ROBERT H. SMITH SCHOOL OF BUSINESS

BUDT758T: Data Mining & Predictive Analytics

Project Title: MENTAL HEALTH CLASSIFICATION MODEL USING ML

Team Members: Drishti Parekh, Mina(Hsin-Yuan) Yen, Nigam Shah, Tanmaya Kompella, Yi-Wei Sun

ORIGINAL WORK STATEMENT

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

Typed Name	Signature
Drishti Sanjay Parekh	D S P
Hsin-Yuan Yen	Mina Y.
Nigam Shailesh Shah	Nigam S. Shah
Sesha Tanmaya Kompella	Tanmaya K.
Yi-Wei Sun	Yi-Wei Sun

MENTAL HEALTH CLASSIFICATION MODEL USING ML

I. Executive Summary

The primary goal of our project was to classify mental illness in employees working in the tech industry. In today's fast-paced world where hustling is the new norm, it is now more important than ever to not overlook mental illness and to assign appropriate importance and resources to it. In order to achieve the same, we should be able to accurately predict/identify risk factors for poor mental health. We started off by putting together, cleaning and exploring the datasets. Then we proceeded to test out several models with variations of the independent variables to try and improve accuracy. Eventually, we were able to detect the presence of poor mental health with high accuracy and low error rates.

II. Data Description

In this study, we aim to develop data mining models for the classification of the risk for poor mental health. We will be focussing on people working in tech roles/industry. The models will be developed with datasets obtained from OSMH/OSMI Mental Health in Tech Survey. This raw dataset spans 5 years from 2017 to 2021 and has 124 variables including demographics (age and gender), geography (country and state of residence), working in a tech company, seeking help in the workplace, and presence of any treatment. Out of 124 variables, we will be using 21 variables, which we found worth including in our study, as mentioned below:

Dependent variable:
MH_disorder (whether this person currently has mental health disorder)
Independent variables:
Personal Information (7):
age, gender, country, past_disorder, family_history, treatment, IT_related. (currently doing IT related positions)

Working Company Information & Provided Healthcare Resources (13):
company_size, HC_coverage (HealthCare coverage), medical_leave (for mental health issue), PH_importance (physical health), MH_importance (mental health), observation, unsupported_space, supported_space, tech_support (how well the tech companies support MH issues), coverage_option, formal_discussion, resources, employer_discussion.

III. Research Questions

According to the research conducted by McKinsey, more than half of the population of middle and high-income countries are likely to suffer from at least one form of mental disorder during their lifetime. Since the start of the COVID-19 pandemic, the stats have only gotten significantly worse. We hope to accurately predict poor/declining mental health of employees by establishing a relationship between different attributes describing employee's mental health conditions and their work life. The goal is to answer the following questions:

- Develop data mining models for the classification of the risk for poor mental health
    - Which factors are related to having poor mental health?
- Focus on people working in tech roles/industry
    - Are these people more likely to have mental health problems?
- Analyze relationship between company and employee's mental health
    - Do employees working in companies providing healthcare resources have better mental health?

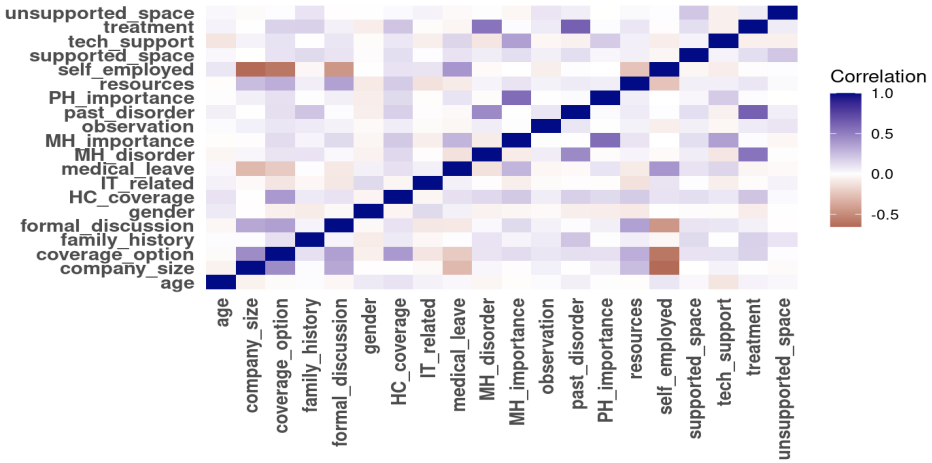
IV. Data Preparation

Initially we had 124 variables which most likely would have resulted in overcomplicated models. Our aim was to develop a simple yet effective model with good accuracy rates and so we started off by identifying the most relevant 21 variables to incorporate into our models. After that, we followed the following data cleaning steps:

- Cleaned the data by trimming down the long column names to improve readability
- Replaced the null values with the mean where required, removed columns with more than 80% null values
- Categorized survey answers for each column/question to make the data uniform
- Converted some important features/variables to factor like HC_coverage, IT_related, past_disorder,etc
- Generated visualizations to gain valuable insight into the relationship between the variables (Exploratory analysis)

V. Data Exploration

1. Heatmap showing relationship between numerical columns : Past disorder and Treatment variables are highly correlated.



2. Health Care Coverage vs Mental Health Disorder
   Most Employees with mental health disorder had mental health benefits from their employers as healthcare coverage. However, many employees who struggled with poor mental health were not eligible for mental health care coverage.

VI. Methodology

The models we used for mental health classification are:

1. Logistic Regression
2. Random Forest
3. Boosting
4. KNN
5. Naive Bayes
6. Classification Tree
7. Clustering
8. Association Rule

After experimenting with different models and variables, we decided to build our models without the "country" variable. The main reason is that most of our models cannot handle text features, and we cannot treat it as a categorical feature since there are too many countries. Also, after sampling, the test data might have some countries that the train data does not have which leads to mismatch of levels. We have used 70/30 partitioning while splitting the train and test data from the dataset.

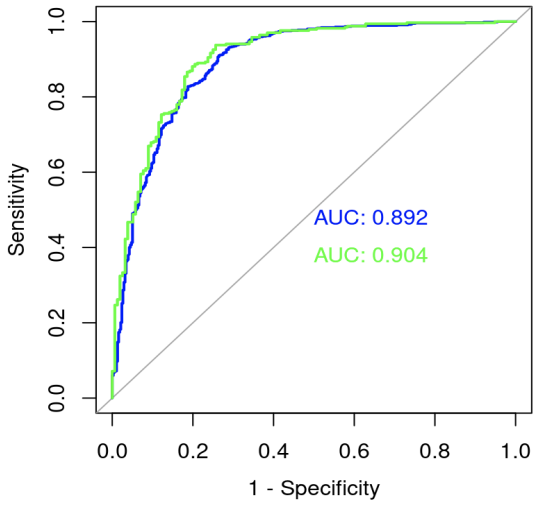
Logistic Regression

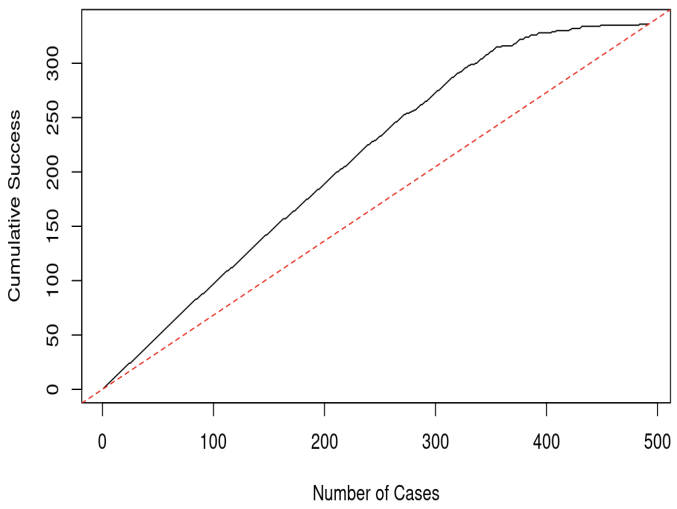
1. Logistic Regression (Using all the variables except 'country')

   Accuracy (test data): 0.871
   Plotting the ROC curve of both training and testing data, we find the AUC for test data is 0.898, which has a chance of improvement.

2. Logistic Regression (Using Backward Elimination method)

   Accuracy (test data) : 0.864
   Using Backward Elimination, we drop the features that do not have a greater effect on prediction of the dependent variable. This method helped us reduce the number of features from 49 to 17. However, the accuracy of this model was not very good compared to the model with all variables.
   We then decided to find the best cutoff value at which the test error rate of the model is the lowest. Previous models used a cutoff of 0.5. New cutoff that we got after plotting the test error rate vs cutoff plot at which the test error rate is the lowest is 0.44.
   At this cutoff value, accuracy of the model increased to 0.872. This model also does a better job at distinguishing between the positive and negative classes, which is indicated by the AUC value of test data at 0.904.
   As we know, lift is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model.The lift chart given below indicates that if, for example, we pick 300 random cases, it gives us 200 successes (baseline), but the classifier gives us 260 successes. Thus, the classifier is approximately 260/200 = 1.3 times better than picking 50 cases at random.

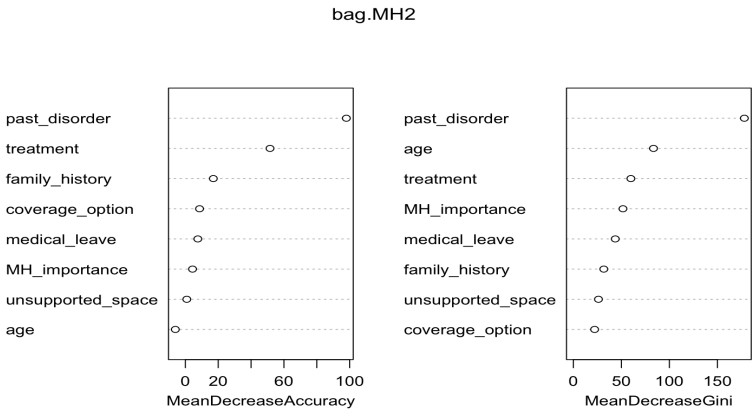




Random Forest

1. Random Forest (Using all the variables except countries)
   We applied the Random Forest algorithm by using all variables except countries as the independent variables. The resulting test accuracy is pretty high- 0.866 upon specifying mtry=5. Upon plotting the chart, we notice that past_disorder, treatment, and family_history are identified to be the most important variables.

2. Random Forest (Using variables from backward elimination - LogReg)
   In our attempts to see if the accuracy can be improved, we built a Random Forest algorithm using the variables identified from the backward elimination method in Logistic Regression. The resulting test accuracy has now decreased slightly 0.866 upon specifying mtry=5. Upon plotting the chart below, we notice that past_disorder, treatment, and family_history are identified to be the most important variables like before.



Boosting

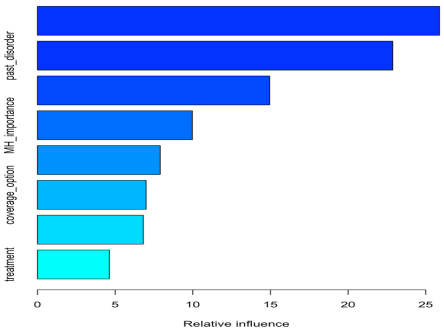
1. Boosting (Using all the variables except 'country')

   We then applied the Boosting algorithm by using all variables except countries as the independent variables. The resulting test accuracy is observed to be 0.854. We also observe that the importance of variables is similar to that of Random Forest.

2. Boosting (Using variables from backward elimination - LogReg)

   To further our attempts to improve accuracy, we built a Boosting algorithm using the variables identified from the backward elimination method in Logistic Regression. The resulting test accuracy has now decreased significantly to 0.821. We can say that what works for one model need not necessarily work for another. We observe that the importance of variables has noticeably changed in this case.

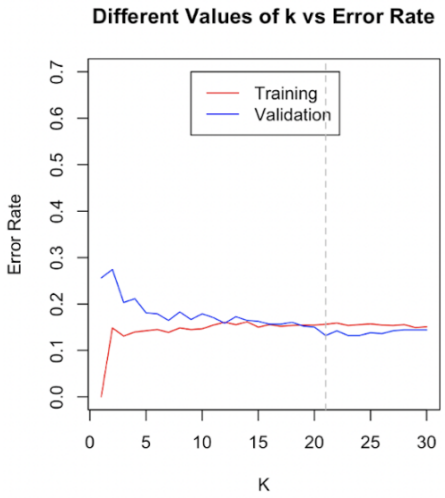
	var	rel.inf
age	age	25.881753
past_disorder	past_disorder	22.872369
medical_leave	medical_leave	14.945457
MH_importance	MH_importance	9.969750
family_history	family_history	7.901243
coverage_option	coverage_option	6.993030
unsupported_space	unsupported_space	6.811917
treatment	treatment	4.624480

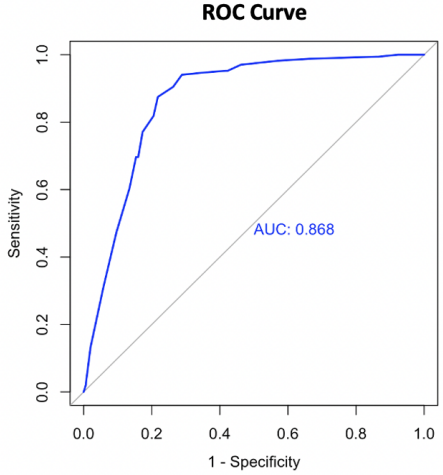


KNN

1. KNN (Using all the variables except 'country')

   We applied the K Nearest Neighbor algorithm by using all variables except countries as the independent variables. In our attempt to find the best k, we plotted different values of k vs the error rate to get 21 as the best k value. The resulting test accuracy is pretty high- 0.8678. Upon plotting the lift chart below, we notice that there is decent lift and the test AUC value from the ROC curve is also fairly high- 0.866.





2. KNN (Using variables from backward elimination - LogReg)

   In our attempts to see if the accuracy can be improved, we built a KNN algorithm using the variables identified from the backward elimination method in Logistic Regression. In our attempt to find the best k, we plotted different values of k vs the error rate to get 26 as the best k value. The resulting test accuracy decreased in this case- 0.8109. Upon plotting the lift chart, we notice that there is decent lift and the test AUC value from the ROC curve has decreased slightly- 0.853.

3. KNN (Using variables from Random Forest)

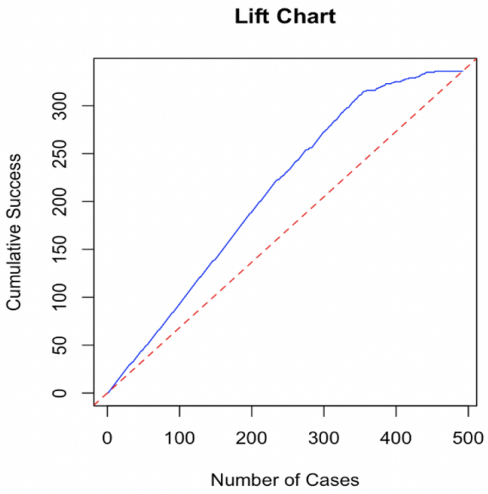
   To further our attempts to improve accuracy, we built a KNN algorithm using the top 8 variables identified from the Random Forest method. In our attempt to find the best k, we plotted different values of k vs the error rate to get 18 as the best k value. The resulting test accuracy is 0.8475. Upon plotting the lift chart, we notice that there is decent lift and the test AUC value from the ROC curve is 0.866.

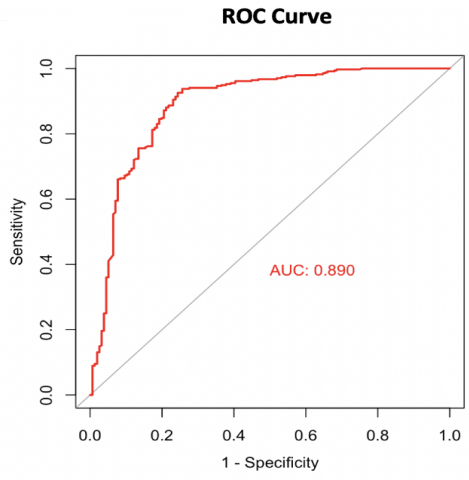
Naïve Bayes

1. Naïve Bayes (Using all variables except 'country')
   We applied Naïve Bayes algorithm by using all variables except countries as the independent variables. The resulting test accuracy is observed to be a decent 0.8434. Upon plotting the lift chart, we notice that there is decent lift and the test AUC value from the ROC curve is also fairly high- 0.889.

2. Naïve Bayes (Using variables from backward elimination - LogReg)
   In our attempts to see if the accuracy can be improved, we applied Naïve Bayes using the variables identified from the backward elimination method in Logistic Regression. The resulting test accuracy slightly increased in this case- 0.8495. Upon plotting the lift chart, we notice that there is decent lift and the test AUC value from the ROC curve has also increased- 0.896.

3. Naïve Bayes (Using variables from Random Forest)
   To further our attempts to improve accuracy, we applied Naïve Bayes using the top 8 variables identified from the Random Forest method. The resulting test accuracy is an improvement at 0.8516. Upon plotting the lift chart below, we notice that there is decent lift and the test AUC value from the ROC curve is fairly high- 0.890.

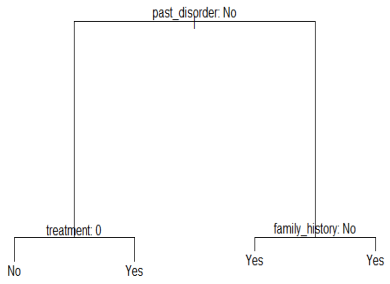




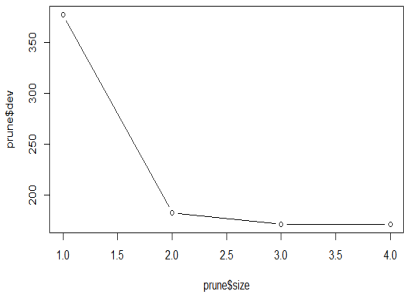
Classification Tree

It's a method for classification and profiling that is simple to interpret for non-technical people.We didn't train the model for different groups of variables since it makes no difference for this model. We also pruned the tree to avoid overfitting, but it doesn't make any difference since the original tree doesn't have many branches as the charts below.

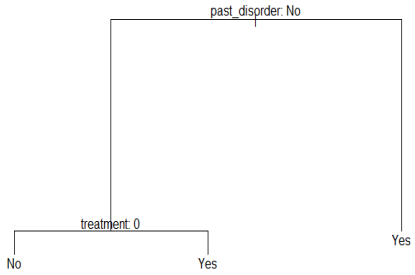
Original tree



Pruned size vs error rate



Pruned tree



Three branches will have the lowest error rate, so we pruned the tree with 3 nodes, and the accuracy for this model will be 86.78%

Clustering

Grouped similar objects together to simplify data for further learning and to establish prototypes. For our project we have used *KMeans Clustering* which is computationally fast for larger datasets like ours. We combined the number of people suffering from mental health disorder issues across different clusters formed using different sets of variables. However, the results were not that insightful to deduce something meaningful out of it.

Association Rule

Expanding from Clustering, we moved to Association Rules which is a study of "what goes with what". The main purpose of this is to categorize the set of common characteristics that people with mental health disorders might possess.

1) Using all variables except 'country':

Here, the rule with the highest support, lift and confidence can be translated into plain english as:
Employees having *somewhat difficulties in getting a medical leave, have a history of past mental health disorders, and have mental health benefits as part of their healthcare coverage* are more likely to have a *mental health disorder*.

2) Using variables from backward elimination - LogReg:

In this case, the rule with the highest support, lift and confidence can be translated into plain english as:
Employees having *somewhat difficulties in getting a medical leave, and have a personal and family history of past mental health disorders* are more likely to have a *mental health disorder*.

3) Using variables from Random Forest:

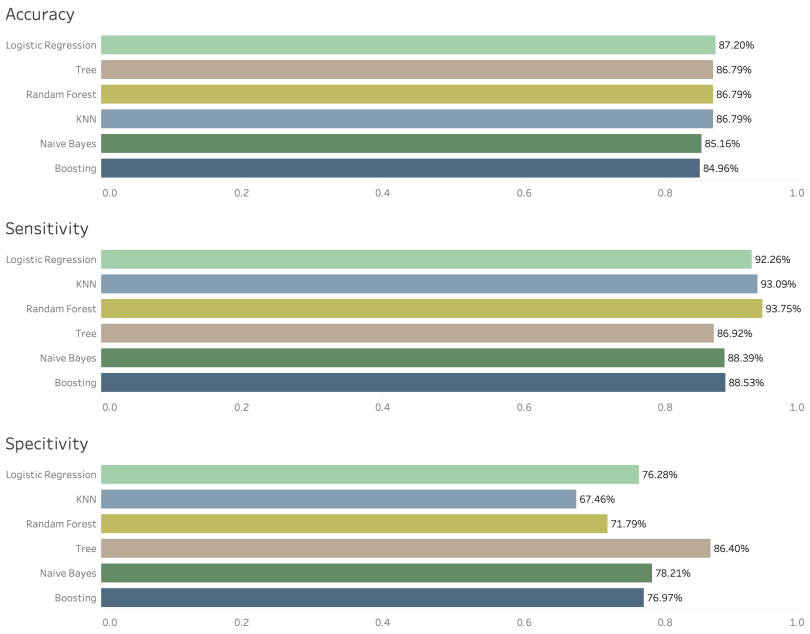
	lhs		rhs	support	confidence	coverage	lift	count
[1]	{HC_coverage_Yes, medical_leave_Somewhat difficult, past_disorder_Yes}	=>	{MH_disorder}	0.03543067	1	0.03543067	1.484134	58
[2]	{HC_coverage_Yes, medical_leave_Somewhat difficult, past_disorder_Yes, treatment_1}	=>	{MH_disorder}	0.03481979	1	0.03481979	1.484134	57
[3]	{medical_leave_Somewhat difficult, past_disorder_Yes, family_history_Yes}	=>	{MH_disorder}	0.03298717	1	0.03298717	1.484134	54
[4]	{medical_leave_Somewhat difficult, past_disorder_Yes, family_history_Yes, treatment_1}	=>	{MH_disorder}	0.03298717	1	0.03298717	1.484134	54
[5]	{HC_coverage_Yes, family_history_I don't know, unsupported_space_Maybe, treatment_1}	=>	{MH_disorder}	0.02810018	1	0.02810018	1.484134	46

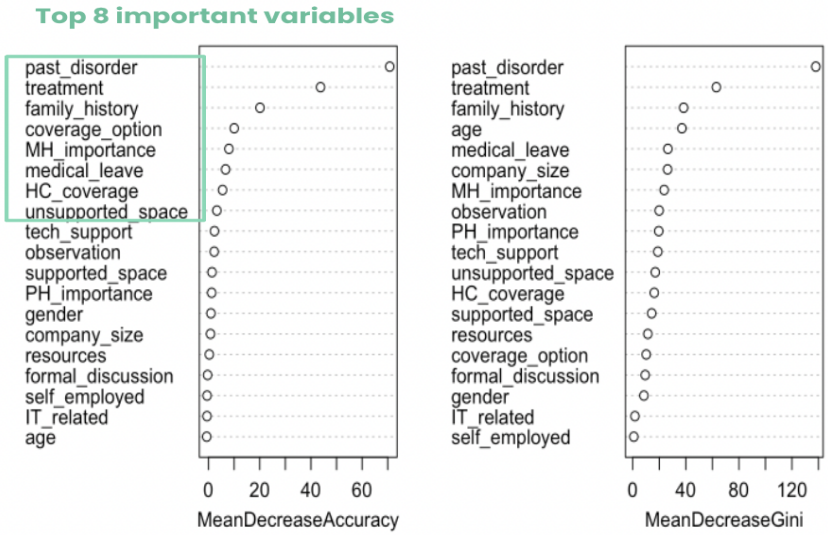
The rule with the highest support, lift and confidence can be translated into plain english as:
Employees having *somewhat difficulties in getting a medical leave, have a history of past mental health disorders, and have mental health benefits as part of their healthcare coverage* are more likely to have a *mental health disorder*.

VII. Conclusion

Even though the initial approach was to select the model with the best Accuracy, we decided to select **Random Forest** as the best model on the basis of **sensitivity**. This is due to the fact that sensitivity represents True Positive Rate (TPR), the ability of the classifier to correctly identify members of the important class. In the case of poor mental health, it is vital to classify true positives correctly to identify who needs help.

Model	Accuracy	Sensitivity	Specificity
Logistic	0.8720	0.9226	0.7628
KNN	0.8679	0.9309	0.6746
Naive Bayes	0.8516	0.8839	0.7821
Random Forest	0.8679	0.9375	0.7179
Boosting	0.8537	0.8853	0.7697
Tree	0.8679	0.8692	0.8640





To answer the research questions posed earlier, we identified the following points.

Factors related to poor mental health:- The following factors stood out in almost all the models implemented:

- Having personal as well as family past history of past mental health disorder
- Availability of mental health care in health coverage
- How much importance is given to mental health in the company

People working in tech roles/industry:-

Professionals working in the field of IT do not correlate to being more prone to mental health issues.

Company and employee's mental health:-

- Ease of asking for a medical leave
- Providing mental health benefits as part of the healthcare package
- Reception of other employees to mental health issues.

VIII. Future Scope

Algorithmic Fairness

- Researching the causes of bias in data and algorithms.
- Building machine learning models with no concept of race, ethnicity, gender or religion, to avoid active discrimination against certain groups.

Text mining

- Perform text mining for the description feature to see the correlation of certain word and their mental health condition.

Expand Dataset

- Add data from other years or other surveys to increase the credibility of our models.

Complex Modeling

- Use complex modeling techniques like Neural Networks, Ensemble Methods etc.

ML Interventions

- Scope for future research to explore extensions of how ML interventions can become valuable tools to support the practices of mental health care experts.

IX. Appendix:

References and Dataset

McKinsey & Company. "*Using digital tech to support employees' mental health and resilience*" https://www.mckinsey.com/industries/life-sciences/our-insights/using-digital-tech-to-support-employees-mental-health-and-resilience

OSMH/OSMI Mental Health in Tech Survey. https://osmhhelp.org/research#