# MENTAL HEALTH PREDICTION

A Project Report

Presented To

Prof. Pramod Gupta

Prepared By:

Sree Swetha Gottumukkala (002924968)

Aamrah Aamrah (002109107)

Kshitija Laware (002984668)

Ghanashyam Vibhandik (002984660)

July 2022

## CONTENTS

## ABSTRACT

We really care about the mental well-being of employees at their workplace as it can make or break their lives. We wanted to explore surveys which would help us understand what factors, both personal and work-related drive a person to require mental healthcare. Features like leaves, benefits, supervisor, coworkers play a significant role in determining the state of mind of the employees. Through this research we wanted to understand what factors play a key role in determining a healthy work environment and find out what needs to be improved to ensure this happens. This research gave us a chance to dive deep into how well employers take care of their employees as well. We have used the knowledge acquired in this course to derive our own conclusions with models we found best suited and methods and techniques to improvise the data.

## 1. Objective

Based on a mental health survey, we aim to find which personal and work-related attributes are the top predictors in determining whether a person will seek treatment for a mental health condition. We used random forest, decision tree, SVM and KNN models for this and created a model that predicts whether a person will require treatment.

## 2. Introduction

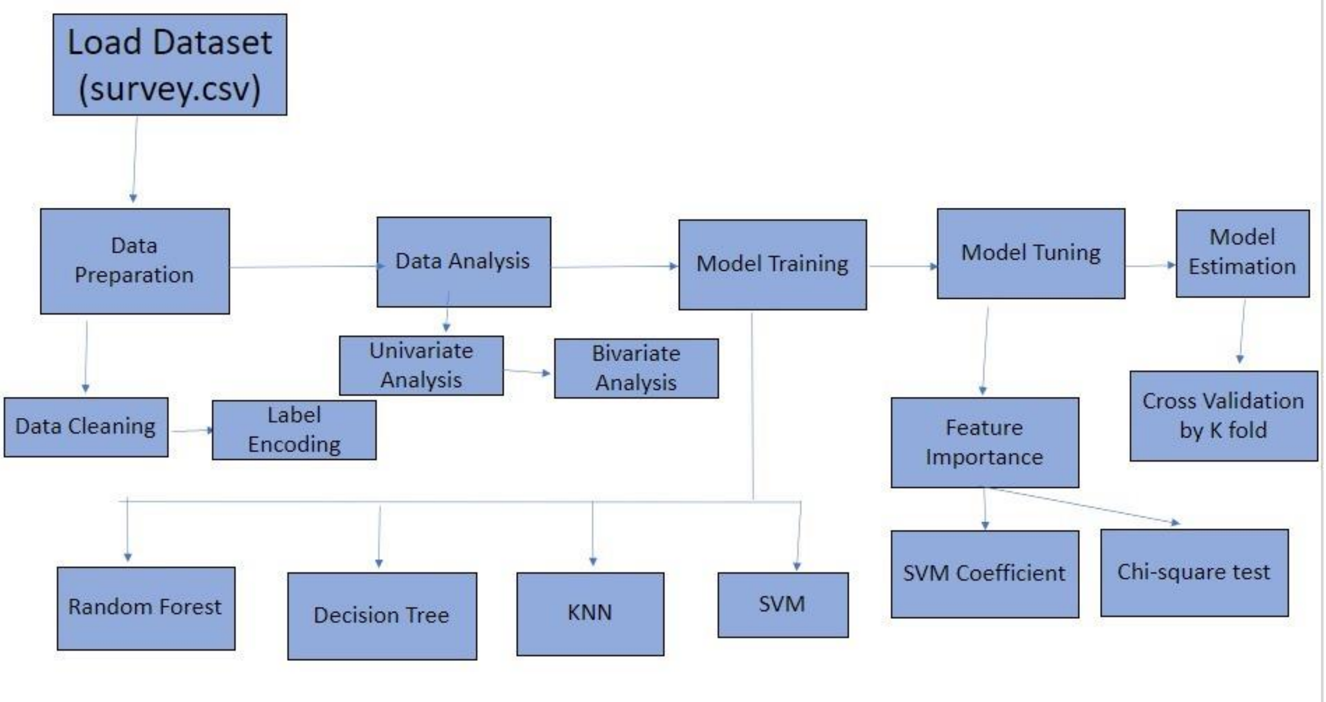
Our dataset is from a 2014 survey by OSMI, LLC that measures attitude towards mental health and frequency of mental health disorders in the tech workplace. It contains information about employees of companies from 48 countries. The dataset has 26 features in total amongst which 23 categorical columns with Yes/No questions that in general seek to find out. Our target variable is the Treatment feature.

**Feature Description**

Timestamp: When the record was submitted
Age: Age of the employee
Country: Country the employee belongs to
Gender: Gender of employee
state: If you live in the United States, which state or territory do you live in?
self_employed: Are you self-employed?
family_history: Do you have a family history of mental illness?
treatment: Have you sought treatment for a mental health condition?
work_interfere: If you have a mental health condition, do you feel that it interferes with your work?
no_employees: How many employees does your company or organization have?
remote_work: Do you work remotely (outside of an office) at least 50% of the time?
tech_company: Is your employer primarily a tech company/organization?
benefits: Does your employer provide mental health benefits?
care_options: Do you know the options for mental health care your employer provides?

wellness_program: Has your employer ever discussed mental health as part of an employee wellness program?
seek_help: Does your employer provide resources to learn more about mental health issues and how to seek help?
anonymity: Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?
leave: How easy is it for you to take medical leave for a mental health condition?
mentalhealthconsequence: Do you think that discussing a mental health issue with your employer would have negative consequences?
physhealthconsequence: Do you think that discussing a physical health issue with your employer would have negative consequences?
coworkers: Would you be willing to discuss a mental health issue with your coworkers?
supervisor: Would you be willing to discuss a mental health issue with your direct supervisor(s)?
mentalhealthinterview: Would you mention a mental health issue with a potential employer in an interview?
physhealthinterview: Would you mention a physical health issue with a potential employer in an interview?
mentalvsphysical: Do you feel that your employer takes mental health as seriously as physical health?
obs_consequence: Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?
comments: Any additional notes or comments

## 3. Program Steps

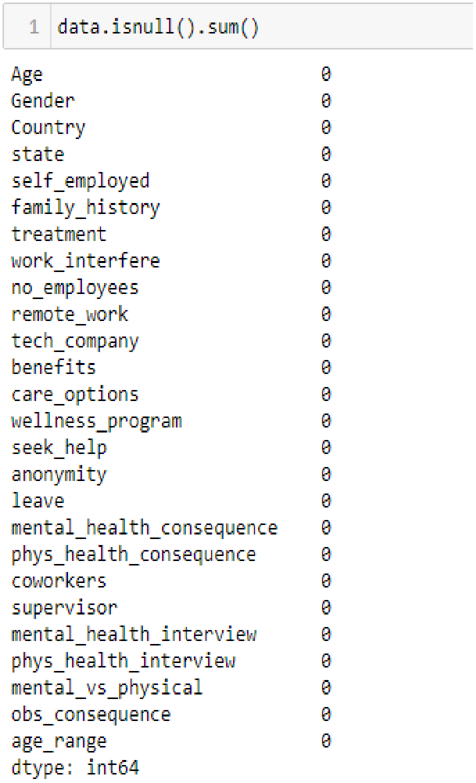


Dataset Link - https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey

# 4. Data Preparation

## 4.1 Data Cleaning

We dropped the columns Comments and Timestamp because they are irrelevant to our prediction. We found state has more than 50% null values. So, we choose not to consider state column while model training. There are null values which we processed according to the datatype of columns to the datatype of columns. Since most of our columns are categorical, we reduced the groups in columns by categorizing the values which are spelling mistakes/synonyms/acronyms to standard values. At the end we dropped 4 duplicate records in our dataset.

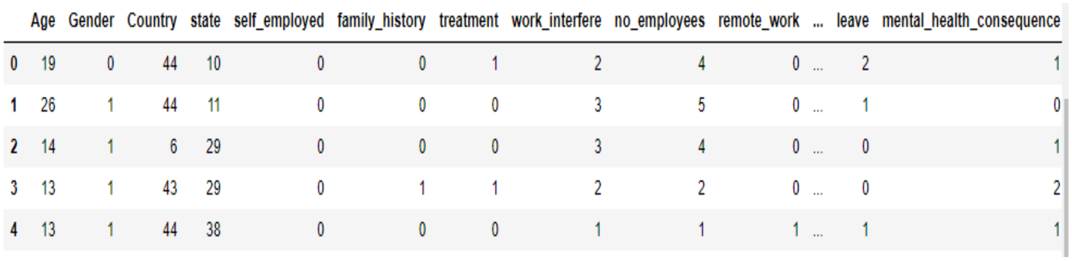


```
data.isnull().sum()
```

```
Age                          0
Gender                       0
Country                      0
state                        0
self_employed                0
family_history               0
treatment                    0
work_interfere               0
no_employees                 0
remote_work                  0
tech_company                 0
benefits                     0
care_options                 0
wellness_program             0
seek_help                    0
anonymity                    0
leave                        0
mental_health_consequence    0
phys_health_consequence      0
coworkers                    0
supervisor                   0
mental_health_interview      0
phys_health_interview        0
mental_vs_physical           0
obs_consequence              0
age_range                    0
dtype: int64
```

## 4.2. Label Encoding

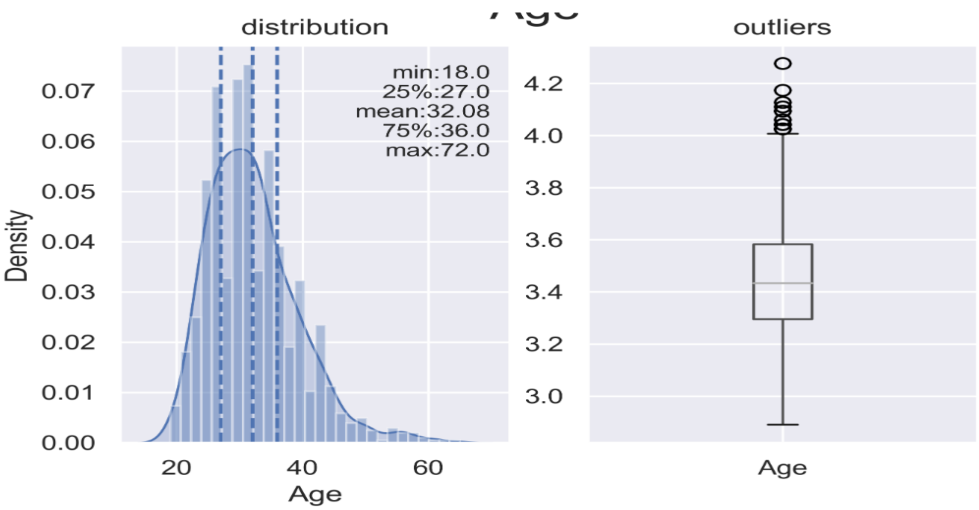
It refers to converting the labels into a numeric form to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. We encode the data for better understanding and checked whether data should be categorized or has too many numeric values.



| | Age | Gender | Country | state | self_employed | family_history | treatment | work_interfere | no_employees | remote_work | ... | leave | mental_health_consequence |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 19 | 0 | 44 | 10 | 0 | 0 | 1 | 2 | 4 | 0 | ... | 2 | 1 |
| 1 | 26 | 1 | 44 | 11 | 0 | 0 | 0 | 3 | 5 | 0 | ... | 1 | 0 |
| 2 | 14 | 1 | 6 | 29 | 0 | 0 | 0 | 3 | 4 | 0 | ... | 0 | 1 |
| 3 | 13 | 1 | 43 | 29 | 0 | 1 | 1 | 2 | 2 | 0 | ... | 0 | 2 |
| 4 | 13 | 1 | 44 | 38 | 0 | 0 | 0 | 1 | 1 | 1 | ... | 1 | 1 |

## 5. Data Analysis

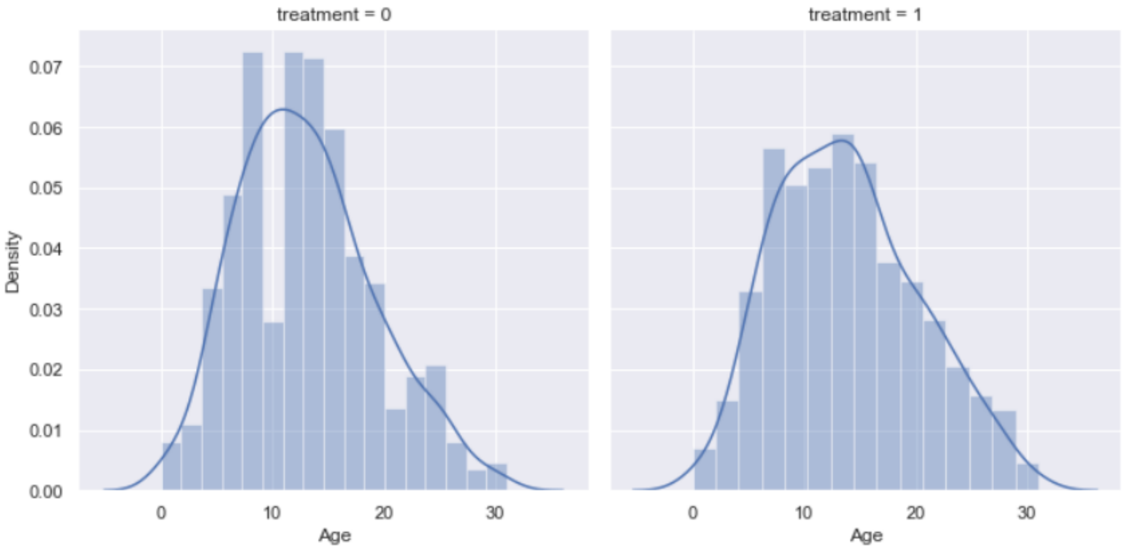
### 5.3.1. Univariate Analysis



From univariate analysis we found outliers in Age which we processed.

### 5.3.2. Bivariate Analysis

We did bivariate analysis of Age and Treatment to check for patterns. The distribution was in the form of bell curve.

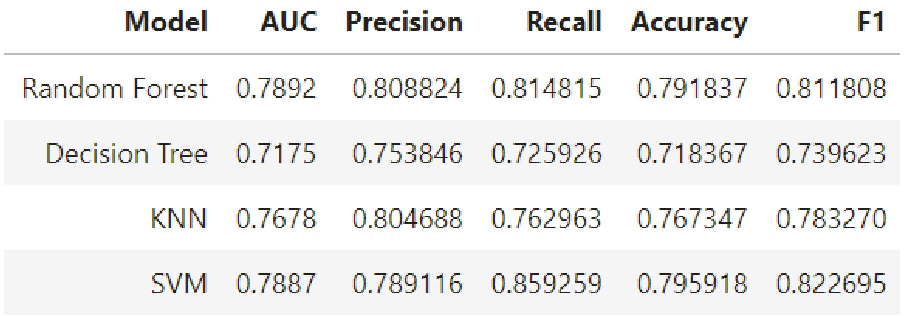


### 5.3.3 Scaling & Fitting

We did Min max scaling to numerical column ‘age’.

## 6. Model Training and Fitting

We used 4 machine learning models namely, Decision Tree, Random Forest, SVM, KNN for our classification problem to predict the treatment value. SVM gave the best accuracy considering all the features. The closest accuracy given is by random forest classifier.



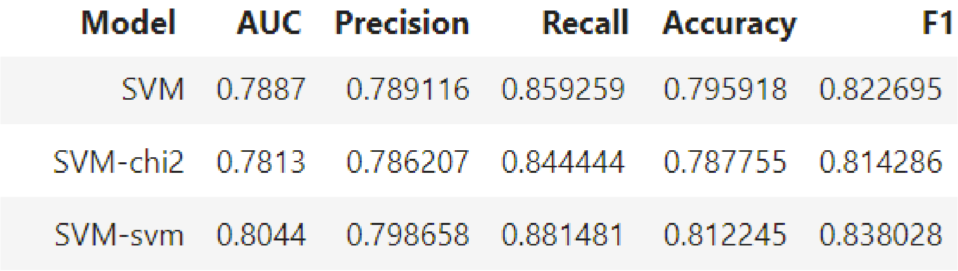
| Model | AUC | Precision | Recall | Accuracy | F1 |
|---|---|---|---|---|---|
| Random Forest | 0.7892 | 0.808824 | 0.814815 | 0.791837 | 0.811808 |
| Decision Tree | 0.7175 | 0.753846 | 0.725926 | 0.718367 | 0.739623 |
| KNN | 0.7678 | 0.804688 | 0.762963 | 0.767347 | 0.783270 |
| SVM | 0.7887 | 0.789116 | 0.859259 | 0.795918 | 0.822695 |

## 7. Model Tuning

We chose SVM to tune our model by selecting key features:  We used 2 ways to check feature importance.

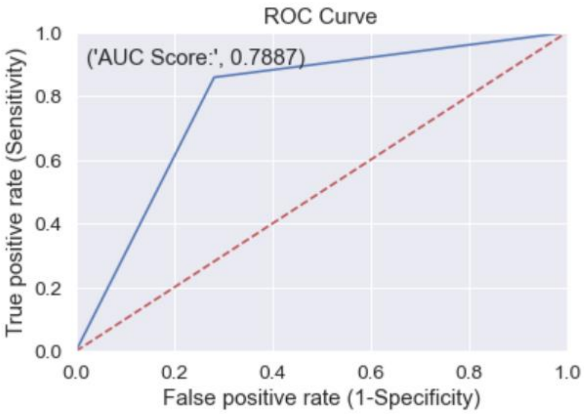
1. Comparing SVM coefficient values : The higher the coefficient value,  the important the feature is
2. Chi-square test : Our predictor and target features are categorical so we used chi-sqaure test as it is best used for these kind of problems.

The model gave better accuracy when we selected 14 most key features based on SVM coefficients. F1 increased by 1.5% with selected features.

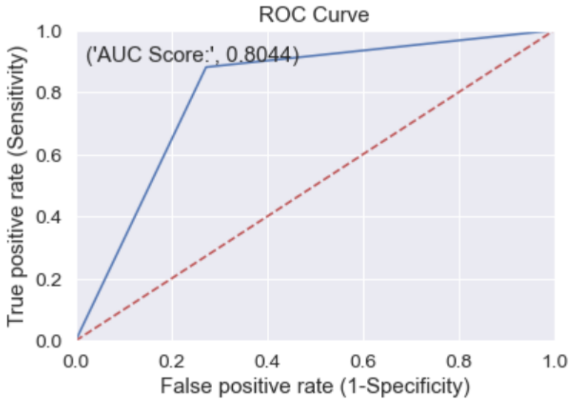


| Model | AUC | Precision | Recall | Accuracy | F1 |
|---|---|---|---|---|---|
| SVM | 0.7887 | 0.789116 | 0.859259 | 0.795918 | 0.822695 |
| SVM-chi2 | 0.7813 | 0.786207 | 0.844444 | 0.787755 | 0.814286 |
| SVM-svm | 0.8044 | 0.798658 | 0.881481 | 0.812245 | 0.838028 |

Our key features as per SVM Coefficients are:
Age, Gender, care options, work interference, family history, number of employees, coworkers, benefits, anonymity, leave, seek help, mental health consequences, wellness program, country.



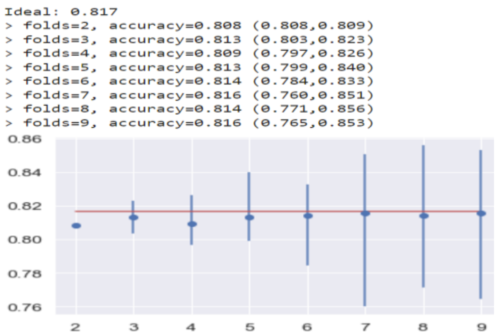
```
Accuracy with reduced features -14
```



## 8. Model Estimation

We used k-fold cross validation to estimate our model.

Model Accuracy based on - fold cross validation fitted on SVM model with notable features gave accuracy for most of the folds close to the mean classification accuracy from LOOCV which is 0.817.



```
Ideal: 0.817
> folds=2, accuracy=0.808 (0.808,0.809)
> folds=3, accuracy=0.813 (0.803,0.823)
> folds=4, accuracy=0.809 (0.797,0.826)
> folds=5, accuracy=0.813 (0.799,0.840)
> folds=6, accuracy=0.814 (0.784,0.833)
> folds=7, accuracy=0.816 (0.760,0.851)
> folds=8, accuracy=0.814 (0.771,0.856)
> folds=9, accuracy=0.816 (0.765,0.853)
```

**Test Cases**

**Case 1**
Age: 23, care_options: Yes, Gender: Trans, no_employees: 26-100, Country: Belgium, family_history: Yes, coworkers: Yes, benefits: No, anonymity: YES, seek_help: YES, mental_health_consequence: NO, wellness_program: YES, leave: DONT know, work_interfere: Never

Output: No

**Case 2**
Age: 28, care_options: Not sure, Gender: Male, no_employees: 6-25, Country: USA, family_history: Yes, coworkers: Yes, benefits: No, anonymity: NO, seek_help: No, mental_health_consequence: Yes, wellness_program: Dont know, leave: Dont know, work_interfere: Never

Output: Yes

## 9. Conclusion

Firstly, we built 4 models and compared their metrics namely, KNN, SVM, Decision Tree and Random Forest to predict "Treatment". From model evaluation SVM model gave high accuracy of 79.59% and F1 score of 82.2% when all features are considered. Random forest gave accuracy of 79.1% close to this.

We conclude that the SVM model performs best for our problem statement. So, for this we made further model tuning by selecting 14 key features using SVM coefficients and after model refitting with selected feature set, we got accuracy of 81.2% and F1 of 83.8%. While performance estimation: of model: Leave-one-out method gave us an ideal test condition of **81.7%**, k-fold cross validation method matched the ideal case with an accuracy of **81.6%** with our model.

An employer can use SVM classifier with the following 14 notable features to predict if employees in their company would need to be treated for mental health and take actions to improve employee's level of satisfaction at their firm; by making changes to work related conditions and providing treatment at the right time to potential ailments.

Important features: Age, Gender, care options, work interference, family history, number of employees, coworkers, benefits, anonymity, leave, seek help, mental health consequences, wellness program, country.

## 10. References

- https://towardsdatascience.com/hands-on-python-data-visualization-seaborn-count-plot-90e823599012
- https://matplotlib.org/stable/gallery/lines_bars_and_markers/bar_stacked.html
- https://www.geeksforgeeks.org/ml-chi-square-test-for-feature-selection/?ref=lbp
- https://careerfoundry.com/en/blog/data-analytics/how-to-find-outliers/
- https://machinelearningmastery.com/how-to-configure-k-fold-cross-validation/