IS 734 FINAL PROJECT REPORT (SPRING 2024)

Predicting Workers Cyber-Resilience Through Mental Health Risk Modeling in Tech Industries

Group Members:

Pranahith Babu Yarra - IW23456
Sai Rajesh Rapelli - HD49179
Ameenur Rahman Khan - DJ46492
Sai Vikas Amaraneni - FS93533
Maram Venkat Kowshik - MO76733

Submitting to: Dr. Faisal Quader

M.S. Information Systems

Department of Information Systems

University of Maryland, Baltimore County

Table of Contents

1. Abstract **3**

2. Introduction **3**

3. Objectives **3**

4. Related Work **4**

5. Implementation **4**

6. Experimental Results **13**

7. Future Work **13**

8. Conclusion **14**

9. References **15**

1. Abstract

This research investigates the severe effect of mental health difficulties such as anxiety, depression, and fatigue on people's productivity, decision-making skills, and overall job performance. We investigate how these concerns can increase the possibility of errors, overlooking critical facts, and engaging in risky activities among employees, weakening cybersecurity security measures. Our research looks at the impact of mental health on humans and organizations, and on understanding its implications for cyber threats and attacks. We aim that this investigation can bring insight into the connections between mental health and cybersecurity, as well as provide ideas on how to build a healthier and more secure workplace.

2. Introduction

Mental health problems, such as stress, anxiety, depression, or burnout, can have a significant impact on an individual's productivity, decision-making, and overall job performance. When employees face mental health issues, they may become more prone to making mistakes, overlooking critical details, or engaging in risky behaviors that jeopardize cybersecurity measures. This project describes the impact of an individual's mental health on both the personal and organizational levels, with a particular emphasis on cyber threats and attacks.

3. Objectives

As part of this project, we will analyze a large survey dataset capturing mental health experiences across the technology workforce to develop predictive models that:

• Identify employees most at risk for mental health issues based on demographics, workplace environment, and policy factors. Algorithms used will include logistic regression, random forest classifiers, and nearest neighbors. Key predictions will assess the overall risk of conditions like depression, anxiety, burnout, and substance abuse.

• Also quantify the potential impacts of poor mental health on employee productivity, resignation rates, and obesity/chronic disease risks. Predictive models will demonstrate how improving mental health could yield cost savings from retaining talented employees.

• Link frequent symptoms like fatigue, trouble concentrating, and irritability to increased cybersecurity risks including vulnerable online behaviors, accidental data exposures, and overlooking security protocols. Finally, quantifying these risks can motivate tech companies to prioritize mental well-being.

With 1 in 5 adults facing mental illness, technology companies must consider employee mental health to build resilient, creative, and cyber-conscious workforces. This project will inform corporate wellness policies and workplace culture changes that support mental health. Enabling professionals to thrive both mentally and physically will strengthen productivity, satisfaction, diversity, and ultimately data protection within the vital tech sector.

4. Related Work

In recent years, there has been a growing focus on the link between employee mental health and cybersecurity risk. Several research has investigated how psychological factors affect an individual's susceptibility to social engineering assaults and unintentional security breaches. Gratian et al. (2018) explored how employee fatigue affects vulnerability to phishing and malware assaults. Their findings revealed that burnout impairs attention and decision-making, leaving people more vulnerable to cyber-attacks. Similarly, Jalava and Mauno (2022) investigated the relationship between employment uncertainty, stress, and cybersecurity compliance. Employees who experienced higher degrees of job instability and stress were shown to be less likely to follow corporate security policies and practices. In addition to individual variables, corporate culture, and leadership have been highlighted as critical components in creating cyber-resilience. Cram et al. (2019) developed a socio-technical model to better understand the relationship between organizational characteristics, employee well-being, and cybersecurity behaviors. Their findings emphasized the need to create a supportive work environment and offer mental health tools in order to decrease cyber dangers. While these studies have provided useful insights, more complete risk models that incorporate a variety of elements, such as individual mental health indicators, organizational culture, and cybersecurity incident data, are still needed. Our research attempts to close this gap by establishing a comprehensive strategy for understanding and mitigating the impact of mental health on cyber-resilience in the technology industry.

5. Implementation

Data Preparation:

The dataset retrieved from Kaggle is first cleaned by removing all the unwanted columns and duplicate rows. Later, the null values are identified and then the respective rows are removed.

Libraries used;

```
[1]  import pandas as pd
     import numpy as np
     from sklearn import preprocessing
     import matplotlib.pyplot as plt
     import seaborn as sns
```

Sample dataset;

```
[4] df.head()
```

	Timestamp	Age	Gender	Country	state	self_employed	family_history	treatment	work_interfere	no_employees	...	leave	m
0	2014-08-27 11:29:31	37	Female	United States	IL	NaN	No	Yes	Often	6-25	...	Somewhat easy	
1	2014-08-27 11:29:37	44	M	United States	IN	NaN	No	No	Rarely	More than 1000	...	Don't know	
2	2014-08-27 11:29:44	32	Male	Canada	NaN	NaN	No	No	Rarely	6-25	...	Somewhat difficult	
3	2014-08-27 11:29:46	31	Male	United Kingdom	NaN	NaN	Yes	Yes	Often	26-100	...	Somewhat difficult	
4	2014-08-27 11:30:22	31	Male	United States	TX	NaN	No	No	Never	100-500	...	Don't know	

Data Exploration:

We explored the data by using various numpy and pandas functions to understand the data. Using the shape attribute we found out there are 1259 rows and 27 columns.

Shape of the dataset;

```
[6] df.shape

    (1259, 27)
```

Feature Engineering:

As we don't have a target attribute, hence we have conducted feature engineering such that, the combination of the 'treatment' and 'work_interfere' attributes closely resembles the amount of cybersecurity risk. So the values combination between these 2 attributes gave rise to the new attribute which is labeled as cybersecurity_risk.

```
[20] def create_cybersecurity_risk(row):
         if row['treatment'] == 'yes' and row['work_interfere'] in ['often', 'sometimes']:
             return 'High'
         else:
             return 'Low'
     df['cybersecurity_risk'] = df.apply(create_cybersecurity_risk, axis=1)
```

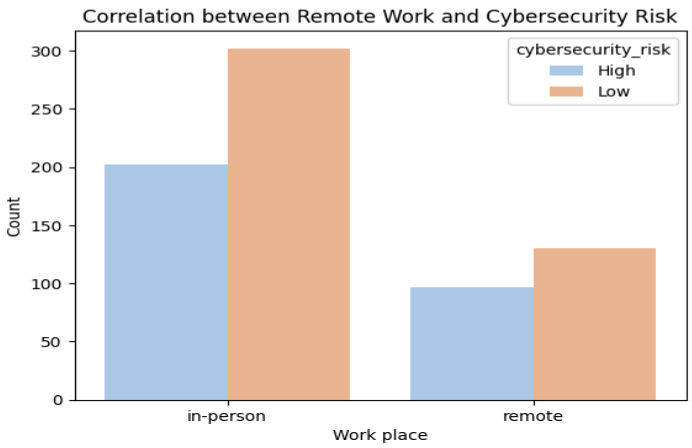
Data Visualization:

For a quick capture of how the dataset is, the below graphs depicts the information about how the data is distributed and also the correlation between different attributes.

Remote Work and Cybersecurity Risk:

Are individuals who work remotely more or less likely to report observed cybersecurity consequences in their workplace?

```
[22] sns.countplot(x='remote_work', hue='cybersecurity_risk', data=df, palette='pastel')
     plt.xlabel('Work place')
     plt.ylabel('Count')
     plt.title('Correlation between Remote Work and Cybersecurity Risk')
     plt.show()
```

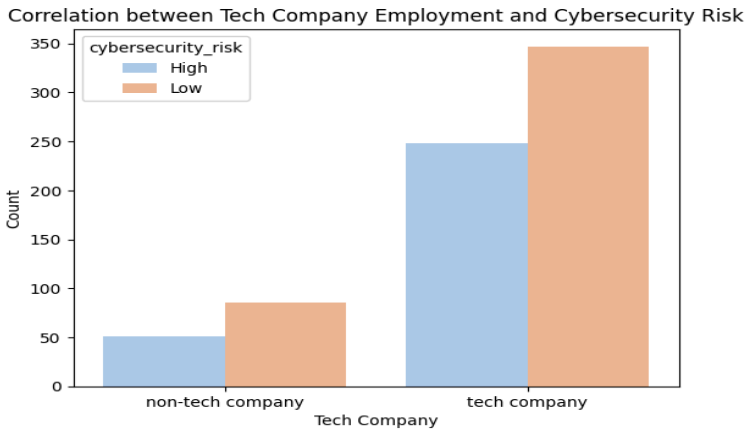


- The graph depicts the association between remote employment and cybersecurity risk. It compares the number of people who face high and low cybersecurity risks when working in-person versus remotely.
- For in-person employment, the number of people with low cybersecurity risk is around 200, while the number of people with high cybersecurity risk is substantially smaller, around 50.
- Individuals with high cybersecurity risk (about 130) are more likely to work remotely than those with low cybersecurity risk (around 100).
- This shows that people who operate remotely are more likely to face cybersecurity dangers than people who work in person. The graph depicts the increasing sensitivity to cyber attacks connected with remote work situations.

Tech Company Employment and Cybersecurity Risk:

Do employees in tech companies perceive higher cybersecurity risks compared to employees in non-tech companies?

```
[23] sns.countplot(x='tech_company', hue='cybersecurity_risk', data=df, palette='pastel')
     plt.xlabel('Tech Company')
     plt.ylabel('Count')
     plt.title('Correlation between Tech Company Employment and Cybersecurity Risk')
     plt.show()
```

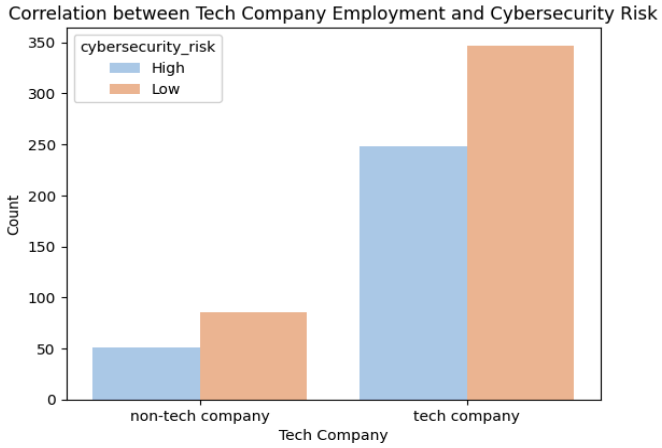


- The graph contrasts employees' perceptions of cybersecurity risk in technology businesses against non-tech companies.
- Employees at non-tech companies perceive a low cybersecurity risk (about 80) slightly more than those who perceive a high cybersecurity risk which is around 50.
- However, in technology organizations, the number of employees feeling a high cybersecurity risk (about 210) is much higher than those perceiving a low cybersecurity risk which is approximately 290.
- This suggests that employees in tech organizations are more likely to perceive greater cybersecurity dangers than employees in non-tech companies. The gap between high and low perceived risk levels is notably bigger for technology business employees, implying that they are more aware of or exposed to potential cyber risks in their workplace.

Tech Company Employment and Cybersecurity Risk:

Do employees in tech companies perceive higher cybersecurity risks compared to employees in non-tech companies?

```
[23] sns.countplot(x='tech_company', hue='cybersecurity_risk', data=df, palette='pastel')
     plt.xlabel('Tech Company')
     plt.ylabel('Count')
     plt.title('Correlation between Tech Company Employment and Cybersecurity Risk')
     plt.show()
```

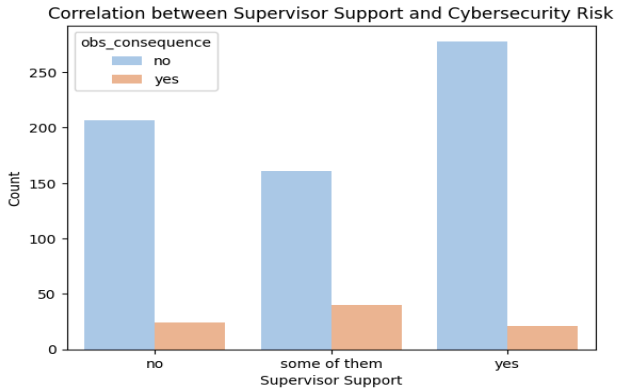


- The graph compares the perceived cybersecurity risk levels of employees in technology and non-tech organizations.
- In non-tech companies, the number of employees who perceive a low cybersecurity risk is around 75, while the number of those who perceive a high cybersecurity risk is lower, around 45.
- In contrast, in technology organizations, the number of employees feeling a high cybersecurity risk (about 210) is much higher than those experiencing a low cybersecurity risk (approximately 290).
- The research clearly reveals that employees in technology organizations are more likely to perceive larger cybersecurity dangers than those in non-technology companies. This perceived gap could be related to the nature of tech organizations working more directly with technology systems and data, making their employees more aware of potential cyber threats.

Supervisor Support and Cybersecurity Risk:

Are employees who are willing to discuss mental health issues with their supervisors less likely to experience cybersecurity incidents?

```
[25] sns.countplot(x='supervisor', hue='obs_consequence', data=df, palette='pastel')
     plt.xlabel('Supervisor Support')
     plt.ylabel('Count')
     plt.title('Correlation between Supervisor Support and Cybersecurity Risk')
     plt.show()
```

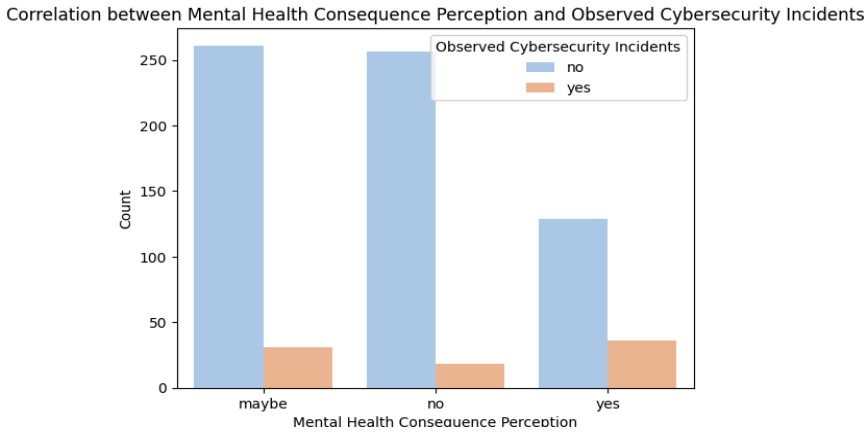


- The bar graph depicts the association between supervisor support levels and cybersecurity incident outcomes, which suggests a possible negative relationship. It implies that when supervisor support increases from 'No' to 'Some of them' and 'Yes,' the number of consequential cybersecurity events declines.
- Specifically, when supervisors provide enough support, the occurrence of consequential incidents significantly decreases. This finding emphasizes the importance of supervisor support in improving organizational cybersecurity resilience. However, while the graph suggests a correlation, more statistical analysis is needed to definitively establish this association.

Effect of Mental Health Consequences Perception on Cybersecurity Incidents:

Do employees who perceive negative consequences for discussing mental health issues with their employer report more cybersecurity incidents?

```
[26] sns.countplot(x='mental_health_consequence', hue='obs_consequence', data=df, palette='pastel')
     plt.xlabel('Mental Health Consequence Perception')
     plt.ylabel('Count')
     plt.title('Correlation between Mental Health Consequence Perception and Observed Cybersecurity Incidents')
     plt.legend(title='Observed Cybersecurity Incidents', loc='upper right')
     plt.show()
```



- The bar graph examines the relationship between employees' perceptions of negative consequences for discussing mental health issues with their workplace and actual cybersecurity incidents.
- Employees who perceive potential bad effects ('maybe') have a higher number of observed cybersecurity issues than those who do not perceive any negative implications, with around 230 incidents reported. Similarly, employees who see definite negative repercussions ('yes') have a greater total of observed occurrences (about 100), albeit less than the 'maybe' group.
- Employees who perceive no negative repercussions have the fewest observed incidences, at 200. This implies a potential relationship between employees' views of repercussions for discussing mental health and their chance of encountering cybersecurity issues, underscoring the need of building supportive environments for mental health in minimizing cybersecurity risks.

Encoding the categorical data:

To do the modeling, the data should contain numerical data rather than non-numerical data. Hence, we conducted label encoding where the columns with Yes/No values are changed to 1/0 respectively and the type of cybersecurity risk columns are of 2 unique values and they are assigned as 0 or 1.

```
[28] from sklearn.preprocessing import LabelEncoder

     # Initialize LabelEncoder
     label_encoder = LabelEncoder()

     # Perform label encoding for each categorical column
     for column in df.select_dtypes(include=['category']).columns:
         df[column] = label_encoder.fit_transform(df[column])

     df.head()
```

	Timestamp	Age	Gender	Country	state	self_employed	family_history	treatment	work_interfere	no_employees
0	0	37	0	3	10	2	0	1	1	4
1	1	44	1	3	11	2	0	0	2	5
4	2	31	1	3	37	2	0	0	0	1
5	3	33	1	3	36	2	1	0	3	4
6	4	35	0	3	18	2	1	1	3	0

Split the dataset:

As we have our dataset completely structured, so now we started splitting the dataset. It is divided into 80% training and 20% testing, where initially the algorithms will be trained on 80% of the dataset and they find a pattern. Later this pattern is tested on the rest 20% of the dataset.

Testing:

As we have split the dataset, we now use them to calculate the accuracy with various algorithms. Based on our problem statement our ultimate goal is to predict if cybersecurity is high or low, hence it is supervised learning as there are labels associated here. Out of which it is a classification type as we are trying to classify if there is cybersecurity risk or not. Now we will investigate different classification algorithms.

1. Logistic regression:

It is a classification algorithm used to predict the type of category with input values. We have tweaked the model by passing various inputs, so max_iter=10000 will iterate the dataset many times to give the final score.

```
from sklearn.linear_model import LogisticRegression
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)
log_reg = LogisticRegression()
log_reg.fit(X_train, y_train)
log_reg_pred = log_reg.predict(X_test)
log_reg_accuracy = accuracy_score(y_test, log_reg_pred)
print("Logistic Regression Accuracy:", log_reg_accuracy)
```

```
Logistic Regression Accuracy: 0.8503401360544217
```

2. Random Forest Classifier:

It is another type of classification algorithm. Each decision tree in a random forest is constructed using a random subset of the input features and the training data. This procedure aids in lowering overfitting and enhancing model correctness. Following the construction of each individual tree, the random forest aggregates the predictions made by the trees by averaging the outputs for regression issues or selecting the majority vote for classification problems.

```
from sklearn.ensemble import RandomForestClassifier
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)
rfc = RandomForestClassifier()
rfc.fit(X_train, y_train)
rfc_pred = rfc.predict(X_test)
rfc_accuracy = accuracy_score(y_test, rfc_pred)
print("Random Forest Classifier Accuracy:", rfc_accuracy)
```

```
Random Forest Classifier Accuracy: 0.9727891156462585
```

3. Support Vector Machine (SVM):

It is used for classification, regression, and outlier detection tasks. The SVM classifier is used while segregating the two classes(hyper-plane/line). They mainly separate data until a hyperplane with a high minimum distance is found and it is used to classify two or more data types.

```
from sklearn.svm import SVC
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)
svm = SVC()
svm.fit(X_train, y_train)
svm_pred = svm.predict(X_test)
svm_accuracy = accuracy_score(y_test, svm_pred)
print("Support Vector Machine Accuracy:", svm_accuracy)
```

```
Support Vector Machine Accuracy: 0.6530612244897959
```

4. K-Nearest Neighbors:

This algorithm works by locating the mentioned K nearest data points in the training dataset to its test data, and then based on the kind of the problem it will predict the output. For a classification problem, it will look for the majority of instances, whereas for the regression problem, it will take the average.

```
from sklearn.neighbors import KNeighborsClassifier
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)
knn = KNeighborsClassifier()
knn.fit(X_train, y_train)
knn_pred = knn.predict(X_test)
knn_accuracy = accuracy_score(y_test, knn_pred)
print("Random Forest Classifier Accuracy:", knn_accuracy)
```

```
K-Nearest Neighbors Accuracy: 0.5714285714285714
```

6. Experimental Results

From the above classification models, we can observe that Random Forest Classifier have the best accuracy compared to other algorithms.

7. Future work

There are several areas of future work that can help to address the issue of mental health in tech space;

1. ***Human-AI Collaboration:*** Explore opportunities for human-AI collaboration in addressing mental health and cybersecurity challenges. Develop AI-powered tools and chatbots to provide personalized support, detect early signs of distress, and promote cyber hygiene practices among employees.
2. ***Ethical Considerations:*** Address ethical considerations related to data privacy, consent, and algorithmic bias in mental health and cybersecurity research. Develop ethical guidelines and frameworks to ensure responsible data use and protect employee rights.
3. ***Industry Partnerships:*** Foster partnerships between academia, industry, and government agencies to co-create solutions and share best practices for promoting mental health and cybersecurity in the tech industry. Collaborate on research projects, knowledge exchange forums, and policy advocacy initiatives.
4. ***Employee Empowerment:*** Empower employees to take ownership of their mental health and cybersecurity by providing training, resources, and support networks. Foster a culture of openness, trust, and collaboration where employees feel empowered to seek help and report security incidents without fear of stigma or retaliation.

8. Conclusion

Our project aimed to analyze the intersection of mental health and cybersecurity within the tech industry workforce. Through comprehensive data analysis and machine learning modeling, we sought to identify patterns, predict cybersecurity risks, and provide insights to support employee well-being and data security measures.

Based on the various accuracy achieved from different classification algorithms, we can conclude that random forest classifier was the best, but has the possibility of the occurence of over-fitting. Hence followed by the logisitc regression would give us better results with it's moderate to better accuracy.

9. References

1. Gratian, M., Bandi, S., Cukier, M., Dykstra, J., & Ginther, A. (2018). Correlating human traits and cyber security behavior intentions. Computers & Security, 73, 345-358.

2. Jalava, J., & Mauno, S. (2022). Employee job insecurity and cybersecurity compliance: The mediating role of psychological strain. Information & Computer Security, 30(1), 58-74.

3. Cram, W. A., Proudfoot, J. G., & D'Arcy, J. (2019). Organizational information security policies: a review and research framework. European Journal of Information Systems, 28(6), 605-641.