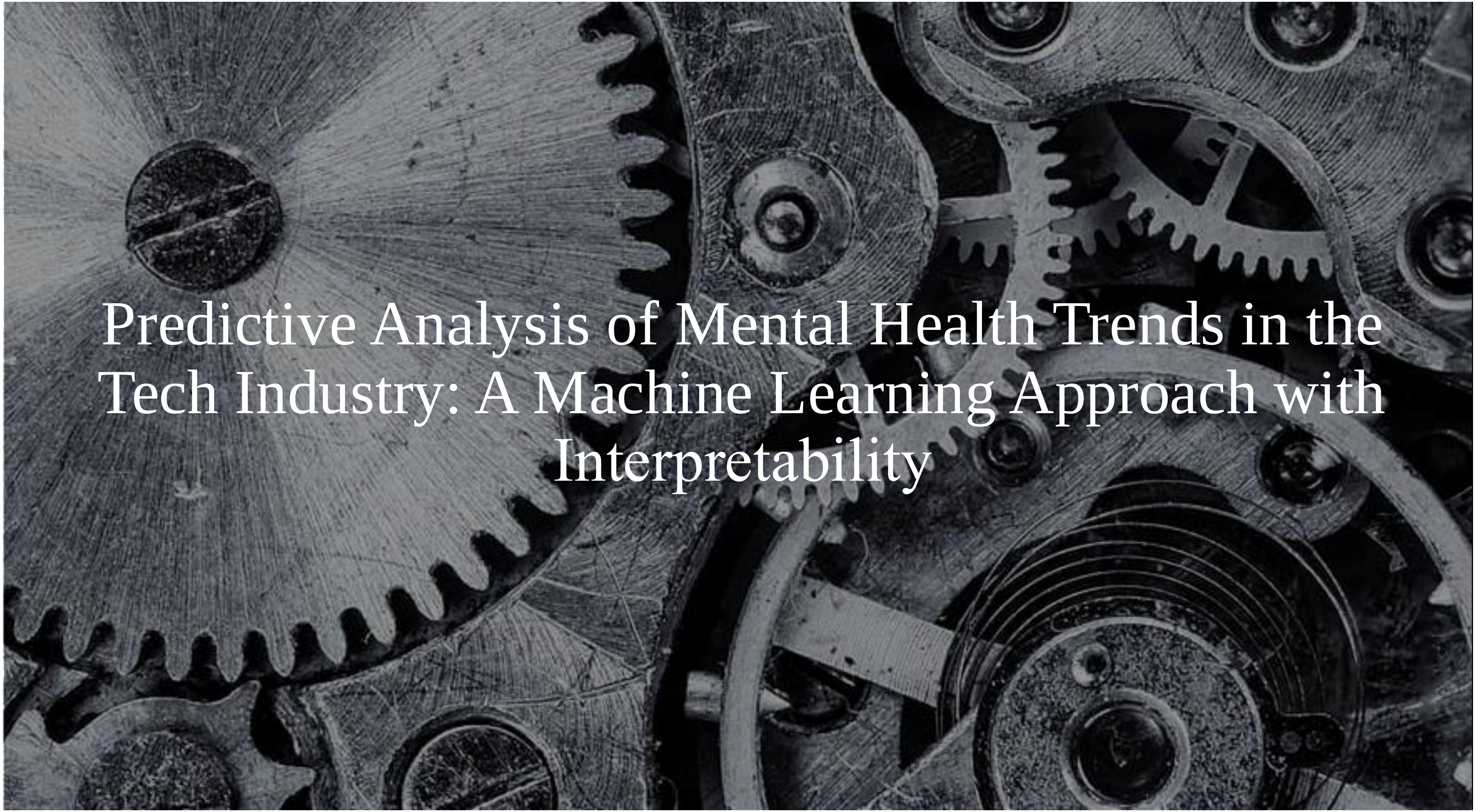


Predictive Analysis of Mental Health Trends in the Tech Industry: A Machine Learning Approach with Interpretability

ABSTRACT

This project aims to develop a tool for predicting mental health trends within the tech industry and providing interpretable insights into those predictions.

Using machine learning techniques on data from the Open-Source Mental Illness (OSMI) survey, the project builds classification models and analyzes them with LIME (Locally Interpretable Model-Agnostic Explanations) to understand the factors influencing predictions.

The project found XGBoost to be the most accurate model with relevant feature usage for classification, making it a valuable tool for identifying potential mental health issues and informing strategies for creating a healthier workplace.

INTRODUCTION

The tech industry faces a significant mental health crisis, with 62% of IT professionals experiencing physical and emotional exhaustion, and 42% considering leaving within six months. To address this, project uses machine learning to forecast mental health trends in the tech industry, aiming to identify potential issues and provide insights into factors influencing these predictions.

- **Previous Studies** on mental health challenges in the corporate sector, particularly in the tech industry, have used machine learning techniques to analyze causes and predict employee attrition and stress levels. However, the need for comprehensive analyses and the integration of tools like LIME is growing, ensuring transparency and interpretability in predictions.
- **Research Methodology** involves a systematic approach, starting with exploratory data analysis using the OSMI survey dataset. Data visualization and label encoding are used, followed by clustering and classification models like Logistic Regression, K-Nearest Neighbours, Decision Tree, Random Forest, and ensemble techniques. LIME is integrated for interpretability.
- **Research Aims** to improve mental health in the tech industry by identifying predictive patterns and using LIME 's interpretability to offer insights for informed decision-making and targeted interventions.

NEED OF STUDY

The tech industry, while at the forefront of innovation, presents unique challenges that significantly impact the mental health of its workforce. The 2022 Burnout Index survey

- 62% of IT professionals experience physical and emotional exhaustion due to work demands.
- 69% of women and 56% of men feel drained after workdays.
- 2 in 5 workers exhibit a high risk of burnout, with 42% considering quitting within 6 months.

Why this study:

- Early identification of at-risk individuals
- tailored interventions
- Cost-effectiveness
- Data-driven decision making

PROBLEM STATEMENT

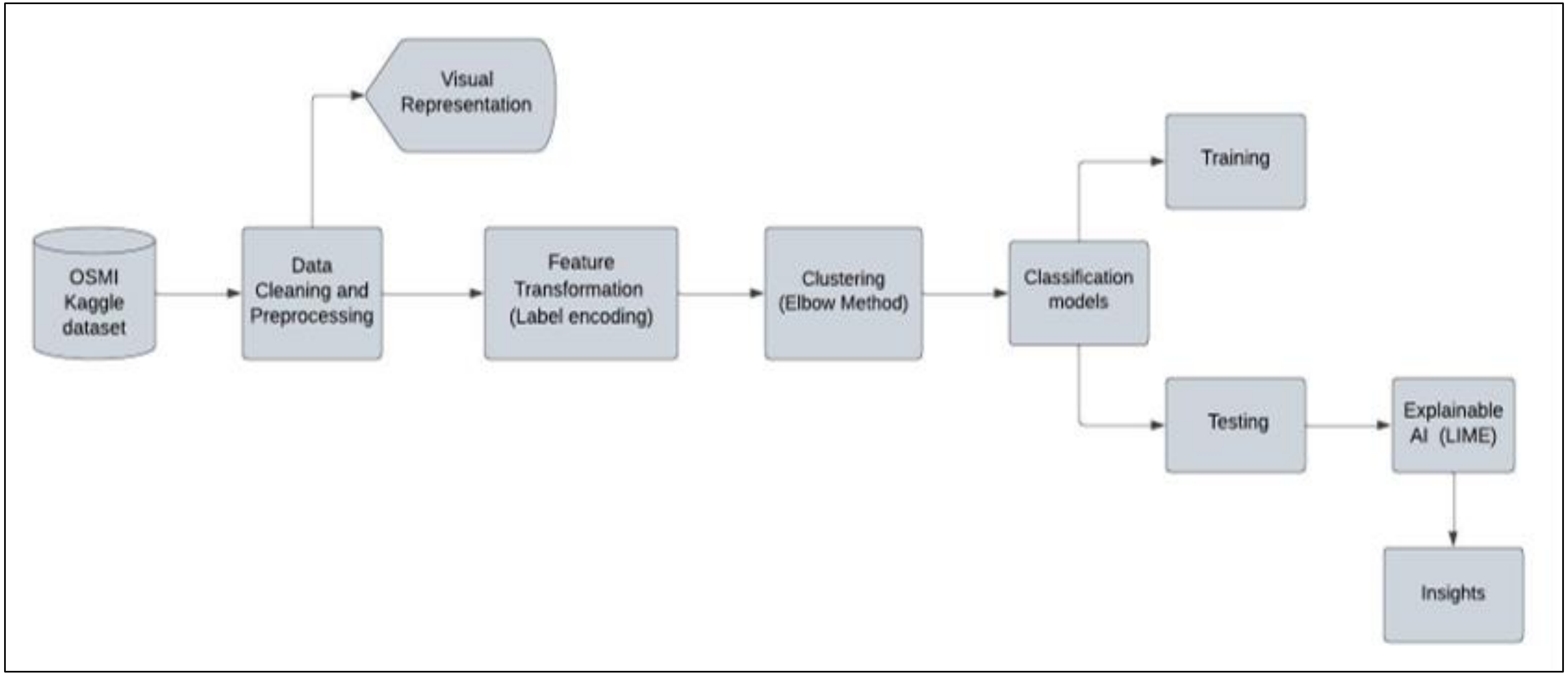
- The tech industry, marked by its rapid pace and demanding work environment, faces an increasing concern regarding the mental health of IT professionals.
- The lack of proactive measures to identify and address mental health issues can lead to severe consequences for both individuals and organizations.
- To tackle this challenge, we propose the development of a machine learning-based predictive tool integrated with interpretability through LIME, aimed at proactively identifying and understanding mental health trends among IT professionals within the tech industry.

OBJECTIVES

The main objective of this study is to analyze and model a given dataset through a comprehensive data science pipeline, encompassing data cleaning, preprocessing, exploratory data analysis (EDA), label encoding, clustering using the K-means algorithm, classification models, and Explainable AI through LIME framework. The study aims to achieve the following specific objectives:

- **Data Collection**: To obtain the dataset for the desired project.
- **Data Cleaning and Preprocessing**: To handle missing data, outliers, and normalize numerical features.
- **Exploratory Data Analysis (EDA)**: To explore data distribution and identify relationships between variables.
- **Label Encoding**: To transform categorical variables into numerical format.
- **K-means Clustering**: To determine optimal clusters using the elbow method and silhouette analysis.
- **Classification Models**: Perform split data, train models, and evaluate performance.
- **Explainable AI with LIME**: Apply LIME to enhance model interpretability.

PROPOSED METHODOLOGY



- Data Preprocessing:
  - **Column Removal**: Removed irrelevant columns such as response ID and response.
  - **Column Renaming**: Renamed columns for enhanced clarity and understanding.
  - **Visualization**: Utilized Power BI for data visualization to extract insights.
- Data Transformation:
  - **Label Encoding**: Converted categorical and Boolean data into numerical format for model compatibility.
  - **Handling Null Values**: Imputed missing values in numeric columns using the median.
  - **Outlier Handling**: Employed the trimming and capping method to address outliers.
- Data Balancing:
  - Checked and ensured the balance of label data to prevent model bias.
- Clustering:
  - Performed clustering using the elbow method to identify optimal clusters for improved model performance.
- Data Splitting:
  - Split the dataset into training and testing sets with a test size of 30 percent.

- Model Selection:
  - ✓ Logistic Regression, Decision Tree, K-Nearest Neighbors, Random Forest, Ensemble Technique (KNN, Random Forest, Decision Tree), Gradient Booster, AdaBoost, XGBoost.
- Model Training and Prediction:
  - ✓ Created and trained each selected model using the training set.
  - ✓ Generated predictions using the test set.
- Model Evaluation:
  - ✓ **Feature Importance**: Determined feature scores using a Random Forest classifier.
  - ✓ **Performance Metrics**: Checked model scores on both the training and test datasets.
  - ✓ **Classification Report**: Generated a classification report for each model.
  - ✓ **Confusion Matrix**: Plotted confusion matrices to visualize model performance.
- Explainable AI (LIME):
  - ✓ Applied Local Interpretable Model-agnostic Explanations (LIME) to the top-performing models (Ensemble, XGBoost, Gradient Boosting) to enhance interpretability and understand model decisions.

DATASET

- *Source of the Dataset:*
  - Kaggle (OSMI Tech survey)
  - OSMI mental health Dataset

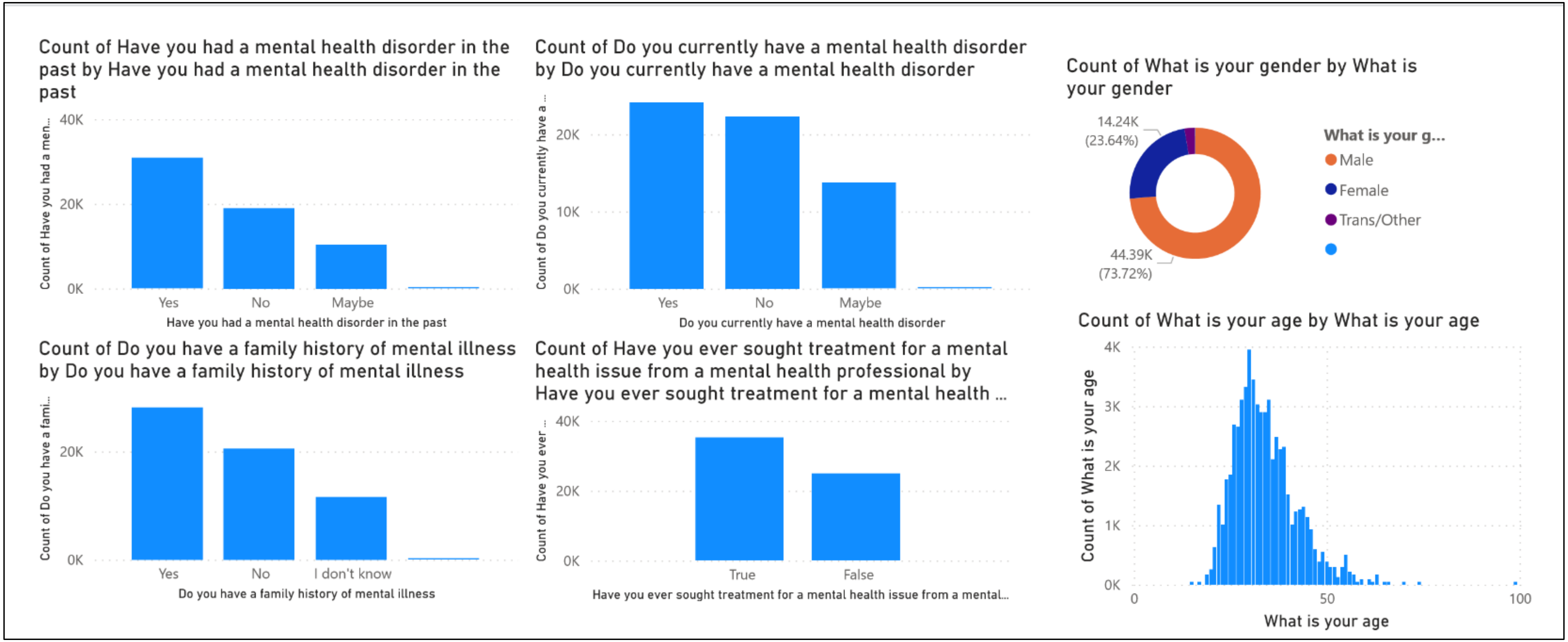
- *No. of Observations:*
  - 60186
- Column Details:
  - 27 columns

- *Details about the columns:*
  - 19 - String
  - 7 - Boolean
  - 1 - Integer

Screenshot of the dataset:

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Are you se	How many	Is your em	Is your pri	Do you ha	Do you ha	Have you	Do you cu	If yes, wha	If maybe,	**Have you been diagnosed with a mental he**	If so, what	Have you	What is yo	What is yo	Age Group
2	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
3	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
4	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
5	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
6	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
7	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
8	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
9	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
10	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
11	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
12	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
13	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
14	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
15	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
16	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
17	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
18	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
19	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
20	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
21	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
22	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
23	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40
24	FALSE	26-100	TRUE		TRUE	No	Yes	No			TRUE	Anxiety Di	FALSE	39	Male	36-40

OSMI Mental Health Dataset



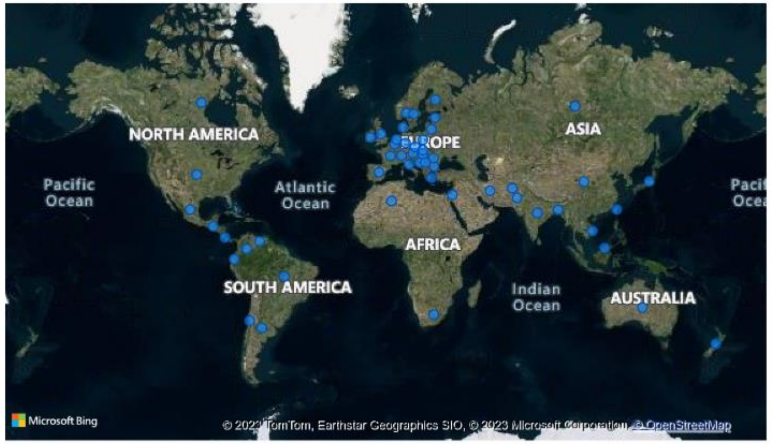
Count of Have you had a mental health disorder in the past by Have you had a mental health disorder in the past
Count of Have you had a men...
40K
20K
0K
Yes
No
Maybe
Have you had a mental health disorder in the past
Count of Do you currently have a mental health disorder by Do you currently have a mental health disorder
Count of Do you currently have a ...
20K
10K
0K
Yes
No
Maybe
Do you currently have a mental health disorder
Count of What is your gender by What is your gender
14.24K
(23.64%)
44.39K
(73.72%)
What is your g...
Male
Female
Trans/Other
Count of Do you have a family history of mental illness by Do you have a family history of mental illness
Count of Do you have a fami...
20K
0K
Yes
No
I don't know
Do you have a family history of mental illness
Count of Have you ever sought treatment for a mental health issue from a mental health professional by Have you ever sought treatment for a mental health ...
Count of Have you ever ...
40K
20K
0K
True
False
Have you ever sought treatment for a mental health issue from a mental...
Count of What is your age by What is your age
Count of What is your age
4K
3K
2K
1K
0K
0
50
100
What is your age

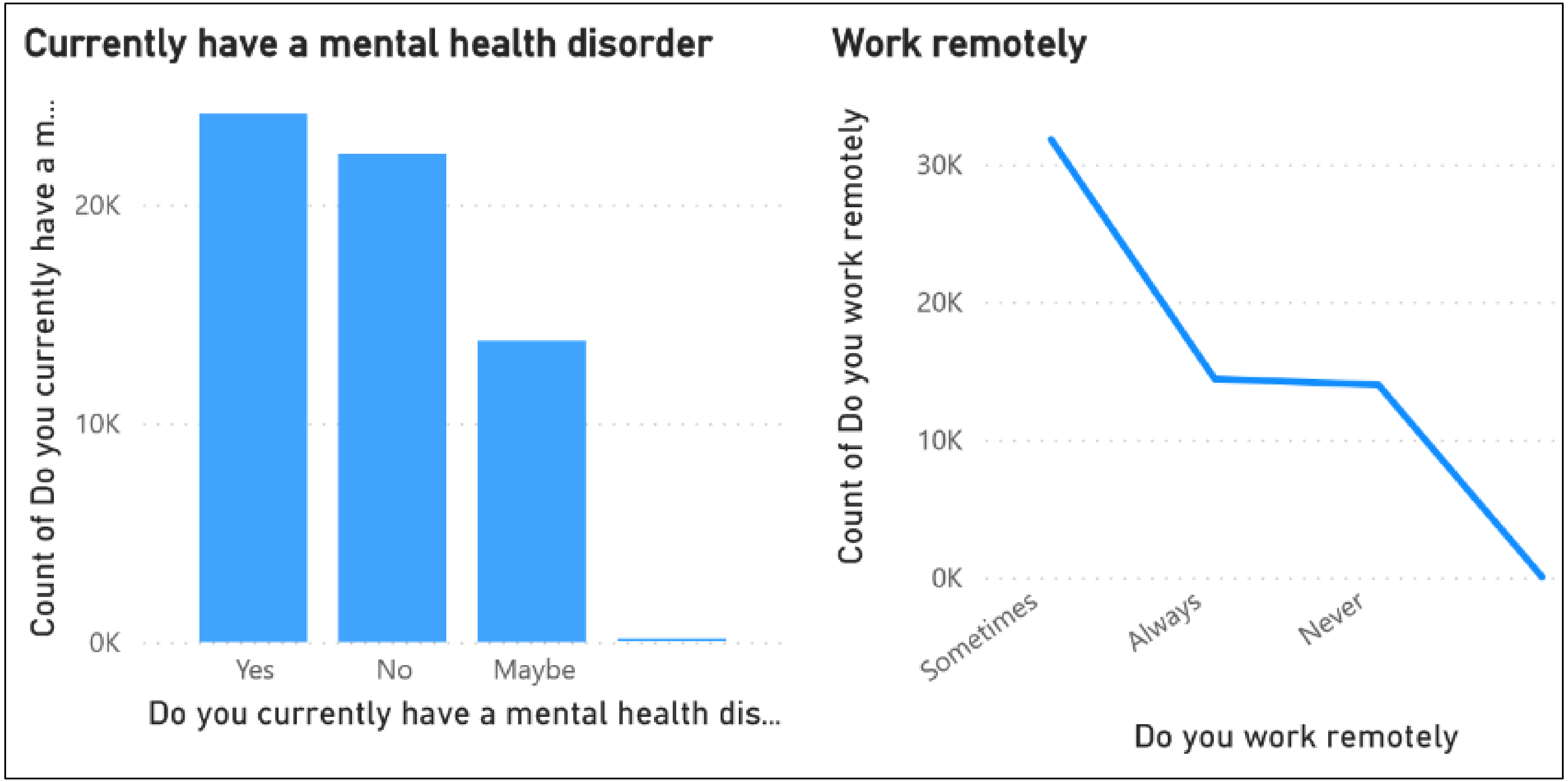
What US state or territory do you live in

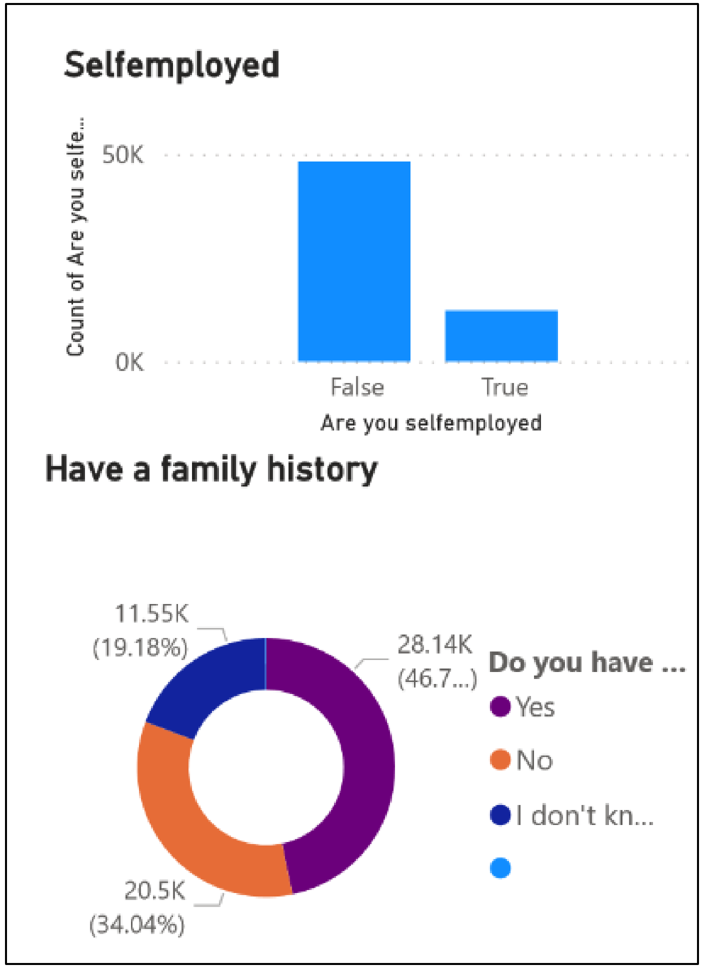
What US state or territory do you work in

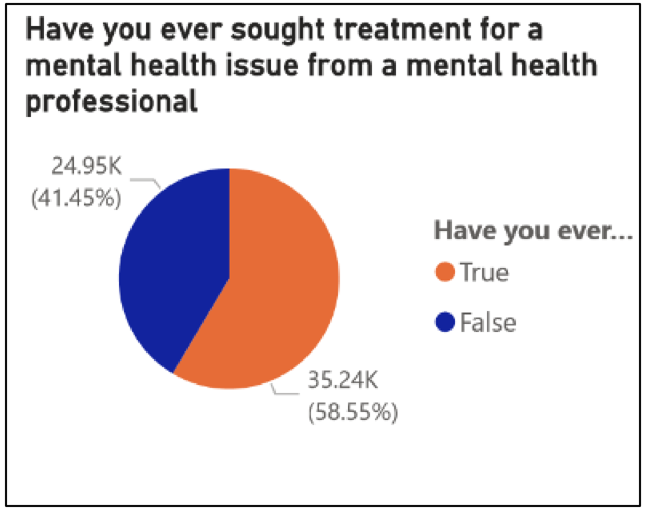
What country do you work in

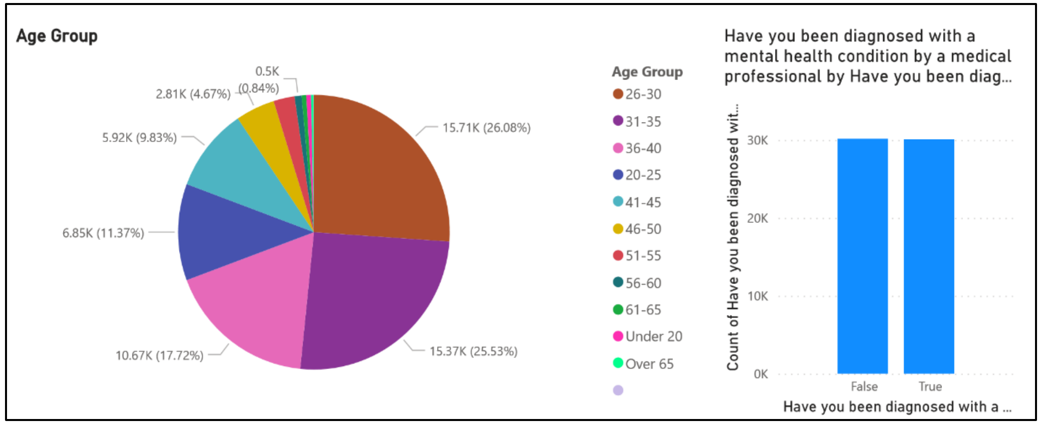
What country do you live in



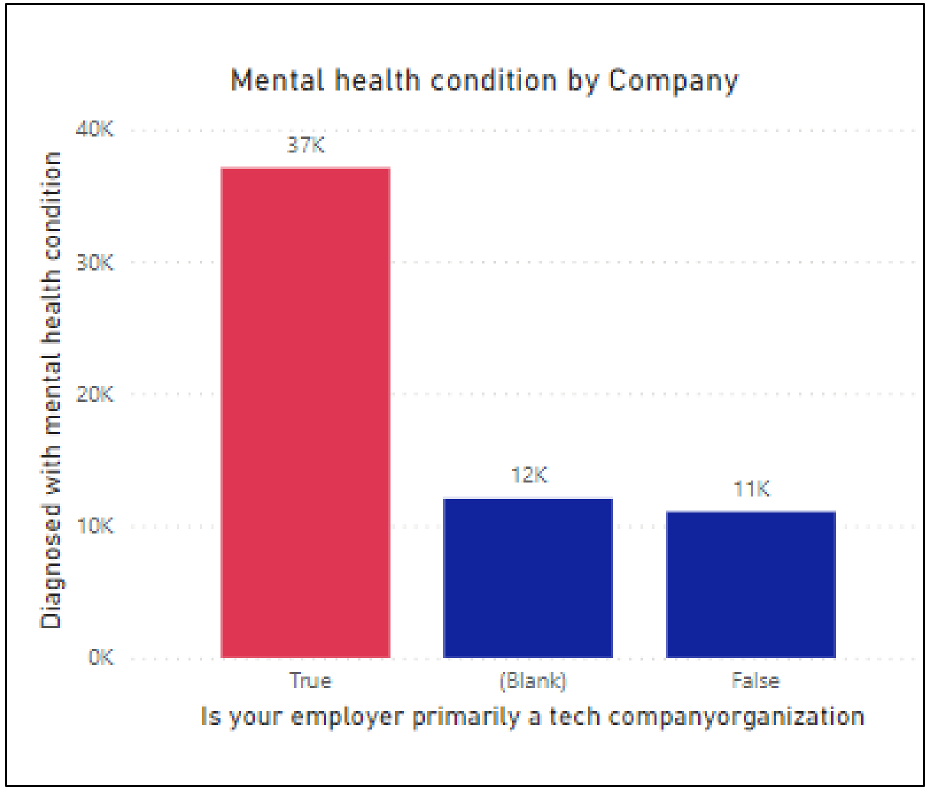

Currently have a mental health disorder
Count of Do you currently have a m...
20K
10K
0K
Yes
No
Maybe
Do you currently have a mental health dis...
Work remotely
Count of Do you work remotely
30K
20K
10K
0K
Sometimes
Always
Never
Do you work remotely

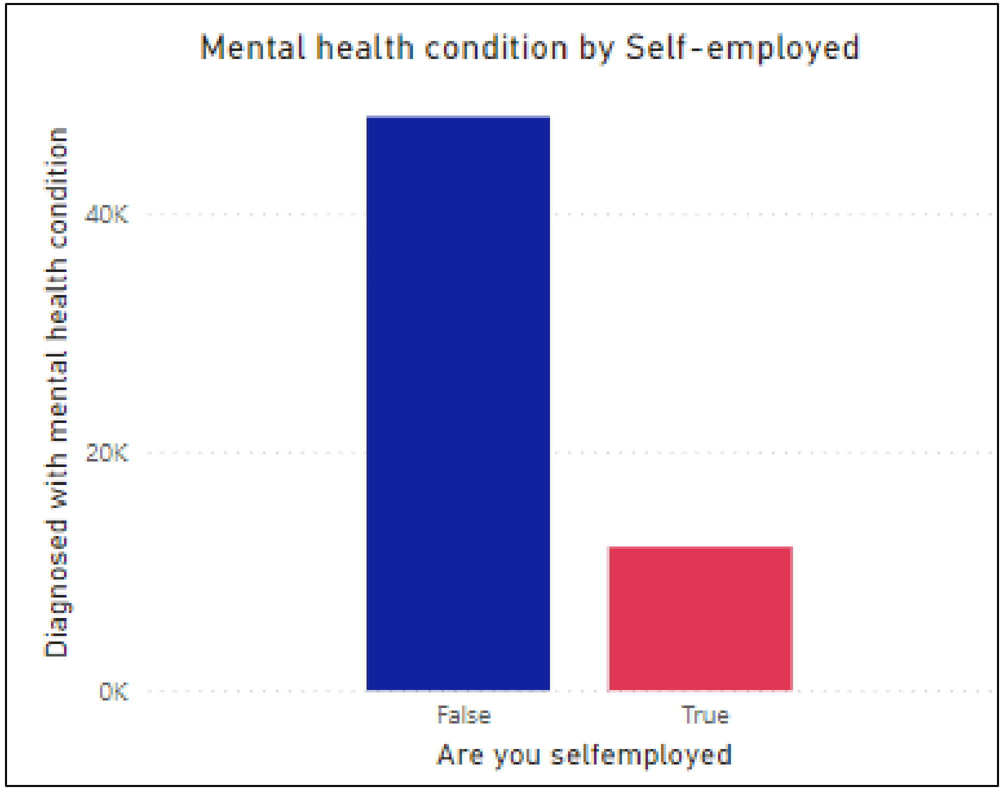

Selfemployed
Count of Are you selfe...
50K
0K
False
True
Are you selfemployed
Have a family history
11.55K (19.18%)
28.14K (46.7...)
20.5K (34.04%)
Do you have ...
Yes
No
I don't kn...


Have you ever sought treatment for a mental health issue from a mental health professional
24.95K (41.45%)
35.24K (58.55%)
Have you ever...
True
False


Age Group
0.5K (0.84%)
2.81K (4.67%)
5.92K (9.83%)
6.85K (11.37%)
10.67K (17.72%)
15.37K (25.53%)
15.71K (26.08%)
Age Group
26-30
31-35
36-40
20-25
41-45
46-50
51-55
56-60
61-65
Under 20
Over 65
Have you been diagnosed with a mental health condition by a medical professional by Have you been diag...
Count of Have you been diagnosed wit...
30K
20K
10K
0K
False
True
Have you been diagnosed with a ...

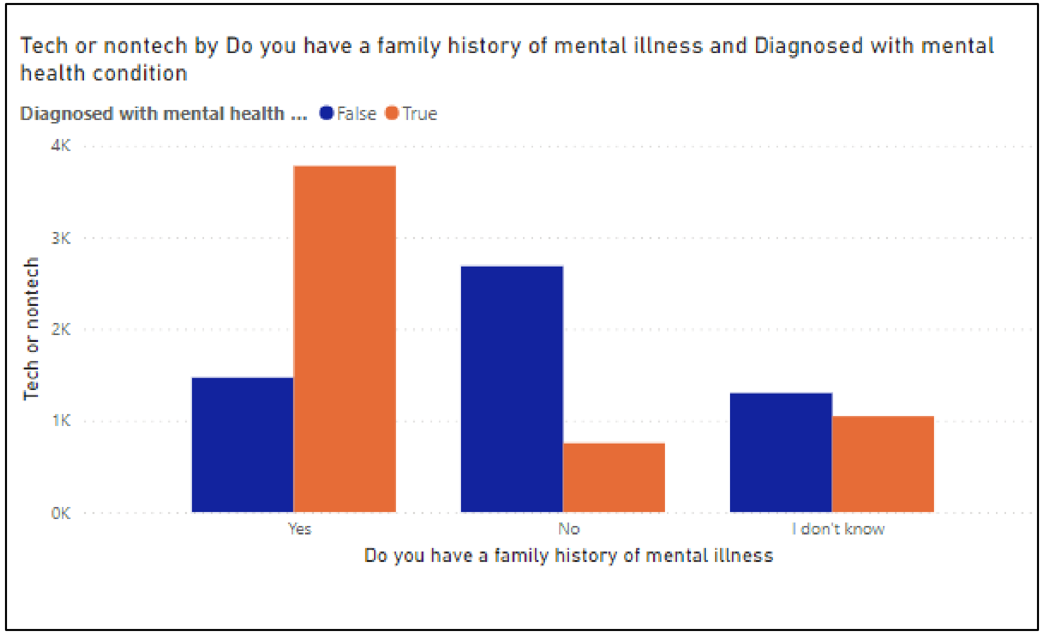
Why has the mental health of people in tech companies been focused for this study

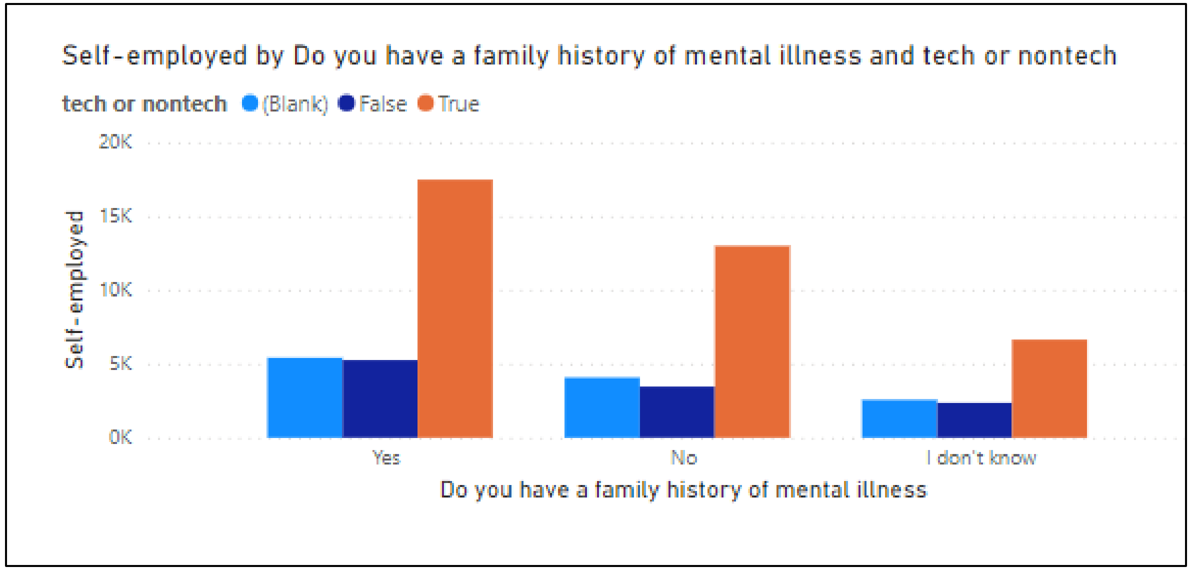




- The prevalence of mental health conditions among employees in the tech industry compared to self-employed individuals. It highlights that a higher percentage of tech employees (approximately 40%) reported having a diagnosed mental health condition, compared to self-employed individuals (approximately 30%).
- Reveals that employees working for tech companies are more likely to have a diagnosed mental health condition (approximately 200) than employees working for non-tech companies (approximately 126)

Family history affecting the mental illness





It compares the responses of tech employees and self-employed individuals to two questions:

1.Do you have a family history of mental illness?

2.Are you diagnosed with a mental health condition?

Chart shows that a higher percentage of tech employees (40%) who answered "Yes" to the first question also answered "Yes" to the second question (200), indicating a stronger association between family history and mental health diagnoses among tech workers.

Self-employed individuals with a family history of mental illness were slightly less likely to have a diagnosed mental health condition

EXPLORATORY DATA ANALYSIS

```python
data=pd.read_csv("OSMI Mental Health Dataset_final.csv",encoding='cp1252')
```

```python
data.shape
```

```
(60186, 27)
```

```python
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 60186 entries, 0 to 60185
Data columns (total 27 columns):
 #   Column                                                                        Non-Null Count  Dtype
---  ------                                                                        --------------  -----
 0   ResponseID                                                                    60186 non-null  object
 1   Are you selfemployed                                                          60186 non-null  bool
 2   How many employees does your company or organization have                     48132 non-null  object
 3   Is your employer primarily a tech companyorganization                         48132 non-null  object
 4   Is your primary role within your company related to techIT                    11046 non-null  object
 5   Do you have previous employers                                                60186 non-null  bool
 6   Do you have a family history of mental illness                                60186 non-null  object
 7   Have you had a mental health disorder in the past                             60186 non-null  object
 8   Do you currently have a mental health disorder                                60186 non-null  object
 9   If yes, what conditions have you been diagnosed with                          23856 non-null  object
 10  If maybe, what conditions do you believe you have                             13524 non-null  object
 11  Have you been diagnosed with a mental health condition by a medical professional  60186 non-null  bool
 12  If so, what conditions were you diagnosed with                                29862 non-null  object
 13  Have you ever sought treatment for a mental health issue from a mental health professional  60186 non-null  bool
 14  What is your age                                                              60102 non-null  float64
 15  What is your gender                                                           60186 non-null  object
 16  Age Group                                                                     60186 non-null  object
 17  What country do you live in                                                   60186 non-null  object
 18  What US state or territory do you live in                                     35280 non-null  object
 19  What country do you work in                                                   60186 non-null  object
 20  What US state or territory do you work in                                     35742 non-null  object
 21  Which of the following best describes your work position                      60186 non-null  object
 22  Do you work remotely                                                          60186 non-null  object
 23  Question Group                                                                51588 non-null  object
 24  Question about speaking openly about mental health vs physical health         60186 non-null  object
 25  Question                                                                      60186 non-null  object
 26  Response                                                                      43846 non-null  object
dtypes: bool(4), float64(1), object(22)
memory usage: 10.8+ MB
```

```python
data.nunique()
```

```
ResponseID                                                                                  1433
Are you selfemployed                                                                           2
How many employees does your company or organization have                                      6
Is your employer primarily a tech companyorganization                                          2
Is your primary role within your company related to techIT                                     2
Do you have previous employers                                                                 2
Do you have a family history of mental illness                                                 3
Have you had a mental health disorder in the past                                              3
Do you currently have a mental health disorder                                                 3
If yes, what conditions have you been diagnosed with                                         128
If maybe, what conditions do you believe you have                                             99
Have you been diagnosed with a mental health condition by a medical professional               2
If so, what conditions were you diagnosed with                                               116
Have you ever sought treatment for a mental health issue from a mental health professional     2
What is your age                                                                              51
What is your gender                                                                            3
Age Group                                                                                     11
What country do you live in                                                                   53
What US state or territory do you live in                                                     47
What country do you work in                                                                   53
What US state or territory do you work in                                                     48
Which of the following best describes your work position                                     264
Do you work remotely                                                                           3
Question Group                                                                                 6
Question about speaking openly about mental health vs physical health                          2
Question                                                                                      42
Response                                                                                    2070
dtype: int64
```

```python
data.describe()
```

	What is your age
count	60102.000000
mean	34.106219
std	8.283055
min	15.000000
25%	28.000000
50%	33.000000
75%	39.000000
max	99.000000

- Basic information of columns.
- OSHI Mental Health Dataset. The data includes information about people's work experiences, family history, and mental health history.
- Summary of the data, including the number of entries, columns, and data types.

LABEL ENCODING AND NULL VALUE

```
newdf=pd.DataFrame(data)
```

```
from sklearn.preprocessing import LabelEncoder
l=LabelEncoder()
for x in newdf:
    if newdf[x].dtypes=='object':
        newdf[x]=l.fit_transform(newdf[x])
```

```
newdf.head()
```

	ResponseID	Are you selfemployed	How many employees does your company or organization have	Is your employer primarily a tech companyorganization	Is your primary role within your company related to techIT
0	0	False	2	1	2
1	0	False	2	1	2
2	0	False	2	1	2
3	0	False	2	1	2
4	0	False	2	1	2

5 rows × 27 columns

```
newdf.isnull().sum()
```

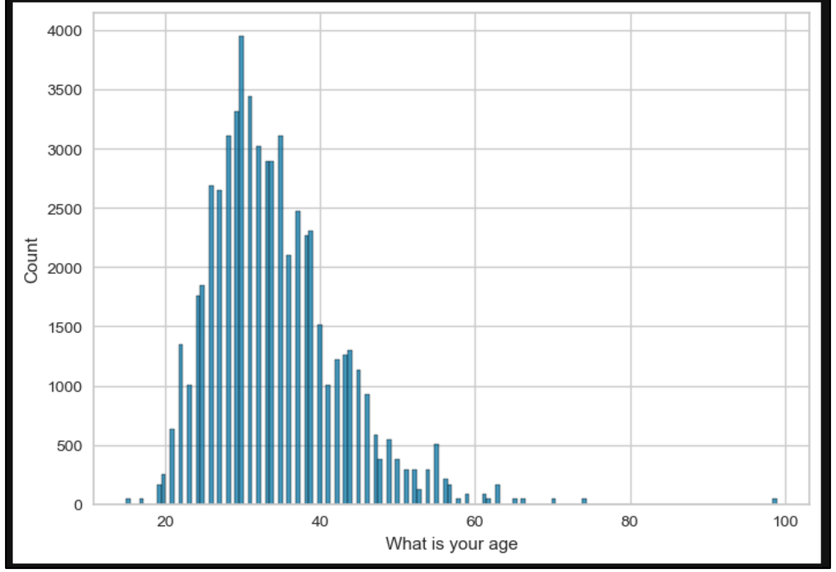
```
ResponseID                                                                          0
Are you selfemployed                                                                0
How many employees does your company or organization have                           0
Is your employer primarily a tech companyorganization                               0
Is your primary role within your company related to techIT                          0
Do you have previous employers                                                      0
Do you have a family history of mental illness                                      0
Have you had a mental health disorder in the past                                   0
Do you currently have a mental health disorder                                      0
If yes, what conditions have you been diagnosed with                                0
If maybe, what conditions do you believe you have                                   0
Have you been diagnosed with a mental health condition by a medical professional    0
If so, what conditions were you diagnosed with                                      0
Have you ever sought treatment for a mental health issue from a mental health professional    0
What is your age                                                                   84
What is your gender                                                                 0
Age Group                                                                           0
What country do you live in                                                         0
What US state or territory do you live in                                           0
What country do you work in                                                         0
What US state or territory do you work in                                           0
Which of the following best describes your work position                            0
Do you work remotely                                                                0
Question Group                                                                      0
Question about speaking openly about mental health vs physical health               0
Question                                                                            0
Response                                                                            0
dtype: int64
```

The label encoding process has been applied to the data in the table. This means that the categorical data has been converted into numerical data. Already existed null values have been converted into Zero

HANDLE NULL VALUE

```
sns.histplot(data["What is your age"])
```

```
<Axes: xlabel='What is your age', ylabel='Count'>
```




```
newdf['What is your age'] = newdf['What is your age'].fillna(newdf['What is your age'].median())
```

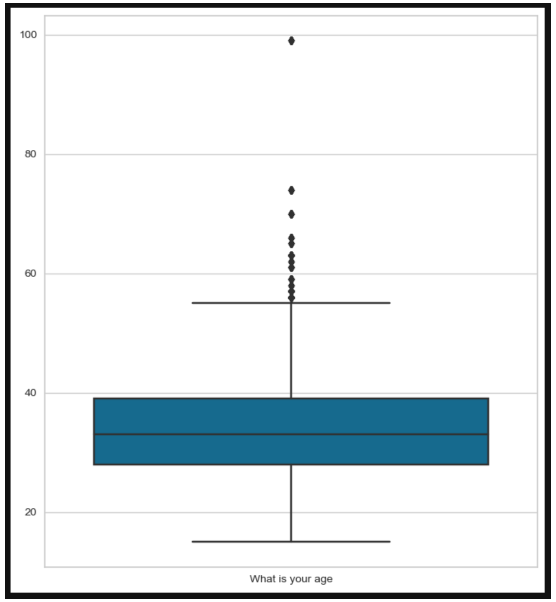
```
newdf['What is your age'].isnull().sum()
```

```
0
```

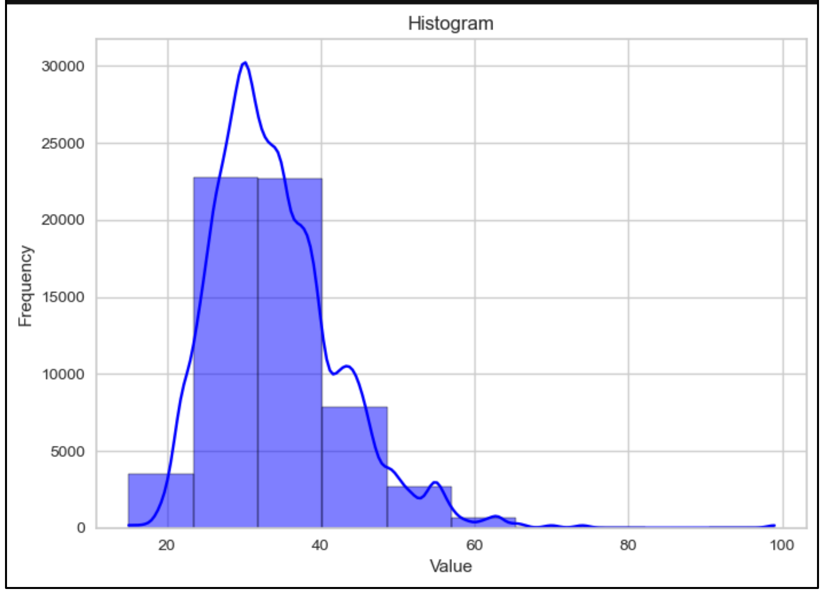
Graph of a numerical column that is skewed to the right. This means that there are more values on the left side of the distribution than on the right side.
The graph also shows a null value on this column.
To handle the null value, the median value of the column was used to replace it.

OUTLIER ANALYSIS

```
plt.figure(figsize=(8, 9))
selected_columns = ['What is your age']
columns = newdf[selected_columns]
sns.boxplot(columns)
```




```
sns.histplot(newdf['What is your age'], bins=10, kde=True, color='blue', edgecolor='black')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.title('Histogram')
plt.show()
```



The box plot shows that there is a long tail on the right side of the distribution, which indicates that there are a number of outliers. The frequency histogram also shows that there are a number of data points that are far away from the main body of the distribution.

Data is Skewed so use Inter-Quartile Range (IQR) proximity rule.

OUTLIER HANDLE

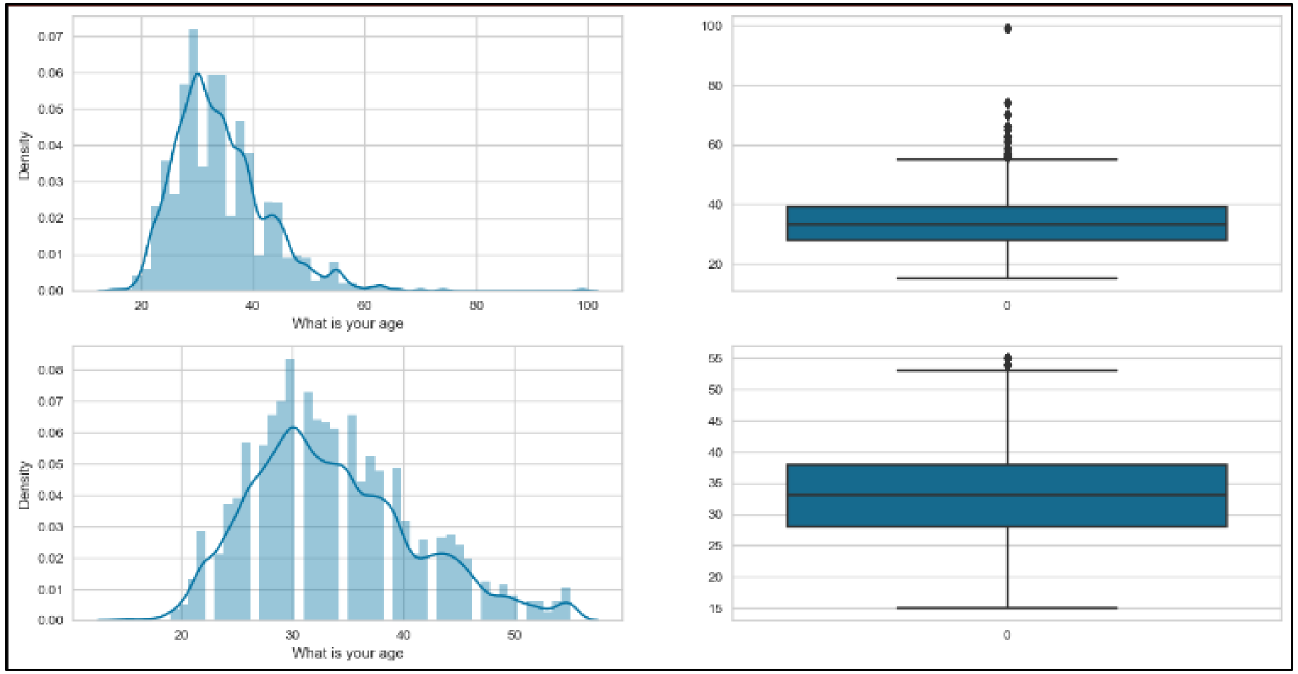
Trimming

```python
Q1 = newdf['What is your age'].quantile(0.25)
Q3 = newdf['What is your age'].quantile(0.75)
IQR = Q3 - Q1
```

```python
upper_limit = Q3 + 1.5 * IQR
lower_limit = Q1 - 1.5 * IQR
newdf[newdf['What is your age'] > upper_limit]
newdf[newdf['What is your age'] < lower_limit]
new_df2 = newdf[newdf['What is your age'] < upper_limit]
new_df2.shape
```

```
(59178, 27)
```

```python
plt.figure(figsize=(16,8))
plt.subplot(2,2,1)
sns.distplot(newdf['What is your age'])
plt.subplot(2,2,2)
sns.boxplot(newdf['What is your age'])
plt.subplot(2,2,3)
sns.distplot(new_df2['What is your age'])
plt.subplot(2,2,4)
sns.boxplot(new_df2['What is your age'])
plt.show()
```

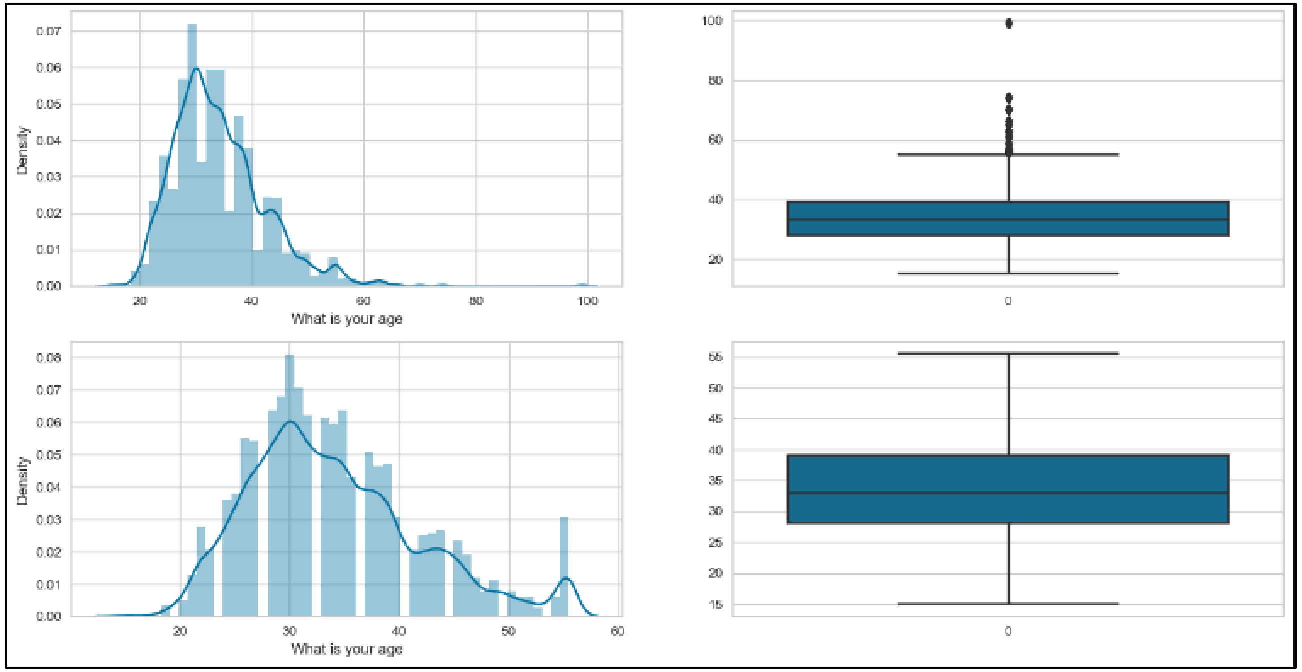


- One graph is a density plot of the age distribution in the dataset "newdf". The other graph is a boxplot of the age distribution in the dataset "newdf2".
- The density plot shows that the age distribution in "newdf" is positively skewed, meaning that there is a longer tail on the right side of the distribution than on the left side.
- The boxplot confirms this, as it shows that there are a few data points that are far away from the rest of the distribution. These outliers are likely to be the cause of the skew in the density plot.
- Removing these two outliers from the "newdf" dataset will reduce the skew in the distribution and make the IQR proximity rule more reliable for identifying outliers.

Capping

```python
new_df_cap = newdf.copy()
new_df_cap['What is your age'] = np.where(
    new_df_cap['What is your age'] > upper_limit,
    upper_limit,
    np.where(
       new_df_cap['What is your age'] < lower_limit,
        lower_limit, new_df_cap['What is your age']))
```

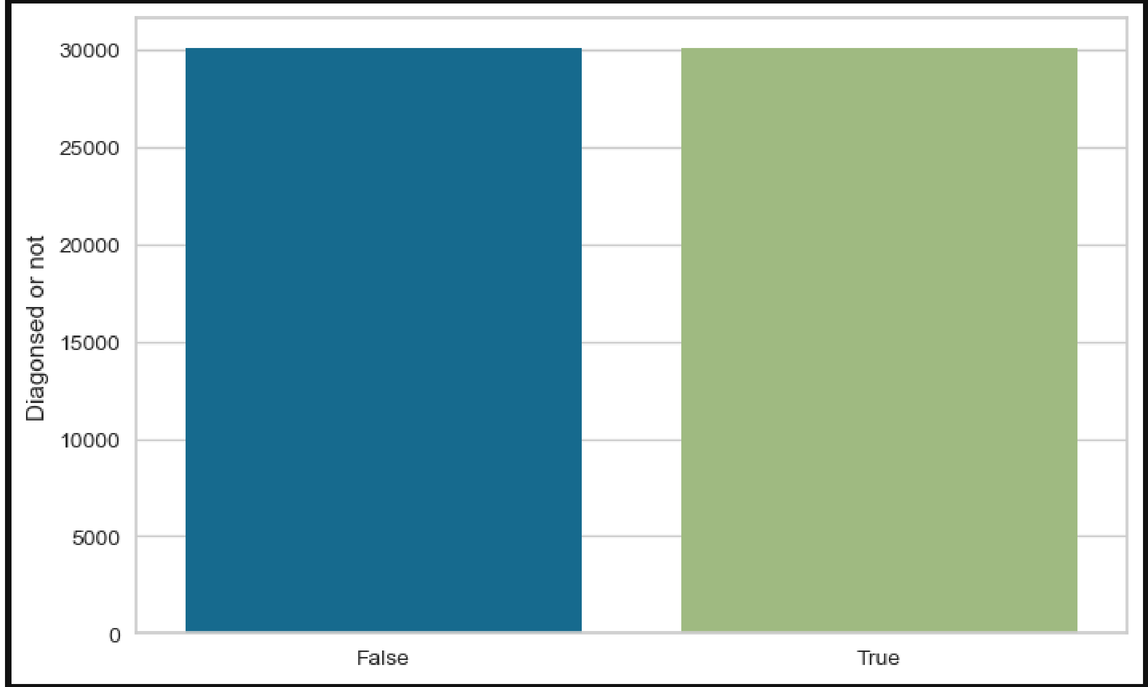
```python
plt.figure(figsize=(16,8))
plt.subplot(2,2,1)
sns.distplot(newdf['What is your age'])
plt.subplot(2,2,2)
sns.boxplot(newdf['What is your age'])
plt.subplot(2,2,3)
sns.distplot(new_df_cap['What is your age'])
plt.subplot(2,2,4)
sns.boxplot(new_df_cap['What is your age'])
plt.show()
```



The density plot shows that the distribution of the data is now more symmetrical, with no significant skew. This suggests that the IQR proximity rule was effective in identifying and removing the outliers.

DATA - BALANCED OR IMBALANCED

```
plt.figure(figsize=(8,5))
sns.barplot(x=newdf['Diagonsed or not'].value_counts().index,y=newdf['Diagonsed or not'].value_counts())
plt.show()
```



The graph shows that the distribution is balanced, meaning that there are approximately equal numbers of data points in each class.

CLUSTERING - DETERMINE OPTIMAL CLUSTER

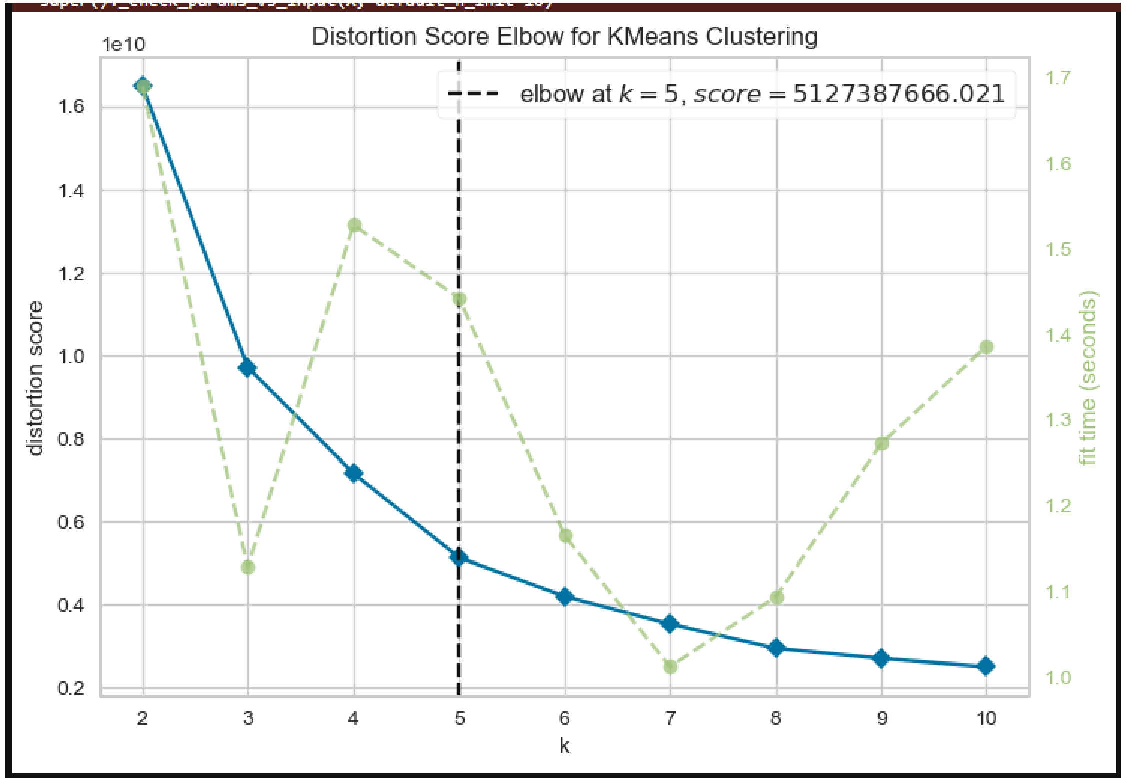
```
df1=newdf.drop('Diagonsed or not',axis=1)
```

```
! pip install yellowbrick
```

```
Defaulting to user installation because normal site-packages is not writeab
Requirement already satisfied: yellowbrick in c:\users\23msp3093\appdata\ro
Requirement already satisfied: matplotlib!=3.0.0,>=2.0.2 in c:\programdata\
Requirement already satisfied: scipy>=1.0.0 in c:\programdata\anaconda3\lib
Requirement already satisfied: scikit-learn>=1.0.0 in c:\programdata\anacon
Requirement already satisfied: numpy>=1.16.0 in c:\programdata\anaconda3\li
Requirement already satisfied: cycler>=0.10.0 in c:\programdata\anaconda3\l
Requirement already satisfied: contourpy>=1.0.1 in c:\programdata\anaconda3
Requirement already satisfied: fonttools>=4.22.0 in c:\programdata\anaconda
Requirement already satisfied: kiwisolver>=1.0.1 in c:\programdata\anaconda
Requirement already satisfied: packaging>=20.0 in c:\programdata\anaconda3\
Requirement already satisfied: pillow>=6.2.0 in c:\programdata\anaconda3\li
Requirement already satisfied: pyparsing>=2.3.1 in c:\programdata\anaconda3
Requirement already satisfied: python-dateutil>=2.7 in c:\programdata\anaco
Requirement already satisfied: joblib>=1.1.1 in c:\programdata\anaconda3\li
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\programdata\anaco
Requirement already satisfied: six>=1.5 in c:\programdata\anaconda3\lib\sit
```

```
from sklearn.cluster import KMeans
from yellowbrick.cluster import KElbowVisualizer, SilhouetteVisualizer

model = KMeans(random_state=42)
elb_visualizer = KElbowVisualizer(model, k=(2,11))
elb_visualizer.fit(df1)
elb_visualizer.show()
```



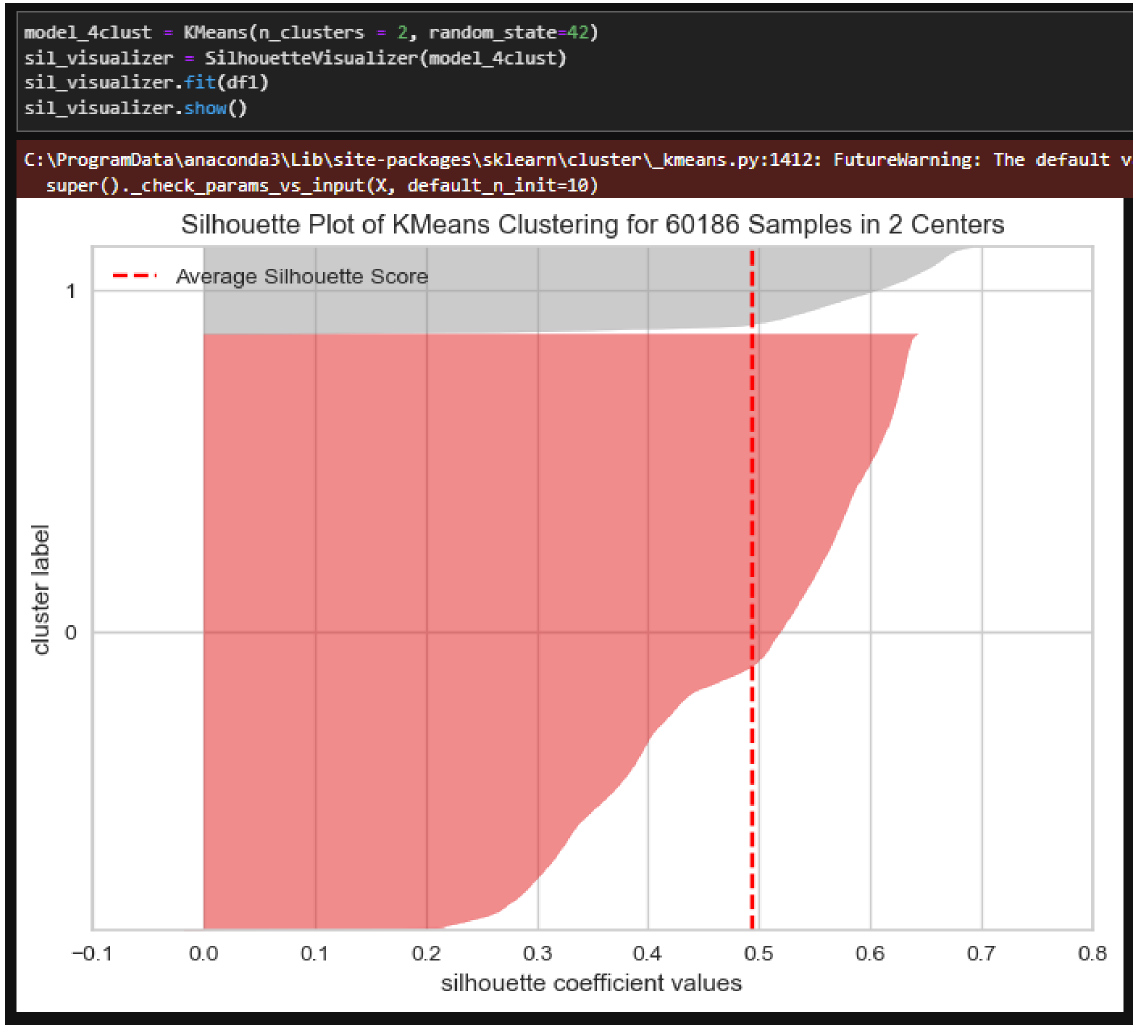

```
<Axes: title={'center': 'Distortion Score Elbow for KMeans Clustering'}, xlabel='k', ylabel='distortion score'>
```

The blue line shows the destruction score for different values of k.

The vertical line at k=5 indicates the optimal number of clusters.

Distortion score elbow for KMeans clustering with the vertical line has 5 as best number of Ks

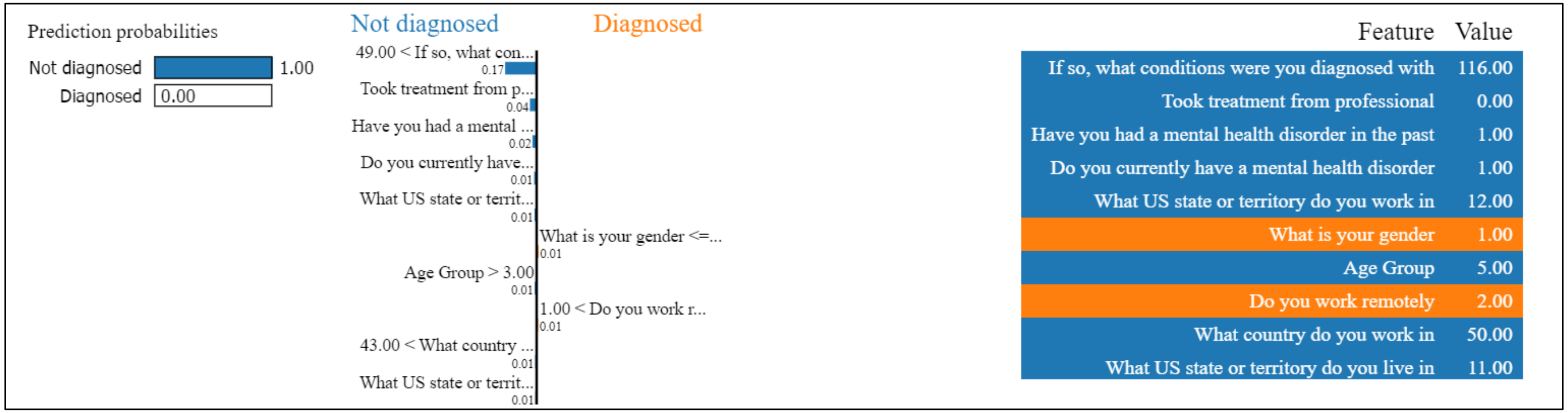
Target feature has 2 clusters, so checks the data is balanced or not through silhouette plot. And it is balanced

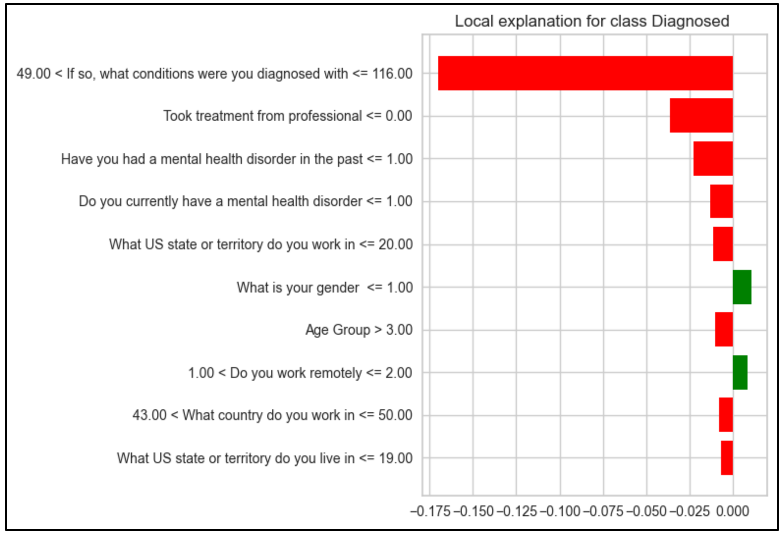


- Silhouette plot of k-means clustering for 60186 samples in 2 centers. This type of plot is used to evaluate the quality of clustering by measuring how well each data point is assigned to its cluster.
- From the silhouette plot the k-means clustering with 2 centers is a good way to group the data points in this dataset.

RESULTS AND DISCUSSION

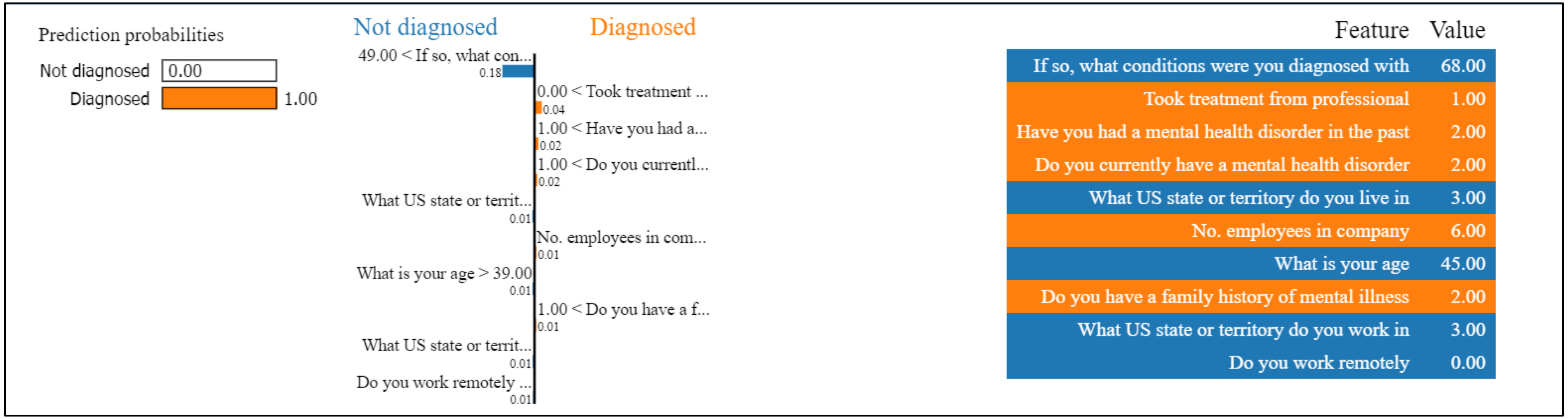
Ensemble model with LIME

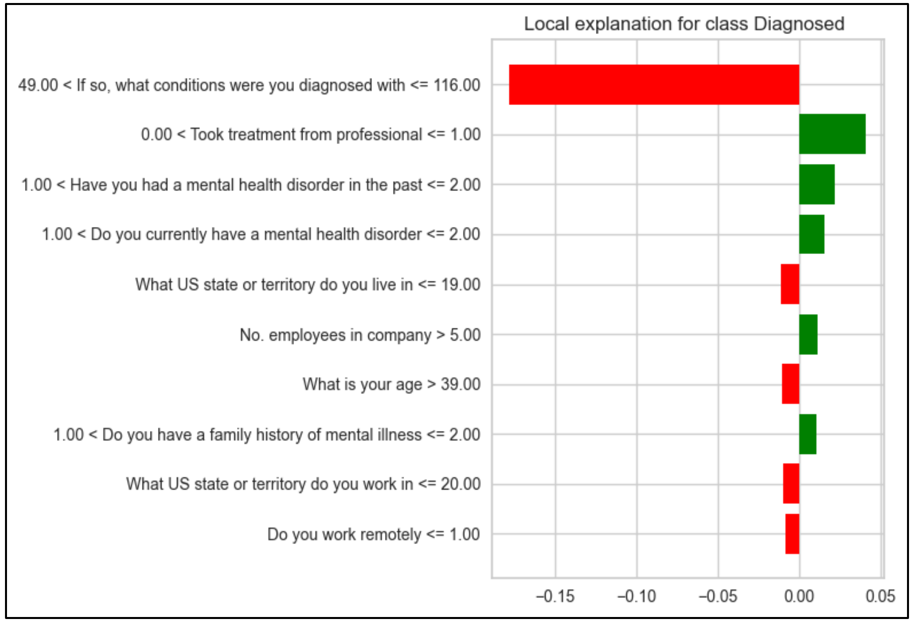





```
Intercept 0.9963890559712774
Prediction_local [0.73734024]
Right: 0.0
```

The person who is not diagnosed with mental health issue because of the condition, treatment, disorder in past, current mental health, state work in, age group, country work in and state live in.

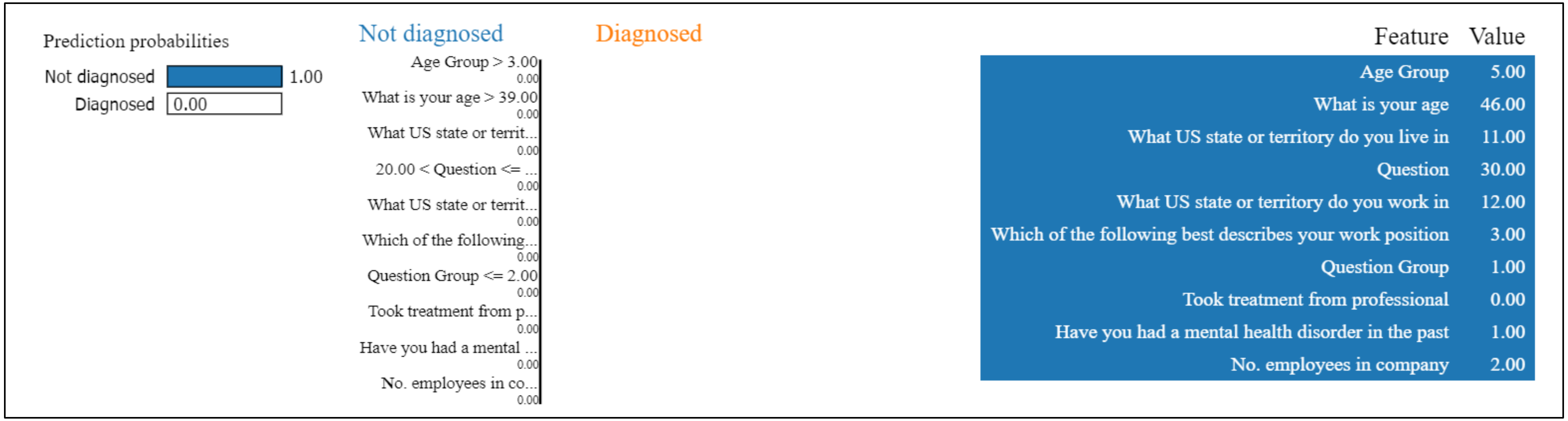


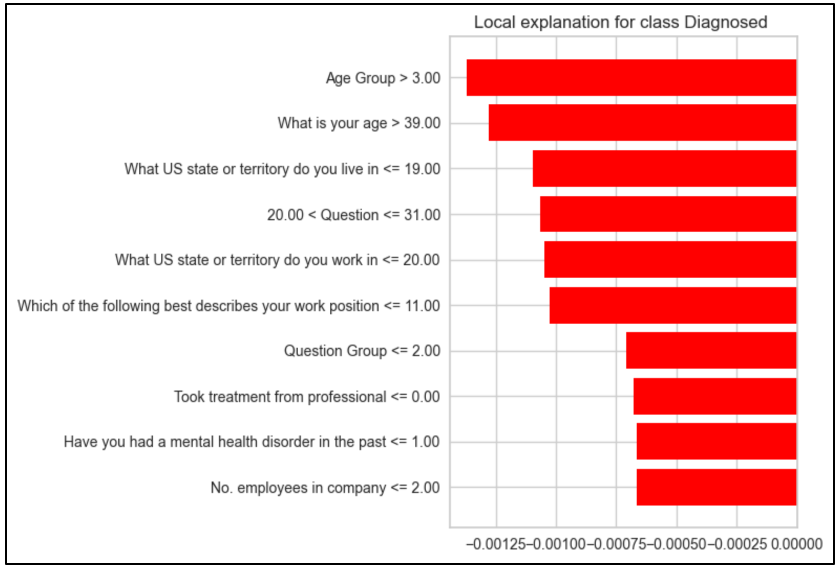



```
Intercept 0.9287602885525325
Prediction_local [0.80889337]
Right: 1.0
```

The person who is diagnosed with mental health issue because of the treatment, disorder in past, current mental health, family history of mental illness.

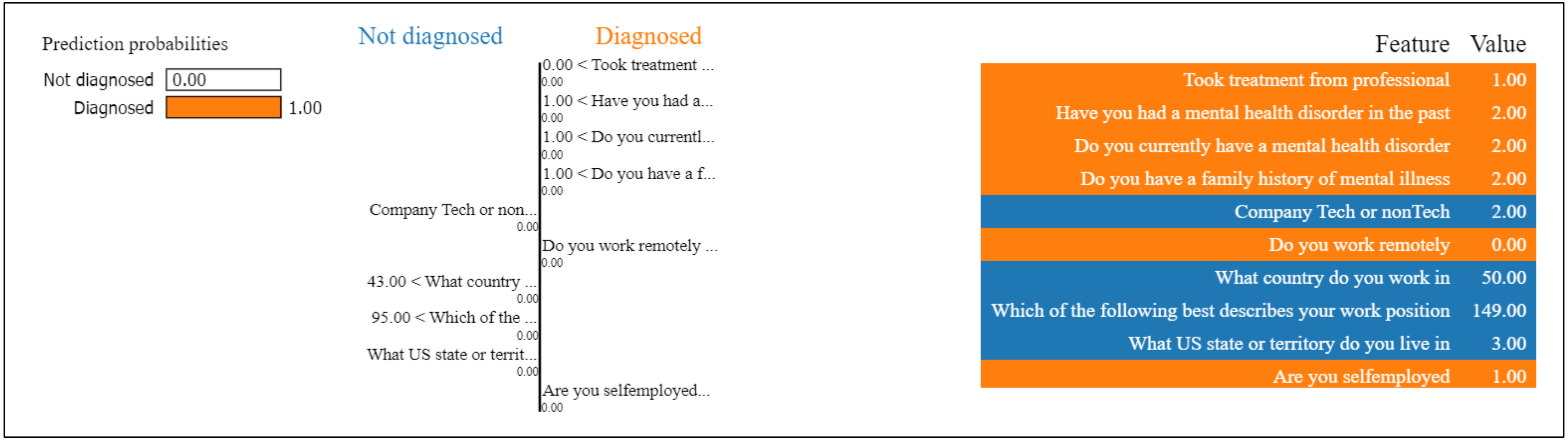
XGBoost with LIME

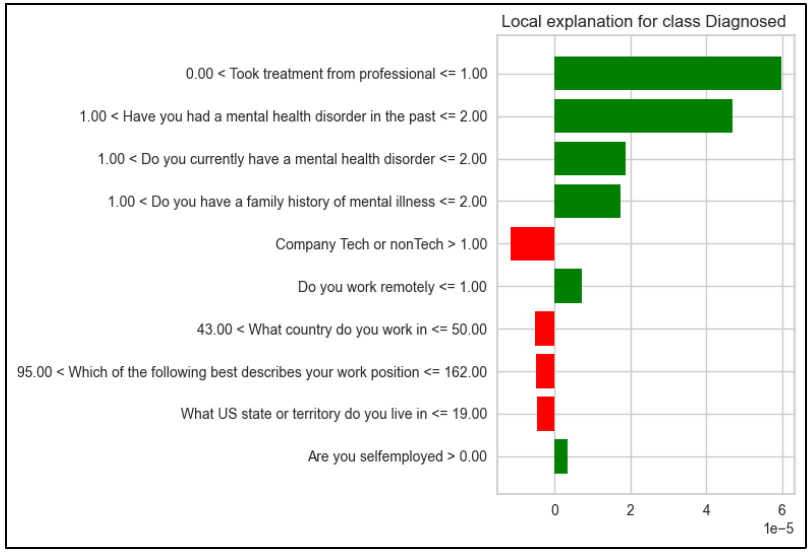





```
Intercept 1.002551749068273
Prediction_local [0.99295334]
Right: 7.666793e-06
```

The person who is not diagnosed with mental health issue because of the age group, age, state live in, state work in, work position, treatment, disorder in past.

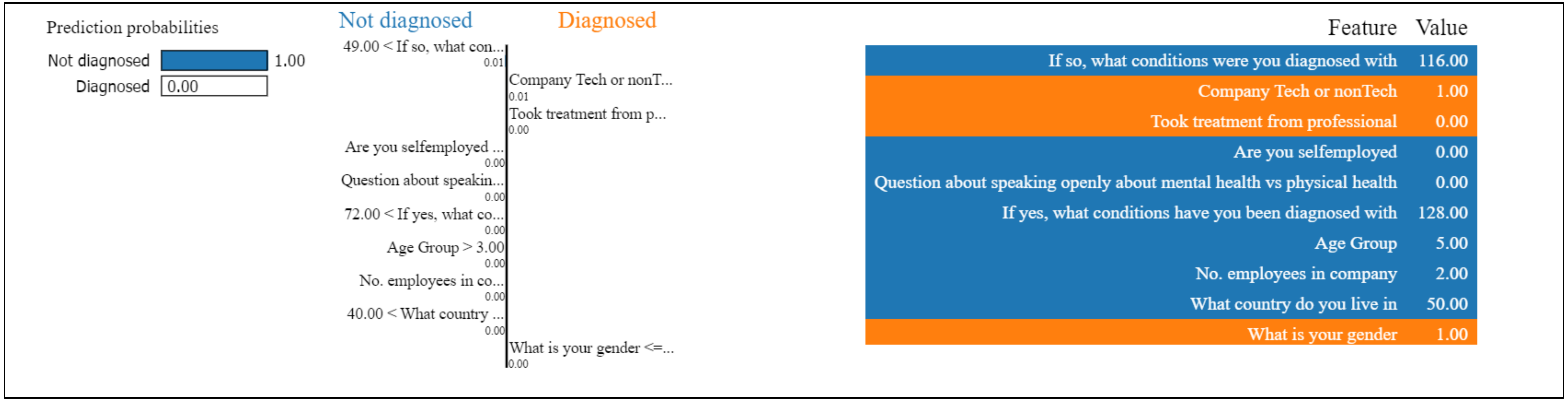


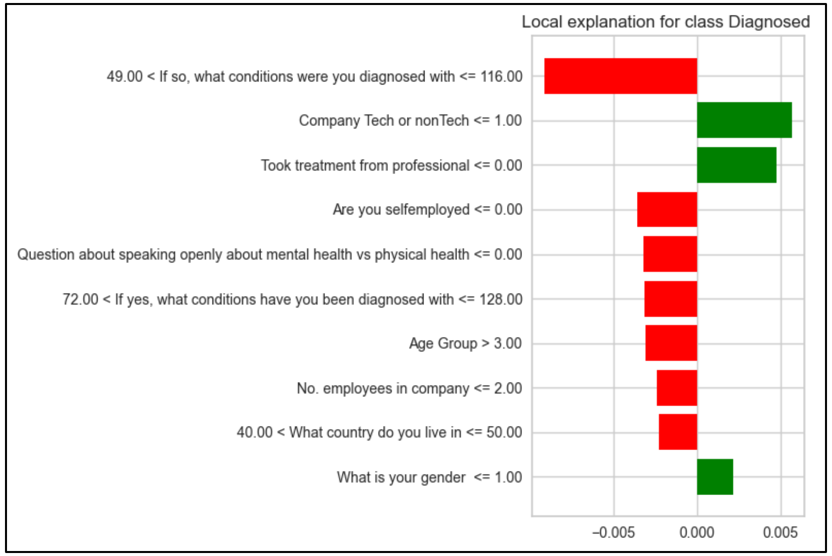



```
Intercept 0.9998614290528125
Prediction_local [0.99998832]
Right: 0.99998283
```

The person who is diagnosed with mental health issue because of the treatment, disorder in past, current mental health, family history of mental illness, work remotely, self-employed.

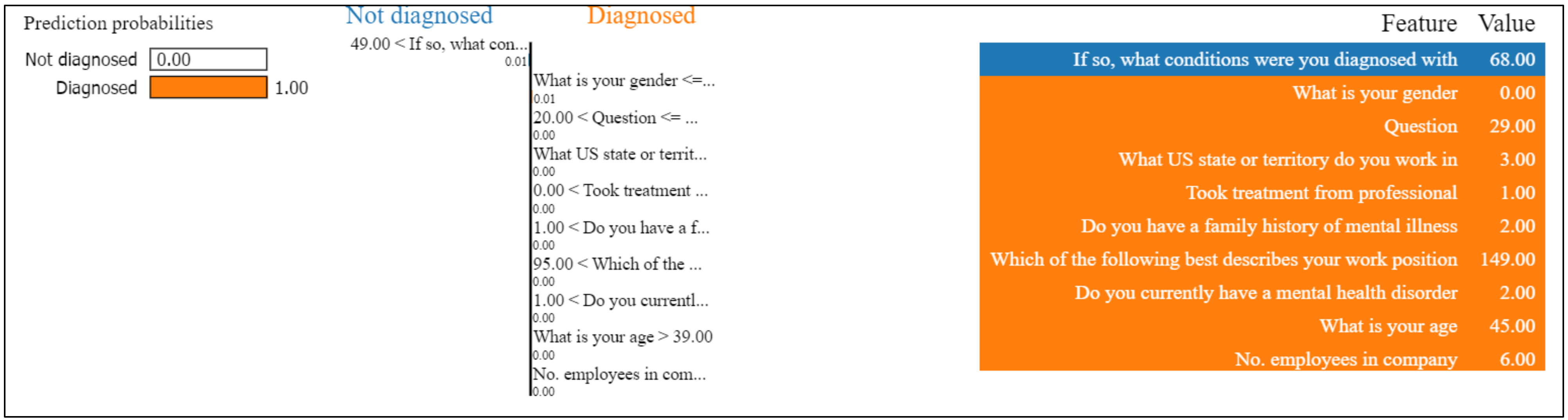
Gradient Boosting with LIME

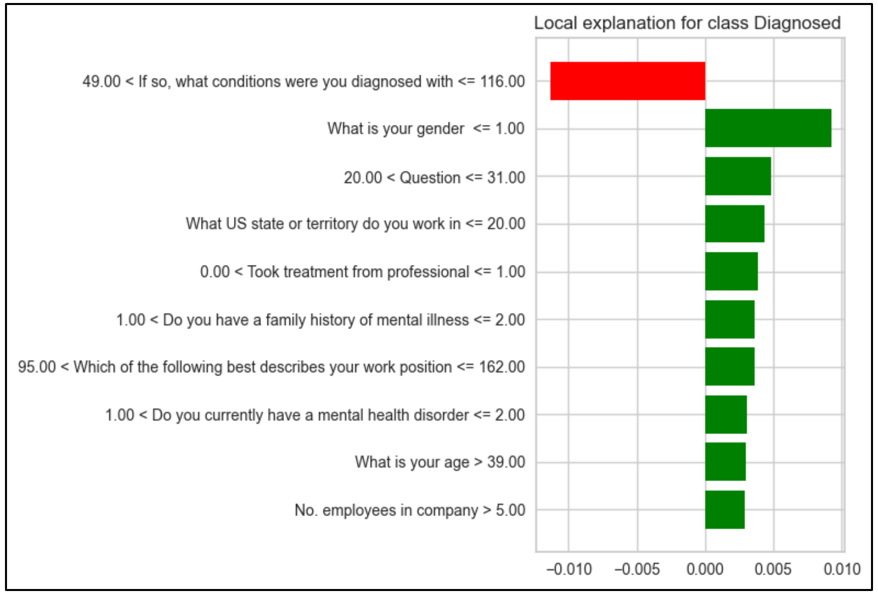





```
Intercept 1.0019850455127222
Prediction_local [0.98785155]
Right: 0.00020206759420826l8
```

The person who is not diagnosed with mental health issue because of the condition, self-employed, if yes then the condition, age group, country live in.






```
Intercept 0.9989980112533737
Prediction_local [0.99486058]
Right:  0.9997293265309739
```

The person who is diagnosed with mental health issue because of the gender, state work in, age, treatment, disorder in past, current mental health, family history of mental illness, work position.

PERFORMANCE ANALYSIS

MODEL	RESEARCH PAPER ACCURACY	IMPROVED ACCURACY
Logistic Regression	83.4	99.0
K - Nearest Neighbours	83.2	100
Decision Tree	76.4	99.6
Random Forest	83.9	100
Ensemble Technique	90.5	100
AdaBoost	88.3	100
XGBoost	93.4	100
Gradient boost classifier	93.9	100

- Data is normally distributed and model score for the train and test is 100%,the most of the data is categorical and bool so model accuracy can be 100%

```
# print the scores on training and test set
print('Training set score: {:.4f}'.format(logreg.score(X_train, y_train)))
print('Test set score: {:.4f}'.format(logreg.score(X_test, y_test)))
```

```
Training set score: 0.9895
Test set score: 0.9901
```

```
# print the scores on training and test set
print('Training set score: {:.4f}'.format(knn.score(X_train.values, y_train)))
print('Test set score: {:.4f}'.format(knn.score(X_test.values, y_test)))
```

```
C:\Users\megha\AppData\Roaming\Python\Python311\site-packages\sklearn\base.py:465:
  warnings.warn(
Training set score: 1.0000
C:\Users\megha\AppData\Roaming\Python\Python311\site-packages\sklearn\base.py:465:
  warnings.warn(
Test set score: 1.0000
```

```
# print the scores on training and test set
print('Training set score: {:.4f}'.format(clf_gini.score(X_train, y_train)))
print('Test set score: {:.4f}'.format(clf_gini.score(X_test, y_test)))
```

```
Training set score: 0.9974
Test set score: 0.9967
```

```
# print the scores on training and test set
print('Training set score: {:.4f}'.format(rfc.score(X_train, y_train)))
print('Test set score: {:.4f}'.format(rfc.score(X_test, y_test)))
```

```
Training set score: 1.0000
Test set score: 1.0000
```

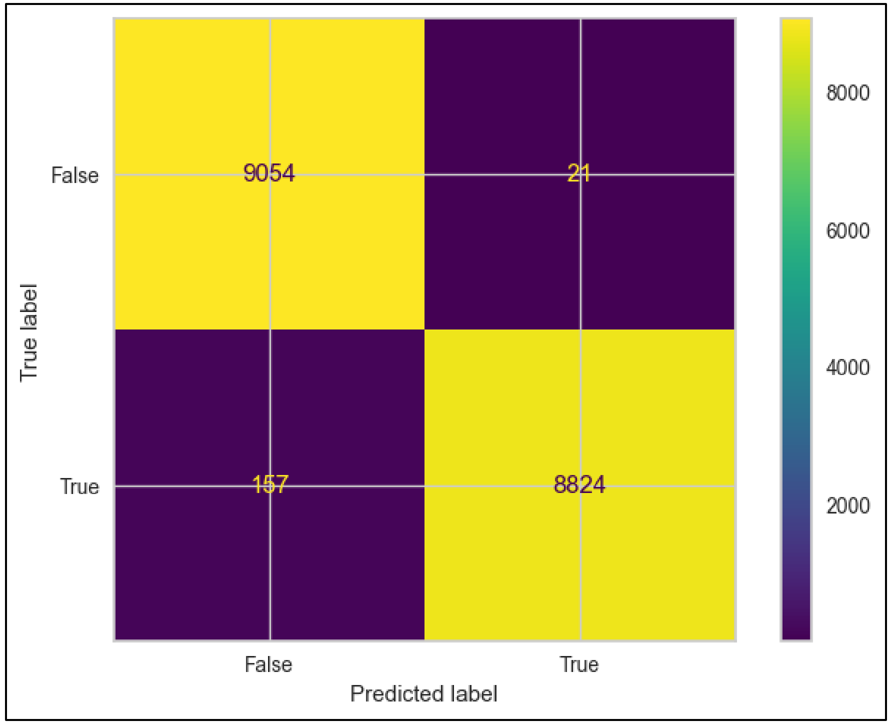
```
# print the scores on training and test set
print('Training set score: {:.4f}'.format(ensemble_model.score(X_train, y_train)))
print('Test set score: {:.4f}'.format(ensemble_model.score(X_test, y_test)))
```

```
Training set score: 1.0000
Test set score: 1.0000
```

```
# print the scores on training and test set
print('Training set score: {:.4f}'.format(xgb_model.score(X_train.values, y_train)))
print('Test set score: {:.4f}'.format(xgb_model.score(X_test.values, y_test)))
```

```
Training set score: 1.0000
Test set score: 1.0000
```

Logistic Regression

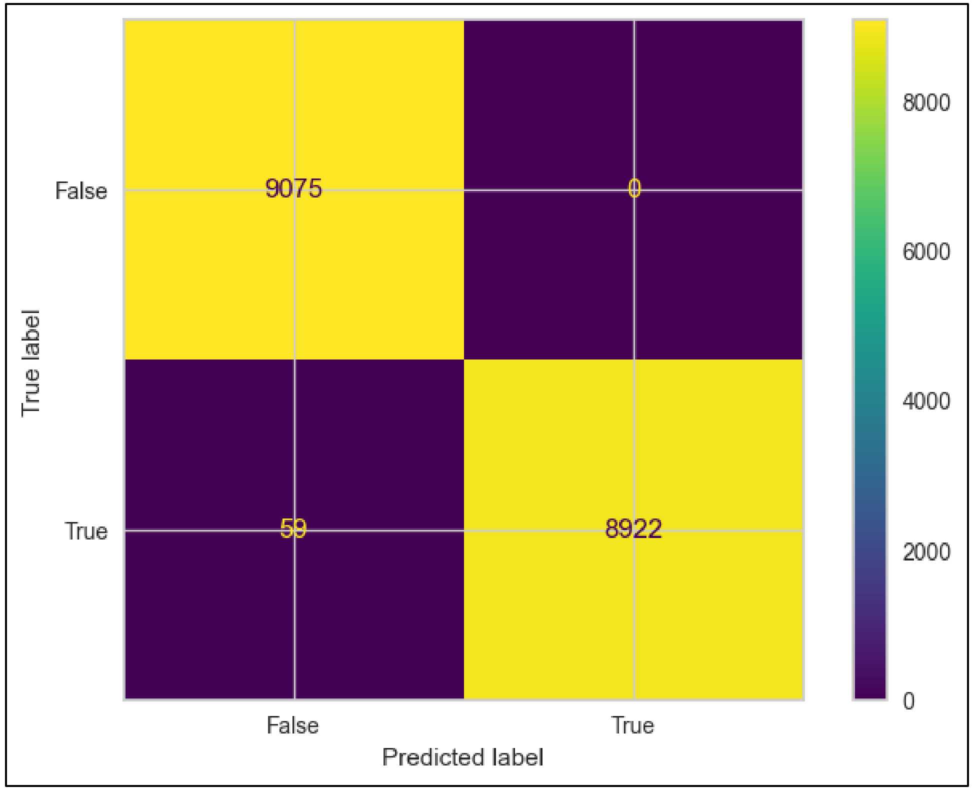


	precision	recall	f1-score	support
0	0.98	1.00	0.99	9075
1	1.00	0.98	0.99	8981
accuracy			0.99	18056
macro avg	0.99	0.99	0.99	18056
weighted avg	0.99	0.99	0.99	18056

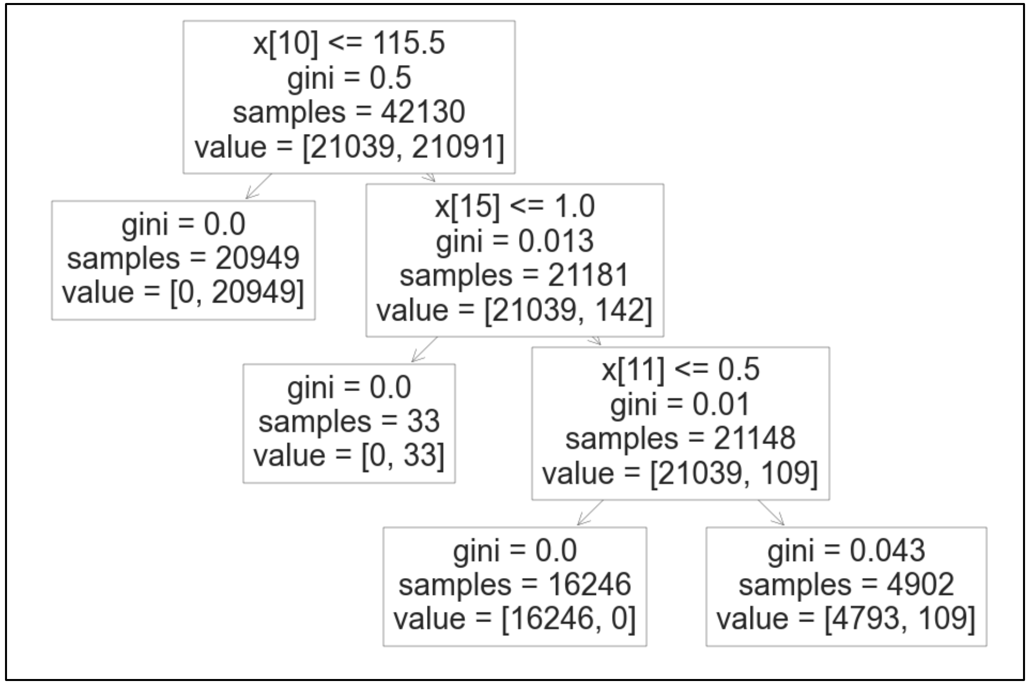
157 true labels and 8824 true predictions, which means that the model correctly classified 98.2% of the true instances.

9054 false labels and 21 true predictions, which means that the model incorrectly classified 1.8% of the false instances.

Decision Tree

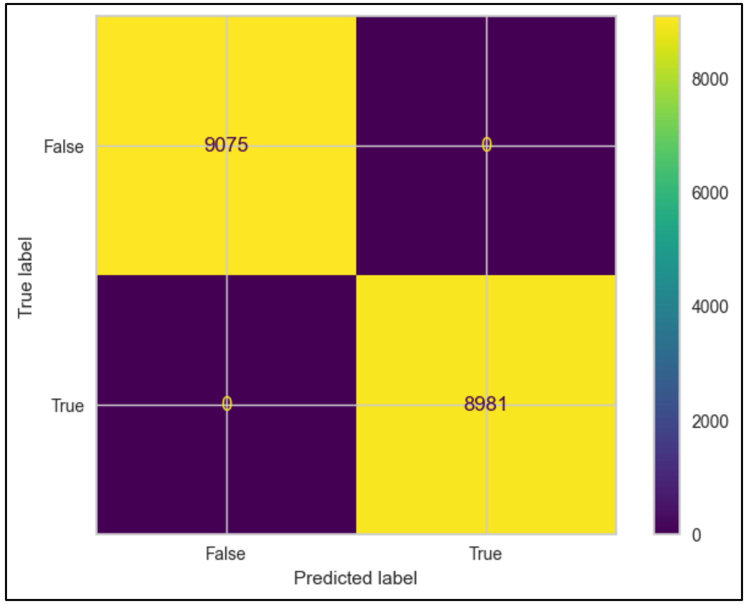


	precision	recall	f1-score	support
0	0.99	1.00	1.00	9075
1	1.00	0.99	1.00	8981
accuracy			1.00	18056
macro avg	1.00	1.00	1.00	18056
weighted avg	1.00	1.00	1.00	18056



The tree starts at the root node, which asks the question "x[10] <= 115.5?". If the answer is yes, the data is classified as value = [21039, 21091]. If the answer is no, the data is passed to the next node, which asks the question "x[15] <= 1.0?".
This process continues until all of the data has been classified into one of the two categories.

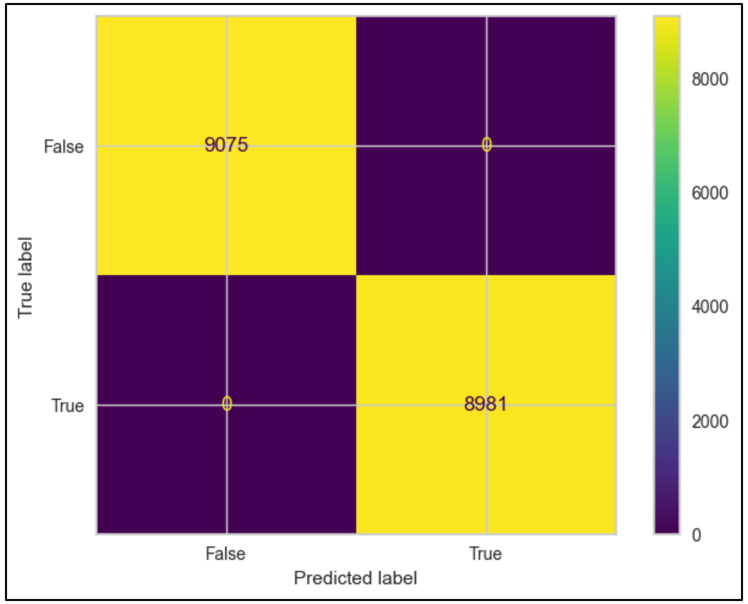
Random Forest



	precision	recall	f1-score	support
0	1.00	1.00	1.00	9075
1	1.00	1.00	1.00	8981
accuracy			1.00	18056
macro avg	1.00	1.00	1.00	18056
weighted avg	1.00	1.00	1.00	18056

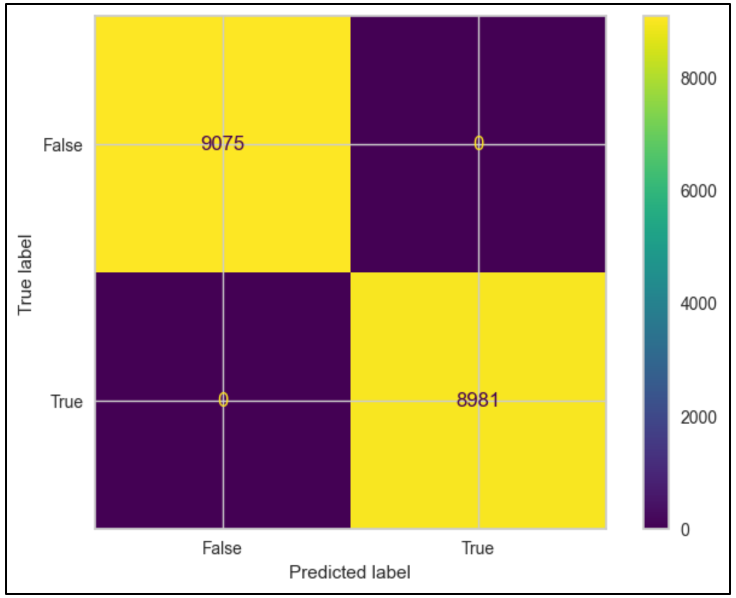
8981 are predicted positive and it’s true which is true positive, 9075 are predicted negative and it’s true which is true negative, 0 false Positive (Type 1 Error) and False Negative (Type 2 Error)

K - Nearest Neighbours



	precision	recall	f1-score	support
0	1.00	1.00	1.00	9075
1	1.00	1.00	1.00	8981
accuracy			1.00	18056
macro avg	1.00	1.00	1.00	18056
weighted avg	1.00	1.00	1.00	18056

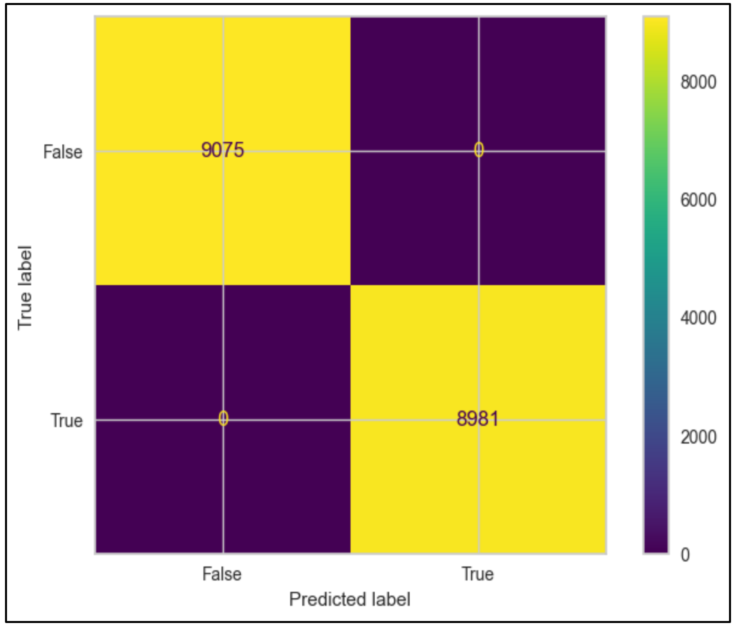
Ensemble technique (KNN, Random forest, Decision tree)



	precision	recall	f1-score	support
0	1.00	1.00	1.00	9075
1	1.00	1.00	1.00	8981
accuracy			1.00	18056
macro avg	1.00	1.00	1.00	18056
weighted avg	1.00	1.00	1.00	18056

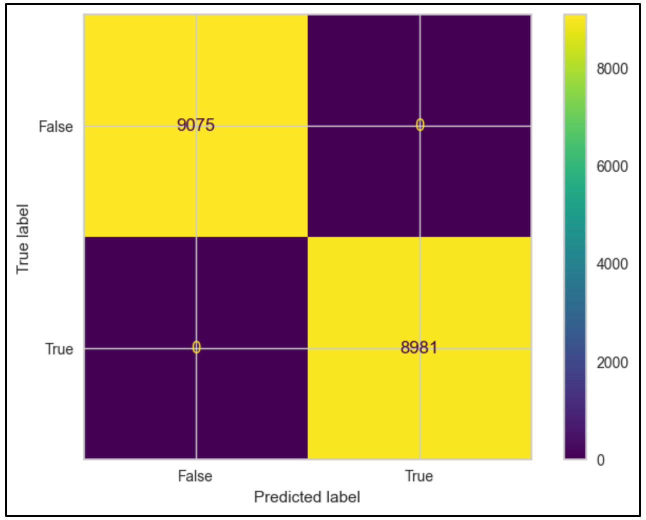
8981 are predicted positive and it's true which is true positive, 9075 are predicted negative and it's true which is true negative, 0 false Positive (Type 1 Error) and False Negative (Type 2 Error)

XGBoost



	precision	recall	f1-score	support
0	1.00	1.00	1.00	9075
1	1.00	1.00	1.00	8981
accuracy			1.00	18056
macro avg	1.00	1.00	1.00	18056
weighted avg	1.00	1.00	1.00	18056

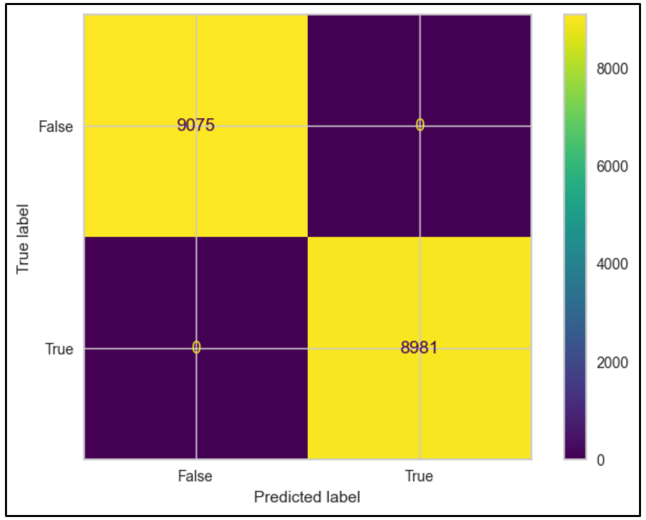
AdaBoost



	precision	recall	f1-score	support
0	1.00	1.00	1.00	9075
1	1.00	1.00	1.00	8981
accuracy			1.00	18056
macro avg	1.00	1.00	1.00	18056
weighted avg	1.00	1.00	1.00	18056

8981 are predicted positive and it’s true which is true positive, 9075 are predicted negative and it’s true which is true negative, 0 false Positive (Type 1 Error) and False Negative (Type 2 Error)

Gradient Boosting



	precision	recall	f1-score	support
0	1.00	1.00	1.00	9075
1	1.00	1.00	1.00	8981
accuracy			1.00	18056
macro avg	1.00	1.00	1.00	18056
weighted avg	1.00	1.00	1.00	18056

CONCLUSION

- The project showcased the use of AI and machine learning techniques, specifically XGBoost with LIME interpretability, to predict mental health trends in the tech industry.

- The model predicts individuals without mental health issues based on demographics and work-related details, including age, residence, work position, and past treatment history.

- The model accurately identifies individuals with mental health issues based on their specific history and circumstances, considering factors such as treatment, past disorders, family history, remote work, and self-employment.

- These findings provide valuable information for stakeholders within the tech industry to develop and create a more supportive and healthy work environment. Additionally, the interpretability provided by LIME allows for greater understanding of the model's predictions and facilitates trust and transparency in its application.