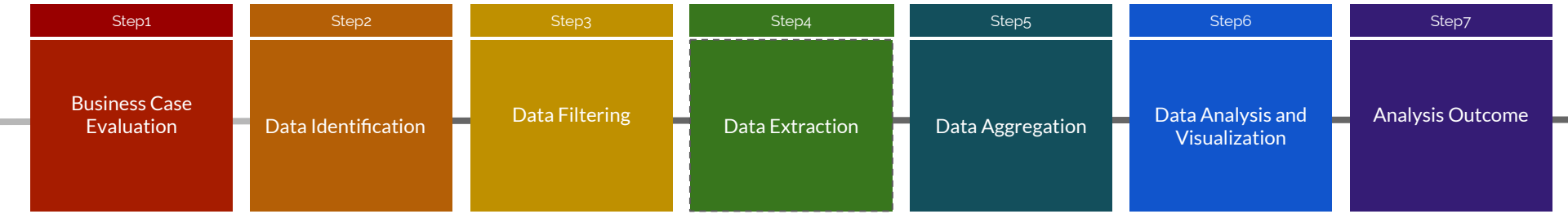
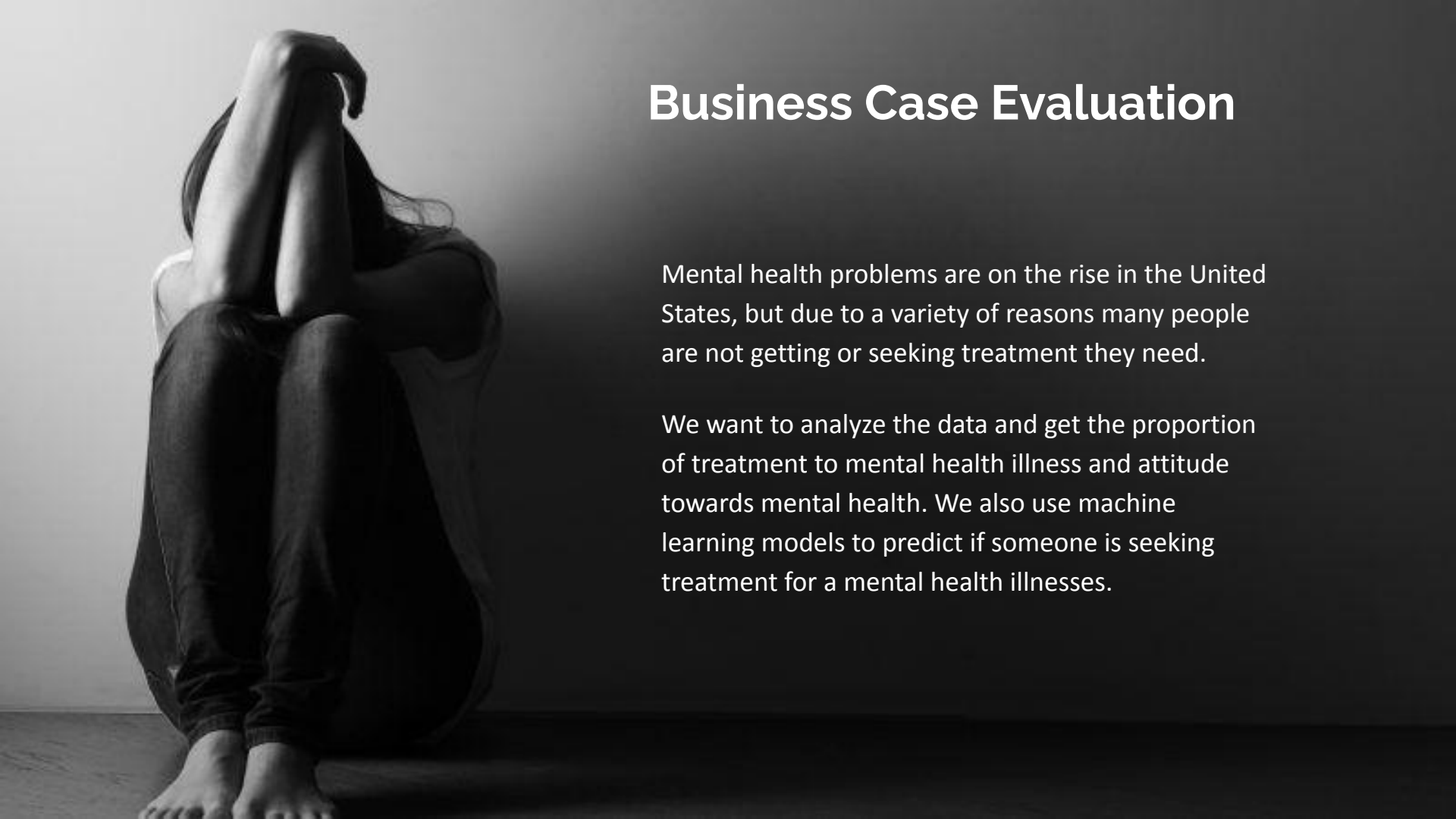
Mental Health Data Analytics Project Ⅲ

By Bibo Gao

Analytics Pipeline





Business Case Evaluation

Mental health problems are on the rise in the United States, but due to a variety of reasons many people are not getting or seeking treatment they need.

We want to analyze the data and get the proportion of treatment to mental health illness and attitude towards mental health. We also use machine learning models to predict if someone is seeking treatment for a mental health illnesses.

Data Identification

This dataset is from a survey that measures attitudes (**27** Columns) towards mental health and frequency of mental health disorders in the tech workplace. The dataset was conducted with **1254** participants, ages range from 5 to 72 years (mean = **32**, standard deviation = **7.375**). **20%** of the participants are women, and **79%** are men. **60%** are from US, 15% from UK, 6% from Canada, with **87%** non-self-employed, **39%** having family mental health issue history, **82%** working in high-tech company.

Dataset Source:https://www.kaggle.com/osmi/mental-health-in-tech-survey

Data Filtering

Normalization
Binning
Missing
Duplicates
Outliers
Filter

1	adults(25-35)	701	0.559
2	adults(35-45)	277	0.221
3	young_adults(18-25)	210	0.167
4	adults(45-55)	42	0.033
5	seniors(55-65)	13	0.010
6	teens(12-18)	7	0.006
7	kids(0-12)	3	0.002
8	pensioners(65+)	1	0.001

- Initial “Gender” attribute has 47 different values. Use Regular Expression to make Gender three values: ‘F’, ‘M’, ‘others’.
- Remove useless data: age < 0 or age > 150
- Remove the column not needed: comments, timestamp
- Discretize age to divide age into different categories.

Data Extraction

Personal Info

Workplace

Attitudes

Others

```
%sql

--drop the table if exists
DROP TABLE IF EXISTS data_mentalhealth2;

--create table
CREATE TABLE  IF NOT EXISTS data_mentalhealth2 (
                    Age string, Gender string,
                    Country string,
                    state string ,
                    self_employed string,
                    family_history string,
                    treatment string,
                    work_interfere string,
                    no_employees string,
                    remote_work string,
                    tech_company string,
                    benefits string,
                    care_options string,
                    wellness_program string,
                    seek_help string,
                    anonymity string,
                    leave string,
                    mental_health_consequence string,
                    phys_health_consequence string,
                    coworkers string,
                    supervisor string,
                    mental_health_interview string,
                    phys_health_interview string,
                    mental_vs_physical string,
                    obs_consequence string)
USING CSV
OPTIONS (path "/FileStore/tables/mentalhealth_data_clean-1.csv", header "true");

/* check results */
select * from data_mentalhealth2 limit 100;
```

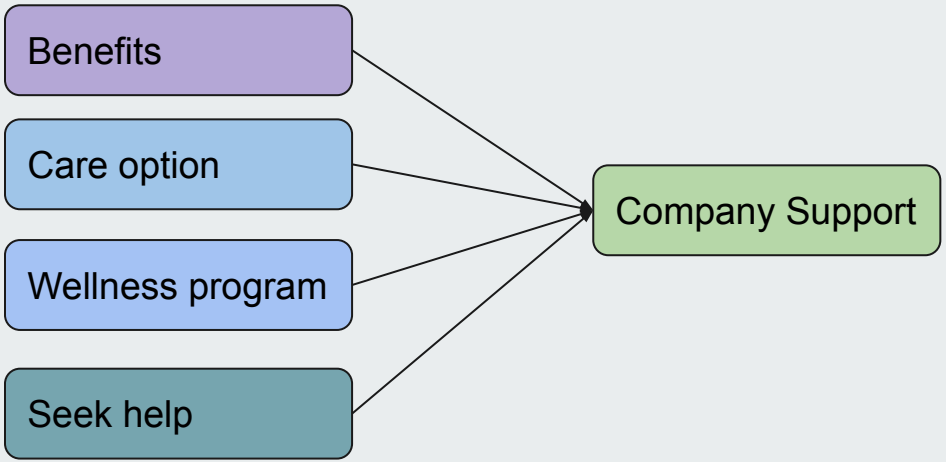
▸ (1) Spark Jobs

	Age	Gender	Country	state	self_employed	family_history	treatment	work_interfere	no_employe
1	adults(35-45)	F	United States	IL	NA	No	Yes	Often	6-25
2	adults(35-45)	M	United States	IN	NA	No	No	Rarely	More than 1
3	adults(25-35)	M	Canada	NA	NA	No	No	Rarely	6-25
4	adults(25-35)	M	United Kingdom	NA	NA	Yes	Yes	Often	26-100
5	adults(25-35)	M	United States	TX	NA	No	No	Never	100-500
6	adults(25-35)	M	United States	TN	NA	Yes	No	Sometimes	6-25
7	adults(25-35)	F	United States	MI	NA	Yes	Yes	Sometimes	1-5

Showing all 100 rows.

Command took 1.41 seconds -- by bgao@luc.edu at 4/27/2021, 12:30:58 AM on comp358

Data Aggregation



Generate a “company_support” attribute which would compose of **benefits, care_option, wellness_program, seek_help** attributes.
Then descretize to High, Medium, Low three levels of company support.

Low	588	0.469
Medium	454	0.362
High	212	0.169

Data Analysis and Visualization

APACHE Spark™

Scopes

1. Factor of Treatment
2. Mental Health vs Physical Health
3. Percentage by States
4. Prediction Model

Factor of treatment

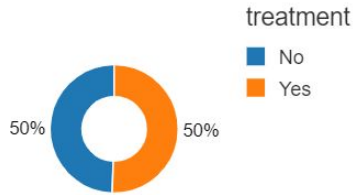
What will make people choose to seek help for their mental illness?

Half

Won’t seek treatment in mental health illness


```
%sql

select
  treatment,
  count(1) as treat_count
from
  data_mentalhealth2
group by
  treatment
```

▸ (5) Spark Jobs



Plot Options...

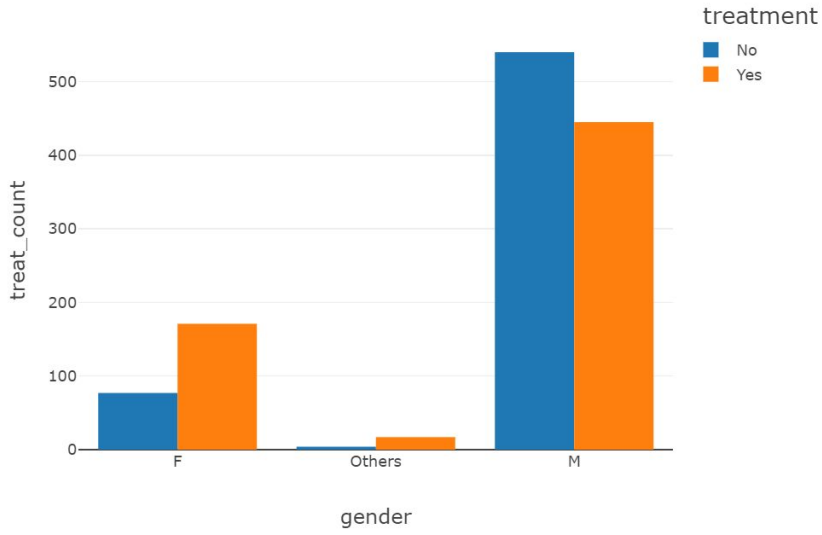
Command took 4.76 seconds -- by bgao@luc.edu at 4/28/2021, 11:46:07 AM on comp358

Gender

The percentage of intention to seek treatment in the female category was higher than in the male category.

```sql
%sql

select
  gender,
  treatment,
  count(1) as treat_count
from
  data_mentalhealth2
group by
  treatment, gender
```

▸ (5) Spark Jobs



Plot Options...

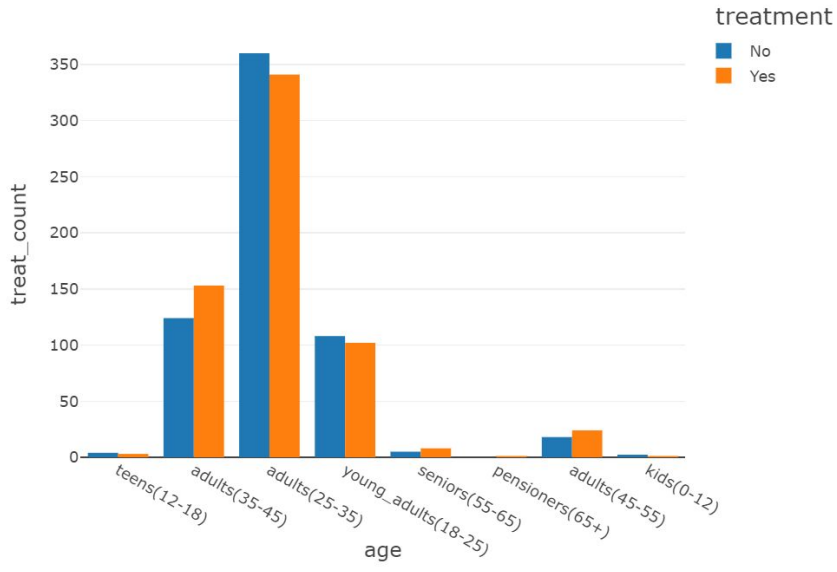
Command took 2.96 seconds -- by bgao@luc.edu at 4/27/2021, 12:32:18 AM on comp358

Age

In general, older adults tend to seek treatment than younger adults.

```sql
%sql

select
  age,
  treatment,
  count(1) as treat_count
from
  data_mentalhealth2
group by
  treatment, age
```

▸ (5) Spark Jobs



Plot Options...

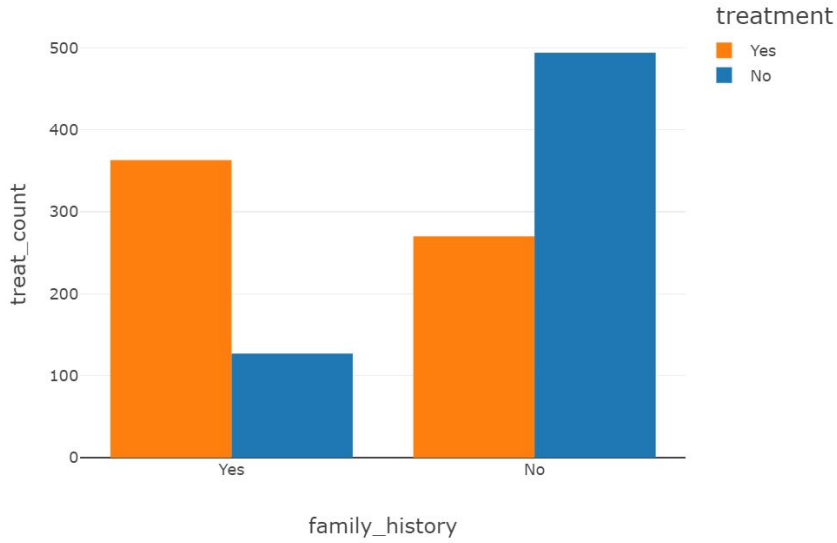
Command took 3.24 seconds -- by bgao@luc.edu at 4/27/2021, 12:37:27 AM on comp358

Family history

If a person has a family history of mental illness, it will encourage him/her to seek mental health treatment.

```
%sql

select
  family_history,
  treatment,
  count(1) as treat_count
from
  data_mentalhealth2
group by
  treatment, family_history
```

▸ (5) Spark Jobs



Plot Options...

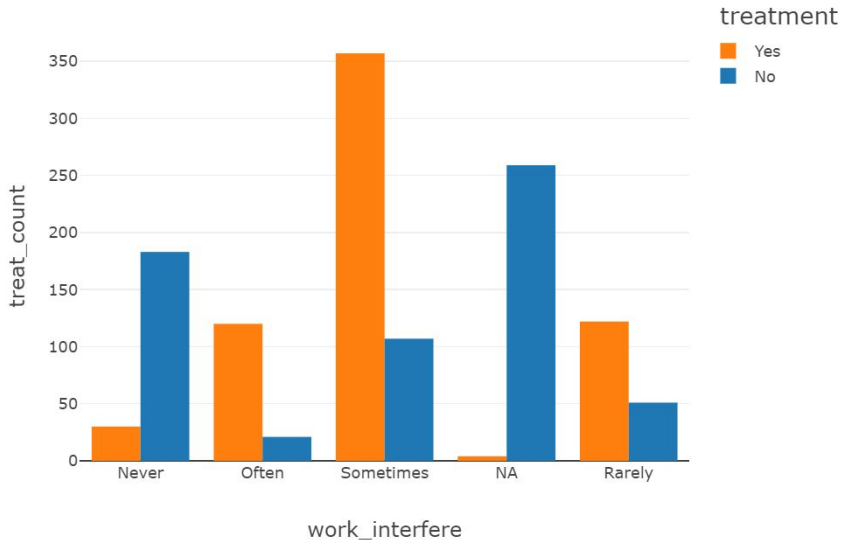
Command took 2.78 seconds -- by bgao@luc.edu at 4/27/2021, 12:41:14 AM on comp358

Work interfere

If a person thinks a mental health condition won't interfere with his/her work, he/she might not seek treatment.

```
%sql

select
  work_interfere,
  treatment,
  count(1) as treat_count
from
  data_mentalhealth2
group by
  treatment, work_interfere
```

(5) Spark Jobs



Plot Options...

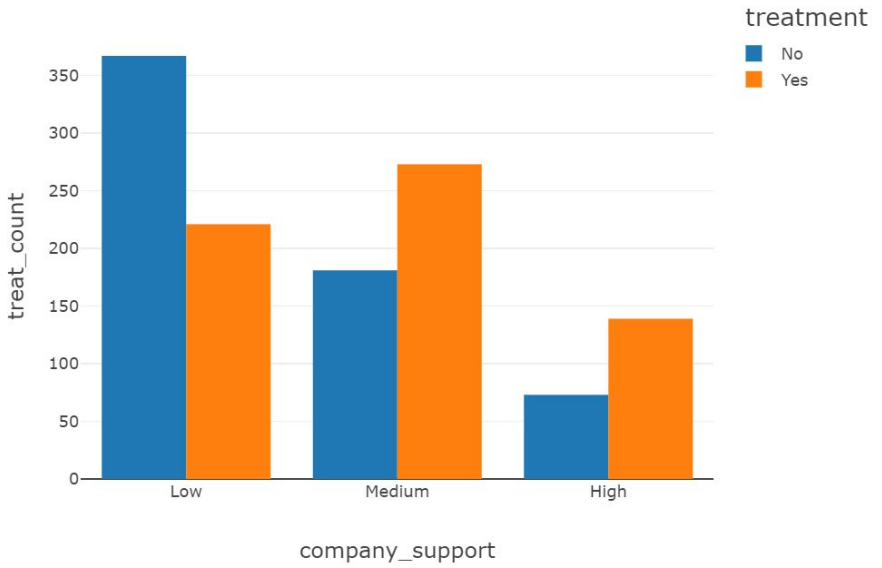
Command took 2.61 seconds -- by bgao@luc.edu at 4/27/2021, 12:44:39 AM on comp358

Company support

Enhancing company support in mental health issue can encourage people to seek treatment.

```
%sql

select
  company_support,
  treatment,
  count(1) as treat_count
from
  mentalhealth_stat_csv
group by
  treatment, company_support
```

(5) Spark Jobs



Plot Options...

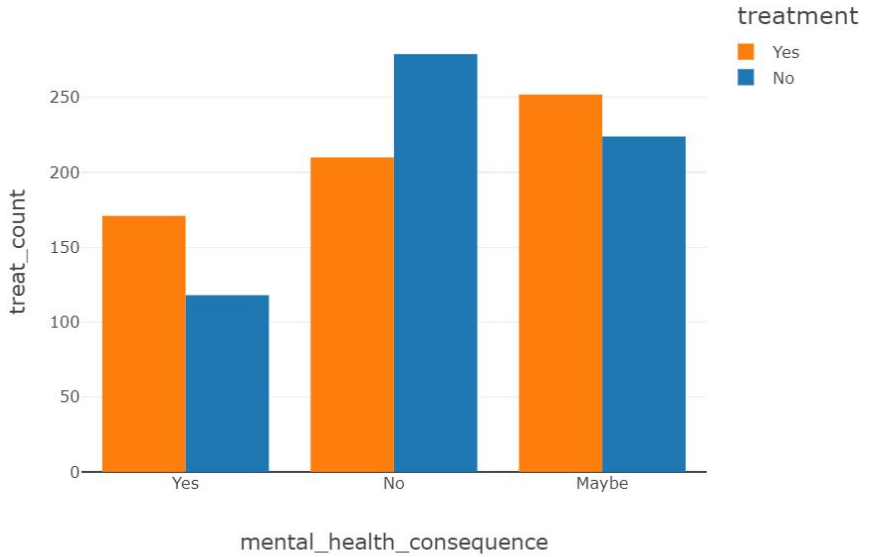
Command took 5.76 seconds -- by bgao@luc.edu at 4/27/2021, 11:33:40 AM on comp358

Mental health consequence

If a person thinks mental health issue have negative consequences, he/she would seek treatment.

```sql
%sql

select
  mental_health_consequence,
  treatment,
  count(1) as treat_count
from
  mentalhealth_stat_csv
group by
  treatment, mental_health_consequence
```

▸ (5) Spark Jobs

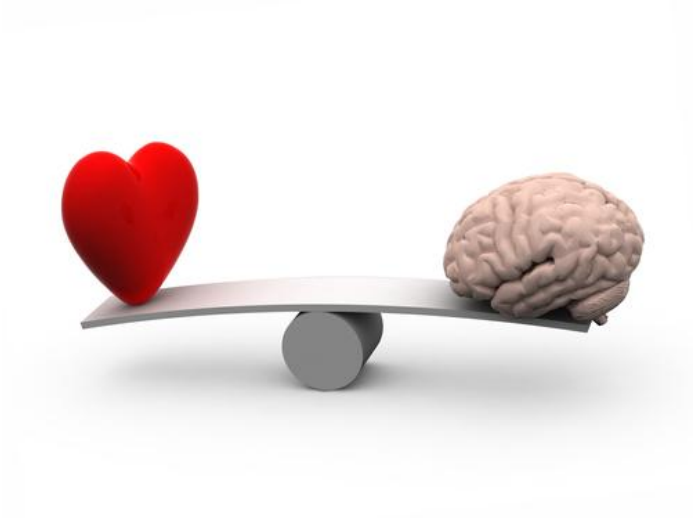


Plot Options...

Command took 3.27 seconds -- by bgao@luc.edu at 4/29/2021, 3:33:37 PM on comp358

Mental Health vs Physical Health

There is a lot of stigma and discrimination associated to such disorders.

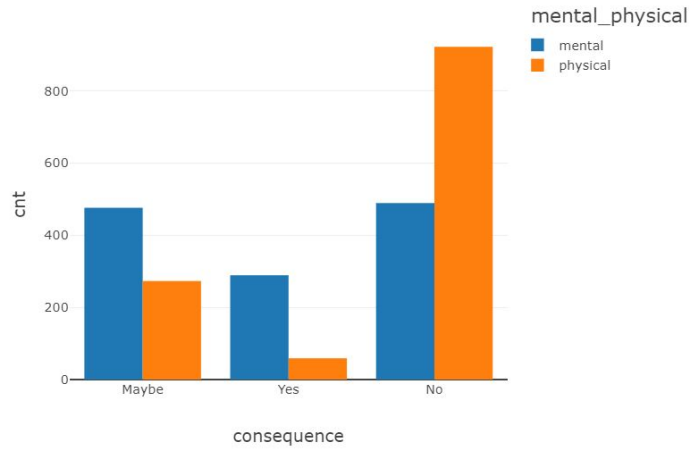


Negative consequences

Comparing to physical health issue, more people think **mental** health issue would have negative consequences.

```sql
%sql

select
  'mental' as mental_physical,
  mental_health_consequence as consequence,
  count(mental_health_consequence) as cnt
from
  data_mentalhealth2
group by
  mental_health_consequence

union

select
  'physical' as mental_physical,
  phys_health_consequence as consequence,
  count(phys_health_consequence) as cnt
from
  data_mentalhealth2
group by
  phys_health_consequence

```

▸ (5) Spark Jobs



Plot Options...

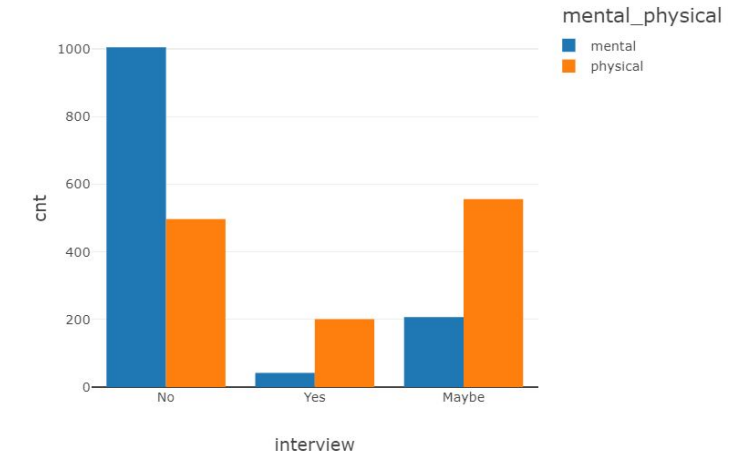
Command took 8.27 seconds -- by bgao@luc.edu at 4/29/2021, 6:03:49 PM on comp358

Talking in an interview

Comparing to physical health, less people would bring up a **mental** health issue with a potential employer in an interview.

```
%sql

select
  'mental' as mental_physical,
  mental_health_interview as interview,
  count(mental_health_interview) as cnt
from
  data_mentalhealth2
group by
  mental_health_interview

union

select
  'physical' as mental_physical,
  phys_health_interview as interview,
  count(phys_health_interview) as cnt
from
  data_mentalhealth2
group by
  phys_health_interview
```

▸ (5) Spark Jobs



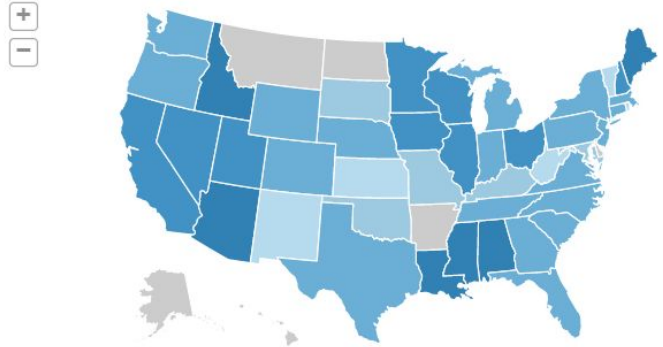
Plot Options...

Command took 8.81 seconds -- by bgao@luc.edu at 4/29/2021, 6:07:09 PM on comp358

Percentage by states

```
1  %sql
2
3    select state, count(IF(treatment = 'Yes', 1, NULL)) / count(*) as treatpercent
4    from mentalhealth_stat_csv
5    where country = 'United States' and state != 'NA'
6    group by state
```

▸ (5) Spark Jobs



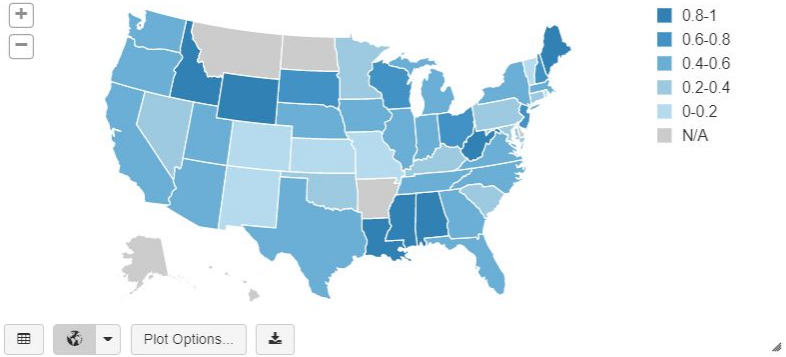
Plot Options...

Command took 2.20 seconds -- by bgao@luc.edu at 4/29/2021, 8:46:42 PM on comp358

Seek treatment percentage

```sql
%sql

select state, count(IF(work_interfere = 'Often' or work_interfere = 'Sometimes', 1, NULL))
      / count(*) as oftensometimespercent
from mentalhealth_stat_csv
where country = 'United States' and state != 'NA'
group by state

```

(5) Spark Jobs

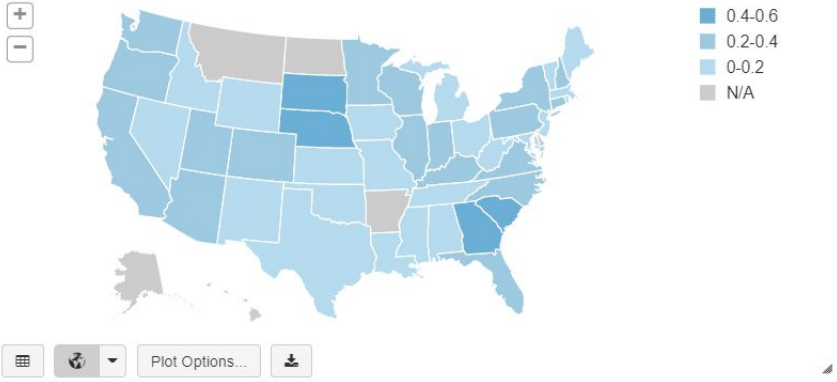


Command took 2.54 seconds -- by bgao@luc.edu at 4/29/2021, 8:42:55 PM on comp358

Often or sometime interfere work percentage

```sql
%sql

  select state, count(IF(company_support = 'High', 1, NULL)) / count(*) as highsupportpercent
  from mentalhealth_stat_csv
  where country = 'United States' and state != 'NA'
  group by state
```

(5) Spark Jobs



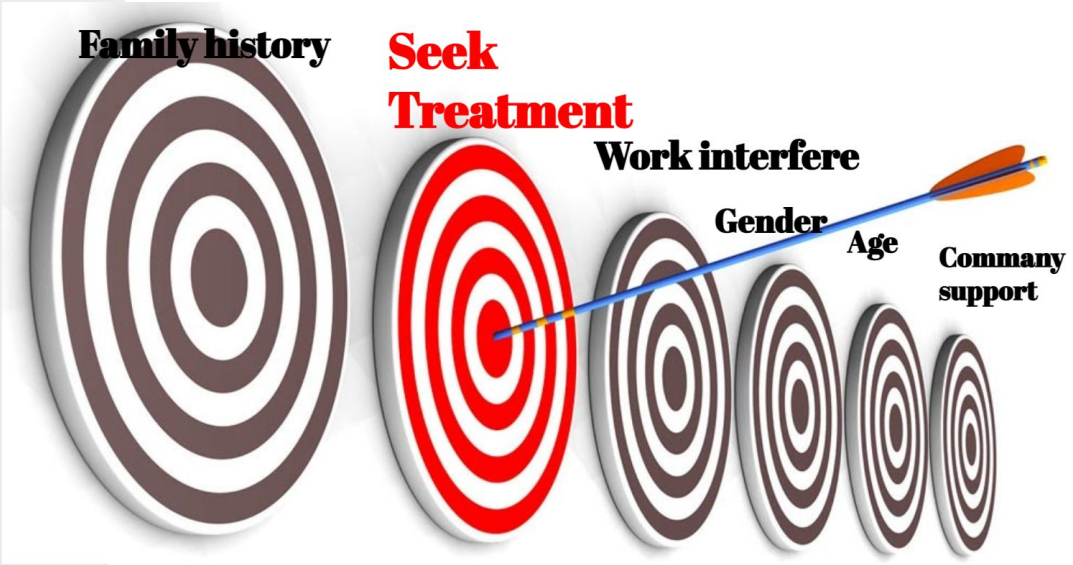
Command took 2.33 seconds -- by bgao@luc.edu at 4/29/2021, 8:48:26 PM on comp358

High company support percentage

Percentage by states

Prediction Model

Select Attribute



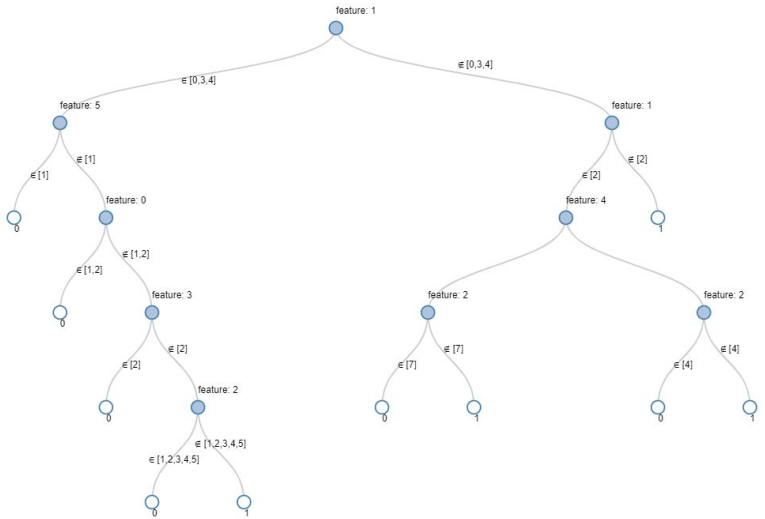
Decision Tree Classification

Cmd 22

```
from pyspark.ml import Pipeline
from pyspark.ml.feature import StringIndexer, VectorIndexer
from pyspark.ml.classification import DecisionTreeClassifier
from pyspark.ml.feature import StringIndexer, VectorIndexer, VectorAssembler, OneHotEncoder
from pyspark.ml.evaluation import MulticlassClassificationEvaluator

# Load the data stored in csv format as a DataFrame.
data = spark.read.format("csv").load("/FileStore/tables/mentalhealth_stat.csv", header=True,
                                     inferSchema = True)
#data.show()
data2 =data.select(data.company_support,data.work_interfere, data.Age, data.Gender, data.mental_health_consequence, data.family_history,
data.treatment.alias('label'))

# Split the data into training and test sets (30% held out for testing)
(trainingData, testData) = data2.randomSplit([0.7, 0.3])

# Index labels, adding metadata to the label column.
# Fit on whole dataset to include all labels in index.
labelIndexer = StringIndexer(inputCol="label", outputCol="indexedLabel").fit(data2)
# string attributes to indexer
company_supportIndexer =StringIndexer(inputCol='company_support', outputCol="indexedcompany_support")
work_interfereIndexer =StringIndexer(inputCol='work_interfere', outputCol="indexedwork_interfere")
AgeIndexer =StringIndexer(inputCol='Age', outputCol="indexedAge")
GenderIndexer =StringIndexer(inputCol='Gender', outputCol="indexedGender")
mental_health_consequenceIndexer =StringIndexer(inputCol='mental_health_consequence', outputCol="indexedmental_health_consequence")
family_historyIndexer =StringIndexer(inputCol='family_history', outputCol="indexedfamily_history")

featureAssembler =
VectorAssembler().setInputCols(['indexedcompany_support','indexedwork_interfere','indexedAge','indexedGender','indexedmental_health_consequence',
'indexedfamily_history']).setOutputCol('features')

# Train a DecisionTree model.
dt = DecisionTreeClassifier(labelCol="indexedLabel", featuresCol="features")

# Chain indexers and tree in a Pipeline
pipeline = Pipeline(stages=[labelIndexer, company_supportIndexer,work_interfereIndexer, AgeIndexer, GenderIndexer, mental_health_consequenceIndexer,
family_historyIndexer,  featureAssembler, dt])

# Train model.  This also runs the indexers.
model = pipeline.fit(trainingData)

# Make predictions.
predictions = model.transform(testData)

# Select example rows to display.
predictions.select("prediction", "indexedLabel", "features").show(15)

# Select (prediction, true label) and compute test error
evaluator = MulticlassClassificationEvaluator(
    labelCol="indexedLabel", predictionCol="prediction", metricName="accuracy")
accuracy = evaluator.evaluate(predictions)

print("Accuracy = %g " % (accuracy))
print("Test Error = %g " % (1.0 - accuracy))

```

Classification Result

```
1
2 tree = model.stages[-1]
3
4 display(tree) #visualize the decision tree model
```




```
+----------+------------+--------------------+
|prediction|indexedLabel|            features|
+----------+------------+--------------------+
|       1.0|         1.0| (6,[0,1],[2.0,1.0])|
|       1.0|         1.0| (6,[0,1],[2.0,1.0])|
|       1.0|         1.0| (6,[0,1],[2.0,1.0])|
|       1.0|         1.0|[2.0,1.0,1.0,1.0,...|
|       1.0|         1.0|[2.0,1.0,1.0,0.0,...|
|       1.0|         1.0|[2.0,2.0,0.0,1.0,...|
|       1.0|         1.0|[2.0,2.0,0.0,0.0,...|
|       1.0|         1.0|[2.0,2.0,0.0,0.0,...|
|       1.0|         1.0|[2.0,2.0,1.0,0.0,...|
|       1.0|         0.0|[2.0,2.0,1.0,0.0,...|
|       1.0|         1.0|[2.0,2.0,3.0,0.0,...|
|       1.0|         1.0|[2.0,2.0,3.0,0.0,...|
|       1.0|         1.0|[2.0,2.0,2.0,0.0,...|
|       0.0|         1.0|[2.0,4.0,0.0,0.0,...|
|       0.0|         0.0|[2.0,4.0,1.0,0.0,...|
+----------+------------+--------------------+
only showing top 15 rows

Accuracy = 0.815029
Test Error = 0.184971
```

```
company_support: string
work_interfere: string
Age: string
Gender: string
mental_health_consequence: string
family_history: string
label: string
```

Cross Validation Result

```
+----------+------------+-----+--------------------+
|prediction|indexedLabel|label|            features|
+----------+------------+-----+--------------------+
|       1.0|         1.0|   No| (6,[0,1],[2.0,1.0])|
|       1.0|         1.0|   No|[2.0,1.0,0.0,0.0,...|
|       1.0|         0.0|  Yes|[2.0,2.0,3.0,0.0,...|
|       0.0|         0.0|  Yes|[2.0,4.0,0.0,0.0,...|
|       0.0|         0.0|  Yes|[2.0,4.0,3.0,1.0,...|
|       0.0|         0.0|  Yes|[2.0,4.0,6.0,2.0,...|
|       0.0|         0.0|  Yes|[2.0,3.0,0.0,1.0,...|
|       0.0|         0.0|  Yes|[2.0,3.0,0.0,0.0,...|
|       0.0|         0.0|  Yes|[2.0,3.0,0.0,0.0,...|
|       0.0|         0.0|  Yes|[2.0,3.0,0.0,0.0,...|
|       0.0|         0.0|  Yes|[2.0,0.0,0.0,1.0,...|
|       0.0|         1.0|   No|       (6,[0],[2.0])|
|       0.0|         0.0|  Yes|       (6,[0],[2.0])|
|       0.0|         0.0|  Yes|       (6,[0],[2.0])|
|       0.0|         0.0|  Yes|[2.0,0.0,1.0,0.0,...|
+----------+------------+-----+--------------------+
only showing top 15 rows

Accuracy = 0.837662
Test Error = 0.162338
```

```
1
2  # Train a DecisionTree model.
3  dt = DecisionTreeClassifier(labelCol="indexedLabel",
4                              featuresCol="features", maxDepth=2)
5
6  paramGrid = ParamGridBuilder() \
7      .addGrid(dt.maxDepth, [2, 5, 10, 20, 30]) \
8      .addGrid(dt.maxBins, [10, 20, 40, 80, 100]) \
9      .build()
10
11 # A CrossValidator requires an Estimator, a set of Estimator ParamMaps, and an Evaluator.
12 crossval  = CrossValidator(estimator=dt,
13                     estimatorParamMaps=paramGrid,
14                     evaluator=MulticlassClassificationEvaluator(
15                       labelCol="indexedLabel",
16                       predictionCol="prediction",
17                       metricName="accuracy"),
18                     numFolds=5)
19
20 # Chain indexers and tree in a Pipeline
21 pipelineCV = Pipeline(stages=[labelIndexer,company_supportIndexer,
22                               work_interfereIndexer,
23                               AgeIndexer, GenderIndexer,
24                               mental_health_consequenceIndexer,
25                               family_historyIndexer,
26                               featureAssembler, crossval])
27 # Run cross-validation, and choose the best set of parameters.
28 modelCV = pipelineCV.fit(train)
29
30 #va = modelCV.stages[-2]
31 treeCV = modelCV.stages[-1].bestModel
32
33 display(treeCV) #visualize the best decision tree model
34
```

Analysis Outcome

This analysis shows that the attitudes and company support could predict intentions to seek treatment for mental illness.

Proposal 1

Increase Awareness

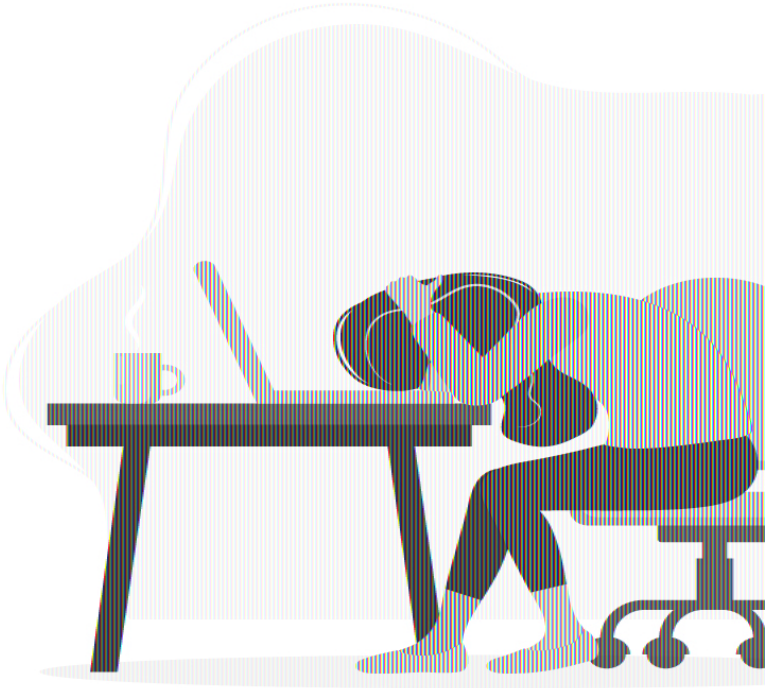
Enhance mental health literacy

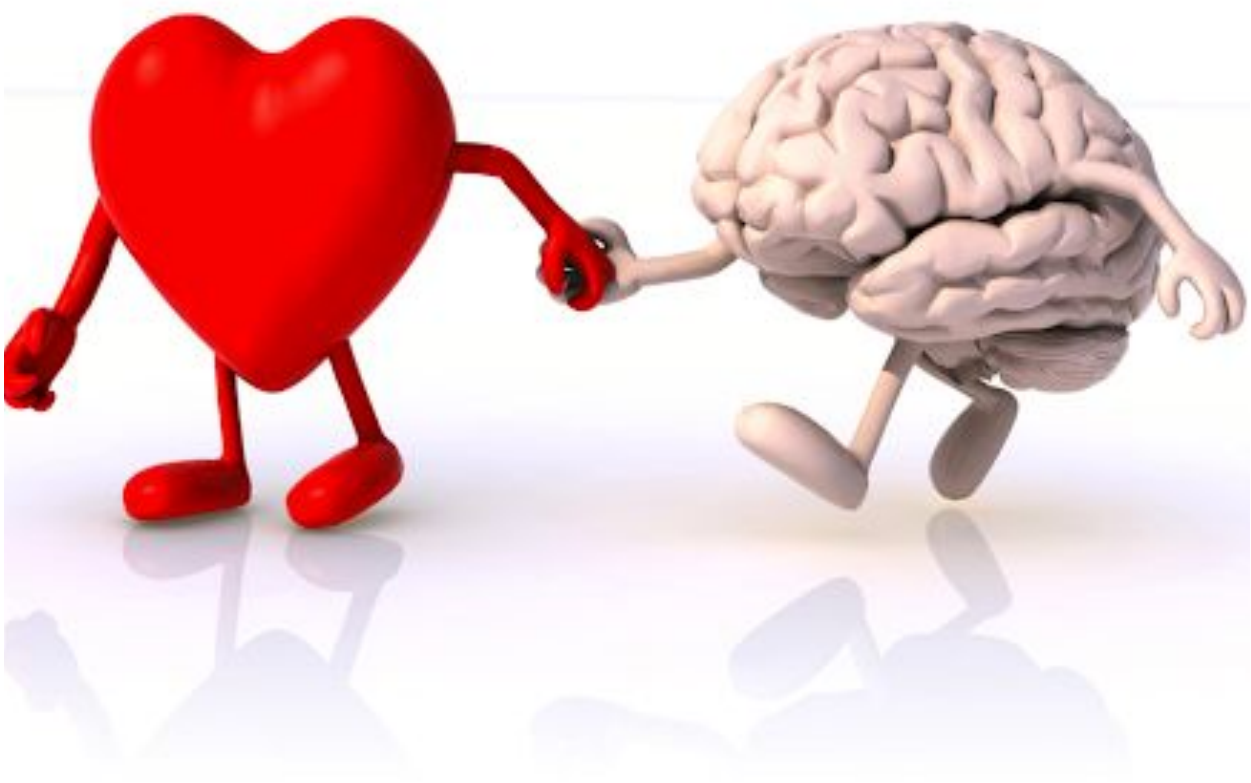
When people realize the impact of mental illness on work, they will be encouraged to actively seek treatment.

Proposal 2

Provide MORE company support and help on mental health issues. It's a win-win.

- **Benefits**
- **Provide resources of seeking help**
- **Treat People Fair**
- **Develop Mental Health Policies**





Health body + Health Mind = Happy life