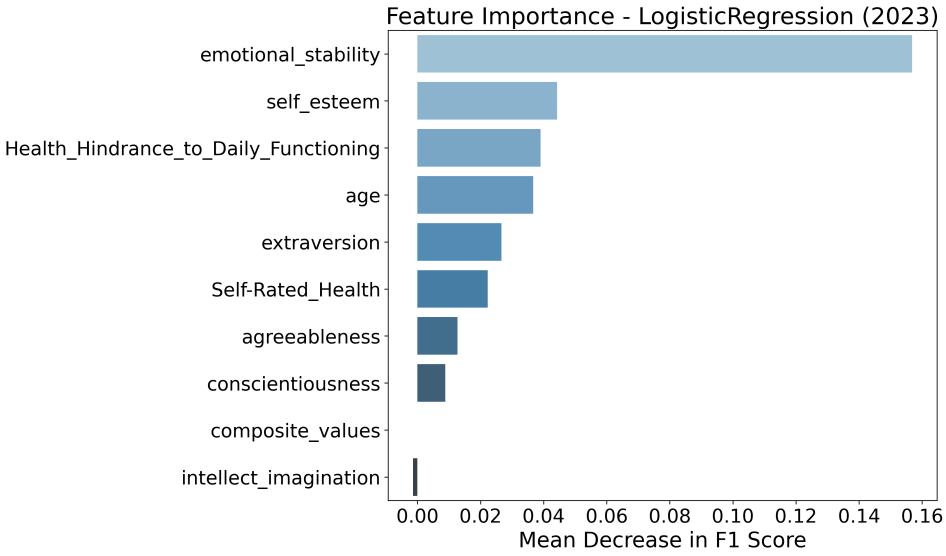

Figure 11: Permutation Feature Importance for 2023

Predictive Stability: 2022 to 2023

To assess the temporal stability of the best-performing model's predictive performance, a paired comparison of nested cross-validation results between matched 2022 and 2023 datasets was conducted. The Logistic Regression model with RF Feature Importance was evaluated using identical 5-fold cross-validation procedures on both years, with folds paired to enable direct statistical comparison of performance metrics.

The paired t-test analysis revealed no statistically significant difference in F1 binary scores between 2022 and 2023 ($t = -0.66$, $p = 0.5461$), indicating that the model's predictive performance remained stable across the two time periods. This stability is visualized in Figure 12, which shows the F1 scores for each corresponding fold pair between years.

The stability analysis suggests that while the model showed some decline when tested on new 2023 data, its overall predictive ability remains stable across time periods.This finding demonstrates that the same predictors can effectively forecast mental health status one year into the future without significant deterioration in performance.

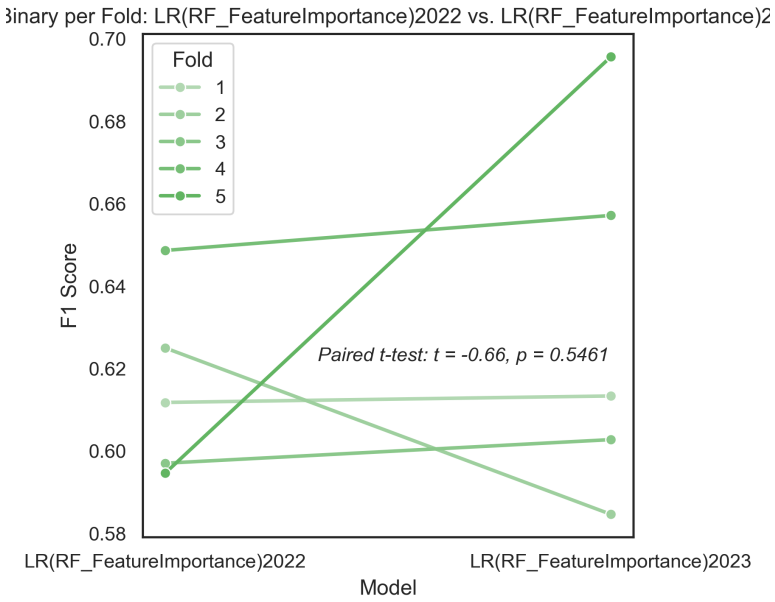

Figure 12: F1 Binary Score Comparison Across Paired Folds: 2022 vs 2023

4.3 *Results research question 3*

4.3.1 *Model Comparison Based on F1 Macro (Three-Class Classification)*

For the overall performance results of the 3-class classification task, see Figure 13 and Figure 14. Among all evaluated models, logistic regression trained on only the top 10 features derived from a Random Forest (RF) feature importance analysis emerged as the best-performing and most robust classifier. This model achieved the highest mean F1 Macro and mean Recall scores among all models tested. It also exhibited the smallest standard deviation in recall across cross-validation folds, indicating low variance and highly consistent performance. This represents an excellent bias-variance tradeoff—the model is complex enough to learn important patterns in the data, yet simple enough to avoid overfitting to noise, as evidenced by its consistent performance across different data subsets.

Using just ten features makes the logistic regression model with the RF Feature Importance set much more efficient and less complex than models like an SVM using the full feature set, thereby reducing the risk of overfitting and improving generalization. Additionally, this reduced feature set translates to substantial computational efficiency gains.

To provide statistical support for the observed differences in model performance, a Friedman test was conducted across all models based on their F1 Macro scores. The result was not statistically significant ($p \geq 0.05$),

indicating that the observed performance differences may be due to random variation across cross-validation folds rather than than one model being consistently better than others.

Despite this, the logistic regression model with RF Feature Importance remains a strong candidate based on its practical advantages: simplicity, consistent cross-validation performance, interpretability, and computational efficiency. While statistical testing did not confirm a significant difference, these practical benefits still position it as a highly suitable model for this classification task.

It is worth noting that the lack of statistical significance may also stem from the limited number of cross-validation folds or sample size. Future work could explore this further by increasing the dataset size or using more extensive cross-validation procedures (e.g., increasing the number of folds or repetitions), which may provide more statistical power to detect meaningful differences between models if they exist.

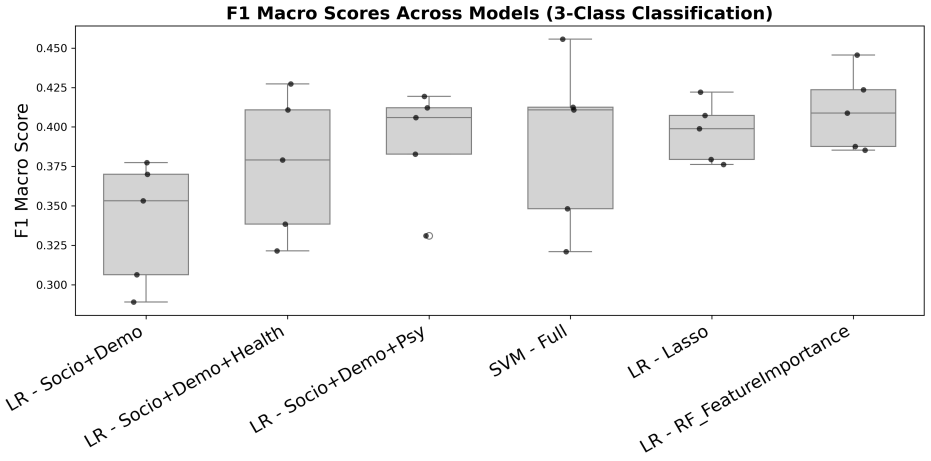

Figure 13: Distribution of F1 macro scores for each model predicting change in MHI-5 (3-class classification), based on 5 outer test folds from nested cross-validation. Each box represents the interquartile range (IQR) of the F1 macro scores across outer folds, with the horizontal line indicating the mean. Black dots correspond to individual fold-level test scores.

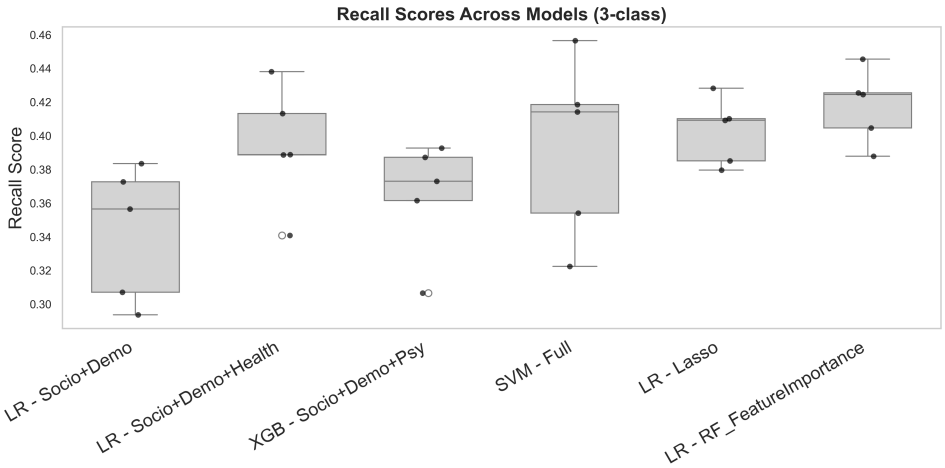

Figure 14: Distribution of recall scores for each model predicting change in MHI-5 (3-class classification), based on 5 outer test folds from nested cross-validation. Each box represents the interquartile range (IQR) of the recall scores across outer folds, with the horizontal line indicating the mean. Black dots correspond to individual fold-level test scores.

4.3.2 *Test Set Results*

Table 12 presents the hyperparameters of the best-performing model, and Table 13 summarizes its performance on the hold-out test set. When evaluating the selected model on the held-out test set, we observed an F1 Macro avg score of 0.35, indicating moderate generalization performance that is slightly lower than the cross-validation result of 0.41.

The model showed varying effectiveness in predicting mental health outcome trajectories, most accurately identifying stable conditions (0.50 recall for class 0), followed by mental health improvements (0.38 recall for class 1), while struggling more with detecting deteriorations (0.20 recall for class -1). This imbalanced performance suggests the model is more sensitive to indicators of stability and improvement than to signals of mental health decline, which may have important clinical implications for early intervention strategies.

Table 12: Best hyperparameters for the Logistic Regression model (3-class prediction task)

Hyperparameter	**Value**
C (Inverse regularization)	0.01
Sampler	RandomUnderSampler (random_state=42)

Table 13: Test set performance of the best model (Logistic Regression with RF-based features) for 3-class classification.

Metric	Value
F1 Macro avg	0.35
Recall (class -1)	0.20
Recall (class 0)	0.50
Recall (class 1)	0.38
AUC-ROC (Macro)	0.52

The confusion matrix in Figure 15 shows that the model most reliably identifies the stable class (class 0), with the highest number of correct predictions (36 instances). In contrast, its classification of the deterioration and improvement classes is less accurate, with frequent confusion between the two. A substantial number of deterioration cases were misclassified as improvements (24 instances), and vice versa (13 instances). Additionally, the model often over-predicts stability, labeling many deterioration (24) and improvement (16) cases as stable. These results suggest that the model struggles to distinguish between negative and positive changes in mental health, potentially because the feature set captures stability-related patterns more effectively than those associated with change.

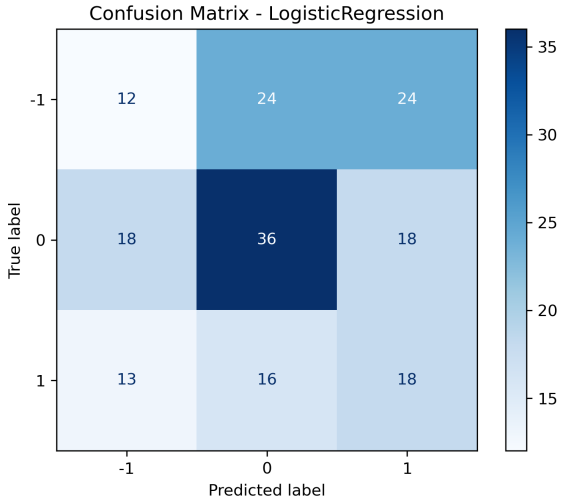

Figure 15: Confusion matrix for the 3-class mental health change model.

Permutation feature importance analysis reveals which features most influenced the model's predictions. The top three features were *intellect_imagination*, *composite_values*, and *emotional_stability*, as shuffling any of

these caused the largest drop in the model's macro-F1 score. This indicates that these traits provided the most useful signal for predicting mental health outcomes. Notably, *conscientiousness* exhibited a slightly negative importance value, implying it did not aid the predictions and is an unreliable predictor in the current model.

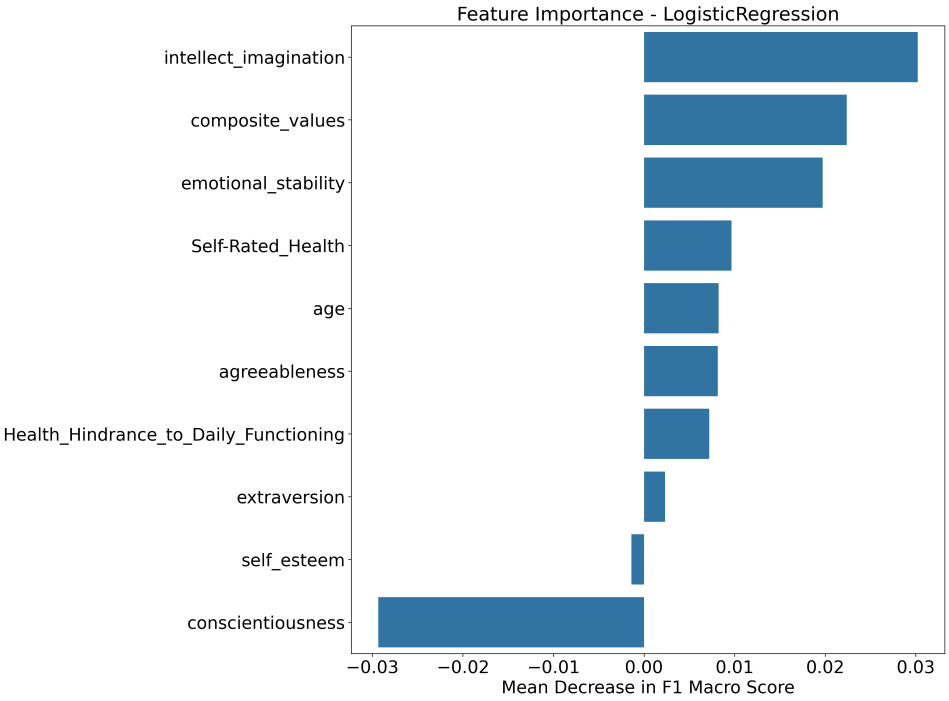

Figure 16: Permutation feature importance for the 3-class mental health change model.

4.3.3 *Error analysis*

To assess the fairness and consistency of the model's performance across subpopulations, error analysis was conducted for the best-performing model across key demographic groups (see Figure 17). The results show limited variability between groups, with F1 macro scores generally low and consistent, reflecting the overall classification difficulty of the 3-class problem.

Across sex, ethnic background, and religion, the model did not appear to systematically favor or disadvantage specific subgroups. Minor variation was observed by education and generation, with slightly better performance for participants with MBO education and those from Generation X.

While there is no strong evidence of systematic bias against major demographic groups, it is important to note that *no strong evidence of bias does*

not equate to proof of fairness. In particular, because the overall performance of the model is low (macro F1 $\approx$ 0.35), the apparent consistency across groups may simply reflect general underperformance. In such cases, a model may appear "fair" only because it is not effectively serving any group.

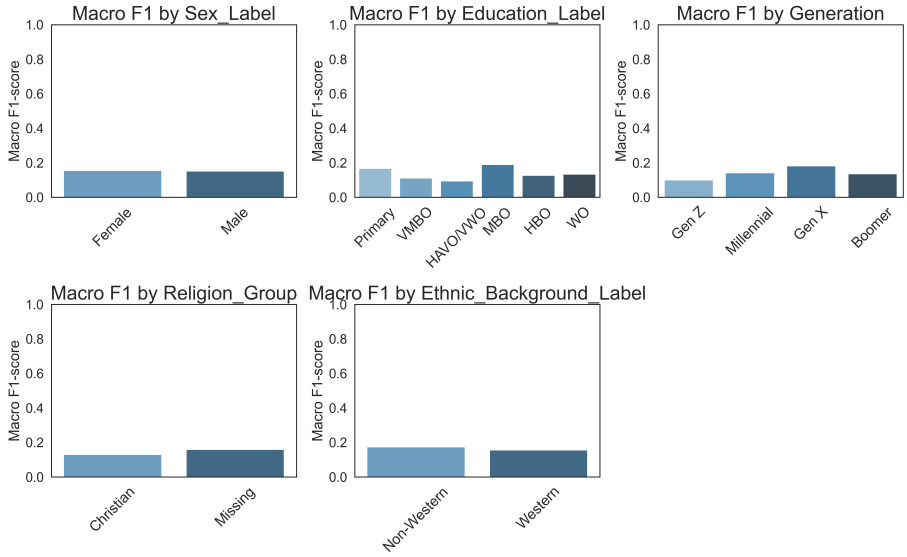

Figure 17: F1 macro scores for the best performing model (Logistic Regression with RF Feature Importance feature set and Random Undersampling).

4.3.4 *Basic Model vs Best Model: statistical significance*

The best-performing model—logistic regression trained on the RF Feature Importance feature set with a random undersampling strategy—achieved a higher mean F1 macro score than the baseline logistic regression model using only socio-demographic features (0.4102 ± 0.0227 vs. 0.365 ± 0.015), as shown in Table 14. A paired sample t-test confirmed that this difference was statistically significant ($t = 4.51$, $p = 0.0107$), indicating that the observed performance improvement is unlikely to be due to chance (see Figure 18).

Table 14: Comparison of F_1 macro scores (mean ± std.) between the baseline and best-performing model for the 3-class mental health change prediction task (nested cross-validation, 5 folds).

Model	**F_1 Macro (Mean ± Std.)**
Logistic Regression (Socio+Demo)	0.365 ± 0.015
Logistic Regression (RF Feature Importance)	0.410 ± 0.023

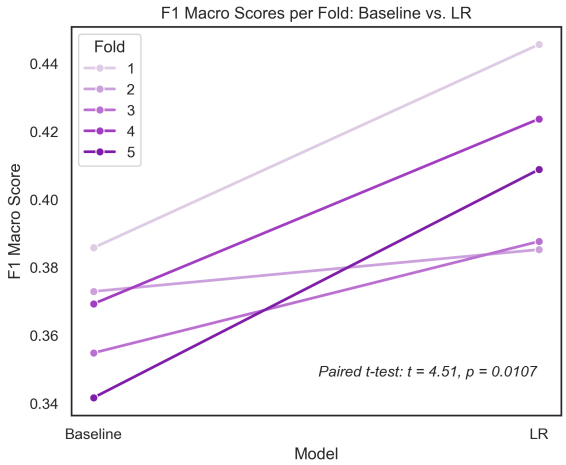

Figure 18: F1 macro scores per fold for the baseline (Logistic Regression with Socio+Demo features) and the best-performing model (Logistic Regression with RF Feature Importance and Random Undersampler). Lines connect paired folds. The logistic regression model using RF-based features significantly outperformed the baseline (paired t-test: $t = 4.51$, $p = 0.0107$).

4.3.5 *Samplers*

Table 15 summarizes the best-performing sampling strategy per feature set for the change prediction task. As shown, random undersampling was the most frequently selected method (3 out of 6 feature sets), indicating its relative effectiveness in handling class imbalance in this setup. This contrasts with the results on the 2022 dataset, where SMOTE-Tomek was the most selected strategy (4 out of 6 sets),highlighting that the optimal sampling method may vary depending on whether the task involves static outcomes (e.g., 2022 classification) or temporal change (e.g., predicting shifts in mental health between 2022 and 2023).

Table 15: Best-performing sampling strategy per feature set (Mean F1 Binary ± STD, 2022 dataset)

Feature Set	No Sampling	Random Undersampling	SMOTE-Tomek
Socio + Demo		0.385 ± 0.051	
Socio + Demo + Health			0.539 ± 0.063
Socio + Demo + Psychology			0.611 ± 0.033
Socio + Demo + Health + Psy			0.676 ± 0.036
Lasso Selected Features			0.657 ± 0.023
RF Feature Importance		**0.646 ± 0.043**	

4.3.6 *Ordinal classification: Predicting ordered mental health outcomes*

To explore whether explicitly modeling the ordinal nature of the mental health outcome variable (*deteriorated* < *stable* < *improved*) could improve performance, the best-performing logistic regression model was re-evaluated using the `mord.LogisticIT()` ordinal classifier. While this approach is designed to reduce extreme misclassifications by leveraging class order, results revealed that the model consistently predicted only the "stable" category. It achieved a recall of 94% for class 1 (stable), but failed to correctly classify any cases in the "deteriorated" or "improved" classes (see Figure 19). This resulted in a macro F1-score of just 0.19. The confusion matrix highlights the model's strong bias toward the majority class, suggesting that although ordinal modeling is conceptually suitable, it did not enhance predictive accuracy in this case.

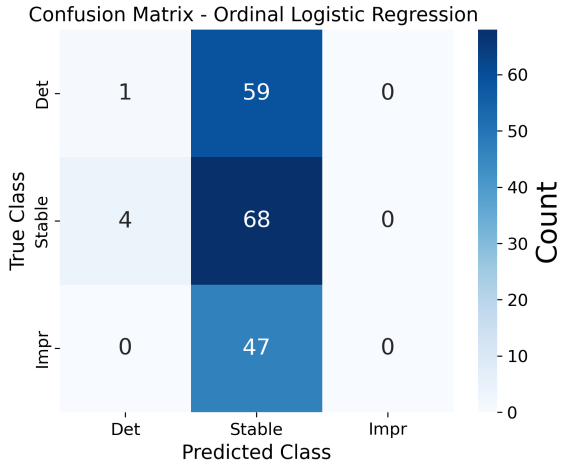

Figure 19: Confusion matrix for the ordinal logistic regression model (`mord.LogisticIT`) predicting change in mental health status. The model shows a strong bias toward the "stable" class, correctly classifying 94% of stable cases but failing to identify any improved or deteriorated cases.

5 DISCUSSION

This chapter interprets the study's key findings in light of the research questions and existing literature, reflects on methodological limitations, and offers recommendations for future work. It then examines the societal impact of these findings and discusses their implications for mental health policy.

5.1 *Research Questions- Answers*

RQ1. *Which machine learning models are most effective in predicting binary classifications of mental health status based on MHI-5 scores in 2022 using LISS panel data?*

The best-performing model for classifying MHI-5 status in 2022 was logistic regression using the top 10 features selected by Random Forest Feature Importance ranking. Despite its simplicity, this model achieved strong predictive performance (F_1 = 0.61, recall = 0.76) and substantially outperformed the baseline model (F_1 = 0.39, based on training set results) that replicated van der Velden, Setti, et al. (2019a)'s predictor set of demographics and social media use ($t = 8.78$, $p = 0.0009$). This demonstrates the added value of a more refined, data-driven feature selection approach and underlines the potential of machine learning over traditional statistical methods in this context.

van der Velden, Setti, et al. (2019a) also used MHI-5 as their outcome measure but applied logistic regression solely to fit the data, without assessing predictive performance or testing generalizability, focusing instead on statistically significant relationships within the training data. By contrast, the current study shows that models trained on a small set of optimized features can generalize well beyond the training data, as evidenced by the minimal generalization gap on the 2022 hold-out set (test set: F_1 = 0.61, recall = 0.76; training set: mean F_1 = 0.65 ± 0.04, mean recall = 0.78 ± 0.07).

Permutation feature importance revealed that emotional stability, health hindrance in daily functioning, and age were the most influential predictors, suggesting that personality traits and health-related factors are stronger predictors of poor mental health than social factors like SNS use, which was central in van der Velden's work. This partially confirms their conclusion that social media has limited predictive power while also extending it by demonstrating which variables most directly drive classification performance.

Importantly, the model performed particularly well in identifying individuals with poor mental health, achieving a recall of 0.76 for the at-risk

class. The confusion matrix shows that the majority of true positive cases were correctly identified, meaning the model successfully flags those with MHI-5 scores below the clinical threshold. This sensitivity is especially valuable for early detection and screening purposes, where the primary goal is to minimize false negatives—that is, to avoid overlooking individuals who may need psychological support. These results highlight the practical utility of the model not only as a predictive tool, but also as a potential aid in mental health monitoring and intervention strategies.

RQ2. *How well do machine learning models trained on 2022 data generalize to predicting mental health status in 2023 based on MHI-5 scores using LISS panel data, and how stable is their predictive performance over time?*

The demonstration of temporal stability ($t = -0.66$, $p = 0.5461$) represents a significant contribution to the sparse literature on longitudinal MHI-5 prediction. Despite a modest decline in 2023 test performance (F_1 = 0.58) compared to 2022 (F_1 = 0.61), the model maintained statistically equivalent F_1 performance when predicting 2023 outcomes using 2022 predictors, with no significant degradation in predictive accuracy across the one-year interval.

This temporal consistency builds upon the work of van der Velden, Das, and Muffels (2019), who examined mental health stability across three cohorts and identified distinct mental health profiles that remained relatively stable over time. Our findings complement their descriptive analysis by demonstrating that predictive models can reliably forecast future mental health status using a similar, though slightly broader, set of predictors.

While this temporal consistency cannot establish causality, it provides evidence for sustained predictor-outcome relationships that extend beyond mere cross-sectional associations, supporting the criterion of temporal precedence as a foundation for future causal investigation. The maintained predictive power across the one-year interval suggests that emotional stability, health hindrance, and age (along with other psychological and health-related predictors) represent enduring mental health determinants rather than transient correlates, aligning with Kraemer et al. (2001)'s emphasis on temporal relationships in causal inference frameworks.

RQ3. *Which machine learning models are most effective in predicting changes in mental health status (improved, stable, or deteriorated) based on MHI-5 score differences between 2022 and 2023?*

The results for RQ3 show that predicting categorical changes in mental health (improved, stable, deteriorated) using MHI-5 scores is significantly more challenging. The best-performing model—a logistic regression using the top 10 features from Random Forest feature importance—achieved a modest macro F1-score of 0.35 on the test set. While the model identified stable trajectories relatively well (recall = 0.50), it struggled to accurately classify both improvement (recall = 0.38) and, particularly, deterioration (recall = 0.20). This limited performance reflects the complexity of modeling dynamic psychological change, consistent with prior research highlighting the difficulty of predicting mental health trajectories over time (Petersen et al., 2018; Zacher & Rudolph, 2021).

Despite this, the study makes a novel contribution by applying machine learning to classify mental health change based on MHI-5 score differences—an approach not seen in existing literature. The use of balanced classes and permutation feature importance further enhances robustness, revealing that psychological variables such as the Big Five's *Emotional Stability* and *Intellect/Imagination* (or *Openness to Experience*), as well as value orientation, contributed most to the predictions.

To account for the ordered structure of the mental health change variable, we also explored ordinal classification using the LogisticIT model. While ordinal modeling can, in theory, reduce extreme misclassifications, in practice, it failed to improve performance in this study. The model systematically defaulted to predicting the *stable* class, yielding no correct classifications for improvement or deterioration. This suggests that while ordinal classifiers offer a conceptually elegant approach, their effectiveness is limited when feature distributions fail to differentiate between adjacent outcome categories. These results reinforce findings from the nominal 3-class analysis: predicting change in mental health remains a difficult task even when temporal structure and class order are incorporated.

5.2 *Sub Questions-Answers*

SQ1: Comparative Model Performance. Although Random Forest and XGBoost are often praised for capturing nonlinearities, our results showed that a parsimonious logistic regression performed on par with these "black-box" models in both accuracy and interpretability—echoing prior work that, with well-chosen predictors, simpler models can match or exceed complex ones (Caruana et al. (2015)).

SQ2: Key Predictors. Permutation importance identified emotional stability, health hindrance to daily functioning, and age as the strongest predictors of both contemporaneous and year-ahead MHI-5 scores. Other personality traits and value-orientation variables also signaled change.

These findings extend Big Five literature by demonstrating their utility in a screening context and highlight the added value of including health-related and value measures alongside personality traits (Malouff et al., 2005; Steel et al., 2008).

SQ3: Theory- vs. Data-Driven Feature Selection. Theory-driven feature blocks (demographics, health, social, psychology) performed respectably, but a data-driven subset based on Random Forest importance yielded higher test F_1 when used in logistic regression. This suggests combining domain knowledge with data-driven pruning to capture panel-specific nuances.

SQ4: Balancing Techniques. Comparing no sampling, undersampling, and SMOTE-Tomek, we found that SMOTE-Tomek produced the best cross-validation F_1 for binary tasks, whereas simple undersampling was optimal for the 3-class task. Practitioners should therefore evaluate multiple resampling methods rather than assume a single best approach.

SQ5: Performance Across Demographics. For 2022 binary classification, only education level varied notably—$F_1 \approx 0.80$ for VMBO and WO groups (and similarly high for Christian respondents)—suggesting more reliable screening in these subpopulations. In contrast, the 3-class macro F_1 remained uniformly low (0.20) across all groups, underscoring the universal difficulty of distinguishing "improved," "stable," and "deteriorated" trajectories.

5.3 *Limitations and Recommendations*

This study employed a comprehensive and systematic approach, implementing multiple feature selection methods, balancing techniques, and machine learning algorithms with nested cross-validation and stability testing. This structured methodology provided and ensured robust and reliable model evaluation. In addition, the analysis approached MHI-5 prediction from multiple angles—cross-sectional, longitudinal, and change-focused—offering a more complete picture. Finally, the inclusion of a wide range of predictors, covering demographic, social, health-related, and psychological domains, allowed for a richer understanding of the complex factors influencing self-reported mental health as measured by the MHI-5.

While this study provides valuable insights, several areas present opportunities for future enhancement. The study focused on self-reported MHI-5 data as the primary mental health screening tool, which provided a solid foundation. However, future research could build upon these findings by incorporating additional validated screening instruments such as the PHQ-9, GAD-7, or CES-D. Additionally, integrating biomarker data

alongside self-reported measures could further enhance predictive validity and provide a more holistic approach to mental health prediction.

The analysis compared four established machine learning algorithms, providing a solid methodological foundation. Future research could explore more complex models such as neural networks and expand hyperparameter optimization ranges. It may also be valuable to investigate additional data balancing techniques and feature selection methods. The random forest feature importance analysis focused on the top 10 features, offering clear interpretability; however, future studies might consider broader feature sets (e.g., top 20).Finally, the theory-driven approach could be strengthened by incorporating contextual variables—such as environmental stressors and life events—to better reflect the multifactorial nature of mental health.

5.4 *societal impact*

By leveraging machine learning to predict MHI-5 scores, we can transform a brief five-item screener into an early-warning system that flags at-risk individuals before they enter crisis. For example, public-health agencies using surveys such as the Dutch LISS panel could run these models each wave to pinpoint communities or demographic segments (e.g., young adults, low-income households) showing increasing risk. These agencies could then deploy targeted interventions—rather than waiting for clinical diagnoses to accumulate—such as SMS invitations to online self-help modules or pop-up screening booths in local clinics. This builds on evidence that proactive screening in primary care reduces untreated depression by up to 30 percent (Thombs & et al., 2018) and aligns with the WHO's call for scalable, data-driven mental-health monitoring to relieve system burdens (World Health Organization, 2022).

5.5 *Conclusion*

This thesis comprehensively and systematically compared multiple machine learning models (logistic regression, random forest, SVM, XGBoost) for predicting MHI-5–based mental health status in 2022 and 2023 using LISS panel data. The best-performing model was a logistic regression trained on a concise set of top predictors selected by a random forest; despite its simplicity, it achieved strong performance and greatly outperformed a baseline model. Importantly, this model's accuracy remained statistically stable when forecasting one year ahead, indicating robust generalization over time. By contrast, predicting categorical change in mental health (improved, stable, deteriorated) proved much more difficult.

Key predictors identified were emotional stability, health-related limitations, and age, highlighting the importance of psychological and health factors over variables like social media use. A feature-selection strategy that combined domain knowledge with data-driven pruning yielded the most efficient models, while performance gains from sampling strategies varied by task, confirming that no single resampling method universally excels.Error analysis showed limited variation in model performance across demographic groups, suggesting overall fairness, though multiclass predictions remained universally challenging.

REFERENCES

Abdul Rahimapandi, H. D., Maskat, R., Musa, R., & Ardi, N. (2022). Depression prediction using machine learning: A review. *IAES International Journal of Artificial Intelligence*, *11*(3), 1108–1118. https://doi.org/10.11591/ijai.v11.i3.pp1108-1118

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631. https://doi.org/10.1145/3292500.3330701

Bhugra, D., & Ventriglio, A. (2023). Political determinants of mental health. *International Journal of Social Psychiatry*, *69*(3), 521–522. https://doi.org/10.1177/00207640231154706

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Burešová, I., Jelínek, M., Dosedlová, J., & Klimusová, H. (2020). Predictors of mental health in adolescence: The role of personality, dispositional optimism, and social support. *SAGE Open*, *10*. https://doi.org/10.1177/2158244020970352

Cardoso, J. S., & da Costa, J. F. (2007). Learning to classify ordinal data: The data replication method. *Journal of Machine Learning Research*, *8*, 1393–1429.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730. https://doi.org/10.1145/2783258.2788613

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 785–794. https://doi.org/10.1145/2939672.2939785

Cohen, P., & Cohen, J. (1995). *Life values and adolescent mental health* (1st ed.). Psychology Press. https://doi.org/10.4324/9780203773994

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Cramer, J. S. (2002). *The origins of logistic regression* (Tinbergen Institute Working Paper No. 2002-119/4). Tinbergen Institute. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=360300

De Vos, K. (2010). *Representativeness of the liss-panel 2008, 2009, 2010* (tech. rep.). CentERdata. Tilburg.

Elovainio, M., Hakulinen, C., Pulkki-Råback, L., & Suvisaari, J. (2020). General health questionnaire (ghq-12), beck depression inventory (bdi-6), and mental health index (mhi-5): Psychometric and predictive properties in a finnish population-based sample. *Psychiatry Research*, *289*, 112973. https://doi.org/10.1016/j.psychres.2020.112973

Géron, A. (2017). *Hands-on machine learning with scikit-learn and tensorflow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Gloster, A. T., Lamnisos, D., Lubenko, J., Presti, G., Squatrito, V., Constantinou, M., Nicolaou, C., Papacostas, S., Aydın, G., Chong, Y. Y., Chien, W. T., Cheng, H.-Y., Ruiz, F. J., Garcia-Martin, M. B., Obando-Posada, D. P., Segura-Vargas, M. A., Vasiliou, V. S., McHugh, L., Höfer, S., … Karekla, M. (2020). Impact of covid-19 pandemic on mental health: An international study. *PLOS ONE*, *15*(12), e0244809. https://doi.org/10.1371/journal.pone.0244809

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1), 84–96. https://doi.org/10.1016/j.jrp.2005.08.007

Harring, J. R., et al. (2006). Latent growth modeling: A structural equation perspective on longitudinal change. *Structural Equation Modeling*, *13*(4), 623–643. https://doi.org/10.1207/s15328007sem1304_6

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., & et al. (2020). Array programming with numpy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hornstein, S., Forman-Hoffman, V., Nazander, A., Ranta, K., & Hilbert, K. (2021). Predicting therapy outcome in a digital mental health intervention for depression and anxiety: A machine learning approach. *Digital Health*, *7*, 1–11. https://doi.org/10.1177/20552076211060659

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*.

Kelly, M. J., Dunstan, F. D., Lloyd, K., & Fone, D. L. (2008). Evaluating cutpoints for the mhi-5 and mcs using the ghq-12: A comparison of five different methods. *Population Health Metrics*, *6*(10). https://doi.org/10.1186/1478-7954-6-10

Kohn, J. N., Jester, D. J., Dilmore, A. H., Thomas, M. L., Daly, R., & Jeste, D. V. (2022). Trends, heterogeneity, and correlates of mental health and psychosocial well-being in later-life: Study of 590 community-dwelling adults aged 40–104 years. *Aging & Mental Health*, *27*(6), 1198–1207. https://doi.org/10.1080/13607863.2022.2078790

Kraemer, H. C., Stice, E., Kazdin, A., Offord, D., & Kupfer, D. (2001). How do risk factors work together? mediators, moderators, and independent, overlapping, and proxy risk factors. *American Journal of Psychiatry*, *158*(6), 848–856. https://doi.org/10.1176/appi.ajp.158.6.848

Krebs, T. S., & Johansen, P. Ø. (2013). Psychedelics and mental health: A population study. *PLOS ONE*, *8*(8), e63972. https://doi.org/10.1371/journal.pone.0063972

Le, T., Lee, M. Y., Park, J. R., & Baik, D. S. (2018). Oversampling techniques for bankruptcy prediction: Novel features from a transaction dataset. *Symmetry*, *10*(4), 79. https://doi.org/10.3390/sym10040079

Levin, J. (2010). Religion and mental health: Theory and research. *International Journal of Applied Psychoanalytic Studies*, *7*(2), 102–115. https://doi.org/10.1002/aps.240

Li, G., & Jiang, L. (2023). Random forest algorithm-based modelling and neural network analysis between social anxiety disorder of childhood and parents' socioeconomic attributes. *2023 IEEE 5th Eurasia Conference on IoT, Communication and Engineering (ECICE)*, 222–225. https://doi.org/10.1109/ECEI57668.2023.10105416

Li, Y. (2023). Application of machine learning to predict mental health disorders and interpret feature importance. *Proceedings of the 2023 International Symposium on Computer Technology and Information Science (ISCTIS)*, 257–261. https://doi.org/10.1109/ISCTIS58954.2023.10213032

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed., Vol. 793). John Wiley & Sons.

Malouff, J. M., Thorsteinsson, E. B., Rooke, S. E., & Schutte, N. S. (2005). The relationship between the five-factor model of personality and symptoms of clinical disorders: A meta-analysis. *Journal of Psychopathology and Behavioral Assessment*, *27*(2), 101–114. https://doi.org/10.1007/s10862-005-3262-0

McFadden, J. (2023). Razor sharp: The role of occam's razor in science. *Frontiers in Human Neuroscience*, *17*, 10952609. https://doi.org/10.3389/fnhum.2023.10952609

McKinney, W. (2010). Data structures for statistical computing in python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th python in science conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Natekin, A., & Knoll, A. (2013). Gradient boosting machines: A tutorial. *Frontiers in Neurorobotics*, *7*, 21. https://doi.org/10.3389/fnbot.2013.00021

pandas development team. (2020, February). Pandas-dev/pandas: Pandas (version latest). https://doi.org/10.5281/zenodo.3509134

Parmar, A., Katariya, R., & Patel, V. (2019). A review on random forest: An ensemble classifier. In J. Hemanth, X. Fernando, P. Lafata, & Z. Baig (Eds.), *International conference on intelligent data communication technologies and internet of things (icici)* (pp. 405–415, Vol. 26). Springer. https://doi.org/10.1007/978-3-030-03146-6_86

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Petersen, R., et al. (2018). Trajectories of cognitive decline in alzheimer's disease. *Nature Reviews Neurology*, *14*(9), 523–535. https://doi.org/10.1038/s41582-018-0047-9

Pollar, F., & Harigovind, F. (2018). Title of the study. *Journal Name*, *10*(Y), ZZZ–ZZZ. https://doi.org/10.xxxx/xxxxxxx

R Core Team. (2023). R: A language and environment for statistical computing. https://www.R-project.org/

Rahimapandi, H. D. A., Maskat, R., Musa, R., & Ardi, N. (2022). Depression prediction using machine learning: A review. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *11*(3), 1108–1118. https://doi.org/10.11591/ijai.v11.i3.pp1108-1118

Rennie, J. D., & Srebro, N. (2005). Loss functions for preference levels: Regression with discrete ordered labels. *Proceedings of the IJCAI multidisciplinary workshop on advances in preference handling*.

Rokeach, M. (1973). *The nature of human values*. Free Press.

Rosenberg, M. (2015). *Society and the adolescent self-image* (Revised). Princeton University Press.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Sasada, T., Liu, Z., Baba, T., Hatano, K., & Kimura, Y. (2020). A resampling method for imbalanced datasets considering noise and overlap. *Procedia Computer Science*, *176*, 420–429. https://doi.org/10.1016/j.procs.2020.09.035

Scherpenzeel, A. C. (2018). "true" longitudinal and probability-based internet panels: Evidence from the netherlands. In V. Vehovar, K. L. Manfreda, & M. P. Couper (Eds.), *Social and behavioral research and the internet* (pp. 77–104). Routledge.

Shaikh Mohammad, B. N., & Siddiqui, K. (2021). Random forest regressor machine learning model developed for mental health prediction based on mhi-5, phq-9 and bdi scale [Available at SSRN: https://ssrn.com/abstract=3867416]. *Proceedings of the 4th International Conference on Advances in Science Technology (ICAST2021)*. http://dx.doi.org/10.2139/ssrn.3867416

Shields-Zeeman, L., & Smit, F. (2022). The impact of income on mental health. *The Lancet Public Health*, *7*(6), e486–e487. https://doi.org/10.1016/S2468-2667(22)00117-4

Smith, A. P. (2011). Snacking habit, mental health, and cognitive performance. *Current Topics in Nutraceuticals Research*, *9*(1), 47.

Spyrou, I. M., Frantzidis, C., & Bratsas, C. (2016). Methodologies of classification compared. In *Control and signal processing in biomedicine* (pp. 118–129). Elsevier. https://doi.org/10.1016/j.bspc.2015.10.006

Steel, P., Schmidt, J., & Shultz, J. (2008). Refining the relationship between personality and subjective well-being: Self- and peer-report analyses. *Journal of Research in Personality*, *42*(3), 743–757. https://doi.org/10.1016/j.jrp.2007.09.003

Strand, B. H., Dalgard, O. S., Tambs, K., & Rognerud, M. (2003). Measuring the mental health status of the norwegian population: A comparison of the instruments scl-25, scl-10, scl-5 and mhi-5 (sf-36). *Nordic Journal of Psychiatry*, *57*(2), 113–118. https://doi.org/10.1080/08039480310000932

Tamilselvi, J. J., & Gifta, C. B. (2011). Handling duplicate data in data warehouse for data mining. *International Journal of Computer Applications*, *15*(4), 7–15.

Thombs, B. D., & et al. (2018). Accuracy of depression screening tools in primary care. *JAMA*, *320*(4), 358–369. https://doi.org/10.1001/jama.2018.10136

Thorstad, R., & Wolff, P. (2019). Predicting future mental illness from social media: A big-data approach. *Behavior Research Methods*, *51*(4), 1586–1600.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC Press.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(1), 1–67. https://doi.org/10.18637/jss.v045.i03

van der Velden, P. G., Setti, I., van der Meulen, E., & Das, M. (2019a). Does social networking sites use predict mental health and sleep problems when prior problems and loneliness are taken into account? a population-based prospective study. *Computers in Human Behavior*, *93*, 200–209. https://doi.org/10.1016/j.chb.2018.11.047

van der Velden, P. G., Das, M., & Muffels, R. (2019). The stability and latent profiles of mental health problems among dutch young adults in the past decade: A comparison of three cohorts from a national sample. *Psychiatry Research*, *282*, 112622. https://doi.org/10.1016/j.psychres.2019.112622

van der Velden, P. G., Setti, I., van der Meulen, E., & Das, M. (2019b). Does social networking sites use predict mental health and sleep problems when prior problems and loneliness are taken into account? a population-based prospective study. *Computers in Human Behavior*, *93*, 200–209. https://doi.org/10.1016/j.chb.2018.11.047

van der Velden, P. G., van der Meulen, E., Das, M., Muffels, R., & Bosmans, M. W. G. (2022). Before and after covid-19: Changes in mental health among dutch adolescents. *Psychiatry Research*, *311*, 114528. https://doi.org/10.1016/j.psychres.2022.114528

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wilhelm, K. A. (2014). Gender and mental health. *Australian & New Zealand Journal of Psychiatry*, *48*(7), 603–605. https://doi.org/10.1177/0004867414538678

World Health Organization. (2022). *World mental health report: Transforming mental health for all*. WHO.

Zacher, H., & Rudolph, C. W. (2021). Individual differences and changes in subjective wellbeing during the early stages of the covid-19 pandemic. *The American Psychologist*, *76*(1), 50–62. https://doi.org/10.1037/amp0000702

APPENDIX A

APPENDIX B: EXPLORATORY DATA ANALYSIS (EDA)

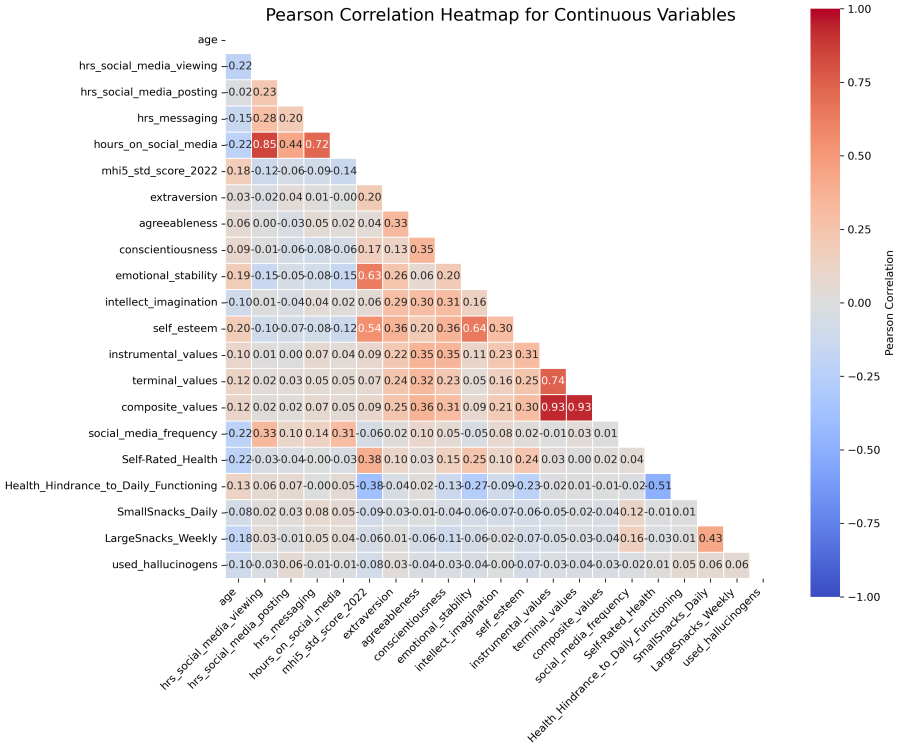

Figure 20: Correlation Heatmap for Continuous Variables (Pearson's r)

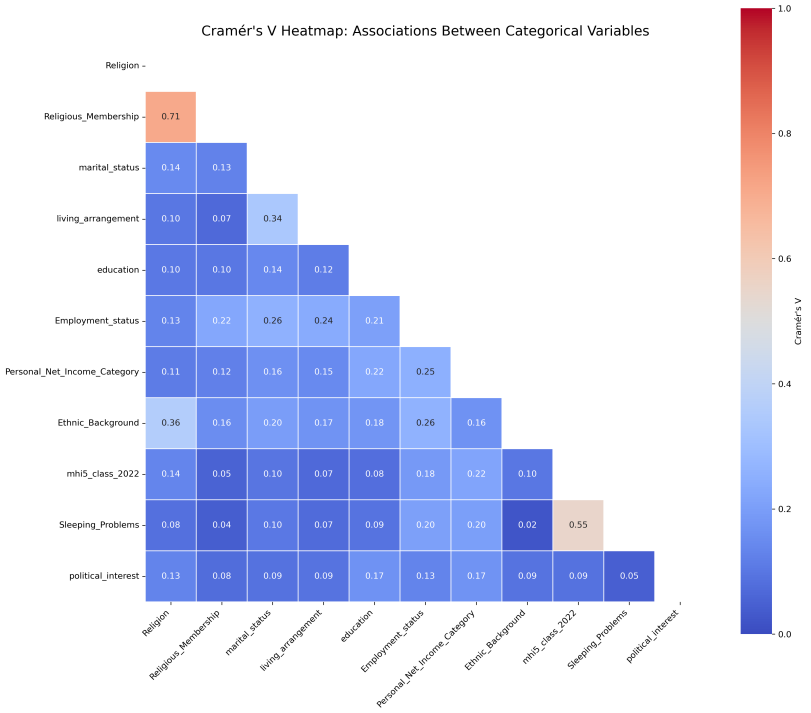

Figure 21: Cramér's V Heatmap: Associations Between Categorical Variables

APPENDIX C: SOFTWARE OVERVIEW

This appendix provides an overview of the software environments and libraries used in this research.

Python Environment

The primary programming language used was Python (version 3.11.4), alongside the following libraries:

- **NumPy** (Harris et al., 2020)
- **Pandas** (McKinney, 2010; pandas development team, 2020)
- **Matplotlib** (Hunter, 2007)
- **Seaborn** (Waskom, 2021)
- **Scikit-learn** (Pedregosa et al., 2011)
- **Optuna** (Akiba et al., 2019)
- **xgboost (XGBClassifier)** (Chen & Guestrin, 2016)

R version 4.3.1 was used for statistical modeling and multiple imputation procedures (R Core Team, 2023).

And this!