Leveraging Predictive Analytics to Overcome Treatment Challenges in Employee Mental Health Support

Md Shezan Ahmed

Date: 01 April 2025

Table of Contents

Executive Summary

The project focuses on identifying and assisting tech sector employees because they need mental health care support. This study utilizes anonymized workplace survey data from Kaggle so

the group could build a machine learning classification model for treatment-seeking behavior prediction among employees. The research pipeline contained several stages such as data cleaning normalization after which came exploratory data analysis and Chi-Squared along with decision tree importance metrics for feature selection and model development through decision trees KNN random forest and logistic regression.

The three key features in the analysis were work_interfere, family_history and care_options. A process of selecting features reduced the database to include only the ten key predictive elements. The researchers divided the collected data into three separate groups for training, validation and testing purposes. The model evaluation occurred through a combination of Accuracy, Precision, Recall and F1 Score measurements.

Model optimization using the Grid Search method occurred as part of the hyperparameter tuning process. Random Forest produced superior results compared to other models with regard to Recall measures and F1 Score performance while the Decision Tree model generated results that were both competitive and easily interpretable. Logistic Regression supplied an uncomplicated yet strong alternative whereas KNN presented both underfitting issues along with reduced generalization performance.

Business Problem

Healthcare professionals working with tech firms now confront rising difficulties when it comes to finding workers who need mental health care. Workers shy away from seeking mental health support because the nature of these matters remains sensitive and counts as a stigmatized activity. The accessibility challenges stop providers from properly distributing resources while

limiting their capability to send targeted outreach. Medical interventions provided too late may worsen health status and reduce employee performance.

Current Approaches to Deal with Employee Mental Health

Currently, many organizations rely on:

- Self-reported surveys or employee wellness check-ins.
- The HR interventions function exclusively through restricted feedback along with monitored actions.
- The delivery of assistance to employees happens solely after they report issues through reactive methods.

Such methods require manual work and produce subjective results while omitting employees who maintain secrets about their struggles. Privacy-related issues substantially diminish the performance of conventional treatment methods.

Role of Predictive Analytics

Predictive analytics solution based on historical survey data helps businesses determine patterns and risk variables that lead people to seek mental health treatment. The analysis of the workplace environment alongside mental health awareness, employee benefits, and past behavior enables the creation of a machine learning predictive model that forecasts individuals treatment-seeking likelihood.

The predictive analysis helps healthcare providers to

- The healthcare provider can send targeted communication to specific populations with high-risk characteristics.

- Proper distribution of counseling together with mental health assistance services would be possible through improved allocation methods.
- Stigma reduction will happen through implementing specific awareness programs that address public perception.
- Health-related operations within organizations and employee welfare show better results.

Analytic Problem

Analytic Problem Statement:

- Our goal is to create a classification model that determines the likelihood of employee sick mental health treatment based on survey answers. The model will predict employee treatment-seeking behavior.
- Target Variable: `treatment`
- The data type consists of two categories (binary: “Yes” or “No”) to indicate prediction outcomes.
- The main purpose is to deploy supervised machine learning models that enable accurate categorization of future respondents between Yes and No.

Data Availability

In this project, we used an anonymized workplace mental wellness survey obtained from tech sector staff members. Following is the link to a dataset:

https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey

Brief Description of Dataset

A mental health survey was designed to collect information from technology industry staff members. The dataset maintains 27 distinct fields that combine demographic statistics (age gender and country) with employment records (staff count and work-from-home capabilities along with managerial backing) as well as mental health metrics (like family background and medical background and work disturbances and leave provisions). Company personnel provided feedback regarding work environment approaches to mental health as well as physical well-being and their preparedness to participate in interviews and the observed effects of mental health problems. The data from the survey report enables researchers to find employee behavioral patterns and workplace elements that identify mental health treatment making choices. The provided dataset helps construct predictive models and recognize vital risk indicators to direct early mental health intervention programs that enhance workplace well-being in tech industry organizations.

Ethical Considerations

The project maintains ethical conduct through data anonymization that both safeguards personal privacy while upholding data protection legislation. The maintenance of anti-discriminatory practices depends on bias-free feature selection and routine fairness assessments for age groups, employment categories, as well as gender to guarantee equal, respectful treatment for all respondents. The research maintained clear transparency through simple explainable models and precise explanation of results. All findings served the purpose of fostering mental health understanding and creating safe comfortable areas that protected people instead of serving as identification or punitive tools.

Data Quality Resolution

The following is an overview of the data preprocessing stages together with fundamental discovery outcomes.

Library Import & Drive Mounting

The necessary Python libraries pandas, NumPy, matplotlib, seaborn, and warnings joined the program after Google Drive successfully mounted for accessing the Google Colab dataset.

Dataset Loading & Overview

A total of 1,259 survey records, including 27 fields for demographic and employment and workplace mental health details, were successfully captured during the loading process.

Data Structure Check

Most of the columns within the dataset contain categorical values (`object`) except for the `Age` column that holds numerical information. The review of initial data showed that all formatting requirements and structural rules were accepted.

Missing Value Analysis

- state has approximately 41% missing values.
- work_interfere has around 21% missing.
- self_employed is missing about 1.4%.
- The most extensive missing data appears in the comments variable which contains 87% gaps.
- All other columns are complete.

Insights & Decisions

- We will eliminate both state and comments from analysis because their unusually high proportion of missing values.
- The variable work_interfere contains important information which demands imputation or transformation.

The analysis required a segmentation of features which resulted in categorical types and numerical types.

Dropping Column not Directly related to Analysis

The analysis part of the research benefited from removing the comments and Timestamp columns because these fields did not directly support the project goals. The column named ‘comments’ contains unstructured free type of text datapoints, and multiple entries are missing,

making direct analysis impossible. The survey submission timestamp, called Timestamp, provides no predictive power or analytical benefits to the study. The processing phase makes the dataset appropriate for modeling by removing unnecessary columns.

Removing Duplicate Raw

During the data cleaning process, 4 rows has been removed from the dataset as those were duplicate. If it stayed in data then it will not predict the desired results and might create unnecessary datapoints for the model training as well.

Categorical Columns Identification:

A complete review of unique values took place for all object data type columns. The evaluation process revealed which variables contained discrete categories instead of continuous quantities.

Findings:

- A total of 24 categorical columns were found throughout the analysis.
- The high cardinality levels in Gender (49 unique values), Country (48) and state (45) warrant either grouping or cleaning procedures.
- Most features, including self_employed , treatment`, and remote_work contain between 2 to 5 unique values which make them good candidates for encoding.

Numerical Columns Identification:

The analysis classed every column except objects as numerical.

Findings:

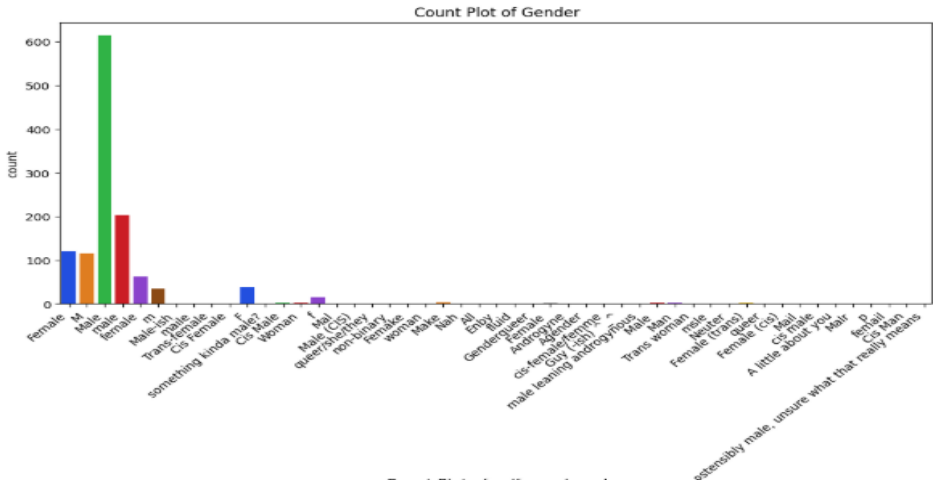
- The sole numerical feature Age among all features contained 53 unique factor values.

The dataset primarily consists of categorical data, according to this analysis, so we should use proper encoding methods for future stages of processing.

EDA (Bird Eye View of Data)

Gender Distribution

Figure 1

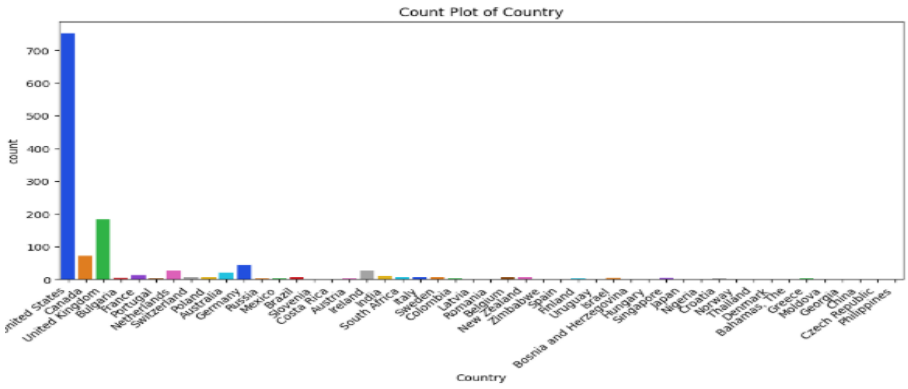


Most participants identify as male yet the distribution follows a long-tailed pattern where female representation comes second and other diverse gender identities form the back end of the distribution. An extreme distribution pattern indicates that standardization of analysis categories must happen for better understanding.

This distribution shows the requirement of standardizing gender categories. It will help create consistent information and eliminate any model imbalances that might occur due to gender disparity.

Country Distribution

Figure 2

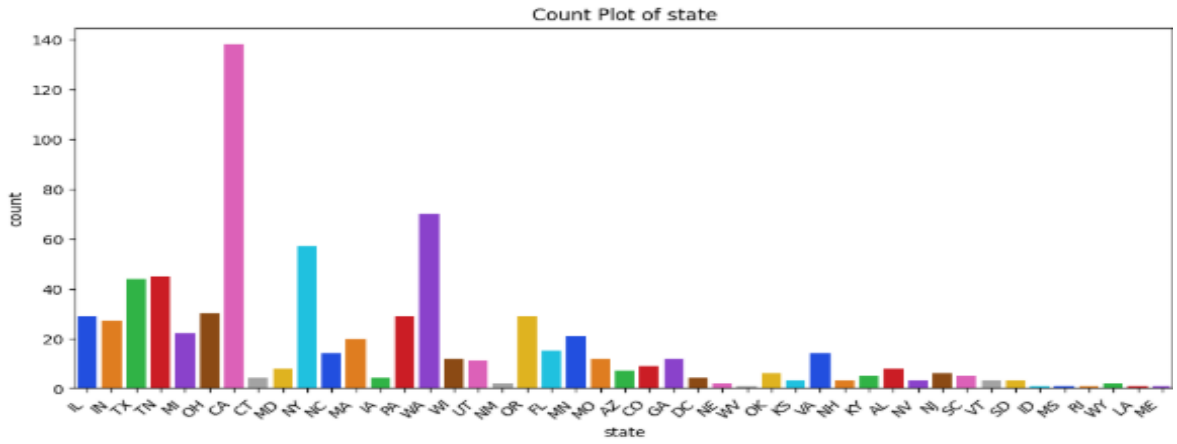


Seven hundred plus responses come from people located in the United States so we have severe sample distribution bias towards this country. A large number of participants in the United Kingdom and Canada together with significant numbers from Germany form the basis of medium-size reaction groups while minority responses exist across most other nations.

This analysis focuses on primary regions through categorization of minor countries into an "Other" category because their responses are not significant.

State Distribution (US States)

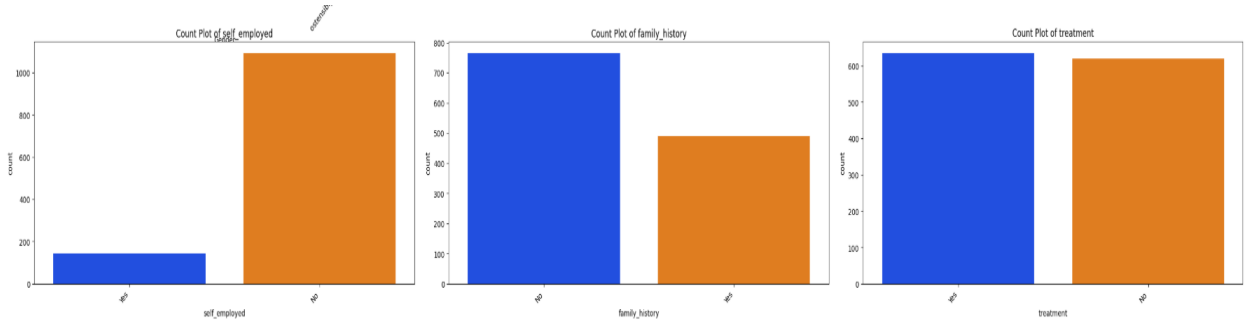
Figure 3



A disproportionate number of respondents came from American states throughout the United States. The highest number of responses comes from California and Washington, followed by a few responses from numerous other states.

We must organize states into regional groupings such as the West, South, and Northeast to improve the analysis, or drop minor representation states since they affect the results.

Figure 4

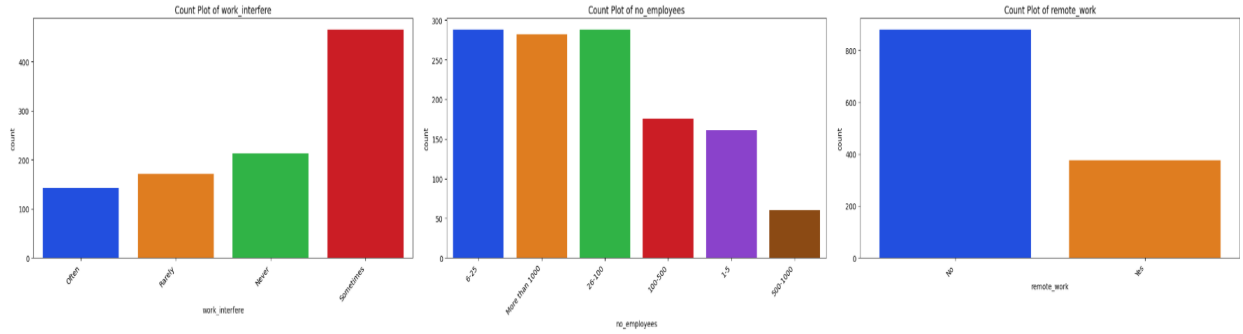


The self-employed feature displayed strong class imbalance because respondents mainly indicated non-self-employment status. By analyzing the count plots, the data reveals that the main employees in the dataset maintain employment through standard work practices.

Most participants during the study disclosed they had relatives who dealt with mental illnesses. The survey results support the belief that this particular population group faces increased danger of developing mental health issues.

According to the treatment variable, approximately equal numbers of respondents have received mental health treatment and have not received it. The equal distribution between classes proves most beneficial for machine learning systems because it results in stable classification while eliminating the necessity for deep resampling.

Figure 5



Work Interference

Most respondents indicated mental health problems create frequent disruptions in their job performance. The selection of "never" and "rarely" and "sometimes" options remained minimal. Data is positively skewed.

Work productivity analysis depends heavily on the prevalence of mental health challenges, which affect a significant percentage of participants, according to the data.

Number of Employees

Data for the no_employees feature shows an almost balanced distribution between organizations with between one and five hundred employees, but it includes decreased responses from businesses with more than five hundred employees. Data is negatively skewed.

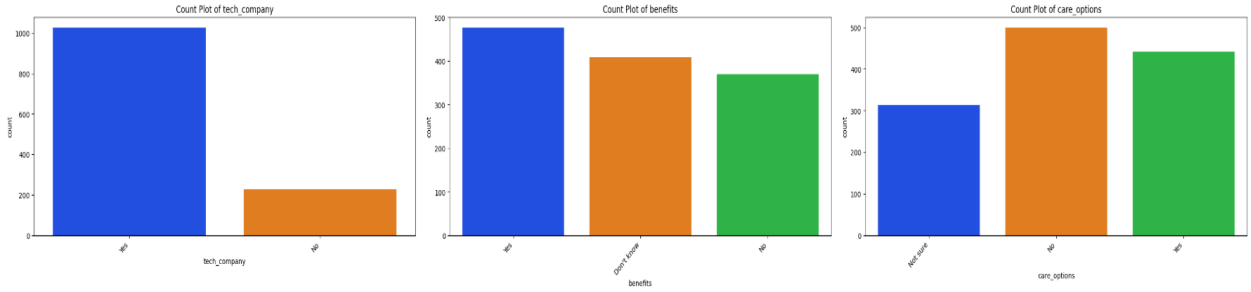
Our data mainly includes workers from smaller enterprises, causing potential limitations in extending conclusions to bigger corporations.

Remote Work

Research participants mainly stated their jobs do not include remote work, even though some individuals reported working remotely. Data is negatively skewed.

The in-person working environment presumably generates stronger mental health effects than other variables present within this dataset. Our analysis of treatment-seeking behavior can benefit from the remote work variable that examines workplace environment effects.

Figure 6



Tech Company

Most employees work in tech companies, while only a minority do not operate from such businesses, according to the distribution data.

The high number of tech professionals in the data set restricts how well the analysis can apply to different economic sectors.

Benefits

The dataset for benefits exhibits equal distribution between "Yes" responses and a slightly larger "No" category. Among the option categories, "Dont know" and "No" maintain almost equal frequency rates.

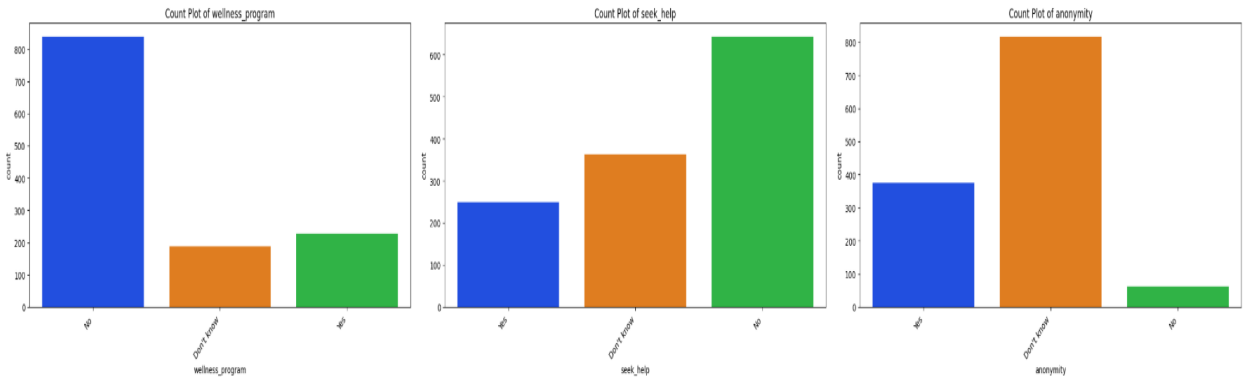
This proper distribution of survey responses demonstrates sufficient coverage about mental health benefits access, thus making it valuable for research purposes.

Care Options

The results from the `care_options` demonstrate a gentle left skew as users most often choose "Yes" then "No" and "Not sure."

The results show that multiple workers have mental health care options through their workplaces but remain unsure about these possibilities that could influence their health treatment behaviors.

Figure 7



Wellness Program

Data is heavily right skewed. Organizational wellness programs exist at a low rate since most respondents selected "No" for this question. The "Dont know" group, along with the "Yes" selection, contains corresponding low yet equivalent response counts.

Analysis indicates that a vast majority of workers neither participate in wellness initiatives nor know about them.

Seek Help

On the distribution of "Seek Help," there exists a balanced left-skew because participants chose "Yes" the most often. The most of survey working individuals answered either "Do not know" or "No" to the question, despite the other options having less responses.

The present data distribution demonstrates sufficient proportion between group classes, making the information usable for evaluation purposes. The variable indicates the extent of workplace receptiveness to mental health support initiatives.

Anonymity

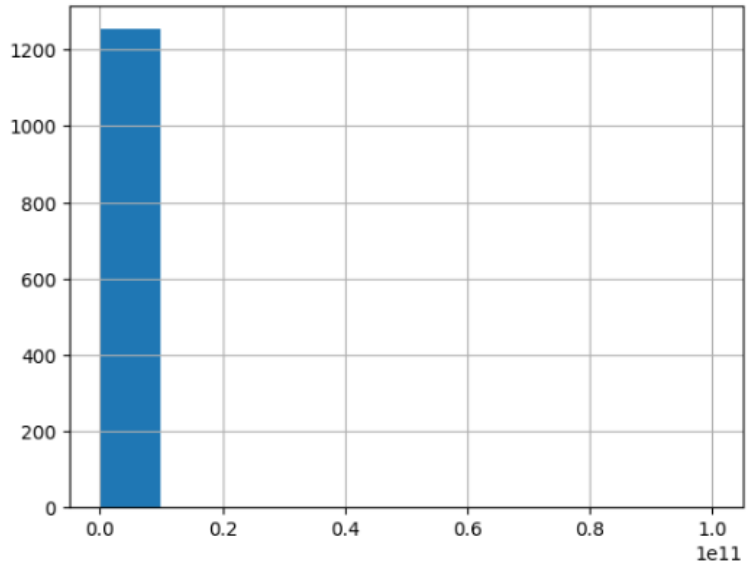
The responses related to the anonymity feature show a clear left-skewed distribution where "Dont know" selection exceeds others by a wide margin. The "Yes" responses are seen at a moderate times, while "No" responses show considerable rate.

Exploration and Cleaning of the `Age` Feature (Including Plotting Challenges)

The significant number of employees who do not know if they have access to anonymous assistance suggests that mental health disclosure behavior could be affected because of the uncertain support conditions.

We assessed the initial `Age` column containing 1259 numerical values. Our examination showed several incorrect and outlier values including the cases of `99999999999`, `-29`, `-1726` as well as young ages that did not fit a working demographic. Our attempt to create a histogram chart failed due to the extreme data points that produced an uninterpretable distribution by pushing most values toward the left section and extending the plot with a massive outlier trail. The interpretation of actual age distribution became impossible because of the data difficulties.

Figure 8

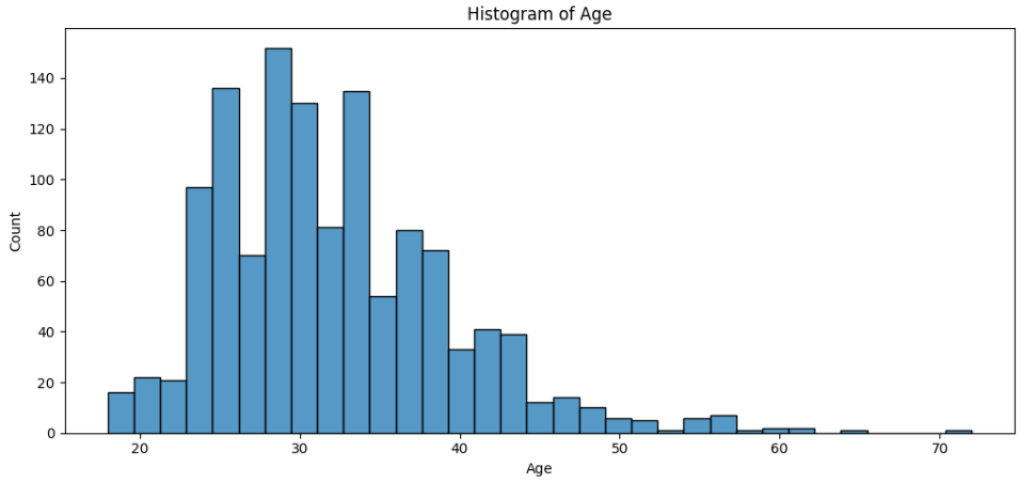


We solved the issue in the `Age` column by deleting three types of entries:

All entries showing age values at 16 or below were removed due to the assumption that these subjects do not belong to the working-age group. The analysis considered age values above 100 as outsider data and potential input errors while processing these records.

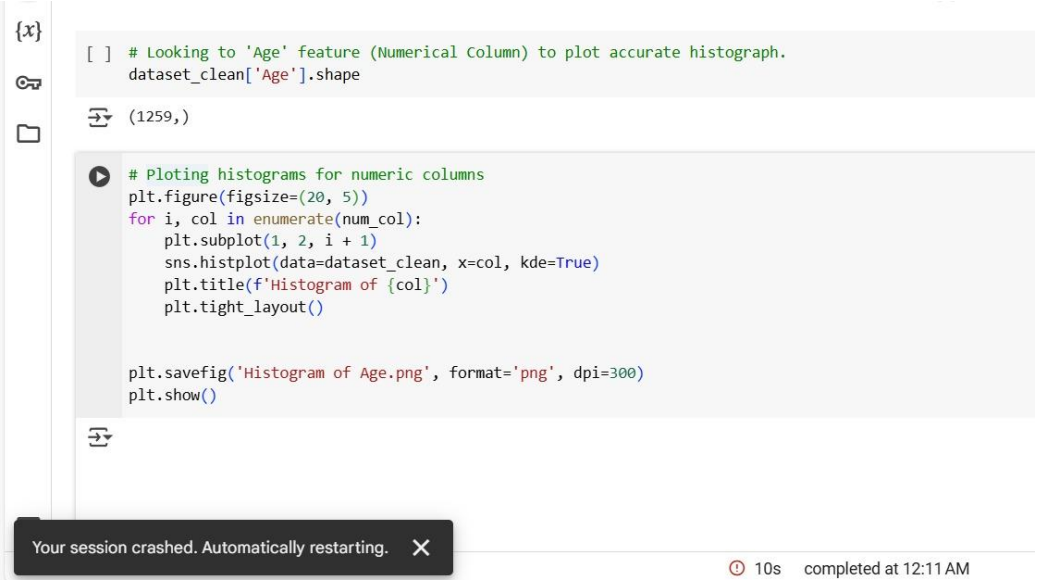
The result of the filtering process left 1247 valid ages between 18 to 72 years old. We managed to create a new histogram from the data which reflected the right-skewed pattern with the highest frequency occurring for people in the age range of 28 to 32. The adult working population characteristics appeared in the majority of survey participants who ranged from 25 to 40 years old.

Figure 9



A runtime problem occurred during plotting which caused the notebook to crash as shown in the provided screenshot. Memory overload or figure rendering conflicts are the likely causes for the program failure during figure save operations with large-sized graphics. The notebook session crashes were prevented in follow-up runs by adjusting the figure size and properly cleaning outlier values.

Figure 10



Handling Missing Values

The data quality required us to eliminate columns containing more than 30% missing value entries.

For columns with minor missing data:

- The missing values in the ‘work_interfere’ feature received the replacement value `"Do not know" during data cleaning.
- The values of ‘self_employed’ that were missing received the most typical input which was `"No" as a replacement.

Gender Column

The ‘Gender’ data passed through cleaning and standardization procedures. The Gender column exhibited diverse entries with inconsistent and misspelled and non-standard data values across the entries. The cleaning process combined both fragmentation reduction with improved clarity.

We converted every male-relevant term such as `M` and `male` and `cis male` to the category `Male`. A single group named `Female` included all female-related terms such as `F`, `woman`, `cis-female/femme`, etc.

We assigned all unique gender identities which include `non-binary`, `queer`, `fluid`, and similar terms into the category named `Other`.

Resulting Distribution:

- Male: 982

- Female: 247

- Other: 18

Figure 11

```
[ ]  # Replacing the invalid values to it's correct form
     dataset_clean['Gender'].replace(['M', 'Male ', 'male', 'm', 'Male', 'Cis Male',
                          'Man', 'cis male', 'Mail', 'Male-ish', 'Male (CIS)',
                           'Cis Man', 'msle', 'Malr', 'Mal', 'maile', 'Make',], 'Male', inplace = True)

     dataset_clean['Gender'].replace(['Female ', 'female', 'F', 'f', 'Woman', 'Female',
                          'femail', 'Cis Female', 'cis-female/femme', 'Femake', 'Female (cis)',
                          'woman',], 'Female', inplace = True)

     dataset_clean['Gender'].replace(['Trans-female', 'something kinda male?',
            'queer/she/they', 'non-binary', 'Nah', 'Enby', 'fluid',
            'Genderqueer', 'Androgyne', 'Agender', 'Guy (-ish) ^_^',
            'male leaning androgynous', 'Trans woman', 'Neuter',
            'Female (trans)', 'queer', 'A little about you',
            'ostensibly male, unsure what that really means'], 'Other', inplace = True)
```

```
dataset_clean['Gender'].value_counts()
```

Gender	count
Male	982
Female	247
Other	18

dtype: int64

Simplifying the `Country` Feature

Response data was collected from more than 40 individual countries when the dataset first began. Model generalization and noise reduction become possible through these actions.

Only the three most common countries, which include the United States and Canada together with the United Kingdom, remained in the dataset.

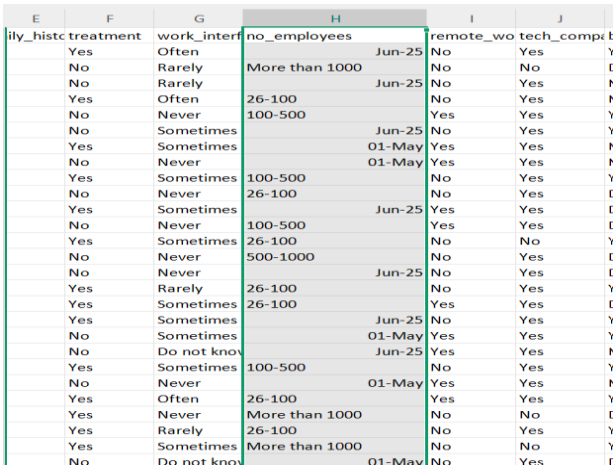
All countries except United States Canada and United Kingdom fell under the category `Other`.

The analysis becomes more reliable by concentrating on regions with enough available data in this preprocessing stage.

Issue with no_employees feature

During no_employees column analysis in the dataset, two distinct results emerged between the original CSV file data and the processed Google Colab format. The original CSV file displayed date type notations (01-May, Jun-25) throughout the no_employees column, which should contain expected categorical ranges ("1-5" or "26-100"). Visual assessment of unique values in the Colab programming environment confirmed that the recorded entries had no issues.

Figure 12



E	F	G	H	I	J
ily_histc	treatment	work_interf	no_employees	remote_wo	tech_compa
	Yes	Often	Jun-25	No	Yes
	No	Rarely	More than 1000	No	No
	No	Rarely	Jun-25	No	Yes
	Yes	Often	26-100	No	Yes
	No	Never	100-500	Yes	Yes
	No	Sometimes	Jun-25	No	Yes
	Yes	Sometimes	01-May	Yes	Yes
	No	Never	01-May	Yes	Yes
	Yes	Sometimes	100-500	No	Yes
	No	Never	26-100	No	Yes
	Yes	Sometimes	Jun-25	Yes	Yes
	No	Never	100-500	Yes	Yes
	Yes	Sometimes	26-100	No	No
	No	Never	500-1000	No	Yes
	No	Never	Jun-25	No	Yes
	Yes	Rarely	26-100	No	Yes
	Yes	Sometimes	26-100	Yes	Yes
	Yes	Sometimes	Jun-25	No	Yes
	No	Sometimes	01-May	Yes	Yes
	No	Do not knov	Jun-25	Yes	Yes
	Yes	Sometimes	100-500	No	Yes
	No	Never	01-May	Yes	Yes
	Yes	Often	26-100	Yes	Yes
	Yes	Never	More than 1000	No	No
	Yes	Rarely	26-100	No	Yes
	Yes	Sometimes	More than 1000	No	No
	No	Do not knov	01-May	No	Yes

```
[17] dataset_clean['no_employees'].unique()
0s
array(['6-25', 'More than 1000', '26-100', '100-500', '1-5', '500-1000'],
      dtype=object)
```

It could be because of Microsoft Excel automatically formatted the categorical strings into dates through incorrect interpretation of strings such as "1-5" or "6-25" as "May 1st" and "June 25th." The conversion of text to date format is an Excel default behavior that occurs when it recognizes contents resembling dates.

Manual data inspections become less accurate because formatting problems deteriorate data quality and result in wrong readouts when these issues remain unaddressed before the data processing stage begins.

Final Dataset Review

The cleaning process succeeded in resolving all missing values so that no column displays null values anymore.

A check was performed on all categorical attributes to verify their standardized processed formats.

The cleaned data contains 24 final dimensions that focus on population characteristics and work environment elements together with mental health assistance measures and employee perception metrics.

Key Features After Cleaning

Cleaned numerical column: `Age` (18–72)

The categories for Gender and Country as well as work_interfere and self_employed are standardized.

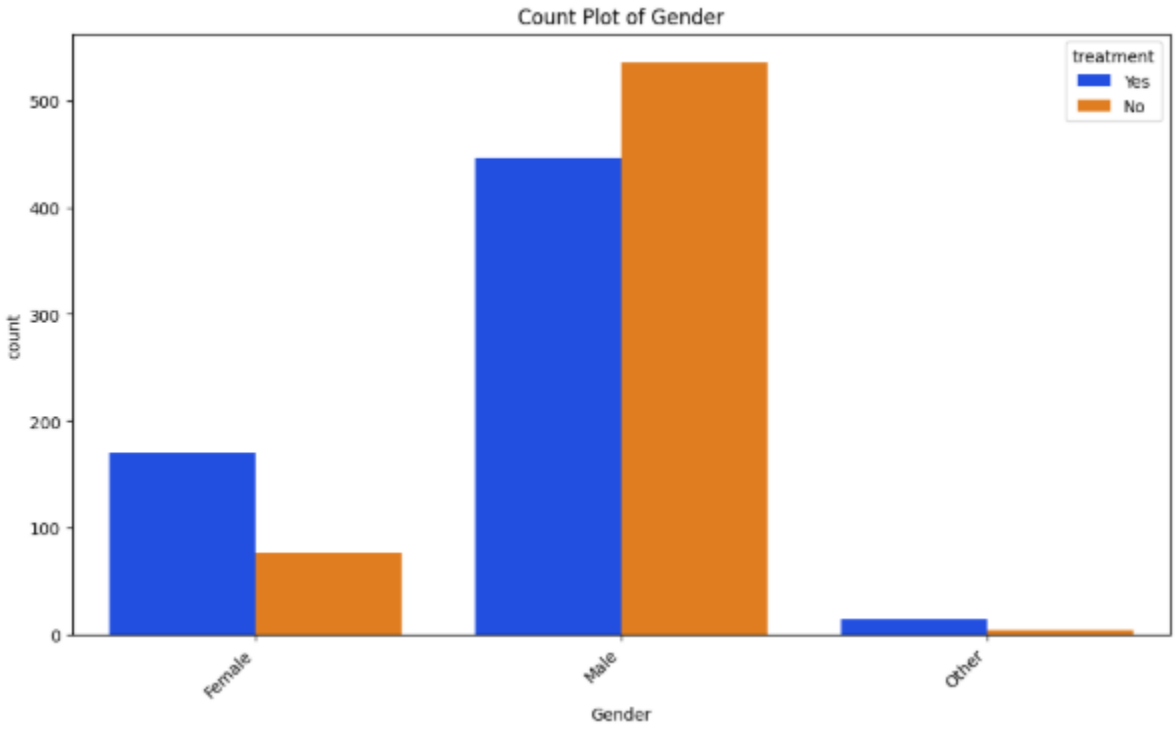
The categorical features have been converted into proper formats to enable encoding and modeling processes.

Normalization Validation Through Visuals

A count plot evaluation demonstrates Data Normalization success throughout the process. The normalization process for all categorical variables resulted in successful standardization which we validated through several count plots that we generated after the data cleaning. The created plots played a dual role to validate both the consistency and modeling potential of the processed data.

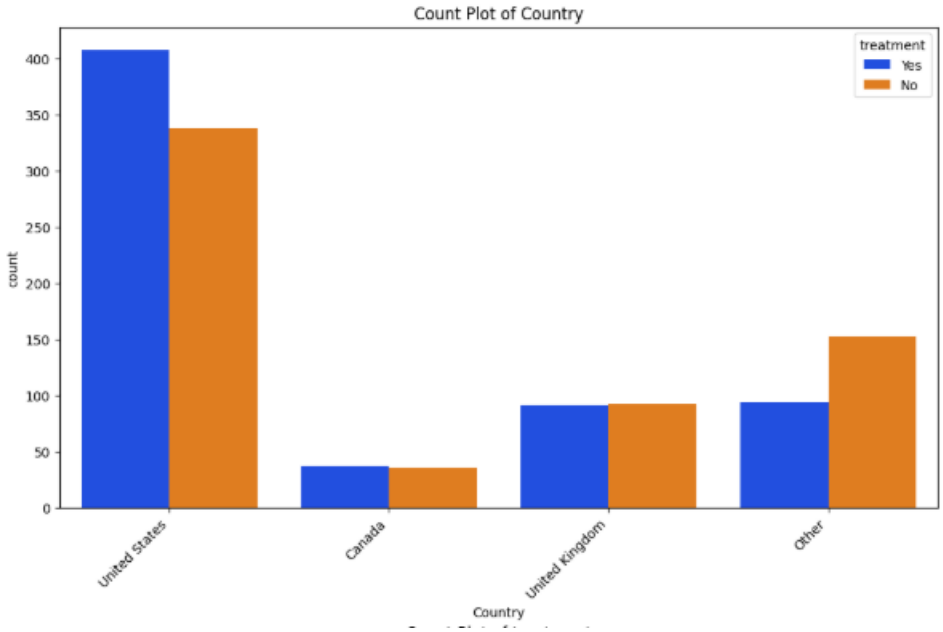
The normalization process combined inconsistent gender labels such as M, f, Cis Male into three options: Male, Female and Other. All standardized categories in the plot display distinct clusters while preserving equal distribution of treatment labels.

Figure 13



The original 40 country values received consolidation into United States, Canada, United Kingdom and Other which made the analysis process easier. The U.S. contains most participants within this dataset yet the analytical groups maintain equal balance between their assigned labels.

Figure 14



The researchers transformed unspecified data points (NaNs) in self_employed, work_interfere and leave by substituting them with either the most common response or “Do not know”. New category labels appear in the standardized format with no missing nor ambiguous information present.

Figure 15

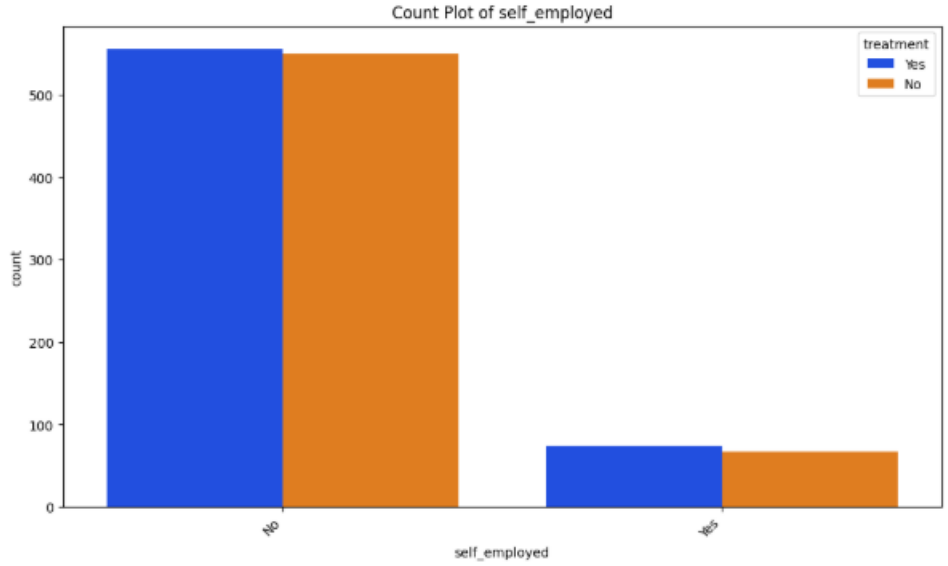
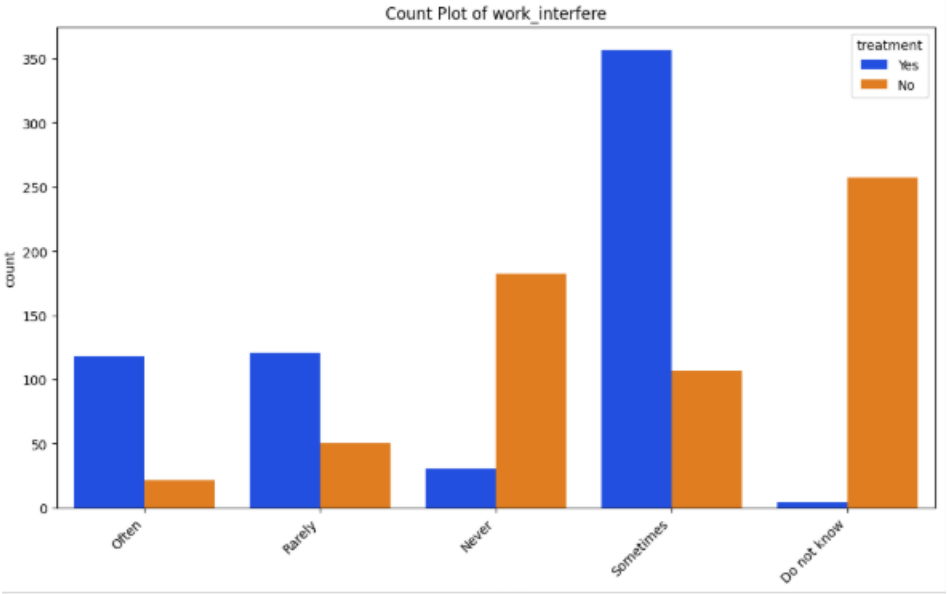


Figure 16



The normalization process resulted in features displaying neat label sets such as Yes, No, Dont know using standardized formatting throughout all options in benefits care_options wellness_program and anonymity categories.

Figure 17

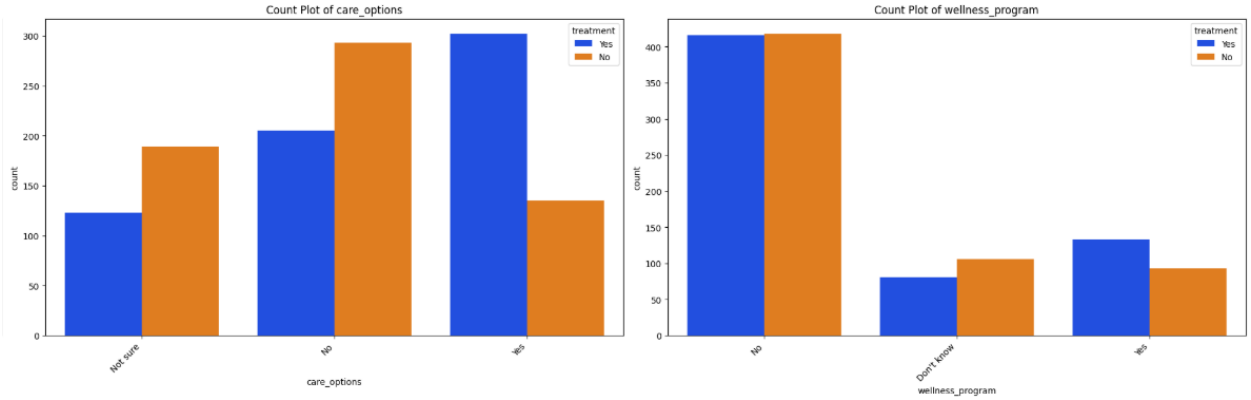
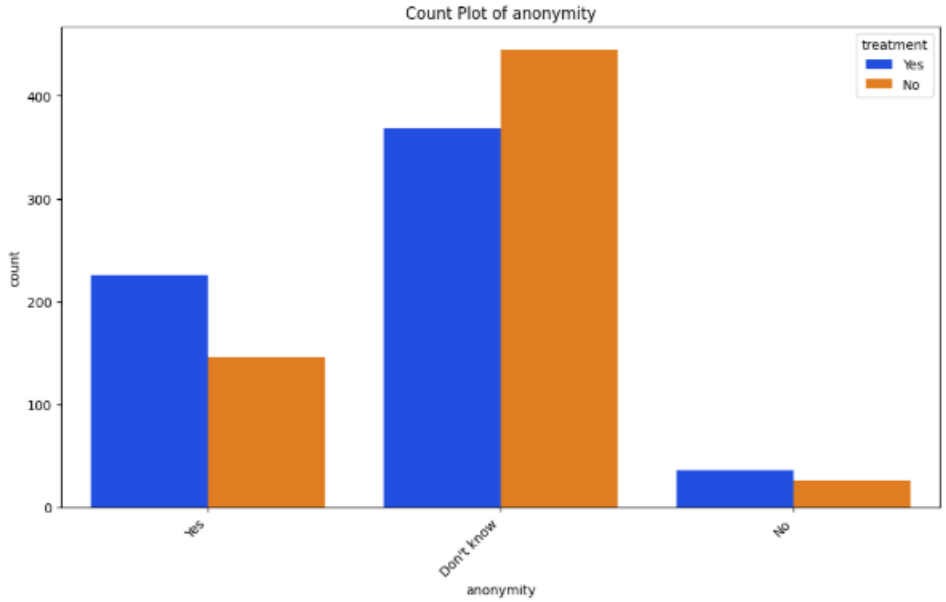


Figure 18



The treatment label appears equally distributed across every plot so viewers can easily contrast the different groups. The consistent format eliminates previous entry errors from continuing through the data set.

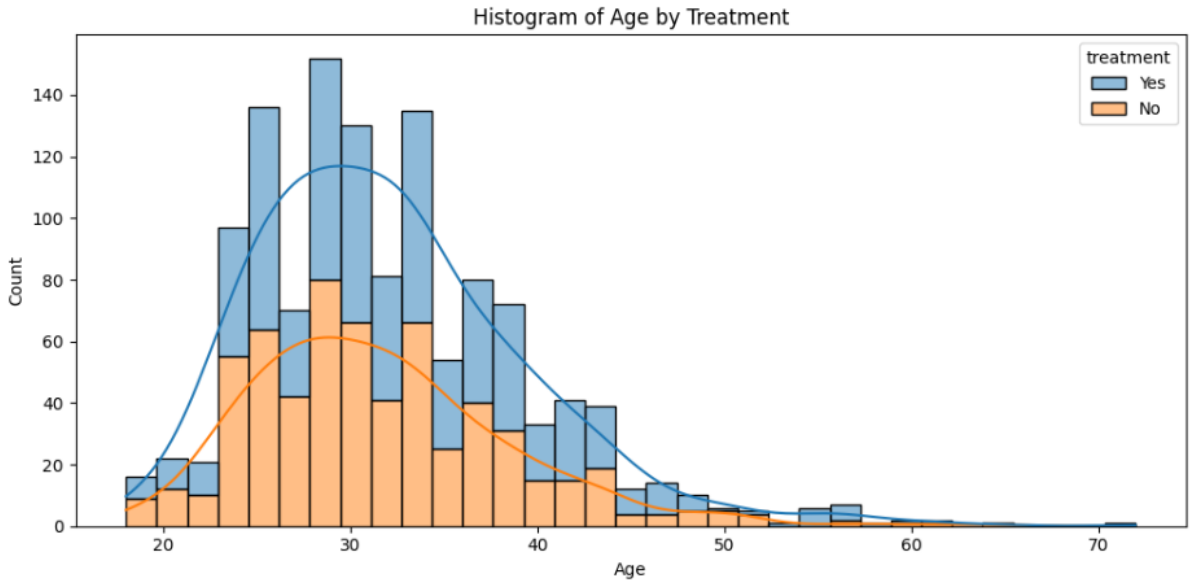
Histogram of Age by Treatment

In order to confirm distribution of age feature and study the link between patient age and their treatment behavior, we generated a histogram of the Age feature where the treatment status colors each bar segment using the hue parameter.

Key findings

The majority of participants belonged to the age group between 25 to 40 and participants reached their peak density at age 30. People from late 20s into early 30s demonstrated increased interest in seeking medical attention. The population size decreased for individuals older than 40 while the difference between Yes/No treatment selection became smaller.

Figure 19



There are 1,247 complete records within the cleaned dataset which contains no missing values throughout its 24 columns. The dataset contains one numerical feature known as Age together with 23 categorical features which have received proper formatting and normalization.

The treatment variable demonstrates good balance between groups since 630 patients reported seeking care but 617 patients did not seek treatment thus creating an even playing field for machine learning model development.

Encoding The Dataset

All 23 categorical features in the dataset achieved successful encoding through label encoding so they transformed into numeric data. The conversion makes numerical values available for machine learning model application. Mapping occurred to generate distinctive integers for each category thus making all columns strictly numerical. The cleaned dataset complies with all essential standards for proceeding with modeling procedures.

Feature Selection

Chi-Squared Test & p-Value Analysis

All categories except age underwent a Chi-Squared test for determining predictive strength in relation to treatment outcome.

Key Findings

Main features

- The main influential attributes for analysis (display high Chi-Squared statistics along with minimal p-Values)
- Main mental health treatment predictors include these features which demonstrate the most direct relationship with therapy needs:
- `work_interfere` ($Chi^2 \approx 525$, $p \approx 0.0000000000$)
- `family_history` ($Chi^2 \approx 106.6$, $p \approx 0.0000000000$)
- Results indicate that four features - `care_options`, `benefits`, `anonymity`, `obs_consequence` show both high significance levels along with minimal p-values.

Moderately Significant Features

- The moderate Chi-Squared scores paired with p-values below 0.05 level indicate these variables may provide meaningful predictive ability: `leave`, `Gender`, `seek_help`, `Country`, and `mental_vs_physical`.

Low Significance Features (High p-Values)

- `self_employed`
- `tech_company`

- `mental_health_consequence`
- `phys_health_consequence`
- The p-values exceeded 0.05 which proves that these variables do not statistically correlate with the target.

Figure 20

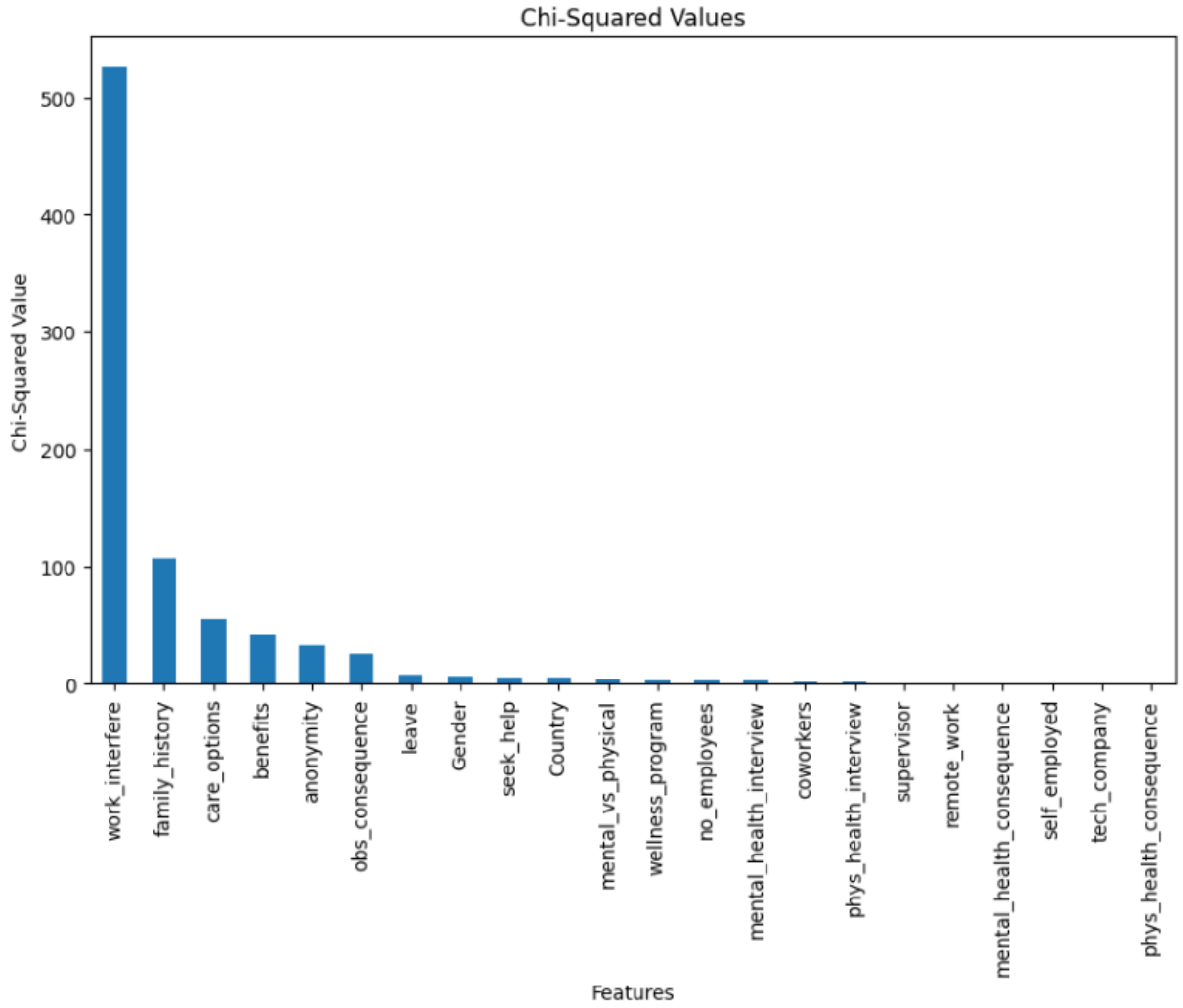
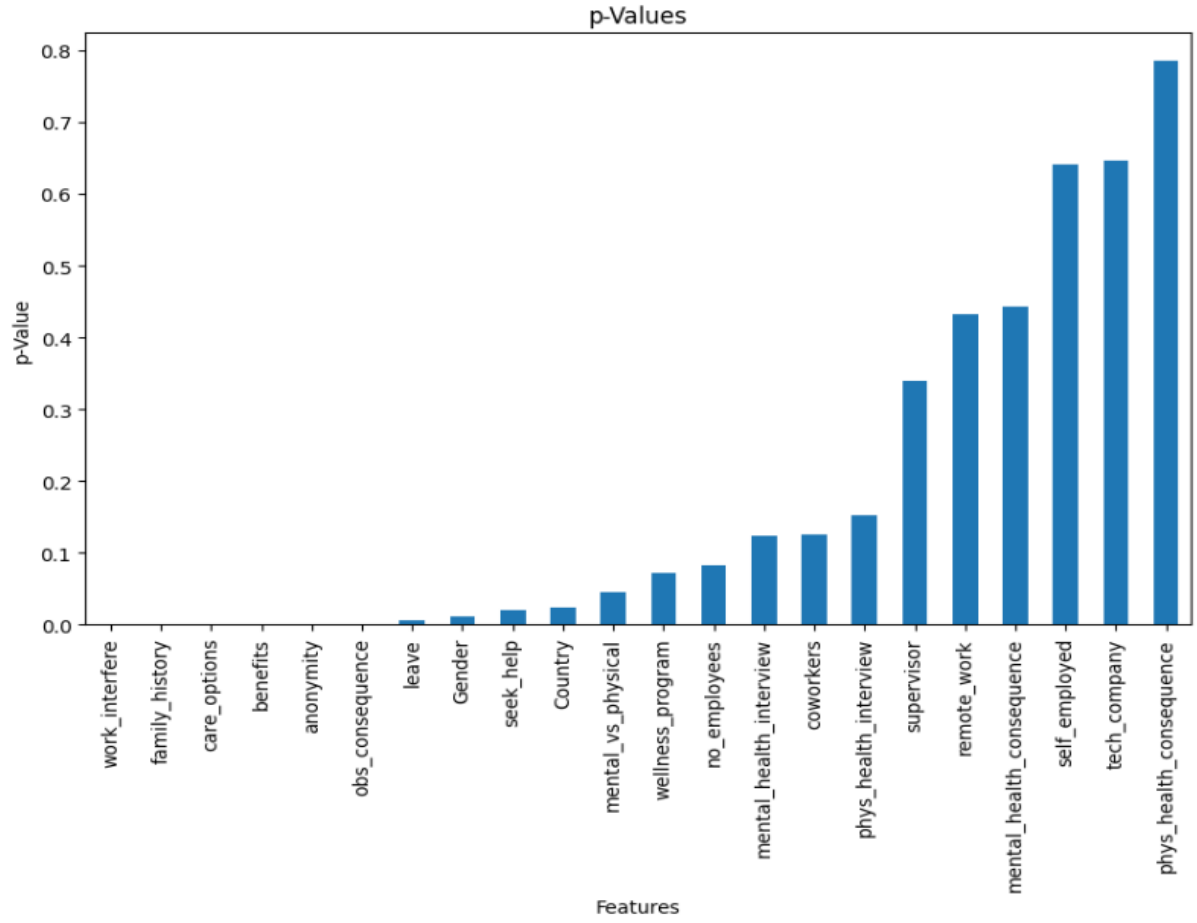


Figure 21

	Chi-Squared Value	p-Value
work_interfere	525.101652	0.0000000000
family_history	106.682849	0.0000000000
care_options	55.106939	0.0000000000
benefits	42.176490	0.0000000001
anonymity	32.402691	0.0000000125
obs_consequence	25.785084	0.0000003816
leave	7.508980	0.0061392119
Gender	6.384193	0.0115141069
seek_help	5.415036	0.0199640408
Country	5.077524	0.0242379645
mental_vs_physical	3.978582	0.0460823462
wellness_program	3.244762	0.0716520955
no_employees	3.005862	0.0829638419
mental_health_interview	2.377829	0.1230684681
coworkers	2.349934	0.1252883458
phys_health_interview	2.044354	0.1527718507
supervisor	0.908614	0.3404823477
remote_work	0.617283	0.4320588050
mental_health_consequence	0.589400	0.4426512684
self_employed	0.216915	0.6414006150
tech_company	0.210151	0.6466493885
phys_health_consequence	0.074553	0.7848198621

Figure 22



Decision Tree Model Summary

The developers created the Decision Tree classifier specifically to analyze individual mental health treatment seeking behavior. The researchers distributed their dataset into three sections for training, validation along with testing purposes. The training with an initial model that featured a maximum depth of 3 performed very well on recall and F1 scores across all datasets particularly in validation metrics.

Our performance booster function came from using `GridSearchCV` to find the best possible hyperparameter combinations.

- `criterion=gini`, `max_depth=5` , `min_samples_leaf=4` , `splitter=random`
- The tuned model achieved:

  Training Accuracy is 85.4% , Validation Accuracy is 80.2% , Test Accuracy is 78.2% and Test F1 Score is 79.6%

The model demonstrates effective generalization capabilities in accurately forecasting patterns of patients who seek treatment.

Decision Tree Feature Selection

During the Decision Tree model analysis the system determined feature importance values to show which features most significantly affected decision processes. Through this assessment we discover major influential variables and find candidate features suitable for eliminating from the model.

The most important features according to importance values consist of these ten components. The predictive power of the model mainly comes from its top 10 features with `work_interfere` being responsible for nearly 49% of predictive decisions. The influence of

workplace mental health on employees proves to be the main factor determining their decision to pursue treatment.

Features which have lower ranked (importance < 0.02) like:

- `Gender` (0.0049)
- `self_employed` (0.0095)
- `mental_health_interview` (0.0043)
- `tech_company` (0.0142)

The model shows good generalization abilities through its ability to correctly predict patterns of patients who need medical attention.

Decision tree importance feature

Decision Trees calculated the most important features through feature importances computation. Among all computed variables `work_interfere` stands out with an extremely high importance rating at 0.49. The values decrease significantly after the first place indicating a strong dominant characteristic. The set of attributes includes `leave` and `no_employees` and `Country` among others.

Figure 23

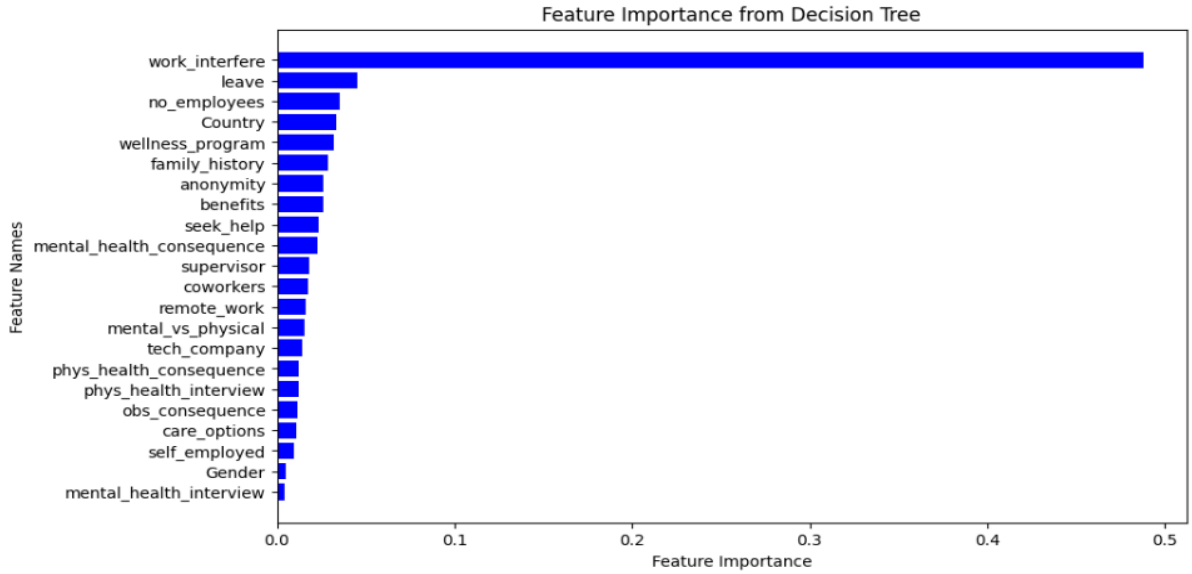


Table 1

	Feature	Importance	Chi-Squared Value	p-Value
0	work_interfere	0.487711	25.785084	0.0000003816
1	leave	0.045533	2.349934	0.1252883458
2	no_employees	0.035484	7.508980	0.0061392119
3	Country	0.033139	55.106939	0.0000000000
4	wellness_program	0.032285	3.244762	0.0716520955
5	family_history	0.028636	32.402691	0.0000000125
6	anonymity	0.026248	2.377829	0.1230684681
7	benefits	0.026041	5.077524	0.0242379645
8	seek_help	0.023186	3.005862	0.0829638419
9	mental_health_consequence	0.022962	2.044354	0.1527718507
10	supervisor	0.018286	0.589400	0.4426512684
11	coworkers	0.017392	0.617283	0.4320588050
12	remote_work	0.016437	6.384193	0.0115141069
13	mental_vs_physical	0.015459	0.074553	0.7848198621
14	tech_company	0.014197	5.415036	0.0199640408
15	phys_health_consequence	0.012384	0.908614	0.3404823477
16	phys_health_interview	0.012276	0.210151	0.6466493885
17	care_options	0.010959	3.978582	0.0460823462
18	self_employed	0.009533	42.176490	0.0000000001
19	Gender	0.004853	106.682849	0.0000000000
20	mental_health_interview	0.004321	0.216915	0.6414006150

Final Features Selection

Mixed top features from both methods.

The set included both domain-related features `Age` and `treatment` alongside those obtained from the initial analysis regardless of their initial absence (because these variables may prove vital for the predicted model).

Selected Features

`work_interfere`

A persons work performance interference due to mental health status is measured on a scale from never to always.

Statistically speaking this variable holds the greatest importance as per Decision Tree analysis and puts forward the most significant ChiSquared test pvalue. The data indicates that mental health treatment seeks or receives a powerful influence in this relationship.

`leave`

The response addresses how comfortable employees feel about taking mental health leave according to their perceived level of difficulty which ranges from very easy to very difficult.

Why selected: Moderate importance in the DT and borderline statistical significance. The organizational support culture together with its cultural dimensions affects the decisions regarding treatment.

`Country`

The country where employees conduct their work serves as the variable.

Why selected: Very high ChiSquared value and importance in the DT. Every country possesses its own unique model of mental healthcare delivery because service availability and public perceptions about mental health and stigma levels substantially differ between nations.

`family_history`

The variable follows whether individuals have experienced mental illness in their direct blood relatives.

Why selected: Statistically very significant and relevant from a health risk perspective. A known family history of mental illness affects personal understanding together with treatment-seeking behavior.

`no_employees`

The variable measures the personnel size for the organization in which the respondent works.

Why selected: Organizational structure as well as resources captures from it. The size of the company affects both availability and quality of mental health programs.

`benefits`

Mental health benefits provided or not provided by employers is the definition of this variable.

The choice of this variable was made because it straightly affects how much treatment is accessible. Benefit-receiving employees tend to search for help and get needed medical attention.

`self_employed`

One of the variables describes whether the person maintains their own business.

A very elevated ChiSquared value occurs with this variable although the Decision Tree importance score remains low. The lack of workplace benefits from employers prevents selfemployed people from getting needed support which impacts their treatment conduct.

`Gender`

The variable captures the gender identity demonstrated by the survey participant.

Why selected: A noticeable high Chi-Squared score. Gender determination plays a key role in defining mental health stigma responses when seeking professional help and shaping how symptoms become visible or registered.

`Age`

The variable measures the age of each study participant.

Age is important both demographically and psychologically so it was included although absent from the DT/Chi² importance tables. People at different ages experience diverse levels of awareness in mental health topics and display different comfort levels with discussion plus their resources to access support differ.

`mental_health_consequence`

The statement pertains to workers beliefs regarding potential detrimental workplace effects from discussing their mental health concerns.

The reason for this choice: It demonstrates the presence of stigma as experienced by working people in their professional environment. The question holds a moderate value while playing an essential role in psychiatric decision processes.

`treatment` (Target variable)

The variable evaluates if someone has sought medical attention because of a mental health issue or if they have received such care.

The target variable selected for prediction constitutes the reason for inclusion. The classification task shows the target outcome which stands as the main focus of the prediction results.

The selection of these features used a combination approach that involved statistical relevance (Chi^2) and modelbased importance (DT) and alignment with domain knowledge about workplace mental health. The selected variables create a solid base for predictive modeling by achieving high interpretability along with strong predictive performance. Feature selection strategies become more reliable through their integration of statistical test (Chi^2) and modelbased importance (DT).

Decision Tree Model After Features Selection

The selected top 10 features from the feature selection step enabled the creation of an advanced model. The data subsets were arranged into training (70%, 872 samples), validation (15%, 187), and test (15%, 188) portions to conduct fair evaluation.

Initial Model Performance

The basic Decision Tree model attained training results by reaching maximum depth of 3 with the chosen features selected through feature selection. It demonstrated:

Table 2

Score	Train	Validation
Accuracy	0.83	0.83
Precision	0.77	0.8
Recall	0.94	0.94
F1	0.85	0.86

The analysis revealed an outstanding capacity for the model to detect the treatment seeking behaviour of patient.

Hyperparameter Tunin

The model optimization process required executing a grid search procedure. The tuned parameters yielded the optimal results when set to Gini criterion and Max Depth = 5 and Splitter = Random and Minimum Samples per Leaf = 4 and minimum sample spit 2.

The tuned model achieved following performance:

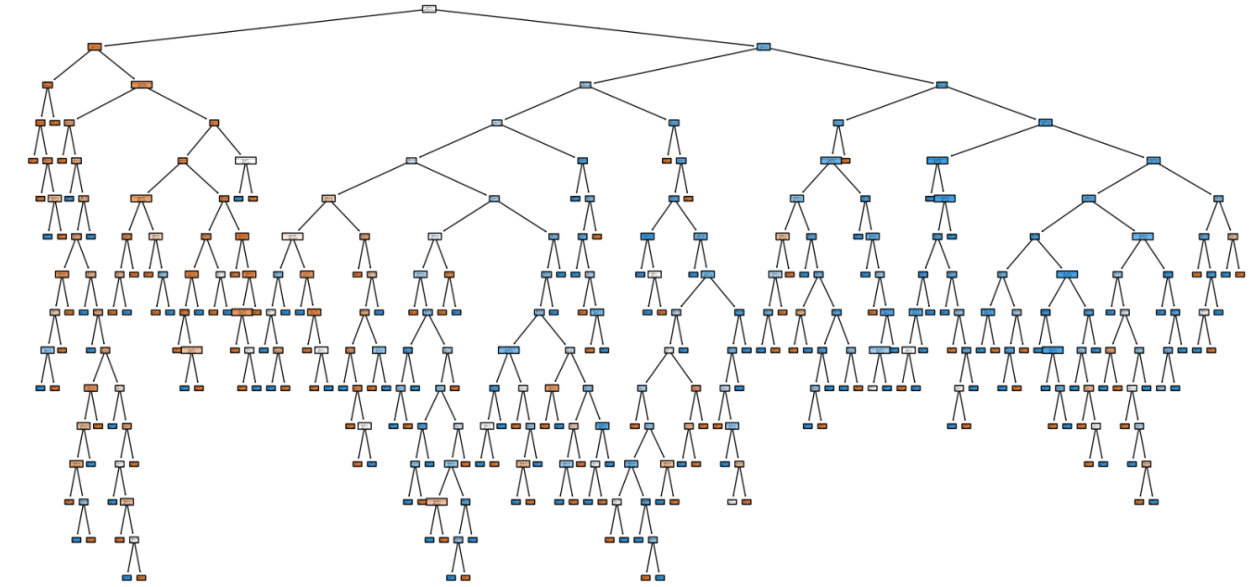
Table 3

Score	Train	Validation	Test
Accuracy	0.85	0.8	0.79
Precision	0.85	0.85	0.78
Recall	0.84	0.77	0.81
F1	0.85	0.81	0.79

The optimized Decision Tree model shows unified performance metrics between all sections of data including training data and validation data, and test data. The model demonstrates excellent generalization characteristics since its accuracy levels reach 0.85 on the training data and 0.80 on validation data, and 0.79 on test data while exhibiting low overfitting. The model maintains a precision at level 0.85 for training and validation, which slightly declines to 0.78 for the test set, showing effective minimization of wrong positive findings. The recall results indicate strong outcome detection abilities with 0.84 on training data and 0.77 on validation data, moving to better performance on the test data with a score of 0.81. The F1 score maintains a balanced performance with a value of 0.85 in training data and 0.81 in validation data, and 0.79 in test data, which demonstrates a proper weighting between precision and recall values. The model shows dependable behavior and maintains data resilience across various datasets which qualifies it for future analysis or deployment.

Decision Tree Visualization

Figure 24



A model based exclusively on the 10 most important features derived from Decision Tree importance produced efficient results. The method successfully detects crucial mental health treatment-seeking behavioral factors through its strong performance together with balanced generalization coupled with interpretable decision paths.

KNN Model

Data Scaling

Initially feature values received MinMaxScaler normalization to range them between 0 and 1 before model training since KNN needs this scale for distance calculations.

Initial KNN Model (k=7)

The training of a KNN model with seven neighbors as the parameter was completed. It performed reasonably well:

Table 4

KNN Model		
Score	Train	Validation
Accuracy	0.83	0.77
Precision	0.79	0.78
Recall	0.87	0.81
F1	0.83	0.8

Predictive performance of the model seemed adequate through balanced precision-retrieval figures yet it pointed to further improvements.

Hyperparameter Tuning with Grid Search

A complete grid search procedure determined the optimal set of KNN parameters. The following parameters were tuned: Number of neighbors (n_neighbors), Distance metric (p and metric), Weighting strategy (uniform or distance)

Best Parameters Identified

n_neighbors: 11 , p: 2 , weights: distance , metric: minkowski , Best CrossValidation Score: ~83.7%

Optimized KNN Model Performance

Table 5

KNN Model			
Score	Train	Validation	Test
Accuracy	0.99	0.77	0.73
Precision	0.99	0.79	0.71
Recall	0.99	0.79	0.77
F1	0.99	0.79	0.73

The above finding of train, validation, and test data score shows K-Nearest Neighbors (KNN) demonstrates substantial overfitting. On the training data, the model achieves flawless scores in all metrics yet its performance declines dramatically when applied to validation and test data. The test set accuracy stands at 0.73 yet the precision measures 0.71 while the recall reaches 0.77 along with an F1 score of 0.73. A wide difference between training and test performance suggests that the model has learned training data examples specifically rather than acquiring broad understanding. The model achieves limited success on new data points because it possesses restricted reliability for practical applications when not properly tuned or regularized.

Random Forest

We started the Random Forest training with default settings and 100 estimators before executing grid search optimization. The optimal parameter combination for this model included `n_estimators=100` together with `max_depth=10` and `min_samples_split=2` and `min_samples_leaf=4`. This resulted in a cross-validated score of about 0.85.

Table 6

Random Forest Model		
Score	Train	Validation
Accuracy	0.99	0.8
Precision	0.99	0.82
Recall	1	0.82
F1	0.99	0.82

Table 7

Random Forest Model			
Score	Train	Validation	Test
Accuracy	0.89	0.83	0.79
Precision	0.85	0.83	0.73
Recall	0.94	0.87	0.92
F1	0.89	0.84	0.81

When we tarin and evaluated Randon forest model on best parameters, it was successful for all dataset, train, validation and test. With 0.89 accuracy of train data, 0.83 accuracy of validation data and 0.79 accuracy of test data model showed high performance with strong generalization. The precision levels in trained and validation and testing data respectively showed

0.85 and 0.83 and 0.73 which indicates the model effectively avoids misclassifying cases. The model showed exceptional performance regarding recall detection particularly on test data where it achieved 0.92. Across all sets the F1 score maintained a balanced state with training reaching 0.89 while validation settled at 0.84 and the test kept 0.81. The tuned Random Forest model shows high reliable practical deployment capabilities because of its solid performance characteristics.

Logistic Regression

The trained Logistic Regression model generated an accuracy of 81% which performed identically well on training data and validation data. The model achieved solid results based on precision and recall statistics and its F1 score to demonstrate a superior balance of actual and incorrect classifications.

Table 8

Logistic Regression Model		
Score	Train	Validation
Accuracy	0.81	0.81
Precision	0.79	0.82
Recall	0.85	0.85
F1	0.82	0.83

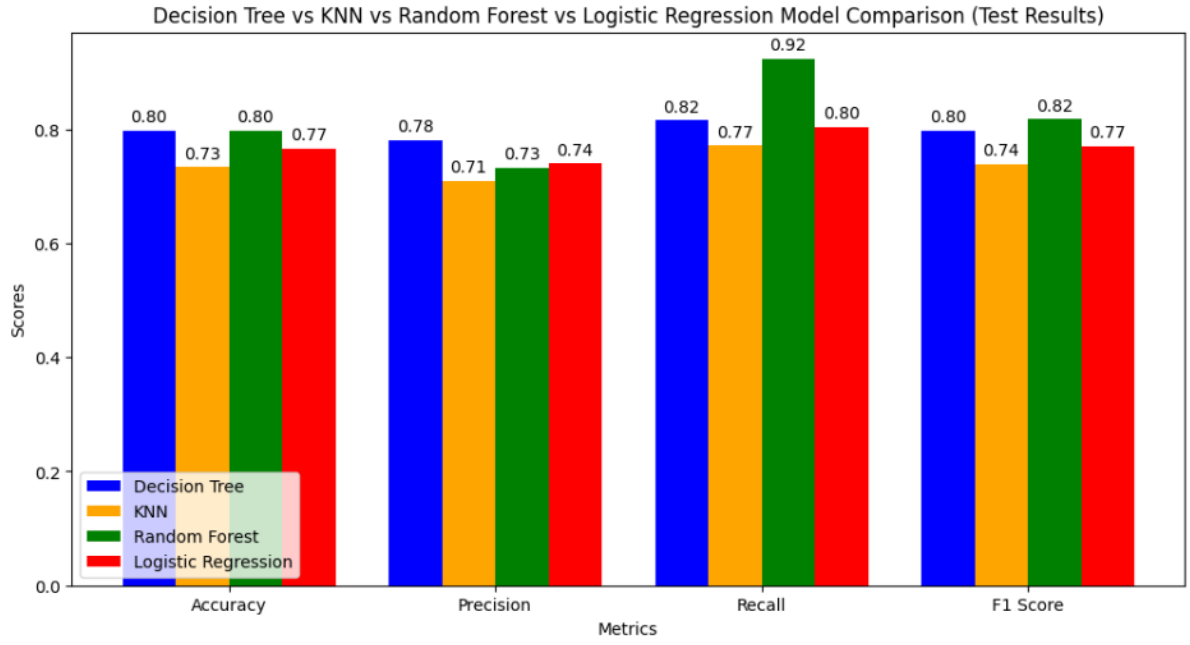
The model went through additional improvements through Grid Search that adjusted regularization parameters `C` along with `penalty` and solver configurations. Under optimized conditions the model maintained strong metrics with a test accuracy of 77% and precision of 74% as well as recall of 80% and a F1 score at 77%. The model demonstrates effective generalization capabilities together with its capability to correctly identify patients who require mental health treatment.

Table 9

Logistic Regression Model			
Score	Train	Validation	Test
Accuracy	0.81	0.81	0.77
Precision	0.79	0.81	0.74
Recall	0.85	0.85	0.8
F1	0.82	0.83	0.77

Comparison

Figure 25



This bar chart shows how Decision Tree together with KNN and Random Forest and Logistic Regression perform during the tests via Accuracy measurements and Precision results and Recall measures and F1 Score metrics.

Random Forest stands out as the superior model among all others. The Random Forest model demonstrates the best performance by attaining a Recall score of 0.92 and an F1 Score of 0.82 because it excels in both positive case detection and precision-recall equilibrium. The model achieves identical Accuracy results (0.80) which matches the results of Decision Tree.

The Decision Tree model demonstrates strong performance in model metrics across all categories which includes a perfect Precision score of 0.78 and a high F1 Score of 0.80 making it an interpretable and competitive choice. Random Forest delivers better recall performance than the Decision Tree but falls short of its results.

The performance of KNN stands as the worst among all models since it achieved 0.73 Accuracy along with 0.71 Precision and 0.77 Recall and 0.74 F1 Score. The model's performance represents its ability to excessively match the training data records that was observed previously as well as its limited capacity to generalize to new test data.

The performance of Logistic Regression falls between KNN and Decision Tree in most evaluation metrics. This baseline model demonstrates satisfactory results with Recall at 0.80 along with a F1 Score of 0.77 because it emphasizes both simplicity and interpretability.

The Random Forest model proves to be the top performer on test data because its recall efficiency and balanced metrics perform exceptionally well, followed by the Decision Tree.

Conclusion

Machine learning presents itself as a method that works effectively and ethically to detect workplace mental health disorders among employees by studying survey results. The Random Forest model achieved optimal predictive power through its exceptional performance in Recall measurements needed to identify persons needing early mental health intervention. The Decision Tree model displayed decent accuracy while providing clear interpretability that makes it valuable for mental health applications. Logistic Regression produced balanced results and simple mechanisms thus becoming an advantageous tool for scenarios with restricted resources. The initial promising results of KNN modeling eventually led to performance levels lower than other tested models.

Our group developed a dependable and interpretation-friendly prediction system able to guide organizations and mental health teams regarding the distribution of resources dedicated to employee mental health care.

References

Kaggle. (n.d.). Mental health in tech survey. Retrieved from https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey