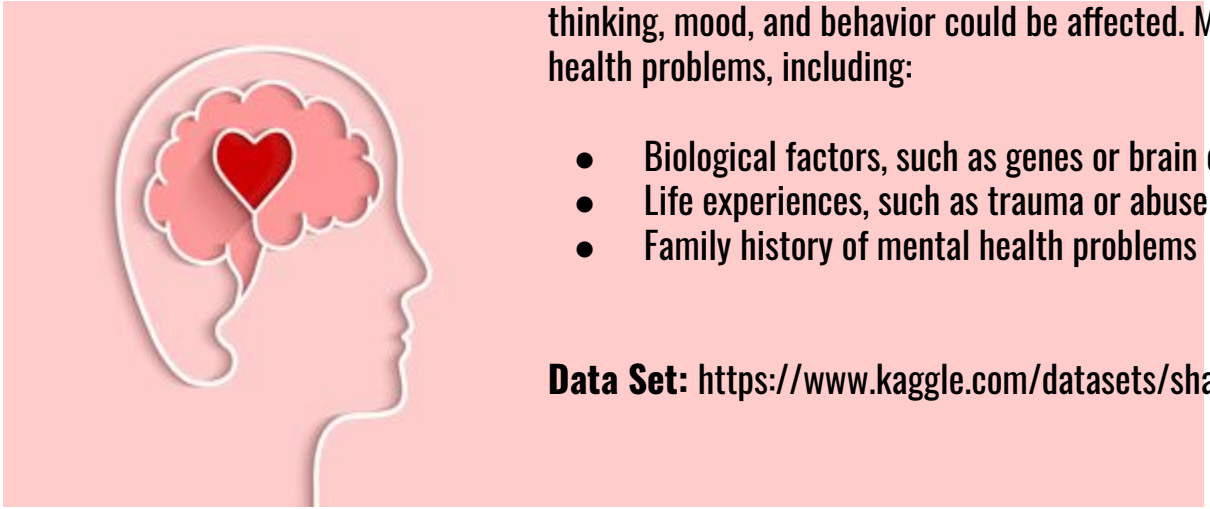
Mental Health in the Work Environment

By: Anh, Annabel, Kierra, Rafael, Robby

Today's Agenda

- Background
- Data Description
- Exploratory Data Analysis
- Research Questions and Analysis
- Challenges
- Conclusion

What is Mental Health?



Mental health includes our emotional, psychological, and social well-being. It affects how we think, feel, and act. It also helps determine how we handle stress, relate to others, and make choices. Mental health is important at every stage of life, from childhood and adolescence through adulthood.

Over the course of your life, if you experience mental health problems, your thinking, mood, and behavior could be affected. Many factors contribute to mental health problems, including:

- Biological factors, such as genes or brain chemistry
- Life experiences, such as trauma or abuse
- Family history of mental health problems

Data Set: https://www.kaggle.com/datasets/shadabhussain/medical-treatment-dataset

Data Description

- Response: treatment Yes/No
- S.no: ID
- Timestamp: Year/Month/Day/Time
- Age
- Gender
- Country
- State
- Self_employed: Yes/No(or NA)
- Family_history: Family with history of mental illness? Yes/No
- Work_interfere: If the participant feels that mental illness interferes with work? Never/Rarely/Sometimes/Often
- No_employees: Number of employees at the participant's workplace? 1-5/6-25/26-100/100-500/500-1000/More than 1000
- Remote_work: If participant works remotely? Yes/No
- Tech_company: If participant's workplace is a tech company? Yes/No
- Benefits: Does employer provide mental health benefits? Yes/No/Don't Know
- Care_options: Does participant know the options for mental health care your employer provides? Yes/No/Not Sure
- Wellness_program: Has employer ever discussed mental health as part of an employee wellness program? Yes/No/Don't know
- Seek_help: Does employer provide resources to learn more about mental health issues and how to seek help? Yes/No/Don't know
- Anonymity: Is participant anonymity protected if he/she chooses to take advantage of mental health or substance abuse treatment resources? Yes/No/Don't know
- Leave: How easy is it for participant to take medical leave for a mental health condition? Don't know/Very easy/Somewhat easy/Somewhat difficult/Very Difficult
- Mental_health_consequence: Does participant think that discussing a mental health issue with his/her employer would have negative consequences? Yes/No/Maybe
- Phys_health_consequence: Does participant think that discussing a physical health issue with his/her employer would have negative consequences? Yes/No/Maybe
- Coworkers: Would the participant be willing to discuss a mental health issue with his/her coworkers? Yes(or Some of Them)/No
- Supervisor: Would the participant be willing to discuss a mental health issue with his/her direct supervisor(s)? Yes(or Some of Them)/No
- Mental_health_interview: Would participant bring up a mental health issue with a potential employer in an interview? Yes/No/Maybe
- Phys_health_interview: Would participant bring up a physical health issue with a potential employer in an interview? Yes/No/Maybe
- Mental_vs_physical: Does the participant feel that employer takes mental health as seriously as physical health? Yes/No/Don't Know
- Obs_consequence: Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace? Yes/No
- Comments: additional comments of participant

Exploratory Data Analysis

Data summary	
Name	MH
Number of rows	1259
Number of columns	27
Column type frequency:	
character	25
numeric	1
POSIXct	1
Group variables	None

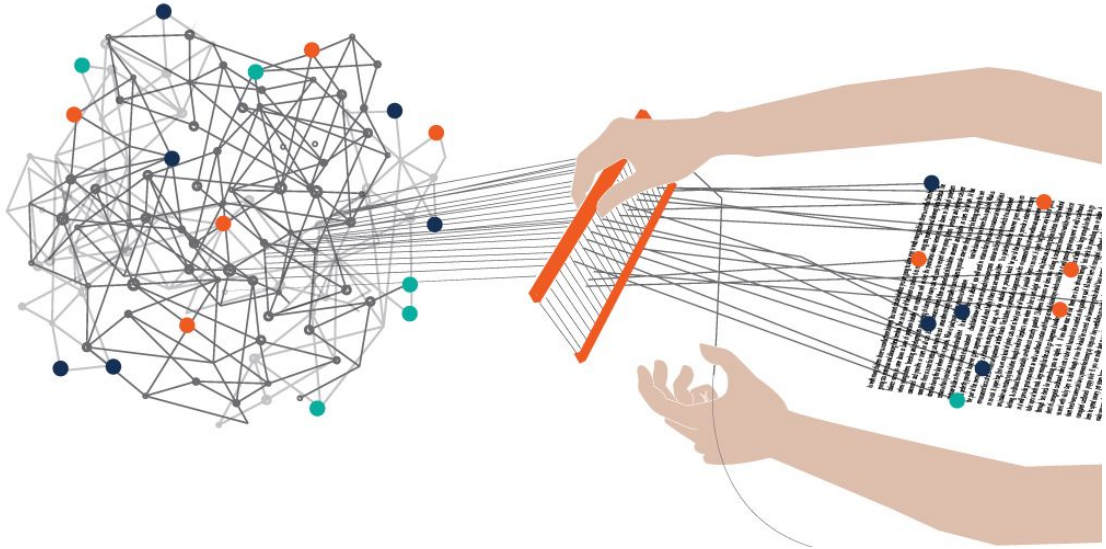
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Gender	0	1.00	1	46	0	44	0
Country	0	1.00	5	22	0	48	0
state	515	0.59	2	2	0	45	0
self_employed	18	0.99	2	3	0	2	0
family_history	0	1.00	2	3	0	2	0
treatment	0	1.00	2	3	0	2	0
work_interfere	264	0.79	5	9	0	4	0
no_employees	0	1.00	3	14	0	6	0
remote_work	0	1.00	2	3	0	2	0
tech_company	0	1.00	2	3	0	2	0
benefits	0	1.00	2	10	0	3	0
care_options	0	1.00	2	8	0	3	0
wellness_program	0	1.00	2	10	0	3	0
seek_help	0	1.00	2	10	0	3	0
anonymity	0	1.00	2	10	0	3	0
leave	0	1.00	9	18	0	5	0
mental_health_consequence	0	1.00	2	5	0	3	0
phys_health_consequence	0	1.00	2	5	0	3	0
coworkers	0	1.00	2	12	0	3	0
supervisor	0	1.00	2	12	0	3	0
mental_health_interview	0	1.00	2	5	0	3	0
phys_health_interview	0	1.00	2	5	0	3	0
mental_vs_physical	0	1.00	2	10	0	3	0
obs_consequence	0	1.00	2	3	0	2	0
comments	1096	0.13	1	3548	0	159	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Age	0	1	30.79	50.83	-1726	27	31	36	329	▁▁▁▁▇

Data Wrangling

- We only used data of participants from the US.
- Encoded categorical columns with 0/1, 0/1/2, 0/1/2/3...
- Deleted columns: S.no, Timestamp, Country, State
- Imputed missing values in self_employed and work_interfere with the highest frequency values("No" and "Sometimes")
- Imputed outliers in Age(Negative, hundred, or illegal working age values) with the column means
- We will save the comments for text mining.



Exploratory Data Analysis cont'd

```
-- Data Summary ------------------
                           Values
Name                       MH_US
Number of rows             751
Number of columns          23
_________________________
Column type frequency:
  factor                   22
  numeric                  1
_________________________
Group variables            None
```

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
Gender	0	1	FALSE	3	1: 562, 0: 185, 2: 4
self_employed	0	1	FALSE	2	0: 695, 1: 56
family_history	0	1	FALSE	2	0: 421, 1: 330
treatment	0	1	FALSE	2	Yes: 388, No: 363
work_interfere	0	1	FALSE	4	2: 433, 0: 125, 1: 111, 3: 82
no_employees	0	1	FALSE	6	6: 216, 3: 170, 2: 134, 4: 113
remote_work	0	1	FALSE	2	0: 513, 1: 238
tech_company	0	1	FALSE	2	1: 611, 0: 140
benefits	0	1	FALSE	3	1: 398, 2: 236, 0: 117
care_options	0	1	FALSE	3	1: 311, 0: 239, 2: 201
wellness_program	0	1	FALSE	3	0: 455, 1: 167, 2: 129
seek_help	0	1	FALSE	3	0: 300, 2: 262, 1: 189
anonymity	0	1	FALSE	3	2: 495, 1: 237, 0: 19
leave	0	1	FALSE	5	0: 385, 2: 137, 1: 108, 3: 68
mental_health_consequence	0	1	FALSE	3	2: 300, 0: 280, 1: 171
phys_health_consequence	0	1	FALSE	3	0: 571, 2: 150, 1: 30
coworkers	0	1	FALSE	2	0: 627, 1: 124
supervisor	0	1	FALSE	2	0: 447, 1: 304
mental_health_interview	0	1	FALSE	3	0: 635, 2: 100, 1: 16
phys_health_interview	0	1	FALSE	3	0: 339, 2: 320, 1: 92
mental_vs_physical	0	1	FALSE	3	2: 363, 1: 201, 0: 187
obs_consequence	0	1	FALSE	2	0: 662, 1: 89

Variable type: numeric

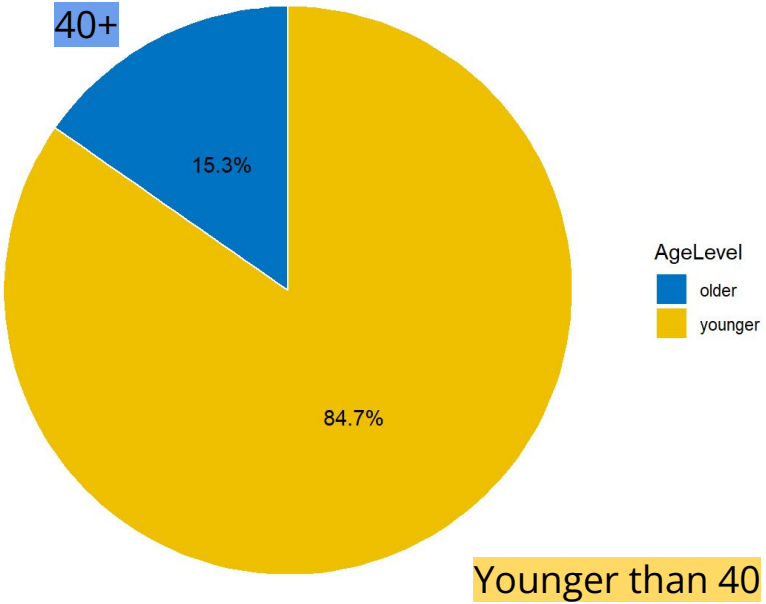
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Age	0	1	33.13	7.62	18	28	32	37	72	▃▇▂▁▁

Research Question 1

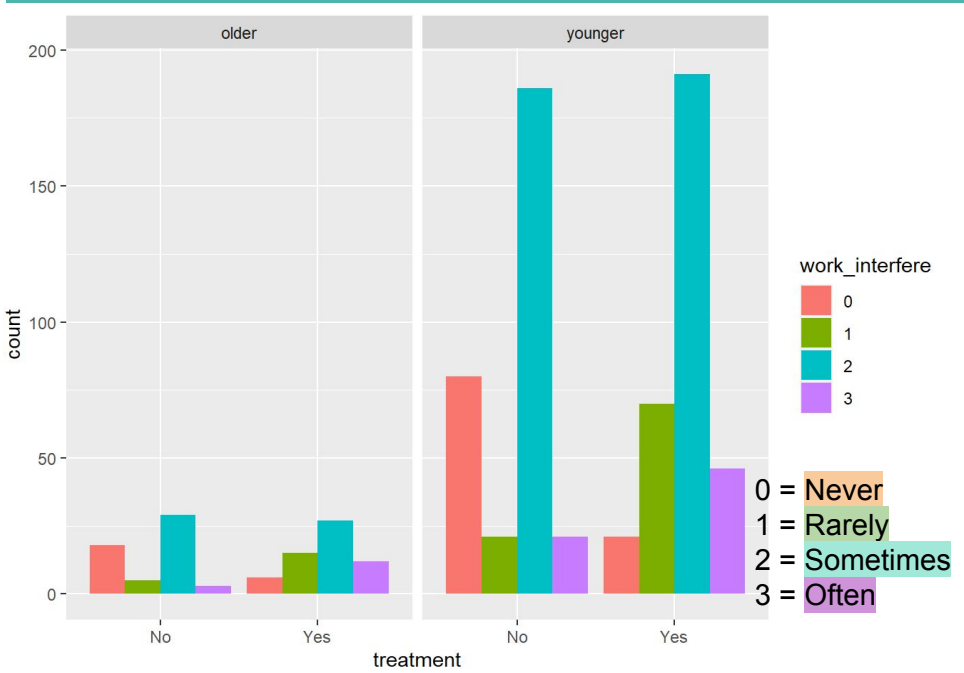
Are employees older than 40 accustomed to stress in the workplace and not seeking mental health treatment as much compared to employees younger than 40?

METHODS USED IN R

Proportion of Workers by Age Group



Work Interference by Age

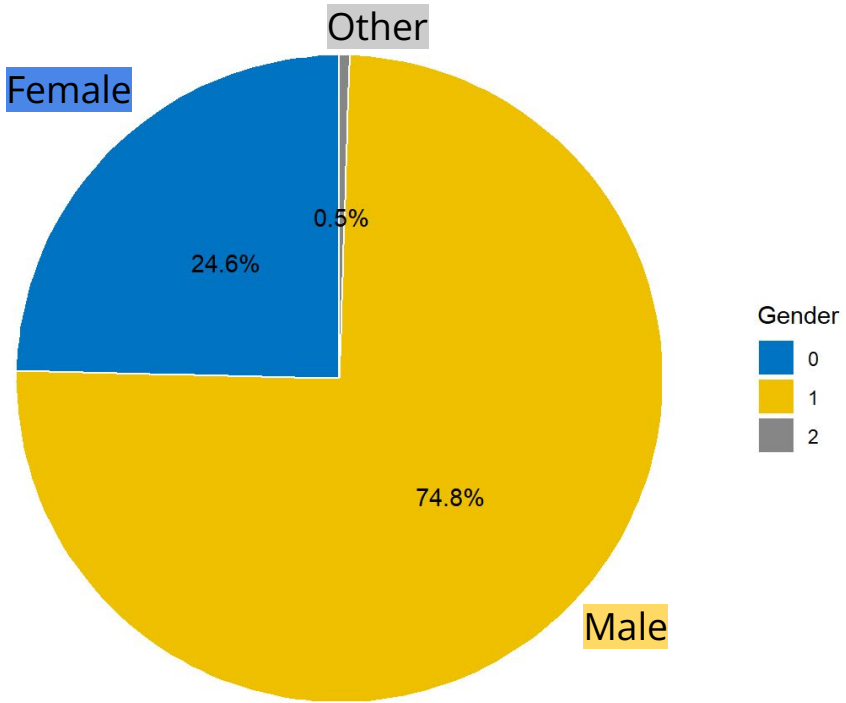


Research Question 2

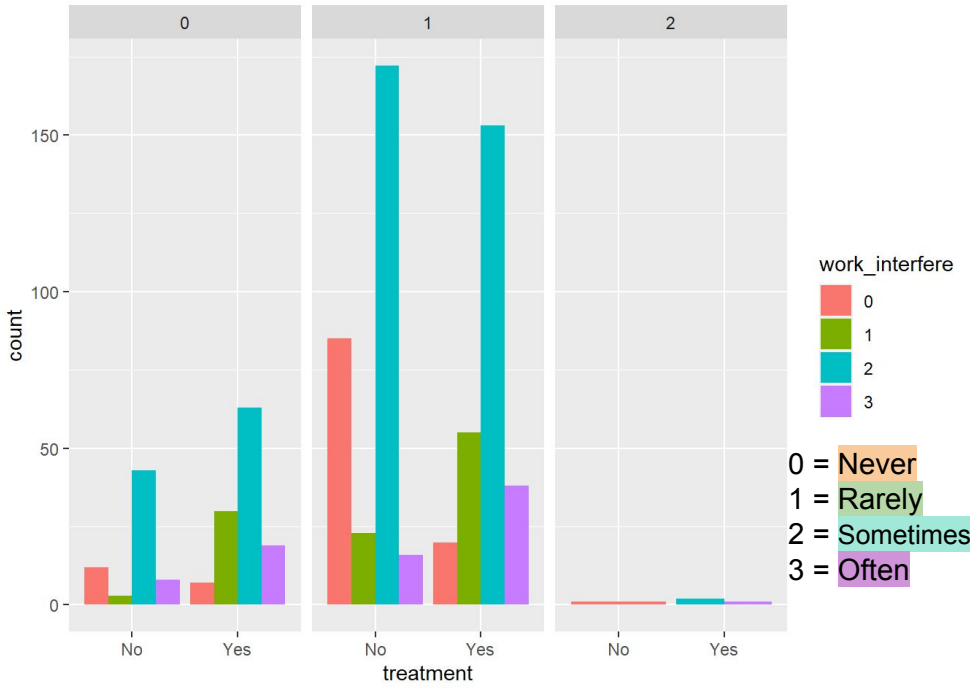
Is one gender seeking mental health services in the workplace more than the other?

METHODS USED IN R

Proportion of Workers by Gender



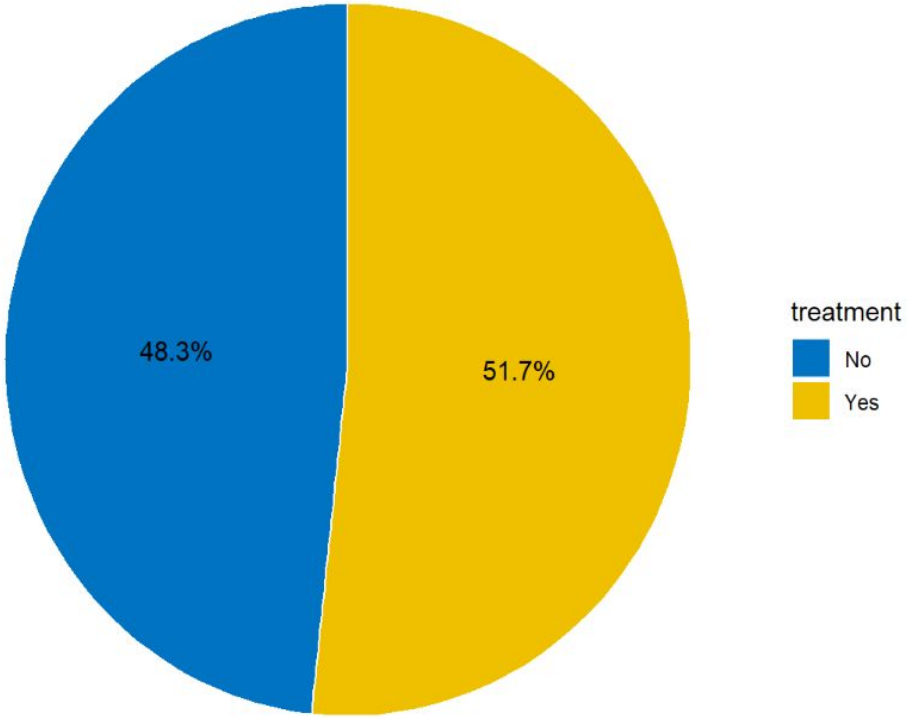
Work interference by Gender



Research Question 3

Which model is best in classifying whether or not the participant needs treatment?

Dependent Variable



- Determine whether or not the participant needs mental health assistance
- Balanced dataset
- Training/Testing - 80%/20%

Method	Confusion matrix	Overall Accuracy
Logistic Regression 1	true pred No Yes No 57 19 Yes 21 53	73.3%
Logistic Regression 2	true pred No Yes No 55 16 Yes 23 56	74.0%
LDA	true pred No Yes No 57 20 Yes 21 52	72.7%
QDA	true pred No Yes No 51 23 Yes 27 49	66.7%
Decision Tree	true pred No Yes No 41 11 Yes 37 61	68.0%
Pruned Tree	pred No Yes No 45 17 Yes 33 55	66.7%

Method	Confusion matrix	Overall Accuracy
Bagging	true pred No Yes No 51 16 Yes 27 56	71.3%
Random Forest	true pred No Yes No 53 14 Yes 25 58	74.0%
Boosting N.Tree = 100 Shrinkage = 0.05, I.D = 3	true pred 0 1 0 53 14 1 25 58	74.0%
SVC Linear	truth pred No Yes No 53 20 Yes 25 52 - best parameters: cost gamma 0.04641589 0.001	70.0%
SVM Radial	truth pred No Yes No 54 22 Yes 24 50 - best parameters: cost gamma 7.742637 0.003593814	69.3%
SVM Poly	truth pred No Yes No 53 25 Yes 25 47 - best parameters: cost degree 100 3	66.7%

Logistic Regression VS Random Forest

- Logistic Regression 2

- Random Forest and Boosting

```
## Overall Accuracy is 0.74
## Sensitivity is 0.778
## Specificity is 0.705
```

```
## Overall Accuracy is 0.74
## Sensitivity is 0.806
## Specificity is 0.679
```

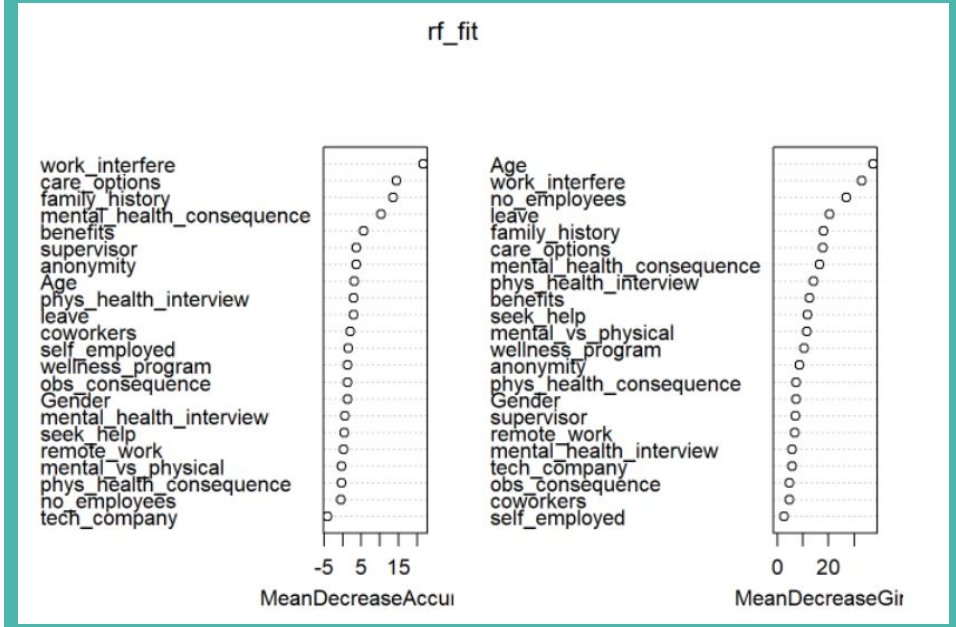
Research Question 4

What are the important/significant factors that determine whether or not the subject needs treatment?

Random Forest

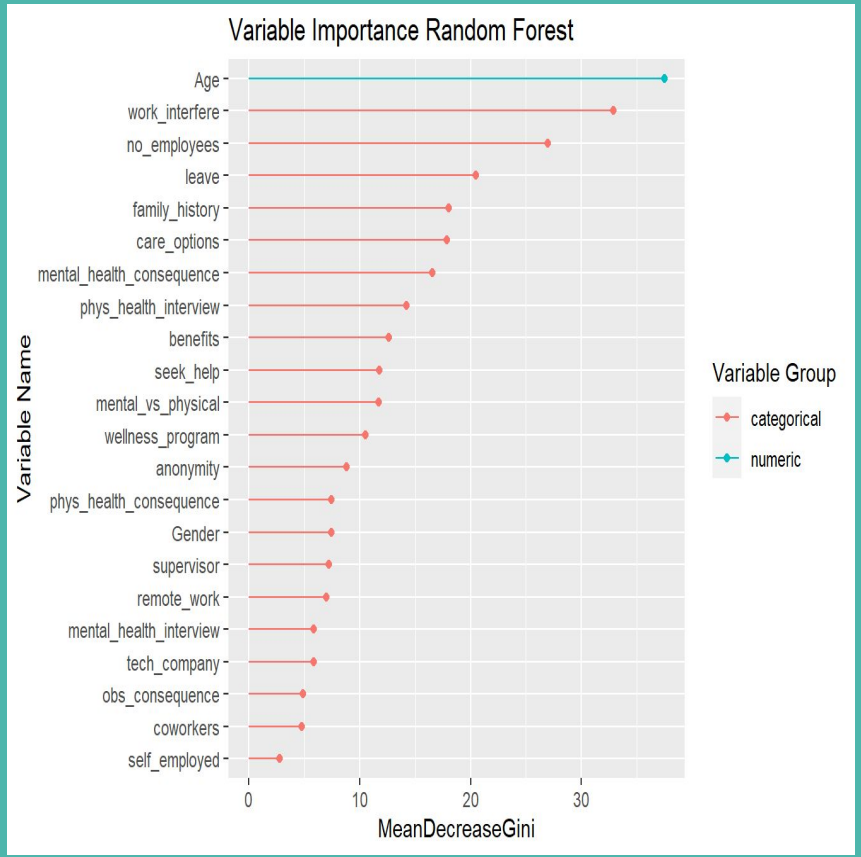
```
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 5

        OOB estimate of  error rate: 31.28%
Confusion matrix:
     No Yes class.error
No  185 100   0.3508772
Yes  88 228   0.2784810
```



Random Forest

- Age
- No employees
- Family history
- Work interference
- Care options
- Leave
- Mental health consequences
- Physical health



Logistic regression 1

- Family history
- Work interference
- Care options
- Leave
- Mental health consequences
- Physical health

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.300121	1.009421	-2.279	0.02269	*
Age	0.001472	0.014057	0.105	0.91661	
Gender1	-0.194866	0.241468	-0.807	0.41966	
Gender2	-1.199522	1.445687	-0.830	0.40669	
self_employed1	0.019208	0.464283	0.041	0.96700	
family_history1	0.856896	0.202312	4.236	2.28e-05	***
work_interfere1	2.531967	0.403016	6.283	3.33e-10	***
work_interfere2	1.223338	0.295232	4.144	3.42e-05	***
work_interfere3	2.104243	0.424017	4.963	6.95e-07	***
no_employees2	-0.413177	0.436256	-0.947	0.34359	
no_employees3	-0.107164	0.467584	-0.229	0.81872	
no_employees4	0.043669	0.490957	0.089	0.92912	
no_employees5	-0.485419	0.625939	-0.776	0.43804	
no_employees6	-0.457819	0.487862	-0.938	0.34803	
remote_work1	-0.129795	0.234450	-0.554	0.57984	
tech_company1	0.149127	0.264029	0.565	0.57220	
benefits1	-0.286940	0.365100	-0.786	0.43191	
benefits2	-0.646993	0.388890	-1.664	0.09617	.
care_options1	0.756571	0.268759	2.815	0.00488	**
care_options2	0.003500	0.264017	0.013	0.98942	
wellness_program1	0.174989	0.335732	0.521	0.60222	
wellness_program2	0.397994	0.311501	1.278	0.20137	
seek_help1	0.212630	0.347620	0.612	0.54075	
seek_help2	0.464986	0.274612	1.693	0.09041	.
anonymity1	0.336043	0.690145	0.487	0.62632	
anonymity2	0.155778	0.675650	0.231	0.81766	
leave1	-0.165537	0.321539	-0.515	0.60667	
leave2	0.231792	0.281106	0.825	0.40961	
leave3	0.860605	0.396552	2.170	0.02999	*
leave4	-0.306077	0.426036	-0.718	0.47249	
mental_health_consequence1	1.027544	0.374187	2.746	0.00603	**
mental_health_consequence2	0.574655	0.285930	2.010	0.04445	*
phys_health_consequence1	-0.062298	0.562721	-0.111	0.91185	
phys_health_consequence2	0.131633	0.278677	0.472	0.63668	
coworkers1	0.392483	0.307919	1.275	0.20244	
supervisor1	-0.118107	0.264746	-0.446	0.65551	
mental_health_interview1	-0.068784	0.805096	-0.085	0.93191	
mental_health_interview2	-0.191185	0.337241	-0.567	0.57078	
phys_health_interview1	0.176080	0.365387	0.482	0.62988	
phys_health_interview2	-0.483388	0.224303	-2.155	0.03116	*
mental_vs_physical1	0.374232	0.357981	1.045	0.29584	
mental_vs_physical2	0.093043	0.269887	0.345	0.73029	
obs_consequence1	0.139963	0.341734	0.410	0.68212	

Logistic regression 2

- Family history
- Work interference
- Care options
- Leave
- Mental health consequences
- Physical health

Given all other variables are constant:

- The estimated odds of needing treatments for people with family history of mental illnesses are 236.077 % of those for people without family history of mental illnesses.
- The estimated odds of needing treatments for participants who claimed mental issues often interfere with their works are 699.398% of those for people who never had issues with mental illnesses at work.

```
# Coefficients:
#                           Estimate Std. Error z value Pr(>|z|)
# (Intercept)               -2.17624    0.35729  -6.091 1.12e-09 ***
# family_history1            0.85899    0.19178   4.479 7.50e-06 ***
# work_interfere1            2.44518    0.38283   6.387 1.69e-10 ***
# work_interfere2            1.14276    0.28302   4.038 5.40e-05 ***
# work_interfere3            1.94505    0.39462   4.929 8.27e-07 ***
# care_options1              1.01634    0.22709   4.476 7.62e-06 ***
# care_options2              0.12968    0.24321   0.533  0.59389
# leave1                    -0.04309    0.29930  -0.144  0.88553
# leave2                     0.27258    0.26351   1.034  0.30094
# leave3                     0.89488    0.37664   2.376  0.01750 *
# leave4                    -0.19528    0.37874  -0.516  0.60613
# mental_health_consequence1 0.77768    0.26605   2.923  0.00347 **
# mental_health_consequence2 0.41364    0.22409   1.846  0.06492 .
# phys_health_interview1     0.16140    0.31055   0.520  0.60325
# phys_health_interview2    -0.51218    0.20178  -2.538  0.01114 *
```

Prediction

Given a person's profile, would the model predict

them seeking help?

Profile

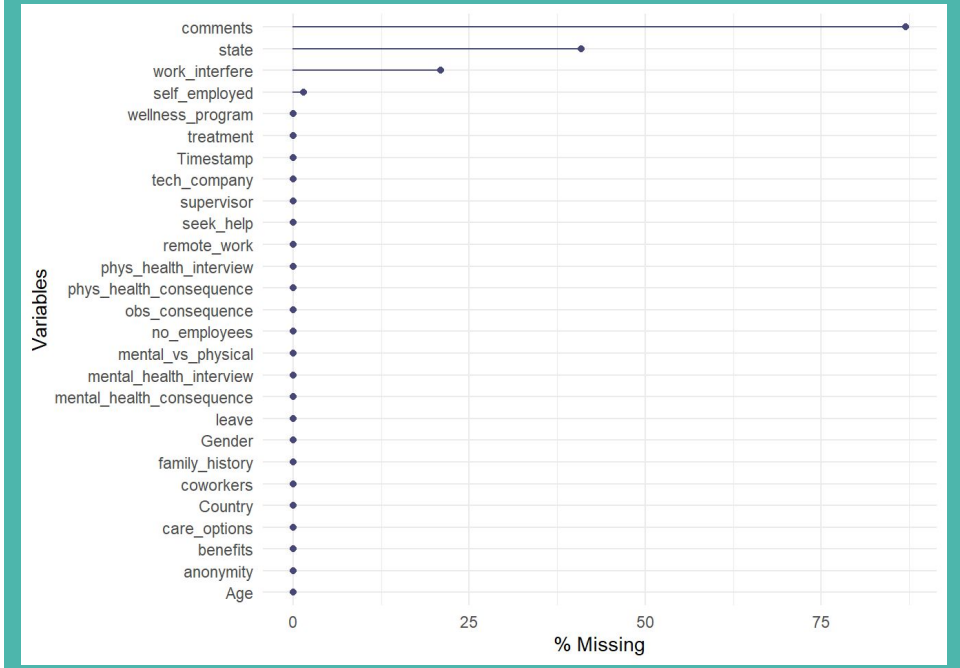
- A Family history of mental health
- Mental illness Sometimes interferes with work
- The worker knows they have care options
- Medical leave is Somewhat difficult
- Thinks discussing mental health issues would lead to negative consequences
- Would not bring up physical health issues to the employer

METHODS USED IN R

Logistic Regression	Boosting
0.9370514	0.8287095

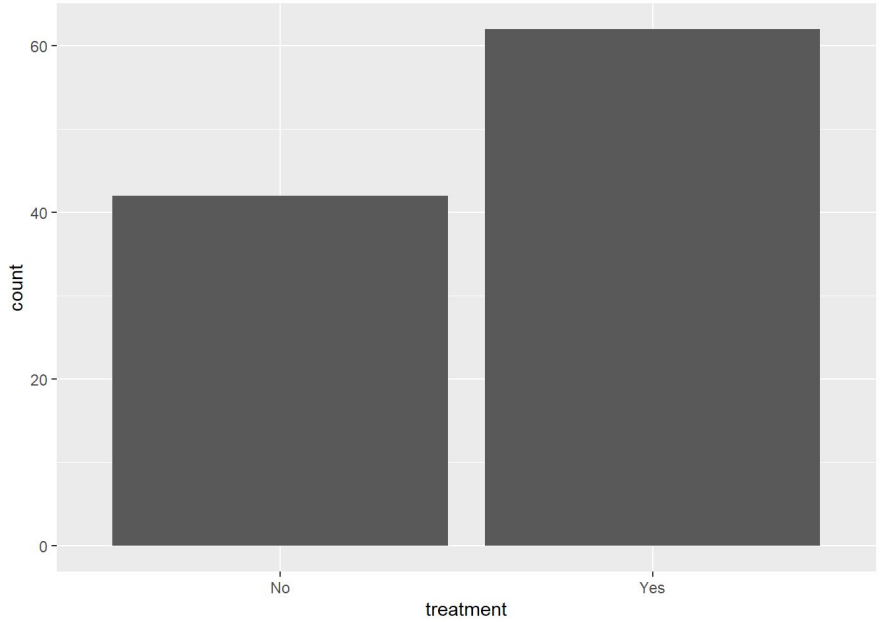
Result: YES

Over 75% of comments are missing



Inference

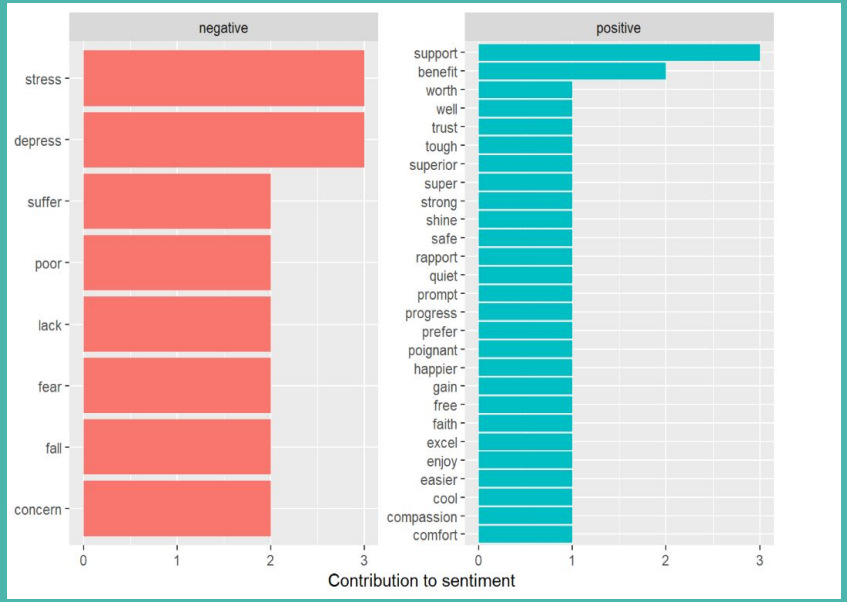
Did people who make comments need help?



60%

Of employees who made comments need treatment

Mental Health boils down to stress and support



Company Issues Affect Our Health

They give us feelings of anxiety, depression and we would like time off or help covered by our insurance

Conclusion

- Logistic Regression 2, Random Forest, and Boosting are the top 3 models with an accuracy of 74% for this dataset.
- There are more employees under 40 seeking mental health services and there are less employees over 40 seeking mental health services.
- Age, No employees, Family history, Work interference, Care options, Leave , Mental health consequences, Physical health are significant predictors in determining whether or not a subject needs treatment.
- Analyzation of the comments using sentiment analysis concluded stress, depress,support, and benefit are the top negative opinions of employees in reference to mental health.
- Logistic Regression and Boosting models had high accuracy rates when making a prediction at 94% and 83%.

CHALLENGES

- Finding the right dataset
- 1 numeric variable in comparison with 22 categorical ones
- Many NA values
- Determining the best ways to process categorical columns
- Choosing to leave a comment may be biased towards those who are more willing to speak or they may be too nervous to leave a comment to express themselves freely
- Profanity
- Misspelling

Future Works

- For future works we would find a data set with more numeric variables.
- Dataset consisting of information from various industries.
- Evaluate how Mental Health affects the quality of work.
- Add a column of salary to see how it affects Mental Health.
- Another column adding of race would be interesting.

MENTAL HEALTH AND WELLBEING RESOURCES

BLACKBURN with DARWEN BOROUGH COUNCIL

Thank You

Professor Liao and STAT 473 Classmates