Mental Health Analysis in the Workforce

Aisha Kothare, Jacqueline Girouard and Joanna Ng

Abstract

Tech culture today is one of the most alluring and promising fields to be working in. However, it comes with its fair share of toxic habits and the industry is well known for causing burn outs and issues with mental health amongst workers in all strata, I.e., right from junior employees to well-experienced heads and executives are prone to experiencing some form of mental health disorder. The pandemic has only worsened problems such as isolation and a lack of work-life balance.

Our goal is to see if we can identify potential predictive factors to a toxic culture within the technology industry to expose potential areas in need of improvement or support. We aim to analyze how factors like gender, remote work, and benefit access at the workplace affect and influence mental health within technology roles.

Using various supervised learning models like Decision Tree, Random Forest and SVM, we review the results to analyze what features give the best way to gauge various fields of our interests, some major questions being:

1. Can we predict if someone saw treatment for a mental health condition?
2. Can we predict if someone anticipates discussing a mental health issue with an employer to have negative consequences?
3. Do qualities relating to mental health indicate a likelihood of being in the tech industry?

From the target factors we identified to evaluate, we could not train a model to confidently predict if someone was working in the technology industry from mental-health and care related features. Nor could we confidently predict negative consequences from discussing mental health with an employer. Even with feature improvements and class imbalance corrections, there does not seem to be a conclusive answer of whether we can predict these two fields. Confidence was like flipping a coin. However, a model was trained to some degree of accuracy for predicting if someone saw treatment for a mental health condition. It would still be improper to generate a conclusion as there remains many questions on data sampling and biases.

We hope there will be a culture of support and understanding within the technology industry for both the employers and employees. The awareness the survey brings can help industry and community leaders to push for an open working environment that can be a safe place to discuss and understand mental health. The inability for our models to discern some of the questions posed highlights the complexity of these topics and warrants further research and investigation. Continued data collection for research such as this is important in striving to build a safe work environment for everyone and setting precedents that measures will be taken to improve the workplace experience and accessibility.

Introduction

In our project, we will be aiming to identify factors that contribute to the mental health of those in the tech field. This survey measured attitudes towards mental health and the prevalence of mental health disorders in the industry. We want to provide some transparency to gauge whether the tech industry can provide better accommodations, support to its employees, and reduce the stigma regarding mental health in the workplace. We have identified a few major questions to answer and explore regarding how likely it is for an employee to seek treatment for a mental health condition and to see whether discussing mental health issues with an employer would have negative consequences. From the results that we analyze, we can see if there are improvements or awareness needed in the tech industry culture.

Poor mental health can contribute negatively to the employee in terms of job performance and productivity, work engagement, functioning daily and having the mental or physical capability, and interactions with coworkers. This could lead to negative impacts for the business with employees' mental health interfering with job responsibilities. The tech industry culture has been talked about a lot especially with the ease of access to social media platforms such as YouTube where those in the industry are boosting up working in tech and promoting to friends and family the benefits of being in this industry. It is extremely critical to discuss and be more open to communication on how mental health is treated in the tech field and whether the workplace has created a positive culture on mental health in the sense where it is safe to speak up about it, seek out resources, and have social networks and programs available to employees.

The mental health background is well-known to many people. Mental health refers to how people think, feel, and behave. This can affect everyday life, relationships, and physical health. Mental health is "more than just the absence of mental disorders or disabilities" according to WHO. Your emotional, social, and psychological well-being is impacted. LinkedIn has written on mental health in the workplace, and they state that it is given the least priority in the workplace and in society (Mentortribes). Tech professionals are more prone to working extra hours which can negatively impact the time used to rest or socialize, which in turn can contribute to mental health disorders (Mentortribes). According to the article, LinkedIn states that 51% of tech professionals have been diagnosed with a mental health condition where 71% said that their productivity takes a hit by a mental health issue with 57% said they have experienced burnout (Mentortribes). The pandemic has only exacerbated issues such as social isolation and lack of work-life balance. Many big tech companies attract some of the brightest talents, however, with incentives such as free food and on-site gyms as well as promises of flexible work hours and unlimited PTO for some cases, the pressure and nature of a competitive industry as the tech field does not allow many employees to take advantage of these rewards and instead takes a toll on employees, as a result, their mental health (Pilsl, Alex). According to BairesDev, "Mental health has been one of the least discussed subjects in the corporate world" and needs to be discussed more about the complexity the tech industry adds (Pilsl, Alex).

Mental health arises in many people in many ways and should not be ignored. Tackling mental health is a continuous process to achieve a healthy work environment that is conducive to the mental wellbeing of employees. As we are involved in tech as well as the tech industry, we are interested in using various supervised learning models to determine the best features that have a profound impact on the tech industry and to understand the factors in the workplace.

Methodology

The data used for this project comes from an open-source database on Kaggle which is in turn sourced from a 2014 survey done by OSMI Mental Health, a non-profit organization "dedicated to raising awareness, educating, and providing resources to support mental wellness in the tech and open-source communities" (*About Osmi)*.

We begin our study by first preprocessing the data. Preprocessing included encoding categorical values and eliminating entries with missing data. We analyzed the summary statistics of the data to identify and remove erroneous datapoints. For example, participants who chose not to disclose their age had a default age of "99999999999" which was removed. Participants also were asked to disclose their gender which was organized into three categories of "male", "female" and "non-binary" based on the responses they wrote (*National Center for Transgender Equality).*

The fields with Nan or null values were replaced with "undisclosed" or 0 based on the feature relevance. Next, the distribution of categories within features was investigated to mitigate under or overrepresented categories. A boxplot of all features showed the features having the most outliers (*Fig 1*). We transformed our dataset to numeric values. Univariate analysis was used to determine feature importance and get the top 12 features from the given 23. Finally, an investigation into feature correlation was made to ensure the independently and identically distributed principle of the predictive models we choose to model was upheld.

Following the statistical investigation and preprocessing of the data, 3 predictive models were trained on the same 80-20 train test split. These models include a Decision Tree, a Random Forest and an SVM. Each model was cross-referenced by analyzing the accuracy, precision, recall and confusion matrix of the model's performance.

Each model underwent parameter experimentation. For example, the SVM experimented with linear, polynomial, and radial basis function kernels. The Decision Tree model experimented with different maximum depths and the Random Forest model was experimented with considering top 12 scored features.

Code

Our codebase was constructed in a google lab file, a sharable jupyter notebook. The notebook requires you to have downloaded the database from kaggle (1) in your working directory. The notebook follows the methodology outlined above. The code is organized according to the following sequence and structure.

Section 1: Load and store the data.

Here we load our dataset from local collab files. It was

originally downloaded from kaggle (1).

Section 2:

2.1: Preprocess Data:

This has various steps, we first perform null checks on the entire dataset to determine bogus fields and came across 3 features that had such values, namely: State, Self-employed, and work interference. These fields were then replaced with undisclosed, assuming certain sections of this survey were uncomfortable revealing this information. Next, we uniformly replaced the gender field with 3 categoric values of male, female and non-binary to condense all values entered in the dataset field. The age field was another such instance that had bad values like negatives and bogus values. These were removed from the dataset. Finally, the refined dataset was replaced with numeric values for all fields.

2.2: Investigate Data:

In this part, we used a heat map for an estimated high-level overview of the feature correlations with each other. We then perform univariate analysis on our data using the Chi-square test to determine the independence of each feature we target i.e.: treatment, mental health consequences, and being in the tech industry against the other lot. This was an excellent test to use since we have a good frequency of each value in our data cells, preventing errors in conclusions when using this specific test. This test helped us to eliminate any features most likely to be independent of class and therefore irrelevant for classification. This gave a promising idea to reveal that all these targets were heavily dependent on the "country" feature. Among the others, we saw that factors like physical health consequences, mental health checks, supportive management, and ability to manage mental and physical health played a key role in determining if someone anticipated discussing a mental health issue with an employer to have negative consequences. We saw a high relationship between the wellness program of a workplace, flexibility of the work environment, mental health of a person, and its routine check to be key factors when it came to determining if a person was related to working in the tech industry. Finally, a person's family history, interference with work on mental health, care options provided by the insurance/ company, and if the person had heard of or observed negative consequences for coworkers with mental health conditions in your workplace were key factors in identifying if they were likely to seek out help for mental health.

```
print(featureScores_mental_health_consequence.nlargest(5,'Score'))
```

	Specs	Score
17	phys_health_consequence	372.830428
20	mental_health_interview	202.389851
2	Country	141.526955
19	supervisor	141.041641
22	mental_vs_physical	131.593515

```
[210] print(featureScores_tech_industry.nlargest(5,'Score'))
```

	Specs	Score
2	Country	76.778030
12	wellness_program	20.487229
9	remote_work	15.496826
16	mental_health_consequence	11.898478
20	mental_health_interview	11.788721

```
print(featureScores_treatment.nlargest(5,'Score'))
```

	Specs	Score
2	Country	265.503527
5	family_history	109.203431
6	work_interfere	93.759412
11	care_options	39.580222
23	obs_consequence	25.602375

To enhance this finding, we then checked for class imbalance potential. This was investigated with histograms. The most significant imbalance we found was related to geographic location. However, the overwhelming input, the United States at 60%, did display geographic diversity then within the "State" category. It is likely the dependency of the country attribute discovered earlier is related to geopolitical issues (*Fig 2*). Significant outliers in Age were identified and removed (*Fig 1*). The Self-Employment category was dropped due to an overwhelming lack of response. As will be discussed in the results section, models were trained with and without class balance correction for tech industry participants.

Section 3: Shared Model Evaluation Code and Metrics

Each Model was evaluated by their accuracy, precision, recall, confusion matrix and ROC Curve.

Section 4: Predictive Models

The models were all trained, evaluated, and predictions were made.

4.1: SVM

Support Vector Machines were evaluated with linear, polynomial and radial basis function kernels. Polynomial kernels were trained with default degree of 3. Gamma values for all kernels were set to the sklearn auto setting which is 1 / n_features.

4.2: Decision Trees

Decision Tree Classifier models were evaluated with varying maximum depths of default, 2, and 5.

4.3: Random Forest

Random Forest models were evaluated with the number of estimators trained to 100. As well as using the features shown important by univariate analysis.

Section 5: Comparative Analysis

In addition to validating models using shared metrics such as accuracy, precision, recall, ROC and confusion matrices, graphics were also created to evaluate accuracies across models directly.

Results

Can we predict if someone saw treatment for a mental health condition?

We can predict if someone saw treatment for a mental health condition. There was not a significant response class imbalance in this case as the target column has pretty much balanced values, thus indicating that we do not need to perform any undersampling or oversampling. The prediction accuracies were at 66-83 % for all the models with all the features analyzed to see if we could predict if someone sought treatment (*Fig 3*). With the top features, we can see a slight overall improvement to all the models with at least 70-80% prediction accuracies (*Fig 3*). For each of the models, the prediction, accuracy, F1-score, and recall values are all relatively high. The Confusion Matrices also depicts high true positives and true negatives identified in the model predictions for all of the models which can also be seen in the ROC curves for all of the models.

Looking at the models individually, the Decision Tree model had the most accurate prediction with all features at 84%; with top features, it tied close with another model, Random Forest, at ~81% whereas SVM results were around the low 70%

range (Fig 3). The Decision Tree Classifier with no maximum depth selected had results in the mid 70% range, where one with a maximum depth of 2 yielded lowest results out of the three decision tree models with 66% with all features and 72% with top features. Overall, the models predicted well in terms of seeing if the person sought treatment for a mental health condition from all the survey participants.

Can we predict if someone anticipates discussing a mental health issue with an employer to have negative consequences?

We cannot confidently predict that there would be negative consequences if someone were to discuss a mental health issue with an employer. The number of survey participants who answered "no" and "maybe" were similar at around 40% whereas 23% of participants said "yes". The models were trained based on the results from the dataset without any correction for response class imbalance even though the classes are not equally balanced, it didn't significantly skew towards a class. Reflecting on the results from the models we have chosen, most of the models were in the 50-60% range with RBF SVM and Poly SVM models towards the 80% mark with all the features included (*Fig 4*). Even with improving the model training with just the top 12 features, the models did not yield any improvement. In fact, the two models from SVM that were originally towards 80% fell to 50-60% (*Fig 4*). This prediction seems to be a toss-up and cannot be used to come to a conclusive result. The F1-score, accuracy, precision, recall, confusion matrix, and ROC curve for each model seems to indicate that the model predictions for this question are not confident enough to say we can predict negative consequences for discussing mental health with an employer (*Fig 4*). The values are close to 50/50, thus, inconclusive.

Do qualities relating to mental health indicate a likelihood of being in the tech industry?

In short, no. Models were trained with and without a correction for response class imbalance for participants who came from the technology industry. When inspecting the confusion matrix for models that did not have corrected class balance (*Fig 5.a-.h2*) with features predicted to be of relevance, the accuracy was around 80% for all models. However, when inspected the confusion matrix, you can observe that the models simply learned to assume that the answer was yes, as 82% of participants belonged to this class (*Fig 2*). When correcting for class imbalance, prediction accuracies across all models were no better than a coin flip at approximately 50-60% (*Fig 6.a-g.1*).

Discussion

When we investigate the results of the questions we want to predict from the above results section, we can try to understand the reasoning behind each of the targeted features that were classified and trained from the data. Having corrected some response class imbalance as well as evaluating the models with top features, we were able to see an improvement in the predictions for the questions.

The first question delves into whether someone saw treatment for a mental health condition. We can predict 70-80% based on the Decision Tree and Random Forest models' statistical results. When considering that the SVM with different kernels, and therefore different decision boundaries, performed similarly, we

postulated that the features being introduced to the model might simply not be descriptive enough. In order to have a more accurate prediction model across all approaches, it is likely that more features would be needed as mental health is a complex and highly personal topic. Features we hypothesize may be helpful in determining this, while remaining focused on our workplace concerns, include time with a company, sense of belonging at the workplace, and identifiable mentorship figures. It is also worth pursuing if being in the technology field meant that employees were more able to exercise their benefits, particularly those offered in an online format.

The second question dives into whether negative consequences would be anticipated from discussing a mental health issue with an employer. All trained models were inconclusive. This may be even if the work environment amongst coworkers is friendly and open to discussing mental health, the employees may not feel comfortable speaking about their mental health to their managers or supervisors. Most respondents did not feel like there would be negative consequences for speaking about their mental health to their employers. It does seem like the tech industry is on the right track where many employees do not feel like they will suffer from potential drawbacks if they do. Although the number of respondents who answered that they anticipate negative consequences seems low, there is more information needed to draw a conclusion such as having a baseline or looking more into the impact of benefits and resources available to employees. Overall, there does not seem to be a firm conclusion as to whether mental health is recognized by the employers to not bear negative results.

The third question investigates into whether qualities related to mental health would indicate a likelihood of being in the tech industry. Based on the results, it does not seem like we can predict this feature. This likely means that either the technology industry experiences mental health similarly to other industries, or that not enough other industries were sampled. When reflecting on the most significant features of the data set, and how location was significantly important, perhaps investigating geographic distributions of mental health and geopolitical issues which factor into mental health access, will shine greater light on how mental health plays into the careers of technology professionals.

Future Work

This study is not conclusive as there are some limitations to the data collected from the survey. We saw that some features like country and state had many distinct outliers, preprocessing them with nan or null or discarding those entire rows would mean loss of a sizable chunk of data which is why we worked without editing them. It is a strong speculation that these outliers must have been a key player in some of the reduced metrics when training our models. Some future work could include diving into geographic mental health research. As this survey stands, many of the participants of the survey are from the United States. There could be an aim to include an equal number of participants represented from many countries and regions across the world for better data sampling. Based on the feature importance, there could be a better way to sample data in the future by eliminating some of the features that were asked as some bear little contribution to the model predictions. The study could dive deeper

into some of the top features such as family history to understand and predict the impacts it could have on the individual where questions stemming from the family background could be relevant and another approach to understanding mental health and its impact in the workforce. Another feature could be work interference to understand how to improve mental health from interfering with work. More questions posed to see where individuals find lacking or support from the workplace could also be beneficial in helping to understand the complexities of mental health and the best ways to approach and give support to the employees. This survey is great to bring mental health awareness in the tech space, however, questions can be formatted and aligned more closely to a specific end goal to help bring the study to more conclusive results. Future work can also entail examining other predictors of mental health conditions and understanding the attitudes regarding them.

Conclusion

Through the data processing and model predictions, we aimed to understand the importance of the features and the relationships between the predictors and target classifiers on treatment, negative consequences, and the tech industry. We did see high relationships between many features with our target classifiers in seeking whether we can predict the responses. We observed a decent predictor of treatment based on the features collected from the survey. There seems to be less confident predictions in terms of anticipating negative consequences for discussions with an employer as well as if the qualities of mental health would indicate someone to seek out the tech industry. There definitely seems to be awareness of mental health in the tech industry. However, there are still stigmas surrounding mental health regarding discussing mental health and the shame some might feel. Over time, we hope there will be more mutual communication amongst the employers and employees to generate a more open and accessible environment conducive to mental health. Companies can establish a mental health system and by bringing awareness to employees the benefits and resources that are available to them. There is a social responsibility for organizations to make a commitment to providing high quality care for their employees, setting the bar where it needs to be to ensure success.

Figures

Fig 1. Feature Outliers

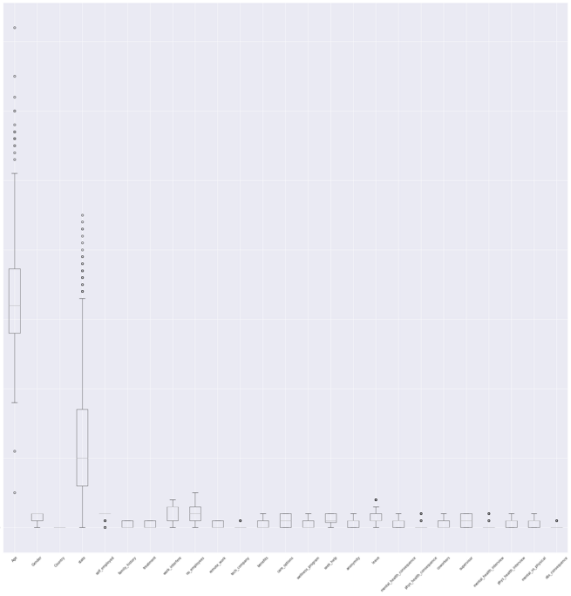


Fig 2. Feature Class Distributions

Fig 3. Results: Can we predict if someone saw treatment for a mental health condition?

Fig 3.a.1 Results Summary, Models with All Features

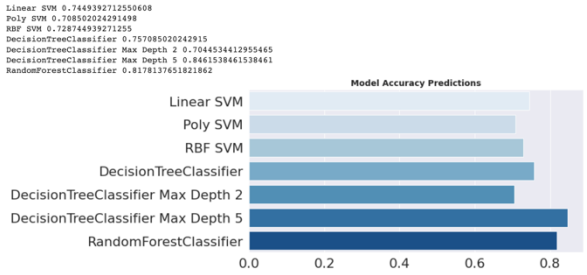


Fig 3.a.2 Results Summary, Models with Top Features

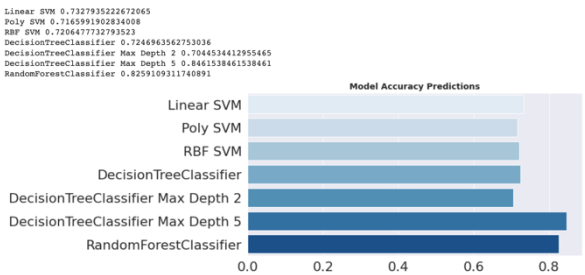


Fig 3.b.1 Linear SVM Results, All Features for Q1

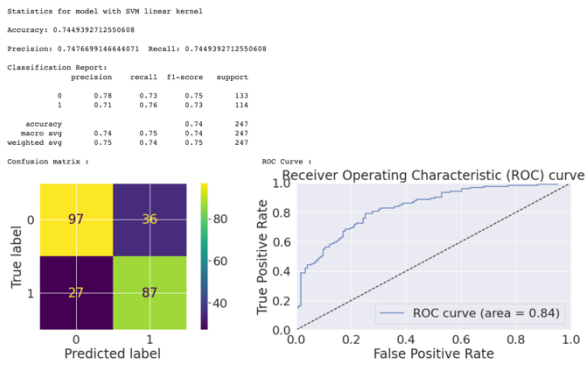


Fig 3.b.2 Linear SVM Results, Top Features for Q1

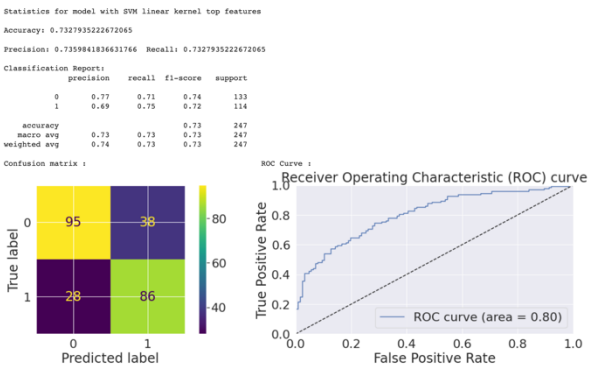


Fig 3.c.1 Polynomial Linear SVM Results, All Features for Q1

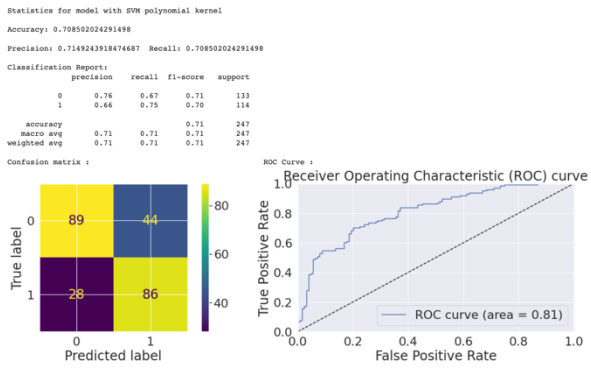


Fig 3.c.2 Polynomial Linear SVM Results, Top Features for Q1

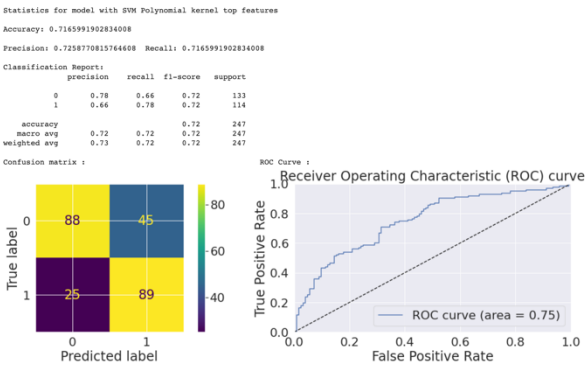


Fig 3.d.1 Radial Basis Function SVM Results, All Features for Q1

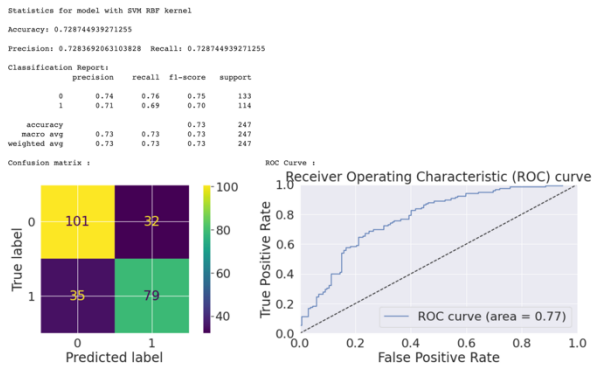


Fig 3.d.2 Radial Basis Function SVM Results, Top Features for Q1

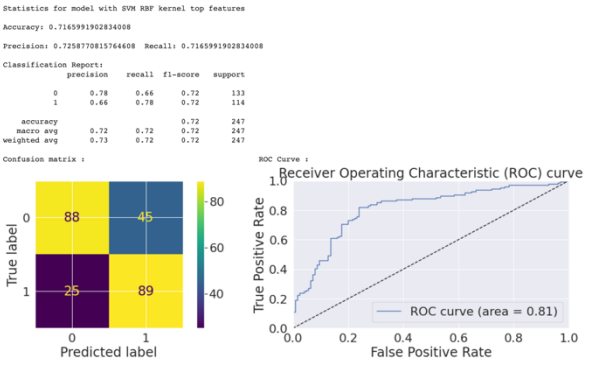


Fig 3.e.1 Decision Tree, Depth of 5 Results, All Features for Q1

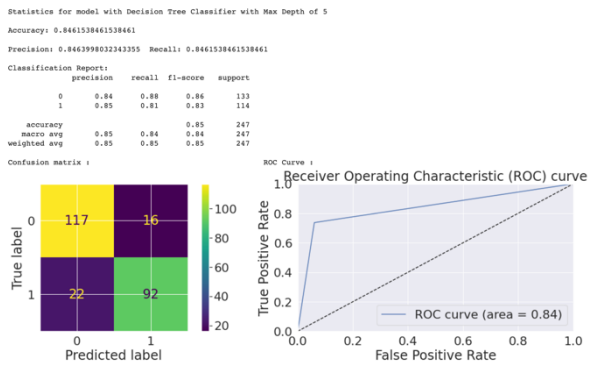


Fig 3.e.2 Decision Tree, Depth of 5 Results, Top Features for Q1

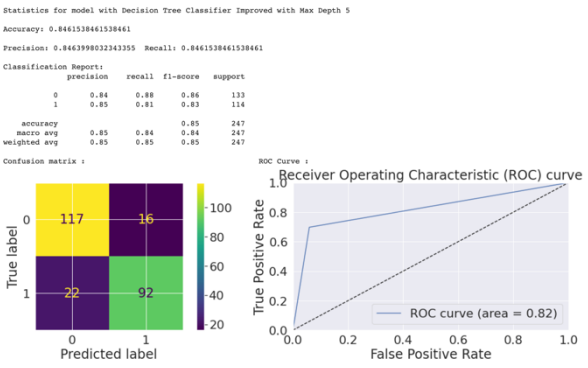


Fig 3.f.1 Decision Tree, Depth of 2 Results, All Features for Q1

Fig 3.f.2 Decision Tree, Depth of 2 Results, Top Features for Q1

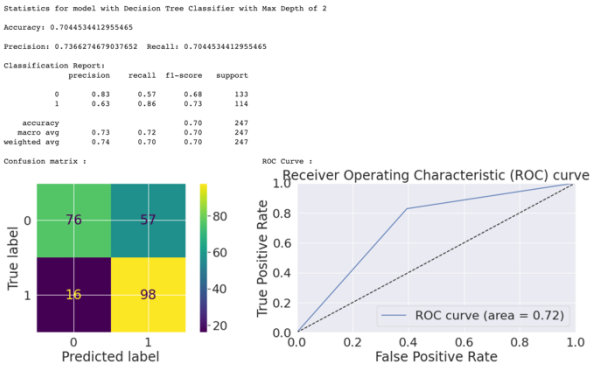


Fig 3.g.1 Decision Tree Results, Top Features for Q1

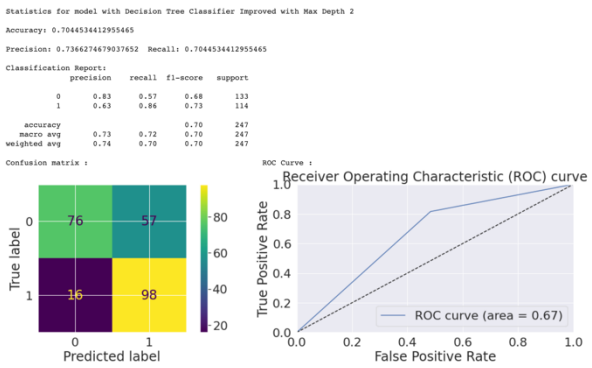


Fig 3.g.2 Decision Tree Results, Top Features for Q1

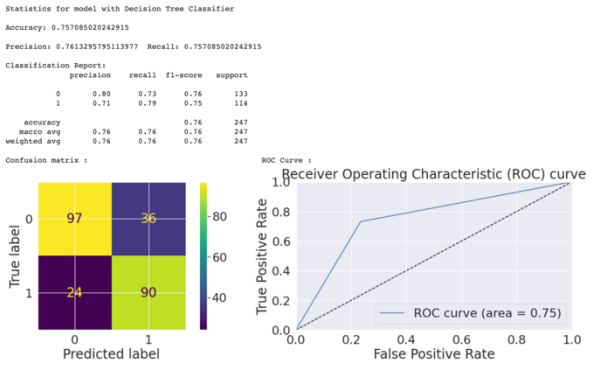


Fig 3.h.1 Random Forest, All Features for Q1

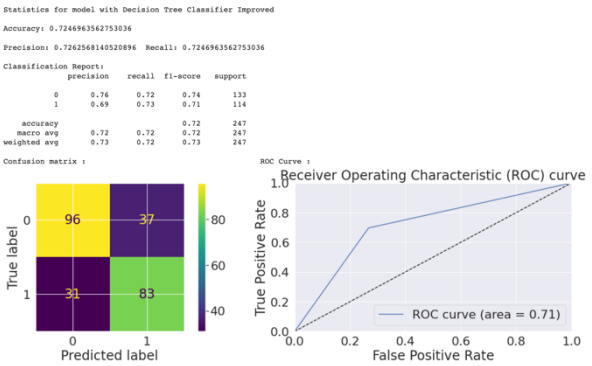
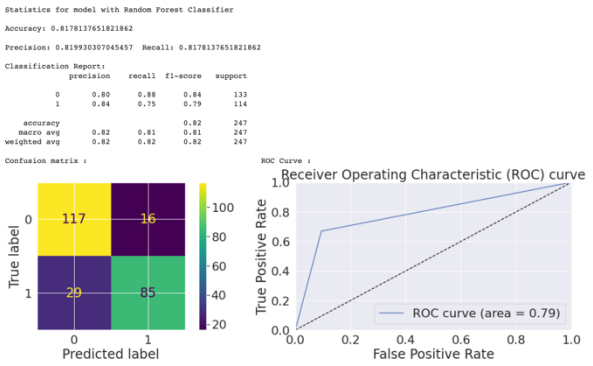


Fig 3.h.2 Random Forest, Top Features for Q1



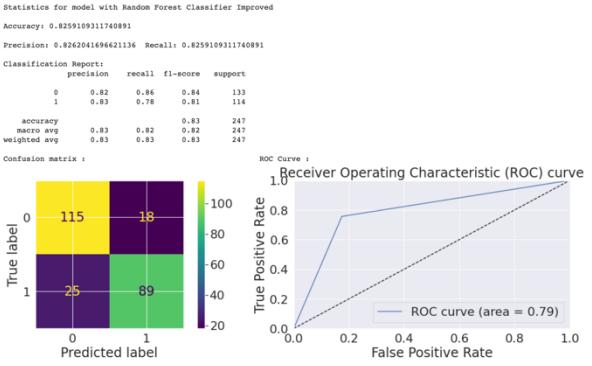


Fig 4. Results: Can we predict if someone anticipates discussing a mental health issue with an employer to have negative consequences?

Fig 4.a.1 Results Summary, Models with All Features

Fig 4.a.2 Results Summary, Models with Top Features

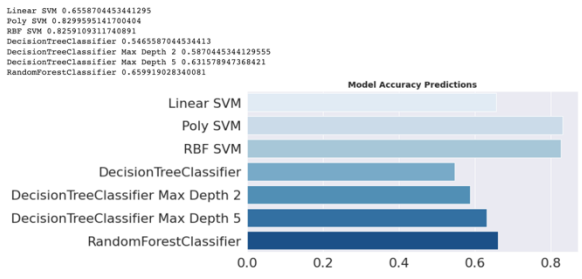


Fig 4.b.1 Linear SVM Results, All Features for Q2

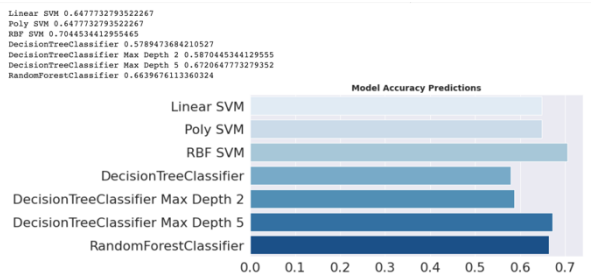


Fig 4.b.2 Linear SVM Results, Top Features for Q2

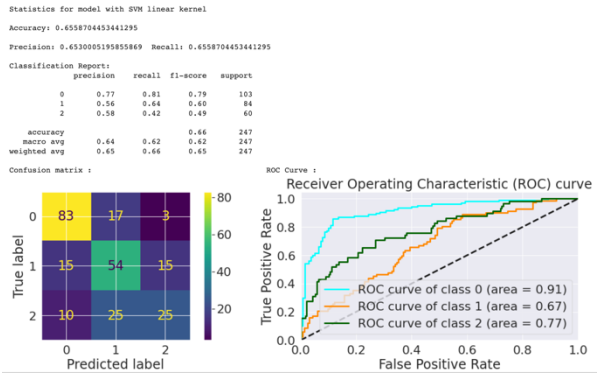


Fig 4.c.1 Polynomial Linear SVM Results, All Features for Q2

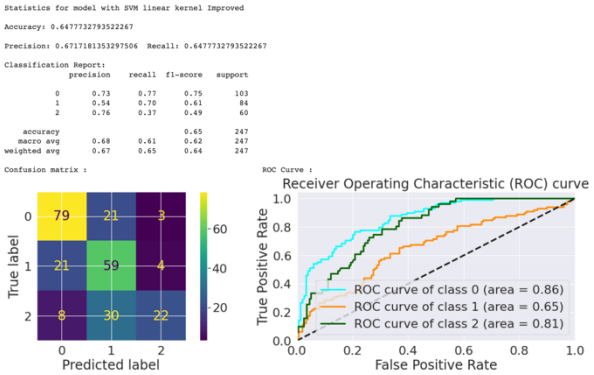


Fig 4.c.2 Polynomial Linear SVM Results, Top Features for Q2

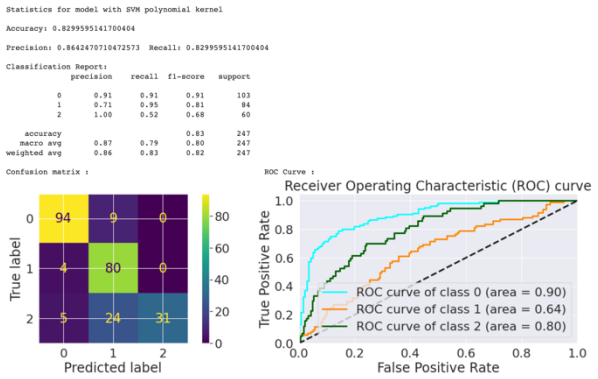


Fig 4.e.1 Decision Tree, Depth of 5 Results, All Features for Q2

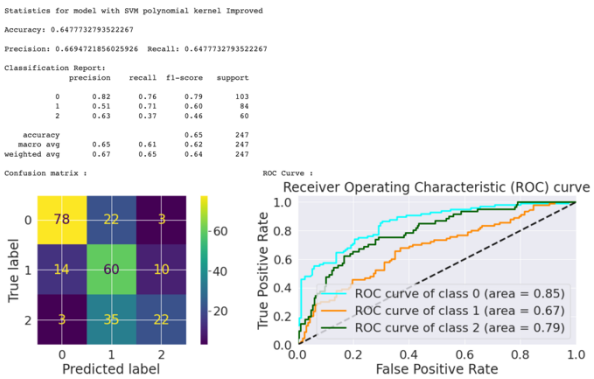


Fig 4.e.2 Decision Tree, Depth of 5 Results, Top Features for Q2

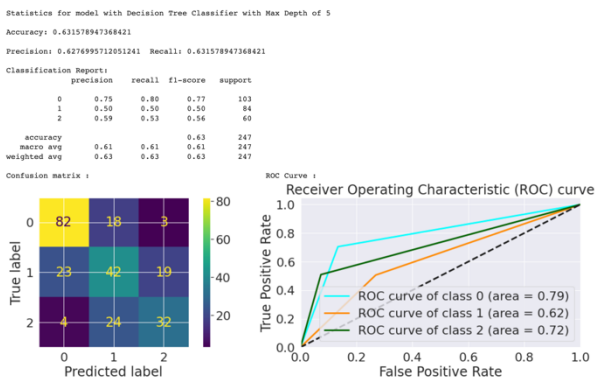


Fig 4.f.1 Decision Tree, Depth of 2 Results, All Features for Q2

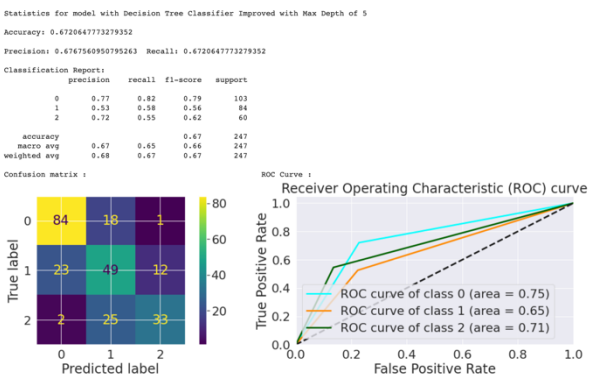


Fig 4.f.2 Decision Tree, Depth of 2 Results, Top Features for Q2

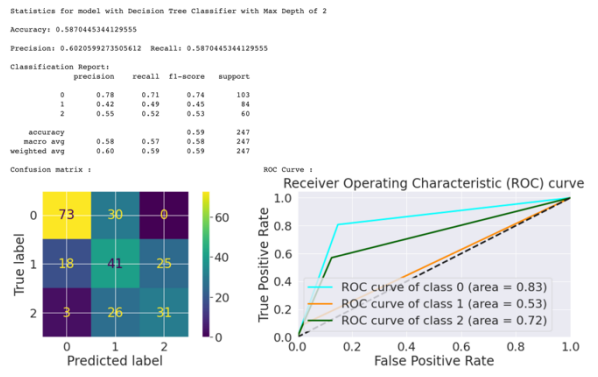


Fig 4.g.1 Decision Tree Results, All Features for Q2

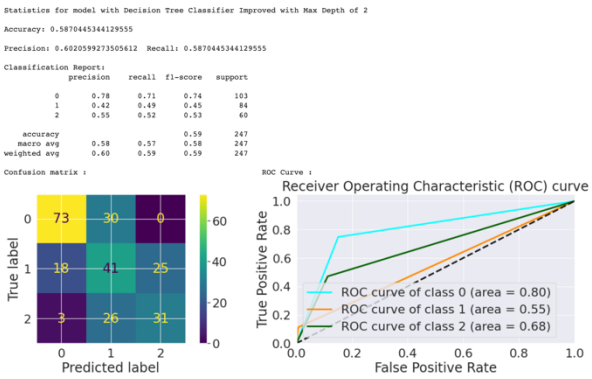


Fig 4.g.2 Decision Tree Results, Top Features for Q2

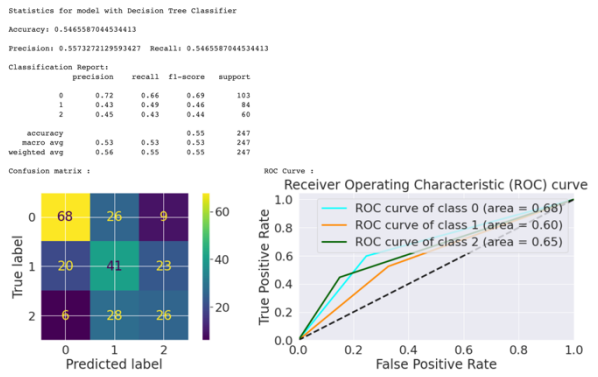


Fig 4.h.1 Random Forest, All Features for Q2

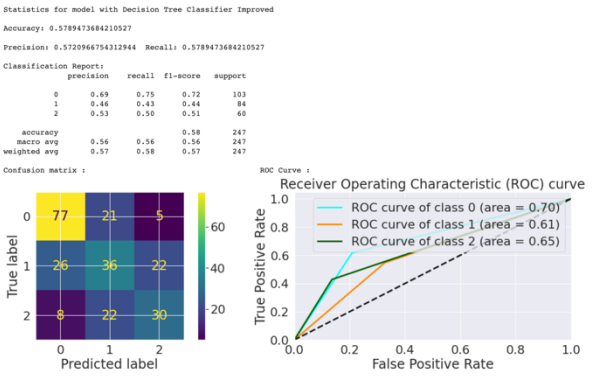


Fig 4.h.2 Random Forest, Top Features for Q2

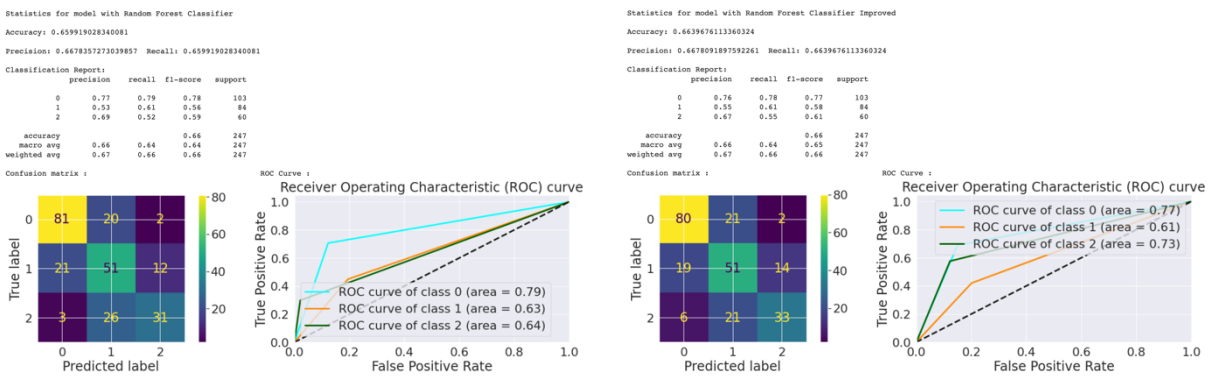


Fig 5. Results: Do qualities relating to mental health indicate a likelihood of being in the tech industry? - NOT CONTROLLING FOR CLASS DISTRIBUTION

Fig 5.a.1 Results Summary, Models with All Features

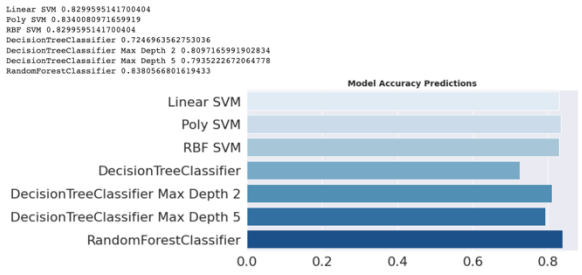


Fig 5.a.2 Results Summary, Models with Top Features

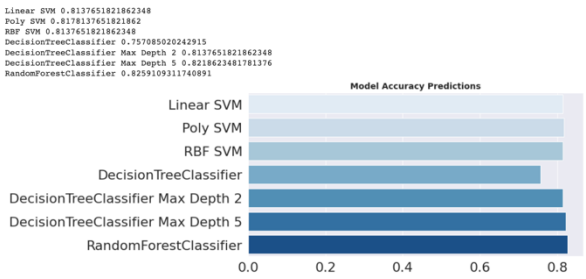


Fig 5.b.1 Linear SVM Results, All Features for Q3

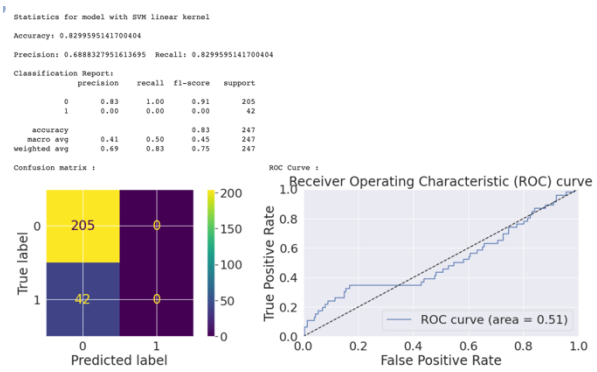


Fig 5.b.2 Linear SVM Results, Top Features for Q3

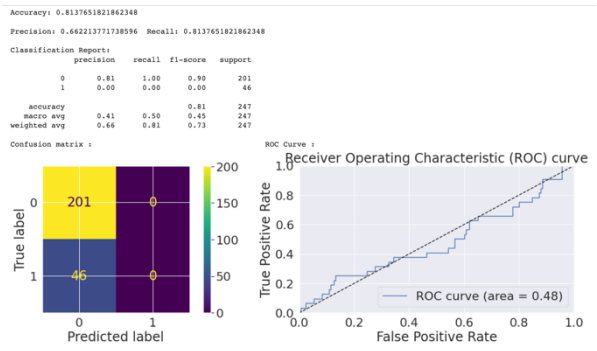


Fig 5.c.1 Polynomial Linear SVM Results, All Features for Q3

Fig 5.c.2 Polynomial Linear SVM Results, Top Features for Q3

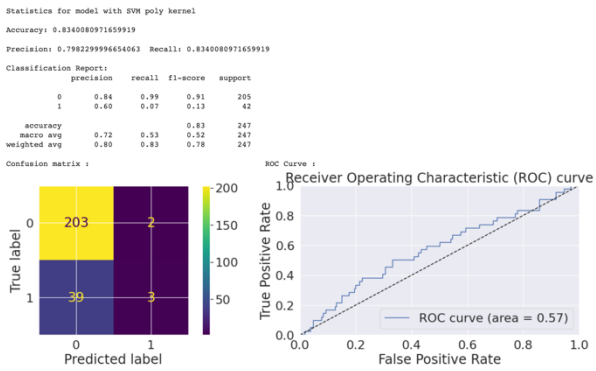


Fig 5.d.1 Radial Basis Function SVM Results, All Features for Q3

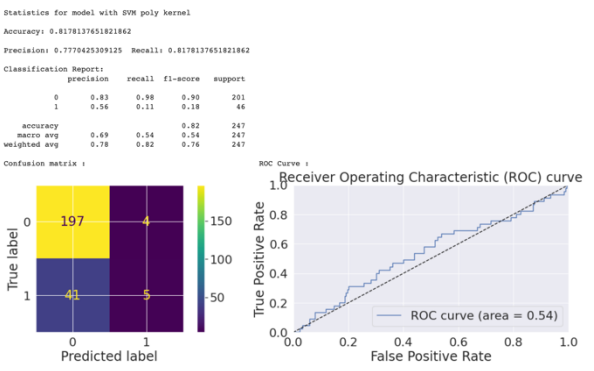


Fig 5.d.2 Radial Basis Function SVM Results, Top Features for Q3

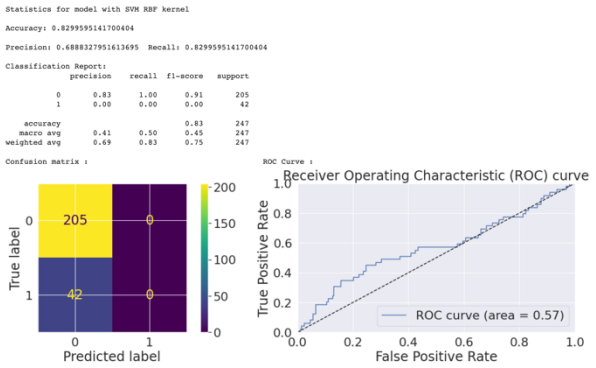


Fig 5.e.1 Decision Tree, Depth of 5 Results, All Features for Q3

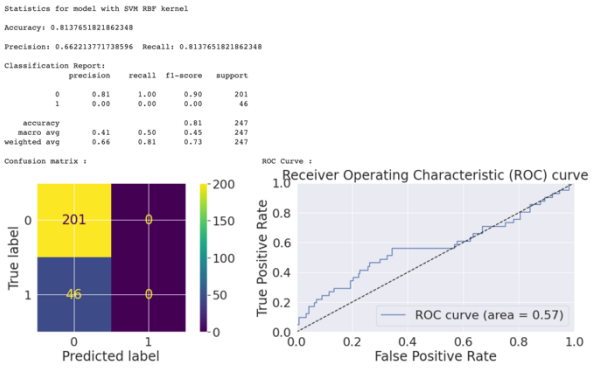


Fig 5.e.2 Decision Tree, Depth of 5 Results, Top Features for Q3

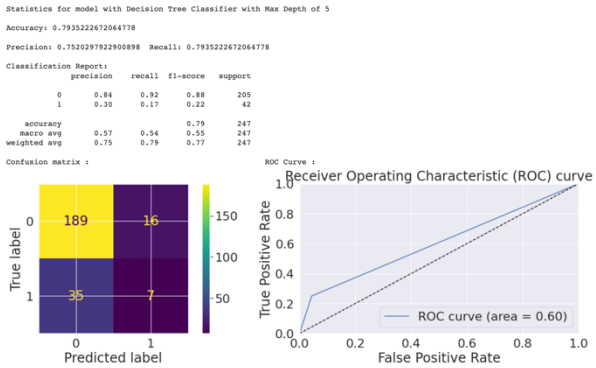


Fig 5.f.1 Decision Tree, Depth of 2 Results, All Features for Q3

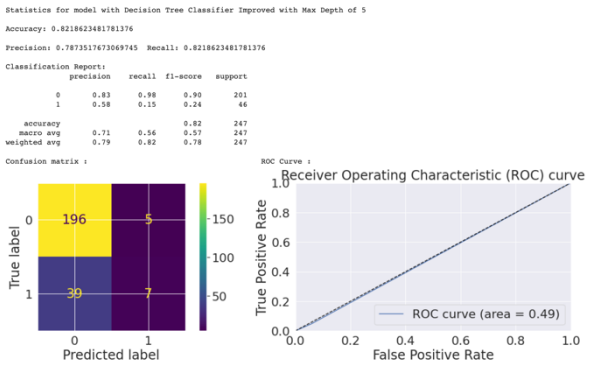


Fig 5.f.2 Decision Tree, Depth of 2 Results, Top Features for Q3

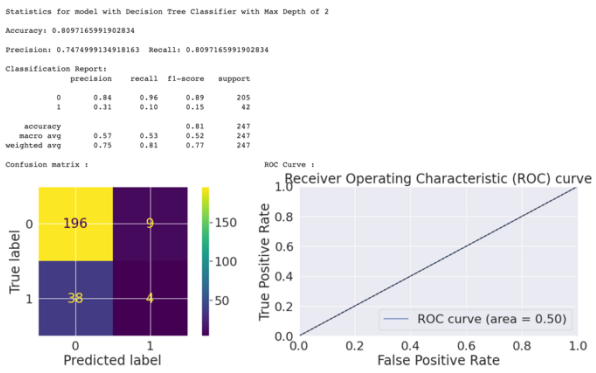


Fig 5.g.1 Decision Tree, All Features for Q3

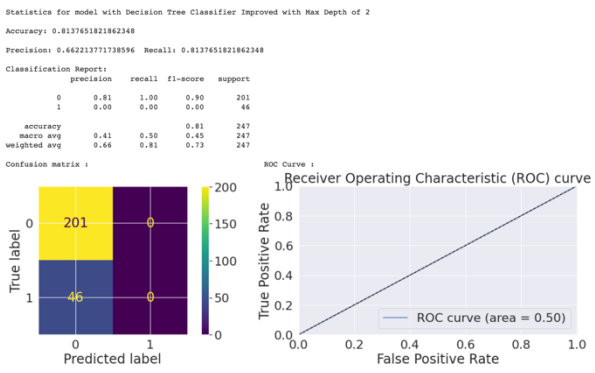


Fig 4.g.2 Decision Tree, Top Features for Q3

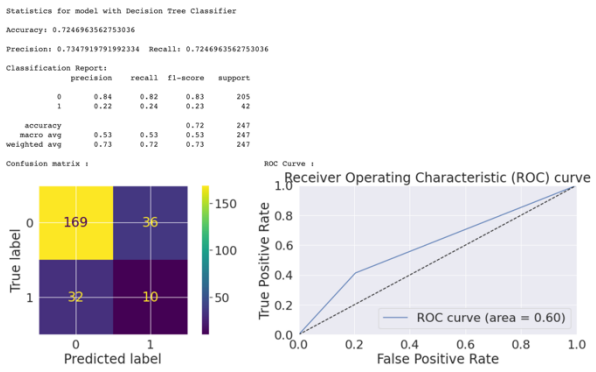


Fig 5.h.1 Random Forest, All Features for Q3

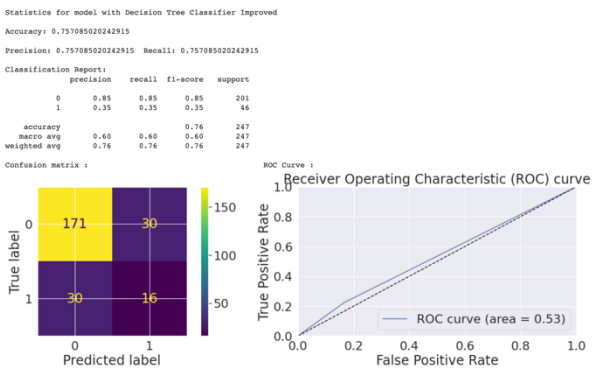
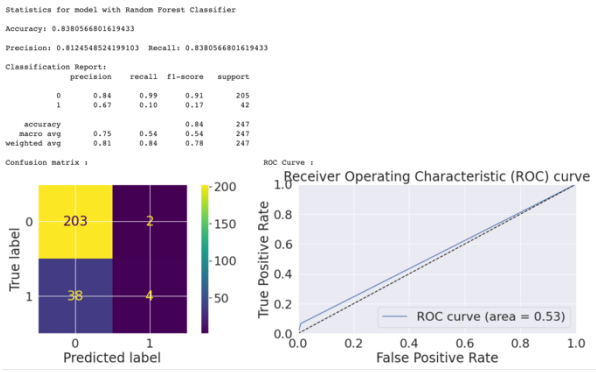


Fig 5.h.2 Random Forest, Top Features for Q3



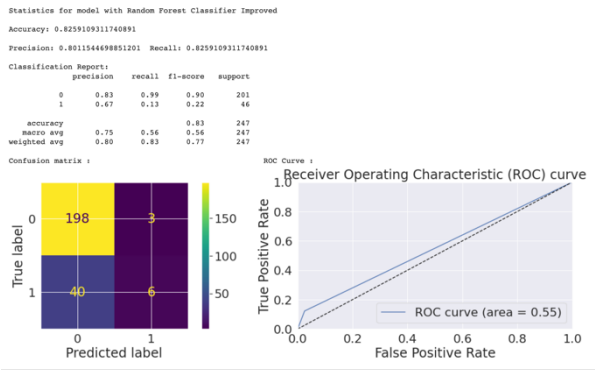


Fig 6. Results: Do qualities relating to mental health indicate a likelihood of being in the tech industry? - CONTROLLING FOR CLASS DISTRIBUTION

Fig 6.a.1 Results Summary, Models with Top Features

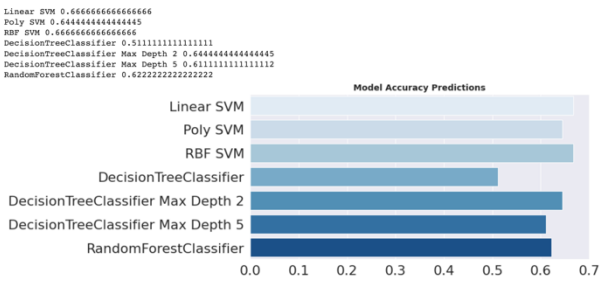


Fig 6.b.1 Linear SVM Results, Top Features for Q3

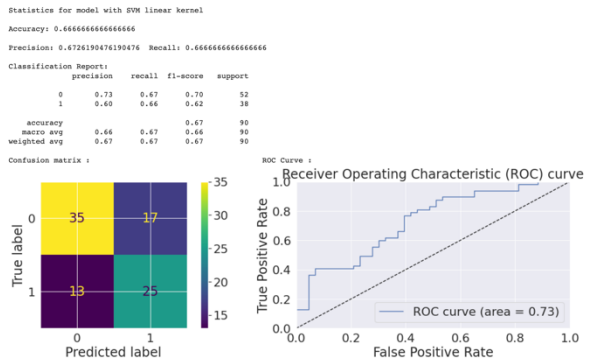


Fig 6.c.1 Polynomial Linear SVM Results, Top Features for Q3

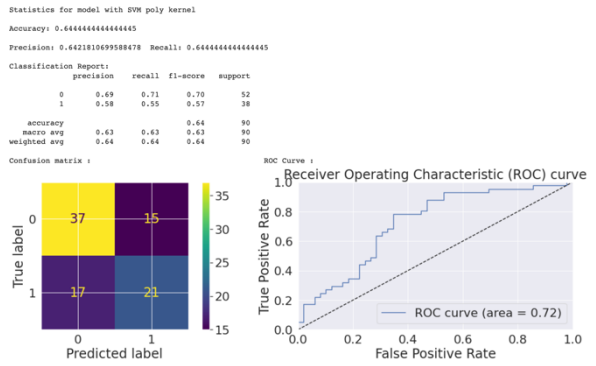


Fig 6.d.1 Radial Basis Function SVM Results, Top Features for Q3

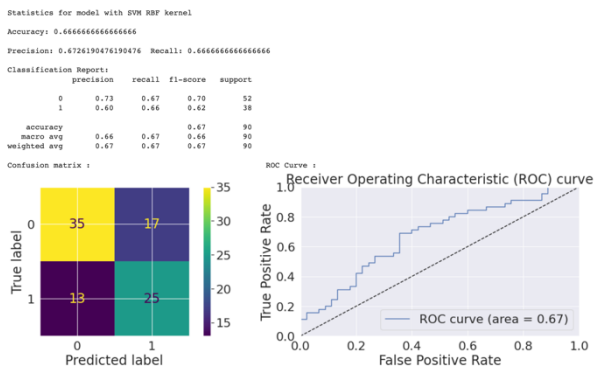


Fig 6.e.1 Decision Tree, Depth of 5 Results, Top Features for Q3

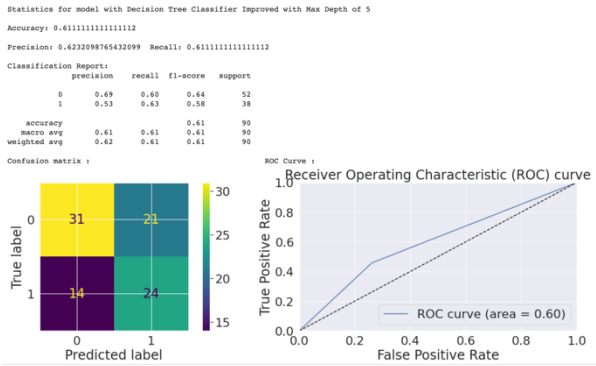


Fig 6.f.1 Decision Tree, Depth of 2 Results, Top Features for Q3

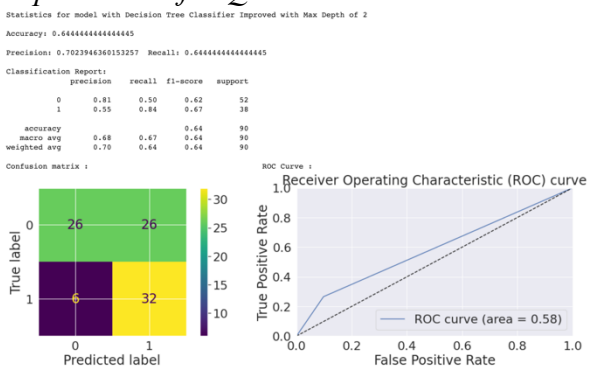


Fig 6.g.1 Decision Tree, Top Features for Q3

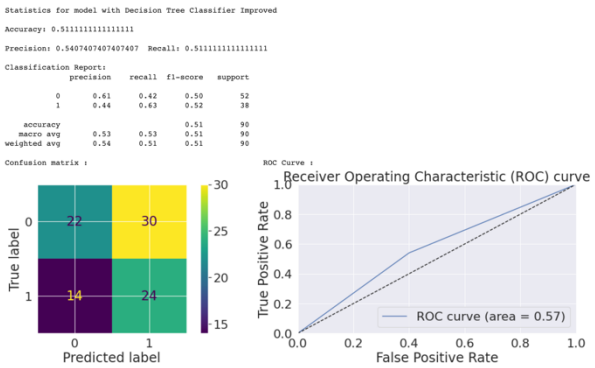
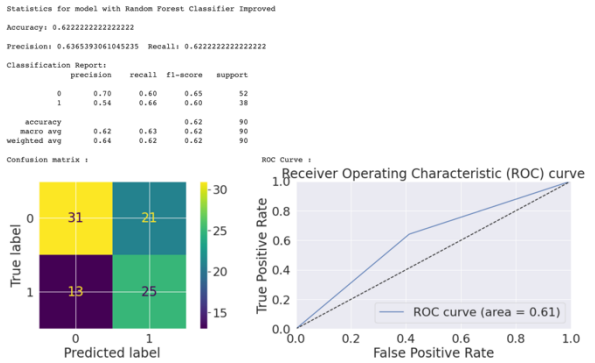


Fig 6.h.1 Random Forest, Top Features for Q3



References

About Osmi : Open sourcing mental health - changing how we talk about mental health in the Tech Community - Stronger Than Fear. OSMI Home. (n.d.). Retrieved December 3, 2022, from https://osmihelp.org/about/about-osmi

Mentortribes. “Mental Health - A Major Concern in the Tech Industry.” LinkedIn, 7 July 2022, https://www.linkedin.com/pulse/mental-health-major-concern-tech-industry-mentortribes/?trk=public_post-content_share-article.

Pilsl, Alex. “Supporting Mental Health in Tech Culture Corp..” *BairesDev*, 11 Nov. 2022, https://www.bairesdev.com/blog/supporting-mental-health-in-tech-culture/#:~:text=More%20than%2090%25%20of%20workers,by%20a%20mental%20health%20issue.

Understanding non-binary people: How to be respectful and supportive. National Center for Transgender Equality. (2018, October 5). Retrieved December 3, 2022, from https://transequality.org/issues/resources/understanding-non-binary-people-how-to-be-respectful-and-supportive

Code Reference

(1) https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey?resource=download

Annex-A

Jacqueline Girouard:

I contributed to sections of the data processing and to the conversion of categorical data into numeric features. I also contributed to various data investigation plots including, pairwise correlation plots, and histogram class distribution plots. I contributed summary statistic metrics such as the confusion matrix and basic model performance statistics. I also trained and evaluated the SVM model for linear, polynomial, and radial basis function kernels.

Aisha Kothare:

I contributed to the sections of data preprocessing which included checking nan values in features and replacing them with accurate values. I also worked on the data investigation section, particularly in determining the feature importance using univariate analysis and boxplot to highlight outliers in features as well as plotting a heatmap to see overall feature correlation. I was responsible for training the random forest model and getting the accurate data splits responsible for using the top 12 key features gathered from my univariate feature analysis.

Joanna Ng:

I briefly touched upon the data preprocessing on the gender feature. I contributed to the ROC curve as part of the model statistics. I also worked on the Decision Tree model training and evaluated different target values based on the maximum depth. Lastly, I have also added the model performance accuracies which summarizes all the models we have used in a bar chart for comparison of the models.