# Predicting mental illness at workplace using machine learning

Table of Contents....................................................................................................................... Figures............................................................................................................................... vii Abstract........................................................................................................................................ viii Chapter 1 - Introduction.................................................................................................................. 1 1.1 Problem Statement................................................................................................................. 1 1.2 Approach to solving the problem .......................................................................................... 1 Chapter 2 - Related Work ............................................................................................................... 2 Chapter 3 - Technical Approach ..................................................................................................... 5 3.1 Project Set up......................................................................................................................... 5 3.2. Technology Stack ................................................................................................................. 6 3.3 Data Exploration.................................................................................................................... 7 3.3.1 Data collection and exploration: ..................................................................................... 7 3.3.2 Feature Engineering ........................................................................................................ 8 3.3.2.1 Cleaning of gender data................................................................................................ 8 3.3.2.2 Data visualization of few attributes.............................................................................. 9 3.3.2.3 Count of mental health disorder in the tech by country ............................................. 10 3.3.2.4 Companies support on mental health issues to their employees................................ 11 3.3.3 General Workflow......................................................................................................... 12 Chapter 4 - Data Preprocessing..................................................................................................... 14 4.1. Missing values.................................................................................................................... 15 4.1.1 Ignore these pairs .......................................................................................................... 15 4.1.2 Complete the missing data ............................................................................................ 15 4.2. Removing outliers............................................................................................................... 15 4.3. Data Analysis...................................................................................................................... 15 4.4. Encoding the categorical data............................................................................................. 15 Chapter 5 – Determining the problem .......................................................................................... 17 Chapter 6 - Metrics & Models ...................................................................................................... 18 6.1 Receiver Operating Characteristic (ROC)........................................................................... 19 6.2 Naïve Bayes......................................................................................................................... 21 6.3 Decision Tree....................................................................................................................... 22 6.4 Random Forest..................................................................................................................... 23 6.5 Logistic regression............................................................................................................... 25 vi 6.6 Metrics................................................................................................................................. 27 6.7 Model Results...................................................................................................................... 28 Chapter 7 – Conclusion & Future

## A B S T R A C T

Mental illness (MI) is a leading cause of workplace absenteeism that often goes unrecognized and untreated. This paper presents a machine learning algorithm for predicting MI at workplace. The dataset consisted of responses from 1259 subjects collected through an online survey using a self-assessed questionnaire on the workplace environment. The responses were used as features for training a support vector machine to predict MI.

### **Chapter 1 – Introduction**

* 1. **Problem Statement:**

Mental health affects your emotional, psychological and social well-being. It affects how we think, feel, and act. It also helps determine how we handle stress, relate to others, and make choices. In the workplace, communication and inclusion are keys skills for successful high performing teams or employees. The impact of mental health to an organization can mean an increase of absent days from work and a decrease in productivity and engagement. In the United States, approximately 70% of adults with depression are in the workforce. Employees with depression will miss an estimated 35 million workdays a year due mental illness. Those workers experiencing unresolved depression are estimated to encounter a 35% drop in their productivity, costing employers $105 billion dollars each year.

* 1. **Approach to solving the problem**

According to [36], Modern data science methodology uses machine learning. Rightfully so. Machine learning contains the algorithm at its core, which gives it a major advantage over all other traditional data science methodologies. These are the guidelines a computer follows to locate a model that as closely matches the facts as feasible. Machine learning differs from conventional data science techniques in that the computer learns how to find the model on its own using the algorithm's instructions rather than receiving instructions from us. In contrast to conventional data science, machine learning requires little human input.

A machine learning algorithm functions similarly to a trial-and-error method, with the added benefit that each subsequent trial is at least as successful as the one before it. But keep in mind that the machine must experience hundreds of thousands of trials and errors to learn effectively, with the frequency of errors decreasing over time. After training is finished, the machine will be able to use the sophisticated computational model it has learnt to analyze new data while still producing extremely accurate predictions.

**Chapter 2 –**

Related Work To comprehend the relationship between mental health treatment and employees' family history, gender, and workplace care alternatives, a study was carried out, according to HDRUK [25]. This has supported our study issue through data exploration and modeling, which has supplied evidence. They were certain that we have chosen a high-quality dataset using the Data Utility Framework, and that our pre-processing techniques have successfully aided in the development of a model that can predict characteristics in the dataset. Due to the project's time constraints, the biggest difficulty was being unable to compare the logistic regression with alternative models including K-nearest neighbors, neural networks, decision trees, and random forest methods.

The project's recommendations included encouraging firms to provide their staff with more mental health benefits packages. The sector should actively investigate the causes of female mental health problems and how to solve them. Finally, companies should provide psychological care for workers with a family history of mental illness priority. Companies must establish a department specifically devoted to mental health counseling and show genuine concern for the issues that affect their workers' mental health, for example, by scheduling regular psychological therapy sessions, hosting group outings, and offering the proper perspective on mental health issues. Particularly since the epidemic, there has been an alarming rise in the number of people suffering from mental illness in the United States and around the world. A few of the stressors that influence employees in the tech sector are long hours, deadline pressure, and failure dread. These stressors can cause burnout, anxiety, depression, and other mental diseases. According [21], CEOs may experience burnout and monotony if they work for the same company for a very long time. However, in the current situation, CEOs are quickly discovered to be burnt out and exhausted. This is not unique to the executive; rather, it has been the norm for the majority of those employed in the tech sector. It is obvious why they are exhausted: they are overworked and fiercely competitive. Another study [42] found that one of the main causes of death in the United States is mental disorder. Like how physical disease costs millions of dollars in direct and indirect costs, mental illness also negatively affects productivity. When a person's mental illness is not adequately treated, both the individual and the wider population suffer detrimental effects. The issue is made 4 worse by the stigma attached to mental illness. Numerous variables can have an impact on mental health. Nowadays, it is well acknowledged that job stress is a significant issue for American business. A growing number of businesses are providing intervention for stress management. However, there hasn't been enough support and direction for individuals looking to incorporate stress management exercises. According to a recent study [43], 50% of tech workers have been diagnosed with a mental disorder, though this number may understate the severity of the issue due to stigmas associated with seeking mental health care in some cultures and groups. All these figures point to a critical need for education and assistance in the tech sector. More significantly, we want to build a forum where people with mental illness may tell their stories so that other participants can understand the difficulties they encounter firsthand.

**Chapter 3 - Technical Approach**

**3.1 Project Set up**

The primary language employed for the project is Python. It was primarily selected because of Python's widespread use in data science projects and its simplicity to comprehend. The project's conda environment is expanded with the addition of the Pandas library. Conda[13] serves as a complete and reproducible virtual environment that makes it simple to set up the project's environment on any system that has Conda installed. It is simple to handle packages, dependencies, and environments for many different languages, including Python, R, Ruby, Lua, Scala, Java, Javascript, etc., when utilizing the conda environment. The language's adaptability makes it simple to use in a variety of tasks. There is just one environment for the project's code now, as shown in Figure 1. Despite this, it is still feasible to set up various environments, such as dev, prod, and test, for various uses. Figure 1 makes clear that the current environment includes a variety of helpful libraries for working on the project, including pandas, NumPy, jupyterlab, etc. Jupyterlab is also used extensively for the project's development. Project Jupyter includes Jupyterlab. A popular tool for creating computation notebooks with visualization is Jupyter Notebook. In addition to ipynb, JupyterLab is an enhancement because it offers terminals, file viewers, and custom components.

**3.2. Technology Stack**

This section discusses the technology leveraged in building this project. The project uses various libraries such as Pandas, NumPy, Seaborn, Matplotlib, etc., and the project also extensively developed on Jupyter Notebook. 1) Jupyter Notebook: A popular tool for creating computation notebooks with visualization is Jupyter Notebook. In addition to ipynb, Jupyter Lab is an enhancement because it offers terminals, file viewers, and custom components. 2) Pandas: Pandas used for data analysis and associated manipulation of Instacart data 7 3) NumPy: NumPy used for performed various operations and calculations on data 4) Matplotlib & Seaborn: Both libraries are used in project for visualization of data in form of a graph.

**3.3 Data Exploration**

**3.3.1 Data collection and exploration:**

Data for the project was gathered via OSMI an open-source data hub [8]. According to the source of the information, OSMI data was gathered from a company that monitors data on mental health. To make an accurate prediction, the two years of data are combined. It includes the 2016. There are 1431 counts of data values in the CSV file. Ultimately, the data is handled as a panda’s data frame, which is perfectly suited for handling the CSV and data analysis. There are a total of 13 customer attributes in the data that we can examine, including gender, state, nation, understanding of mental health, supervisor support for mental health, difficulty conveying mental health issues, etc. If their career is included in the data would be affected because of this mental issue untreated or uncommunicated. Figure 3 represents the data attributes of this data set. Many of the publicly accessible datasets I could discover that were good for prediction were smaller than the Mental health in the tech data [8], which was larger. Additionally, there were not many articles that used this dataset and specifically addressed this issue.

Along with investigating the length and properties, it was also determined whether there were many null values in the data, rendering them useless. Fortunately, there aren't a lot of null values in the available data. There were missing values out in a few columns. The threshold was set to remove columns with more than 50 percent of missing values of the data being absent, reducing the columns from 63 to 42.

**3.3.2 Feature Engineering**

***3.3.2.1 Cleaning of gender data***

The gender field in this dataset is a string that could have been filled with any value. As there were numerous ways to handle the non-binary genders, I divided the gender into three groups to clear up all the data. 9 I made them and placed them in 3 categories: • Male • Female • Genderqueer/ other

***3.3.2.2 Data visualization of few attributes***

The Employee Mental Health Benefits at the Current and Previous Employers column was another intriguing one to read. This enables us to determine whether the employee is aware of the advantages to mental health provided by their employer. Figure 5 demonstrates whether their employer offers mental health benefits as a component of the medical coverage.

A group of bars with different colored squares

Description automatically generated with medium confidence

**3.3.3 General Workflow**

If we want to build a machine learning system, then there are mainly three steps that need to be followed. The first step is collecting data, the second is to train the machine learning model and the third is to iterate the model many times before it starts to perform correctly.

The machine learning workflow in 3 stages.

• Gathering data

• Data pre-processing

• Researching the model that will be best for the type of data

• Training and testing the model

• Evaluation

A diagram of a machine learning life cycle

Description automatically generated

The machine learning model is simply a piece of code that an engineer or data scientist trains to become intelligent. Therefore, if you feed the model bad data, it will return bad data in the form of inaccurate or erroneous predictions from the trained model.

**Chapter 4 - Data Preprocessing**

Applying data mining methods to these noisy data would not result in high-quality findings since the algorithms would not be able to successfully discover patterns. Data processing is necessary to raise the data's overall quality. Duplicate or missing numbers can alter the data's overall statistics. Outliers and inconsistent data points frequently interfere with the model's general learning process, producing poor predictions.

Figure 9 represents step wise methods show how a data is prepared, analyzed, and communicated.

A white rectangular sign with orange text

Description automatically generated

A diagram of data processing

Description automatically generated

Data Preparation is an integral aspect of data preprocessing, and its purpose is to clean the data by completing tasks such as resolving inconsistencies, smoothing noisy data, filling in missing values, and removing outliers.

**4.1. Missing values**

Here are some solutions to this problem:

**4.1.1 Ignore these pairs**

This method should be considered when the dataset is large, and a tuple contains numerous missing values.

**4.1.2 Complete the missing data**

There are numerous ways to accomplish this, including manually filling in the values, predicting the missing values using the regression method, and numerical methods such as attribute mean.

**4.2. Removing outliers**

Clustering techniques group similar data points together. The tuples that are not contained within the cluster are outliers or inconsistent data.

**4.3. Data Analysis Schema integration and object matching:**

The data can exist in a variety of formats and attributes, which can make data integration challenging. Eliminating redundant attributes from all available data sources. Conflicts in data value detection and resolution.

**4.4. Encoding the categorical data**

Data that is divided up into various groups is referred to as categorical data. The majority of machine learning models are built using equations. Therefore, it should be intuitively clear that include categorical data in the equation will lead to certain issues. as the formulae simply requireintegers. The output from the machine learning model can be erroneous since it may assume that there is some association between the three variables. To Using Dummy Encoding, we can fix this problem right away. Dummy variables are ones that accept the values 0 or 1, respectively, to signify the lack or presence of a value. In this case, a variable's value of 1 denotes its presence in a particular column, whereas a value of 0 is given to all other variables.

**Chapter 5 – Determining the problem**

I was extremely interested in determining whether the survey allows us to predict whether a participant has sought treatment for a mental health condition. I believe it is crucial to comprehend this prediction, as the tech industry has recently placed a greater emphasis on mental health than other traditional industries. When a mental health issue arises, this would assist their employees in receiving assistance and gaining awareness.

**Chapter 6 - Metrics & Models**

A machine learning model is a program that uses a dataset that has never been seen before to find patterns or make decisions. For example, in natural language processing, machine learning algorithms may examine and precisely pinpoint the intent underlying previously unheard utterances or word combinations. You may train a machine learning model to recognize objects in photos, such cars or pets. A machine learning model that has been "trained" on a huge dataset can complete these tasks.

The machine learning algorithm is tuned during training to uncover patterns or outputs from the dataset, depending on the task. This approach produces a machine learning model, which is typically a computer program with specific rules and data structures.

**Below is a list of metrics & models used in the project:**

• Receiver Operating Characteristic (ROC)

• Naïve Bayes

• Decision Tree

• Random Forest

• Logistic Regression

**6.1 Receiver Operating Characteristic (ROC)**

Formally, each instance I is mapped to one of the positive and negative class labels in the collection p, n. A mapping from occurrences to anticipated classes is what makes up a classification model (or classifier). Various thresholds may be used to some classification models' continuous outputs (such as an estimate of an instance's class membership probability) in order to predict class membership. Other models merely produce a discrete class label that indicates the instance's expected class. We use the labels "Y, N" for the class predictions generated by a model to differentiate between the actual class and the predicted class. When a classifier and an instance are provided, there are four different outcomes that could occur. If the circumstance is good and it is categorized as such, then a true positive is recorded; nevertheless, if the categorization is bad, then a false negative is reported. In the event that the occurrence is not negative and is not labeled as such, it is regarded as a false positive; in contrast, if the occurrence is negative and is labeled as positive, it is seen as a real negative. Given a classifier and a number of examples, it is possible to construct a two-by-two confusion matrix, which is also known as a contingency table. This matrix can be used to convey the outcomes of a set of cases (the test set). This matrix serves as the basis for the construction of a great many standard metrics.

A graph with a line and a line

Description automatically generated with medium confidence

**6.3 Decision Tree**

A decision tree is a formal representation of certain mappings. A tree is either a class-labeled leaf node or a structure composed of a test node connected to two or more subtrees. A test node computes a result depending on the instance's attribute values, where each potential result is associated with a subtree. Classifying an instance begins with the root node of the tree. If this node is a test, the instance's conclusion is decided, and the process continues using the corresponding subtree. When a leaf is encountered, its label indicates the expected class of the instance. Using a divide-and-conquer approach, one can construct a decision tree from a set of instances. If all instances belong to the same class, the tree is a leaf labeled with that class. Otherwise, a test with distinct outcomes for at least two examples is selected, and the instances are partitioned based on this outcome. A node that specifies the test serves as the tree's root, and for each result, the appropriate subtree is created by applying the same procedure to the subset of instances that have that result.

A diagram of confusion matrix

Description automatically generated

**6.4 Random Forest**

The decision tree is a supervised learning method that can be used for either regression or classification. Each internal node in a tree represents a decision that partitions the dataset into subsets and each leaf node represents a subset of the dataset [26] Starting with the root node, internal nodes are added to a decision tree based on some predetermined criteria such as information gain [31]. Information gain is a measure of how important a given feature is for discriminating between classes to be learned. For example, in the case of sentiment analysis, the feature the presence of the word “terrible” would likely tell us more about the polarity of the review than the word “plastic” and thus would likely have a higher information gain. We keep adding internal nodes until the dataset has been completely partitioned into subsets by the internal nodes or another stopping condition is met. After training, a decision tree can be used to classify new data points using the internal nodes.

A diagram of a confusion matrix

Description automatically generated

Rand. Best Score: 0.8399817136886103

Rand. Best Params: {'criterion': 'entropy', 'max\_depth': 3, 'max\_features': 13, 'min\_samples\_leaf': 6, 'min\_samples\_split': 5}

[0.84, 0.842, 0.847, 0.841, 0.843, 0.84, 0.838, 0.84, 0.84, 0.84, 0.84, 0.84, 0.84, 0.843, 0.84, 0.84, 0.84, 0.84, 0.846, 0.842]

The Random Forest Classifier is a type of ensemble learning algorithm that builds upon the ideas of the decision tree. Ensemble learning algorithms are a type of algorithm that combines multiple machine learning methods into a single predictive model. The goal of ensemble methods is to overcome the shortcomings of individual methods.

The three types of ensemble methods are bagging, boosting and stacking [27]. Bagging algorithms take the average of multiple models in order to reduce the variance of the individual models [27]. Random forest is a type of bagging algorithm that creates a series of random decision trees to classify a training dataset [27]. A common drawback of decision trees is their propensity to overfit data [28]. By creating multiple decision trees, the random forest algorithm can substantially mitigate this issue. The random forest model is trained by creating multiple decision trees where each is based on a random subset of features and training data points. To classify an example, the model takes the individual predictions of the decision trees and selects the class based on a majority vote [26]. For this project the Random Forest classifier was implemented using the scikit-learn library.

**6.5 Logistic regression**

A machine learning approach called logistic regression is used to forecast binary outcomes given a set of independent factors [32]. The linear regression model is transformed nonlinearly via logistic regression [33]. The output of the linear regression model is transformed by applying the sigmoid function to it. A function called the sigmoid converts any real value to a different integer between 0 and 1. It is possible to think of the resulting number as a probability value that can be applied to categorize a data point [34]. For instance, a data item will be classified as class 1 if the likelihood that it belongs to that category is greater than 0.5. One of the most well-liked machine learning algorithms is logistic regression because of its effectiveness and simplicity. In this project, scikit-learn was used to implement logistic regression.

It is possible to compare a logistic regression model to a linear regression model, but logistic regression makes use of a cost function that is more complex and can be referred to as either the sigmoid function or the "logistic function" instead of a linear function. The logistic regression hypothesis states that the cost function is often limited to the range between 0 and 1. According to the logistic regression hypothesis, it is not possible for it to have a value greater than 1 or less than 0, hence linear functions cannot effectively represent it. Figure 17 : Logistic regression hypothesis expectation We used the hypothesis' formula while applying linear regression, i.e. hΘ(x) = β₀ + β₁X We will slightly change it for logistic regression, i.e. σ(Z) = σ(β₀ + β₁X) Our hypothesis was predicted to produce values between 0 and 1. Z = β₀ + β₁X H = sigmoid at x. (Z) For example, h(x) = 1/(1 + e-(0 + 1X)

Accuracy: 0.726063829787234

Classification Error: 0.27393617021276595

False Positive Rate: 0.25555555555555554

Precision: 0.7513513513513513

**6.7 Model Results**

A graph of different colored rectangular bars

Description automatically generated with medium confidence

**Chapter 7 – Conclusion & Future work**

Final F1 and ROC scores incline toward the possibility of the prediction. More work is required for feature importance testing and additional modeling. Additionally, far more attention needs to be paid to factors like firm size, management style, and the availability of mental health support. While certain aspects of the mental health technology field have been successful, others still require improvement. Awareness needs to be provided by the employers and should be honored the same way as a physical pain. Though this data helps us predict if a mental health treatment is needed or not, work must be done in terms of understanding the root cause for it at work. No information or proof exists on what occurs to a fresh graduate who enters the tech industry or the peer pressure that goes along with it. To better understand the pulse of young graduates who enter the IT industry and to compare them with more experienced professionals in the field, I intend to collaborate with psychologists to develop a qualitative survey.