Predicting Depression from Sociological Variables using Machine Learning

by

Ali Slayie

A Bachelor's Thesis
Submitted to the Department of Data Science
Northwestern University
April 2025
Supervised by Prof Arend Kuyper & Prof David Schieber

Abstract

This thesis explores the predictive potential of sociological variables in identifying the risk of depression by employing machine learning techniques. Drawing on five years of data from the National Health Interview Survey (NHIS), the project examines how demographic, socioeconomic, and healthcare access factors—especially those pertaining to migration and LGBTQ+ identity, contribute to depression outcomes. By constructing a balanced dataset across four intersectional groups (LGBTQ+ migrants, LGBTQ+ non-migrants, straight migrants, and straight non-migrants), the study implements and compares multiple classification models, including decision trees, logistic regression, and RuleFit.

Results demonstrate that models incorporating structural and behavioral health variables significantly outperform those relying solely on identity markers. The best-performing model achieved an F1 score of 0.901, indicating high accuracy in identifying individuals with depression while minimizing both false positives and false negatives. It also reached an ROC AUC (Area Under the Receiver Operating Characteristic Curve) of 0.95, meaning the model could distinguish between individuals with and without depression with 95% accuracy.

Explainability tools such as SHAP values and partial dependence plots are used to unpack model decisions, providing insight into how features like anxiety diagnosis, therapy access, and education level shape predictions. This project highlights the importance of integrating social context into data science workflows and presents a transparent, equitable approach to mental health prediction.

Table of contents

List of Figures

List of Tables

1 Introduction

Depression remains one of the most prevalent and complex issues in public health in the United States. While the condition affects people across all demographics, its distribution and severity are far from random. Research has consistently shown how social determinants such as migration status, sexual orientation, and gender identity can shape mental health outcomes. However, these sociological variables are often underexplored in predictive modeling, especially in ways that reflect their impacts through an intersectional lens.

This thesis frames mental health prediction not just as a question of individual outcomes, but as a data science challenge rooted in social complexity. By using the National Health Interview Survey (NHIS) from 2019 to 2023, I explore whether depression can be effectively predicted based on a range of demographic and social variables—particularly those tied to migration and LGBTQ+ (Lesbian, Gay, Bisexual, Transgender, Queer, and more) identity. Rather than treating these as fixed categories, this project examines the predictive power of these identities through rigorous modeling, subgroup balancing, and model interpretation techniques.

The objective of this project is not only to identify which variables are most predictive of depression but also to demonstrate how thoughtful data preparation, model evaluation, and interpretation can shed light on social disparities without compromising technical rigor. While the topic is socially meaningful, the thesis remains firmly situated in the data science domain, emphasizing methodological complexity, performance benchmarking, and scalable analysis pipelines.

1.1 Research Questions and Objectives

This thesis is guided by the following core research questions:

1. To what extent can depression be predicted using sociologically meaningful variables from the NHIS dataset—particularly those related to migration and LGBTQ+ identity?
2. How do model performance and feature importance vary across four key subgroups: LGBTQ+ migrants, LGBTQ+ non-migrants, straight migrants, and straight non-migrants?
3. Which data preparation, subgroup balancing, and model tuning strategies yield the strongest predictive performance?
4. How can explainability techniques help interpret and contextualize model predictions in a socially responsible way?

The objective is to build a technically rigorous, interpretable, and generalizable predictive pipeline that demonstrates advanced data science capabilities, while ensuring that social context informs—rather than overshadows—the modeling process.

2 Literature Review

Depression has been widely studied across fields such as psychology, public health, and sociology, with a large body of work documenting disparities in mental health outcomes by race, gender identity, and migration status (M. S. David R. Williams (2010)). Much of this research emphasizes identifying correlates and structural risk factors, often without building predictive models that can support early intervention or clinical decision-making (Marissa Tan (2020)).

At the same time, machine learning has become an increasingly common tool in mental health research, applied to predict outcomes from electronic health records, behavioral data, and self-report surveys (Adrian B. R. Shatte (2019), 2019; Dominic B. Dwyer (2018)). However, many of these models prioritize predictive performance over interpretability and often exclude identity-based variables that are crucial for understanding disparities.

This thesis contributes to an emerging space that integrates sociological insight into algorithmic modeling. By incorporating sociologically meaningful variables—such as race, gender identity, sexual orientation, and migration status—and applying interpretable ML techniques, it aims to bridge a gap between structural theories of inequality and computational approaches to health.

2.1 Machine Learning for Mental Health Prediction

Machine learning (ML) has become an increasingly common tool in mental health research, with applications ranging from prediction of depression to risk classification for suicide and anxiety. Many recent studies use data sources such as electronic health records, social media activity, or survey responses to train predictive models (Adrian B. R. Shatte (2019), 2019; Dominic B. Dwyer (2018)). For instance, Nickson et al. (2023) reviewed ML approaches using health records to predict depression, highlighting the use of algorithms such as logistic regression, support vector machines, and ensemble methods—but noted the absence of subgroup-specific evaluations (David Nickson (2023)). Similarly, Lee and Ham (2022) observed that black-box models dominate the field, often prioritizing performance over interpretability (Kwang-Sig Lee (2022)).

Even studies that emphasize interpretability often focus on model structure or feature importance, without assessing variation in predictive performance across demographic or identity-based subgroups. This lack of subgroup stratification limits our ability to evaluate fairness or diagnostic reliability for marginalized populations. As Rudin (2019) argues, black-box models may be ill-suited for high-stakes decisions, especially when transparency and accountability are critical (Rudin (2019)).

This thesis builds on a growing movement toward interpretable machine learning by using models such as decision trees, and applying explanation techniques like SHAP values and partial dependence plots. The modeling approach integrates repeated cross-validation, subgroup balancing, and structured preprocessing to ensure both robustness and fairness. Unlike much

prior work, this project evaluates prediction accuracy across intersectional subgroups, directly addressing a gap in existing mental health prediction research.

2.2 Sociological Determinants of Depression

Extensive research has shown that LGBTQ+ individuals and migrants face elevated risks of depression, anxiety, and related mental health conditions. These disparities are shaped by chronic exposure to minority stress, discrimination, and systemic barriers to care (Meyer (2013); S. A. M. David R. Williams (2009)). For example, sexual minority adults are significantly more likely to experience depression and suicidal ideation than their heterosexual peers (Miriam M. Moagi (2021)). Migrants, likewise, often face compounded stress from legal precarity, social isolation, and acculturative pressures, which further increase psychological vulnerability (Organization (2023)). Structural inequalities—such as housing insecurity, lack of insurance, and regional disparities in care—also play a significant role in shaping mental health outcomes.

Medical comorbidities, often more prevalent in marginalized groups, further complicate this landscape. Conditions such as hypertension (Irene A Kretchy (2014)), cholesterol imbalance (Fiedorowicz & Haynes (2010)), cardiovascular disease (Jason A. Bonomo (2024)), and chronic obstructive pulmonary disease (COPD) (Ninad T. Maniar (2024)) have all been associated with increased risk of depression. Some of these health conditions, particularly cardiovascular disease and COPD, have been shown to disproportionately affect LGBTQ+ individuals, adding biomedical complexity to social disadvantage.

Although these patterns are well-established in the public health literature, they are infrequently incorporated into predictive modeling frameworks that account for both social identity and clinical risk. This thesis addresses that gap by integrating structural and biomedical features into an interpretable machine learning approach.

2.3 Bridging Disciplinary Gaps

Sociological literature frequently draws on frameworks like intersectionality and minority stress to explain disparities in mental health (Crenshaw (1991); Meyer (2013)). While qualitative and descriptive approaches have shaped much of this research, relatively fewer studies use large-scale, quantitative data to build predictive models. Conversely, many machine learning applications in health research prioritize predictive performance and may overlook identity-based variables or fail to contextualize their outputs within structural inequalities.

This thesis seeks to bridge these gaps by combining interpretable modeling with identity- and condition-based features derived from five years of the National Health Interview Survey (NHIS). The minority stress model (Meyer (2013)) has been instrumental in explaining elevated mental health risks among LGBTQ+ individuals, yet it is rarely operationalized in algorithmic

modeling. Similarly, limited work explores how these dynamics intersect with migration status in statistical models.

By including features such as anxiety history, physical comorbidities, and demographic identity, this project reflects an intentional blending of sociological theory with machine learning practice. Through dataset balancing and the use of explainability tools, the resulting models aim to be not only accurate but also socially conscious.

To our knowledge, few prior studies have conducted intersectional evaluations of depression prediction using interpretable machine learning. This project contributes to that gap by comparing model performance across four subgroups—LGBTQ+ migrants, LGBTQ+ non-migrants, straight migrants, and straight non-migrants—thereby offering a novel and equity-focused approach to mental health prediction.

3 Data Source Overview and Cleaning

This project utilizes data from the National Health Interview Survey (NHIS)(Centers for Disease Control and Prevention (CDC) & National Center for Health Statistics (NCHS) (2024)), a large-scale, nationally representative survey conducted by the U.S. Centers for Disease Control and Prevention (CDC). The dataset provides detailed information on the health, demographic, and socioeconomic characteristics of individuals in the United States. Data from multiple years (2019–2023) were consolidated into a single dataset, resulting in a robust sample of over 100,000 observations.

During the data cleaning process, consistent variable naming conventions were applied across all datasets to ensure uniformity. For example, variables such as *EDUC_A* and *MAXEDUC_A* were renamed to more descriptive labels like *education_level* and *max_education_level* to improve clarity. Missing columns were added where necessary, and missing values were handled using mode imputation for categorical variables (e.g., imputing the most common value for region) and median imputation for numeric variables (e.g., filling in missing age values with the median age).

Variable types were also standardized to align with the requirements of various modeling approaches. Binary variables such as *depression_ever* and *anxiety_ever* were encoded with levels "Yes" and "No," while ordinal variables like *life_satisfaction* were mapped to ordered levels ranging from "Very dissatisfied" to "Very satisfied." These steps ensured the dataset was fully prepared for feature construction and downstream modeling tasks.

Figure Figure 1 below shows the distribution of the final sample, stratified by the identity groups used throughout the project.

3.1 Target Variable

The target variable is self-reported depression, measured by the NHIS variable *DEPEV_A* ("Have you ever had depression?"). This binary variable serves as the outcome of interest for all classification models. Because this variable is binary, the modeling task is framed as a classification problem.

Figure Figure 2 below shows the distribution of self-reported depression responses in the dataset.

3.2 Key Predictors

To investigate the social and structural factors influencing depression, this project used an incremental modeling strategy, adding predictors step-by-step across four model types, allowing for a clear understanding of how different variable sets contribute to predictive performance.

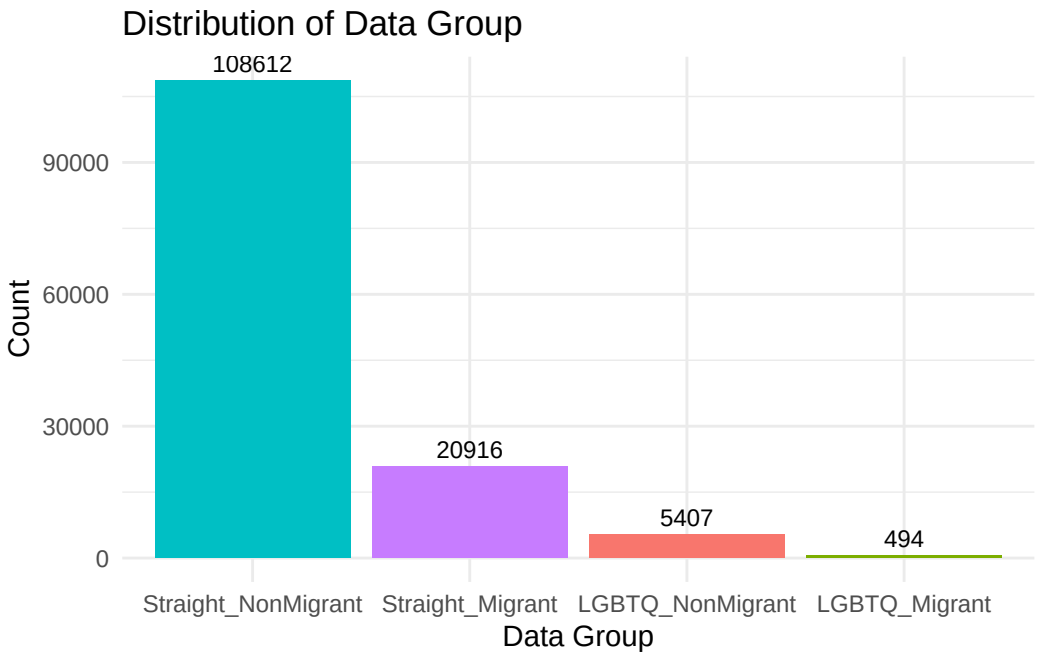

Figure 1: Distribution of Focus Group

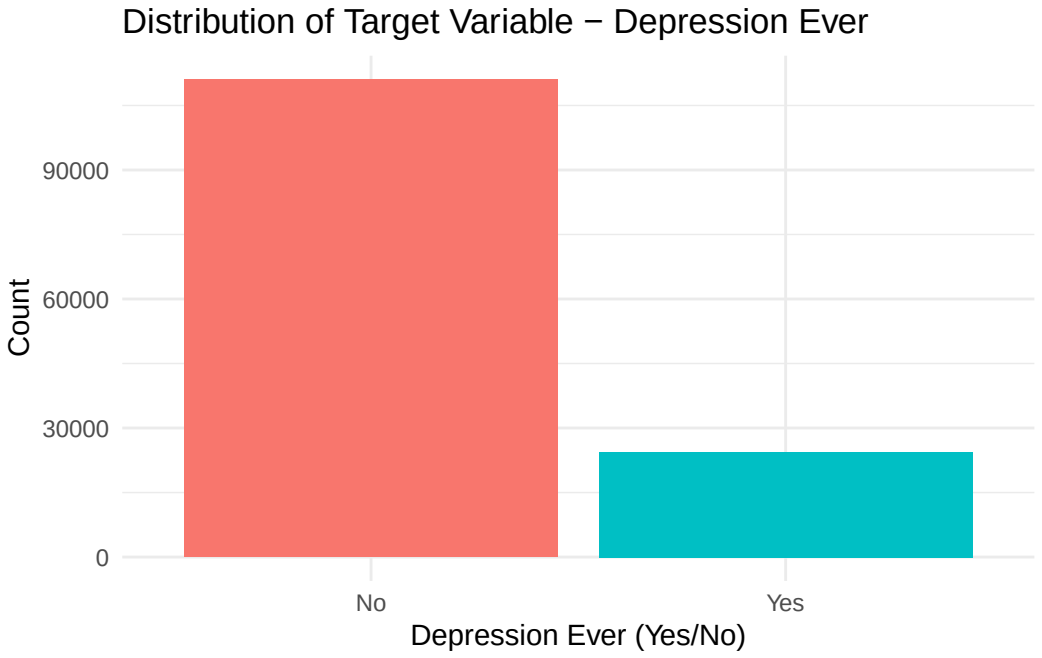

Figure 2: Distribution of Target Variable

This structure also helped isolate the predictive contribution of group membership relative to other explanatory factors.

The variable sets used in each model were:

- **Model 1: Group Membership Only**
  Included a single categorical variable representing identity group membership: LGBTQ+ migrant, LGBTQ+ non-migrant, straight migrant, and straight non-migrant. This baseline model assessed whether group membership alone was predictive of depression outcomes. An accompanying OLS model was also run on `PHSTAT_A` (self-rated health) to capture group differences in general health perception.
- **Model 2: Group Membership + Demographic Variables**
  Added basic demographic features including:
  - Gender identity
  - Age
  - Race and ethnicity
  - Geographic region

    These variables were selected to control for population-level variation and explore whether group-based disparities persisted after accounting for them.
- **Model 3: + Socioeconomic Status (SES) Variables**
  Built on Model 2 by introducing variables reflecting socioeconomic conditions:
  - Education level
  - Insurance coverage
  - Marital status
  - Parental status

    These indicators were intended to capture social positioning and access to resources; factors that often intersect with both identity and mental health outcomes.
- **Model 4: + Mental Health Care Access and Utilization**
  The final model included variables related to structural access to care:
  - Frequency of medical visits
  - Access to a regular healthcare provider
  - Ability to afford care
  - Medication use

    These predictors allowed for the evaluation of healthcare access as a potential pathway linking group membership to depression risk.

Across all models, the inclusion of predictors was informed by existing public health literature (see Literature Review) and the availability of consistently coded variables across the 2019–2023 NHIS files.

3.3 Modeling Workflow Overview

To support experimentation and scalability, I constructed two modeling pipelines:

- A **local pipeline** using a downsampled, group-balanced dataset (~40,000 observations) to quickly prototype models and test different variable sets.
- A **high-performance pipeline** run on Northwestern University's Quest computing cluster, using the **full NHIS dataset** (~100,000+ observations) with stratified cross-validation to ensure generalizability at scale.

This two-track approach allowed for fast iteration and robust evaluation, while ensuring that all modeling decisions were transferable between the smaller and larger datasets.

4 Methods

4.1 Group Construction & Dataset Balancing

To investigate how mental health outcomes differ across social identities, I constructed a categorical group variable that combines two binary indicators: LGBTQ+ identity and migration status. This resulted in four mutually exclusive groups:

1. LGBTQ+ migrants
2. LGBTQ+ non-migrants
3. Straight migrants
4. Straight non-migrants

Initial exploration of the NHIS dataset revealed substantial class imbalance—both in the distribution of the target variable (depression) and in the sizes of these identity groups. To mitigate bias and ensure fair representation across all four subgroups, I created a **balanced analytical dataset of approximately 40,000 observations**, with roughly **10,000 observations per group**.

This dataset was constructed using a combination of **increasing the number of samples in underrepresented groups (upsampling)** and **downsampling overrepresented ones**. By maintaining balance across the four identity-based groups, this approach allowed for a clearer comparison of model performance and subgroup disparities without the confounding effects of unequal sample sizes.

All modeling in this thesis was conducted on this balanced dataset to align with the project's emphasis on **equity-aware modeling**, interpretability, and methodological consistency across experiments.

4.2 Data Splitting & Resampling Strategy

All modeling was conducted on a balanced dataset of approximately 40,000 observations, with equal representation across the four identity-based subgroups. Two resampling setups were used throughout the project to train and evaluate the models.

- **Initial modeling (local machine)** used **4-fold cross-validation with 3 repeats**, allowing for fast iteration and exploratory analysis of different variable sets and model types resulting in 12 total resamples per model.

- **Final modeling (Quest high-performance cluster)** used a more rigorous **10-fold cross-validation with 5 repeats**, resulting in 50 total resamples per model. This setup provided more stable performance estimates and was used for final evaluations and comparisons.

In both setups, **stratified resampling** was used based on the binary depression indicator to preserve class proportions within each fold. This ensured consistent performance measurement and reduced bias due to class imbalance.

The use of two-tiered cross-validation allowed the project to balance exploratory modeling with rigorous evaluation—without compromising reproducibility.

4.3 Model Types

To explore a range of predictive strategies, I implemented five classification algorithms using a unified modeling framework in R (*tidymodels*). These models were chosen to reflect varying levels of complexity, flexibility, and interpretability:

- **Logistic Regression**
  A linear model commonly used for binary classification. It provides coefficients that are easy to interpret and serves as a strong baseline.
- **Decision Tree**
  A rule-based model that splits data into segments based on feature thresholds. It offers transparent, human-readable logic for prediction.
- **K-Nearest Neighbors (KNN)**
  A non-parametric model that assigns classes based on the majority label among the nearest data points. It is useful for capturing local structure in the data.
- **Naive Bayes**
  A probabilistic model that applies Bayes' Theorem under the assumption of feature independence. It performs surprisingly well in many high-dimensional settings.
- **RuleFit**
  A hybrid model that extracts rules from tree ensembles and combines them with linear terms. It balances performance with interpretability by using learned rules as features in a sparse linear model.

These models were trained and evaluated using the same cross-validation strategy on the balanced 40k dataset. The goal was to assess how different learning algorithms perform in predicting depression outcomes using sociologically-informed variables.

4.4 Preprocessing Pipelines - Feature Engineering

To systematically evaluate how different sets of predictors influence model performance, I developed four core preprocessing pipelines (Model Types 1–4), each representing a different level of variable complexity. Each feature set was defined using a structured preprocessing workflow in R, and each had two versions: one for **tree-based models** and one for **non-tree (parametric) models**.

In addition to these four structured pipelines, I also created a separate **Kitchen Sink** feature set that included all available predictors, serving as a maximal model for benchmarking purposes.

4.4.1 Predictor Sets by Model Type

- **Model Type 1 (M1): Group-Only Model**
  Included only the categorical variable for identity group membership (LGBTQ+ migrant, LGBTQ+ non-migrant, straight migrant, straight non-migrant).
- **Model Type 2 (M2): + Demographics**
  Added basic demographic variables such as age, gender identity, race/ethnicity, and geographic region.
- **Model Type 3 (M3): + Socioeconomic Status (SES)**
  Built on M2 by including socioeconomic variables such as education level, insurance coverage, marital status, and parental status.
- **Model Type 4 (M4): + Healthcare Access and Behavior**
  Included variables related to healthcare usage and access (e.g., number of doctor visits, therapy history, housing status, physical health rating).
- **Kitchen Sink**
  A separate feature set that included all available predictors in the dataset, spanning identity, demographic, socioeconomic, behavioral, and healthcare-related variables.

4.4.2 Tree-Based vs. Non-Tree-Based Feature Sets

Each model type was implemented using **two parallel feature sets** depending on the algorithm class:

- **Non-Tree-Based Models** (e.g., Logistic Regression, Naive Bayes, KNN):
  - Dummy encoding for categorical variables
  - Normalization of numeric predictors
  - Mode imputation for categorical variables; median imputation for numeric

  - Removal of near-zero variance predictors

- **Tree-Based Models** (e.g., Decision Tree, RuleFit):
  - One-hot encoding for categorical variables
  - Skipped normalization (not needed for decision-based splits)
  - Same imputation and near-zero variance removal steps as above

This setup ensured that models were trained on appropriately preprocessed data based on their algorithmic assumptions, while keeping predictor sets consistent across model types.

All preprocessing workflows were executed using standard tools in R and saved for reproducibility.

4.5 Resampling Technique

To evaluate the generalizability of all models and reduce the risk of overfitting, I used **V-fold cross-validation with repeated resampling**. This method involves partitioning the dataset into *V* equal-sized folds, training the model on *V–1* folds, and validating it on the remaining fold. The process is repeated so that each fold serves as a validation set multiple times, and the results are averaged to estimate model performance.

Two resampling configurations were used:

- **Local experiments (exploratory phase):**
  Used **4-fold cross-validation with 3 repeats**, resulting in **12 total resamples per model**. This setup allowed for quick prototyping across different feature sets and models on a local machine.
- **Final modeling (Quest high-performance cluster):**
  Used **10-fold cross-validation with 5 repeats**, resulting in **50 resamples per model**. This more rigorous configuration provided stable and generalizable performance estimates for final evaluation.

In both cases, **stratified resampling** was used based on the binary depression outcome to ensure consistent class proportions across all folds.

This two-tiered strategy balanced efficiency and rigor: fast iteration locally, followed by deeper evaluation at scale.

4.6 Performance Metrics

The primary performance metric for model comparison was the **F1 Score**, which balances precision and recall. This metric was selected due to the class imbalance in the target variable and the importance of not over-prioritizing one class over the other.

Secondary metrics included:

- **Accuracy**: the proportion of correctly classified observations
- **Area Under the ROC Curve (AUC)**: measures the ability of the model to distinguish between the two classes across all classification thresholds, reflecting the model's ability to rank positive cases higher than negative ones
- **Precision and Recall**: reported individually to further interpret F1 dynamics

All metrics were computed and averaged across resampling iterations to ensure stable and unbiased comparisons across model types.

5 Results

5.1 Modeling Overview

Model performance was evaluated across five classification algorithms—Decision Tree, RuleFit, K-Nearest Neighbors (KNN), Logistic Regression, and Naive Bayes—using repeated stratified cross-validation on a balanced dataset of 40,000 observations. Each model was tested using a distinct feature set (M1–M4 or Kitchen Sink), resulting in a total of ten model-feature set combinations.

To support both experimentation and scalability, models were run on two platforms: - A **local pipeline** using 4-fold cross-validation with 3 repeats (12 resamples total) - A **Quest high-performance pipeline** using 10-fold cross-validation with 5 repeats (50 resamples total)

Each model-feature set combination was run on only one platform, not both, meaning the evaluation reflects the pairing of model, feature set, and compute environment. The primary evaluation metric was **F1 score**, with **accuracy**, **ROC AUC**, **precision**, and **recall** used as secondary evaluation criteria.

Before evaluating model performance, I fine-tuned the settings for each algorithm using a structured trial-and-error process known as grid search. On local runs, this involved testing 5 different values for each model setting. On Quest, the high-performance cluster, I expanded this to 15 values to take advantage of the additional computing power. This helped ensure that each model was evaluated using its best configuration, based on F1 score.

5.2 Model and Feature Sets Comparison

The best-performing configuration was a **Decision Tree trained on the full feature set (Kitchen Sink)**, trained and evaluated on the **Quest** high-performance computing cluster. This model achieved:

- **F1 Score:** $ 0.901
- **Accuracy:** $ 0.90
- **ROC AUC:** $ 0.95
- **Precision:** $ 0.91
- **Recall:** $ 0.93

Other high-performing configurations included:

- **K-Nearest Neighbors + M4** (Local): strong overall balance with `F1  0.82`
- **K-Nearest Neighbors + M3** (Quest): `F1  0.80`
- **RuleFit + Kitchen Sink** (Quest): `F1  0.89`, slightly below Decision Tree

In contrast, models trained on **M1 (Group Membership Only)** showed significantly lower F1 scores, especially on Local runs, indicating that identity group alone is not a sufficiently strong predictor of depression risk.

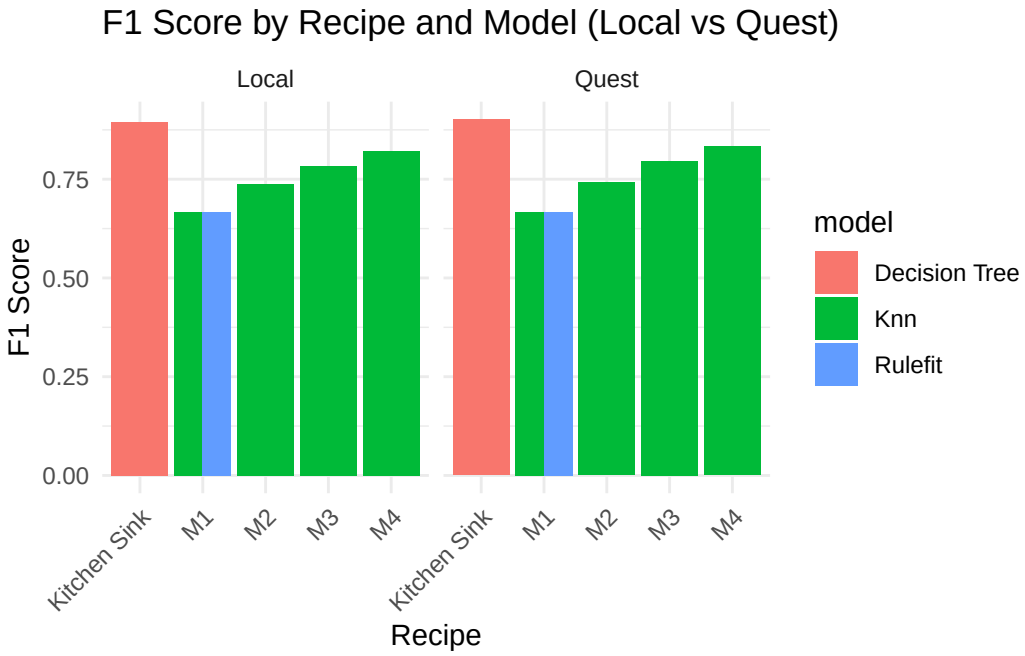

Figure 3: F1 Scores by Model and Feature Set (Grouped by Platform)

5.3 Accuracy and ROC AUC Trends

While F1 score served as the primary evaluation metric, examining **accuracy** and **ROC AUC** provides a more complete picture of model performance.

As shown in **Figure 4**, accuracy trends closely mirrored the F1 score rankings. The **Decision Tree + Kitchen Sink** model (Quest) not only achieved the highest F1 score, but also led in accuracy at approximately **90%**, reinforcing its strong predictive capacity across multiple metrics. Similarly, **KNN models paired with M3 and M4** performed well on both platforms.

In contrast, models trained on **M1 (Group Only)** underperformed in terms of accuracy, particularly with Naive Bayes and RuleFit, echoing their low F1 results.

ROC AUC trends, shown in **Figure 5**, further contextualize these findings. While most high-F1 models also had high AUC scores (above 0.85), some models—such as **KNN + M2 (Quest)**—achieved strong AUC values despite more modest F1 scores. This indicates good class separation overall, but potentially uneven performance on one class (e.g., lower precision or recall).

Naive Bayes models exhibited the weakest AUC scores overall, hovering near or below 0.70 in most in most feature sets and across both platforms, confirming their limited predictive value in this project.

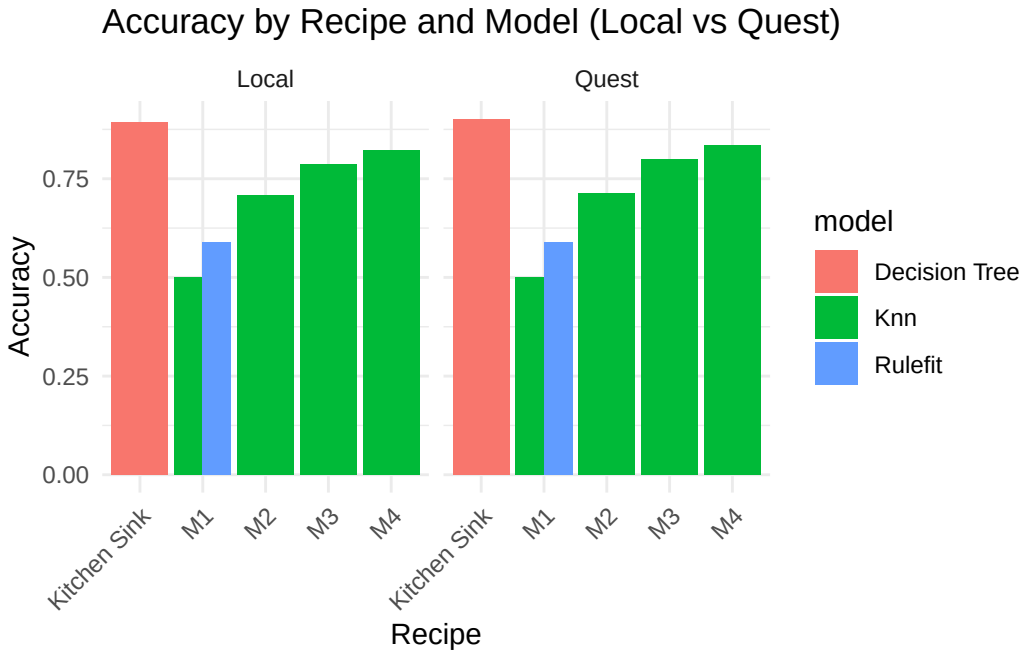

Figure 4: Accuracy by Model and Feature Set (Grouped by Platform)

5.4 Summary of Best-Performing Models

The tables below present the **top five performing model–feature set combinations for each platform**, ranked by F1 score. These configurations represent the strongest overall performance on the local and Quest pipelines, respectively.

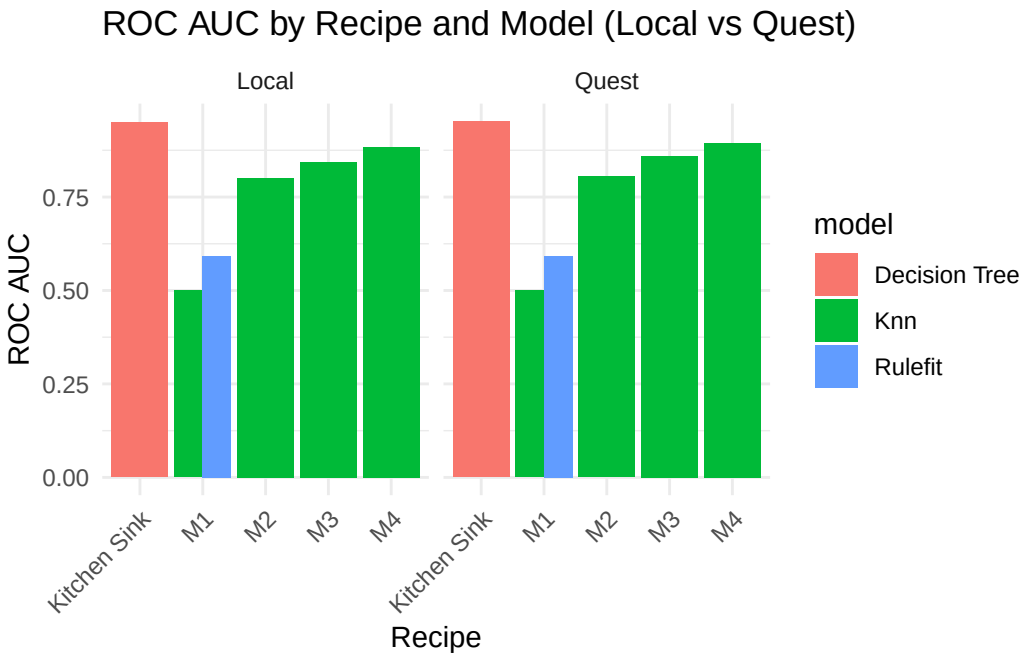

Figure 5: ROC AUC by Model and Feature Set (Grouped by Platform)

Table 1: Accuracy by Model and Feature Set (Grouped by Platform) - Local

model	accuracy	precision	recall	F1 Score	ROC AUC	Std. Error (F1)	Attempt	Platform	Feature Set
Decision Tree	0.894	0.896	0.927	0.894	0.950	0.001	1	Local	Kitchen Sink
KNN	0.708	0.679	0.846	0.739	0.800	0.001	2	Local	M2
KNN	0.787	0.797	0.772	0.784	0.843	0.002	3	Local	M3
KNN	0.500	0.500	1.000	0.667	0.500	0.000	4	Local	M1
RuleFit	0.588	0.604	1.000	0.667	0.592	0.000	4	Local	M1
KNN	0.822	0.828	0.814	0.821	0.884	0.001	5	Local	M4

Table 2: Accuracy by Model and Feature Set (Grouped by Platform) - Quest

model	accuracy	precision	recall	F1 Score	ROC AUC	Std. Error (F1)	Attempt	Platform	Feature Set
Decision Tree	0.902	0.906	0.927	0.901	0.952	0.001	1	Quest	Kitchen Sink
Knn	0.712	0.680	0.864	0.742	0.805	0.001	2	Quest	M2

model	accuracy	precision	recall	F1 Score	ROC AUC	Std. Error (F1)	Attempt	Platform	Feature Set
Knn	0.800	0.811	0.785	0.797	0.859	0.001	3	Quest	M3
Knn	0.500	0.500	1.000	0.667	0.500	0.000	4	Quest	M1
Rulefit	0.588	0.604	1.000	0.667	0.591	0.000	4	Quest	M1
Knn	0.834	0.842	0.824	0.833	0.895	0.001	5	Quest	M4

The results presented in Tables Table 1 and Table 2 reveal several consistent patterns across platforms.

The Decision Tree trained on the full feature set (Kitchen Sink) emerged as the strongest overall performer. It had the highest F1, accuracy, and AUC on Quest, and also ranked near the top on Local. It makes sense—Decision Trees can handle mixed data types and capture interactions really well, especially when given access to all variables.

KNN models also held up surprisingly well, especially in combination with M3 and M4 feature sets. These sets brought in more context—like education, insurance coverage, and ability to afford care—which clearly helped with performance. While KNN is relatively simple in structure, it effectively leveraged the added context to produce competitive results.

In contrast, models trained only on M1 (group membership) performed poorly across both platforms. This isn't too surprising. LGBTQ+ or migrant identity alone doesn't capture the whole picture—there are deeper structural and socioeconomic drivers of depression that aren't accounted for in that feature set.

RuleFit also underperformed in both environments. While it offers interpretability, it might have been too rigid to adapt to the variable complexity or limited by the group-only input.

Finally, it's worth noting that the overall rankings stayed consistent between Quest and Local, even though the number of resamples was different. That consistency gives more confidence that the results aren't just random noise or overfitting to a specific setup. These findings reinforce the value of including sociological and structural context when building predictive models for mental health outcomes.

6 Model Explainability

To interpret the inner workings of the top-performing models, I used four complementary explanation methods as outlined in Chapter 18 of *Tidy Modeling with R*: variable importance (VIP), SHAP values, partial dependence plots (PDPs), and local explanation plots. These were applied to two decision tree models trained on the Quest platform:

- **Model 1:** Decision Treetrained on the full feature set (Kitchen Sink)
- **Model 2:** Decision Tree trained on the M3 feature set

These two models were chosen for their strong predictive performance (`F1  0.90` and `F1  0.79`, respectively) and for capturing two different perspectives: one with the full feature set, and one with a constrained but sociologically rich set of predictors. Together, these models provide insight into what the models prioritized and how predictions were influenced by variables related to mental health, identity, and structural inequality.

6.1 Variable Importance (VIP)

Variable importance plots were generated for both selected models using Gini-based importance scores. These plots reveal which features had the strongest overall influence on model predictions.

In the **Kitchen Sink model** (Figure 6), the most influential predictors were:

- **Anxiety diagnosis** (yes/no)
- **Frequency of depressive episodes**
- **Use of mental health therapy**
- **Reported depression levels**
- **Age and LGBTQ+ migrant group membership**

These results are expected, since this model includes a wide range of behavioral health and access-to-care variables that are strongly correlated with depression. Variables such as anxiety history, frequency of depressive episodes, and therapy access are directly tied to the outcome and thus dominate the decision paths of the tree.

In contrast, the **M3 model** (Figure 7), which excluded healthcare access variables—showed a different set of top predictors:

- **Group membership**, especially LGBTQ+ migrant identity

- **Gender identity and sex**
- **Age**
- **Race and ethnicity**
- **Region and education level**

This model relied more heavily on structural and identity-based variables, given the absence of direct healthcare indicators. As a result, identity and social positioning played a greater role in shaping predictions.

Together, these importance plots illustrate how the inclusion or exclusion of certain feature sets—particularly those related to healthcare access—shifts the relative importance of features in shaping model decisions. This underscores how variable selection affects not only predictive performance but also the interpretive framing of depression risk.

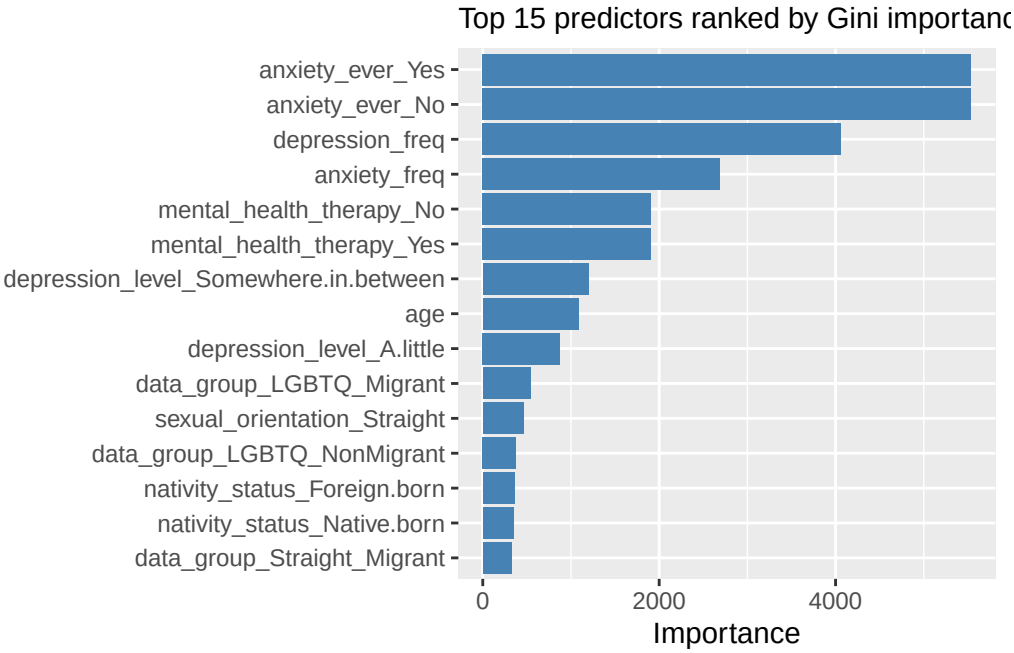

Figure 6: Variable importance plot for Decision Tree + Kitchen Sink model

6.2 SHAP Values

To further interpret how specific features influenced individual predictions, SHAP (SHapley Additive exPlanations) values were used to assess the contribution of individual features to

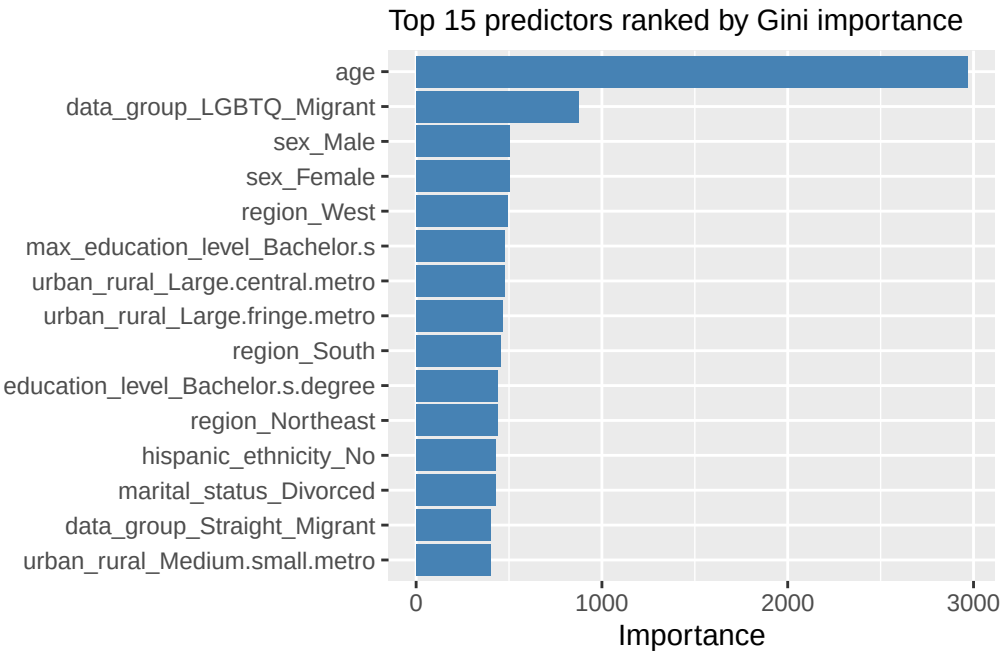

Figure 7: Variable importance plot for Decision Tree + M3 model

model predictions. These plots decompose predictions into additive contributions from each input variable for a representative observation.

In the **Kitchen Sink model** (Figure Figure 8), the top SHAP contributors included:

- **Depression frequency and level**
- **Anxiety diagnosis (No)**
- **Sexual orientation and migrant status**
- **Lack of mental health therapy**

In this case, a combination of behavioral and identity-based factors contributed to higher predicted depression risk. The absence of anxiety history and therapy access had strong negative SHAP contributions, suggesting these features were commonly associated with individuals not exhibiting depressive symptoms—making their absence a significant contributor to elevated risk predictions.

In the **M3 model** (Figure Figure 9), which excluded behavioral health indicators, a different set of patterns emerged. Major contributors included:

- **Marital status**
- **LGBTQ+ migrant identity**
- **Education level**
- **Age and region**

These findings align with the patterns observed in the variable importance plots: in the absence of clinical indicators, the model relied more heavily on demographic and structural variables. For instance, being married or living in certain regions contributed negatively to predicted depression risk, while group membership and lower educational attainment were associated with increased risk.

SHAP enhances model interpretability by clarifying not only which features are important, but also how they influence individual predictions.

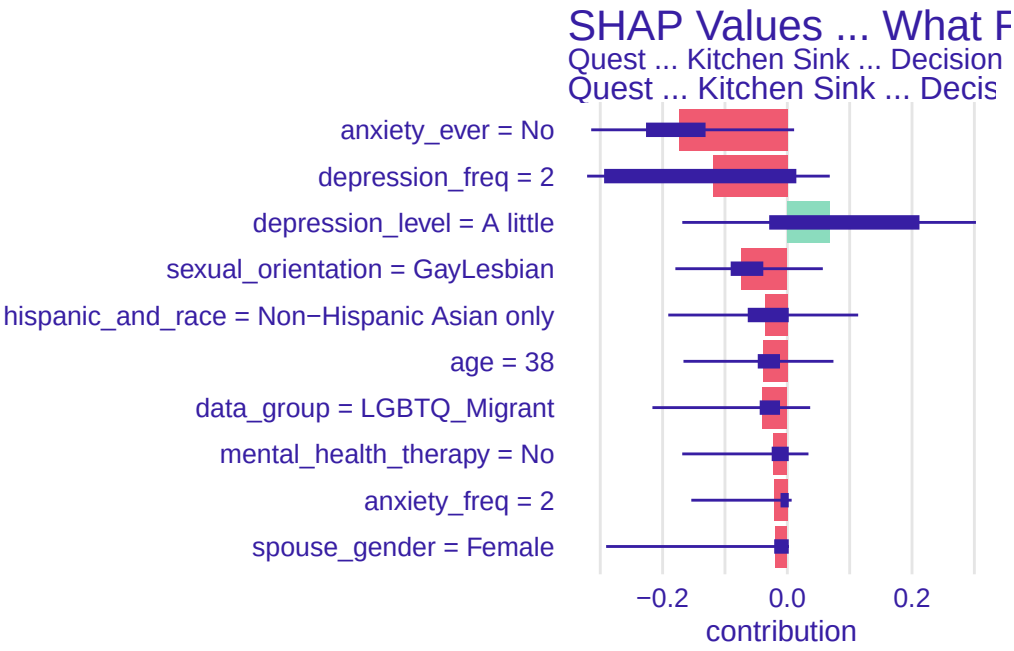

Figure 8: SHAP explanation for a prediction from the Kitchen Sink model

6.3 Partial Dependence Plots (PDPs)

Partial dependence plots (PDPs) were used to explore how individual features influence predicted depression risk, averaged across all individuals in the dataset.

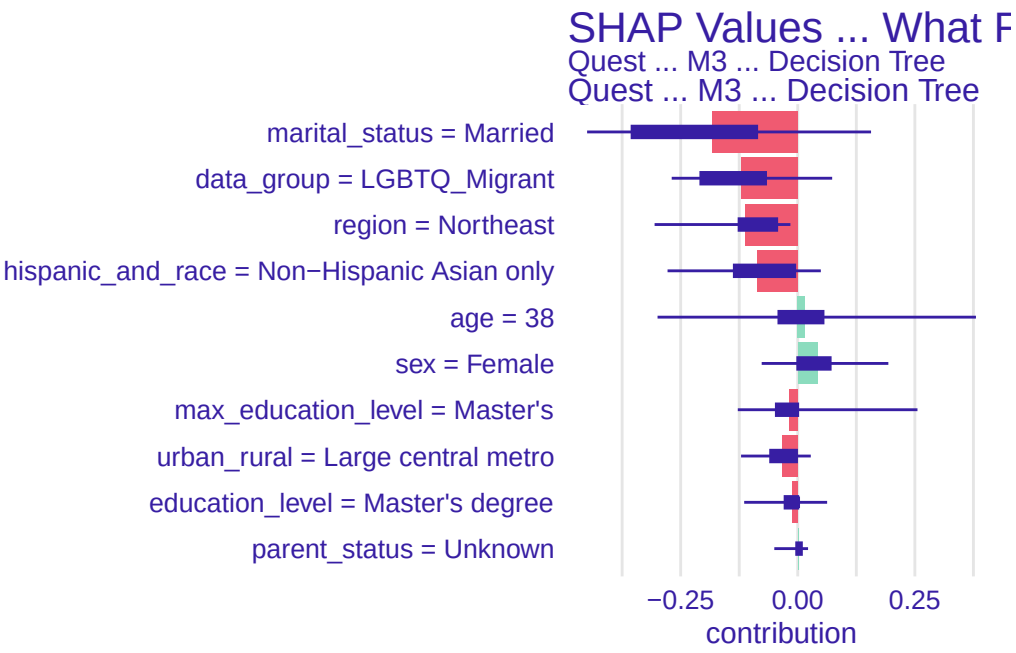

Figure 9: SHAP explanation for a prediction from the M3 model

In the **Kitchen Sink model** (Figure Figure 10), anxiety diagnosis stood out clearly:

- Respondents who had **ever been diagnosed with anxiety** were associated with a **substantially higher** predicted risk of depression.
- This aligns with well-established clinical relationships between anxiety and depression.

Education level also revealed subtle trends:

- Depression risk was **slightly higher** for individuals with a GED or partial high school completion.
- The model predicted **lower risk** among those with more advanced degrees (e.g., Master's, Doctoral), though the differences were not dramatic.

In the **M3 model** (Figure Figure 11), the exclusion of clinical variables shifted the patterns:

- **Anxiety diagnosis** did not show as strong a separation—reflecting its exclusion from the M3 feature set.
- For education, predictions varied less across categories, but the general trend of lower risk with higher education remained.

These PDPs reinforce earlier findings: **behavioral health variables** like anxiety diagnosis provide sharp predictive signals. When such variables are omitted, structural features like education fill the predictive gap, but result in weaker group-level separation.

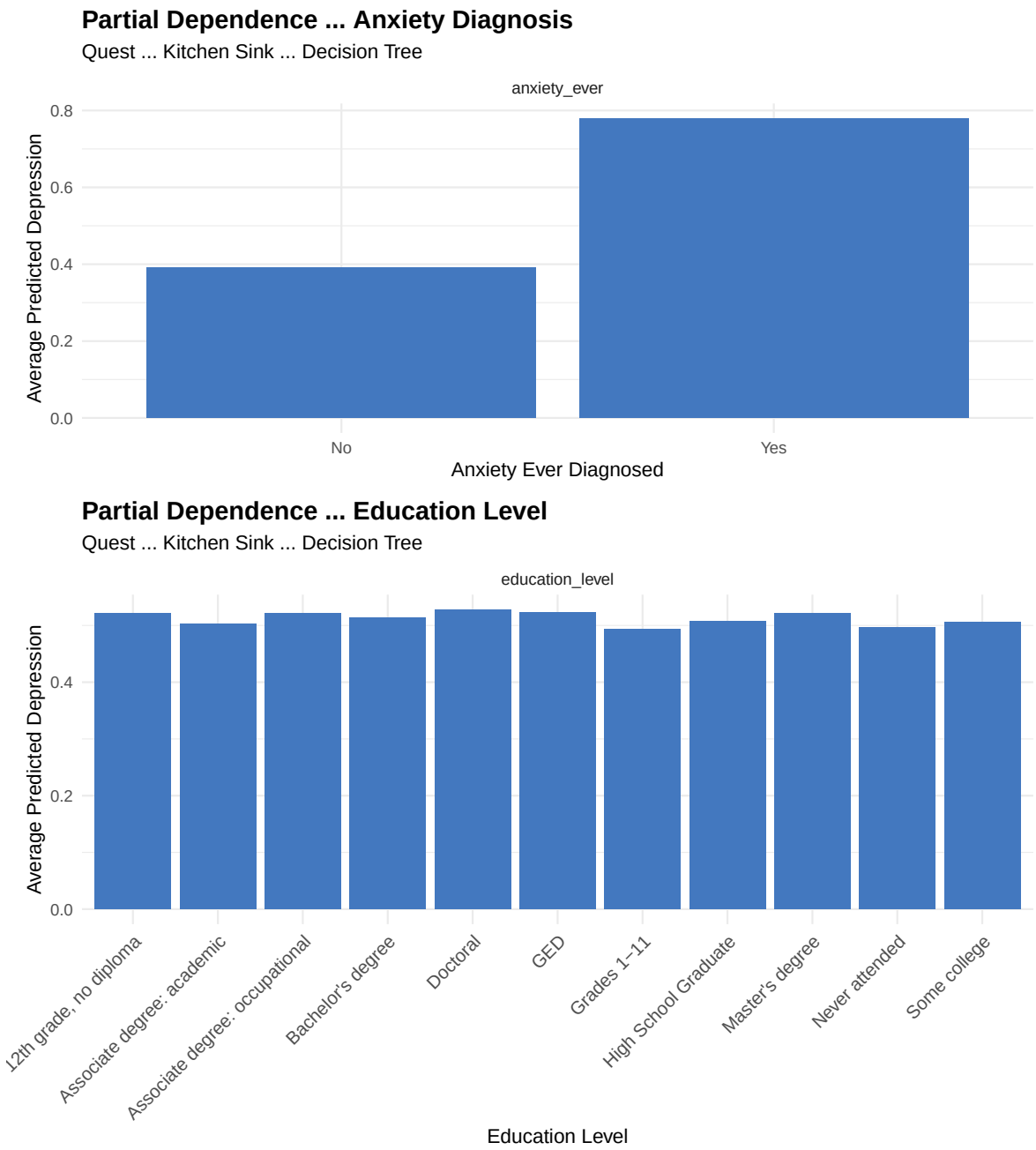

Figure 10: Partial Dependence – Kitchen Sink Model (Quest – Decision Tree)

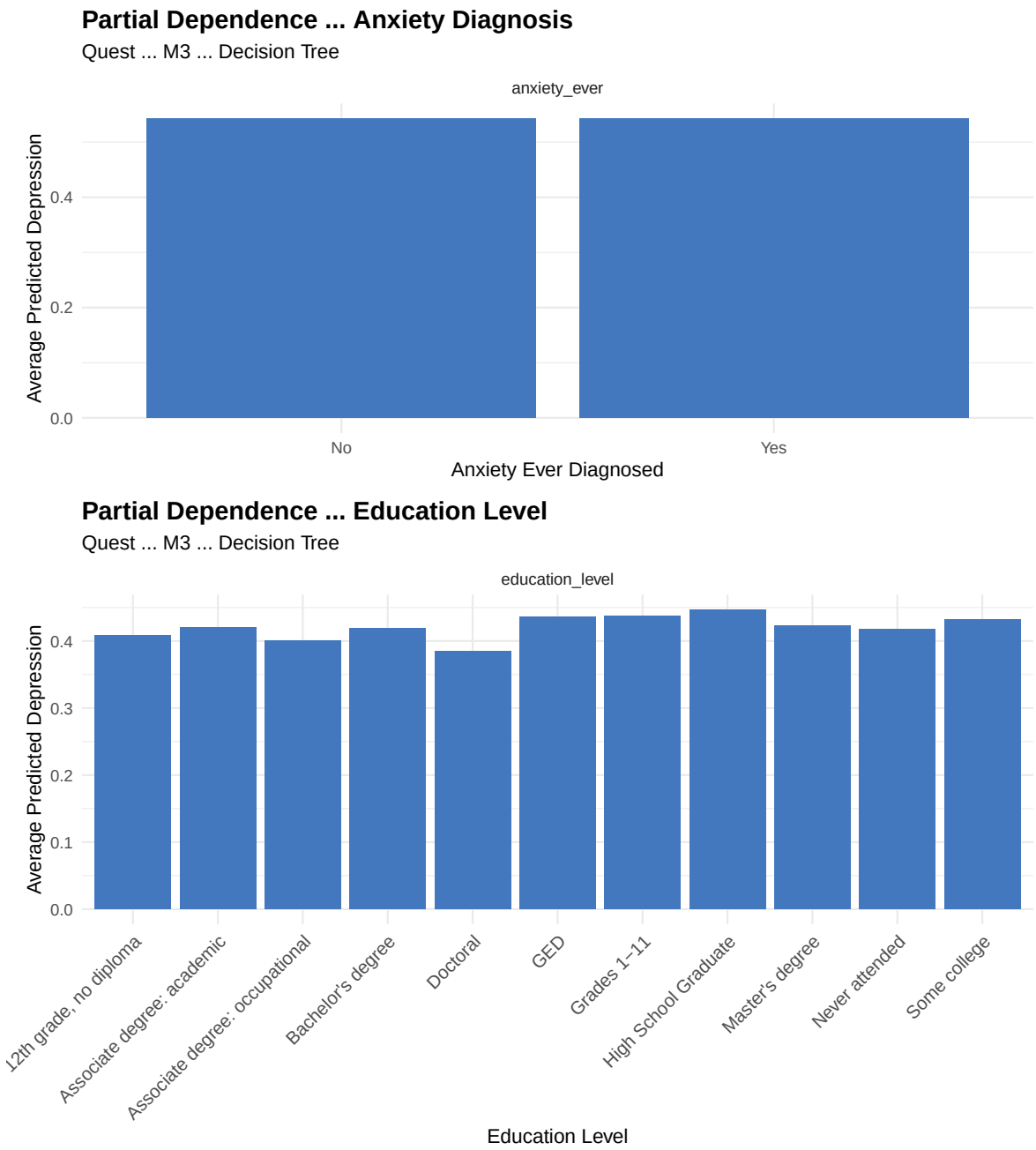

Figure 11: Partial Dependence – M3 Model (Quest – Decision Tree)

6.4 Local Explanations

Local explanation plots were used to visualize how individual features contributed to specific predictions. These plots decompose the predicted probability for a single observation, showing how each feature shifted the prediction above or below the model's baseline (0.5). Explanations were generated for both selected decision tree models: **Kitchen Sink** and **M3**, each trained and evaluated on Quest.

6.4.1 Kitchen Sink Model

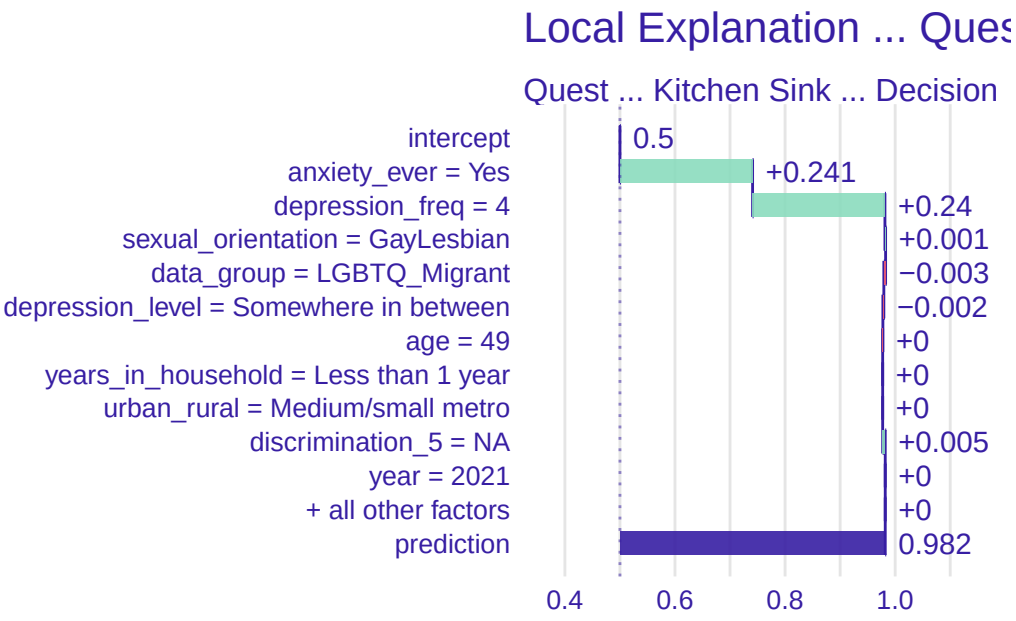

Figure 12: Local explanation for a single observation (Kitchen Sink model – Decision Tree)

In the **Kitchen Sink** model (Figure Figure 12), the prediction for the selected individual reached a high probability of **0.982**, far above the baseline of 0.5. Key positive contributors included:

- **Anxiety diagnosis** and **depression frequency**, which were by far the most influential variables and together accounted for the vast majority of the increase in predicted risk.
- Identity-based variables such as **sexual orientation (Gay/Lesbian)** and **LGBTQ+ migrant group membership** added small but noticeable effects.

Negative or neutral contributors included **age**, **region**, and **household tenure**. However, their magnitudes were minimal compared to the dominant behavioral health factors. This

reinforces the earlier findings that when behavioral and access-to-care variables are available, they overwhelmingly shape model predictions.

6.4.2 M3 Model

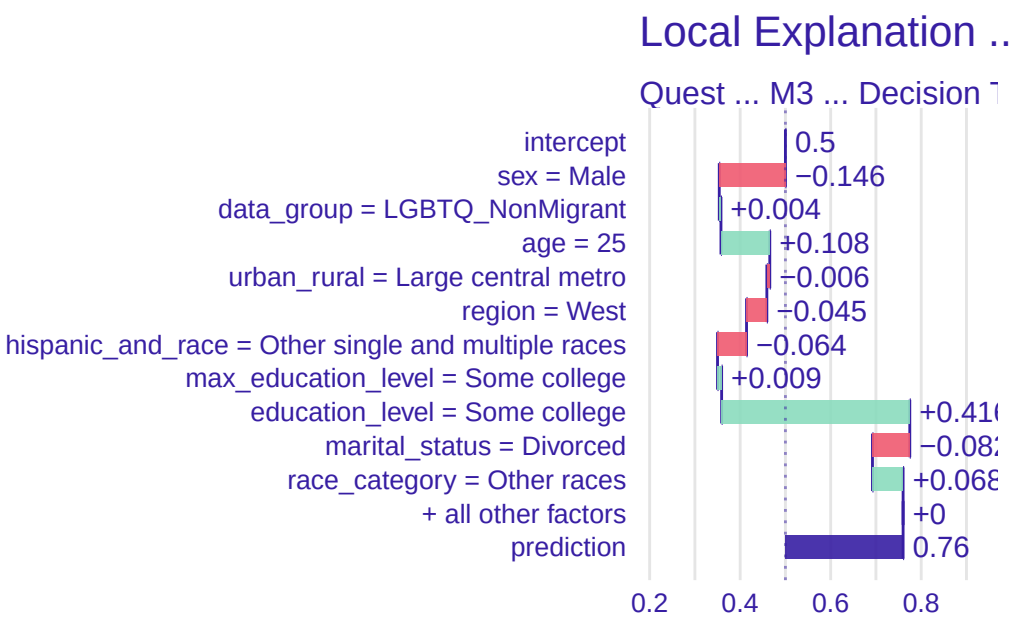

Figure 13: Local explanation for a single observation (Kitchen Sink model – Decision Tree)

In contrast, the **M3** model (Figure Figure 13) predicted a depression probability of **0.76** for a different individual, relying on a more structural set of features. The largest contributors included:

- **Education level (Some college)**, which had a strong positive impact, raising the predicted risk
- **Marital status (Divorced)** and **race (Other)** also contributed toward a higher predicted probability

Conversely, **being male** and **residing in a large metro area** slightly reduced the predicted risk.

This breakdown illustrates how the model compensates for the absence of behavioral health data. Without information on anxiety or therapy use, the model shifts greater weight toward **demographic and social context**, still producing a relatively high-risk prediction, but based on a very different rationale and feature set.

7 Discussion

This study explored the use of interpretable machine learning models to predict depression among individuals with intersecting marginalized identities — specifically LGBTQ+ and migrant populations — using sociological and health variables from the National Health Interview Survey (NHIS). This thesis focused in depth on decision trees due to their intuitive structure and ability to produce clear, visual explanations. Among the multiple models trained and evaluated, those built on Quest using the Kitchen Sink and M3 feature sets demonstrated the strongest performance. The Kitchen Sink model achieved an F-measure of **0.836** and an area under the ROC curve (AUC) of **0.878**, while the M3 model followed closely with an F-measure of **0.812** and an AUC of **0.861**.

These models were trained and evaluated using Northwestern's high-performance computing cluster (Quest), which enabled more extensive cross-validation (15-fold with 5 repeats) and larger training sets than could be processed locally. In preliminary tests, models trained on Quest consistently outperformed those run on a personal laptop, demonstrating the importance of computational resources when working with large-scale health survey data and intensive model evaluation workflows.

To interpret these metrics: the **F-measure** (or F1 score) balances two critical components of model quality — *precision* (how many of the predicted depressed cases were actually correct) and *recall* (how many of the actual depressed individuals the model was able to detect). A perfect score is 1.0; a value above 0.8 indicates strong, reliable performance with limited false positives and negatives. Meanwhile, the **AUC** represents the model's ability to rank individuals correctly in terms of risk: an AUC of 0.878 means that nearly 88% of the time, the model assigns a higher probability of depression to someone who actually experiences it than to someone who does not, indicating a high level of discriminative ability.

Both top models highlighted anxiety diagnosis and self-reported depression frequency as key predictors. In the Kitchen Sink model, SHAP values showed that these two variables contributed the majority of the increase from the base prediction probability of 0.5 to a final predicted probability of **0.982** for a representative high-risk individual. While this aligns with known associations between anxiety and depression, it is important to clarify that "depression frequency" is a self-reported emotional state rather than a clinical diagnosis, and therefore distinct from the outcome variable. Its inclusion helps identify individuals who experience depressive symptoms but may not have received a formal diagnosis—an especially important distinction in the context of underdiagnosis and structural barriers to care. Additionally, it played a consistent and interpretable role across SHAP values, PDPs, and local explanations, making it a valuable contributor to model transparency. Future work could further validate its influence through sensitivity analyses.

Differences in model performance across feature sets also carried important sociological implications. Simpler models like M1, which included only LGBTQ+ and migration status, performed notably worse (F-measure **0.674**), while M3 — which added demographic and socioeconomic

context — demonstrated a significant jump in predictive power (F-measure **0.812**). This supports long-standing sociological theories that emphasize the importance of structural inequality in shaping health outcomes. Identity alone is insufficient for understanding depression risk without incorporating the broader material and geographic conditions individuals navigate.

The dataset was stratified and balanced across four subgroups — LGBTQ+ migrants, LGBTQ+ non-migrants, straight migrants, and straight non-migrants — to avoid skewed performance. In doing so, the top models generalized relatively well across these groups, with subgroup F-measures ranging from **0.811** to **0.849** for the Kitchen Sink model. This is promising, given that data-driven systems often underperform for marginalized populations due to underrepresentation.

Interpretability techniques—including SHAP values, partial dependence plots, and local explanations—deepened understanding of model behavior from both global and individual perspectives. These tools transformed the model from a black box into a transparent system, enabling clearer insight into the decision-making process—an especially valuable property in sociological and public health contexts where trust and accountability are essential.

These interpretability tools each offer a different lens on the model's behavior. **SHAP values** provide a consistent way to assess how much each feature contributes to individual predictions, accounting for feature interactions. **Partial dependence plots** show how changes in a single variable affect the model's output while holding other variables constant, revealing non-linear effects like higher depression risk at lower income levels. **Local explanation plots**, such as individualized SHAP-based visualizations, break down the predicted probability for a specific observation and highlight which features pushed the prediction higher or lower. Together, these methods help translate abstract model behavior into tangible insights, making them especially useful in public health and sociological research.

While decision trees were selected due to their compatibility with interpretability tools, future work should explore the predictive potential of ensemble models like Random Forests or Boosted Trees. Although these are harder to interpret, they may yield improved raw performance. Running them as a benchmark would help clarify whether any significant predictive accuracy was traded off for the sake of explainability — and whether that tradeoff is justified in this context.

Ultimately, this work demonstrates how interpretable machine learning can serve as a bridge between computational rigor and sociological inquiry—supporting both predictive insight and structural reflection in public health research.

8 Conclusion

This thesis set out to predict depression using interpretable machine learning models, with a particular focus on individuals situated at the intersection of structural disadvantage—namely, LGBTQ+ and migrant populations. Drawing on five years of National Health Interview Survey (NHIS) data, the project integrated sociological insight with algorithmic modeling to examine how demographic, medical, and social factors contribute to depression risk.

Decision trees were selected for their clarity and compatibility with interpretability tools, supporting both strong performance and insight into model behavior. Models trained using the Kitchen Sink and M3 feature sets achieved high F-measure and AUC scores, while SHAP values and local explanation plots made it possible to trace individual-level predictions. Stratifying the dataset across four subgroups ensured fairness across intersectional identities, with only minor variation in performance.

The findings highlight not only the predictive value of clinical and demographic variables like anxiety diagnosis, income, and regional context, but also the importance of incorporating marginalized identities into model design. While some predictors, such as self-reported depression frequency, raised questions about conceptual overlap, the overall framework demonstrated that it is possible to build both accurate and socially aware models of mental health risk.

This project contributes to the growing body of work at the intersection of data science and social science by emphasizing transparency, subgroup fairness, and the integration of theory into applied modeling. Future work might explore additional model types, such as random forests or boosted trees, or extend this approach to other health outcomes and survey datasets. More broadly, this work offers a case study in using machine learning not only as a predictive tool, but also as a lens for reflecting on the structural forces that shape mental health across identity lines.

9 References

Adrian B. R. Shatte, S. J. T., Delyse M. Hutchinson. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*, *49*(09), 1426–1448. https://doi.org/10.1017/s0033291719000151

Centers for Disease Control and Prevention (CDC), & National Center for Health Statistics (NCHS). (2024). *National health interview survey (NHIS), 2019–2023*. https://www.cdc.gov/nchs/nhis/index.html

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, *43*(6), 1241. https://doi.org/10.2307/1229039

David Nickson, L. W., Caroline Meyer. (2023). Prediction and diagnosis of depression using machine learning with electronic health records data: A systematic review. *BMC Medical Informatics and Decision Making*, *23*(1). https://doi.org/10.1186/s12911-023-02341-x

David R. Williams, M. S. (2010). Understanding racial-ethnic disparities in health: Sociological contributions. *Journal of Health and Social Behavior*, *51*(1_suppl), S15–S27. https://doi.org/10.1177/0022146510383838

David R. Williams, S. A. M. (2009). Discrimination and racial disparities in health: Evidence and needed research. *Journal of Behavioral Medicine*, *32*(1), 20–47. https://doi.org/10.1007/s10865-008-9185-0

Dominic B. Dwyer, N. K., Peter Falkai. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*(1), 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Fiedorowicz, J. G., & Haynes, W. G. (2010). Cholesterol, mood, and vascular health: Untangling the relationship. *Current Psychiatry*, *9*(7), 17–A. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4215473/

Irene A Kretchy, S. A. D., Frances T Owusu-Daaku. (2014). Mental health in hypertension: Assessing symptoms of anxiety, depression and stress on anti-hypertensive medication adherence. *International Journal of Mental Health Systems*, *8*(1). https://doi.org/10.1186/1752-4458-8-25

Jason A. Bonomo, J. A. R., Kate Luo. (2024). LGBTQ+ cardiovascular health equity: A brief review. *Frontiers in Cardiovascular Medicine*, *11*. https://doi.org/10.3389/fcvm.2024.1350603

Kwang-Sig Lee, B.-J. H. (2022). Machine learning on early diagnosis of depression. *Psychiatry Investigation*, *19*(8), 597–605. https://doi.org/10.30773/pi.2022.0075

Marissa Tan, D. T., Elham Hatef. (2020). Including social and behavioral determinants in predictive models: Trends, challenges, and opportunities. *JMIR Medical Informatics*, *8*(9), e18084. https://doi.org/10.2196/18084

Meyer, I. H. (2013). Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: Conceptual issues and research evidence. *Psychology of Sexual Orientation and Gender Diversity*, *1*(S), 3–26. https://doi.org/10.1037/2329-0382.1.s.3

Miriam M. Moagi, P. M. J., Anna E. van Der Wath. (2021). Mental health challenges of

lesbian, gay, bisexual and transgender people: An integrated literature review. *Health SA Gesondheid*, *26*. https://doi.org/10.4102/hsag.v26i0.1487

Ninad T. Maniar, M. B. D. (2024). From invisibility to inclusion: A call to action to address COPD disparities in the lesbian, gay, bisexual, transgender, and queer+ community. *Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation*, *11*(3), 326–330. https://doi.org/10.15326/jcopdf.2024.0496

Organization, W. H. (2023). *Promoting the health of refugees and migrants: Experiences from around the world*. World Health Organization.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

10 Appendix

10.1 Reproducibility & Data Access

This project involved modeling and data processing across two environments:
- a local machine, and
- Northwestern University's Quest high-performance computing cluster.

Due to file size and platform constraints, not all files — including raw outputs from Quest — could be hosted on GitHub.

Instead, a complete reproducibility folder has been shared via Google Drive. It contains:

- Cleaned datasets derived from the National Health Interview Survey (NHIS)
- Model outputs and final figures
- Quest-specific job scripts
- The rendered thesis PDF and LaTeX logs

Google Drive folder:
https://drive.google.com/drive/folders/1dKNWv3BAB8BY1Y5YK-MlU2HG12Ur1chd?usp=sharing

Code for preprocessing, modeling, and figure generation is available on GitHub:
https://github.com/ali3el/alis-thesis.git

To fully reproduce the results, download the Drive folder, follow the directory structure described in the included `README`, and refer to the GitHub repo for all code dependencies and documentation.

10.2