**Mental Health Prediction in Tech**

Data Science Course Project

By:-

Purva Nartam (D1 - 04)

Satviki Pathak (D1- 11)

Sumit Rajgure (D1 - 21)

Prasad Saraf (D1 - 30)

## 

# Abstract

# The present study aims at understanding the dimensions contributing to the well-being of software professionals and mechanical professionals. The study aims to find out the level of anxiety and mental health of software and mechanical professionals. And as we know high levels of anxiety may lead to mental illnesses. Software professionals’ environment of job is highly time-bound, client-concerned and technology concentrated. The trends in turn, attached with many factors, contribute to high anxiety. These factors are extremely diverse, including change of technology, client communication, fear of uselessness, family support, long working hours, and work overload etc. The sample for the present study includes 100 professionals, 50 software and 50 mechanical professionals of both sexes. The age of the sample group was between 20 to 40 years. But significant relations were found in only the psychological dimension of mental health not in physical dimension.

# KEYWORDS : anxiety, mental health, software professionals, ROC Curve, binary classifier

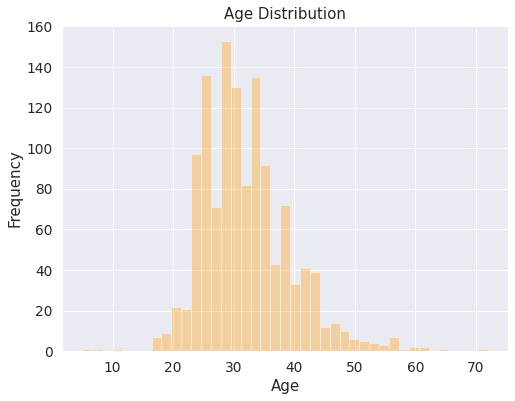
# Introduction

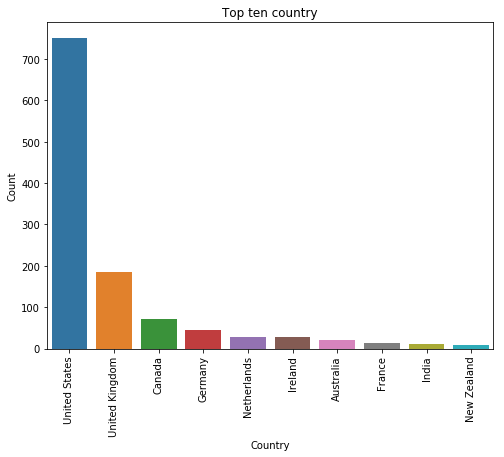
Occupational anxieties such as severe resentment, impractical expectations from superiors, being achievement oriented, lack of job safety, and the inability to accept failure have led to a host of psychological difficulties among software professionals. Clinical psychologists say that, people working in the field of information technology (IT) go through a lot of anxiety, depression and loneliness because of their work environment and often display feelings of inadequacy, lowered self-esteem and dissatisfaction. Distinctly from working with the individual to treat the psychological concerns of IT professionals, it is also essential to address the organization of a company.Some of the methods that can be used to help professionals overcome stress and help them lead a balanced life are relaxation training, cognitive therapy, and assertiveness training. Long working hours, stress and pressure at work, night shifts, and lack of sleep can lead to various mental health and physical problems. Anxiety is an unpleasant state of inner turmoil, often accompanied by nervous behavior, such as pacing back and forth, somatic complaints and rumination. Anxiety is not considered to be a normal reaction to a perceived stressor although many feel it occasionally. Anxiety disorders are psychological disorders that feature motor tension (jumpiness, trembling, inability to relax), hyperactivity (dizziness, racing heart, or possible perspiration), and expectations and thoughts. The most common five types of anxiety disorders are generalized anxiety disorder, panic disorder, phobic disorders, obsessive compulsive disorder and post-traumatic stress disorder. By measuring the level of anxiety of the professionals will help us to provide intervention programmes to them and make them feel better and perform well at their job effectively and efficiently. Mental health describes a level of psychological well-being, or an absence of a mental disorder. From the perspective of 'positive psychology' or 'holism', mental health may include an individual's ability to enjoy life, and create a balance between life activities and efforts to achieve psychological flexibility. In an organization or company if a person has to meet the goals of a job, he should have a sound mental balance. He/she should have well-balanced emotions to deal with the pressures of the work life. A person with good mental health will perform efficiently at work and also in life as a whole.

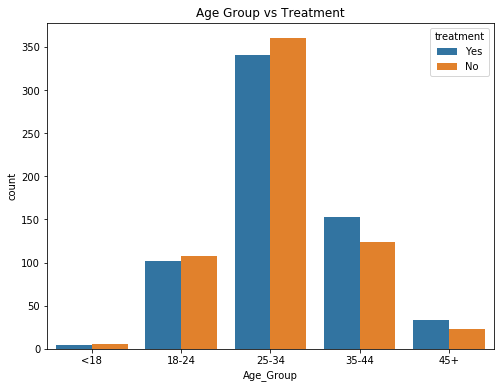
# Review Of Literature

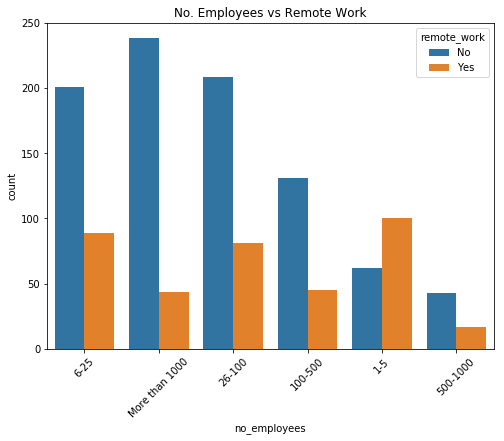
The enduring streams of information technology revolutions are transforming the business world (Laudon and Laudon, 2007) from traditional work processes to IT enabled integrated environment. The impact of this change has brought many challenges to software professionals and developers, working in organizations as in-house programmers and developers. The rise in software demand to business and industry, beyond the capacity of MIS professionals, who cater to the needs of organizations, has given birth to software houses (Rajeswari and Anantharaman, 2003). These software houses are fulfilling the demand of industry and providing customized software according to the need and requirements of the client organizations, by using latest available technology and skills in the market. The technology is changing so swiftly that it is becoming difficult for the professionals to keep abreast with the upcoming technology along with the daily chores of the workplace. Software industry is a human capital intensive industry (Rajeswari and Anantharaman, 2003) and largely based on knowledge workers with technology concentrated environment. Also, the software development process is a learning and communication process (Glass, 1997); hence, it requires greater interaction with the clients, deep understanding of the nature and business processes, clear and timely communication with people involved in the development process, and insight into technological innovations. This situation puts pressure on professionals involved in the process of software development in software houses and results in anxiety among them. There is a strong reason to believe that software professionals, working either in a software house or in any organization for in-house development and maintenance, are prone to more serious risks as compared to people involved in such jobs two or three decades ago (Brod, 1984). It has been pointed out that „high performance (requirements) with high technology can exercise a dangerous influence on the human personality ... anyone who is constantly working or playing with computers is at risk‟ (Kaluzniacky, 1998). The constant use of computers affects the users in terms of fatigue, eye strain, arm and shoulder pain, and backache. Khosrowpour and Culpan (1989) published a stress-related study applied to individuals working in computer-related fields. They observed: „Information processing professionals see change in technology as a prerequisite for their existence, yet the speed of this change can have profound psychological and physiological effects.‟

# Visualizing The Dataset

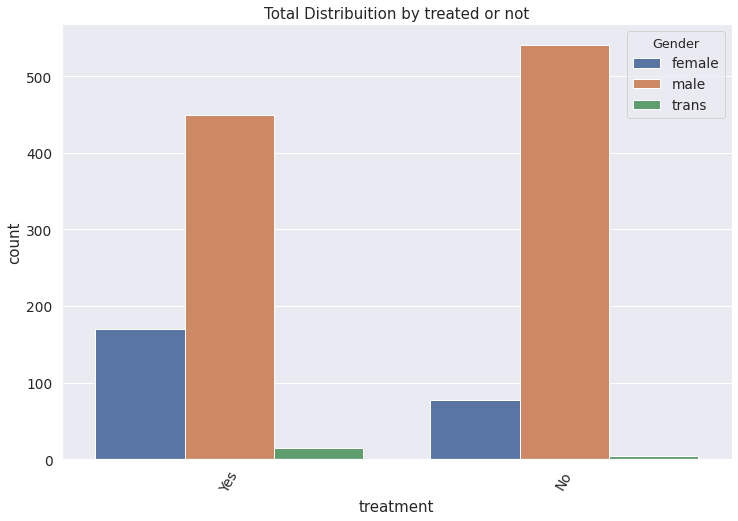
The dataset we have considered can be found [here](https://www.kaggle.com/osmi/mental-health-in-tech-survey). Considering the number of attributes considered for the prediction, various data comparisons were calculated to study the relation between them. This gives the idea about the parameters which are essential for determining causes of stress for software developers.The dataset consists of developers having age in the range of 20-45. It can be inferred from the above graph that very few people stay in the IT industry after 40-45 years of age. Most of the people are developers and do the actual coding.

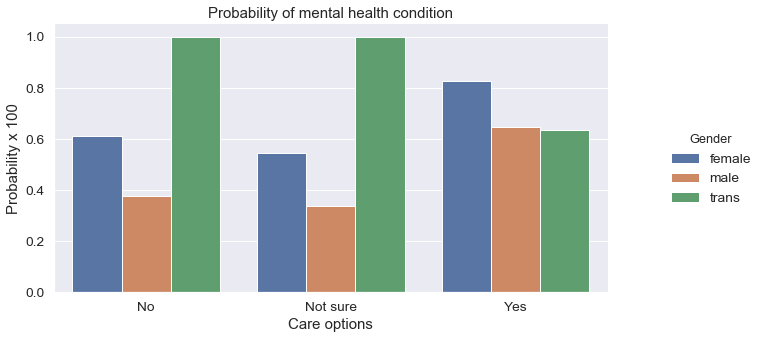
The dataset consists of 1249 developers across different countries. But most of the developers(more than 700) work in the USA. So, it can be said that the predictions done on this dataset will be more applicable to the developers working in USA. But predictions will still hold valid for developers working in other countries.

Most developers not seeking any treatment for their mental health are below 35. Also, most of the developers who are seeking treatment regarding mental health fall in the age group of 25-34. Even the developers in their early 20’s are seeking a treatment for mental health. But as age increases more and more developers opt for a mental health treatment.

Developers who work in large teams of size more than 100 don’t prefer to do remote work. The count of remote work slightly increases for smaller team sizes such as 1-5 or 6-25. For smaller there less dependencies involved and hence the developers are free to work on their own terms.

This graph shows the comparison of age groups seeking a treatment for mental health. The developers in the age group of 35-40 are actively seeking for a treatment. But still it can be inferred that developers not seeking any treatment are more in number.

This graph shows whether the developers are able to recover from the mental stress even after taking a treatment. Women are better in recovering from any mental distress as compared to men.

The graph shows that while suffering from any mental stress, developers are aware of the care options that their employer provides. The trans-gender people very unsure of the companies policies on care-options. It seems that women are very well aware of the care-options provided by their employer.

# Data Preprocessing

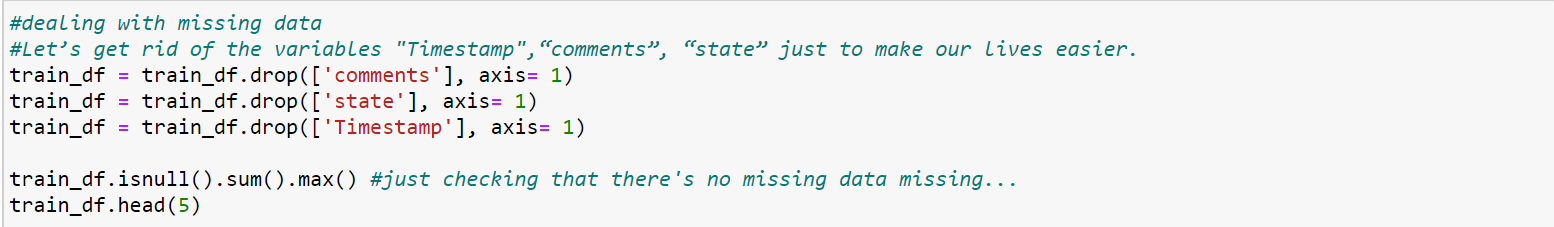
# Data preprocessing is an integral step in Data Science as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn; therefore, it is extremely important that we preprocess our data before feeding in into our model.

In this project we have used following techniques in data preprocessing,

1. Dropping attributes which are not useful.
2. Cleaning NaN data.
3. Removing inconsistency from categorical data.
4. Filling median values where values are missing

# **1. Dropping attributes which are not useful**

→ In the given dataset there are few attributes like comments, state and timestamp which are not useful in our model so we are dropping columns of those attributes.



# **2. Cleaning NaN data.**

→ In the given dataset, every attribute has some NaN(Not available) values. So in this step we are replacing every NaN value with default values according to datatype.



# **3. Removing inconsistency from categorical data.**

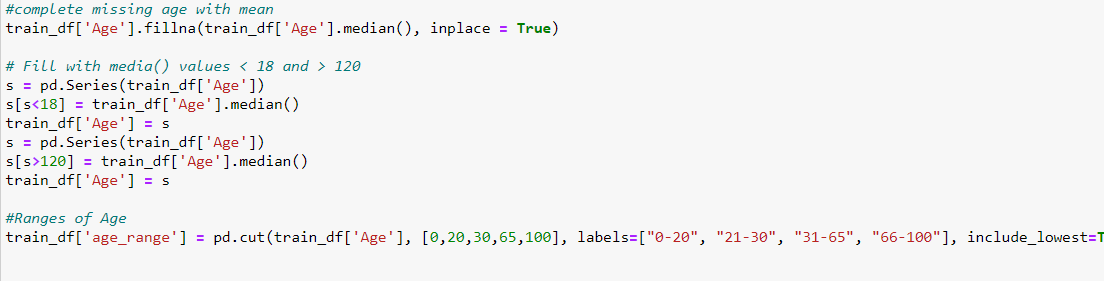
→ In the gender category there was inconsistency in the categories so in this step we have removed all inconsistent categories and kept only three categories i.e male, female and trans



# 

# **4. Filling median values where values are missing**

→ In the age attribute there were some missing values so we replaced missing values with median of age.



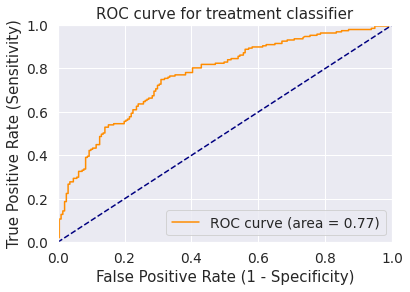
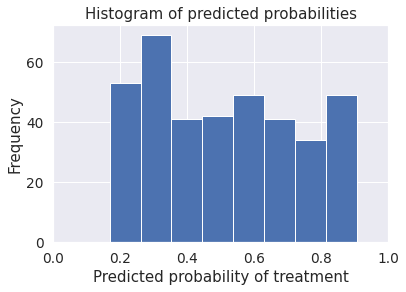
# Models

The mental fitness prediction models for potential need of treatment requires classification models. The supervised learning algorithms considered are

1. Logistic Regression
2. KNN Classifier
3. Decision Tree
4. Random Forest

ROC Curve - Receiver Operating Character Curve - is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is the graph of true positive rate vs. false positive rate. For a good classifier true positive values should be greater than false positives, false negatives or true negative values .The ROC curve for each model graphically shows the area covered by the curve , where the more the area covered, the better is the ability of the classifier.

* **Logistic Regression**
  + Advantages: performs well when the dataset is linearly separable.
  + Disadvantages: performs weakly with high dimensional datasets because of a tendency for overfitting. Assumption of linearity between dependent and independent variables is a limitation.

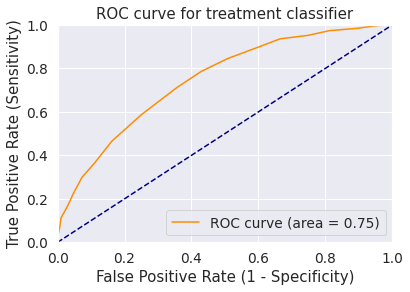
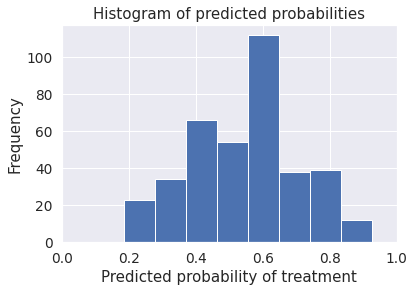


ROC Curve area for Logistic regression as a binary classifier is **0.77**

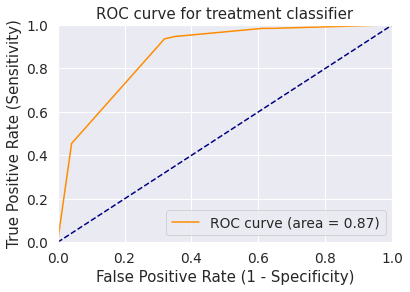
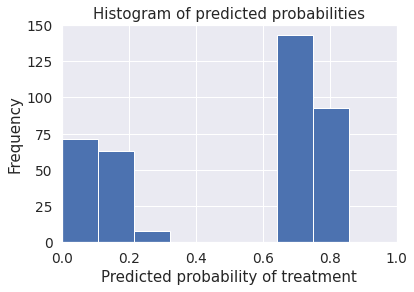
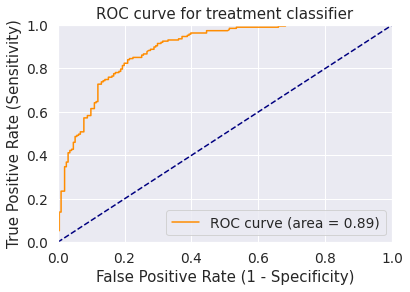
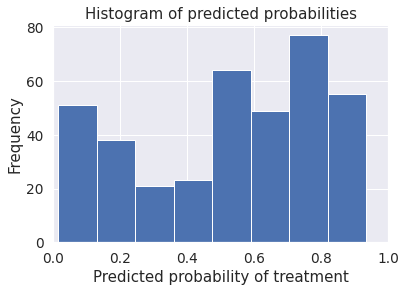
* **KNN Classifier (K nearest neighbor)**

Advantages : Performs well with less number of features and small value of K.

Disadvantages: Lazy learner. Sensitive to magnitude of distance. Large memory requirement. Complexity increases due to sorting of distance. All features are not of the same scale hence to calculate distance accurately the features need to be normalized.

* + 

ROC Curve area covered for KNN Classifier is **0.75.**

* **Decision Tree Classifier**
  + **Advantages:** Unlike KNN, data does not require normalization or scaling of data. Missing values do not affect the ability of the decision tree. Initially whole data is at the root; later it is recursively split considering each feature at each level of the tree.
  + **Disadvantages**: For a Decision tree sometimes calculation can go far more complex compared to other algorithms. It is inadequate for applying regression and predicting continuous values. It often involves higher time to train the model, especially for a high number of features.
  + 
  + The ROC Curve area covered for Decision tree is **0.87**.
* Random Forest Classifier
  + **Advantages**: Improved version of Decision tree for data high dimensions. Coalition of multiple decision trees. Can be used to decide the importance of a feature in classification.
  + **Disadvantages**: For data with different values, attributes with more values ​​will have a greater impact on random forests, so the attribute weights generated by random forests on such data are not credible.
  + 
  + ROC Curve area for Random Forest is **0.89.**
  + Not much improvement is seen from Decision trees ROC curve area. Although it is better with respect to training time and reduced complexity. The ROC curve is more smooth and continuous than Decision Tree’s ROC Curve which suggests that it tracks small parts parallely.

# Conclusion

Medical fitness is as important as physical fitness. As software jobs keep rising in the wake of digitization, the study of mental fitness with respect to the considered 27 or more factors proves beneficial for providing a solution to the job lifestyle of employees in the tech industry.

The Data visualizations provide an idea that most individuals mentally frustrated and need treatment or care options ,are male in the age ranges 30 - 40 who are not provided any care options from their organizations for stress busting.

As seen the ROC curve areas have:

**A(**Random Forest**)** > **A(**Decision Tree**)** > **A(**Logistic Reg**)** > **A(**KNN Classifier**)**

**0.89 > 0.87 > 0.77 > 0.75**

**Hence,** Performance of model of a binary classifier for a data with many features

Random Forest > Decision Tree > Logistic Regression > KNN