Examining Mental Health Outcomes in the Age of Remote Work

University of Missouri – Saint Louis

By:

Kevin Kpankou, Jack Rhodes, Jill Sisson, and Amaan Syed

November 30, 2024

# Abstract

With remote work increasingly viewed as a viable option, especially given the current economic conditions, we aim to use the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology to analyze the dataset and identify patterns among features. Techniques such as correlation analysis, regression, and machine learning will be used to uncover these patterns. We hope these findings provide greater insight and draw attention to potential issues associated with remote work, helping to mitigate mental health risks.

# Report Summary

This report examines the correlation between remote work and changes in mental health status. We selected the "Remote Work and Mental Health" dataset for this project, which contains 5,000 records of features related to mental health, capturing aspects such as age, gender, and stress level. Important fields like hours worked per week, industry, and job roles are all included. According to Glassdoor (2018), “Understanding the nuances of workplace culture across different countries is essential for fostering effective international collaborations and managing diverse teams.” The dataset effectively captures data from various countries, providing a broad perspective on work cultures across the globe.

For our model, we chose the DecisionTreeClassifier. We selected this model because it is interpretable and transparent, offering intuitive reasoning for decisions made within the model. This approach aligns well with the features we have chosen, as it allows us to understand and visualize how each factor influences the outcomes.

During the data preparation phase, we focused on our target variable, “Mental Health Condition,” which initially contained a significant number of missing values. These missing values were addressed before applying the decision tree classifier. Many features in this dataset reflect daily life factors, making feature importance a priority for our model. Additionally, the decision tree’s ease of visualization made it an ideal choice.

The parameters chosen for the decision tree classifier were entropy for the criterion, a maximum depth of 8, and a random state of 42. These parameters were determined using GridSearchCV to identify the best combination for accuracy, ultimately improving our model’s performance.

# 

# Introduction

The rise of remote work was sudden and disruptive. Pre-2021, remote work was seen as something that technology would eventually allow us to do; the ability to work efficiently from outside office buildings was expected to go hand in hand with technological advancements, making our lives more flexible. However, the grim state of remote work that emerged as a result of the COVID-19 pandemic was not what we had hoped for. It was sudden and unpredictable, leaving many people unable to fend for themselves. Businesses shut down, and people had to get creative with how they would ration and preserve basic goods and necessities. COVID-19 had many byproducts; due to the infectious nature of the disease, quarantining became common and necessary at that time. The need to maintain distance between individuals made it virtually impossible to work within an office.

As a result, COVID-19 brought about the rise of online conference platforms, with Zoom and Google Meet being the most prominent during this time. These platforms allowed people to meet and video call from the comfort of their home or place of shelter, enabling a certain degree of productivity during the pandemic. With the ability to communicate effectively, companies affected by the pandemic had a way to keep their doors open. Remote work became the solution that many companies adopted; it not only allowed employees to communicate effectively but also helped avert the risk of transferring COVID-19 to others nearby.

Eventually, as the pandemic worsened, most parts of society had to take the same route as companies did. While this was a metaphorical Band-Aid on the problem of COVID-19, working from home, or remote work, also brought with it many new challenges. For many people, the isolation associated with remote work negatively affected mental health.

With our study, we aim to examine the effects of remote work on mental health and how it impacts individuals psychologically.

# CRISP-DM Framework

The CRISP-DM framework is an industry-wide standard process for data mining that allows teams and companies to navigate the process of data mining, as well as providing a more interpretable process from the viewpoint of stakeholders. The first step of CRISP-DM is business understanding, and in this phase, the goal is to identify problems and understand how we can define project goals to mitigate or potentially solve those problems, as well as how we can use data mining techniques to help solve them. In our remote work and mental health research project, our problems are those that affect human health, specifically mental health, and how it translates into the workplace due to new ways of working, like remote work, which often leads individuals to spend large amounts of time in isolation.

The second step of CRISP-DM is data understanding, which involves gathering, describing, and exploring the data. For this, we used Kaggle, a popular site for data analysts and data scientists. Kaggle has hundreds, if not thousands, of datasets that are easily accessible. In the data preparation stage, we focused on cleaning and preparing data for modeling, which primarily affected our target variable, which was mental health condition in the dataset. Using Pandas, we found that almost all, if not most, of our missing values came from the mental health condition column.

In our modeling phase, we chose to use the decision tree classifier because of its interpretability, ability to capture non-linear relationships, and its flexibility in handling outliers and missing data.

Evaluation is the fifth step in the CRISP-DM process. In this stage, we evaluate our model's performance and see if it achieves the goals established during the business understanding phase. We used metrics like accuracy, and precision to assess the model's performance after creating a decision tree classifier to examine the connection between remote work and mental health outcomes. These measures gave us insight into how well the model could categorize people's mental health issues. In order to determine which factors had the most effects on mental health in the context of remote work, we also looked at feature importance. By means of this assessment, we made certain that the model not only demonstrated strong statistical performance but also yielded significant insights that are consistent with our research objectives. Any flaws discovered during the evaluation process, such as overfitting or low predictive power in certain categories, were carefully documented and resolved, as model dependability is critical for making meaningful suggestions.

Deployment, the last phase, includes putting the results of the study to use. We thought about how these results might be used in practical situations, even though this research effort does not entail implementing the model in a live environment. Businesses might utilize the information from our model, for example, to create stronger mental health support systems for remote workers or customize regulations for remote work based on social isolation and work-life balance, two characteristics that have an impact on mental health. Businesses may make data-driven decisions that support employee well-being by knowing how remote work affects mental health. With more data, the model might be further improved in an applied context and used to track changes in mental health over time, giving organizations a continuous resource for dealing with the unique challenges of remote work.

## Business Understanding

From a business standpoint, the ability for employees to work remotely was more of a reflex to the pandemic. However, post-pandemic, some companies still allow certain employees to work remotely. Also, some fields are more likely to work remotely, such as, for example, technology fields like software development. Remote work helps save resources for businesses, as well as allowing them to communicate with employees through various means. According to Maryville University (2023), “only 7% of workers with jobs that could be done remotely reported working off-site; as of March 2023, that percentage had increased fivefold.”

From a business standpoint, certain roles don't require an employee to be in the office at all times; therefore, the switch to remote work has been an obvious one. The problem we try to tackle is from the employee side. We, as humans, are naturally social creatures, and we need constant social interaction to function properly. With that being said, interacting with different people every day allows us to learn and grow from each other. According to Khan (2024), “Working from home comes with its own set of stressors that are often underestimated. The merging of personal and professional spaces can be mentally taxing.”

The lack of social sustenance that remote work brings can impact us differently depending on our current situation. People who have more support from family members or friends, or who enjoy a more social and outgoing lifestyle, can still see the effects of remote work on mental health, but not as much as someone who relies on their coworkers not just as coworkers but as friends.

There also comes up the topic of not being able to focus properly in an area with which you are comfortable. Various sources detail the dichotomy between being able to work productively and being in a space where you feel comfortable and relaxed. Just the visual cues, like having your bed close to you, can affect your ability to be productive. According to Apartment Therapy (2020), “Why? Your room is a place for rest. If you bring your work into that space, your brain and body might associate it with productivity, which could make it harder to sleep. On the flip side, you might be tempted to relax or even doze off in your room when you’re supposed to be getting things done.”

Without the increased pressures of an office building, such as other people keeping you accountable when you’re not productive, and the desire to be seen as a valuable contributor to the workplace, thoughts like taking a nap during the workday, getting to tasks whenever you feel like it, or focusing on household tasks instead of the job at hand are challenges that we as humans struggle with

## Data Understanding

Our dataset came from Kaggle, a popular platform for data analysts and scientists. There are several publicly accessible datasets on Kaggle, and this one fits in nicely with our research goal of comprehending the effects of remote labor on mental health. Age, gender, job role, industry, stress level, work location, mental health condition, and satisfaction with remote work are just a few of the characteristics included in the 5,000 entries that make up the dataset. These aspects offer a thorough understanding of personal traits, workplaces, and mental health metrics. All of which are essential to our investigation.

We used summary statistics to perform an exploratory study in order to better comprehend the data. Several important revelations were made by the df.info() and df.describe() outputs. First, the dataset has seven numerical fields (such as Age, Years of Experience, Hours Worked per Week, and Work Life Balance Rating) and thirteen categorical elements (such as Gender and Job Role). According to the numerical breakdown, the workforce is both youthful and seasoned, with ages ranging from 22 to 60 and a mean of 41. A significant percentage of the sample appears to be composed of seasoned experts, as indicated by the mean Years of Experience, which is roughly 18.

We also observed intriguing results in factors that are directly associated with work-life balance and mental health. For example, on a scale that typically ranges from 1 to 5, the average Work Life Balance Rating is approximately 2.84, indicating a generally moderate work-life balance. Indicating varying experiences with remote work assistance and feelings of isolation, the Social Isolation Rating and Company assistance for Remote Work scales similarly showed average values close to the midpoint. A further indication of possible outliers or different work arrangements among people was the vast range of Hours Worked per Week, which ranged from a minimum of one hour to a maximum of sixty hours.

When evaluating the dataset's quality, we found that most columns had no missing values, making it generally robust. To guarantee consistency, we will need to fix a few missing entries in crucial categories like Physical Activity and Mental Health Condition during the data preparation stage. With mean values, standard deviations, and range information that seemed realistic and in line with expectations from the actual world, the summary statistics further supported the dependability of the data.

Stress Level, Mental Health Status, Work-Life Balance Rating, and Social Isolation Rating are important variables of importance for our analysis. These characteristics are essential for analyzing relationships and possible effects since they directly represent the mental health and the general well-being of those working remotely. By concentrating on these factors, we hope to find connections.

Our dataset is of high quality and well-suited for examining mental health outcomes in the context of remote employment, according to our exploratory analysis. We are ready to go forward with the data preparation and other stages of our investigation now that we have a firm grasp on the structure and properties of the data.

## Data Preparation

We performed a number of essential procedures to guarantee the dataset's quality and usefulness for investigating the connection between remote work and mental health outcomes prior to analyzing it. We began by identifying and eliminating fields that were unnecessary, particularly the "Employee\_ID" column. We were able to condense the dataset and concentrate on characteristics that directly contribute to our analysis of mental health in a remote work situation by eliminating this field because it was irrelevant to the study's objective.

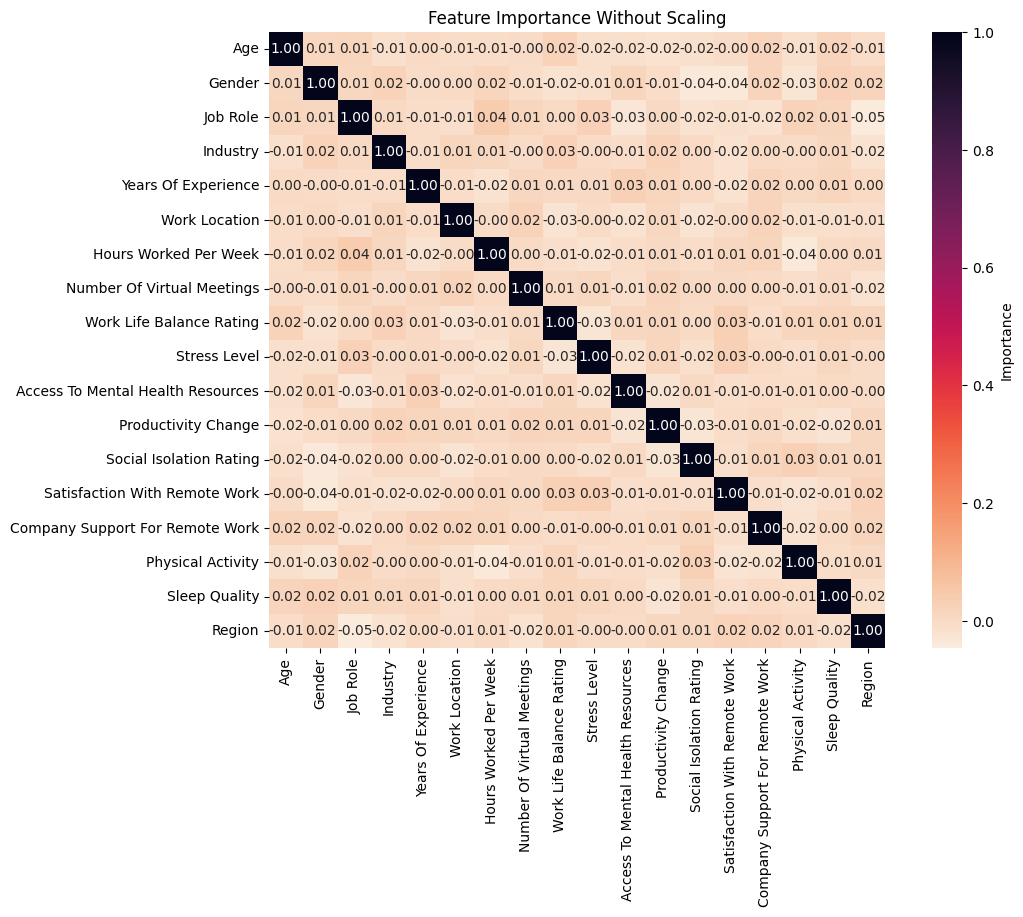
One of the most important aspects of the data preparation process was dealing with missing values. To fill in the blanks for important columns like "Physical\_Activity" and "Mental\_Health\_Condition," we used the placeholder value "N/A." We were able to preserve the data structure while retaining a larger percentage of the dataset because of our method's good handling of missing data. We preserved a strong dataset for analysis without sacrificing its integrity by filling in these gaps rather than eliminating entire rows. Then, in order to find any last missing values, we performed a general check over all columns. To guarantee a clean, high-quality dataset, we decided to eliminate rows containing null entries after assessing the magnitude of these missing values. This step was essential for maintaining the data's correctness and consistency, particularly for analyses that are sensitive to missing data.

In order to determine category qualities, we lastly looked at the distinct values found in each column. We were able to discern between numerical and categorical variables by examining unique entries in each column, which provided us with information about the structure of our dataset. We were able to identify variables that could need special processing, like encoding for modeling reasons, and gain a better understanding of the data distribution thanks to this step.

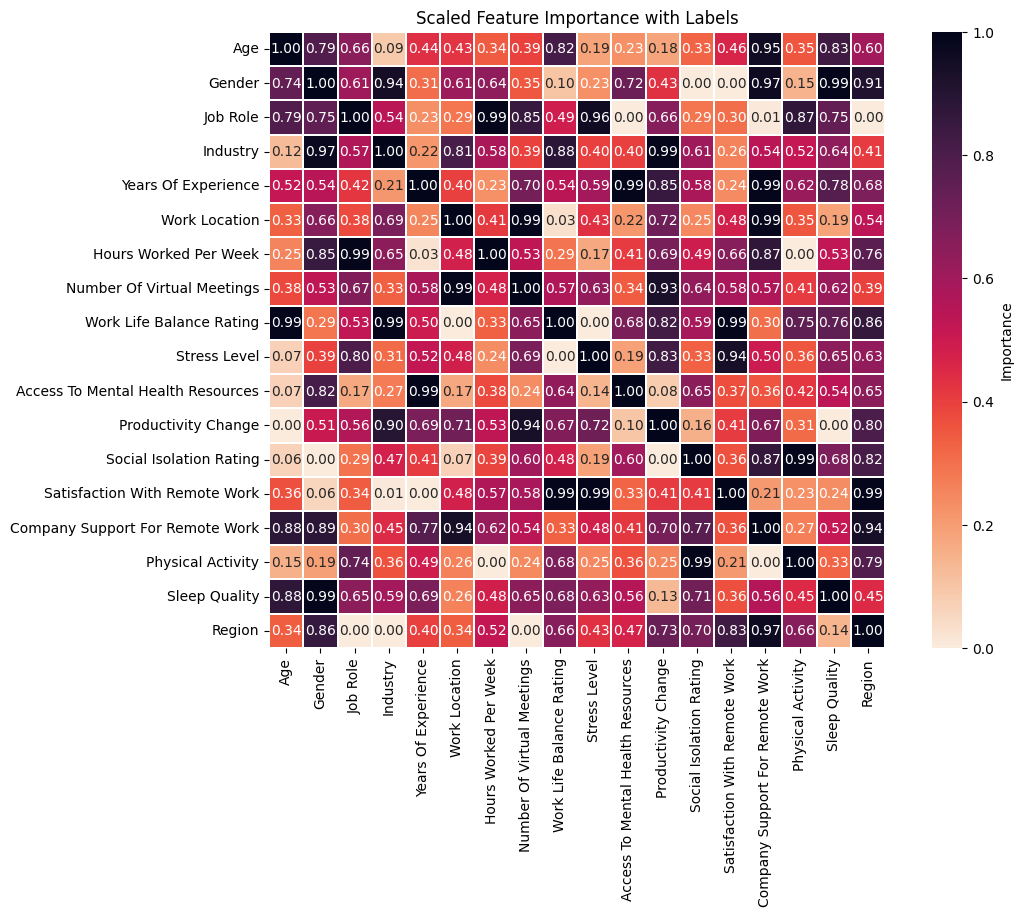
Converting categorical variables into a numerical format appropriate for machine learning models was an essential aspect of our preparation. We found category categories that reflect important aspects of the remote work experience and mental health setting, including "Gender," "Job\_Role," "Industry," and others. We were able to preserve the data in these variables in a format that our models could use by converting these categorical columns into numerical codes using an ordinal encoder. To make it easier to understand, we also saved the mappings of these categories to their encoded values. This allowed us to keep track of the changes made to each categorical variable.

Through these thorough data preparation methods, we guaranteed that our dataset was well-organized, clean, and ready for in-depth analysis. This careful preparation gave us the confidence to investigate the effects of remote work on mental health outcomes, laying a solid basis for the modeling and analysis phases.There were also a lot of necessary steps taken to enhance the usefulness and soundness of the dataset for examining the association between remote work and mental health before proceeding with the analysis. The next thing we did was to remove any fields that were not relevant, such as the “Employee\_ID” column. However, because its relevance to the aim of the study was not quite there, we were allowed to remove it and focus on the aspects which would help us in understanding the core of our research, which was concerned with mental health and occupation in a remote context during this stage. During the course of the project, our focus was on our model’s accuracy to performance ratio and the insights that were based on the performance metrics were reliable. Variables such as max\_depth, and cross validation techniques were applied in order to optimize the model so that it was neither too complex nor too simplistic and principally provided reliable outcomes.

We implemented Spearman’s Rank Correlation to assess the influence each feature had over the others. From the first result, it was clear that the feature importance was very low.

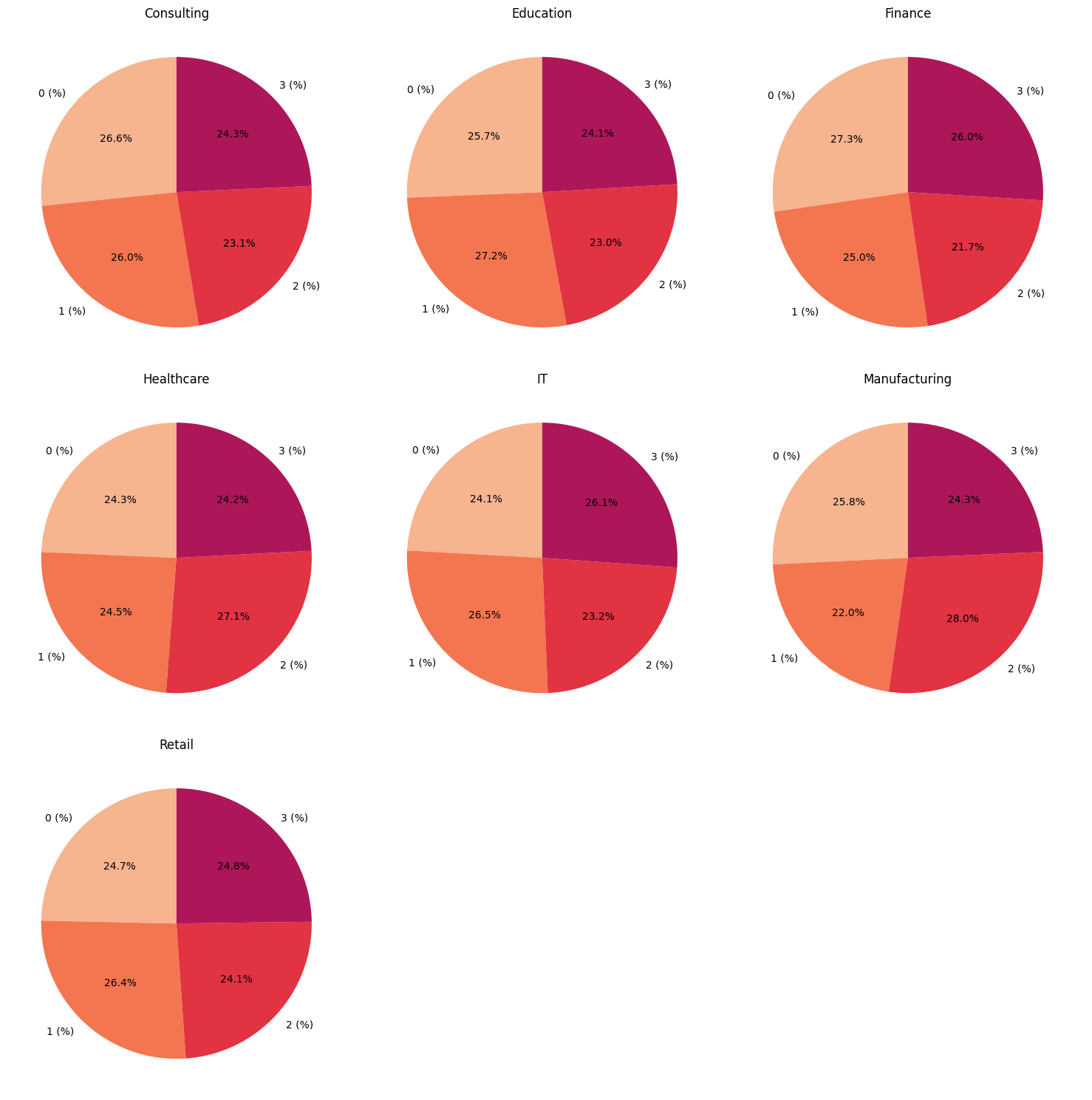


To fix this, we removed the self-referencing features from consideration when ranking correlation. Then, the rest of the features were scaled to be between 0 and 0.99. The following shows the results of this scaling.



The graph above shows the correlation between sets of features. The colors indicate positive and negative correlations, with pale orange representing low correlations and navy blue to almost black representing high positive correlations. For example, on the X-axis, we can examine the age feature. Age has a high correlation with sleep quality, as shown by its dark blue color. This suggests that as age fluctuates, sleep quality fluctuates inversely. The graph helps visualize the inverse relationship between these two features.

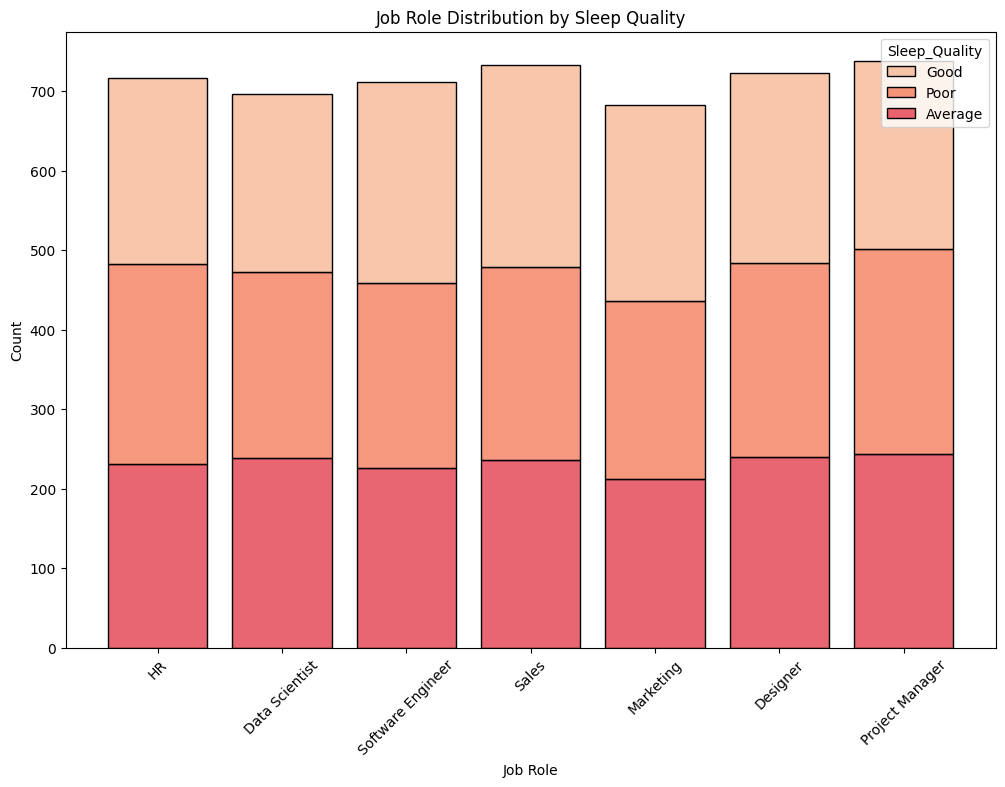
When looking at the relationship between gender and industry, for example, it appears to have a relatively high importance of 0.94 - 0.97. However, when looking at the data, the features are very evenly split. This is highly unlikely to occur naturally and leads us to consider the possibility that the dataset we are using may be synthetically generated.



This is just an example of how negligible the effects each feature has on the others. While we may doubt the validity of the data, we move forward with analysis and shift our expectations to a more hypothetical evaluation as we proceed.

An interesting correlation found is between years of experience and access to mental health resources. These two features exhibit a negative correlation, similar to the relationship between sleep quality and age. This could suggest a phenomenon where, as people work at a specific company for a longer period, they are less likely to use mental health resources. However, further research would be needed to confirm this.

One of the more intriguing dynamics found was between sleep quality and job role. It was somewhat surprising, as it was assumed that these two features would show a more pronounced dichotomy compared to other sets of features but they are not as inversely correlated in this dataset as originally thought. Notable positive correlations include the relationship between the region's features and sleep quality, job role and company support for remote work, and, finally, physical activity and sleep quality.



Some occupations might have a greater share of "Poor" or "Average" quality sleep compared to other occupations. Suppose software engineers have a higher percentage of "Poor" sleepers as compared to other job positions, such as HRs and Project Managers; this may indicate that more demanding or irregular hours of work affect sleep quality a lot more.

On the other hand, higher "Good" in Project Managers or Marketing kinds of roles may suggest that these jobs allow for better work-life balance, thus positively impacting sleep.

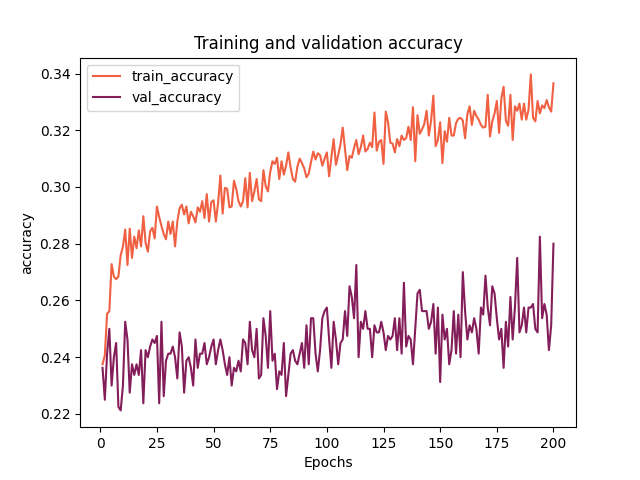
## Modeling

### Decision Tree

The first model we created was the decision tree model. First, the categorical values were encoded using one hot encoding. Then, We used the GridSearchCV function to optimize the tree using the best possible parameters that would have the highest accuracy score. It was determined that the splitting criterion be based on gini impurity, and that the tree would have a max depth of 9. This gave us a best accuracy score of 0.0857. 

As visualized in the tree above, this is a very complex and difficult to classify dataset. It’s difficult to identify any patterns or draw any conclusions. After using spearman’s rank correlation, we discovered that many of the features that influence each other, don’t have a very significant impact. This is reflected in the results of our decision tree. Our testing accuracy score is consistently around 0.13.

### Neural Network

Because the decision tree model did not yield any substantial results, we decided to see if a neural network would be more successful at training on this dataset. We started with four hidden layers, the first with 128 nodes and the following three with 64 nodes. The output layer has four nodes representing our mental health condition target variable (anxiety, depression, burnout, or n/a). We found this combination of nodes, along with using the ReLU activation function for the internal hidden layers, sigmoid function for the first layer, and softmax function for the output layer yielded the highest testing accuracy score of 26.7.

Plotting our accuracy in both training and validation, we can see that while the neutral network is a marked improvement over the decision tree model, it is still having trouble with performing on new data after it has been trained.

## Evaluation

During the evaluation stage, we thought about how our research might be applied to establish a more wholesome remote workplace. Although we are not implementing the model in a real-world environment, our findings provide insightful advice for businesses wishing to assist remote workers' mental health.

Designing policies for remote work to better suit the demands of employees is one noteworthy aspect. According to our data, job role, work hours, and company support levels can all have an impact on mental health outcomes. Businesses could use this information to develop responsive policies that take into account various jobs and preferences. For example, jobs that require a lot of teamwork can benefit from a hybrid approach that permits face-to-face contacts to preserve social ties. However, more independent positions could continue to operate entirely remotely in the interim, with additional virtual support systems in place to reduce feelings of isolation.

Using our findings to track and monitor trends in mental health over time is another relevant application. Organizations could utilize our model as a tool to spot new trends in employee mental health by frequently updating the data. By serving as an early warning system for mental health hazards, this continuous monitoring may assist businesses in making proactive adjustments to their support networks. Companies can monitor the effects of any changes they make by tracking trends over time, which enables them to adjust policies in light of actual results.

All things considered, employing these insights helps businesses make data-driven decisions that support resilience and employee well-being. Businesses may create a productive and encouraging work environment and strengthen and improve the health of their remote workforce by addressing certain mental health factors that influence remote employment. There are several advantages for both employees and businesses when there is a mentally healthy workforce. Employees are more likely to be motivated, engaged, and productive in their jobs when they feel supported and have good mental health.

# References

Ali, Waqar. “Remote Work & Mental Health 🌍🧠.” *Kaggle*, 22 Sept. 2024, [www.kaggle.com/datasets/waqi786/remote-work-and-mental-health/data](http://www.kaggle.com/datasets/waqi786/remote-work-and-mental-health/data).

“Remote Work and Mental Health: Maintaining Your Well-Being While Working from Home.” *Maryville University Online*, 27 Sept. 2023, [online.maryville.edu/blog/remote-work-mental-health/](http://online.maryville.edu/blog/remote-work-mental-health/).

Khan, Saniya Ahmad. “Why Office Work Boosts Mental Health over Remote Work.” *YourStory*, [yourstory.com/2024/11/mental-health-benefits-office-work](http://yourstory.com/2024/11/mental-health-benefits-office-work) . Accessed 8 Nov. 2024.

Abramson, Ashley. “The One Room You Should Avoid from 9 to 5.” *Apartment Therapy*, Apartment Therapy, 4 Oct. 2023, [www.apartmenttherapy.com/working-from-home-bedroom-advice-36828053](http://www.apartmenttherapy.com/working-from-home-bedroom-advice-36828053).

*10 Ways Workplace Culture Differs around the World - Glassdoor Blog*, [www.glassdoor.com/blog/workplace-differences-around-the-world/](http://www.glassdoor.com/blog/workplace-differences-around-the-world/) . Accessed 10 Nov. 2024.