Affect on Mental Health in the Tech Industry

Istihad Ahmed ID22241144, Abeda Zahid Chandrica ID22341061

***Abstract*—As more and more people are joining the technical industries over time, it is a matter of huge concern how this shift from the world of education to the fast-paced corporate world has been affecting people’s mental health. The corporate sector is often very unforgiving; if one cannot perform enough, there are hundreds of alternate options ready to replace him. The main transformation for someone coming freshly out of the academic nurture is that, until then all their fights were firmly focused on dragging themselves through but now everything one does is for someone else’s benefit - to keep the ones with more power happy. Such competitive approach almost always takes a heavy stress upon people and secretly they end up developing mental health issues. For this paper, we tried to draw a classification model on whether a random tech industry employee has been suffering from mental health issues or not.**

I. Introduction

IN this project, the core objective is to be able to judge if a person is having any mental health issues currently which might have links to his early life, family background, work environment etc. Here we used a dataset from a survey carried by OSMI, a non-profit and open source mental health community which has tried to reach as many people as possible over the last decade. The survey was designed to get information from the employees of the tech industry about themselves, their family background, how they feel about their mental state and if they had consulted any professional to seek help, etc. The survey had 6 different question groups each originating from different aspects that ensured responses to be more distributed - ’Speaking openly about a mental health disorder with employer’, ’Safe and supportive workplace for those with mental health issues’, ’Speaking openly about a mental health disorder’, ’Resources for employees with mental health disorders’, ’Impact of mental disorders on worker productivity’, ’Negative experience after speaking openly about mental health disorder’.

We trained this classification problem, as the response of the model would be ’True’ or ’False’ if a person has mental issues, by the massive pool of data from the above mentioned survey. Using classification models like, K-Nearest Neighbor (KNN), Decision Tree and Logistic Regression we planned on recovering the best-fitting model for the problem. We also demonstrated our findings through heatmaps, bar charts and other visual stances.

II. Pre-processing

OUR project used the dataset of “OSMI Survey Data” that contains all the 60186 responses of people who were surveyed, with a total of 28 features in the dataset such as: ’Are you self employed’, ’Do you have a family history of mental illness’, ’Have you been diagnosed with a mental health condition by a medical professional’, ’Do you work remotely’. In total we had a total of 60186x28 data points.

There were few missing data for the question ’What is your age’ and several missing responses for ’Is your employer primarily a tech company organization’ which was a categorical feature. To get more accurate results using the model, we had to delete the responses with these missing data.

Out of the total 28 features, most of which are categorical data, the questions like ”index”, ”ResponseID”, ”How many employees does your company or organization have”, ”Is your primary role within your company related to techIT”, ”What country do you live in”, ”What US state or territory do you live in”, ”What country do you work in”, ”What US state or territory do you work in”, ”Question Group”, ”Question”, ”Response”, etc. had no direct relevance to the idea of our project. This made us remove these features. We used heatmap to help us decide some of which features to remove.

Few other questions, which are basically extensions of the previous questions of the survey detailing out the mental health conditions, ”If yes, what conditions have you been diagnosed with”, ”If maybe, what conditions do you believe you have”, ”If so, what conditions were you diagnosed with”, ”Which of the following best describes your work position” had too many unique responses that even encoding them would not be effective enough for them to have the required impact on the model. Therefore we had no other options but to get rid of them as well, leaving us with a total of 48048x10 datapoints.

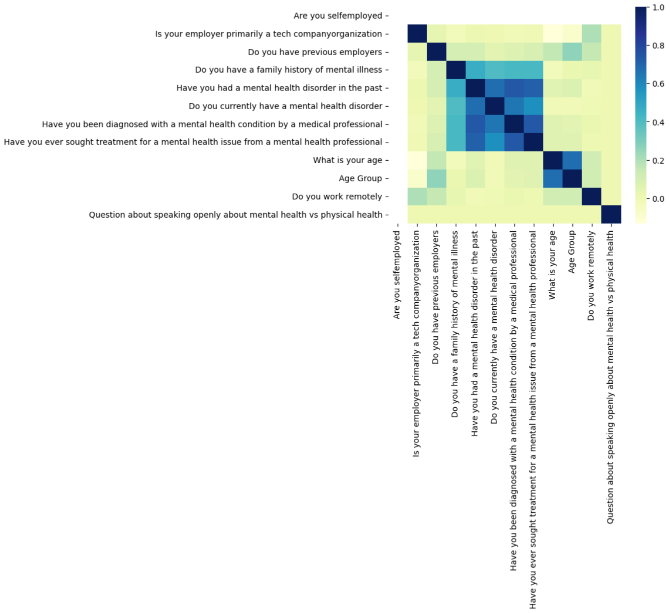

Fig. 1. Heatmap of input and output features

Finally, before proceeding to apply models, we encoded the categorical features like, ”Is your employer primarily a tech company organization”, ”Do you have a family history of mental illness”, ”Have you had a mental health disorder in the past”, ”Do you currently have a mental health disorder”, ”Age Group”, ”Do you work remotely” into integer values (assigning weights) from strings. Due to lack of quantitative values in the dataset, we did not require to use scaling.

III. Methodology

AMONG the 10 features we finally settled into 9 are our input features and 1, “Do you currently have a mental health disorder”, is our output feature. So to split this dataset we have used the Stratified splitting method to ensure that the data splitting goes unbiased. Here we kept the training and testing set to a 80-20 percent ratio respectively. Here X_train and y_train are the training input and output set whereas the X_test and y_test are the test set consisting of 20% of the total data.

Now, since almost all of the features in the dataset were quantitative or boolean, following some categorical data being encoded to numbers, we could attempt to apply the data models. As stated before, our required output is discrete and categorical, which means we were looking for ’Yes’, ’No’ or at best a ’Maybe’ as the conclusion. Therefore, we needed data models that would accurately predict categorical data. Thus, our choices were made with respect to which ones fit best with this case. The data models we implemented are all Classification Models:

1. Decision Tree: Decision tree is the first model we applied. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. The output is based on the features that give us the highest information gain. The Decision Tree was mainly used due to its ability to predict categorical outputs. This model had the highest accuracy in comparison to the other data models we attempted.

2. KNN - (K-Nearest Neighbors): The K - Nearest Neighbour Algorithm is our next model. In this model, the algorithm identifies the nearest neighbors to a given point and classifies them into a label fitting with its nearest neighbours to help make a prediction. This is done based on the calculated distance (Euclidean Distance) of neighbor nodes from the given node and then it is classified by a vote of its neighbours and assigned to to a class of its k nearest neighbors.

3. Logistic Regression: The last model we used is the Logistic Regression. This is used when the dependent variable is dichotomous or binary. It is used for predicting the categorical dependent variable using a given set of independent variables. In Logistic regression, instead of fitting a regression line, we fit an “S” shaped logistic function, which predicts the minimum and maximum values. For this dataset, this model had the lowest accuracy in comparison to the other data models we attempted.

IV. Results

AFTER applying the models using our train dataset, we used each of them to predict the output. From there, we measure the performance metrics of the data model in terms of accuracy - which measures how close the test outputs were in comparison to pre-existing the train dataset. Confusion matrix was shown in the colab file, as a means to represent the values.

As per the models we applied, accuracy scores were 78% for Decision Tree, 71% for KNN and lastly, 68% for Logistic Regression.

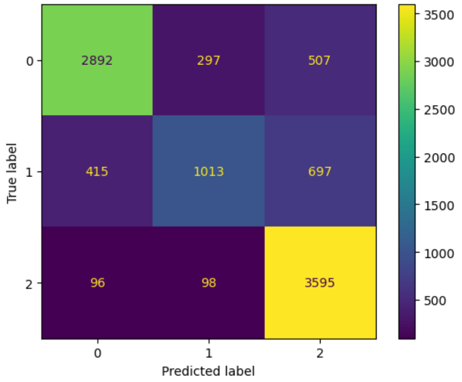

Fig. 2. Decision Tree Confusion Matrix

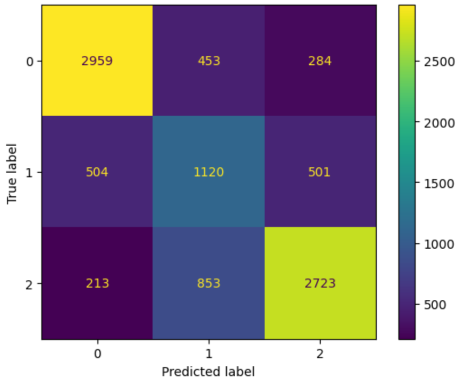

Fig. 3. KNN Confusion Matrix

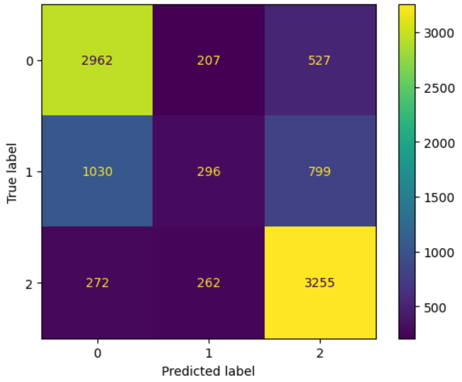

Fig. 4. Logistic Regression Confusion Matrix

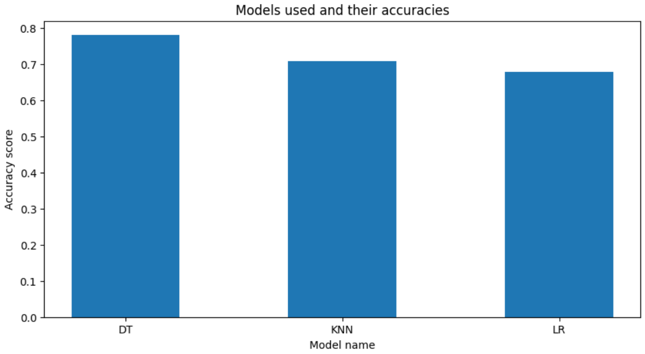

Fig. 5. Bar chart showcasing prediction accuracy of all models.

V. Conclusion

OUT of the three tested models, the Decision Tree model had the highest accuracy in comparison to all other models with logistic regression being the least accurate. Overall lack of quantitative data in the dataset might have effected the accuracy scores. Therefore, we found that the Decision Tree model was best suited for our given dataset.