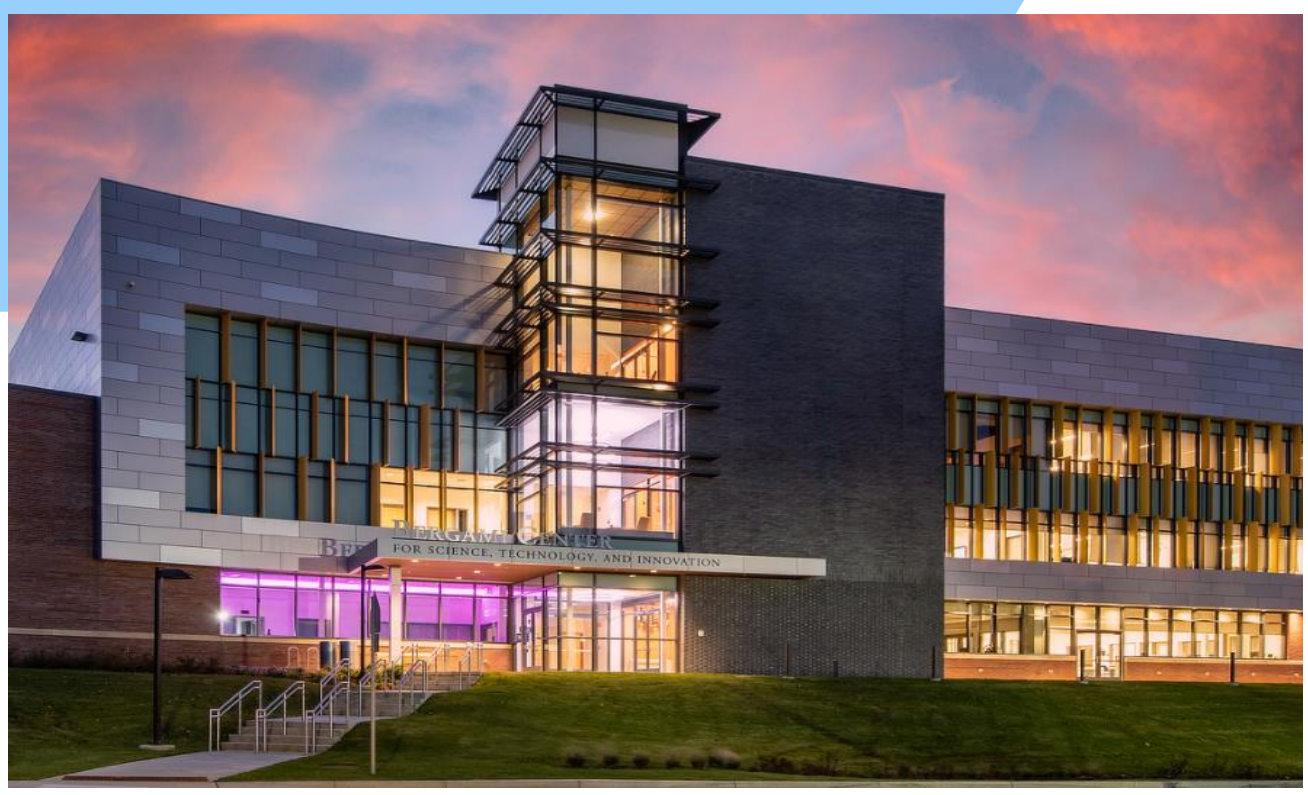
University of New Haven
TAGLIATELA COLLEGE OF ENGINEERING
Electrical & Computer Engineering and Computer Science

INTRODUCTION TO DATA SCIENCE

TECHNICAL REPORT



Team - 7: The Unstoppable
Sangeetha Mydam
Sireesha Chimbili
Raghuram Santhoshi Reddy
Naveen Kumar Nettem
Venkata Sai Harsha Linga

CONTENTS

Abstract:

Mental health is now-a-days a topic which is most frequently discussed when it comes to research but least frequently discussed when it comes to the personal life. The wellbeing of a person is the measure of mental health. The increasing use of technology will to a lifestyle of less physical work. Also, the constant pressure on an employee in any industry will make more vulnerable to mental disorder. These vulnerabilities consist of peer lead pressure, anxiety attack, depression, and many more. Here we have taken the dataset of the questionnaires which were asked to an IT industry employee. Based on their answers the result is derived. Here output will be that the person needs an attention or not. Different machine learning techniques are used to get the results. This prediction also tells us that it is very important for an IT employee to get the regular mental health check up to tract their health. The employers should have a medical service provided in their company and they should also give benefits for the affected employees.

The Tech Industry is one of the sectors where employees face many Mental disorders with the increasing Stress level. Due to COVID-19, the impact has been grown leading to even more mental health issues in the employees. This is the main reason to choose this specific challenge in predicting the health conditions of the people in Tech Industry.

According to OSMI data:

- 51% of tech professionals have been diagnosed with a mental health condition
- 71% of tech workers said their productivity is affected by a mental health issue
- 57% of tech industry employees reported burnout

Therefore, the subject of mental health in the tech industry should not be ignored.

INTRODUCTION:

Increased digital use can have a potentially detrimental mental health impact on you, particularly in areas of reduced social skills, increased cyber bullying and heightened isolation. When you use technology in excess, it can result in depression and more.

Each year, there are millions of people in the United States who are struggling with a mental health condition. Below are some statistics regarding the impact and prevalence of mental illness as well as digital use involving mental health:

- Each year, there are around 43.8 million American adults, or 18.5 percent, that suffer with a mental illness. That's around one in five adults.
- The health care industry now has over 165,000 mobile apps, with mental-health disorders being the biggest category, helping people with a wide range of issues from depression to addiction.

Mental health is a measure of handling stress and decision making with every step in life. Mental Health is very important factor in every stage in life whether it be childhood or an adult. Mostly mental health is something which never discussed publicly, and no proper awareness is there in society. People would generally not talk about it in public. Mental health could affect one's thinking and behavior. Some common reasons of instable mental health could be:

• Past life experiences, such as ragging or bullying

• Biological factors, such as genes

• Hereditary problem from family Mental health problems are very common if people accepts it and they can get better over it. If someone is experiencing few below mentioned symptoms, then they might have mental health problem.

• Too less or too much of sleep

• Going away from usual routine and people

• Sudden increase in intake of drugs

• Severe mood swings

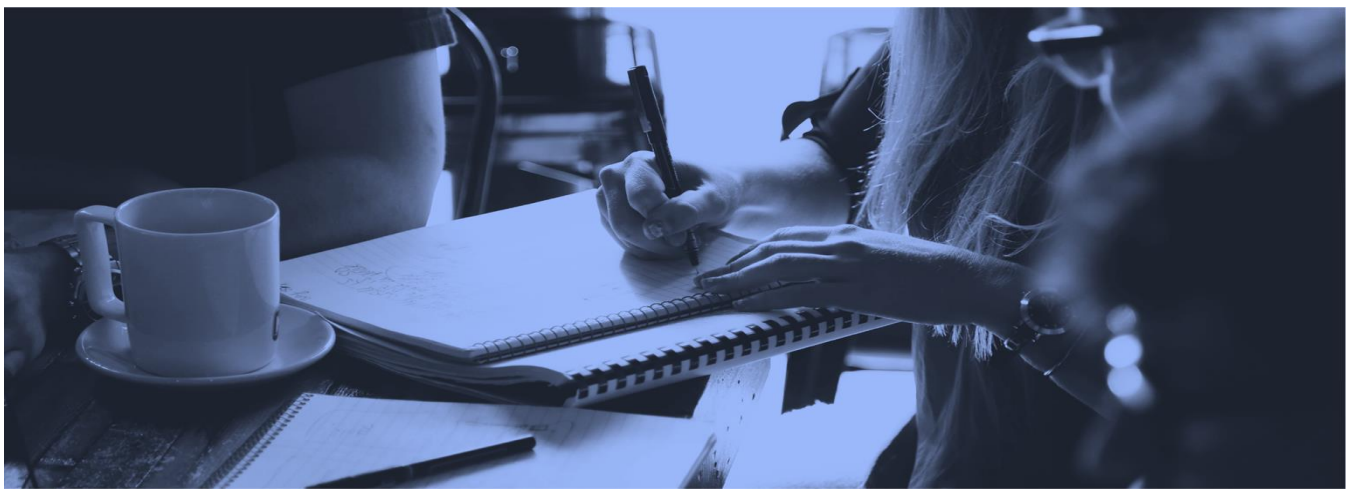
• Thinking to harm yourself Positive mental health will support person to have full potential. It will also him to come up with the stress at home and workplace. It will increase productivity of the people. Now to maintain this positive mental health one should get help from others, connect with others, help others, have usual routine and develop coping skills.

Executive Summary:

- Our main objective is to develop a model to predict the chances of having the disorders of a person working in the Tech Industry based on their inputs.
- Handled missing values.
- Replaced Null values with the most frequent values or average of the values.
- Transformed data in a consistent way by removing the features having less number of responses.
- Worked on identifying Outliers of the features.
- Encoded the data of Categorical values to Numerical Values Using One hot encoder, Label Encoder.
- Used Heat map of correlations to identify the most important features.
- Used Data Visualization to analyse the impact of the people having past disorders and the impact of the mental health on their work life.
- Introduced a feature called “tech_flag” to categorize if the employee is tech/non-tech using the work position of the employee.
- split the data to training and testing with 70% of the training data and 30% of the data to testing data.
- Scaled the data using standard scalar to improve the accuracy of the model.

Highlights of the project:

- Mental Health Dataset has been taken from the Kaggle.
- We have used CRISP methodology as it is satisfying the business challenge.
- For modelling we have used random forest classifier and Logistic Regression to build the model and accuracy score and Confusion Matrix to evaluate the model.
- We have deployed the model and app into the web using Flask.

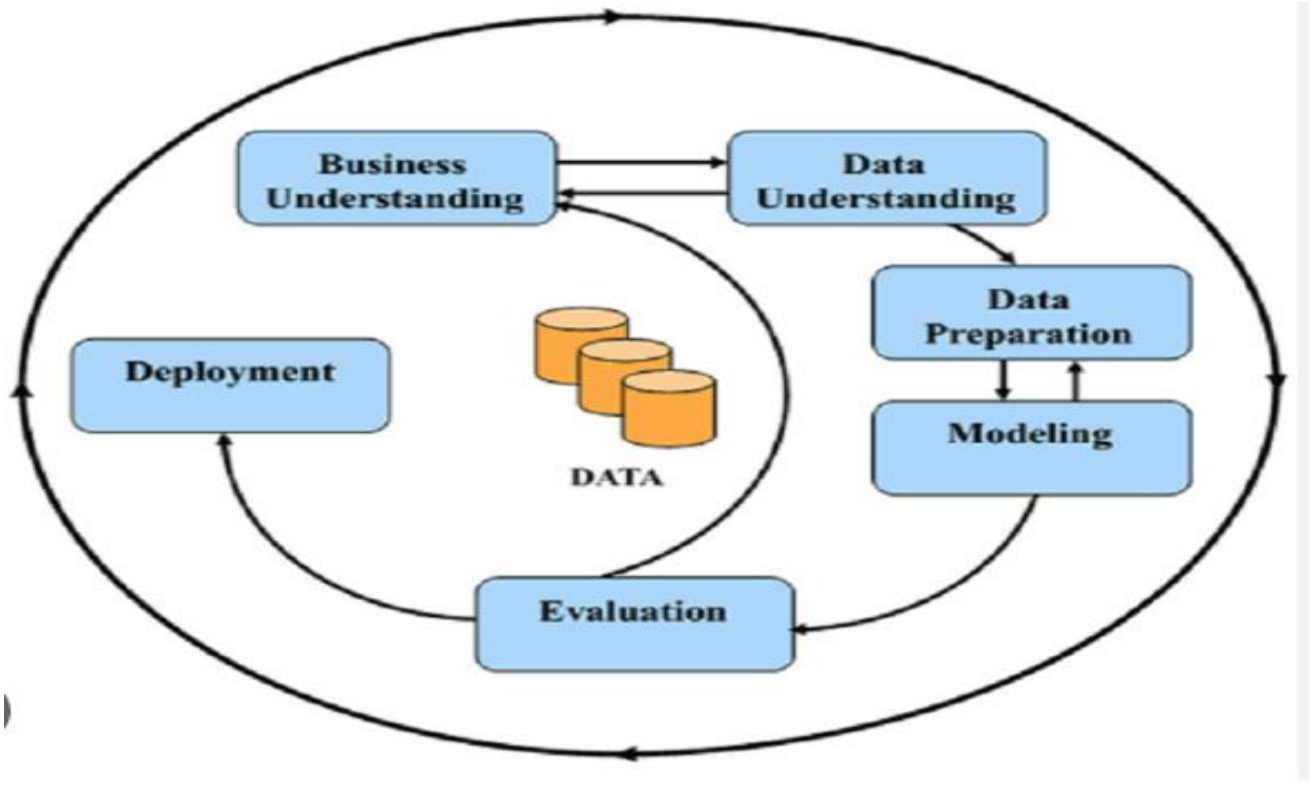


Methodology:

The project involved CRISP methodology, which includes:

- Business understanding
- Data understanding
- Data preparation
- Modelling
- Evaluation
- Deployment

This is the CRISP methodology, where it involves the understanding of the challenge and the data then, building the model using the data through which it is evaluated as if it is satisfying the business challenge and making it deployed.



Business Understanding

Through the survey from the employees, we have noticed that there are different Mental disorders for a different set of people. So, this model is to predict the chances of having the disorders based on the inputs provided by the user.

Data Understanding:

We have collected the data from Kaggle and we used the relevant attributes like tech company status, gender, work Position, Health Coverage etc… to build the model.

Data Preparation:

We have prepared the data like replacing the null values to appropriate values, segregating the data according to their age groups and maintaining the consistent data with the data types etc.

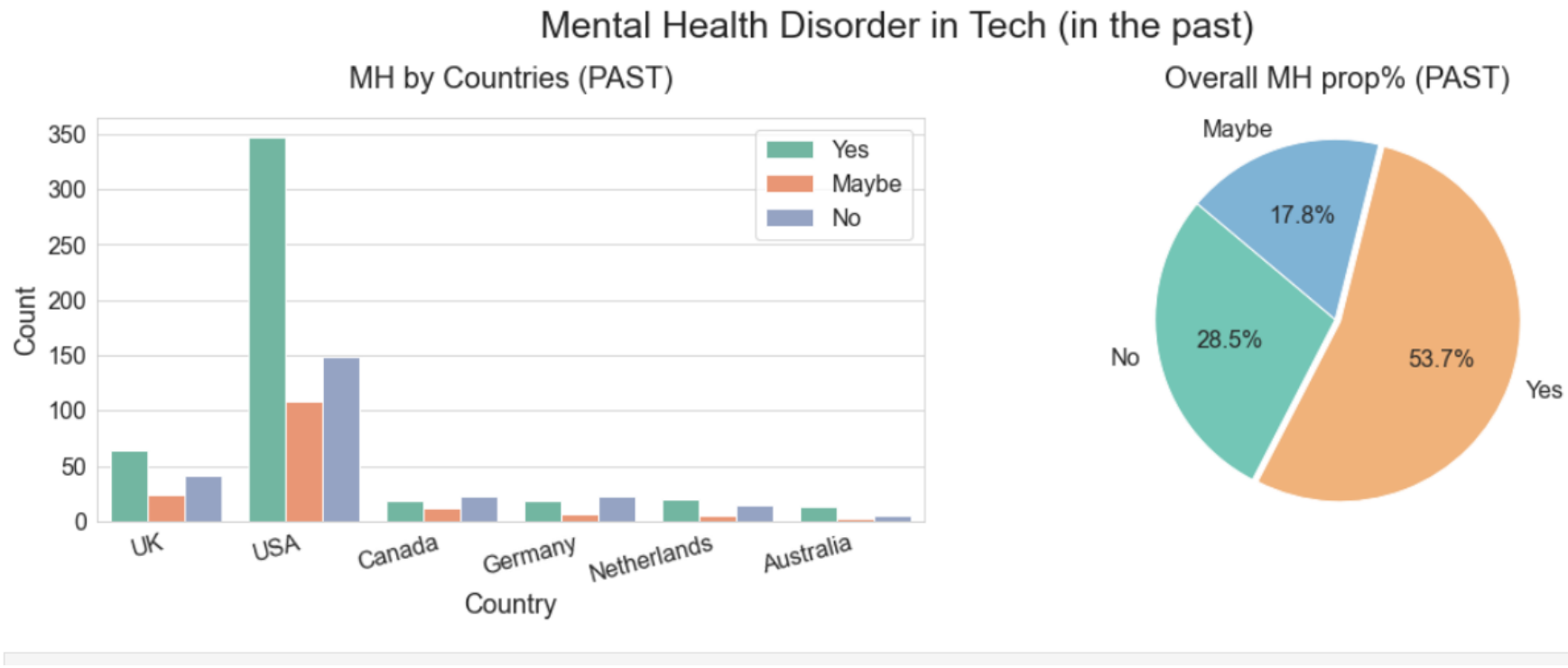


Fig: Visualized the disorders in the Tech Industry in the past based on the survey data

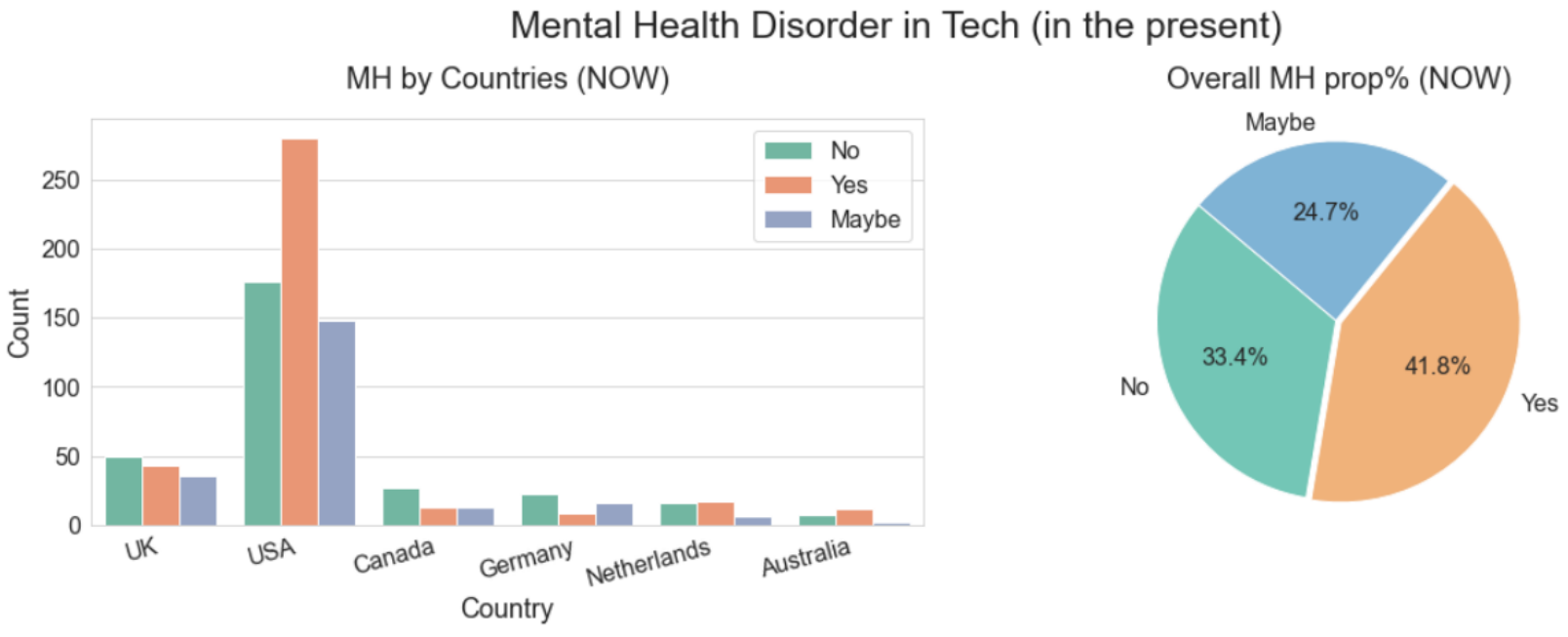


Fig: Visualized the disorders in the Tech Industry in the present based on the survey data

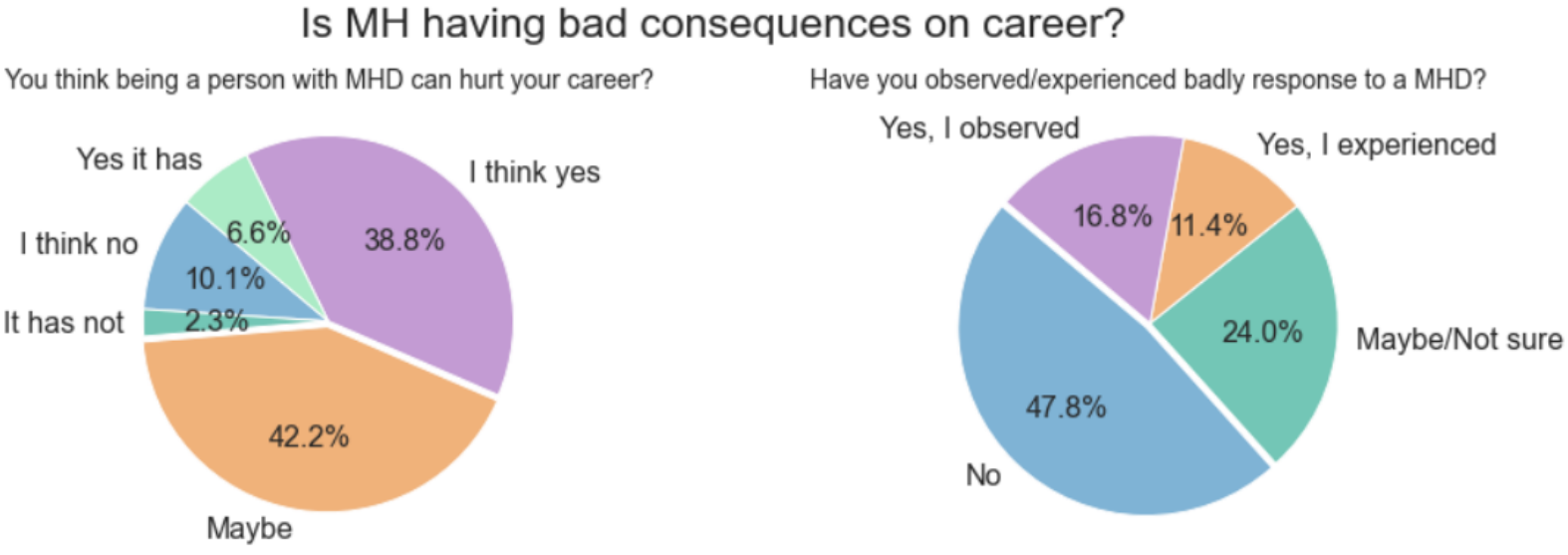


Fig: Visualized the impact of the employees' disorders on the career.

Modeling:

We have used Logistic Regression and Random Forest Classifier to build the model and our target is to achieve the chances of having the mental disorders from the input data.

Evaluation:
The accuracy score has been used to evaluate the model. 70% of the data has been used for training and 30% is to test the model and we have achieved 74.369% accuracy and it can be improved by collecting the new inputs.

Deployment:
Flask has been used to deploy the model and can be accessed through the link provided
http://127.0.0.1:5000/disorderPredict

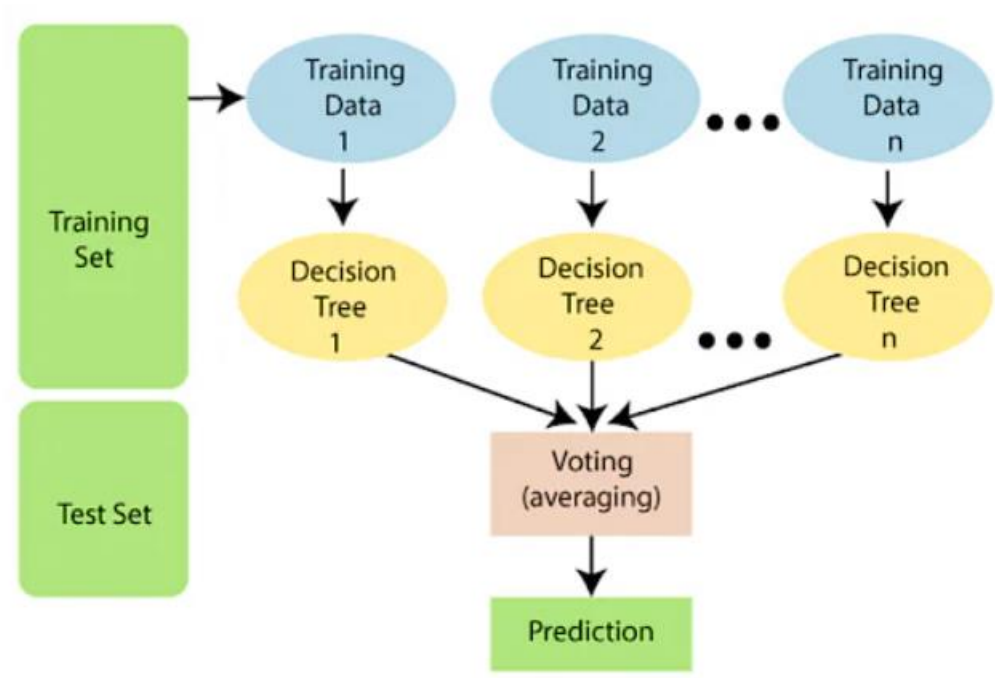
Models used:

- Random Forest Classifier.
- Logistic Regression.

Why have we used random forest algorithm?

- There are a lot of benefits to using Random Forest Algorithm, but one of the main advantages is that it reduces the risk of overfitting and the required training time. Additionally, it offers a high level of accuracy.
- Random Forest algorithm runs efficiently in large databases and produces highly accurate predictions by estimating missing data.
- It can be used for both regression and classification types of problems. It is easy to use.
- Overfitting of the dataset is not a problem in the random forest algorithm.

Working of random Forest algorithm:



The following steps explain the working Random Forest Algorithm:

Step 1: Select random samples from a given data or training set.

Step 2: This algorithm will construct a decision tree for every training data.

Step 3: Voting will take place by averaging the decision tree.

Step 4: Finally, select the most voted prediction result as the final prediction result.

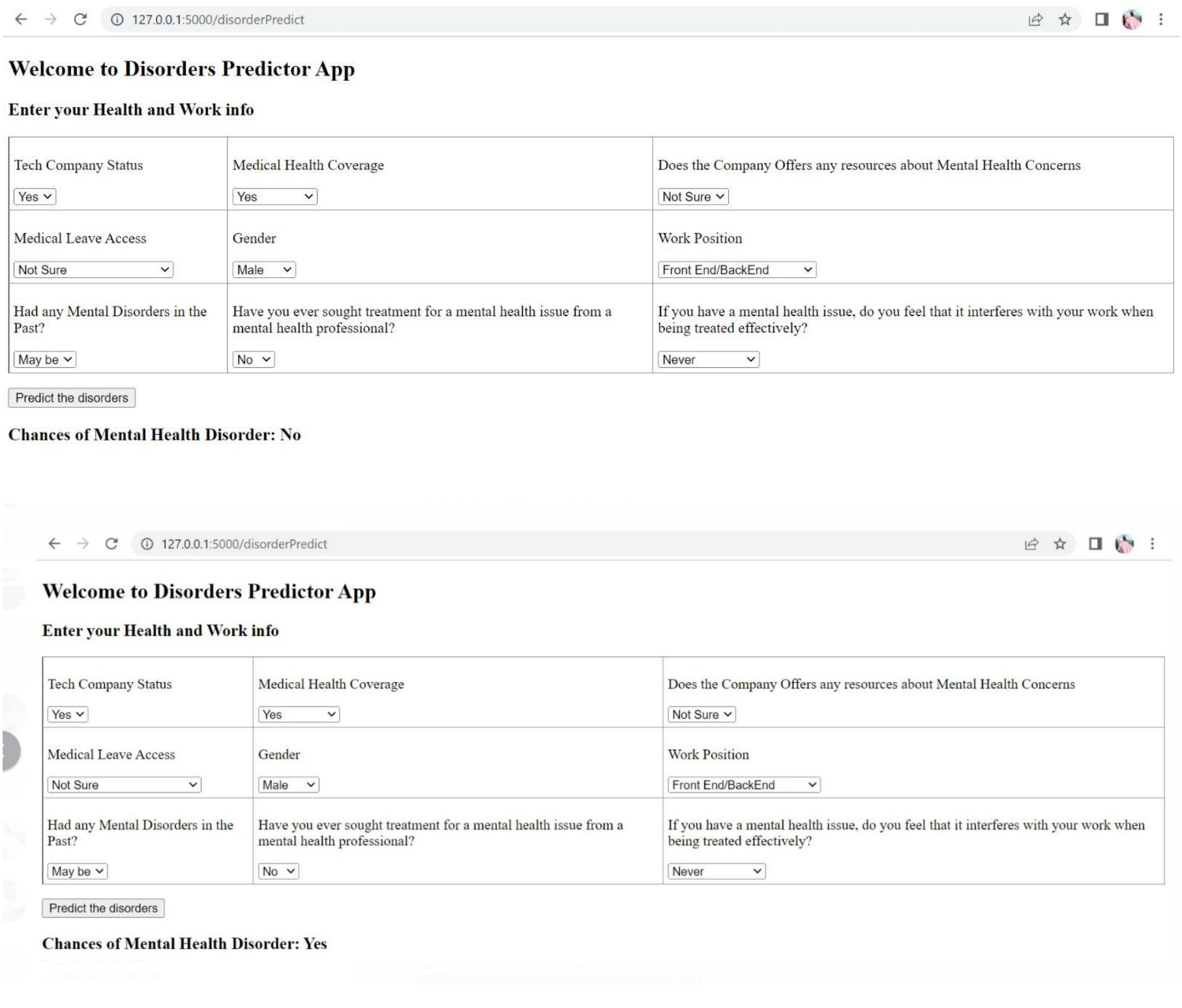
Results:

Here while applying these techniques to dataset, we also evaluate methods by following metrics:

- Classification Accuracy: it is a measure of correct predictions in percentage
- Confusion Matrix: Tabular format to describe performance of model o True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN)
- Precision: it measures how often correct positive value is predicted.
- AUC Score: it is area under ROC (Receiver Operating Characteristic) curve in percentage.

Model	Accuracy (%)	Confusion Matrix									
		00	01	02	10	11	12	20	21	22	
Logistic Regression	72.776	24	26	20	12	108	23	9	11	138	
Random Forest	74.394	33	13	24	16	105	22	8	12	138	

Here are few snapshots of working of the app using the model designed.



Conclusion:

A prediction model was developed to predict the chances of having the disorders of a person working in the Tech Industry based on their inputs. Using the data set, the proposed model can predict the chances of having disorders with more accuracy. This result showed better performance comparing to the other existing approach.

A user-friendly application has been developed for end users based on the proposed prediction model so that any individual can use the application to predict the chances of having disorders in Tech Industry easily. This outcome indicated an extension of many other existing works, since most of the existing works mainly focus on developing and comparing the performance of prediction model or techniques and did not expend to develop any mobile application for end users.

In sum, the outcome of this research provides an effective and efficient approach to detect the disorders of Employees in Tech Industry. This type of model could be used to detect metal health progress among employees and also could lead to policy changes. Employees could talk to colleagues and their managers about their problem freely. Hence upper management could help them to get correct aid with beneficiaries like work from home, flexible timings, more leaves, many more. Employees should know health benefits provided by their organization participate in any wellness programs. Proper feedback should be provided to employee when they resign from their job. This could help them to improve their health.

Contributions/References:

1. G. Azar, C. Gloster, N. El-Bathy, S. Yu, R. H. Neela and I. Alothman, "Intelligent data mining and machine learning for mental health diagnosis using genetic algorithm," 2015 IEEE International Conference on Electro/Information Technology (EIT), Dekalb, IL, 2015, pp. 201-206.

2. B. Saha, T. Nguyen, D. Phung and S. Venkatesh, "A Framework for Classifying Online Mental Health-Related Communities with an Interest in Depression," in IEEE Journal of Biomedical and Health Informatics, vol. 20, no. 4, pp. 1008-1015, July 2016.

3. T. Simms, C. Ramstedt, M. Rich, M. Richards, T. Martinez and C. Giraud-Carrier, "Detecting Cognitive Distortions Through Machine Learning Text Analytics," 2017 IEEE International Conference on Healthcare Informatics (ICHI), Park City, UT, 2017, pp. 508-512.

4. A. R. Subhani, W. Mumtaz, M. N. B. M. Saad, N. Kamel and A. S. Malik, "Machine Learning Framework for the Detection of Mental Stress at Multiple Levels," in IEEE Access, vol. 5, pp. 13545-13556, 2017.

5. R. M. Khalil and A. Al-Jumaily, "Machine learning based prediction of depression among type 2 diabetic patients," 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Nanjing, 2017, pp. 1-5.

6. Srividya, M & Subramaniam, Mohanavalli & Natarajan, Bhalaji. (2018). Behavioral Modeling for Mental Health using Machine Learning Algorithms. Journal of Medical Systems. 42. 88. 10.1007/s10916-018-0934-5.