Final Project Research Report
Mental Health of Tech Company Workers Analysis

By Jia-Hung Wu, Sean Chen, Kexin Cui

Introduction

“Move fast and Break things" is a famous mantra of one of the most leading social media company, Facebook. Most experts in tech/IT industries reckon creation speed as one key feature to build successful products.The stress and cut-throat competition takes a toll on the mental well-being.

Many of us will be working at Tech/IT company, and we all know how intensive the working environment is going to be. In order to respond quickly and fast to the market’s need, employees in IT workplace have to face a lot of challenges and heavy workload. In this case, mental health is extremely important, both for the employee him or herself, but also for the company since it has been proved by researches that mental health could affect work productivity. And this is our motivation of performing this project. We want to be able to predict whether an employee should be treated of his mental illness or not based on the values obtained in the dataset.

If we could find a prediction model from worker's self-evaluated survey to suggest whether his or her mental health is in red light, it could raise awareness and improve conditions for those affected by mental health issues. This could help workers to adjust their workload and seek for appropriate medical treatments and also help companies to build a better work environment for employees.

Literature Review

We have seen some researches that used machine learning techniques and ubiquitous sensors to understand mental health [1]. But first of all, they did not mainly focus on Tech companies, which our project focuses on. Also, the dataset they used for machine learning techniques is different from ours.

One of the Kaggle Projects did something similar that relates to mental health in IT workplace and used the dataset from OSMI survey, but we performed more data analysis before diving into the actual machine learning training part. We also calculated f1 measures for different models in our project.

In another paper we read, researchers mentioned that there’s difficulties sensors to capture users’ cognitive context, such as mood and well-being states [2]. Our project, in contrast, collected data from an online survey from people all around the globe, including several questions to measure their self-reported mental health conditions. The paper used GPS location information collected from smartphones to investigate level of stress from mobility patterns, which is really inspiring for future work of our project.

Problem Formulation

The input for our model is the dataset, which is from the 2014 survey that measures attitudes towards mental health and frequency of mental health disorders in the tech/IT workplace. This dataset is from OSMI (Open Sourcing Mental Illness). Open Sourcing Mental Illness is a non-profit organization dedicated to dealing with mental disorders in Tech workplace. They collected and gathered data through web surveys. What they do in support of this goal includes providing e-books on mental wellness in the workplace, hosting a forum on conversations on mental health, and holding talks at developer conferences about mental health in the community. OSMI offers a survey on mental health in IT industry. This survey/dataset consists of various questions including the mental health of the respondents, the demographics of the respondents and how employer views on mental health in the workplace.

But we need to keep in mind that there may be some bias in our dataset. Because this was an opt-in, there is a high possibility that those who are more concerned with mental health would be more likely to participate in the survey. We need to consider this bias in our analysis of the dataset.

Proposed Solution

Consider the dataset size and type of this problem (binary classification problem), our group decided to perform three machine learning models: KNN, Decision Tree and Random Forest.
In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. This is the most convenient and basic model which we can apply for our research. See **Fig. 1**.

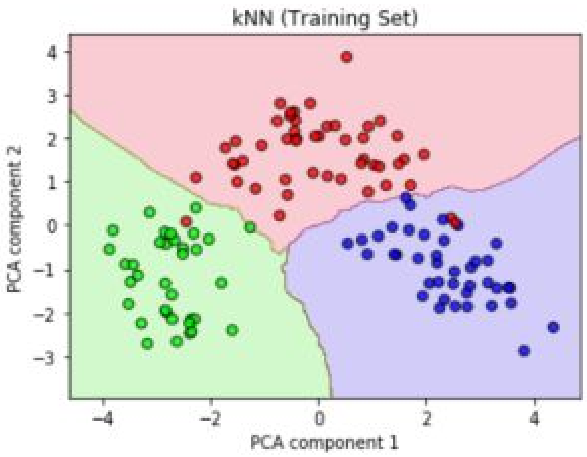
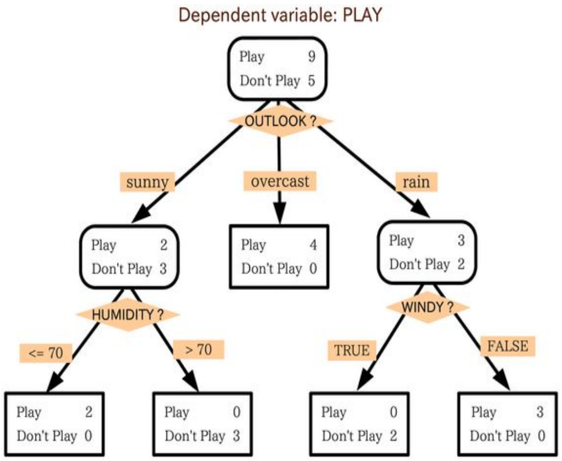


Fig. 1 KNN Classifier

Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). And this model is also more explanatory and easier to interpret the result. See **Fig. 2**.



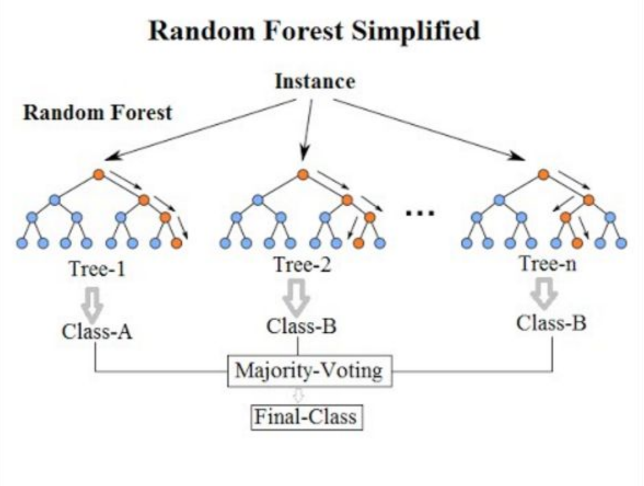


Fig. 2 Decision Tree Classifier

Fig. 3 Random Forest Classifier

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. We use random forests to make up for the deficiency of decision trees in overfitting and having too complicated model. Major-voting between numbers of decision tree could effectively solve these problem. See **Fig. 3**.

First, we extracted features of the highest correlation with the “treatment” feature, and then we added other features to enhance the capability of representing the completeness of data.

Data Description

OSMI survey consists of questions like age and gender of the participants, the location they work, type of their work, family history of mental illness, if they are working remotely, if they are self-employed, self-reported mental health conditions, and so on. This survey is also concerned with how certain demographic and work-life balance of participate impact the mental health conditions in Tech industries.

Our dataset consists of survey responses of 1257 participants, each filling out the survey to 27 questions. We did some data cleaning including some trivial features such as timestamps, state the participant was from if the participant was in the US, and comments. And also since there were missing values in some rows that were not answered by participates, we removed those rows.

Then we did some exploration with the data. First of all, we performed the data cleaning step: removing all the Null values and remove some irrelevant features. After some calculation, we got covariance matrix and "treatment" covariance matrix, as shown in **Fig. 4**. We are able to observe the relationship among all the variables. When the covariance between two features is low, it is darker on the heat maps, and it means that they don't share much in common. In other words, each of them consists max amount of information.

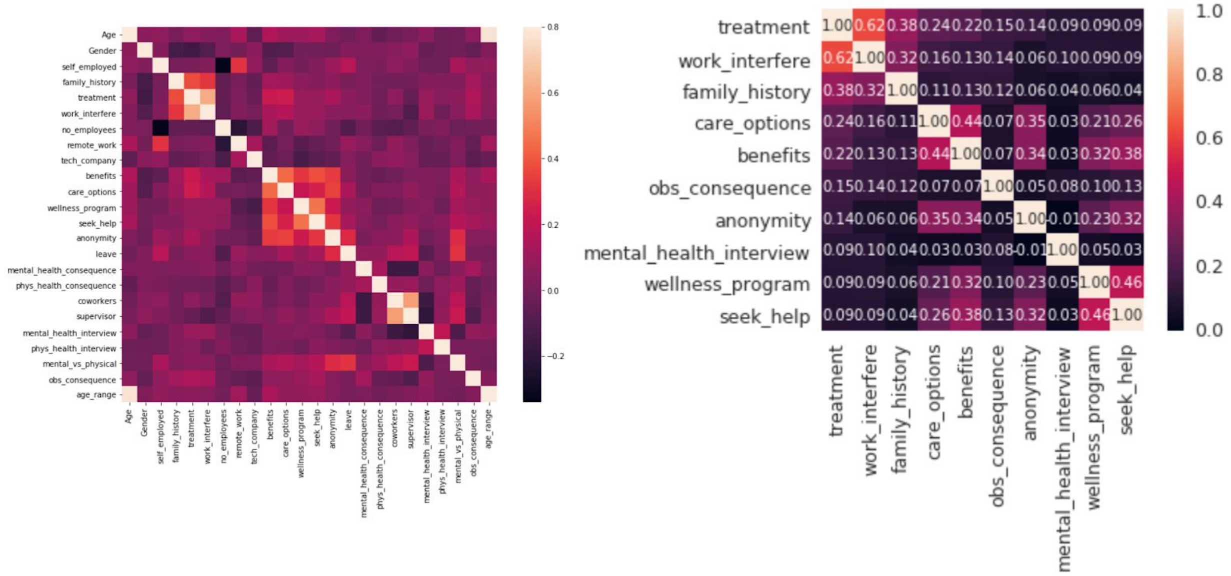


Fig. 4 Covariance Matrix and "Treatment" Covariance Matrix

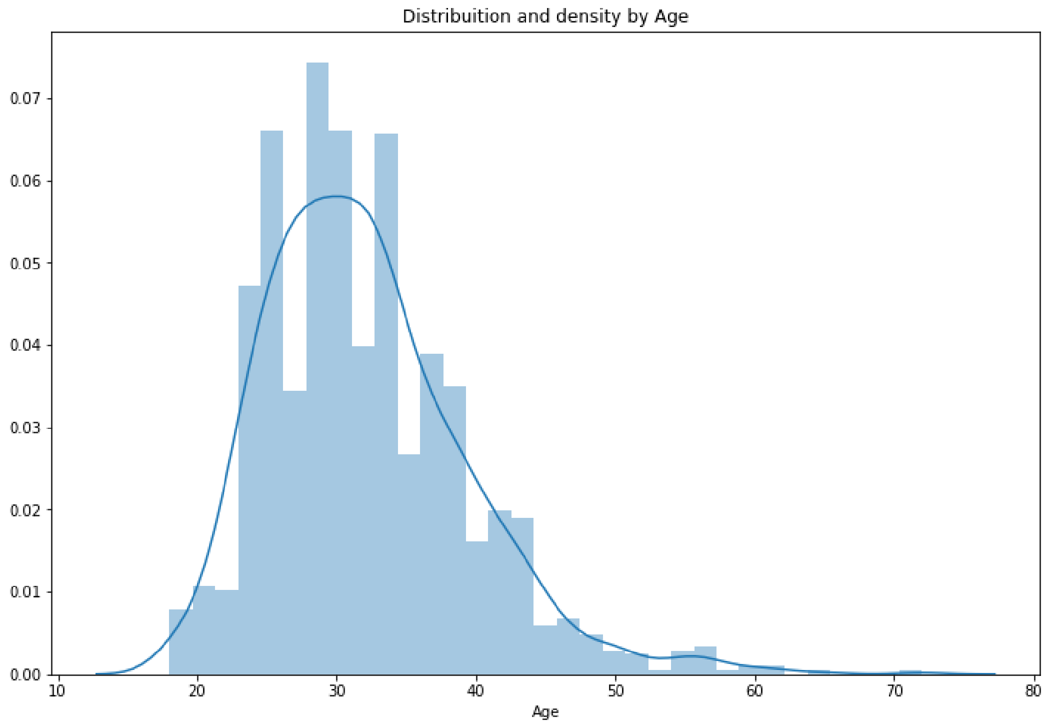


Fig. 5 Distribution and density by Age

From **Fig. 5** we can tell that the age range from 18-72 years old. The average age is 32 years old. Also we can observe that most participants are in the range of 24 to 35, which is younger if we compare this range to that in industries like consulting.

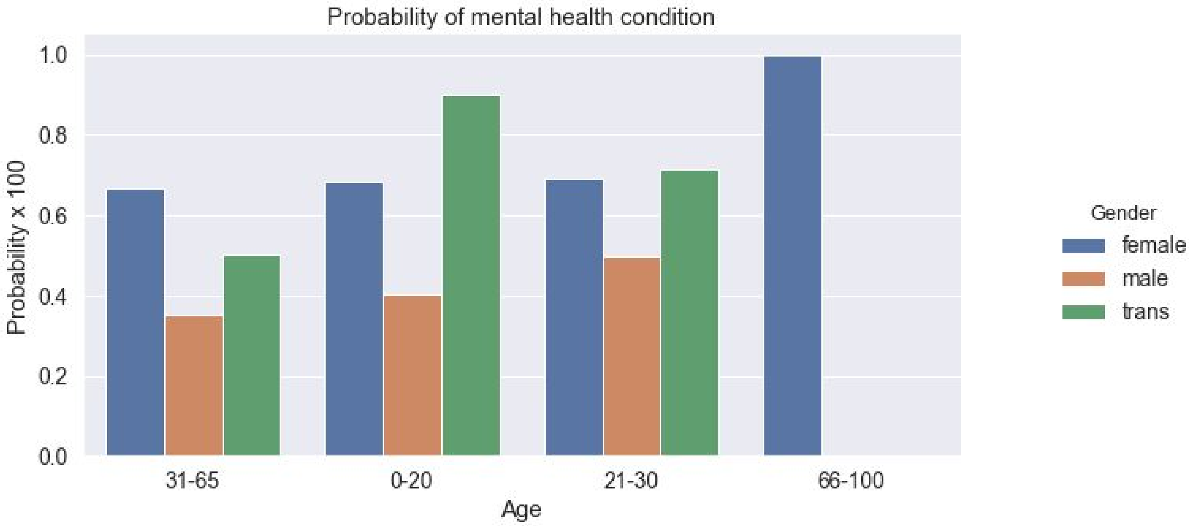


Fig. 6 Probability of mental health condition

If we look into the data, by grouping the participants by age range, we found in **Fig. 6** that for all groups female and transgender participants have had higher probabilities to seek for treatments than male participants do. Notice that since there was only 1 female participant in the age range from 66 to 100, it showed the highest probability to seek for treatments.

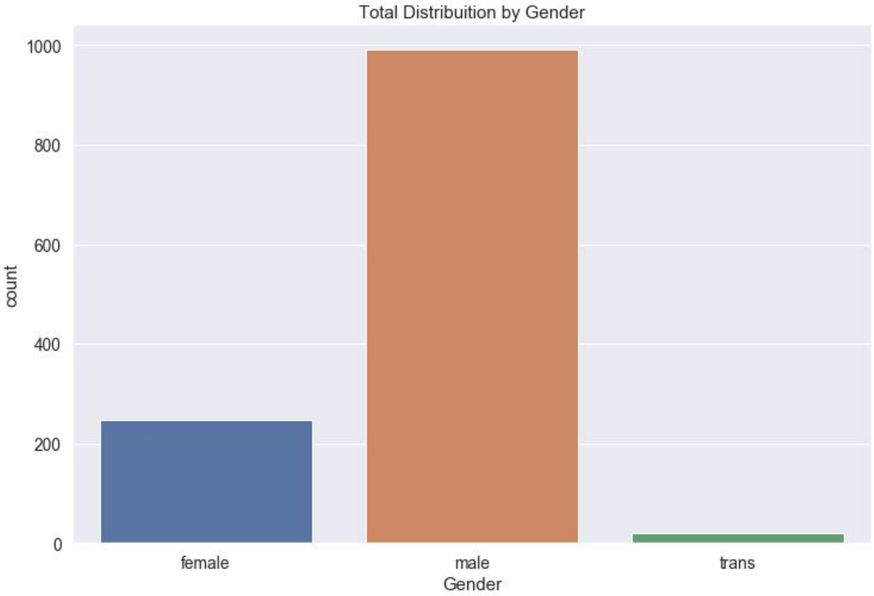


Fig. 7 Total Distribution by Gender

We can observe from **Fig. 7** that the participants are predominantly male (almost 1000), which is typical of the gender imbalance in the IT/tech industry. Around 220 participants are female. We also need to take this into consideration when we do our data analysis that the gender distribution is unbalanced.

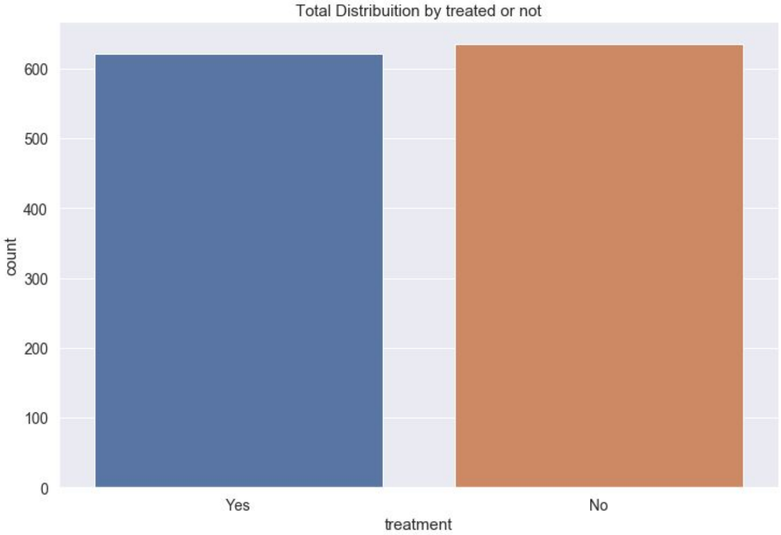


Fig. 8 Total Distribution by Treated or Not

We can see from the survey the total distribution of whether the participants have had treatments or not, as shown in **Fig. 8**. The question is “Have you sought treatment for a mental health condition?”. The distribution of this question shows that around less than half of participates have had treatments before, and a little bit more than half have not had treatment before.

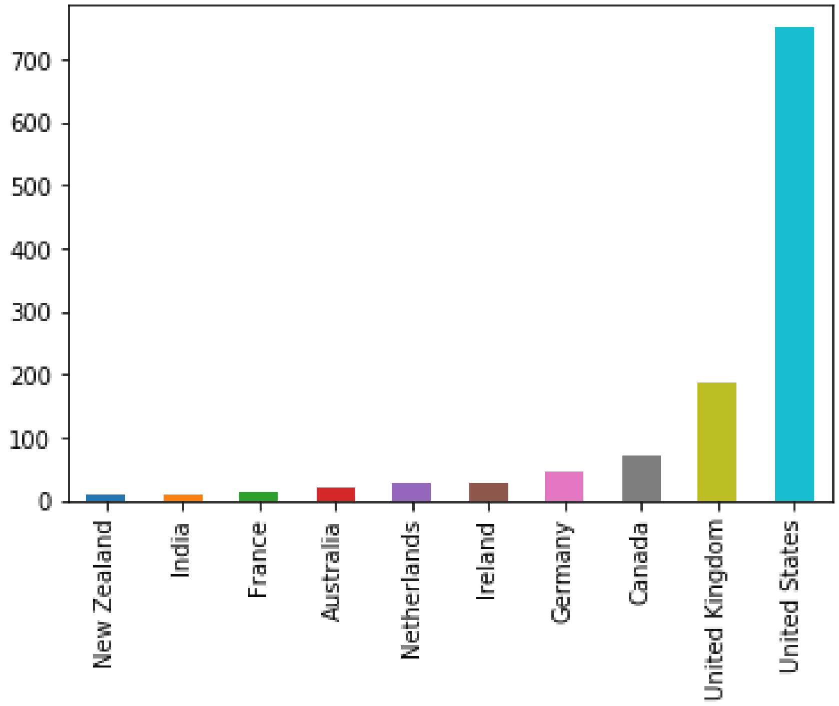


Fig. 9 Distribution by Country

The participants are from 47 countries around the world. But we can see from **Fig. 9** that there are around 60% of the participants from the US and around 15% of the participants from the UK, which shows a pretty skewed distribution geographically.

Results

After data extraction procedure, we choose age, gender, family_history, benefits, care_options, anonymity, leave,work_interfere, obs_consequence, mental_health_interview, seek_help and wellness_program as our input feature, while the treatment feature is our ground truth feature.

To test the performance of our models, in all the following machine learning models we split our data set into training dataset (70%) and testing dataset (30%). First, we run grid search algorithm to find the best hyperparameters which would give us the best performance scores. Then we use these hyperparameters in our training process. To evaluate the results, we use precision, recall rate, F1-measure and the receiver operating curve (ROC) of the classifiers. By comparing these scores, we know which model performs the best on the prediction. The followings show our training results.

1. **KNN model**

   Here we present the result of our KNN model in **Table 1**, using k = 19. The precision rate is around 0.75. The recall rate is much higher, almost 0.9. We can see the confusion matrix in **Fig. 10** below. The number of false positive prediction (56) is much higher than that of false negative prediction (20).

Table 1 Performance of KNN

Evaluation Index	Performance
Precision	0.74887
Recall	0.89304
F1-measure	0.79894
ROC-AUC	0.87

In **Fig. 9**, we test our KNN classifier and draw the ROC curve. The AUC is about 0.85, which seems to perform pretty well.

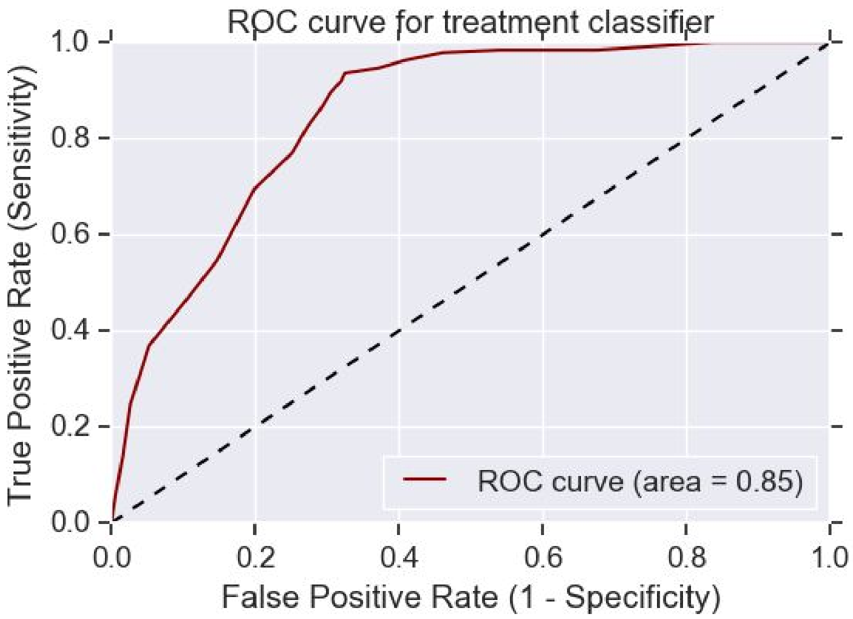


Fig. 9 ROC of KNN

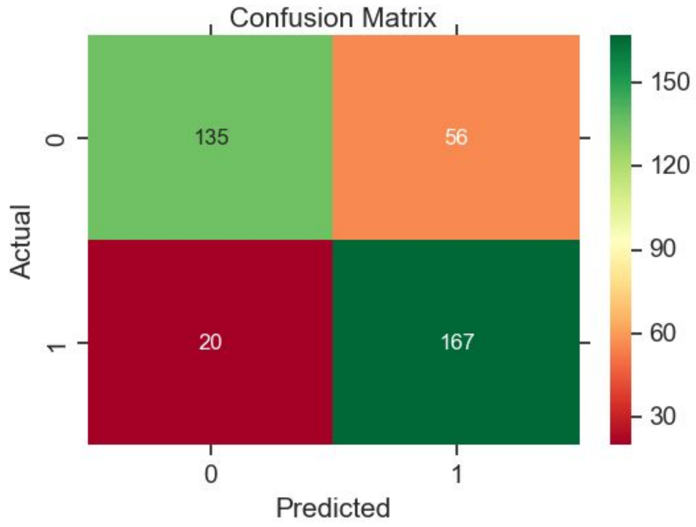


Fig. 10 Confusion matrix

2. **Decision Tree**

   Here we present the result of our decision tree model in **Table 2**. We can see that all three performance indices are near 0.75. Our maximum depth of decision tree is 3. We can see the confusion matrix in **Fig. 12** below.

Table 2 Performance of Decision tree

Evaluation Index	Performance
Precision	0.77472
Recall	0.75401
F1-measure	0.76984
ROC-AUC	0.82

In **Fig. 11**, we test our decision tree classifier and draw the ROC curve. The AUC is about 0.82.

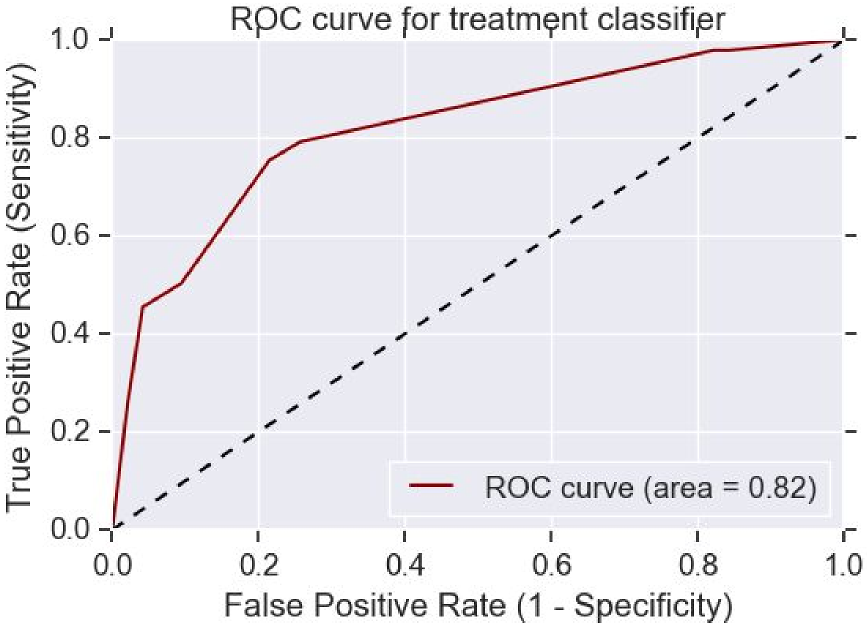


Fig. 11 ROC of Decision Tree

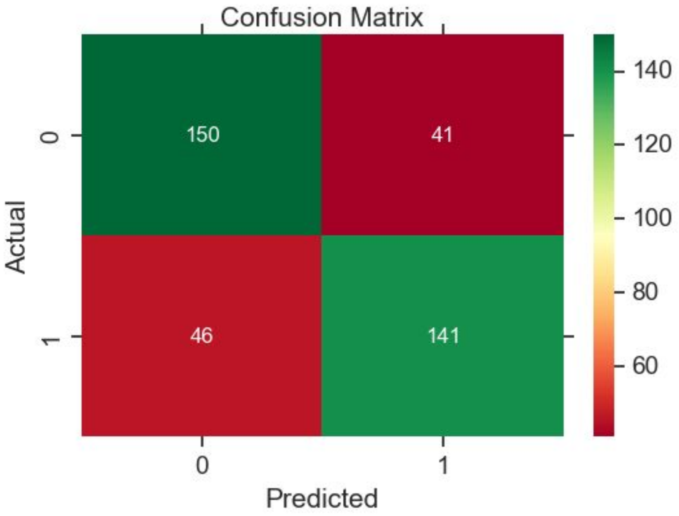


Fig. 12 Confusion Matrix of Decision Tree

3. Random Forest

Here we present the result of our random forest model in **Table 3**. We use 20 estimators for the major vote. We can see that while the precision rate is only 0.74, the recall rate is much higher, almost 0.9. The confusion matrix is in **Fig. 14** below. We can see that the number of false positive prediction (59) is much higher than false negative prediction (15).

Table 3 Performance of Random Forest

Evaluation Index	Performance
Precision	0.74458
Recall	0.91978
F1-measure	0.80423
ROC-AUC	0.89

In **Fig. 13**, we test our random forest classifier and draw the ROC curve. The AUC is about 0.89. Our classifier seems to outperform the random guess line.

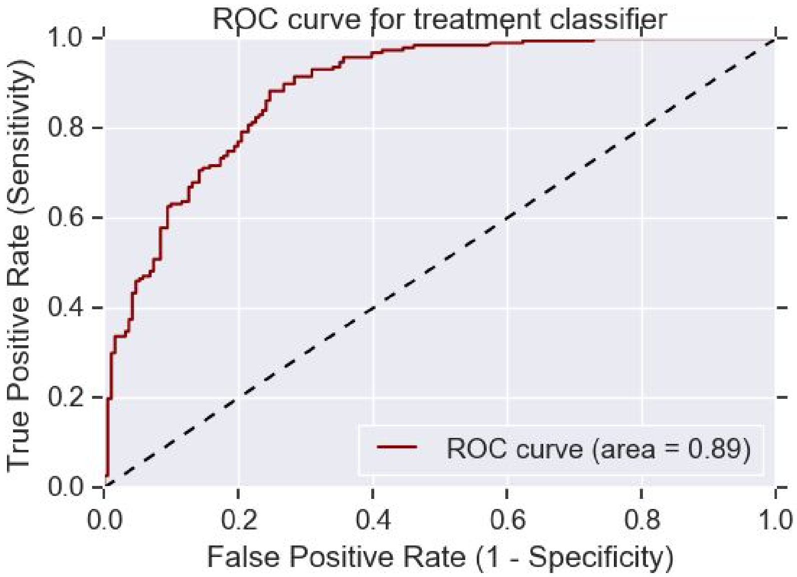


Fig. 13 ROC of random forest

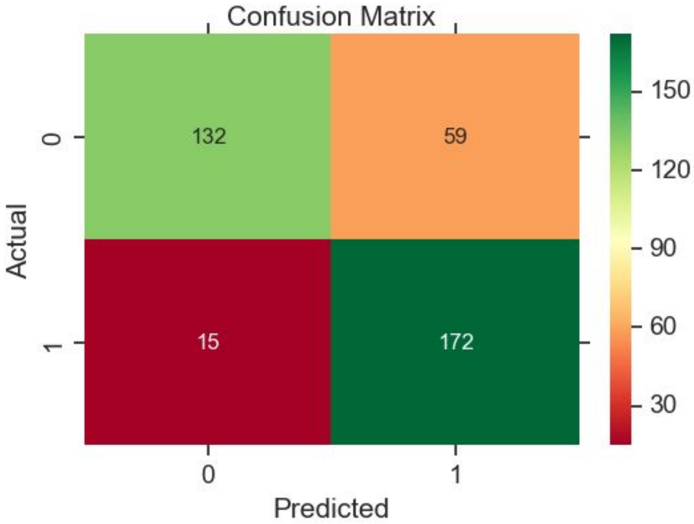


Fig. 14 Confusion Matrix of Random Forest

Conclusions

We believe the models we designed and implemented are going to be useful for workers at Tech company. The best model we have for this 2014 survey dataset is the random forest model, with an 80.4% accuracy.

We have three points to address for our future plan. First of all, the dataset we used in training our models was collected in 2014. There are surveys conducted in 2016, 2017 and 2018. It could useful if we can perform a yearly analysis of all these surveys and observe if our results change over this period of time.

And then as we mentioned before, there might be some bias in our dataset. Since it was an opt-in survey, it is likely that those who are more concerned with mental health disorders may be more likely to take the survey. Therefore, we should use a larger and less biased dataset in the future. Also, it could be useful if the survey includes more information like the working culture of the specific tech company. This also means that we can find a more objective way to collect the data rather than only using subjective self-reported data for analysis. This includes exploiting mobile phones, monitors and different devices to detect biological activities.

At the end, our group thinks that it could be interesting if we finalized a great model and make an application out of the results. So employees can use it with ease and convenience, anytime and anywhere they want.

Reference

[1] Mohr, David C. et al. “Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning.”: Annual review of clinical psychology 13 (2017): 23-47 .
[2] G. Mikelsons, M. Smith, A. Mehrotra, M. Musolesi, “Towards deep learning models for psychological state prediction using smartphone data: Challenges and opportunities”, 2017.