**Mental Health in Tech Workspace**

**Project Report**

**Introduction:**

In this project we are determining how technology impact our mental health like social comparison, feelings of missing out, and cyberbullying all stem from the content we see online. These negative impacts lead to **more depression and anxiety**. Our bodies experience the negative effects of technology, too. More screen time can disrupt sleep, especially if this screen time is before bed. We were motivated by the Connecting Communities theme to create a project that addresses a subject of importance to college students throughout the world who are studying in technology and seeking employment in the IT sector. The treatment of mental health issues in the tech workplace is much more challenging because most people are reluctant to talk about their issues. We want to raise as much awareness as we can about this important subject and give new employees confidence that there is someone who will listen to their difficulties and offer support.

**Methods:**

The method we performed in our project is first we did data cleaning in which we

* Checked for the null values in the dataset but we found four variables have null values; comments, state is contributing to a greater percentage of null values.
* Dropped two columns (Timestamp & comments) to ease our lives while predicting our target. We choose to drop these two columns as 75% of cells are null values also it will be of no use in the prediction of the disease.
* Then, we checked for duplicates and removed them.
* We found outliers in Age variable in our dataset, and we removed certain rows of age based on the criteria specified as this is the most important column for prediction target.

Performed the Exploratory Data Analysis (EDA) to summarize the main characteristics of the data visually Plotted the Bar graph, Distribution plot and Count plot.

**Logistic Regression:**

Although the name of the model includes the word "regression," Logistic Regression is a supervised classification algorithm. This classifier works well with binary class datasets.

The logistic regression model is imported and fitted using the training data for the 80:20 splits of the data. Then, metrics like recall, recall accuracy, precision, and f-measure are calculated. For the model created with an 80:20 split of the data, a confusion matrix is also plotted.

**KNN Classifier:**

KNN Classifier is a straightforward supervised machine learning technique that classifies data points based on the Euclidean distance between the test data point and each individual data point.

We have considered 9 neighbours for the KNN classifier model, and the model is fitted using the training data. The testing set is then given to the model so it can make a prediction. A confusion matrix is then presented for the model created using the default split (80:20) of the data after accuracy, precision, recall, and f-measure are computed for both divides of the data.

**Random Forest:**

An ensemble machine learning approach called Random Forest Classifier creates a series of decision trees from data samples that are randomly chosen.

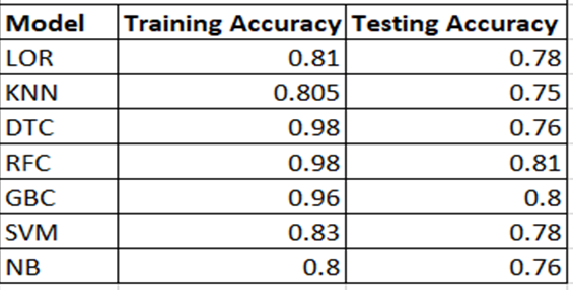
The model is first trained using the training set after importing the random forest classifier. For both the default and 80:20 splits of the data, accuracy, precision, recall, and f-measure of the model are computed, and a confusion matrix is plotted. Then, using gridsearchcv method, a parameter grid and 10 cross folds are passed to both the random forest models to implement hyper parameter tuning. Calculating the model accuracies, obtaining the optimal parameters, and Confusion matrices for the models are displayed.

**Navie Bayes Classifier:**

One of the supervised machine learning methods that uses the conditional probability concept is the naive bayes classifier.

This classifier performs at minimal compute cost and utilises large datasets effectively. The models were created with data splits of 80:20 using the procedures listed below. A training set of data is imported and fitted to the Naive Bayes classifier. The testing set is then given to the model so it can make a prediction. Following that, the model's accuracy, precision, recall, and f-measure are calculated, and a confusion matrix is presented for the model created using the standard (80:20) split of the data.

**Results:**



The dataframe shown in the image above contains the names of the models we have created and their degrees of accuracy for the 80:20 splits of the data.

**For 80:20 split:**

For 80:20 split: For this split, the accuracies for logistic regression classifier model with best parameters and default parameters are 82% and 78% respectively. The accuracies for KNN Classifier model with best parameters and default parameters are 81% and 75% respectively. The accuracies for the models; Decision tree, Random Forest, Gradient boosting, SVM Classifier, KNN Classifier, and Naïve Bayes classifier, are 98%, 98%, 96%, 83% and 80% respectively.

**Discussion:**

The Connecting Communities theme inspired us to develop a project that speaks to a topic important to college students worldwide who are majoring in technology and looking for work in the IT industry. Most people are unwilling to communicate about their concerns, making it far more difficult to manage mental health issues in the tech sector.

In order to prepare our data, we removed null values, reduced some columns, looked for outliers, looked for duplicates, and eliminated them altogether.

Carried out the Exploratory Data Analysis (EDA) to visually summarise the key aspects of the data Bar graph, distribution plot, and count plot were all plotted.

A bar graph was plotted to determine the relationship between the features. examined the effects of work interference on the various age groups.

To visualize our result properly we plot some graph:

Distribution and Density by Age: We discovered that accurately reflects the age distributions for all the survey samples. Most of the people, as far as we can tell, are in their early 30s or late 20s.

Work interfere and count by Age: The conflict between job and mental illness is illustrated in this graph and we can see, the median age for all groups is in their 30s, and occasionally, they do feel like their disease is harming their ability to work.

Respondents by Gender: This graph displays the gender distribution of survey responses. Respondents who are men predominate. Because there are twice as many female respondents as male respondents, survey results may be skewed.

Does work from home effects mental health: The association between persons who work from home and how they perceive receiving mental health treatment is examined in this bar plot. People who work from home plainly demonstrate that this has very little bearing on their mental health, yet the majority of respondents concurred that this does not apply to them.

Does self employed effects mental health: This plot shows that being self-employed has no impact on mental health. A very tiny percentage of respondents to this question said "yes," indicating that they are dealing with mental health concerns.

**Conclusion:**

After comparing our machine learning models' accuracy on the data set before and after normalisation, we found that they performed quite well. Highest accuracy with a good training results are obtained for Logistic regression with best parameters on both before and after normalization process. So, we consider this model is the best classifier model to this kind of dataset. Future developments in machine learning models will enable the automatic early detection of most diseases. Additionally, it will improve illness detection's effectiveness and precision.

**Contributions:**

|  |  |  |
| --- | --- | --- |
| **Student** | **ID** | **Contributions** |
| Deekshith | 0796575 | I have carried out data preprocessing (looked for any missing data).Exploratory Data Analysis, including Detecting Duplicates and Outliers,  contributed remarks to the Jupyter notebook and Related Work the project report's sections |
| Pavan | 0796575 | I created KNN and SVM classifier models. I helped with dataset documentation, data preparation, and report's EDA parts |
| Chaitanya Pasupulate | 0796910 | I created the gradient boosting and random forest classifiers and also performed  hyper parameter adjustment  to increase accuracy and achieve the ideal parameters. I participated in the report Methods section's documentation. |
| Dilip Reddy sane | 0790735 | I developed the Nave Bayes classifier and logistic regression models. I have contributed to the results, documentation.  The project report's Conclusion and Discussion sections. |
| Shweta | 0795669 | I participated in pre-processing part like removing duplicate value, identified and dropped two columns and also in Exploratory Data Analysis. I have build Decision tree model also. |

Deekshith (0796575) – Data analysis, visualizations & ML models

Pavan (0796575) – ML models

Chaitanya Pasupulate (0796910) – Data Cleaning, EDA & Modelling

Dilip Reddy sane (0790735) – Data cleaning, EDA & Feature Importance

Shweta (0795669) – ML models

**Reference:**

**NYC data science:** [**https://nycdatascience.com/blog/student-works/data-survey-on-mental-health-in-tech-industry/**](https://nycdatascience.com/blog/student-works/data-survey-on-mental-health-in-tech-industry/)

**Kaggle:** [**https://www.kaggle.com/datasets/anth7310/mental-health-in-the-tech-industry**](https://www.kaggle.com/datasets/anth7310/mental-health-in-the-tech-industry)

**Blog:** [**https://osmhhelp.org/about/blog**](https://osmhhelp.org/about/blog)

**EDA:** [**https://towardsdatascience.com/exploratory-data-analysis-eda-python-87178e35b14**](https://towardsdatascience.com/exploratory-data-analysis-eda-python-87178e35b14)

**Appendices:**

* DAB 304 - Healthcare Analytics Final Project - Healthcare\_project.ipynb
* The above file consists of all the code for reading the data, EDA, Data Pre-processing, and models building and evaluation.
* Dataset File:
* It has all the features and target data for Mental health in tech world.