# Predicting Mental Health Treatment-Seeking Behavior in the Tech Industry Using Machine Learning: A Model Evaluation, Fairness Assessment, and Explainability Study

*A Project Report Prepared By:*

**MUKISA Vicent**
**2025/HD05/26353U**

**Makerere University**

**Master of Science in Computer Science (MSCS)**

**Course Unit MCS 7103: Machine Learning**

**2025/2026**

**25th November 2025**

# Table of Contents

# Abstract

This project investigates the factors influencing mental-health treatment-seeking among employees in the technology industry by developing a predictive machine-learning model using the 2014 OSMI Mental Health in Tech Survey. After cleaning and preprocessing the data, several classification models were trained, including Logistic Regression, RidgeCV, LassoCV, Naïve Bayes, and a Decision Tree. RidgeCV emerged as the most reliable model, demonstrating strong accuracy, high recall, and stable generalization. The analysis showed that work interference, family history, and workplace support features, such as benefits and care options, were the strongest predictors of treatment-seeking behavior, while demographic factors played a relatively minor role. Fairness evaluation revealed moderate performance differences across gender and regional groups, reflecting underlying data imbalances rather than demographic dependence. Model interpretability, achieved through SHAP and LIME, confirmed that predictions were primarily driven by meaningful behavioral and workplace-related features. Overall, the study demonstrates that machine learning, when paired with fairness checks and transparent interpretability, can help identify mental-health risk patterns in organizational settings. The findings emphasize that such models should support, not replace, human judgment, and highlight opportunities for future work using richer datasets, expanded fairness metrics, and more expressive models.

# 1 Introduction

Mental-health challenges are increasingly recognized as a significant concern within the technology industry, where high workloads, tight deadlines, remote work pressures, and cultural stigma can shape whether employees seek professional support. Despite growing awareness, many individuals still avoid or delay treatment due to workplace perceptions, personal hesitation, or lack of supportive structures. Understanding the factors that influence help-seeking behavior is therefore important for designing better organizational policies and promoting employee well-being.

This project uses data from the **2014 OSMI Mental Health in Tech Survey**, a widely referenced dataset capturing demographic characteristics, workplace conditions, and perceptions of mental-health support among employees in the tech sector. The survey provides a rich foundation for modeling because it includes both personal attributes (age, gender, family history) and workplace factors (benefits, supervisor support, anonymity, work interference), allowing for a holistic understanding of treatment-seeking patterns.

The goal of this study is to build a predictive machine-learning model that estimates whether a respondent is likely to seek mental-health treatment, based on the features provided in the dataset. Beyond prediction, the project seeks to identify the most influential factors driving these decisions and evaluate how consistently the model performs across demographic groups.

To achieve this, the analysis focuses on three core machine-learning objectives:

i. **Classification:** Train and compare models, including Logistic Regression, RidgeCV, LassoCV, Naïve Bayes, and Decision Trees, to determine which approach best predicts treatment-seeking behavior.
ii. **Fairness:** Assess whether the model performs equitably across key demographic groups, particularly gender and region.
iii. **Interpretability:** Use SHAP and LIME to explain global and local model behavior, ensuring transparency in how predictions are made.

By combining predictive modeling with fairness checks and interpretability, this project aims to provide both accurate insights and responsible use of machine learning in a sensitive mental-health context.

# 2 Methodology

## 2.1 Dataset Overview

This project uses the 2014 OSMI Mental Health in Tech Survey (survey.csv), sourced from Kaggle. The raw dataset contained 1,259 records and 27 variables, covering demographic information, workplace conditions, and mental-health perceptions within the tech industry. After cleaning and removing non-informative fields, the final working dataset consisted of 1,251 respondents and 25 variables.
The target variable, treatment_Yes, indicates whether a respondent has sought professional mental-health treatment. The remaining features fall into four main categories:

i. **Demographics**: age, gender, country, and state.
ii. **Employment attributes**: self_employed, remote_work, number of employees, tech_company.
iii. **Workplace factors**: benefits, care options, anonymity, supervisor support, coworker support, leave difficulty.
iv. **Perceptions and consequences**: reported mental-health consequences, physical-health consequences, and observed stigma.

Initial inspection showed that most variables were categorical, age was the only numeric feature, and several columns contained missing or inconsistent entries (for example, more than 40 variations of gender labels, missing values in state and work_interfere, and unstructured text in comments). These findings informed the cleaning and preprocessing steps completed during data preparation.

## 2.2 Data Preparation

Data preparation involved cleaning the raw survey responses, transforming categorical variables, and creating a standardized dataset suitable for machine-learning models. These steps ensured consistency, preserved data integrity, and produced a high-quality feature set for modeling.

### 2.2.1 Data Cleaning

Initial inspection revealed several inconsistencies and missing values. The following corrections were applied:

i. **Dropped irrelevant fields**: The Timestamp and comments columns were removed because they contained no predictive or structured information.
ii. **Handled missing values:** Missing entries in state, *work_interfere*, and *self_employed* were replaced with "Unknown", retaining respondents rather than discarding incomplete records.

iii. **Age correction:** Unrealistic age values were removed by restricting the range to 15–80 years, producing a cleaned age range of 15–79.

iv. **Gender normalization:** More than 40 inconsistent gender entries were consolidated into three standardized categories: *male*, *female*, and *other*.

After cleaning, the dataset contained 1,251 respondents and 25 variables, with no remaining missing values.

### 2.2.2 Feature Transformation for Modeling

To make the dataset compatible with machine-learning algorithms, several preprocessing steps were completed.

#### 2.2.2.1 One-hot encoding

All categorical variables were converted to numeric form using `pd.get_dummies()` with `drop_first=True`. This expanded the dataset to **1,251 rows × 137 columns**, transforming categorical responses into binary 0/1 indicators such as:

- Gender_Female
- family_history_Yes
- benefits_Yes
- work_interfere_Sometimes

This ensured that no information was lost while enabling algorithms to interpret categorical attributes correctly.

#### 2.2.2.2 Train–Test Split

The encoded dataset was split into:

- Training set: 1,000 samples & testing set: 251 samples

The split was stratified on the target variable (treatment_Yes) to preserve the original class distribution:

- Sought treatment: 632 and did not seek treatment: 619

This balance supports unbiased model evaluation.

#### 2.2.2.3 Feature Scaling

Numeric features were standardized using StandardScaler, mainly affecting the Age variable.

- Before scaling: mean ≈ 32.03, std ≈ 7.23
- After scaling: mean ≈ 0.00, std ≈ 1.00

The fitted scaler was saved for reproducibility.

Scaling ensures that all numeric features contribute proportionally to model learning and prevents dominance by features with larger ranges.

#### 2.2.2.4 Final Dataset Structure

At the end of preprocessing:

i. The dataset was fully cleaned, encoded, and normalized.
ii. The final feature matrix contained 136 independent variables and one binary target.
iii. The dataset (OSMI_encoded_phase4_ready.csv) provided a stable foundation for modeling, fairness analysis, and interpretability work.

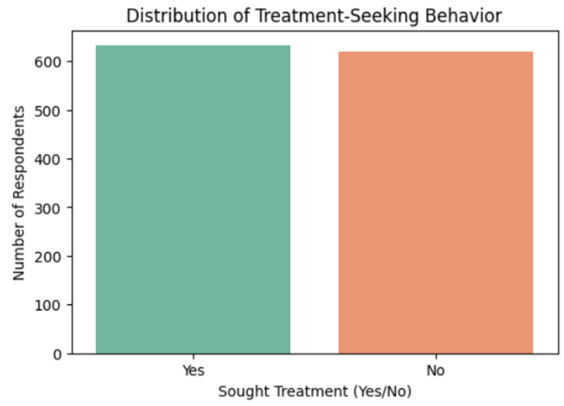
## 2.3 Exploratory Data Analysis (EDA)

An exploratory analysis was used to understand demographic patterns, behavioral indicators, and workplace factors associated with mental-health treatment-seeking.

### 2.3.1 Target Variable Distribution

The cleaned dataset contained 1,251 respondents, with 56 percent reporting treatment and 44 percent not seeking treatment, resulting in a balanced target suitable for classification.

***Figure:*** *Distribution of Treatment-Seeking Behavior*




| treatment | count |
|---|---|
| Yes | 50.52 |
| No | 49.48 |

*Figure 1: This chart shows that the dataset is nearly balanced, with 50.52% of respondents reporting that they sought mental-health treatment and 49.48% reporting they did not.*

### 2.3.2 Demographic Patterns

Age was concentrated between 25 and 40 years, which reflects the typical working population in the technology sector.

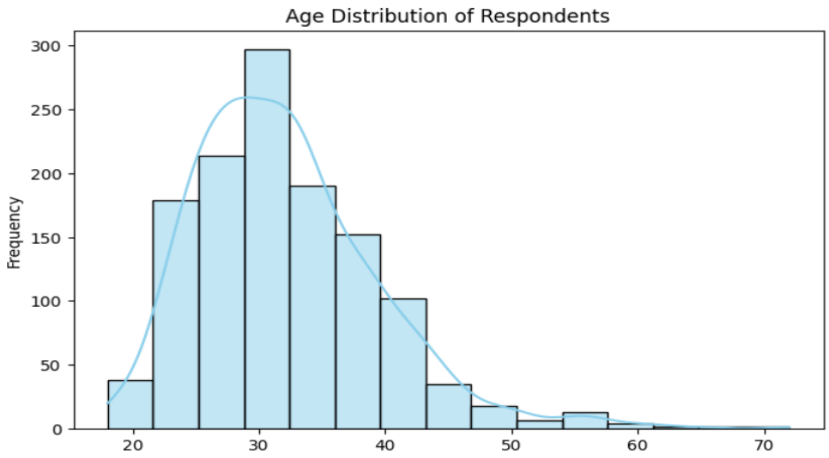



*Figure 2: Histogram showing the age spread of respondents, with most participants clustered between 25 and 35 years.*

Gender distribution was highly imbalanced, with approximately 80 percent male, 20 percent female, and very few identifying as other. This imbalance motivated the fairness analysis conducted later.

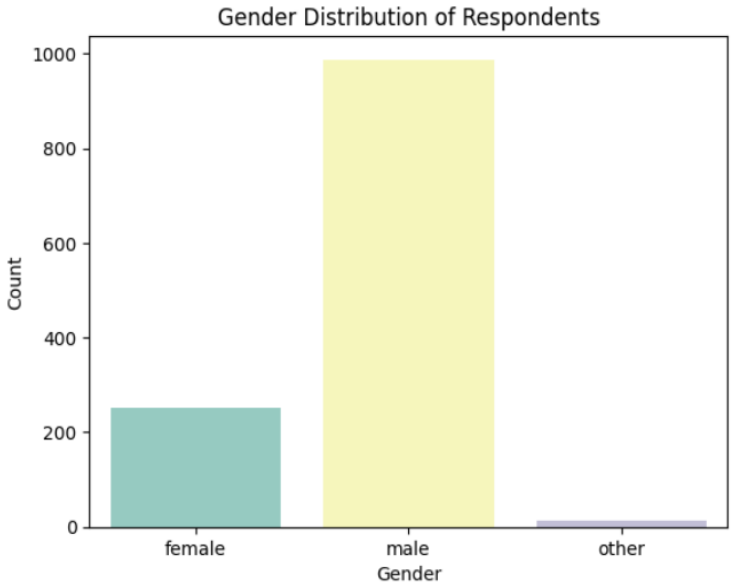



*Figure 3: Most respondents identify as male, with female respondents forming a smaller portion and "other" gender identities being very few. This imbalance reflects what we observed during the EDA phase.*

### 2.3.3 Treatment-Seeking Across Demographic Groups

Female respondents were more likely to seek treatment (about 68 percent) compared to males (46 percent), suggesting gender differences in openness and help-seeking behavior.

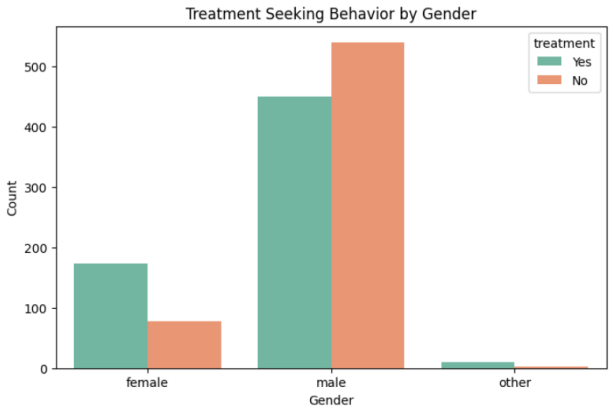



*Figure 4: Females show a higher rate of seeking treatment compared to males, even though males form the largest group in the dataset.*

Respondents with a family history of mental illness were also much more likely to seek treatment (~73 percent vs ~35 percent without family history), making this a strong behavioral predictor.

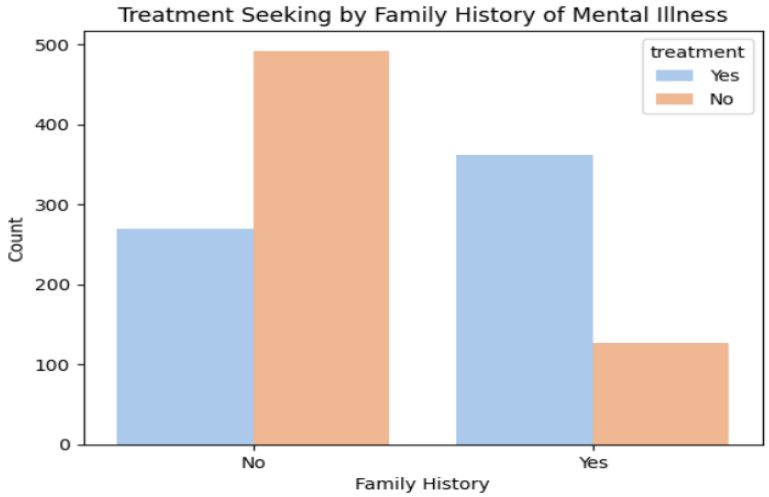



*Figure 5: Respondents with a family history of mental illness are far more likely to seek treatment compared to those without such a history.*

### 2.3.4 Workplace Factors

Work interference ("Sometimes" or "Often"), employer mental-health benefits, and access to care options showed visible associations with increased treatment-seeking. These factors reflect the influence of organizational support and daily stress on mental-health decisions.

### 2.3.5 Correlation Overview

A correlation heatmap of the encoded dataset showed weak to moderate correlations and minimal multicollinearity. Strongest associations with treatment included:

- work_interfere levels
- family_history_Yes
- benefits_Yes

Demographic attributes contributed less.

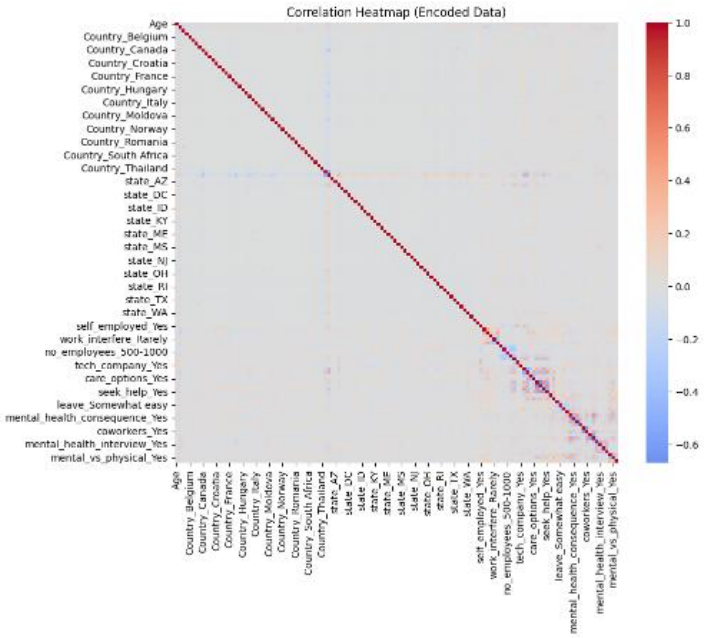



*Figure 6: Shows correlations among all encoded features, with stronger clusters appearing around work-interference and care-related variables.*

### 2.3.6 Key Insights

| Category | Finding | Implication |
|---|---|---|
| Target distribution | Balanced | Good for classification |
| Age | 25–40 dominates | Reflects the tech workforce |
| Gender | Male-heavy | Requires fairness analysis |
| Family history | Strong influence | Key behavioral predictor |
| Work interference | Strong influence | High explanatory power |
| Workplace factors | Moderate to strong | Organizational role |
| Correlations | Weak–moderate | No multicollinearity |

### 2.3.7 EDA Conclusion

The EDA highlighted work interference, family history, and workplace support as the most influential factors associated with treatment-seeking. No major statistical issues were found, and relationships were consistent with expected behavioral patterns. This provided a solid foundation for model development and interpretability in the next phases.

## 2.4 Modelling Approach

### 2.4.1 Overview of Modeling Strategy

The goal of this project is to predict whether a respondent is likely to seek mental-health treatment, which makes the problem a binary classification task. Because the dataset contains a mix of demographic, behavioral, and workplace features, the modeling strategy focused on evaluating both linear and non-linear approaches to capture different types of relationships in the data.

A baseline Logistic Regression model was first developed to establish initial performance and to determine whether the target was linearly separable. Two regularized linear models, RidgeCV (L2) and LassoCV (L1), were then trained to improve generalization and handle the large number of one-hot encoded features. Finally, a Decision Tree and a Naïve Bayes classifier were included as non-linear and probabilistic baselines to contrast performance across learning paradigms.

Model performance was evaluated using five standard metrics:

i. **Accuracy** to measure overall correctness
ii. **Precision** to assess false-positive control
iii. **Recall** to capture sensitivity to treatment seekers
iv. **F1-Score** to balance precision and recall
v. **ROC-AUC** to evaluate discrimination between classes across thresholds

These metrics provide a comprehensive view of each model's strengths and trade-offs and support the selection of a final model for interpretability and fairness analysis.

### 2.4.2 Baseline Model, Logistic Regression

Logistic Regression was selected as the baseline classifier to provide a simple and interpretable starting point for the modelling process. As a linear model, it estimates the relationship between the features and the likelihood of seeking treatment through a weighted combination of inputs. The model was trained on standardized features and evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. These metrics were chosen because they offer a balanced assessment of classification performance, especially in datasets with near-balanced class distributions.

### 2.4.3 Regularized Linear Models: RidgeCV and LassoCV

To improve generalization and address the high-dimensional one-hot encoded feature space, two regularized linear models were trained: RidgeCV (L2) and LassoCV (L1). Regularization penalizes large coefficients, reducing overfitting and enhancing stability. Ridge shrinks coefficients proportionally while retaining all predictors, whereas Lasso can reduce some coefficients to zero, effectively performing feature selection. Both models use cross-validation to automatically determine the optimal regularization strength (alpha).

### 2.4.4 Additional Baseline Models

#### 2.4.4.1 Naïve Bayes Classifier

The Naïve Bayes classifier was included as a lightweight probabilistic baseline model. It applies Bayes' theorem under the assumption that all features are conditionally independent given the target label. Although this assumption does not always hold in real-world data, Naïve Bayes is computationally efficient, easy to interpret, and often performs well on high-dimensional datasets. Its inclusion provides a useful contrast to regularized linear models by offering a fundamentally different modelling perspective.

#### 2.4.4.2 Decision Tree Classifier

The Decision Tree classifier was used to explore potential non-linear relationships and feature interactions that linear models may not capture. Decision Trees construct a hierarchy of "if–then" decision rules, allowing the model to partition the feature space based on splits that best reduce impurity. This structure provides clear interpretability and makes it possible to understand how specific features influence predictions. Hyperparameters such as maximum depth and splitting criteria were applied to control model complexity and reduce the risk of overfitting.

## 2.5 Methodology Summary and Conclusion

In summary, the methodology combined systematic data preparation, exploratory analysis, and a diverse set of modelling techniques to ensure a robust prediction framework. The dataset was cleaned, encoded, and scaled to support reliable learning across a high-dimensional feature space. Multiple models, ranging from simple linear baselines to regularized regressions, probabilistic methods, and a non-linear decision tree, were implemented to capture different types of relationships in the data. Cross-validation, standardized evaluation metrics, and interpretability tools were incorporated to promote fairness, transparency, and model stability. This foundation sets the stage for the next section, which presents and analyzes the results produced by each model.

# 3 Results

## 3.1 Model Results

This section presents the performance of all trained models on the test dataset. Each model was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC to provide a balanced view of its predictive ability. The results for each classifier, Logistic Regression, RidgeCV, LassoCV, Naïve Bayes, and the Decision Tree, are summarized below, together with their confusion matrices, ROC curves, and key predictive insights. These outcomes form the basis for the comparison and analysis in the subsequent sections.

### 3.1.1 Logistic Regression

Logistic Regression was used as the baseline model to establish a reference point for evaluating more advanced approaches. It provides a simple, interpretable linear classifier that is well-suited for binary prediction tasks such as determining whether an individual is likely to seek mental-health treatment. The results below present its performance on the test set, followed by its confusion matrix, ROC curve, and the key features influencing its predictions.

#### 3.1.1.1 Model Performance

| Metric | Score | Interpretation |
|---|---|---|
| Accuracy | 0.805 | The model correctly classified about 80.5% of all respondents. |
| Precision | 0.805 | When predicting that a respondent will seek treatment, it is correct 80.5% of the time. |
| Recall | 0.811 | The model successfully identifies 81.1% of actual treatment seekers. |
| F1-Score | 0.808 | Indicates a strong balance between precision and recall. |
| ROC-AUC | 0.883 | Shows excellent ability to distinguish between treatment seekers and non-seekers. |

### 3.1.1.2 Confusion Matrix

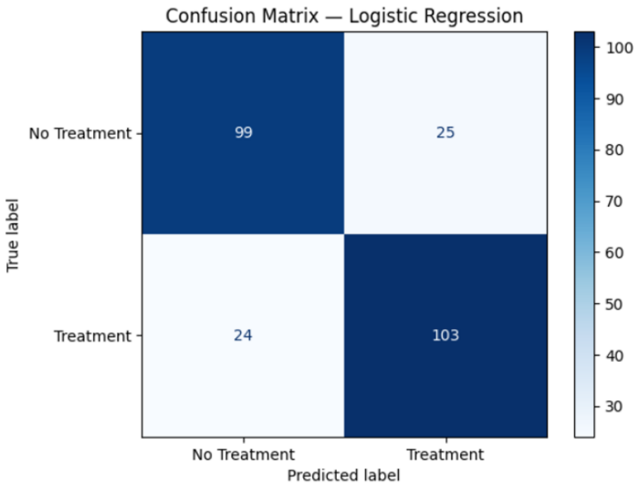



*Figure 7: Confusion matrix showing how the Logistic Regression model classified treatment vs. non-treatment cases, with most predictions correctly aligned to the true labels.*

The confusion matrix shows:

- True Negatives (99): Correctly identified non-treatment cases.
- True Positives (103): Correctly identified treatment seekers.
- False Positives (25): Predicted treatment, but the person had not sought it.
- False Negatives (24): Missed treatment seekers.

The low number of false classifications indicates that the model performs reliably across both classes.

### 3.1.1.3 ROC Curve

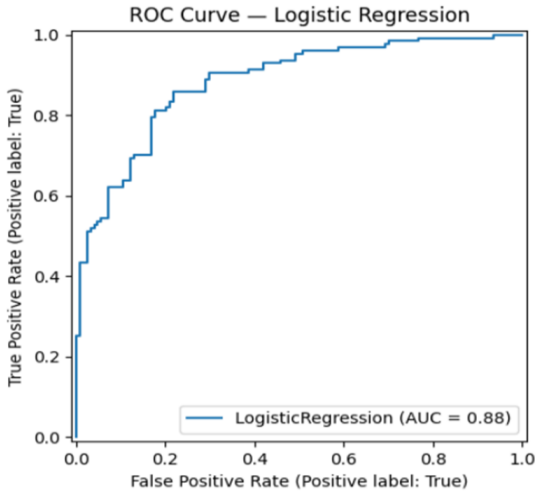



*Figure 8: The ROC curve shows that Logistic Regression achieves strong discrimination ability, with an AUC of 0.88, indicating the model reliably distinguishes between those who seek treatment and those who do not.*

The ROC curve rises sharply toward the top-left corner, showing strong separability between treatment and non-treatment cases.
The AUC of 0.88 confirms the model's ability to maintain a high true-positive rate while controlling false positives across threshold settings.

#### 3.1.1.4 Key Predictive Features

Coefficient analysis revealed the factors that most strongly increased or decreased the probability of seeking treatment.

Top predictors increasing treatment-seeking:

- work_interfere_Sometimes (1.47)
- work_interfere_Often (1.19)
- work_interfere_Rarely (0.89)
- family_history_Yes (0.54)
- benefits_Yes (0.47)
- coworkers_Yes (0.42)
- care_options_Yes (0.38)
- anonymity_Yes (0.32)

These features highlight that work interference, family history, and workplace support are major drivers of help-seeking.

Top predictors decreasing treatment-seeking:

- work_interfere_Unknown (−1.15)
- seek_help_No (−0.47)
- seek_help_Yes (−0.41)
- supervisor_Yes (−0.27)
- Country/State indicators such as Austria, Italy, OK, MD, PA (−0.26 to −0.35)

Negative coefficients mainly reflect silence, uncertainty, or low disclosure, suggesting that respondents who do not report interference or who show hesitation toward seeking help are less likely to have engaged with treatment.

#### 3.1.1.5 Interpretation and Summary

The Logistic Regression model provides a strong, balanced baseline, correctly predicting most treatment seekers and non-seekers while maintaining a high ROC-AUC. The confusion matrix and coefficient patterns align with insights from the EDA phase work interference, family history, and clear workplace support systems consistently encourage treatment-seeking, while nondisclosure and perceived stigma suppress it.

These results establish a solid benchmark for evaluating more advanced models such as RidgeCV, LassoCV, Naïve Bayes, and Decision Trees in the next sections.

### 3.1.2 RidgeCV (L2 Regularization)

RidgeCV applies an L2 penalty to smooth coefficient estimates and reduce variance, making it well-suited for datasets with many correlated or redundant predictors. In this project, RidgeCV provided more stable and slightly stronger performance than the baseline Logistic Regression while maintaining interpretability.

#### 3.1.2.1 RidgeCV Model Performance

| Metric | Score | Interpretation |
|---|---|---|
| Accuracy | 0.825 | Correctly predicts 82.5% of respondents, slightly higher than the baseline model. |
| Precision | 0.794 | When predicting treatment, 79.4 percent of positive predictions are correct. |
| Recall | 0.882 | Identifies 88.2 percent of actual treatment seekers, reducing missed cases. |
| F1-Score | 0.836 | Balanced blend of precision and recall, indicating consistent performance. |
| ROC-AUC | 0.898 | Excellent discrimination between treatment and non-treatment cases. |

**Regularization Strength:**
The cross-validated optimal alpha was 100.0, indicating a strong L2 penalty. This suggests the presence of multicollinearity and redundant predictors in the dataset. RidgeCV handles such conditions effectively by shrinking coefficients without removing variables.

#### 3.1.2.2 Key Predictive Features

RidgeCV reinforced the same core predictors identified in the baseline model:
Strong positive predictors (increase the likelihood of seeking treatment):

- work_interfere_Sometimes
- work_interfere_Often
- family_history_Yes
- benefits_Yes
- care_options_Yes

These highlight the importance of workplace strain, family history, and organizational support.
Negative predictors (reduce the likelihood of treatment-seeking):

- work_interfere_Unknown
- seek_help_No

- Gender_male

These reflect patterns of nondisclosure, hesitation, and gender-related differences.

RidgeCV improved generalization, increased recall, and achieved the highest discrimination among linear models (ROC-AUC ≈ 0.90). Its stable coefficient patterns and strong predictive performance make it the final model selected for fairness and interpretability analyses. RidgeCV offers a reliable balance between accuracy, robustness

### 3.1.3 LassoCV (L1 Regularization)

LassoCV applies an L1 penalty that drives less influential coefficients to zero, allowing the model to perform automatic feature selection. This makes Lasso especially valuable for high-dimensional datasets such as this one, where one-hot encoding produces many sparse predictors. The objective was to simplify the model, reduce overfitting, and highlight the most important factors influencing mental-health treatment-seeking.

#### 3.1.3.1 LassoCV Model Performance

| Metric | Score | Interpretation |
|---|---|---|
| Accuracy | 0.821 | Correctly classifies 82.1 percent of respondents, very close to RidgeCV. |
| Precision | 0.773 | About 77.3 percent of predicted treatment seekers are correct. |
| Recall | 0.913 | Identifies 91.3 percent of actual treatment seekers, the highest recall among all models. |
| F1-Score | 0.838 | Strong balance between precision and recall. |
| ROC-AUC | 0.909 | Excellent ability to discriminate between the two classes. |

#### 3.1.3.2 Regularization Strength and Model Simplicity

The optimal regularization strength was $\alpha = 0.01$, indicating a mild penalty that was sufficient to remove noise without dramatically altering the model's structure. Lasso retained 52 active features out of 136, eliminating approximately 62 percent of weaker predictors. This reduction improved interpretability while preserving strong predictive performance, making Lasso a compact but expressive model.

#### 3.1.3.3 Key Predictive Features

Top positive predictors (increase likelihood of treatment):

- work_interfere_Sometimes, Often, Rarely
- family_history_Yes
- care_options_Yes
- benefits_Yes

- anonymity_Yes
- coworkers_Yes
- state_NV, Country_United Kingdom

These patterns indicate that work interference, family background, and supportive workplace structures remain the strongest motivators for help-seeking.

Top negative predictors (decrease the likelihood of treatment):

- work_interfere_Unknown (–0.06)
- Gender_male (–0.02)
- Country_Colombia, Hungary, Czech Republic
- state_PA, state_SD, state_VA

These features reflect nondisclosure, hesitation, and region-specific or cultural differences in mental-health openness.

LassoCV achieved the highest recall (0.91) and one of the highest ROC-AUC scores (0.91), making it particularly effective at identifying individuals who are likely to seek treatment. By reducing the feature set to 52 variables, it delivered a lean, interpretable, and generalizable model. The model consistently emphasized the same behavioral and workplace drivers found in RidgeCV and the baseline analysis, while exposing clear barriers such as nondisclosure and lower male help-seeking tendencies.

Although RidgeCV was selected as the final model for fairness and interpretability analysis due to its stability, LassoCV offers valuable insight into feature importance and model simplification.

### 3.1.4 Naïve Bayes Classifier

Naïve Bayes is a simple probabilistic classifier based on Bayes' theorem and the assumption that features are conditionally independent given the target class. It is computationally efficient and easy to interpret, making it a useful benchmark to contrast with linear and regularized models. In this project, Naïve Bayes was applied to assess how a generative modeling approach performs on the mental-health survey data.

#### 3.1.4.1 Naïve Bayes Model Performance

| Metric | Score | Interpretation |
|---|---|---|
| Accuracy | 0.518 | Correctly predicts only about 52 percent of cases, close to random guessing. |
| Precision | 0.513 | Half of the positive predictions are incorrect, indicating many false positives. |
| Recall | 0.961 | Identifies 96.1 percent of treatment seekers, but at the cost of overpredicting the positive class. |
| F1-Score | 0.668 | Driven largely by the inflated recall rather than balanced performance. |
| ROC-AUC | 0.513 | Very poor discrimination between classes; essentially no better than chance. |

#### 3.1.4.2 Interpretation

Although Naïve Bayes achieves extremely high recall, it suffers from very low accuracy, precision, and discriminative power. The model predicts "treatment" for almost everyone, producing many false positives. This behavior suggests that the independence assumption underlying Naïve Bayes does not hold in this dataset. Key variables such as work interference, benefits, and family history are correlated, and the algorithm is too simplistic to capture these relationships.

Naïve Bayes achieved high sensitivity but poor overall predictive reliability. Its low accuracy and near-random ROC-AUC (≈ 0.51) indicate that it cannot meaningfully distinguish between treatment seekers and non-seekers. This reinforces that mental-health behaviors depend on multiple interacting factors, which Naïve Bayes cannot model. For this reason, it is not suitable as a final model but remains valuable as a baseline comparison.

### 3.1.5 Decision Tree Classifier

A Decision Tree Classifier was trained to evaluate how a non-linear, rule-based model performs compared to the linear and regularized models. Decision Trees partition the feature space using hierarchical if–then rules, allowing the model to capture feature interactions and threshold effects that linear models may miss.

### 3.1.5.1 Model Performance

| Metric | Score | Interpretation |
|---|---|---|
| Accuracy | 0.801 | Correctly classifies about 80 percent of respondents, comparable to Logistic Regression, RidgeCV, and LassoCV. |
| Precision | 0.742 | Slightly lower precision due to more false positives. |
| Recall | 0.929 | Very high recall, identifying most treatment seekers. |
| F1-Score | 0.825 | Balanced overall performance. |
| ROC-AUC | 0.832 | Good discrimination ability, though weaker than RidgeCV and LassoCV. |

The tree demonstrates strong sensitivity (recall) and provides interpretable decision rules, though its precision and discrimination are slightly weaker than those of the regularized models.

### 3.1.5.2 Decision Tree Feature Importance

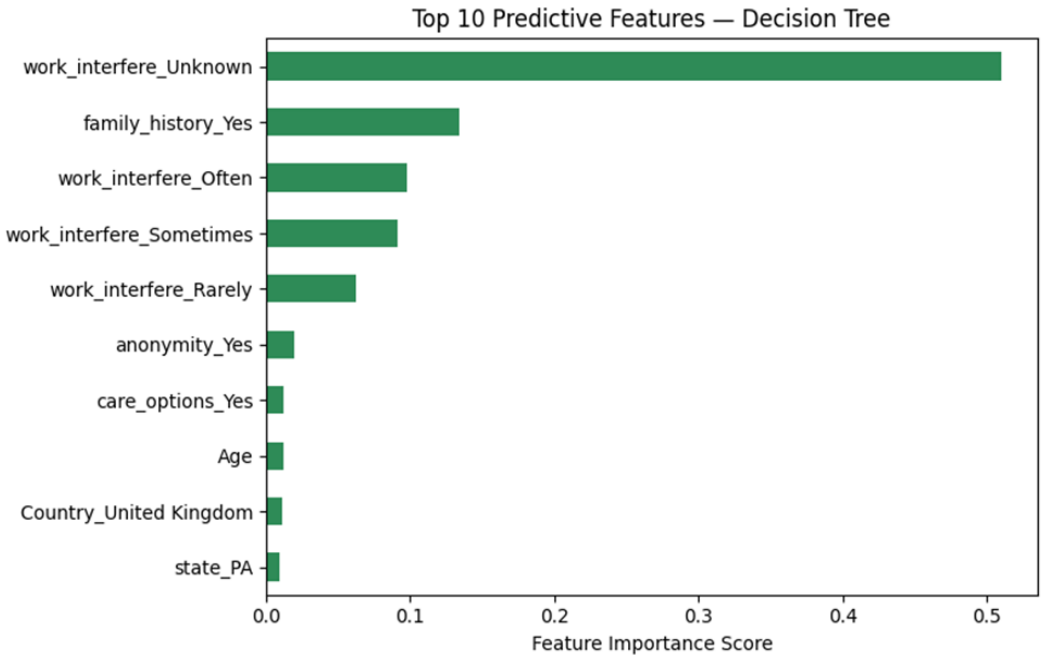



*Figure 9: Shows which features the Decision Tree relied on most when making predictions.*

The top-ranked features were:

i. work_interfere_Unknown: strongest predictor; nondisclosure is itself highly informative.
ii. family_history_Yes: strong positive influence on treatment-seeking.
iii. work_interfere_Often / Sometimes / Rarely: confirms that workplace interference strongly impacts help-seeking.
iv. anonymity_Yes, care_options_Yes, Age, Country_United Kingdom, state_PA: smaller but meaningful contributions.

The prominence of *work_interfere_Unknown* suggests that avoiding reporting mental-health interference may indicate hidden or unacknowledged distress.

### 3.1.5.3 Decision Tree Diagram

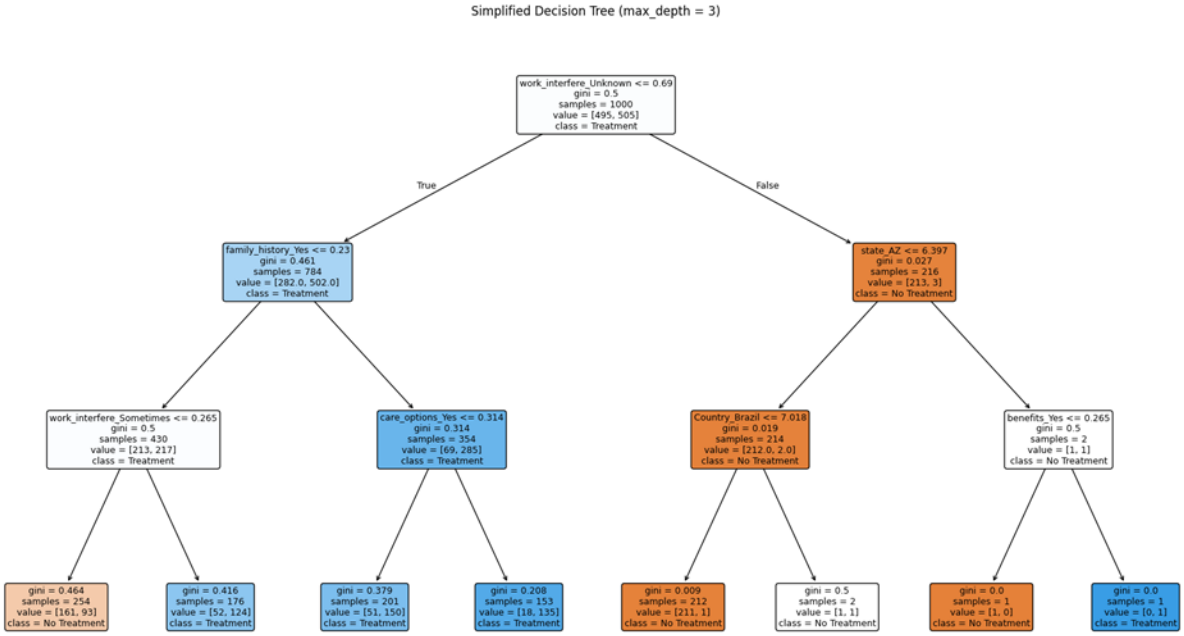



*Figure 10: A simplified decision-tree (max_depth = 3) showing key splits the model uses to predict treatment-seeking, mainly driven by work interference, family history, care options, and location-based features.*

A simplified tree was generated to illustrate the model's decision logic. Several patterns emerge:

i. Work interference is the first and strongest split, confirming its central importance.
ii. Family history and care options form the next major splits, aligning with the EDA and SHAP findings.
iii. Geographical indicators (e.g., Brazil, AZ) appear in deeper branches, showing mild regional differences.
iv. Leaf nodes with highly skewed outcomes (e.g., [211, 1] or [18, 135]) reveal localized but uneven class distributions.

The simplified tree demonstrates how the model combines behavioral and workplace features into clear rule-based pathways.

#### 3.1.5.4 Decision Tree Summary

The Decision Tree performs competitively with the linear models, offering strong recall and intuitive interpretability. It captures non-linear interactions, such as the combined effect of work interference and family history, which validates patterns identified earlier through EDA.
However, despite its interpretability, ***the Decision Tree was not selected*** as the final model because:

i. It is more prone to overfitting with high-dimensional one-hot encoded data.
ii. It produces unstable decision structures (small data changes can alter the tree).
iii. Its precision and ROC-AUC are weaker than RidgeCV and LassoCV.
iv. Regularized linear models generalize more reliably across groups.

Thus, the Decision Tree serves primarily as an interpretable baseline model, supporting insight into feature interactions rather than functioning as the final predictive model.

## 3.2 Models Comparison

This section compares the performance of all trained models, Logistic Regression, RidgeCV, LassoCV, Naïve Bayes, and Decision Tree, to identify the most effective approach for predicting treatment-seeking behavior. Each model was evaluated using Accuracy, Precision, Recall, F1-Score, and ROC-AUC to provide a balanced view of predictive power, reliability, and discrimination ability. The results highlight trade-offs between linear, regularized, probabilistic, and non-linear models.

### 3.2.1 Model Comparison Summary

| Model | Accuracy | Precision | Recall | F1-Score | ROC-AUC | Key Strength / Observation |
|---|---|---|---|---|---|---|
| Logistic Regression | 0.805 | 0.805 | 0.811 | 0.808 | 0.883 | Balanced baseline; good generalization and interpretability |
| RidgeCV (L2) | 0.825 | 0.794 | 0.882 | 0.836 | 0.898 | Strongest overall model; excellent ROC-AUC and generalization |
| LassoCV (L1) | 0.821 | 0.773 | 0.913 | 0.838 | 0.909 | Highest recall; compact model due to feature selection |
| Naïve Bayes | 0.518 | 0.513 | 0.961 | 0.668 | 0.513 | Very high recall but poor precision and near-random discrimination |

| | | | | | | |
|---|---|---|---|---|---|---|
| Decision Tree | 0.801 | 0.742 | 0.929 | 0.825 | 0.832 | High recall and interpretability; captures non-linear relationships |

### 3.2.2 Interpretation

The regularized models, **RidgeCV and LassoCV**, delivered the strongest results overall. RidgeCV achieved the best balance of accuracy, stability, and discrimination, making it the most reliable general-purpose model. LassoCV achieved the highest recall and strongest ROC-AUC, while offering valuable feature-selection benefits.

Naïve Bayes produced extremely high recall but suffered from poor precision and very weak ROC-AUC, suggesting that its independence assumption is not suitable for this dataset. The Decision Tree performed competitively and offered clear interpretability, but its precision and overall stability were weaker compared to the regularized linear models.

Logistic Regression performed well as a baseline, confirming that the dataset has a strong linear structure and acts as a useful benchmark for evaluating other models.

### 3.2.3 Conclusion

RidgeCV emerged as the best overall model, offering the strongest combination of predictive performance, generalization, and interpretability. LassoCV provided complementary insights through feature selection and high recall, while the Decision Tree contributed interpretive value by revealing non-linear decision rules. Together, these results guided the selection of RidgeCV as the final model for fairness and interpretability analysis.

## 3.3 Fairness Analysis

Fairness analysis assesses whether the final model (RidgeCV) performs consistently across key demographic groups. Because the OSMI dataset exhibits notable demographic imbalances, particularly by gender and region, and because EDA revealed group-level differences in treatment-seeking behavior, it is important to evaluate whether the model's predictions remain stable and equitable. This ensures the responsible application of machine learning in a sensitive mental health context.

### 3.3.1 Gender Fairness

#### 3.3.1.1 Performance by Gender

| Gender | Accuracy | Precision | Recall | FPR |
|---|---|---|---|---|
| Female | 0.9375 | 0.9722 | 0.9459 | 0.0909 |
| Male | 0.7980 | 0.7333 | 0.8556 | 0.2478 |

#### 3.3.1.2 Interpretation

The model performs substantially better for female respondents. Female accuracy (93.75 percent) and precision (97.22 percent) are both extremely high, and recall remains strong at 94.59 percent. In contrast, male performance is notably lower, with accuracy dropping to 79.80 percent and precision to 73.33 percent. The higher false-positive rate for males (24.78 percent) indicates that the model overpredicts treatment-seeking among male respondents.

These outcomes mirror the EDA findings:

i. Female respondents showed clearer, more consistent help-seeking patterns (~68 percent treatment rate).
ii. Male respondents displayed more mixed behavior and higher proportions of "Unknown" responses (~45 percent treatment rate).

#### 3.3.1.3 Gender Fairness Conclusion

The model is more reliable for females than for males. This fairness gap is driven by data representation and behavioral variation, not algorithmic bias. However, the disparity highlights the need for cautious interpretation of predictions for male respondents.

### 3.3.2 Regional Fairness

#### 3.3.2.1 Performance by Region

| Region | Accuracy | Precision | Recall | FPR |
|---|---|---|---|---|
| **US** | 0.8487 | 0.8556 | 0.8851 | 0.2000 |
| **Europe** | 0.7813 | 0.6452 | 0.8696 | 0.2683 |
| **Other** | 0.8000 | 0.7500 | 0.8824 | 0.2778 |

#### 3.3.2.2 Interpretation

Model performance is strongest for US respondents, who form the largest regional group in the dataset. More data points enable the model to learn clearer and more stable patterns from this population.

Performance is lower for Europe and Other regions, where precision drops (Europe: 64.52 percent) and false-positive rates rise (Europe: 26.83 percent; Other: 27.78 percent). These results reflect a combination of smaller sample sizes, cultural diversity, and more varied treatment-seeking behaviors.

#### 3.3.2.3 Regional Fairness Conclusion

The model generalizes best to US respondents, with weaker performance for Europe and Other regions. These disparities stem from uneven data distribution and cultural heterogeneity, rather than intentional bias.

### 3.3.3 Combined Fairness Summary

| Group Type | Group | Accuracy | Precision | Recall | FPR |
|---|---|---|---|---|---|
| **Gender** | Female | 0.9375 | 0.9722 | 0.9459 | 0.0909 |
| **Gender** | Male | 0.7980 | 0.7333 | 0.8556 | 0.2478 |
| **Region** | US | 0.8487 | 0.8556 | 0.8851 | 0.2000 |
| **Region** | Europe | 0.7813 | 0.6452 | 0.8696 | 0.2683 |
| **Region** | Other | 0.8000 | 0.7500 | 0.8824 | 0.2778 |

This summary shows that disparities exist across both gender and region groups, with the widest performance differences between female and male respondents.

### 3.3.4 Overall Fairness Conclusion

Although the RidgeCV model performs strongly at an overall level, it does not perform uniformly across demographic groups. It is significantly more accurate and reliable for female respondents and somewhat more stable for US respondents than for European or Other regions.

These differences are primarily driven by:

i. Uneven sample representation
ii. Behavioral variation identified in EDA
iii. Cultural differences in mental-health reporting
iv. Greater clarity and consistency in some groups compared to others

While the model is not intentionally discriminatory, its reliability varies across demographic groups, emphasizing the importance of fairness evaluation in mental-health prediction tasks. Predictions for underrepresented groups should therefore be interpreted with caution and contextual awareness.

## 3.4 Model Interpretability (SHAP & LIME)

As mental-health prediction may influence sensitive workplace or policy decisions, interpretability is essential to ensure transparency, trust, and ethical use. Although RidgeCV was selected as the final model for its strong generalization and stability, linear models can still produce complex interactions that are not immediately intuitive. To address this, the project applies SHAP (SHapley Additive exPlanations) for both global and local interpretability, complemented by optional LIME for additional local explanation.

SHAP assigns each feature a contribution value for a prediction based on cooperative game theory principles. This allows us to quantify how much each feature pushes the model toward predicting "Treatment = Yes" or "No Treatment," both overall (global importance) and for individual respondents (local explanation). The results validate whether model behavior aligns with the patterns discovered during EDA.

### 3.4.1 SHAP Global Interpretability

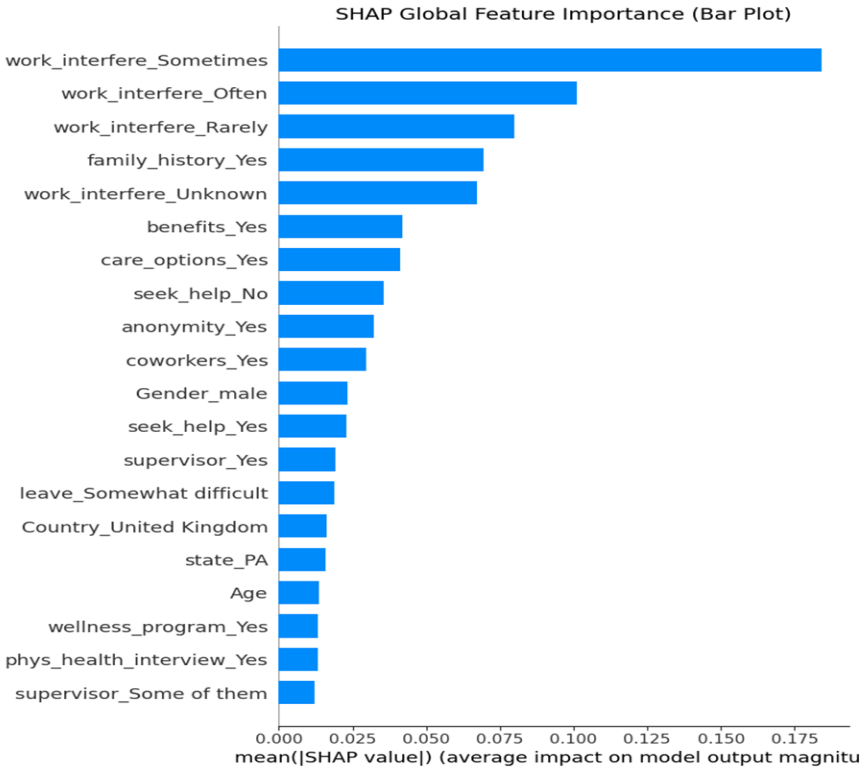



*Figure 11: This plot shows the average absolute SHAP value for each feature, indicating its overall influence on the model's predictions. Higher bars correspond to stronger predictive impact.*

The SHAP global importance plot shows that work interference features dominate the model's predictive power. Specifically:

- work_interfere_Sometimes, work_interfere_Often, work_interfere_Rarely, work_interfere_Unknown

together form the strongest contributors to predicting treatment-seeking.

Other consistently influential predictors include:

- family_history_Yes, benefits_Yes, care_options_Yes, seek_help_No
- anonymity_Yes, coworkers_Yes

Demographic attributes such as gender, age, and country/state appear in lower positions, indicating that the model relies far more on behavioral and workplace-related features than demographic identity. This aligns with the ethical requirement to avoid demographic over-reliance.

#### 3.4.1.1 SHAP Summary Plot (Beeswarm)

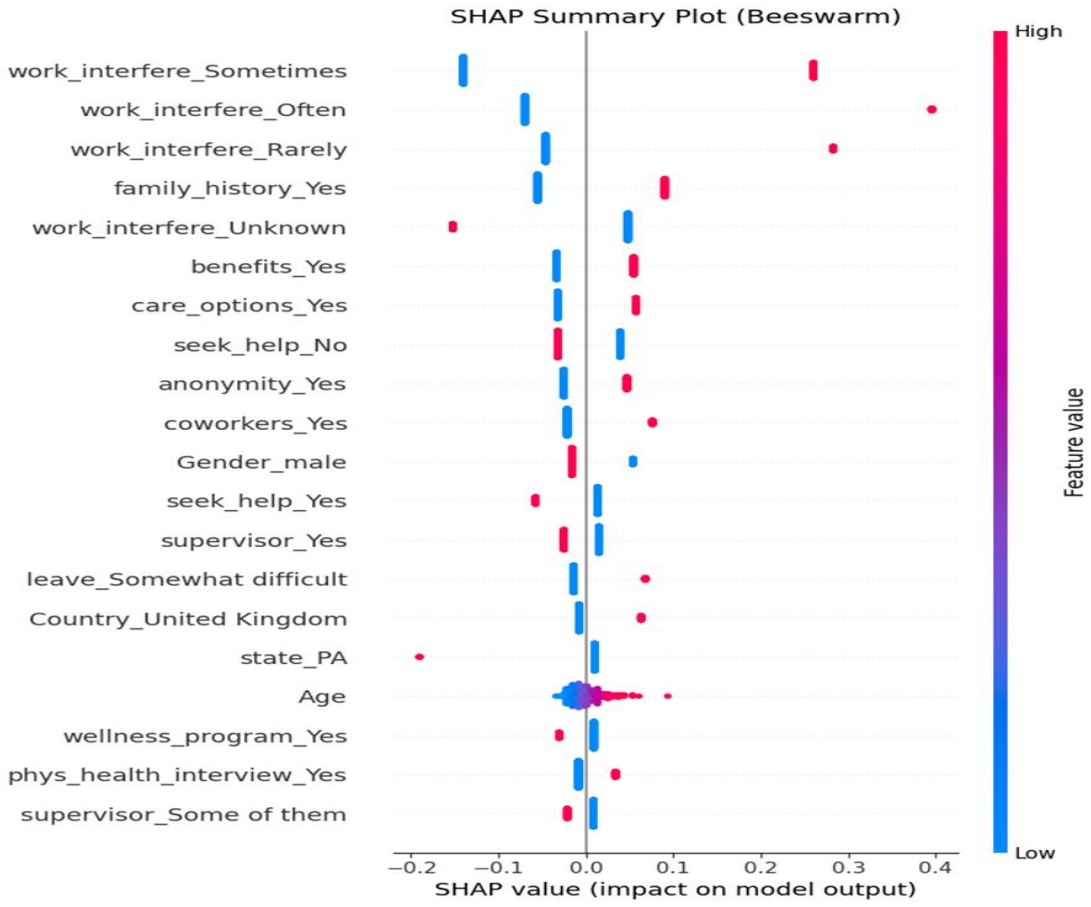



*Figure 12: This plot shows the distribution of SHAP values for each feature, indicating both magnitude and direction of influence on the model's predictions. Red points represent high feature values, blue points represent low values.*

The SHAP beeswarm plot provides deeper insight into how feature values affect predictions.

**Key observations:**

- High values of work interference ("Sometimes", "Often", "Rarely") consistently push the prediction toward Treatment = Yes, reflecting increased workplace strain.
- Positive risk indicators such as family history, available benefits, and care options also shift predictions upward, reflecting awareness and organizational support.
- Negative behavioral indicators, such as seek_help_No, push predictions toward No Treatment, reflecting stigma or lack of workplace openness.
- Demographic variables (gender, age, region) exhibit much smaller SHAP impacts, confirming that the model is not driven by sensitive demographic attributes.

Overall, the bees' warm results validate that the model is appropriately relying on behavioral and contextual workplace features.

### 3.4.1.2 SHAP Local Interpretability (Force Plot)

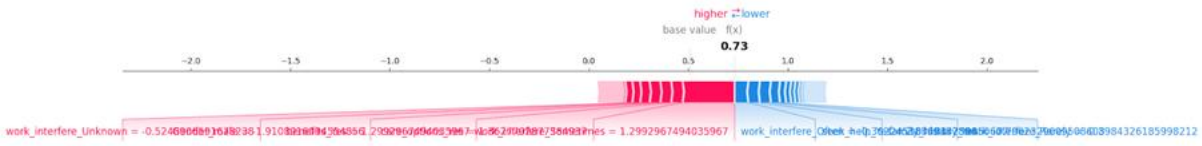



*Figure 13: Red features push the prediction toward "Treatment = Yes", while blue features push toward "No Treatment". The length of each bar represents the magnitude of that feature's contribution.*

The SHAP force plot explains how the model arrived at a single prediction by showing how each feature pushes the output above or below the model's baseline probability.

In this example:

- **Positive contributors** (in red), such as reporting that work sometimes interferes with mental health, having employer benefits, or available care options, push the prediction toward *Treatment = Yes*.
- **Negative contributors** (in blue), such as lower work interference, lack of family history, or limited help-seeking behavior, pull the prediction toward *No Treatment*.

The final prediction lands above the baseline value, meaning the model determines that this individual is likely to seek treatment.

Local interpretability confirms that the model's decisions for individual cases reflect meaningful and context-appropriate patterns rather than random or biased influences.

### 3.4.2 Interpretability Summary

Across global and local SHAP explanations, the model consistently highlights:

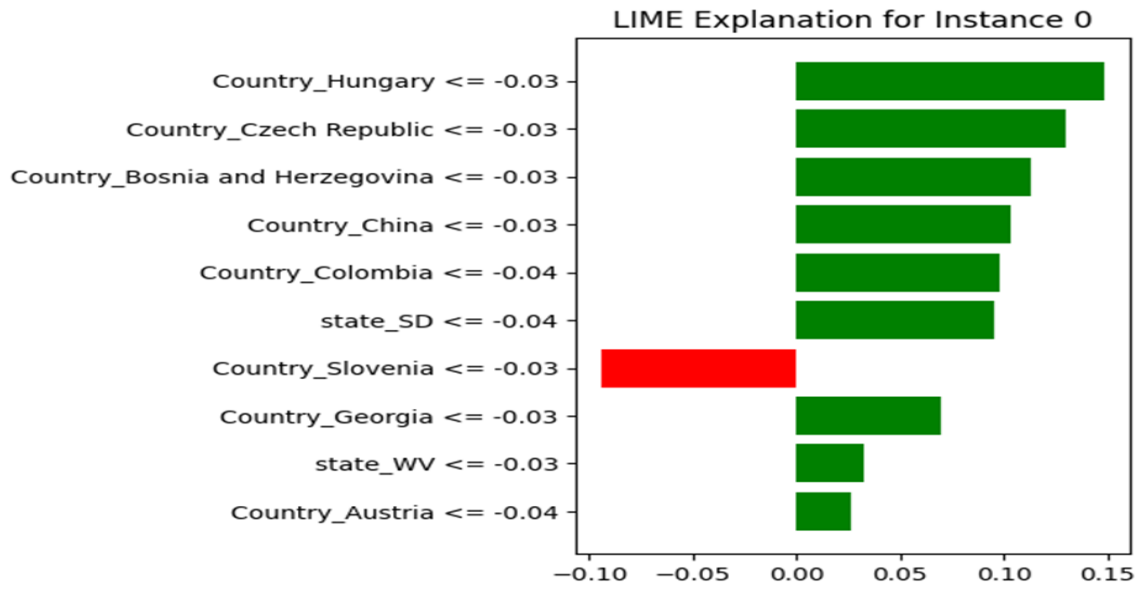
- Work interference as the strongest driver of treatment-seeking.
- Family history, benefits, care options, and anonymity as major supporting factors.
- Behavioral hesitation (e.g., seek_help_No) as a barrier to help-seeking.
- Demographic attributes as minor contributors, reducing fairness concerns.

These findings align with EDA patterns and real-world expectations in workplace mental-health behavior, confirming that the RidgeCV model is both interpretable and ethically grounded.

### 3.4.3 LIME Local Interpretability

LIME (Local Interpretable Model-Agnostic Explanations) provides case-specific interpretability by approximating the model's behavior around a single prediction. While SHAP explains overall model logic, LIME helps clarify why the RidgeCV model produced a particular output for one individual in the test set.

#### 3.4.3.1 LIME Explanation Plot




In the plot:

- Green bars push the prediction toward Treatment = Yes.
- Red bars push it toward No Treatment.
- Bar length shows the strength of each feature's influence.

For this instance, the most influential factors were country and state dummy variables. This occurred because the individual had few active behavioral or workplace indicators (such as work

interference or family history), so LIME relied on the available one-hot encoded features to approximate the local decision boundary.

#### 3.4.3.2 LIME Interpretation Summary

- Several countries (e.g., Hungary, the Czech Republic, Bosnia and Herzegovina) pushed the prediction upward.
- Country_Slovenia was the strongest downward influence.
- These effects are local only and reflect the sparse active features for this individual

#### 3.4.3.3 Consistency with SHAP

LIME's explanation does not contradict the SHAP findings:

- SHAP identifies the true global drivers of treatment-seeking (work interference, family history, benefits, care options).
- LIME explains a single prediction, using whichever features were active for that specific respondent.

Together, SHAP and LIME confirm that the RidgeCV model behaves consistently and transparently at both global and local levels.

### 3.4.4 Interpretability Conclusion

SHAP and LIME together confirm that the model behaves transparently and relies on meaningful factors rather than sensitive demographics.

SHAP (global) shows that treatment-seeking is driven mainly by work interference, family history, and workplace support (benefits, care options, anonymity). Demographic features such as age, gender, and region have only a minor influence, supporting the fairness findings.

LIME (local) explains a single prediction. For the chosen instance, few behavioral features were active, so LIME relied on available country/state dummy variables. This is normal in high-dimensional one-hot encoded data and does not contradict SHAP's global insights.
Together, SHAP and LIME show that the model is consistent, interpretable, and grounded in real behavioral and workplace patterns, with demographic features playing only a small role.

# 4 Conclusion

This project developed and explained a machine-learning model to predict mental-health treatment-seeking using the OSMI Mental Health in Tech dataset. After cleaning the data, exploring key patterns, training several models, and assessing fairness and interpretability, the study found clear behavioral and workplace factors influencing help-seeking.

RidgeCV was the strongest overall model, showing consistent performance and stable generalization. Across all models, work interference, family history, and workplace support (benefits and care options) were the most influential predictors, while demographic attributes such as age, gender, and region played only minor roles.

Fairness analysis showed some differences across groups, especially between males and females and across regions, but these reflected underlying data patterns rather than demographic reliance. Interpretability methods (SHAP and LIME) confirmed that the model's decisions were driven mainly by meaningful behavioral and workplace factors, not sensitive demographics.

Overall, the project demonstrates that machine learning, when paired with fairness checks and clear interpretability, can help highlight mental health risk patterns in workplace contexts. Because the topic is sensitive, such predictions should support, not replace, human judgement. Future work can explore richer datasets, debiasing methods, or more expressive models to improve fairness and reliability.

# 5 Ethical and Responsible AI Considerations

Predicting mental-health treatment-seeking involves sensitive personal information, so ethical, fairness, and privacy considerations are essential. The following principles guided the development and interpretation of this project.

## 5.1 Privacy and Data Sensitivity

The OSMI dataset contains highly sensitive mental-health and workplace information. Although anonymized, features such as family history, work interference, and personal perceptions require careful handling. All processing in this project used anonymized, encoded data with no identifiable information. Any real deployment of such a model would need to comply with data-protection regulations (e.g., GDPR, HIPAA, Uganda's Data Protection and Privacy Act, 2019) and implement secure storage, restricted access, and informed-consent procedures.

## 5.2 Fairness and Bias Monitoring

Because mental-health outcomes can differ across demographic groups, fairness analysis was performed across gender and region. While some performance differences were observed, demographic variables contributed minimally to predictions, reducing the risk of demographic bias. Nevertheless, fairness should be monitored continuously in practice, supported by methods such as re-weighting, threshold adjustments, or debiasing if disparities increase.

## 5.3 Risks of Misclassification

Incorrect predictions can have real implications.

- False negatives may delay help for individuals who need support.
- False positives may inappropriately flag low-risk individuals.

For this reason, model predictions should not be used in isolation. They should complement broader organizational wellness strategies and be interpreted by qualified mental-health or HR professionals.

### 5.3.1 Transparency Through Interpretability

To support accountability, SHAP and LIME were used to explain both global and local model behavior. These methods ensure that predictions can be understood and justified, reducing the risk of opaque or unexplained automated decisions, an essential requirement in mental-health contexts.

### 5.3.2 Avoiding Misuse and Stigmatization

Mental-health models can be misused for exclusion, monitoring, or unfair evaluation. This project positions the model strictly as a tool for support, not for performance assessment, job eligibility, or disciplinary decisions. Responsible use requires policies ensuring predictions are applied only to enhance mental-health support, early intervention, and benefits planning.

# 6 Limitations of the Study

This project provides useful insights, but several limitations should be acknowledged.

**1. Self-reported survey data.**
The OSMI dataset is based on self-reported responses, which may contain recall errors, social desirability bias, or inconsistencies in how individuals interpret questions. These issues can affect the reliability of the target labels used for model training.

**2. High-dimensional and sparse features.**
One-hot encoding, especially for country and state variables, produced a large and sparse feature space. This influenced model behavior and interpretability, particularly for LIME, which is sensitive to the distribution of active features around individual instances. Geographic imbalance, mainly the dominance of U.S. respondents, also limits broader generalizability.

**3. Linear model constraints.**
Although RidgeCV performed well, it remains a linear model and may miss complex non-linear relationships inherent in mental-health behavior. More expressive models could capture these patterns, but would require additional interpretability measures to remain transparent.

**4. Limited fairness scope.**
Fairness analysis was restricted to gender and region because these were the available demographic indicators. Other important fairness dimensions, such as ethnicity, socioeconomic status, or job role, were not included and may affect model reliability across different populations.

**5. Cross-sectional data only.**
The dataset captures a single point in time and does not reflect longitudinal changes in mental-health status or treatment-seeking behavior. As a result, the model cannot infer temporal trends or causal relationships.

# 7 Future Work

There are several ways this work can be extended and strengthened.

**1. Use richer and longitudinal data.**
Future studies could incorporate datasets that track mental-health indicators over time or include workplace productivity, engagement, or real-time well-being measures. This would support modelling of long-term risk patterns rather than single-point outcomes.

**2. Explore more expressive models.**
Non-linear models such as gradient boosting, random forests, or neural networks could capture deeper behavioral patterns. If used, they should be paired with advanced interpretability methods, such as DeepSHAP, Integrated Gradients, or counterfactual explanations, to maintain transparency.

**3. Broaden fairness evaluation.**
Fairness analysis could be expanded to additional protected attributes (e.g., ethnicity, job role, socioeconomic factors) where available. Techniques such as re-weighting, adversarial debiasing, or threshold adjustments could further reduce disparities in model performance.

**4. Improve local interpretability stability.**
LIME's sensitivity to high-dimensional one-hot encoded features could be addressed by grouping related variables, applying dimensionality reduction, or using more robust surrogate models that cluster similar features.

**5. Develop real-world decision-support tools.**
Future work could build prototype dashboards or APIs integrating human oversight, regular fairness monitoring, clear model communication, and strong privacy controls. Such systems would help organisations use insights responsibly to support employee mental health without causing unintended harm.