Mental Health in Tech. Industry

Amogha Shettar, Amoolya Shettar, Ankush Kesri

Abstract

With the arrival of technology and start-ups in the current industry, the demand for innovation and resilience has been increased. Because of the current competition among rival companies working in the same field, the emphasis on high productivity has grown significantly. Due to the high focus on profits and productivity, the founders of these companies are usually under a lot of stress, and as a result, employees of these companies face the same stress due to tight deadlines, unusual hours, and late nights. This has resulted in a decline in mental health as well as an increase in absenteeism and long-term stress, which is exactly the opposite of what the current industry desires. Using exploratory data analysis, we focused on mental health, its predictors, and the effects it has on the productivity of the tech industry. The study has been done on use of data mining and apply some of the techniques including Support Vector Machines (SVM), Logistic Regression (LR) and Random Forest (RF) to predict whether the employee who is dealing with mental health issues has sought treatment for the same and if so then what factors influenced this decision.

Keywords

Mental health, data mining, work life, mental health issues, feature selection, modelling

1. Introduction

Mental health and physical health have a direct relationship. For instance, depression is linked with asthma, diabetes, and hypertension. Employees experience high morale when they work in a harmonious and healthy workplace. In any organization, the staff is its largest asset. Without a mentally healthful place of work, the group will revel in low morale, humans turn into cynical, harassed, and anxious; bodily fitness troubles will increase, illness degrees will rise, productiveness will drop, and organizational weather may be affected. A mentally healthful place of work is one that is visible as a satisfied and pleasant region to work. It has excessive productiveness degrees and is efficient and is open to discussions about intellectual fitness issues.

Mental health is negatively related to absenteeism. Long-term stress and traumatic events at work can lead to psychological problems, which can cause the worker to be absent from work and prevent the worker from working again. When we're under pressure, it can be hard to maintain a balance between work and non-work life. The experience of work stress is a challenge to the health and safety of workers and to the healthiness of their organizations.

For this study, we have applied data mining techniques to identify strong predictors of mental health and to determine how likely it is that the patient suffering from mental illness will sort the treatment or not, including what factors influenced the patient in doing so.

1.1 Significance

The current state of knowledge about mental health and its main predictors is insufficient. Because mental health is a direct contributor to a person's well-being as an employee and the advancement of industry,

industries should place a greater emphasis on this area. This study contributes to the understanding of the relationship between mental health and a variety of factors included in this dataset, such as age, gender, and so on. There is still a lot of work to be done, and the results can be used by the industry to improve the mental health of the employees, which will boost the industry's productivity.

2. Literature Review

This section provides a brief review of previously published literature of mental health and data mining. A brief explanation of mental illness is provided. Mental Health is "a state of well-being that allows individuals to experience their abilities, cope with the stress of normal life, do productive and fruitful work, and contribute to the community." According to the World Health Organization (WHO). Mental health includes, among other things, subjective well-being, perceived self-efficacy, autonomy, ability, intergenerational dependence, and self-actualization of one's intellectual and emotional potential. From a positive psychological or holistic point of view, mental health may include the ability of an individual to enjoy life and balance life's activities with efforts to achieve psychological resilience. Cultural differences, subjective assessments, and competing expert theories all influence the way "mental health" is defined. Some early signs associated with health problems are thinking of stimulating sleep, lack of energy, and harming oneself and others. Various scholars have used machine learning to understand and predict mental health issues of individuals, like predicting the factors affecting mental health. Tian et al. used a multilayer perceptron algorithm to build a suicide recognizer for Weibo social media. Carpenter used machine learning and psychiatric assessment data to assess the risk of anxiety.

Data Mining has been used in mental health diagnosis using genetic algorithms by Ghassan Azar and his co-authors and the study was published in 2015 IEEE International Conference on EIT. Other studies include using data mining to do the research on the relationship between Covid-19 and mental health. M. P. Dooshima et. al. has made use of biological, environmental, psychological and demographic factors to make prediction. In some study, before applying data mining techniques, various mental health experts were consulted to validate the features and required parameters. M. Srividya et. al. has made use of questionnaire to obtain values for various features that can be used to predict the factors for mental health.

Shahidul Islam Khan and colleagues examined data from 500 prescription reports, which included a description of the patient, personal and family health history, and mental illness problems, including symptoms. They compared the accuracies of Random Forest, KNN, and SVM in predicting patients' mental disorders given all other features. The models' final accuracies were 85.9, 84.4, and 85.1 for the RF, SVM, and KNN models, respectively. They discovered that the Random Forest model performed the best among the models tested in predicting the patient's mental disorder.

3. Problem Definition

Given the dataset that contains the possible factors that could affect the attitude of a person at the workplace, this project will address two primary goals:

1. To identify the strong predictors of mental health illness or attitude towards mental health at the workplace and predicting whether a person has sought treatment for a mental health condition or not.

2. If the person has not sought treatment, then what are the factors that are letting him/her to not have a positive attitude towards mental health.

4. Methodology

The data is collected from Kaggle. The dataset has 27 columns and 1259 rows. The dataset has 1 numerical attribute, 3 categorical variables, 19 nominal variables, and 2 ordinal variables. The attributes cover the geographic and demographic of the people like age, gender, family history, state and country, information about the workspace like the number of employees, type of employment, type of work, and attributes related to the support given at the workspace for mental health like care options, leave, wellness programs and benefits.

4.1 Preprocessing

'Timestamps' column was found to be irrelevant as time is not taken into account in this analysis. 'Comments' was dropped as it contained too many null values and also, they were subjective. The dataset contained a column called 'gender' that has 49 unique values. We categorized all the entries into three categories for the sake of simplicity. They were categorized as 'Male', 'Female' and 'Other'.

```
df_mhealth['Gender'].unique()
```

```
array(['Female', 'M', 'Male', 'male', 'female', 'm', 'Male-ish', 'maile',
       'Trans-female', 'Cis Female', 'F', 'something kinda male?',
       'Cis Male', 'Woman', 'f', 'Mal', 'Male (CIS)', 'queer/she/they',
       'non-binary', 'Femake', 'woman', 'Make', 'Nah', 'All', 'Enby',
       'fluid', 'Genderqueer', 'Female ', 'Androgyne', 'Agender',
       'cis-female/femme', 'Guy (-ish) ^_^', 'male leaning androgynous',
       'Male ', 'Man', 'Trans woman', 'msle', 'Neuter', 'Female (trans)',
       'queer', 'Female (cis)', 'Mail', 'cis male', 'A little about you',
       'Malr', 'p', 'femail', 'Cis Man',
       'ostensibly male, unsure what that really means'], dtype=object)
```

Fig. 1: Gender unique values

'State' column was dropped as it contains states of the United States. Missing data existed in the columns: state, self_employed, work_interfere and comments.

```
df_mhealth.isna().sum()
```

```
Timestamp                      0
Age                            0
Gender                         0
Country                        0
state                        515
self_employed                 18
family_history                 0
treatment                      0
work_interfere               264
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
comments                    1095
dtype: int64
```

Fig. 2: Null count

The age column had three rows with negative ages. They were removed as outliers. The age range was set between >18 and <90.

Except for Age feature, all other features are categorical, therefore Label encoding has been performed on all other features. Label encoding is an encoding technique for handling categorical data. In this technique each label is assigned a unique integer based on alphabetical ordering. Feature scaling is a method used to normalize the range of independent variables or features of data. This method is generally used for normalization of data. For our dataset, feature scaling has been used on Age feature.

4.2 Data Mining

4.2.1 Logistic Regression

Logistic regression is a classification model rather than a regression model, despite its name. Logistic regression model is an efficient and simple method for linear and binary classification problems. This model is easy to realize and performs very well with linearly separable classes. Logistic regression model is a statistical method that can be generalized to multiclass classification. Logistic regression can be thought of as a linear regression but for classification problems. The logistic regression range is between 0 and 1. There is no requirement of a linear relationship between input and output variables in logistic regression. For large datasets, the logistic regression model is really good. They are relatively fast to implement. Through dummy variables, logistic regression can be modified to handle categorical explanatory variables. Logistic regression can be used for both class probability estimation and classification as it is tied with logistic data distribution. The output variable is binary in the basic version of logistic regression; however, it can be extended to multiple classes.

Logistic regression is much easier than models such as neural networks, however it is not as easy as the interpretation of linear regression or kNN models. Logistic regression is prone to complete separation and restrictive expressiveness. output of the logistic regression is reported in terms of numerical odds of the binary and the maximum likelihood method is intensive computationally. Since logistic models are heteroskedastic, the maximum likelihood method does not minimize variance so there exists no measure of fit. However, there are various pseudo-R2 statistics that convey goodness of fit.

4.2.2 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning technique that is used widely in classification problems and pattern recognition. By constructing a multidimensional hyperplane SVM performs a classification that discriminates two classes by maximizing the margin between data clusters. Transformation from input space into a multidimensional space is done by using special nonlinear functions called kernels. Two parameters for the kernel that are gamma and C need to be selected before to develop an optimal SVM model. The degree of non-linearity is controlled by parameter gamma and over-fitting of the model is controlled by the parameter C. The idea behind using SVM technique is to build a n-1 dimensional separating hyperplane that discriminates two classes in n-dimensional space. Example: the separating hyperplane will be a straight line dividing the space into half for two variables in a dataset that create a two-dimensional space. SVM searches for something called maximum-margin separating hyperplane which is an optimal hyperplane when there are more dimensions involved. Distance between the nearest data point and hyperplane is maximized. The fundamental difference between SVM and multiple logistic regression is that SVM tends to classify entities without giving the estimates of the classes in the dataset.

4.2.3 Random Forest Classifier

Random Forest Classifiers is an ensemble-based learning method. They are fast in operation, can be implemented in a simple way, and they have proven to be successful in various domains. Decision tree is the core unit of random forest classifiers. Using the features of a dataset, a hierarchical structure, a decision tree is built. Random forest is basically a collection of decision trees that are associated with a set of bootstrap samples that originated from the original dataset. The size of the sunset that are created from the original dataset is the same as that of the original dataset. The approach consists of building a number of decision trees in the training step and the majority vote amongst them in the classification step. In the training step, a technique called bagging is applied by random forests to individual trees in the ensemble. A random sample with replacement is selected repeatedly by bagging from the training set and it fits trees to those samples. Without pruning, the tree is grown. Using out-of-bag errors, the number of trees in the ensemble is learned automatically. Random Forest Classifier is very popular mainly because of good performance and its simplicity. However, random forests show a degree of unpredictability regarding the final trained model. To calculate feature importance, random forests have methods to introspect which are based on the decision process the random forest used to build the model. Features are dropped to estimate their importance on the model performance in feature importance. Gini impurity is used to determine the split between classes at each node. The gini impurity helps in measuring how informative a feature can be in a model.

5. Results

5.1 Exploratory Analysis

Before employing the data mining techniques, a brief exploratory analysis was conducted on the data.

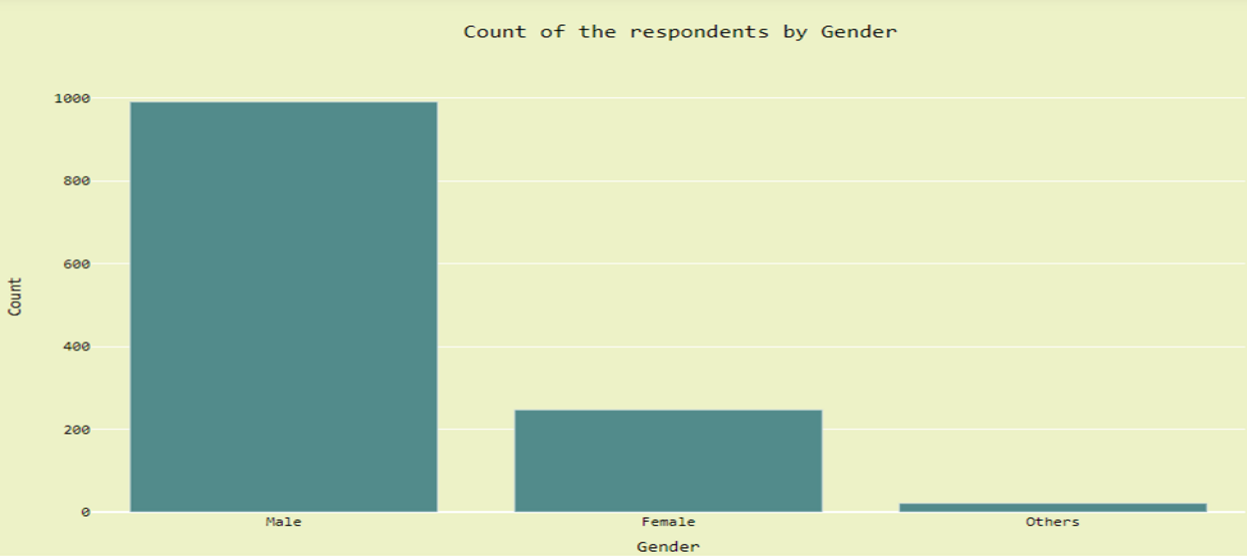

Fig. 3: Count of respondents by gender

In figure 3, we can see that respondents are largely male. The male to female ratio is almost 4:1.

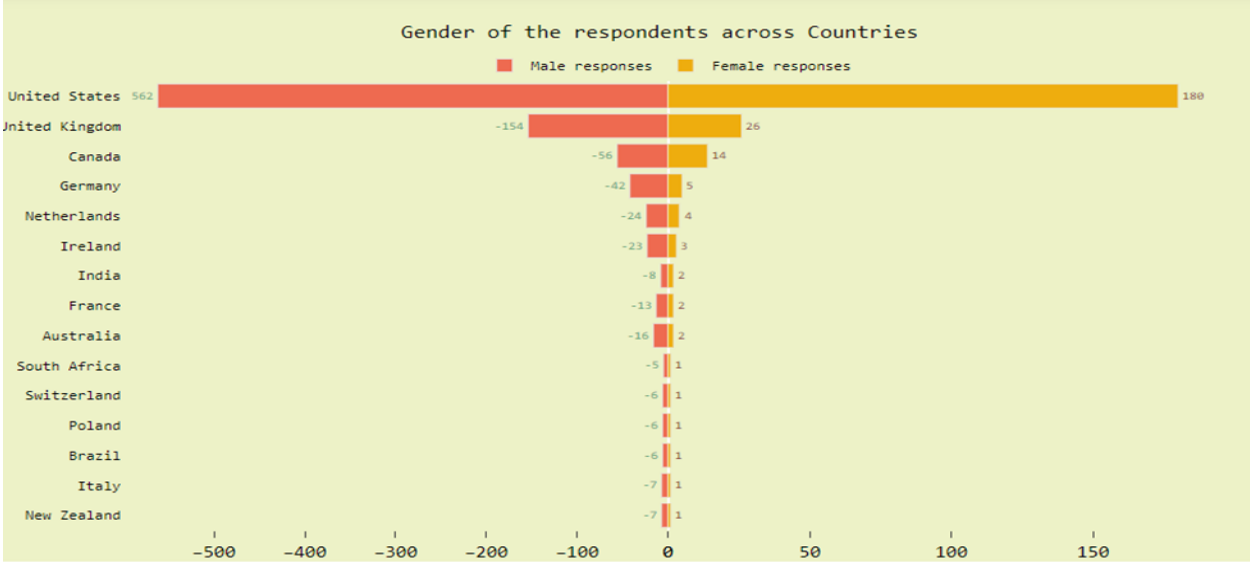

Fig. 4: Gender of respondents across countries

In figure 4, we can see that there are many countries with a very few responses. Most of the responses are from the U.S. The percentage of responses from the U.S is almost 65%. We observed that the U.S., U.K. and Canada had the most correspondents accounting for about 82% of all responses. Therefore, it does not make sense to conclude any demographic findings.

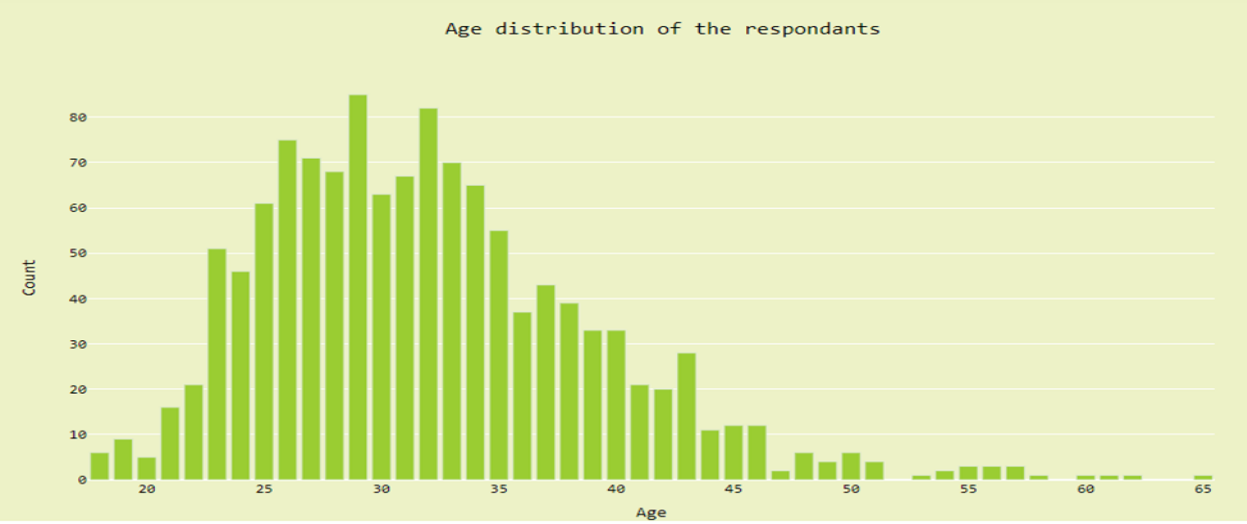

Fig. 5: Age distribution of the respondents

In figure 5, we can observe that the median age is 31 and the age is centered between 25 and 38.

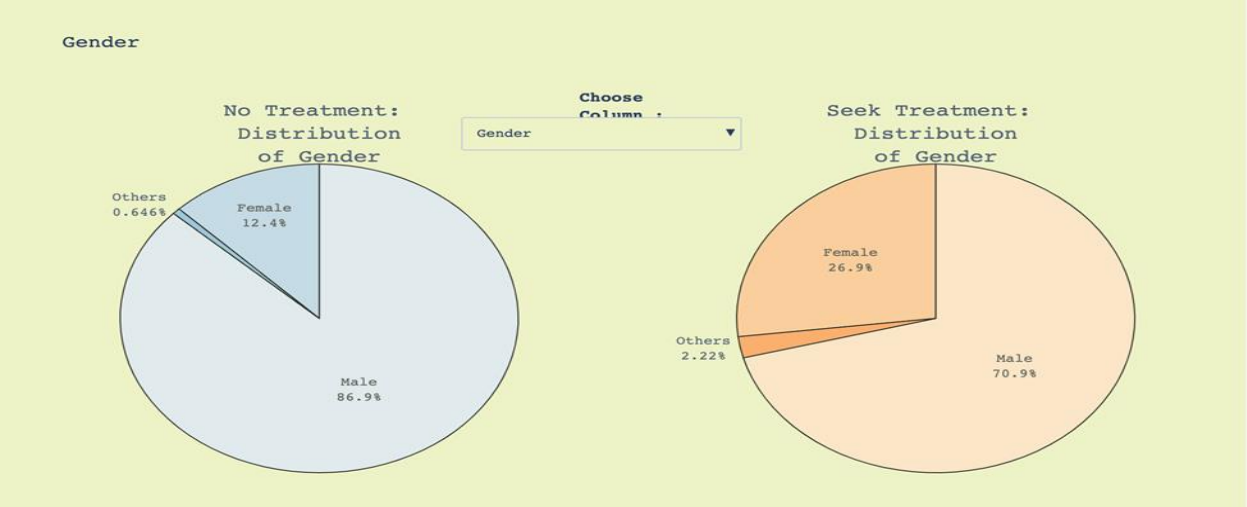

Fig. 6: Seek treatment or not based on features

Figure 6 shows us if employees seek treatment or not based on various factors like gender, company size, benefits, work interference, employer score, care options, family history, remote work, leave, and wellness program. These pie graphs helped us in selecting various factors that really impacted the decision of an employee to seek treatment or not.

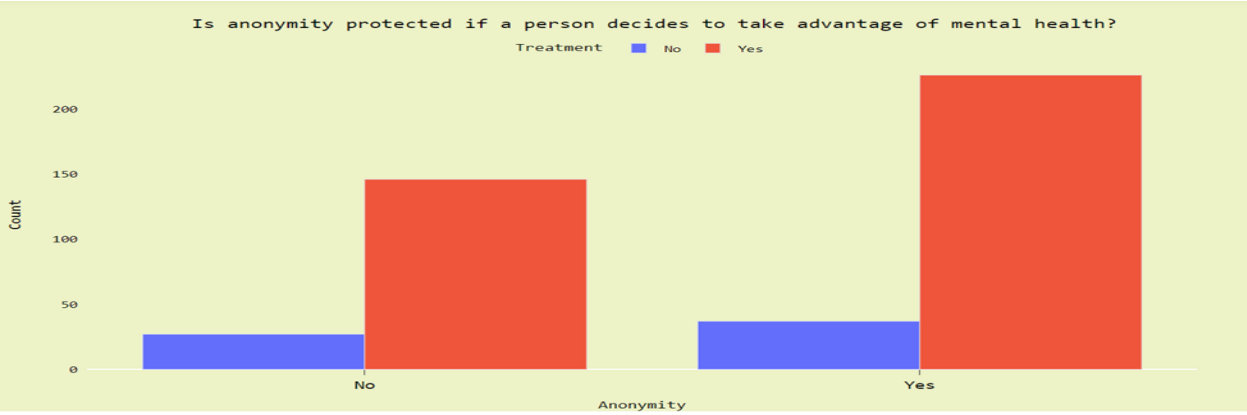

Fig. 7: Anonymity and mental health

Since mental health is stigmatized in today's world, promising anonymity can be helpful. It will create a barrier for those who might want to seek treatment if anonymity is not promised. In figure 7, we can see that there is a 25% rise in the number of treatments for those people whose anonymity is protected whereas, there is no change in the number of treatments for those people whose anonymity is not protected.

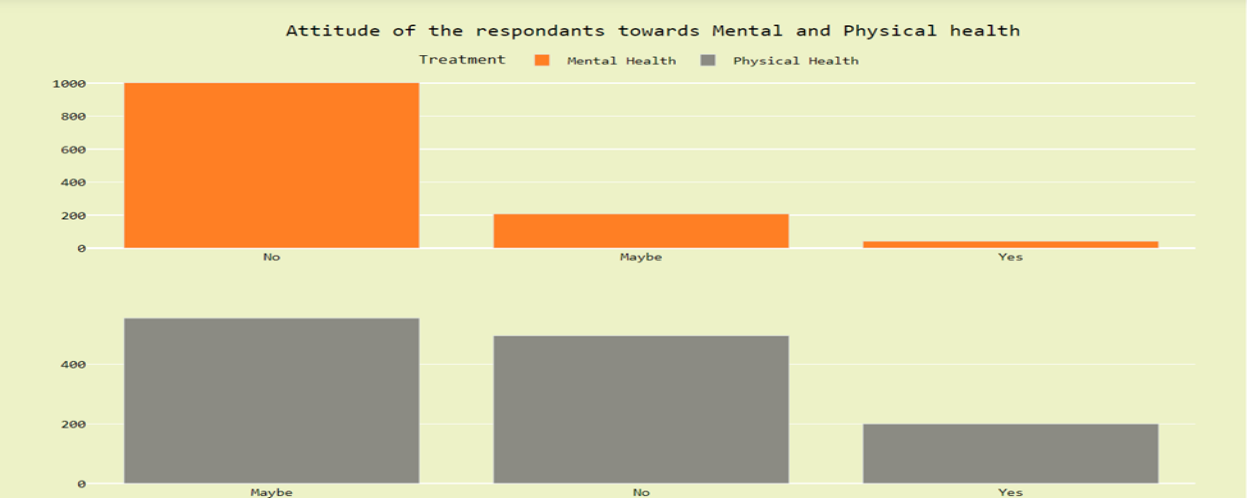

Fig. 8: Attitude of respondents towards Mental health and Physical health

In figure 8, we are trying to understand how the attitude varies towards mental health when compared to physical health. We found that approximately 80% of the respondents are not comfortable speaking about their mental health whereas only 5% people feel comfortable speaking about their mental health. An ambiguous choice called 'maybe' spikes while talking about physical health but remains very little while talking about mental health.

5.2 Data Mining

The data was randomly divided into a training set (70%) and a testing set (30%). Table. 1 shows the accuracy, error, precision, AUC score, and CV AUC of the classification for the basic models like logistic regression model, support vector machine, and ensemble model - random forest classifier.

Table. 1: Evaluation scores for the classification models

Model	Accuracy	Error	Precision	AUC Score	CV AUC
Logistic Regression	0.7340	0.2659	0.7790	0.7363	0.7635
Support Vector Machine	0.7952	0.2047	0.7716	0.7922	0.8614
Random Forest Classifier	0.7819	0.2154	0.7601	0.7813	0.8726

Since the accuracy was not very good in building a model that classifies the original data, and since there were too many features that were contributing in building a model, we decided to perform feature selection using Extra Tree Classifier. This classifier is an ensemble learning technique which aggregates the result of multiple de-correlated trees collected in a "forest" to output its classification result. This is used to do feature selection; each feature is ordered in descending order according to the Gini importance of each feature and the user selects the top k features according to his/her choice.

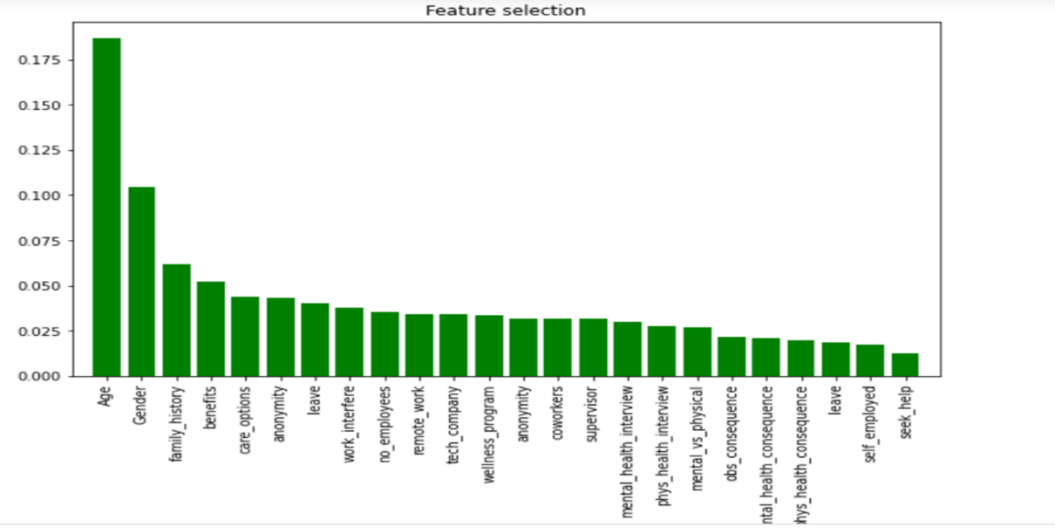

Fig. 9: Feature selection using Extra Tree Classifier

Fig. 9 shows a bar graph that displays the features after feature selection is performed. From the graph, it is quite evident that Age, gender are really important features that contribute largely. After feature selection, the accuracies have improved, and the results are displayed below in the Table. 2

Table. 2: Evaluation scores for the classification models after Feature Selection

Model	Accuracy	Error	Precision	AUC Score	CV AUC
Logistic Regression	0.7181	0.2819	0.7419	0.7187	0.7547
Support Vector Machine	0.8085	0.1941	0.7510	0.7999	0.8495
Random Forest Classifier	0.7845	0.2181	0.7544	0.7783	0.8903

6. Discussion

Based on the analysis done, and the results derived, we can say that there is an improvement in work performance by nearly 86% and rate of absenteeism is low after receiving the treatment for depression. The companies can begin to create a culture of understanding and compassion at the tech company by providing access to mental health benefits to the employees.

Also, it is quite evident that gender, age, and family history greatly influence the decision to get treatment for the employees. The company can provide support by making an assessment of the employee's personality as various characters show various needs. Work interference also influences the employees on whether they need to get the treatment or not. The company must also make sure that they provide good benefits for the employees so that the employees will be able to maintain and balance their mental health.

7. Conclusion

Mental health has a significant effect in the work environment and is directly linked to the performance of the employees and is inversely proportional to absenteeism. Mental health, if left unchecked, can lead to a rise in depression, anxiety, and loneliness which as a result causes loss of motivation at work and reduces work performance. Poor mental health can negatively affect not only job performance but also communication and daily functioning. So, it is important that employees get the support and the treatment they need for their mental health. Age, gender, family, benefits, care options, anonymity, leave, and work interfere are the major factors that affect the mental health of an employee at work space.

The study sheds light on best strategies to improve employee's well-being and to encourage them to seek treatment when necessary. Future research could include a larger dataset that have additional features that could determine and give an insight into the tech world with a different perspective. Additionally, apart from the tech industry, this study can also be performed on other industries as well.

References:

1. Wikipedia contributors. (2021, December 16), https://en.wikipedia.org/wiki/Mental_health

2. Luo, M. (2021). Research on Students' Mental Health Based on Data Mining Algorithms. *Journal of Healthcare Engineering*, *2021*.

3. Barkved, K. (n.d.). Let's Talk: It's Time to Get Serious About Mental Illness in Tech.

4. Vu, V. (2021, March 22). *Mental Health in Tech Industry*.

5. Alonso, S. G., de la Torre-Díez, I., Hamrioui, S., López-Coronado, M., Barreno, D. C., Nozaleda, L. M., & Franco, M. (2018). Data mining algorithms and techniques in mental health: a systematic review. *Journal of medical systems*, *42*(9), 1-15.