The Isolation Equation: Remote Work and Mental Health Post COVID-19

AARYA NAGRANI, ALEX FENG, ANIKA AHMED, CADE VINEY, KIMBERLY BROWN

0. Abstract

The global transition to remote work, accelerated by the COVID-19 pandemic, has reshaped professional and personal dynamics, giving rise to unique challenges for individuals adopting a digital nomad lifestyle. This study investigates the factors contributing most significantly to mental health outcomes among this new wave of digital nomads in a variety of industries. Using Naïve Bayes and Support Vector Machine machine learning models we analyzed a dataset of self-reported mental health indicators, workplace conditions, and demographic information to determine if, based on a subset of these values, we could adequately classify whether a respondent has a mental illness. Our findings highlight the pivotal role of industry, job role, gender, and satisfaction with remote work in predicting mental health risks. These insights offer valuable guidance for organizations and policymakers aiming to support mental well-being in the rapidly growing remote workforce of the post-pandemic era.

1. Introduction

1.1. Motivation

The adoption of remote work as a mainstream professional model has brought flexibility and freedom to the workforce. However, it has also introduced significant stressors and loneliness in a continually online world. Previous studies have shown that the increased prevalence of remote work may "undermine mental health" and result in changes to habits and behavior due to increased isolation. (Islam et al., 2023) As the remote workstyle gains traction among young professionals in tech-driven industries, understanding the mental health implications of this lifestyle is increasingly critical.

This study seeks to unravel the key factors influencing mental health outcomes in remote workers across a variety of fields and experience levels. We hope that with the knowledge of the major contributing factors, those directly impacted will be more aware and purposeful in addressing what work-life balance truly looks like in today's digital age.

1.2. The Data

Our dataset comprises information collected from over 5,000 participants worldwide, focusing on those with remote work arrangements. Key attributes include:

- Mental Health Indicators: Self-reported stress levels, symptoms of anxiety and depression, and overall satisfaction scores.
- Workplace Conditions: Work-life balance, job stability, working hours,

and access to mental health resources.
- Demographic and Lifestyle Data: Age, gender, geographic mobility, and frequency of social interactions.

The dataset integrates survey responses, anonymized health records, and publicly available information. Data collection spanned 2021-2023 to capture evolving trends in the post-pandemic workforce.

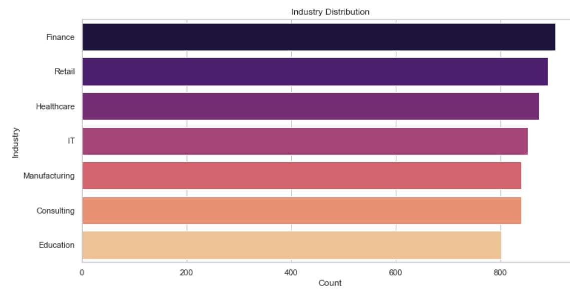
1.3. The Plan

This paper is structured as follows: In Section 2, we outline the data preprocessing steps employed to prepare the dataset for analysis, including handling missing values and feature engineering. Section 3 presents the machine learning models applied and evaluates their performance. Section 4 discusses key findings, with Section 5 offering conclusions and recommendations for stakeholders. Finally, Section 6 identifies avenues for future research.

2. Pre-Processing the Data

2.1. A Closer Look at the Data

We conducted initial visualizations using Python's built-in libraries (including pandas, numpy, matplotlib, etc.) to explore the dataset which revealed that our data was uniformly distributed across participant age, gender-identity, job role, industry, experience, work-life balance rating, remote work satisfaction rating, and mental health condition. These were the primary features we wanted to ensure consistency across as we were primarily interested in the effects of working remotely on anxiety, depression, and burnout. Industry domain will play a large role in our data so the distribution should be acknowledged:



One limitation of our data we would like to give specific declaration of is the proportion of non-mentally ill to mentally-ill respondents. This will later play a large role in the inability to effectively label non-mentally ill subjects.

2.2. Feature Engineering/Eliminating Redundancies

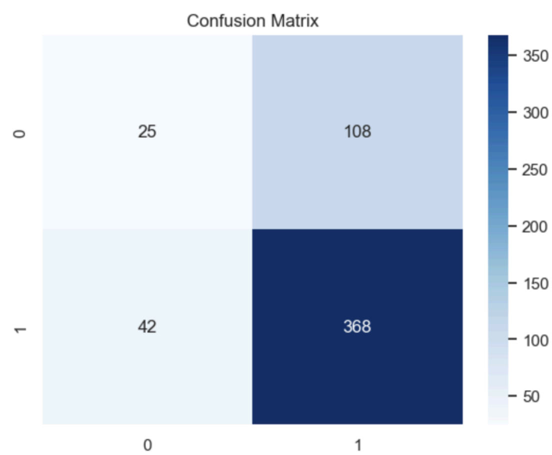
We addressed null or missing values within the dataset with a placeholder value of "None" to indicate the absence of data. We dropped columns with no correlation to our analysis's purpose such as "Employee_ID" and ensured rows with missing data were exempt. We one-hot encoded relevant categorical variables, converted and scaled ordinal values, and prepared our data to be split into training and validation sets.

3. Models

3.1. Naïve Bayes

The Naïve Bayes classification algorithms are a collection of algorithms and "one of the most effective text classification methods" because they are able to efficiently classify data with independent features. (Huang and Li, 2011) Naïve Bayes is effective for large-scale and small-scale datasets, making it a good model for us to understand the impacts of remote work on mental health. In our project, we opted for the Multinomial Naïve Bayes classifier due to its proficiency in handling categorical data – aligning well with our one-hot encoded features.

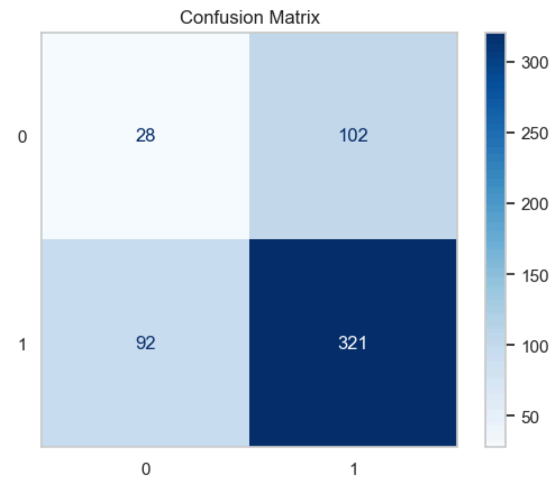
After we trained the model on 25 features, the Multinomial Naïve Bayes classifier predicts the testing set with an accuracy of 72.38%.



3.2. SVM (Support Vector Machine):

We decided to use a Support Vector Machine (SVM) with a linear kernel to classify instances into two categories: those with and without a mental health condition. We used SVM because it performs well for classification in a high-dimensional feature space. (Mammone et al., 2009) To address class imbalance, we created synthetic minority class samples by applying SMOTE to the training data. (Blagus and Lusa, 2013)

After we trained the model on 35 features, the SVM classifier predicts the testing set with an accuracy of 64.27%.



4. Discussion

4.1 SVM

Upon evaluation using the validation set, the SVM model achieved an overall accuracy of 64.27%. The classification report revealed that for the negative class, the model attained a precision of 23% and a recall of 22%, resulting in a F-1 score of 22%. Conversely, for the positive class (presence of mental health condition), the model achieved a precision of 76%, a recall of 78%, and an F1-score of 77%. These performance metrics indicate that while the model demonstrates a reasonable capacity to identify instances with a mental health condition, its efficacy in accurately detecting instances without such conditions is markedly limited.

4.2 Naïve Bayes:

We evaluated the Naïve Bayes model using the validation set which achieved a true positive rate of 77.31%, indicating that it accurately identified mental health conditions with a false positive rate of

22.69%. Yet, the true negative rate of 37.31% should be noted when considering the models overall accuracy. With a false negative rate of 62.69%, the model underscores correctly identifying positive cases. These metrics suggest that while the model exhibits a reasonable ability to identify positive instances, its capacity to reliably detect and classify negative cases is significantly constrained, limiting its utility in balanced or high-stakes scenarios.

5. Summary

5.1. Best Models and Accuracy

Table 2 is a table of the models we used and their accuracy.

5.2. Conclusion

We used machine learning algorithms to analyze factors that impact mental health conditions among remote workers. From our Naïve Bayes model, we identified that industry, gender, and satisfaction with remote work impacted the mental health condition of workers. From our SVM model, we identified that job role, industry, and region impacted the mental health condition of workers. Both models are a powerful tool for prediction of true positive cases of mental illness but should not be considered for cases in which they predict no mental illness. We hope these models serve new respondents with the means of understanding if they are forming habits or introducing lifestyle changes that will veer them towards mental illness with increased certainty.

6. Future Work

The data we obtained from Kaggle was relatively limited and we believe, with a larger sample size, more reliable trends could have been discovered. In addition to a larger dataset, it would be interesting to see how the link between the conditions of remote work relate to a greater variety of mental illnesses and conditions. The dataset only provided the distinction between depression, anxiety and burnout, which really scratches the surface of the possibilities the post-pandemic virtual workplace has brought about. We implore future researchers to make a concerted effort in defining the nuances of controllable factors specific to the remote lifestyle that have a unique bearing to emotions and conditions such as loneliness, lack of purpose, and energy.

References

[1] Islam, N., Baun, K., Racette, R., *Effects of telework on anxiety and depression across the United States during the covid-19 crisis*. PLoS One. *18*(1)

[2] Mammone, A., Turchi M., Cristianini N., *Support vector machines.* WIREs Computational Statistics *1*(3) pp. 283-289

[3] Huang, Y., Li, L., *Naive Bayes classification algorithm based on small sample set*. 2011 IEEE International Conference on Cloud Computing and Intelligence Systems pp. 34-39

[4] Blagus, R., Lusa, L. *SMOTE for high-dimensional class-imbalanced data*. BMC Bioinformatics *14*(106)