MENTAL HEALTH ISSUES IN THE TECH INDUSTRY

Dharmik Shah

Computer Science
College of Natural Sciences
The University of Texas at Austin

ABSTRACT

Mental health issues in the tech industry are an ongoing concern, fueled by causes like high stress, tight deadlines, and overall challenging work. Studies frequently label tech as one of the worst sectors for developing mental health issues. In 2014, Open Sourcing Mental Health (OSMH) surveyed tech workers, asking them whether they had previously sought treatment for mental health issues while working in tech. The survey asked about personal and workplace-related factors that could affect one's mental health. These include individual factors like age, gender, and family history and company-related factors like whether the company offers benefits, whether it is remote-friendly, etc. This survey aimed to determine whether an individual is likely to have mental health problems due to various personal and work-related causes while working in tech. This paper examines the classification task of determining whether a tech worker will likely experience mental health issues due to a combination of personal and workplace-related factors using machine learning models, specifically regression, ensemble, and neural networks. We scale our features using PCA analysis and comment on important research questions posed in the introduction.

1. INTRODUCTION

Mental health is an umbrella term under public health that encompasses emotional, psychological, and social well-being. The state of one's mental health can influence how well one thinks, behaves, handles stress, and lives [1]. Maintaining a positive level of mental health is essential, but unfortunately, there are many professional industries where that is easier said than done. According to estimates, the healthcare, manufacturing, and tech sectors are the worst hit in mental health cases. The tech industry specifically is notorious for having a high number of mental health cases among its workers. Studies estimate that at least 50% of tech workers will experience some mental health issues in their working lives [2]. In 2014, a survey was issued to tech workers asking them to comment on whether they had sought treatment for what they perceived was a mental health issue due to working in tech. It also asked the workers to indicate their personal and company-related information.

Through this survey, we can try to understand what human and company factors can likely lead to someone developing mental health issues in tech. In this paper, we will perform machine learning analysis on this survey and attempt to answer the following research questions:

1) Can mental health issues arise at any age?
2) Which gender is more likely to experience mental health issues?
3) What are the most likely features that serve as a predictor for someone who has a mental health issue?
4) Evaluate dataset on a variety of machine learning models to see how well they can classify a tech worker with having or not having a mental illness

2. BACKGROUND

Open Sourcing Mental Health (OSMH) is a non-profit corporation dedicated to raising awareness and educating people about mental wellness in tech and open-source communities. In 2014, they released a survey to actively working tech workers, asking them to comment on whether they sought treatment for what they perceived was a mental health illness due to working in tech [3]. Studies show that the tech industry is one of the worst in terms of the number of mental health cases it holds among its workers. Constant work, tight deadlines, and always keeping up with the latest tech can often lead to burnout and unstable mental health issues, such as depression, anxiety, and more. These external factors, however, are common to the entire tech industry and thus were not the target of this survey. Instead, OSMH decided to look at mental health from an internal perspective. They wanted to determine what factors related to the individual themselves and the company they work for can result in someone developing mental health issues and seeking treatment. The individual factors included age, gender, family history of mental illness, etc. The company factors included whether the workplace offered any mental health assistance, whether the candidate believes their condition (if any) impacts their work, whether their work is remote-friendly, etc. Through these responses, OSHM could understand mental health in tech from an internal perspective. Moreover, as company factors were also listed, this survey allowed OSHM to determine what companies could do to mitigate the rate of mental health issues developing among its workers. The survey gathered more than 1200 responses, which at that time, could have been considered the largest survey done in tech-related to mental health. It was then published on Kaggle for public analysis [4].

3. DATA & VARIABLES

3.1 FEATURES

The survey recorded the responses of 1259 individuals, amongst which 50% reported having a mental illness and 50% reported that they did not. The data was effectively presented as a classification task. There were 27 columns, with 26 being the features (split between personal and work-related) and the remaining column being the output class. Individual features included age, gender, country, etc. Company features included benefits (does the company provide mental health benefits), remote work (does the company allow remote work), work interfere (if you have a mental health illness, do you feel it interferes with your work), etc. The output class was called treatment and indicated whether the individual sought treatment for their perceived medical condition. Note that it is impossible for someone to self-diagnose whether they have a mental health issue or not. The survey considers that someone has a mental health illness if they have visited a doctor to discuss their problem. This survey does not list whether or not it is truly a mental health illness such as depression.

	Timestamp	Age	Gender	Country	state	self_employed	family_history	treatment
0	2014-08-27 11:29:31	37	Female	United States	IL	NaN	No	Yes
1	2014-08-27 11:29:37	44	M	United States	IN	NaN	No	No
2	2014-08-27 11:29:44	32	Male	Canada	NaN	NaN	No	No
3	2014-08-27 11:29:46	31	Male	United Kingdom	NaN	NaN	Yes	Yes
4	2014-08-27 11:30:22	31	Male	United States	TX	NaN	No	No

5 rows × 27 columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1259 entries, 0 to 1258
Data columns (total 27 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Timestamp                  1259 non-null   object
 1   Age                        1259 non-null   int64
 2   Gender                     1259 non-null   object
 3   Country                    1259 non-null   object
 4   state                      744 non-null    object
 5   self_employed              1241 non-null   object
 6   family_history             1259 non-null   object
 7   treatment                  1259 non-null   object
 8   work_interfere             995 non-null    object
 9   no_employees               1259 non-null   object
 10  remote_work                1259 non-null   object
 11  tech_company               1259 non-null   object
 12  benefits                   1259 non-null   object
 13  care_options               1259 non-null   object
 14  wellness_program           1259 non-null   object
 15  seek_help                  1259 non-null   object
 16  anonymity                  1259 non-null   object
 17  leave                      1259 non-null   object
 18  mental_health_consequence  1259 non-null   object
 19  phys_health_consequence    1259 non-null   object
 20  coworkers                  1259 non-null   object
 21  supervisor                 1259 non-null   object
 22  mental_health_interview    1259 non-null   object
 23  phys_health_interview      1259 non-null   object
 24  mental_vs_physical         1259 non-null   object
 25  obs_consequence            1259 non-null   object
 26  comments                   164 non-null    object
```

Figure 1: High level overview of features and their values

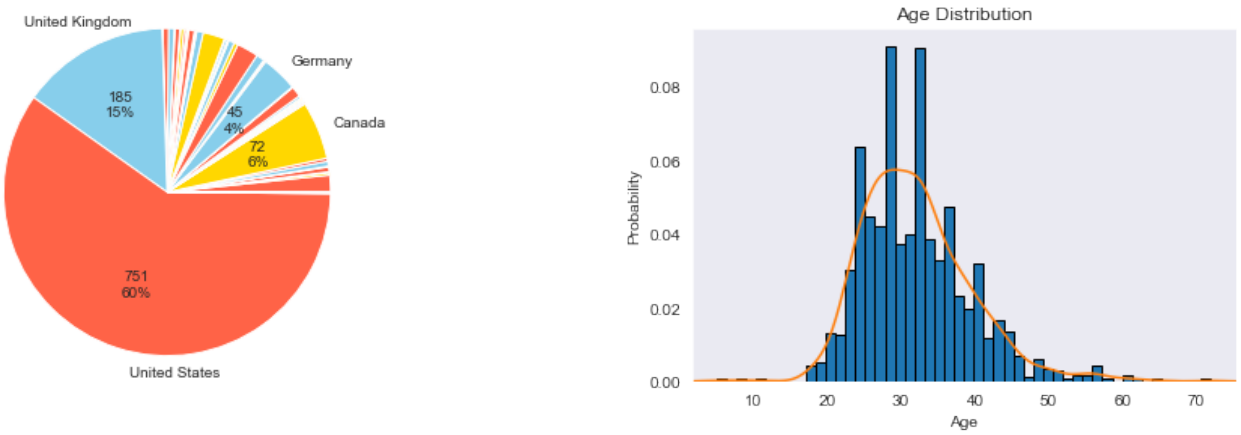


Figure 2: Age and country distributions (notice invalid values of age and majority of countries)

3.2 PRE-PROCESSING

After understanding the features, the next step in our analysis was preprocessing the data. For example, we found that the country column had a significant bias towards the United States (60%) and the United Kingdom (15%), with the remaining countries appearing far less often than the majority. Thus, we decided to drop this feature since most participants were from 1st world countries. Using a similar approach, we found that we could drop additional columns like timestamp, state, and comments. We also found that participants in this survey weren't entirely truthful, as there were a few cases where the age column was negative or had a very high value. We accounted for this by removing these outliers. For the gender column, we noticed a high variance in the number of answers given. Although most responses fell into the Male or Female genders, the values themselves had to be filtered independently, as candidates used "male", "M", "man", "Cis man" and more all to refer to the same gender. We repeated this process for participants that were females. Although we saw other genders reported, it was a tiny minority compared to Male and Female, so we tried our best to merge them into the main ones. After these significant changes, the remaining task involved converting categorical values into numeric ones so that our models could understand them.

After preprocessing our data, we went from 26 feature columns to 22. We felt that this was still too many and would be difficult for the models to train against, so we opted to use PCA. PCA, short for principal component analysis, is a technique used to reduce a dataset's dimensionality while still keeping most of the variance amongst the features [5]. There is further analysis of PCA and how we determined the number of features to keep in a later part of this paper. For now, we conclude that we were able to transform 22 feature columns into 8 while still maintaining more than 95% of the variance in the data.

3.3 VARIABLES

The features used are listed below, with the additional treatment column serving as our predicting class. We can see our breakdown of individual and company-related factors. For instance, age, gender, and family history are all factors about the individual. The number of employees, benefits and supervisors are factors about the company. We converted the features to a categorical representation and scaled them. We showcase detailed statistics below.

	Age	Gender	family_history	treatment	work_interfere	no_employees	benefits	supervisor	phys_health_interview	mental_vs_physical
count	1235.000000	1235.000000	1235.000000	1235.000000	1235.000000	1235.000000	1235.000000	1235.000000	1235.000000	1235.000000
mean	32.085020	0.800000	0.388664	0.500405	2.275304	2.785425	1.050202	1.100405	0.715789	0.808907
std	7.341854	0.400162	0.487644	0.500202	1.597302	1.737688	0.837187	0.845212	0.723109	0.835255
min	5.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	27.000000	1.000000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
50%	31.000000	1.000000	0.000000	1.000000	3.000000	3.000000	1.000000	1.000000	1.000000	1.000000
75%	36.000000	1.000000	1.000000	1.000000	4.000000	4.000000	2.000000	2.000000	1.000000	2.000000
max	72.000000	1.000000	1.000000	1.000000	4.000000	5.000000	2.000000	2.000000	2.000000	2.000000

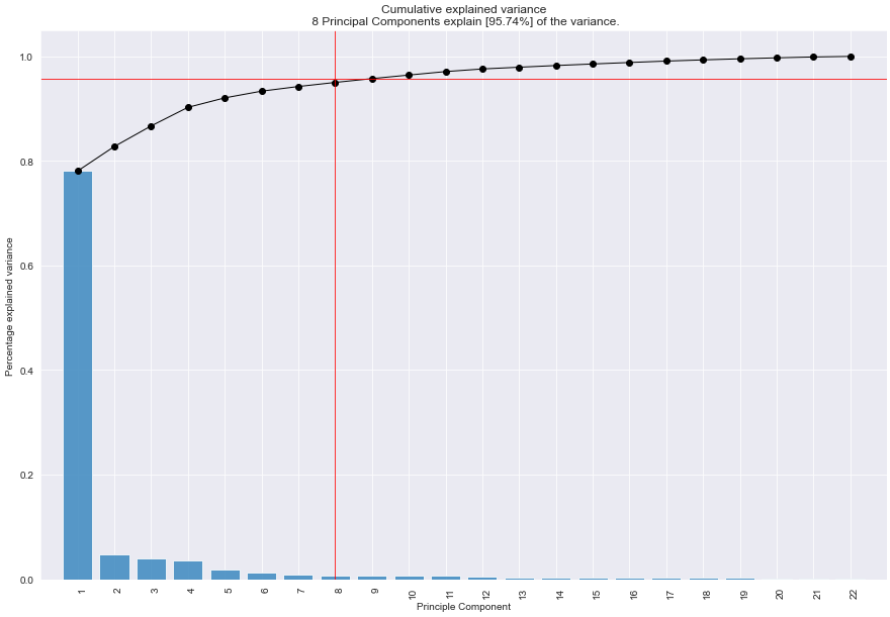
Figure 3: Detailed statistics of 8 columns with treatment label after being scaled

4. METHODS

We used a total of 8 features determined from PCA to perform a classification task. The task was to determine whether the individual likely had a mental illness based on the input features. In the dataset, this is analogous to predicting whether the candidate sought out treatment, where 0 is the negative case, and 1 is the positive case.

We trained our data using regression, ensemble, and deep learning models. The simplest models are regression models, which estimate the relationship between a dependent variable against multiple independent variables. We went with Logistic Regression for our regression model, as it is excellent with binary predictions, which is our task exactly. Based on the ending value between 0 and 1, we mapped it to which class is most likely [6]. We expected it to perform well. For our next model, we used a decision tree, which is useful for classification and regression. A decision tree works by creating rules through leaves and internal nodes that allow it to represent a function of discrete attributes [7]. For something like detecting mental illness, we felt that the rules could be pretty complicated, and although it may be a good model, we did not expect it to be the top one in terms of accuracy. We now discuss ensemble models, that function by running multiple models simultaneously and combining their results [8]. We used the Random Forest as our ensemble model. A random forest consists of numerous decision trees, and we chose 300 as our hyperparameter [9]. We expected it to perform much better than the decision tree as it is a stronger version of that algorithm. To expand on ensemble methods, we also included XGBoost, which does gradient boosting. Boosting is a type of machine learning algorithm focused on reducing bias and variance, thus converting weak models into stronger ones [10]. Our final model was a deep neural network, which we believe should be a good estimator due to its ability to estimate complex functions. We decided to include a simple DNN with two hidden layers, optimized using Adam and its loss calculated through binary cross-entropy. As only eight features exist, we did not require a neural network with convolutions or pooling [11]. We also trained it over 100 epochs. In summary, we used Logistic Regression, a Decision Tree, Random Forest, XGBoost, and a DNN to perform the classification task of determining whether an individual has a mental illness.

As we had over 1200 samples, we decided to split the training and testing splits by 70% training and 30% testing.



	0	1
0	PC0	Age
1	PC1	no_employees
2	PC2	Gender
3	PC3	work_interfere
4	PC4	benefits
5	PC5	supervisor
6	PC6	mental_vs_physical
7	PC7	phys_health_interview
8	PC8	family_history

Figure 4: PCA components and the variance chart

5. RESULTS

5.1 PCA ANALYSIS

Before we dive into how each model performed, let us first discuss which features PCA chose. The graph above illustrates the principal component count against the amount of variance it explains. We can see that through 8 principal components (from a starting of 22), we could account for 95% of the variance amongst the input features. On the other image, we can see which input feature that individual components correspond to, starting from the strongest and working our way to the weakest. It is a healthy mix of personal features such as age, gender, family history, and company features like the number of employees, benefits and supervisors. We will examine these components and try to understand the reasoning behind their selection.

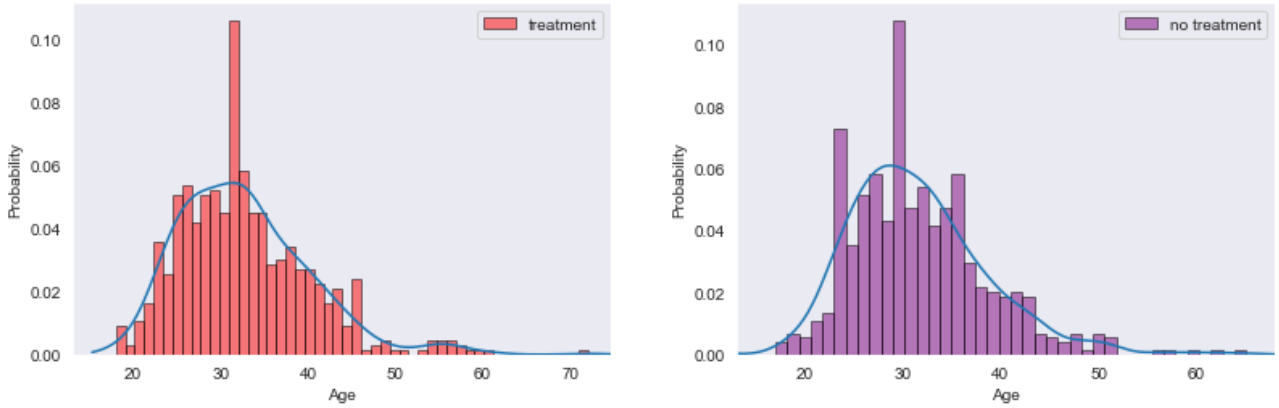


Figure 5: Age and its impact on whether someone gets treatment

Selecting age as the top principal component suggests that age may impact whether someone can develop a mental health condition or not. We now answer our first research question, where we want to determine if mental health issues arise at any age. Realistically, although mental health issues can come about at any age, the fact that it is the top component suggests that a select range of ages are more likely or less likely to experience these issues. Above, we present two histograms, mapping ages against whether the individual has a mental health illness. Viewing the data through a histogram, it is clear that the probability of having a mental health illness is low when you are new to the industry and low when you are on your way out. The ages of 25 – 45 seem to be the most impactful, and it is in this age group you may or may not experience a mental health issue. We can see this through the histogram as there are spikes in both of them during this age range. This means that many individuals will experience mental health issues during this time, but many will not as well. To answer our research question, although mental health issues can appear at any age, you are less likely to have them when you are younger and much older but may experience them during your major working years

Skipping over the workplace-related components temporarily, we want to understand why gender and family history were considered by PCA. We want to attempt to answer our second research question of which gender is more likely to experience mental health issues. Unfortunately, 80% of the survey participants were males, and thus the data is skewed in that direction. However, when analyzing the dataset, we come across a surprising finding.

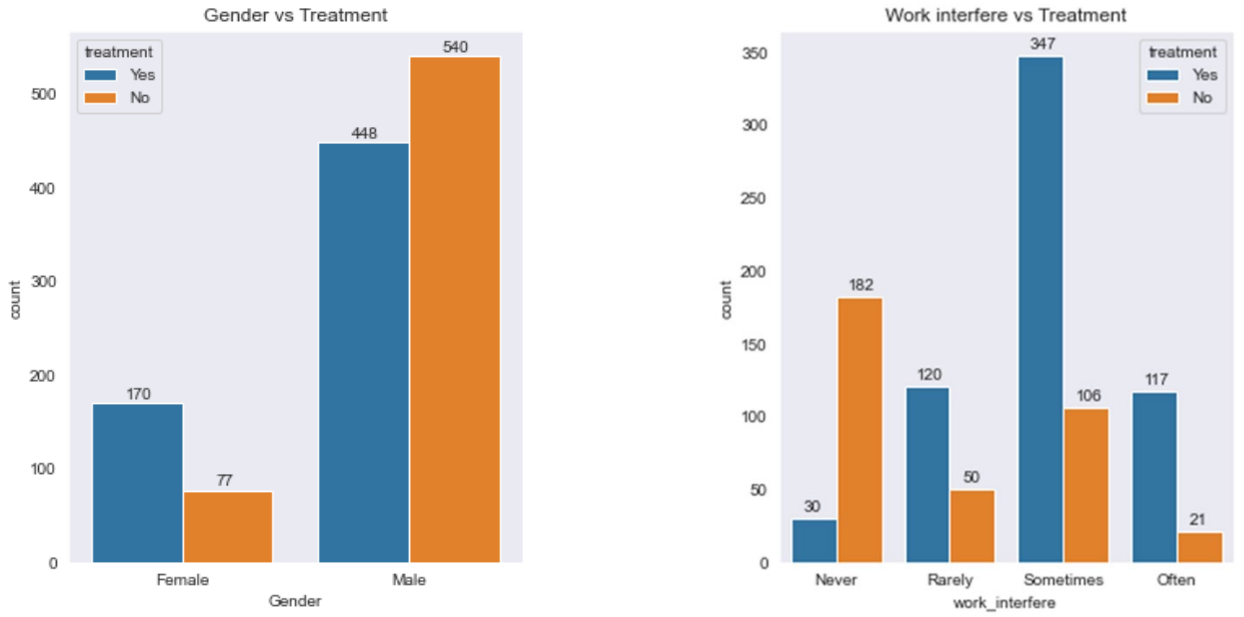


Figure 6: Gender and Work interference impact on seeking treatment

Approximately 69% of females mentioned that they sought treatment, compared to 45% of males who said the same. This fact is surprising because there are four times as many males as females, yet the treatment percentage in males is still lower. This fact suggests that gender plays a role in whether you choose to receive mental health treatment; more specifically, females are more likely to experience this. We are not entirely confident in our finding since there are fewer female responses in the dataset; however, based on the data given, we can conclude that females likely experience more mental health issues and seek treatment than males.

We will now discuss the company-related components, specifically work interference and benefits. Work interference records how much an individual perceives their work to be impacted if they believe they have a mental health issue. The data goes from not affecting (the leftmost bar) to having a significant impact on the right. As the interference with work increases, so does the number of people seeking mental health treatment. This fact suggests that employees can develop mental health issues if they feel their abilities are diminishing at work. Something interesting here is that employees can develop mental health issues from the slightest impact on their work.

In many sectors, performing poorly at work may be unfavorable, but it does not always lead to mental health issues. It is clear from the data that in tech, the same rules do not apply and that performing poorly at work can often lead to higher levels of stress and, thus, mental health issues.

We now consider the benefits feature, which details whether the individual's company offers mental health benefits. We know that employees generally will use their medical benefits, and if companies add mental health benefits, employees are more likely to use them on even the slightest doubts. In this survey, 50% of candidates that received treatment reported that 48% of their respective companies offered some benefit. On the other hand, 50% of candidates that did not receive treatment mentioned that only 27% of their respective companies have mental health-related benefits. From this analysis, we claim that whether a company offers mental health benefits or not does not have a significant impact on whether an individual will develop these issues in the first place. We can see that even if companies do not provide a benefit, 50% of individuals will still develop some mental health issues, potentially from other factors. We can conclude that benefits cannot be an influencing factor like age or gender. We still believe it is in the company's best interest to provide mental health benefits, as treating them can be costly, especially in the United States. If companies do not offer this, many candidates with potential mental health issues may not seek treatment even though they need to. Other company factors may result in someone experiencing mental health issues, such as the number of employees (startups vs. big corps), supervisor (not getting along with manager), etc. These make sense intuitively, as startups work extremely fast, and having a bad manager is never a good sign. However, we will not dive deep into these and thus conclude our discussion on PCA.

5.2 IMPORTANT FEATURES

PCA has effectively given us our most important features that can determine whether a tech worker is likely to develop mental health issues and receive treatment or not. We now can answer our third research question, where we tried to determine the most impactful features. According to PCA, the most important individual features are age, gender, family history, and the most critical work-related features are the number of employees, work interference, and benefits. We are only listing the ones we have discussed, even though there are more components.

5.3 MODEL RESULTS

We trained our dataset using logistic regression, a decision tree, a random forest, XGboost, and a deep neural network. We split our data into train and test with a ratio of 70% train and 30% test.

Model	Accuracy (%)	Overall Rank
Logistic Regression	81.43	4
Decision Tree	73.47	5
Random Forest	83.55	3
XGBoost	84.35	2
DNN	84.98	1

Table 1: Model accuracy with overall rank

Overall, all our models had good evaluations on the test set. The decision tree had the worst accuracy by a large margin, just as we initially predicted. We know that there are many steps and decisions can impact whether someone will seek treatment for a mental health issue, and thus, a simple decision tree was not expected to perform well. Logistic regression performed much better than the decision tree and crossed the 80% accuracy mark. We knew that logistic regression is excellent for classification tasks, and we could use this model well by applying PCA. Since we predicted that the decision tree would perform poorly, we also opted to use a random forest, where we selected 300 as the number of trees in the forest. We found this number to be a good stopping point. Fewer trees resulted in lower accuracies, and more trees improved accuracy by a tiny amount. The concern is that we would be overfitting, so we stopped at 300 [12]. As expected, it performed much better than the decision tree as it was able to model more complex rules, and it was able to classify 10% of additional examples correctly. For our other ensemble model, we chose the gradient-boosting model XGBoost. Gradient boosting can be used when decision trees are not predicted to perform the best, and it generally outperforms random forests [13]. We can see that XGBoost exceeds random forest by a small percentage point. Our final model was a deep neural network with two hidden layers, all activated by reLU, and the final output determined by sigmoid. It is a simple network, but its accuracy puts it in the top position. It is only slightly ahead of XGBoost.

We can conclude that the neural network performed the best out of all the models. We now answered our fourth and final research question regarding the performance of each of the machine-learning models. Before we complete this section, we want to see the impact of PCA and whether it resulted in higher accuracies for our models. Without PCA, we would have 22 features instead of 8. We summarize the results below.

Model	Accuracy (%) [PCA]	Accuracy (%) [no PCA]
Logistic Regression	81.43	**81.94**
Decision Tree	73.47	**77.09**
Random Forest	**83.55**	83.32
XGBoost	**84.35**	83.28
DNN	**84.98**	80.32

Table 2: Model accuracy with and without PCA applied

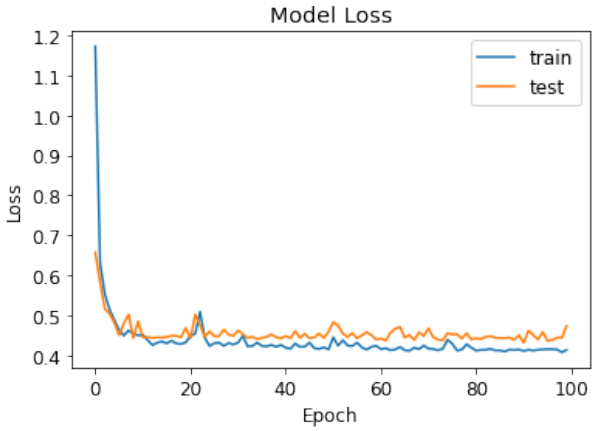


Figure 7: DNN training and test loss

We can see that applying PCA has its advantages and disadvantages. The disadvantage of PCA is that we cover 95% of the variance, but the remaining 5% is enough for logistic regression and a decision tree to perform better. The advantage is that our top 3 models (DNN, XGBoost, Random Forest) remain the best regardless of whether PCA is applied or not. The DNN is explicitly 4% more accurate through the utilization of PCA, and similarly, the ensemble methods of random forest and XGBoost are as well compared to their non-PCA counterpart. We can conclude that PCA is overall a success, as it improves the accuracy of our best models by a significant margin.

6. CONCLUSION

In this paper, we analyzed a survey dataset of around 1200 examples, indicating whether a person has sought mental health treatment while working in tech. The survey asked the individual to comment on their personal and workplace-related factors. By understanding the dataset, we performed PCA analysis to drop the number of features from 26 to 8, and trained across various machine learning models. We then evaluated our models with and without PCA, and concluded that PCA is an important step that allowed our top models to perform significantly better. In the order of lowest to highest accuracies, we have a decision tree, logistic regression, random forest, XGBoost, and finally, a deep neural network, with a peak accuracy of 84.98%. In this paper, we also determined the answers to our research questions. We specifically saw that the most likely features to predict someone with a mental health issue were age, gender, and family history. The most essential work-related are the number of employees, work interference, and benefits. Moreover, we determined that mental health could realistically arise at any age, but it was most prevalent during the 25 – 45 age range. We also concluded that women are more likely to develop mental health issues in tech than men. A shortcoming of this dataset was that it was predominantly male-skewed, and the number of participants was low. For future work, it is worth looking into a recent dataset with a good mix between the genders and surveys of many more people.

REFERENCES

[1] Prince, Martin, et al. "No health without mental health." The lancet 370.9590 (2007): 859-877.

[2] Barkved, Kirsten. “Let's Talk: It's Time to Get Serious about Mental Illness in Tech.” IQmetrix, https://www.iqmetrix.com/blog/lets-talk-its-time-to-get-serious-about-mental-illness-in-tech.

[3] “Open Sourcing Mental Health - Changing How We Talk about Mental Health in the Tech Community - Stronger Than Fear.” OSMH Home, https://osmhhelp.org/.

[4] Open Sourcing Mental Illness, LTD. “Mental Health in Tech Survey.” Kaggle, 3 Nov. 2016, https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey?datasetId=311.

[5] Abdi, Hervé, and Lynne J. Williams. "Principal component analysis." Wiley interdisciplinary reviews: computational statistics 2.4 (2010): 433-459.

[6] Lemon, Stephenie C., et al. "Classification and regression tree analysis in public health: methodological review and comparison with logistic regression." Annals of behavioral medicine 26.3 (2003): 172-181.

[7] Song, Yan-Yan, and L. U. Ying. "Decision tree methods: applications for classification and prediction." Shanghai archives of psychiatry 27.2 (2015): 130.

[8] Sagi, Omer, and Lior Rokach. "Ensemble learning: A survey." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8.4 (2018): e1249.

[9] Speiser, Jaime Lynn, et al. "A comparison of random forest variable selection methods for classification prediction modeling." Expert systems with applications 134 (2019): 93-101.

[10] Schapire, Robert E. "The boosting approach to machine learning: An overview." Nonlinear estimation and classification (2003): 149-171.

[11] Montavon, Grégoire, Wojciech Samek, and Klaus-Robert Müller. "Methods for interpreting and understanding deep neural networks." Digital signal processing 73 (2018): 1-15.

[12] Ying, Xue. "An overview of overfitting and its solutions." Journal of physics: Conference series. Vol. 1168. No. 2. IOP Publishing, 2019.

[13] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.