Mindfulness, Gender & Online Teaching: An Exploratory Data Analysis & Predictive Modeling Project

2023-12-04

Executive Summary

This section describes the dataset and variables, and summarizes the goal of the project and key steps that were performed.

Overview of the Project

This project aims to explore the dynamics of mindfulness teaching, with a focus on gender and online teaching modalities. Utilizing a comprehensive dataset obtained through a nationwide online survey of `768` participants, the project report employs quantitative statistical analysis using `R`. Key steps include data tidying, exploratory data analysis, sentiment analysis, and predictive modeling using tree-based algorithms.

Key Findings

The exploratory data analysis revealed that mindfulness teachers are predominantly **middle-aged females**.

- **Exploratory data analysis** of gender and online teaching preferences showed significant gender-based differences in online teaching practices.
- **Statistical tests** confirmed the importance of gender in online teaching engagement, particularly in the preference for online teaching and the number of students taught online.
- **Advanced machine learning models**, including Random Forest and XGBoost, were employed to further investigate these relationships, indicating moderate accuracy in predicting online teaching based on gender.
- **Sentiment analysis** of open-ended survey responses highlighted a predominantly positive outlook among mindfulness teachers regarding the future of the sector.

The analysis provided valuable insights into the challenges and opportunities perceived by mindfulness teachers.

Conclusion and Implications

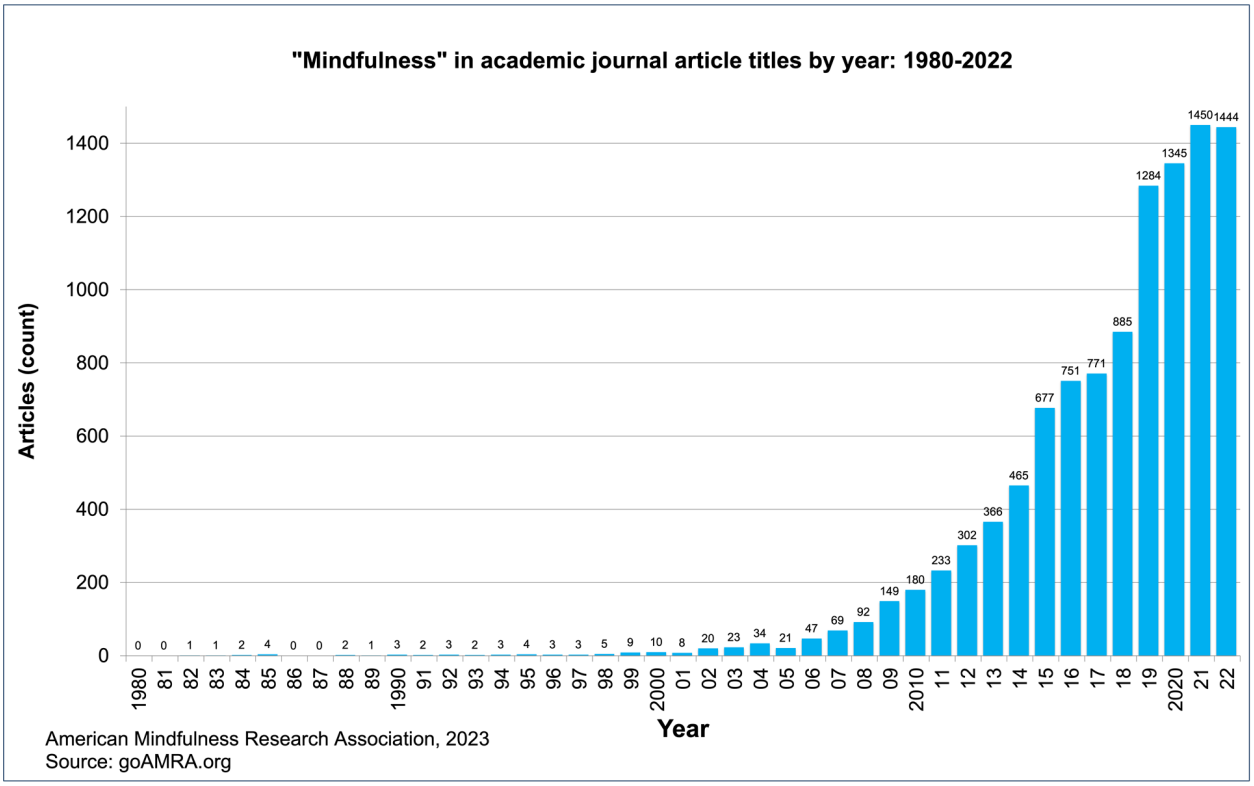
The project successfully demonstrates the significant role of gender in online mindfulness teaching and offers a nuanced understanding of the current state and future prospects of the mindfulness sector.

The findings have implications for how mindfulness is taught and received in online settings, emphasizing the need for further research and targeted strategies to address gender disparities and optimize teaching methodologies for diverse audiences.

The study's limitations, primarily its reliance on self-reported data and the need for broader demographic representation, pave the way for future research to build upon these findings and explore other influential factors in mindfulness teaching.

1. Introduction

Mindfulness - a mind-body practice to enhance awareness of the present moment - has become a global phenomenon. A recent survey suggests 15% of adults in Britain - almost 8 million people - have learnt how to practice mindfulness meditation (Simonsson *et al.*, 2021). Yet, despite the remarkable rise and popularity of mindfulness, and the many studies of its therapeutic effectiveness (see figure 1), studies of the movement's spread and significance are few.



For the purpose of this project report, we focus on the following specific research questions:

Research Questions

- Who are mindfulness teachers?
- What factors influence online teaching of mindfulness?
- How does gender relate to online mindfulness teaching?
- What are mindfulness teachers' outlooks on the future?

2. Methods/Analysis

This is a methods/analysis section that explains the process and techniques used, including data cleaning, data exploration and visualization, insights gained, and the modeling approach.

We adopted a **mixed-methods** design using **quantitative** and **qualitative** methods.

Our data collection included a nationwide online survey with **768** participants.

This project report focuses on the quantitative analysis of the survey data. We use `R` and specifically packages within the `tidyverse` and `tidymodels` ecosystem, due to their ease of use, and accessibility. We followed the guidance provided in *Tidy Modeling with R* by Kuhn and Silge (2022).

To address our research questions, we engage in data tidying, preprocessing, exploratory data analysis, and machine learning tree-based algorithms.

2.1 Data Tidying & Exploratory Data Analysis

We used packages from base `R` and the `tidyverse` ecosystem to tidy the data. We renamed variables, replaced empty strings with `NA`, and converted variables to their appropriate data types.

We performed Exploratory Data Analysis using descriptive statistics to quantitatively measure patterns in the survey data.

2.2 Tree-Based Modeling

Description of the Approach

The tree-based modeling approach in this project utilized two primary algorithms: **Random Forest** and **XGBoost** (eXtreme Gradient Boosting).

- **Random Forest**, known for its robustness, creates a 'forest' of decision trees, outputting the mode of the classes from individual trees for predictions.
- **XGBoost**, an efficient implementation of gradient boosting machines, builds sequential trees where each corrects the predecessor's errors, enhancing prediction accuracy.

The implementation involved the `ranger` and `xgboost` engines in R, respectively, and was facilitated by the `tidymodels` framework, integrating packages like `parsnip`, `recipes`, `rsample`, and `yardstick` for a comprehensive modeling workflow.

Justification for the Choice of Algorithms

The selection of Random Forest and XGBoost was motivated by their ability to effectively manage complex, non-linear relationships within the data.

- **Random Forest** reduces overfitting through averaging multiple decision trees, making it ideal for generalizing findings across populations.
- **XGBoost** is favored for its quick processing and accuracy in classification tasks, sequentially improving trees for more precise modeling. Both algorithms excel in handling missing values and providing predictor importance scores, offering valuable insights, especially crucial for educational research where clear, actionable results are sought.

Implementation and Evaluation

1. The implementation process started with data preprocessing, where categorical variables were converted into dummy variables, and numeric ones were normalized, a critical step for the effectiveness of tree-based models.
2. The models were then specified with parameters apt for the dataset: tree numbers and variable splits for Random Forest, and boosting rounds for XGBoost.
3. Cross-validation, especially stratified sampling based on gender, was employed to ensure robustness and minimize bias.
4. The models' performance was evaluated using metrics like accuracy, ROC AUC, and log loss, providing a well-rounded understanding of their predictive capabilities in the context of mindfulness teaching practices.

This thorough evaluation confirmed the models' applicability and reliability in real-world educational settings.

2.3 Sentiment Analysis

The process detailed in the provided text involves a sentiment analysis of survey text data. Here is a summary of the steps followed:

Data Preparation: The raw text data from a survey was converted into a data frame format, which is necessary for processing in R, a programming language often used for data analysis.

Tokenization: The text data was tokenized, meaning it was broken down into individual words or "tokens". This is a common first step in text analysis to enable word-level processing.

Sentiment Scoring: Each tokenized word was then matched with a sentiment score using the `bing` lexicon, which classifies words as either positive or negative.

Sentiment Summary: After scoring, the data was summarized to count the frequency of each sentiment category. This summary was then used to calculate the net sentiment by subtracting the number of negative sentiment words from positive sentiment words.

Visualization: The top words by sentiment frequency were visualized in a plot, showing the most frequent words categorized by sentiment. This helps in understanding the overall sentiment of the text data by highlighting the most common positive and negative words.

3. Results

This section answers our research questions by presenting exploratory data analysis, the modeling results, and discusses the model performance.

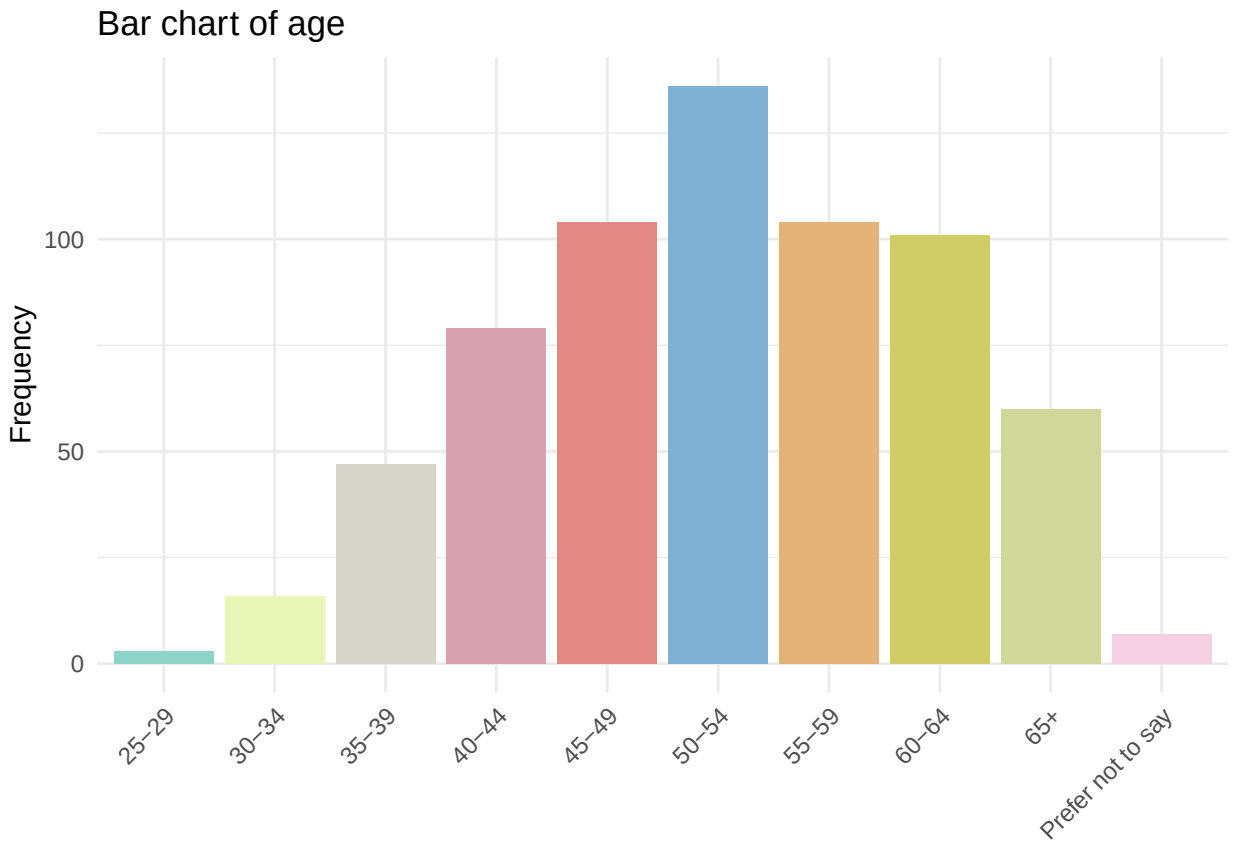
3.1 Who are mindfulness teachers?

Gender & Age

Mindfulness teachers tend to be female and middle-aged.

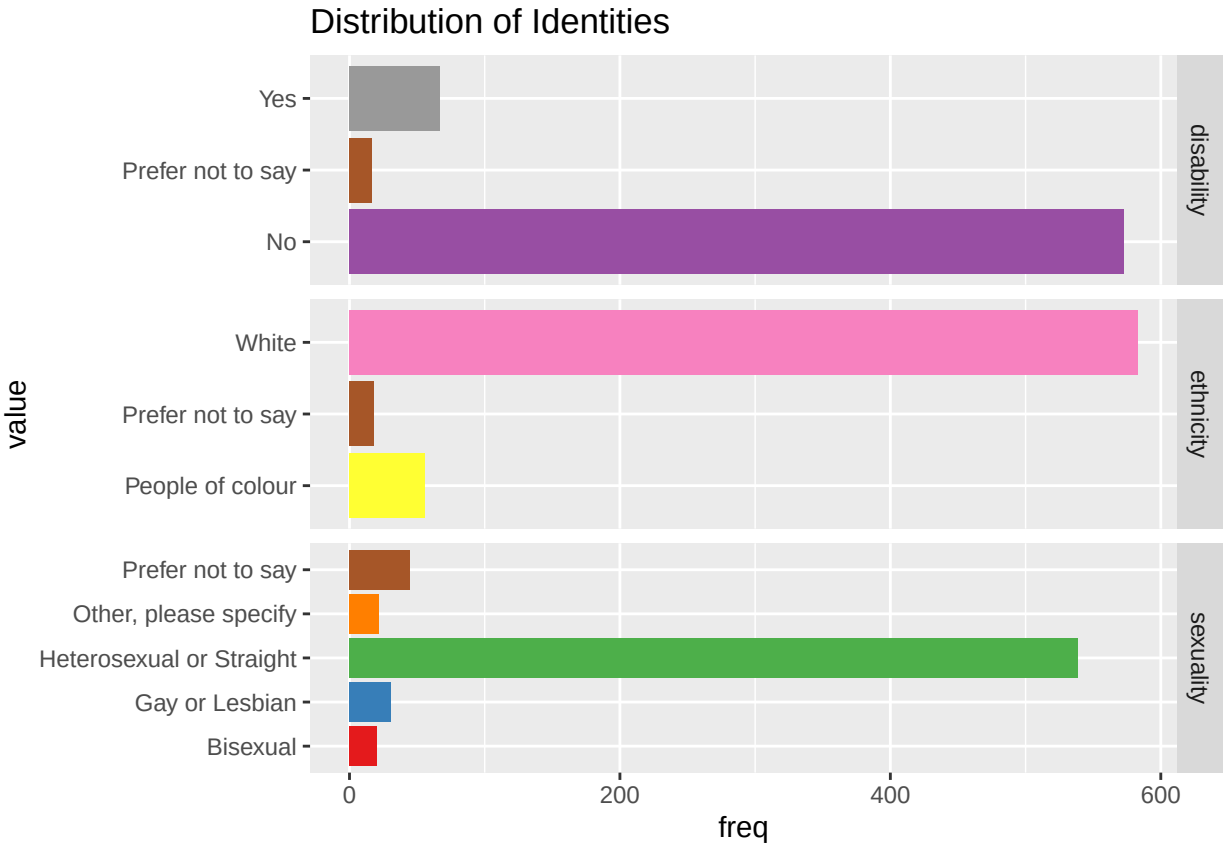
Table 1: Gender Distribution

Gender	Count
Female	461
Male	185
Other, please specify	5
Prefer not to say	6



Disability, Ethnicity & Sexuality

Our participants tended to be white, heterosexual or straight, and non-disabled.



3.2 What factors influence online teaching of mindfulness?

We narrowed down a larger dataset named `survey` to focus on four key aspects: age, gender, online teaching experience, and the year each respondent started teaching mindfulness. This selection, done for analytical clarity, is crucial for targeting relevant data.

The code then addresses a common issue in data processing: empty entries. It transforms these blanks into a recognizable format (labeled `NA` for `Not Available`) in `R`, which is a standard approach for dealing with missing information.

The next significant step is the filtration of the dataset, achieved using a functionality from the `dplyr` package, known for its data manipulation capabilities. This process excludes any records with missing information in the key areas, ensuring the final dataset is comprehensive and reliable for analysis.

Data Preparation for Modeling

We prepared dataset for tree-based machine learning modeling, with a specific focus on gender representation. We employed a stratified splitting approach, where the dataset, `survey_filtered`, was divided into training and testing sets, ensuring both sets have a representative balance of genders. This is particularly important given that over 70% of the dataset's participants are female, raising concerns about potential bias in a randomly split dataset.

Stratification, achieved through the `initial_split` function with a `strata = "gender"` argument, ensures that both training (80% of the data) and testing (20%) sets mirror the overall gender distribution of the original dataset. Such a balanced split is crucial for developing unbiased and accurate tree-based models, because prevents the model from being skewed towards the majority class, thereby enhancing the model's generalizability and fairness.

Random Forest Classification Model

We conducted data preprocessing, model specification, and evaluation for a **Random Forest** classification model.

We started with data preprocessing using the `recipes` package, where categorical variables are converted into dummy variables, and numeric variables are normalized. The Random Forest model, specified using the `rand_forest` function from the `parsnip` package and implemented via the `ranger` engine, is designed for classification tasks.

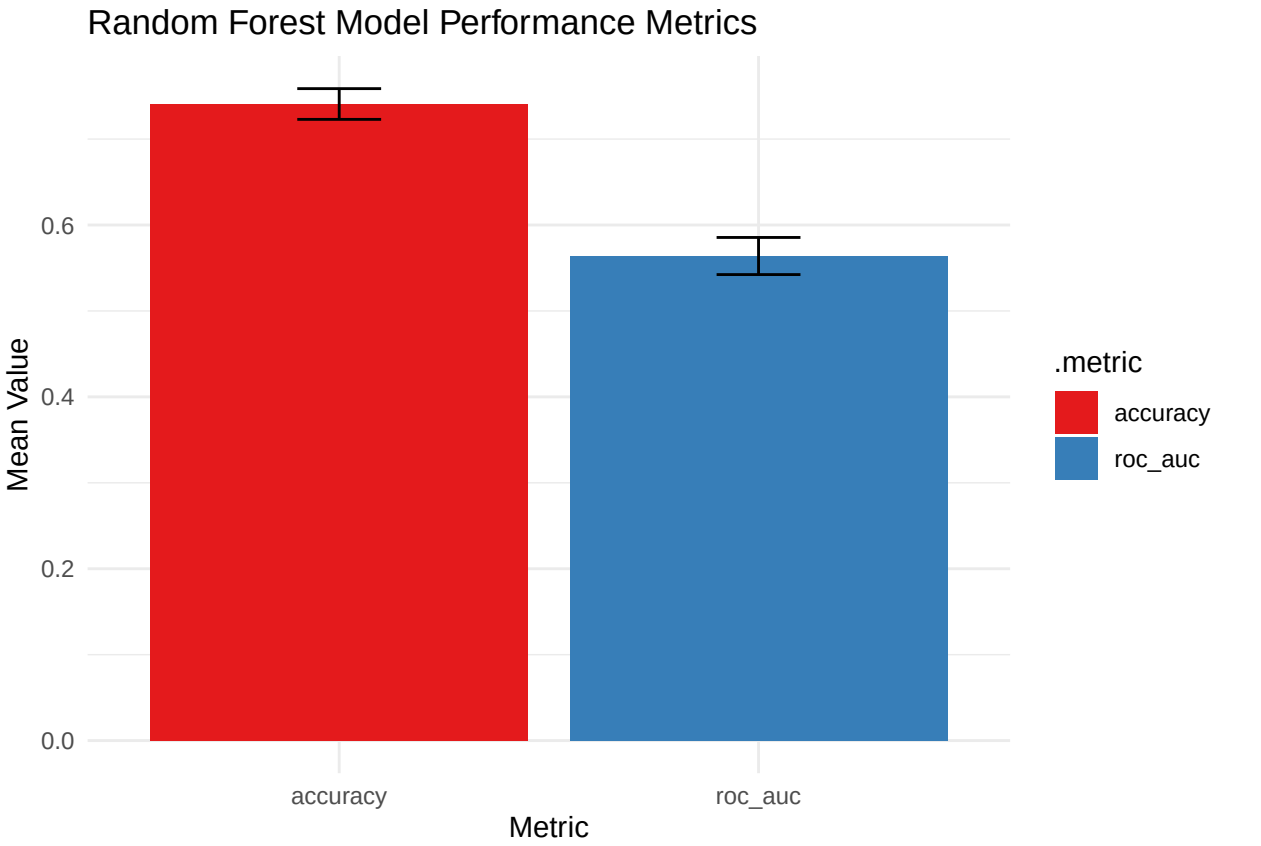
Cross-validation is set up with `vfold_cv` from the `rsample` package, ensuring a balanced representation of genders in each fold. The workflow combines the preprocessing steps and the model, which is then fit and evaluated on the training data with a focus on robustness and error handling.

This comprehensive approach, integrating specific `R` functions and packages, ensures the creation of an effective and fair classification model, considering the importance of gender balance in the training process.

We created a bar plot visualizing the performance metrics of a Random Forest model using the `ggplot2` package in R. The metrics plotted are accuracy and **ROC AUC** (Area Under the Receiver Operating Characteristic Curve), two common measures for evaluating classification models.

The plot shows two bars, with the height of each bar representing the mean value of the respective metric. Error bars are included to represent the standard error around the mean, giving a sense of the variability or uncertainty in the metric estimates. The graph is styled with a minimal theme and uses color coding to distinguish between the metrics.

The model exhibits a moderate level of accuracy and ROC AUC, with the values of both metrics being close to or above `0.5`, suggesting that the model performs better than random chance. The error bars indicate some variation in the model's performance across different folds or iterations of the cross-validation process.



We constructed a summary table for the performance metrics of a Random Forest model. The `kable` function

from the 'knitr' package is used to format the data frame into a `LaTeX` table, which is further styled with `kable_styling` to have striped rows and a fixed position.

The resulting table, as shown in the code's output, concisely displays the Random Forest model's classification performance, indicating that the model has a **relatively high mean accuracy** and a **lower, yet above chance-level**, mean ROC AUC, with small standard errors suggesting precise estimates from the cross-validation process.

Table 2: Random Forest Model Performance

Metric	Estimator	Mean	N	Standard.Error
accuracy	binary	0.7408159	5	0.0179004
roc_auc	binary	0.5639125	5	0.0216070

3.3 How does gender relate to online mindfulness teaching?

I tested whether there is a relationship between gender, technology use, online teaching, and the number of students taught online.

> **p-value** or **probability value** helps us find out if results are statistically significant. A common threshold is 0.05. If the p-value is below this, we can say the results are likely to be significant.

Our results show that gender does not seem to be significantly related to use of technology. But gender does influence both whether a person teaches mindfulness online and the number of students taught online.

Table 3: Results for Pearson's Chi-Squared Tests

Tested.Relationship	X.squared..Test.statistic.	Degrees.of.Freedom..df.	p.value	Significance
Gender and Technology	1.7838	3	0.618500	Not Significant
Gender and Online Teaching	14.2040	3	0.002640	Significant
Gender and Online N	35.3230	18	0.008605	Significant

The Chi-Squared Test for independence only tells us if there is a significant difference. It does not tell us the nature of that difference. To find out what difference there is, we need to look at the descriptive statistics.

When we looked at whether different gender categories have a preference for online teaching, we found some interesting patterns.

- Females were **less inclined** to teach online, with only about 23% saying 'Yes' compared to 77% who said 'No'.
- Males, on the other hand, showed a slightly **higher inclination** towards online teaching, with 32% saying 'Yes'.
- The category "Other, please specify" had the highest inclination towards online teaching, with a massive 80% saying 'Yes'.

Table 4: Gender and Online Teaching Preference

Gender	Yes	No	Yes.Percentage	No.Percentage
Female	105	356	22.78	77.22
Male	60	125	32.43	67.57
Other, please specify	4	1	80.00	20.00
Prefer not to say	2	4	33.33	66.67

Table 5: Gender and Number of Students Taught Online

Gender	Zero	X1.5	X6.10	X11.20	X21.50	X51.100	More.than.100	Zero.Percentage	X1.5.Perc
Female	303	53	21	10	21	19	34	65.73	
Male	106	11	10	14	11	11	22	57.30	
Other, please specify	1	0	1	1	1	0	1	20.00	
Prefer not to say	4	0	0	1	0	0	1	66.67	

When I explored how many students each gender category has taught online, I found some interesting differences:

- Females predominantly (about 66%) have taught 0 students online. However, there's a small proportion (around 7%) who have taught more than 100 students online.
- Males show a pattern where a majority (57%) have taught 0 students online, but interestingly, close to 12% have taught more than 100 students online.
- For the "Other, please specify" category, the data is spread out, suggesting variability in teaching online, with 20% having taught more than 100 students online.

Our statistical tests confirmed that there are significant differences in the preferences and patterns of online teaching across gender categories. Specifically, while gender doesn't seem to significantly influence the use of technology, it does play a role in the preference for online teaching and the number of students taught online.

To put it simply, gender seems to influence whether someone engages in online teaching and how many students they've taught online, but it doesn't necessarily dictate their use of technology in general.

We conducted data preprocessing steps for the `survey` dataset focused on the variables `gender` and `online_teaching`, which are of particular interest due to previously identified significant differences between genders in online teaching practices, as determined by **Chi-Squared** tests.

The code first cleans the data by converting empty strings to `NA` (missing values) in the `gender` column and temporarily converts `online_teaching` to a character type to handle missing values similarly.

After dropping all rows with missing values in these two key columns, `online_teaching` is converted back to a factor, which is suitable for categorical data analysis in `R`.

The dataset is then filtered to include only the relevant variables, ensuring a clean and focused dataset for subsequent modeling.

This focused preprocessing is justified as it prepares the dataset for modeling that will further investigate the relationship between gender and online teaching, based on prior significant findings.

Random Forest and XGBoost

We split the dataset `survey` into training and test sets, with stratification on the `gender` variable to ensure representative samples, and sets a seed for reproducibility.

Two machine learning models, **Random Forest** and **XGBoost**, are prepared for a classification task to predict `online_teaching` based on `gender`.

Both models are known for handling categorical data and are suitable for classification problems where the relationships between variables may be complex and non-linear.

> **Random Forest** is a robust and widely-used ensemble learning method that can provide good performance with less risk of overfitting due to its bagging approach.
>
> **XGBoost** is another powerful ensemble technique that uses a gradient boosting framework and is often praised for its performance in classification tasks, handling various types of data, and its

speed and efficiency.

We create workflows for each model, fit them to the training data, and evaluate predictions against the test data.

This approach allows for a comparative assessment of two leading algorithms in predicting how gender influences online teaching engagement, potentially offering insights into the structure and strength of this relationship.

Table 6: Model Performance Metrics

Model	Metric	Estimate
Random Forest	Accuracy	0.7010
Random Forest	Kappa	0.0313
XGBoost	Accuracy	0.7010
XGBoost	Kappa	0.0313

Based on the results provided:

The **Random Forest** model predicted correctly that 100 observations would be 'No' and correctly predicted 1 observation as 'Yes', but it incorrectly predicted that 43 observations would be 'No' when they were actually 'Yes'. This model did not predict any 'Yes' observations incorrectly as 'No'.

The **XGBoost** model produced the same confusion matrix results as the Random Forest, with 100 true negatives, 43 false negatives, and 1 true positive. There is no mention of false positives, which could be a missing value in the output provided.

Both models achieved the same accuracy of `0.701` and a Kappa statistic of `0.0313`. The accuracy indicates that approximately 70.1% of predictions were correct, while the Kappa statistic, which accounts for the accuracy that could occur by chance, suggests that the agreement between predictions and actuals is only slightly better than chance.

Given that both models show identical performance metrics, it might indicate that they are either reaching their performance limit given the data or that there could be an error in the model configuration or the data itself. The identical confusion matrices further suggest that both models are performing similarly on this particular dataset.

XGBoost Model Performance

We evaluated the XGBoost model's performance over ten iterations using a metric called **log loss**, which measures the accuracy of a classifier by penalizing false classifications.

A data frame, `xgb_performance`, is defined to hold the iteration numbers and their corresponding training log loss values. These values show a trend of **decreasing log loss over the iterations**, indicating the model is improving and becoming more accurate as it learns.

The `kable` function from the `knitr` package is utilized to neatly format this data into a `LaTeX` table, which is styled to have a striped pattern and fixed position in the document.

We generated a line plot using `ggplot2`, which visually depicts the **decline in training log loss across iterations**, reinforcing the numerical data with a graphical trend that signifies the model's increasing predictive performance. The minimal theme of the plot ensures focus on the data points and the trend line, providing a clear visual representation of the model's training progress.

Table 7: XGBoost Training Log Loss

iter	training_logloss
1	0.6241134
2	0.5892745
3	0.5706919
4	0.5608741
5	0.5554729
6	0.5526112
7	0.5511099
8	0.5503296
9	0.5499500
10	0.5497339

3.4 What are mindfulness teachers' outlooks on the future?

We asked our participants "What do you see as the main challenges and/or opportunities for the future of the mindfulness sector?"

This is an open-ended question that allows for a wide range of responses. Participants were asked to provide their responses in a text box, which provided character data.

To analyse the data, we followed the procedure described in the Methods/Analysis section.

Sentiment Summary

A summary of these sentiment scores was created to determine the frequency of each sentiment category.

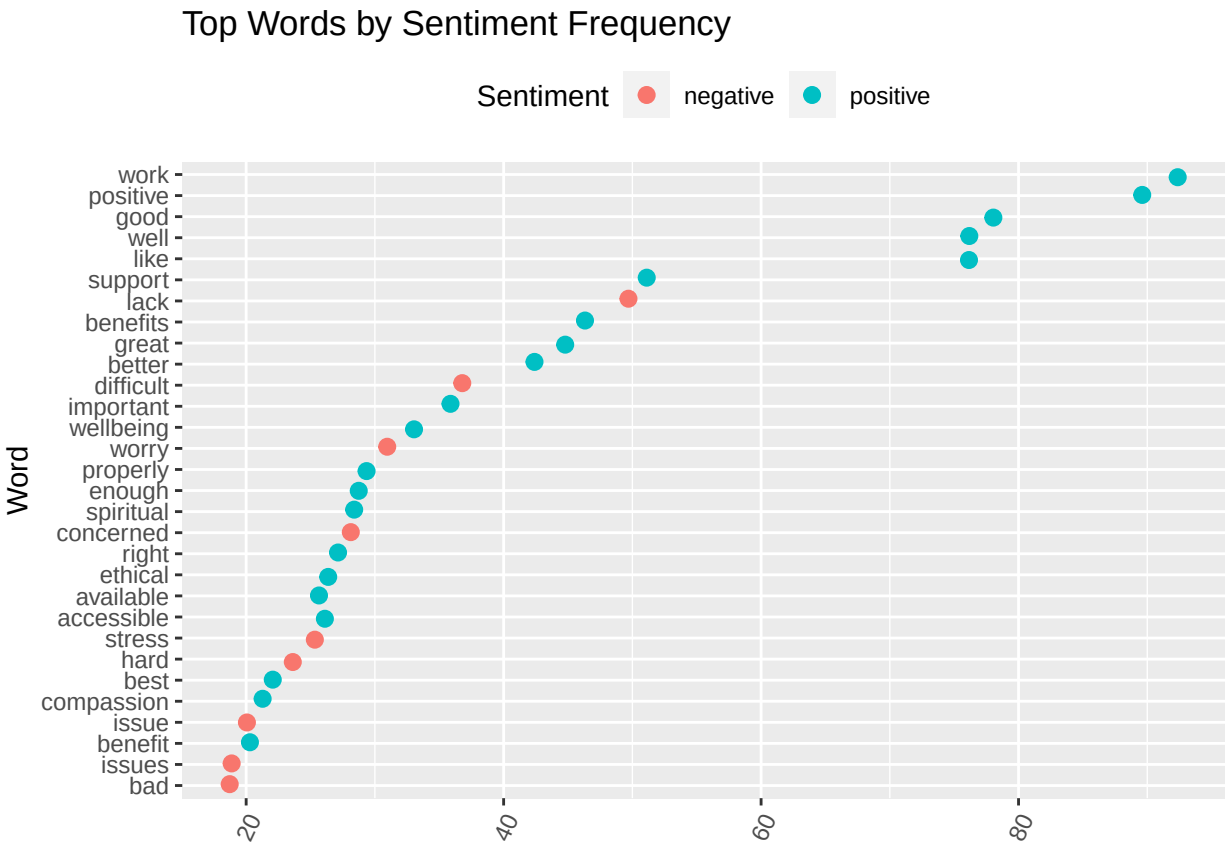
Table 8: Net Sentiment Summary

negative	positive	net_sentiment
1119	1927	808

Net Sentiment Value: This was followed by calculating net sentiment by offsetting negative words against positive ones. A net sentiment value of 808 indicates a predominantly positive sentiment in the text data analyzed.

Top Word Frequencies: Finally, the most frequent words in each sentiment category were visualized in a plot, aiding in comprehending the overall sentiment of the survey data.

The visualization provided shows a scatter plot with words on the y-axis and their frequency on the x-axis, colored by sentiment. Words like "work", "good", "like", and "well" appear most frequently with a positive sentiment, while negative words are less frequent. The plot is flipped for better readability.



The process is justified as it provides a structured approach to quantify and visualize sentiments expressed in text data, allowing for objective analysis of qualitative data.

4. Conclusion

This section gives a brief summary of the report, its potential impact, its limitations, and future work.

Overview

The **Mindfulness, Gender & Online Teaching** project provides a comprehensive exploration into the dynamics of mindfulness teaching, particularly examining the intersections of gender and online teaching modalities. Through a robust analysis of data from a nationwide survey, key insights were uncovered about the demographic characteristics of mindfulness teachers, their preferences, and practices in online teaching.

Key Findings

> **A gendered landscape in mindfulness teaching**, with females being the predominant demographic. There's a notable variation in online teaching preferences and the number of students taught, influenced significantly by gender.

These insights are crucial in understanding the current trends and future directions in the mindfulness sector.

> **The advanced statistical and machine learning models**, including Random Forest and XGBoost models, have provided valuable predictions about online teaching practices based on gender, contributing to a deeper understanding of the sector's dynamics.

However, the project has its limitations.

Limitations

The reliance on **self-reported data** may introduce biases, and the sample may not comprehensively represent the entire mindfulness teaching community. This limitation underscores the need for broader and more inclusive data collection in future studies.

Future Directions

Looking forward, this project lays a foundation for further research in the field.

Future studies should aim to include a more diverse sample to capture a wider range of experiences and perspectives.

There's also a need to explore additional factors that might influence mindfulness teaching practices, such as cultural backgrounds, socioeconomic status, and geographic locations. The integration of these aspects will offer a more holistic understanding of the mindfulness sector, facilitating the development of more inclusive and effective teaching strategies.

This research has the potential to significantly impact the mindfulness community, guiding mindful teaching practices to be more attuned to the diverse needs of learners in an increasingly digital world.

5. References

This section lists sources for datasets and/or other resources used.

Kuhn, M., & Silge, J. (2022). *Tidy Modeling with R.* O'Reilly Media, Inc. ISBN: 9781492096481.

Simonsson, O., Fisher, S., & Martin, M. (2021). Awareness and Experience of Mindfulness in Britain. *Sociological Research Online, 26* (4), 833-852.