Introducción

En este trabajo se pretende desarrollar, mediante el proceso KDD, un patrón que permita averiguar si un empleado cualquiera de una compañía tiene una enfermedad mental a partir de una serie de preguntas sobre cómo lleva la empresa el trato de las enfermedades mentales.

Hoy en día, las enfermedades mentales siguen siendo tema tabú en la sociedad, sobre todo en el entorno laboral. Mediante este proyecto quiero investigar si fuera posible saber si una persona tiene una enfermedad mental dependiendo de las respuestas que da a ciertas preguntas sobre cómo es el ambiente en el trabajo respecto a la salud mental.

#### Recursos utilizados

Para las partes que he necesitado escribir código, por ejemplo, la adecuación de la base de datos y la transformación de estos, he usado el lenguaje de programación Python. Los scripts que he escrito están adjuntos a este proyecto, y explicados brevemente más adelante.

También he modificado un programa con el algoritmo KNN para adaptarlo a mi dataset, igualmente en Python.

Una vez pasada la fase de transformación, decidí usar una página web, BigML, para realizar todo el proceso de Data Mining. Esta página me ha permitido lanzar varios modelos y evaluarlos rápidamente, para ver cuál de ellos es el que mejor se adapta a mi dataset.

Además, cuenta con una serie de gráficos que facilitan mucho la tarea de interpretar los datos y decidir qué hacer, si fuera necesaria, en la siguiente iteración del proceso KDD.

Proceso KDD

# Primera iteración

## Seleccionar el conjunto de datos

Esta es la primera fase del proceso KDD donde se ha elegido una base de datos, que es la siguiente: <https://www.kaggle.com/ekwiecinska96/mental-health-in-techology-survey-2014-and-2016#survey_2016.csv>

Al ser una base de datos de *Kaggle*, no es necesario convertir su formato ni realizar una tarjeta de datos, por lo que podemos comenzar a trabajar en la segunda fase inmediatamente.

A continuación, se procede a explicar la base de datos y el trabajo que se realizará sobre ella.

Esta base de datos contiene 63 preguntas realizadas a 1433 personas sobre cómo son tratadas las enfermedades mentales en su empresa, cómo creen que reaccionarían los compañeros de trabajo ante una enfermedad mental y si ellos tienen una enfermedad mental, entre otras preguntas del mismo estilo.

## Adecuación de la base de datos

Las columnas, inicialmente, que tiene, son las siguientes.

|  |
| --- |
| **LEYENDA:**  Columna borrada Columna que se quiere calcular Columna a tener en cuenta en posteriores ejecuciones Columna que contiene el dataset final |

1. Are you self-employed?
2. How many employees does your company or organization have?
3. Is your employer primarily a tech company/organization?
4. Is your primary role within your company related to tech/IT?
5. Does your employer provide mental health benefits as part of healthcare coverage?
6. Do you know the options for mental health care available under your employer-provided coverage?
7. Has your employer ever formally discussed mental health (for example, as part of a wellness campaign or other official communication)?
8. Does your employer offer resources to learn more about mental health concerns and options for seeking help?
9. Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources provided by your employer?
10. If a mental health issue prompted you to request a medical leave from work, asking for that leave would be…
11. Do you think that discussing a mental health disorder with your employer would have negative consequences?
12. Do you think that discussing a physical health issue with your employer would have negative consequences?
13. Would you feel comfortable discussing a mental health disorder with your coworkers?
14. Would you feel comfortable discussing a mental health disorder with your direct supervisor(s)?
15. Do you feel that your employer takes mental health as seriously as physical health?
16. Have you heard of or observed negative consequences for co-workers who have been open about mental health issues in your workplace?
17. Do you have medical coverage (private insurance or state-provided) which includes treatment of mental health issues?
18. Do you know local or online resources to seek help for a mental health disorder?
19. If you have been diagnosed or treated for a mental health disorder, do you ever reveal this to clients or business contacts?
20. If you have revealed a mental health issue to a client or business contact, do you believe this has impacted you negatively?
21. If you have been diagnosed or treated for a mental health disorder, do you ever reveal this to coworkers or employees?
22. If you have revealed a mental health issue to a coworker or employee, do you believe this has impacted you negatively?
23. Do you believe your productivity is ever affected by a mental health issue?
24. If yes, what percentage of your work time (time performing primary or secondary job functions) is affected by a mental health issue?
25. Do you have previous employers?
26. Have your previous employers provided mental health benefits?
27. Were you aware of the options for mental health care provided by your previous employers?
28. Did your previous employers ever formally discuss mental health (as part of a wellness campaign or other official communication)?
29. Did your previous employers provide resources to learn more about mental health issues and how to seek help?
30. Was your anonymity protected if you chose to take advantage of mental health or substance abuse treatment resources with previous employers?
31. Do you think that discussing a mental health disorder with previous employers would have negative consequences?
32. Do you think that discussing a physical health issue with previous employers would have negative consequences?
33. Would you have been willing to discuss a mental health issue with your previous co-workers?
34. Would you have been willing to discuss a mental health issue with your direct supervisor(s)?
35. Did you feel that your previous employers took mental health as seriously as physical health?
36. Did you hear of or observe negative consequences for co-workers with mental health issues in your previous workplaces?
37. Would you be willing to bring up a physical health issue with a potential employer in an interview?
38. Why or why not?
39. Would you bring up a mental health issue with a potential employer in an interview?
40. Why or why not?
41. Do you feel that being identified as a person with a mental health issue would hurt your career?
42. Do you think that team members/co-workers would view you more negatively if they knew you suffered from a mental health issue?
43. How willing would you be to share with friends and family that you have a mental illness?
44. Have you observed or experienced an unsupportive or badly handled response to a mental health issue in your current or previous workplace?
45. Have your observations of how another individual who discussed a mental health disorder made you less likely to reveal a mental health issue yourself in your current workplace?
46. Do you have a family history of mental illness?
47. Have you had a mental health disorder in the past?
48. Do you currently have a mental health disorder?
49. If yes, what condition(s) have you been diagnosed with?
50. If maybe, what condition(s) do you believe you have?
51. Have you been diagnosed with a mental health condition by a medical professional?
52. If so, what condition(s) were you diagnosed with?
53. Have you ever sought treatment for a mental health issue from a mental health professional?
54. If you have a mental health issue, do you feel that it interferes with your work when being treated effectively?
55. If you have a mental health issue, do you feel that it interferes with your work when NOT being treated effectively?
56. What is your age?
57. What is your gender?
58. What country do you live in?
59. What US state or territory do you live in?
60. What country do you work in?
61. What US state or territory do you work in?
62. Which of the following best describes your work position?
63. Do you work remotely?

Los motivos por los que he decidido quitar las columnas en rojo son los siguientes:

* Muchas de ellas están relacionadas con el tema de cómo ven los empleados que se llevan las enfermedades mentales en la empresa *(“Is your employer primarily a tech company/organization?”*, por ejemplo, o las relacionadas con la familia)
* Las preguntas de *“¿Por qué no?”* son de texto libre, y aparte de la imposibilidad de interpretarlas para Data Mining, no interesan para este ejercicio.
* Están relacionadas con una empresa anterior (hay aproximadamente 20 preguntas referidas a experiencias con empresas anteriores), y eso no interesa para el tema del ejercicio, que se tiene que centrar solo en la empresa actual.
* La geolocalización de los encuestados no tiene tampoco cabida para el tema principal del ejercicio, así como su puesto de trabajo.

Hay algunas columnas que, a pesar de parecer a primera vista que no tienen relación con el tema principal, puede ser interesante tenerlas en cuenta. Estas columnas son:

* Teletrabajo. Pienso que el hecho de trabajar remotamente puede tener impacto en el ambiente de trabajo (en este caso, si trabajaría esa persona sola o en la oficina), y me gustaría ver si tiene consecuencias en el hecho de pensar que tiene una enfermedad mental.
* Edad y género. Por mera estadística, creo que estas dos columnas pueden ser útiles al final, sobre todo la de edad.
* Número de empleados de la compañía. creo que es interesante saber lo que piensa la persona del ambiente en la compañía dependiendo de los empleados que tenga.

Obtengo la tarjeta de datos en esta fase.

## Pre-procesamiento

#### Tratamiento de datos nulos, blancos y outliers

Aquí se limpia, ordena y se prepara la base de datos para su posterior utilización por el algoritmo. Hay tres puntos básicos que voy a tratar:

1. Conversión del formato de la base de datos

Dado que Kaggle me ha permitido descargar la base de datos en formato csv, no he tenido que convertir la base de datos.

1. Datos en blanco

Los datos en blanco del dataset se han convertido en datos nulos al pasarlo a Python. Estos datos son respuestas de la encuesta que se han dejado en blanco. En la mayoría de los casos, las he decidido interpretar dependiendo del contexto de la pregunta, asignándoles el valor “quizá”, puesto que he considerado que si se deja una pregunta de sí/no en blanco en una encuesta es porque la respuesta sería “quizá”. Sin embargo, en algunos casos he decidido borrar directamente las entradas con datos en blanco, si no eran muchos.

1. Datos nulos, o que se salen de los parámetros

He filtrado toda la tabla dependiendo de la primera columna, *“Are you self-employed?”,* porque si el empleado no pertenece a una empresa, su encuesta es inútil para el ejercicio.

Dado que es una encuesta y la mayoría de respuestas son binarias o están dentro de ciertos grupos cerrados, solo he encontrado un par de “outliers”. Estos son, por ejemplo, los valores “3”, “323” y “99” en el campo de edad. He decidido borrar las entradas enteras de la base de datos, porque es imposible saber la edad real de esas personas.

En el caso de la pregunta *“Have your observations of how another individual who discussed a mental health disorder made you less likely to reveal a mental health issue yourself in your current workplace?”,* los datos nulos los he tratado como “no aplica”, porque si no han contestado es probable que sea porque no han observado nada aplicable a esa pregunta.

1. Datos mal escritos

En esta sección explicaré cómo he tenido que modificar la tabla para poder pasarla por los scripts de Python.

He tenido varios problemas:

* Los nombres de las columnas estaban entre comillas… Pero no todos. He solucionado este problema quitando las comillas dobles a toda la tabla.
* En los nombres de algunas columnas había comas. Por ejemplo, en la pregunta *“If you have a mental health issue, do you feel that it interferes with your work when being treated effectively?”* hay una coma, lo que genera una nueva columna al ser el csv un archivo delimitado por comas. Para ello he decidido reemplazar en uno de los scripts los nombres de esas columnas por el mismo nombre, pero sin coma.

#### Scripts en Python

Para poder realizar todo lo anteriormente mencionado, he escrito unos scripts en Python que hay que ejecutar para adecuar la base de datos para el siguiente paso. Estos son los archivos:

* Archivo quitarColumnas

Este archivo contiene las columnas que se van a quitar de la base de datos. Es el primero que debe ejecutarse.

Aquí soluciono también el problema del carácter extraño ‘Â’, que no fue muy complicado ya que la columna que lo contenía, igualmente, había que borrarla en el proceso de selección. Para esto, simplemente he copiado y pegado el nombre de la columna en el archivo que las borra.

* Archivo quitarComillas

El archivo csv tenía dentro varios errores de escritura en los identificadores de las columnas: un carácter extraño (‘Â’) y dos comillas dobles ‘ “ ‘ al final de una de las cadenas. Para solucionar esto, he escrito un script en Python para quitar todas las comillas de los identificadores de las columnas.

En este script también se suprimen las comas dentro de los nombres de las columnas que generaban más columnas de la cuenta.

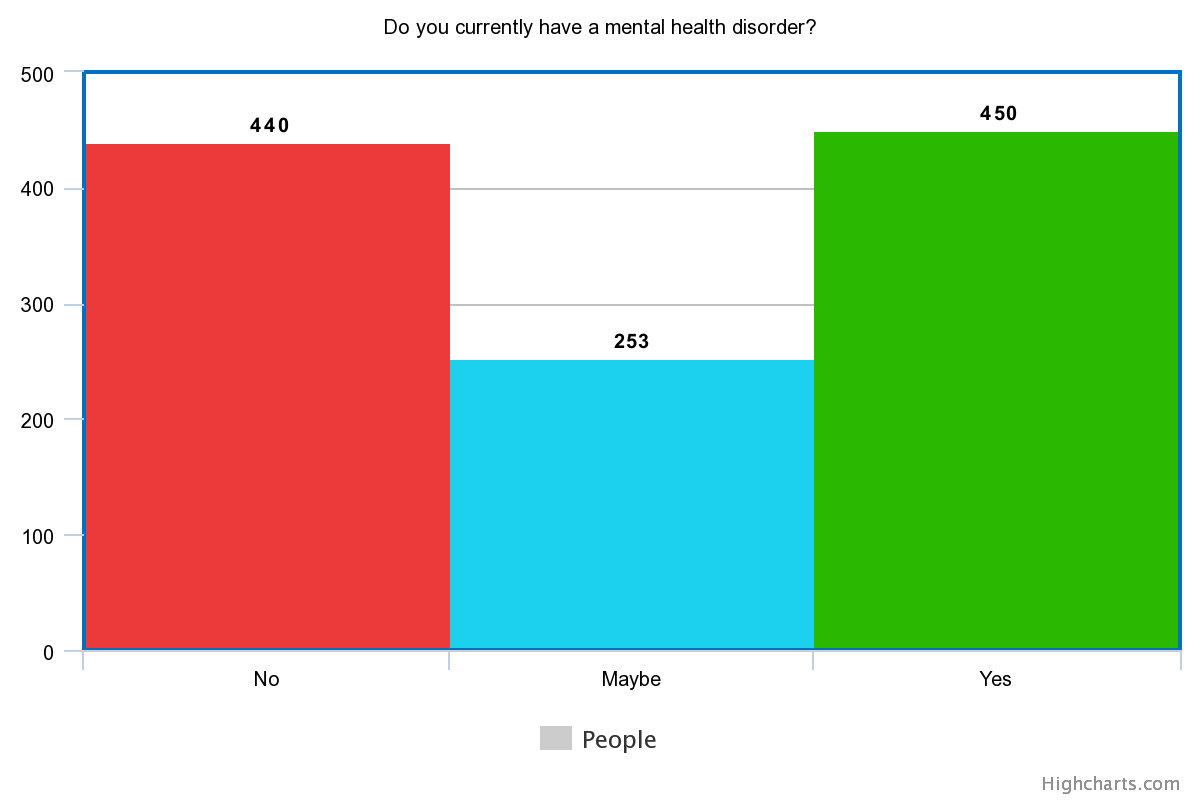
* Archivo limpiarFilas

Este archivo quita de la base de datos las filas con los outliers y los datos nulos que no se han podido rellenar.

* Archivo reemplazarColumnas

Este archivo es muy importante: es el que contiene todo el proceso de adecuación de la base de datos. En él, se pasan a números y grupos los datos de la encuesta, para posterior interpretación por el algoritmo de Data Mining.

Tras el pre-procesado, la columna *label* se ha quedado así:



## Data Mining

Para realizar esta fase he decidido hacer uso de la página BigML. Esta herramienta me permite subir mi base de datos limpia y transformada, y evaluar diferentes modelos en ella.

Lo primero que hice fue cargar el dataset y apartar los datos para entrenamiento y testing, en proporción de 90% / 10%. Ahora, para los datos de entrenamiento, elegí un modelo supervisado PORQUEEEEEE……………………

#### Modelos y evaluaciones

He realizado cuatro modelos diferentes para esta iteración, a fin de encontrar el que mejor resultados diera:

1. Árbol de decisión

El árbol mostró la pregunta más decisiva para la clasificación: *“If you have a mental health issue do you feel that it interferes with your work when NOT being treated effectively?”,* seguida por *“If you have a mental health issue do you feel that it interferes with your work when being treated effectively?”.* Estas dos preguntas son importantes, porque muestran la importancia que la gente da a las enfermedades mentales en su trabajo, y la respuesta que den podría denotar la experiencia que tienen con enfermedades mentales.

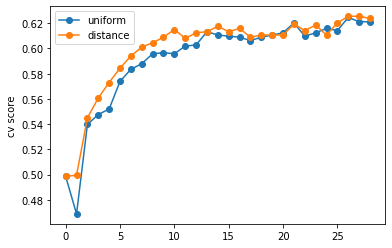
Al evaluar el árbol, he obtenido los siguientes resultados:

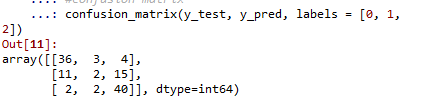


Muestran un 51.3% de precisión del modelo escogido.

1. Algoritmo KNN

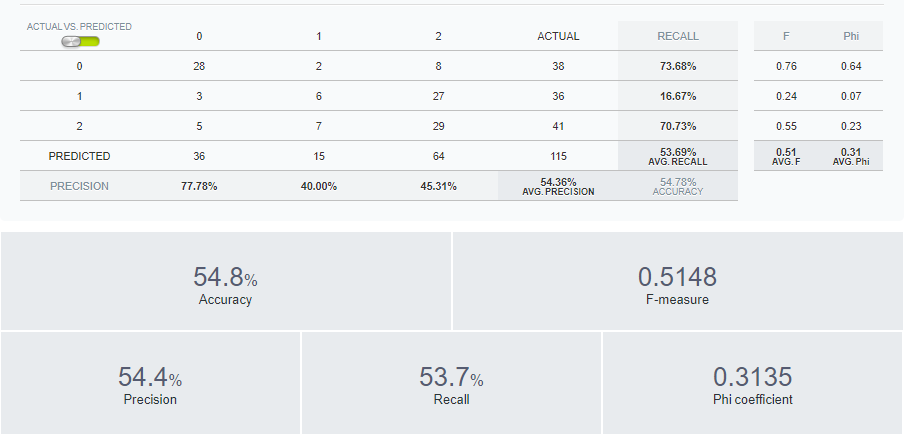
También realicé el algoritmo KNN con este conjunto de datos. Estos son el gráfico y la matriz de confusión que salieron:





1. Random forest

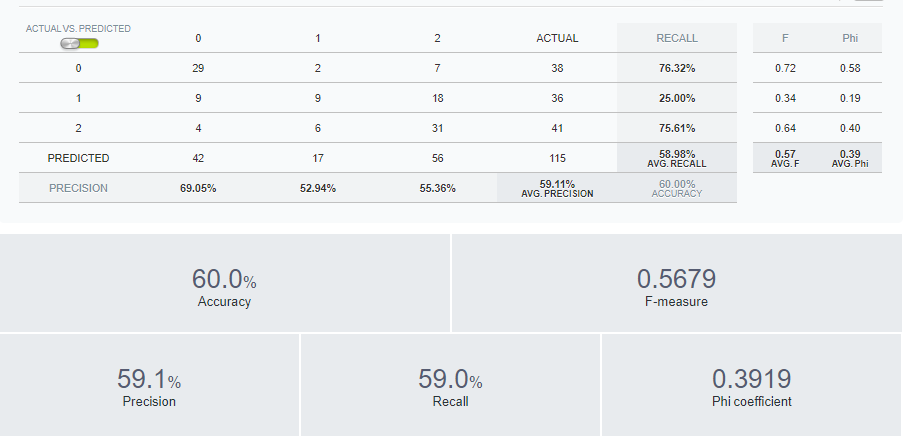
Los resultados del random forest fueron ligeramente mejores que los del árbol de decisión, con una precisión del 54.8%.



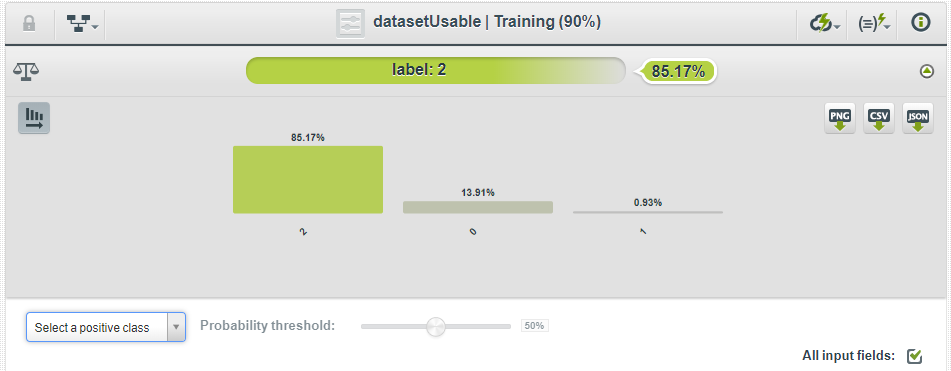
En este modelo se puede observar una reducción general de los fallos en la predicción de la clase 0 (el “No”), pero un aumento considerable de los fallos al evaluar la clase 2 (“Sí”) como clase 1 (“Quizás”).

1. Regresión logística

BigML tiene la opción de evaluar el dataset con todos los algoritmos que ofrece, y mostrar el que ha tenido mejores resultados. El mejor modelo para mi dataset fue el de regresión logística:



#### Testing

He pasado un test con las respuestas que yo daría a las preguntas, y me ha salido un “Sí” con un 85% de precisión, y sí, ha acertado.

#### Scripts en Python

Para esta primera iteración he modificado un script en Python con el algoritmo KNN para adaptarlo a mi dataset. Este script se llama knn.

#### Conclusiones de la iteración

Al ver estos resultados me di cuenta de lo siguiente: el modelo ha fallado demasiado prediciendo como “quizás” las personas respondieron que “Sí” creen que tienen una enfermedad mental (clase 2 predicha como clase 1). Esto me hizo reflexionar sobre la propia pregunta de la encuesta: si a una persona se le pregunta “¿Crees que tienes una enfermedad mental?” y responde “quizás”, es que siente que la tiene, pero no está completamente segura. Considero entonces que el algoritmo, al predecir los “sí” como “quizás”, no se ha equivocado del todo, porque una persona completamente sana siempre responderá rotundamente “no”, pero si ve que tiene algunos síntomas y no está segura de ello dirá que “quizás”.

Decidí entonces hacer otra iteración del KDD, juntando los valores de “Sí” y “Quizás” en un solo valor: el 1, que ahora representaría a las personas que, mentalmente, no se encuentran sanas.

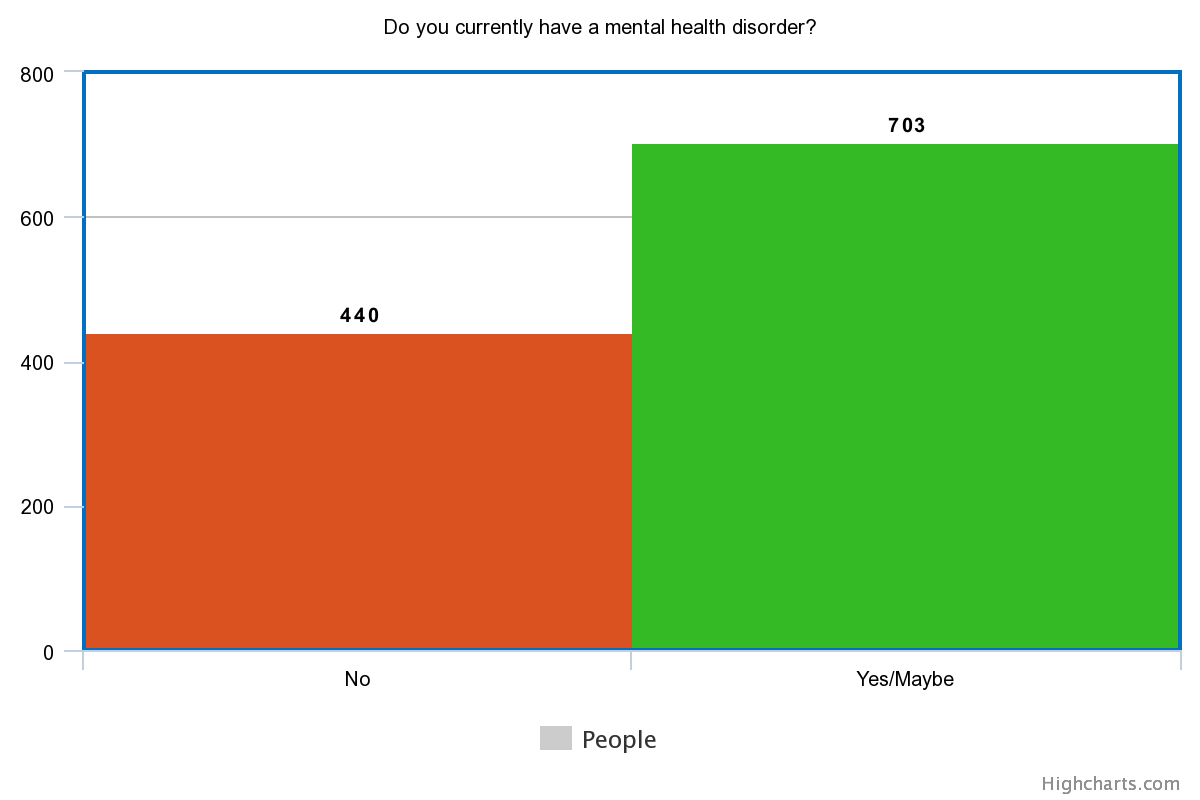
# Segunda iteración

## Pre-Procesado

El cambio propuesto para esta segunda iteración es el siguiente:

* Juntar los valores 1 y 2 de la label, correspondientes a las respuestas “Sí” y “Quizás” de la encuesta.

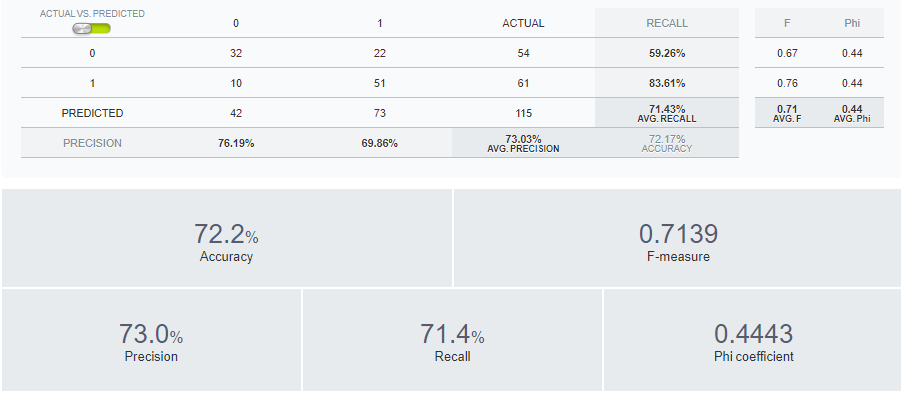
Tras el cambio, los valores de la columna label han quedado así:



## Data Mining

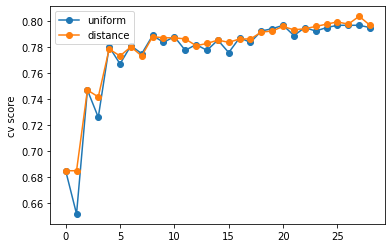
#### Modelos y evaluaciones

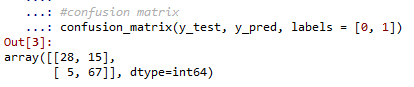
1. Árbol de decisión



1. Algoritmo KNN

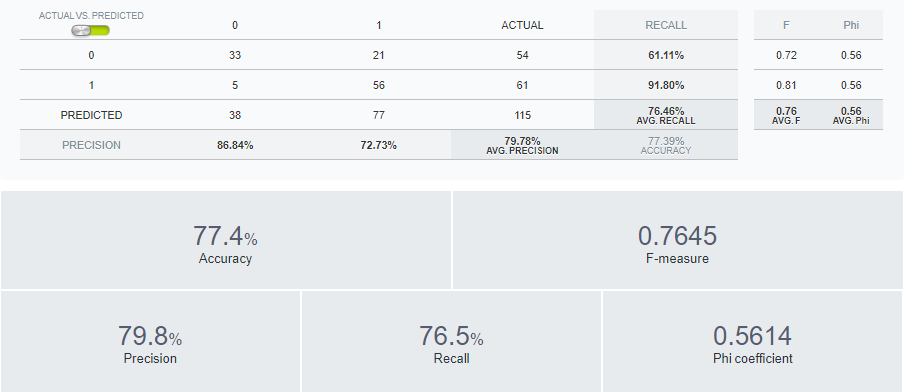
A continuación, se muestran la gráfica y la matriz de confusión del algoritmo KNN con dos clases:





Podemos observar que el algoritmo tiene problemas clasificando los valores “No”, puesto que ha clasificado 15 valores “No” como “Sí”.

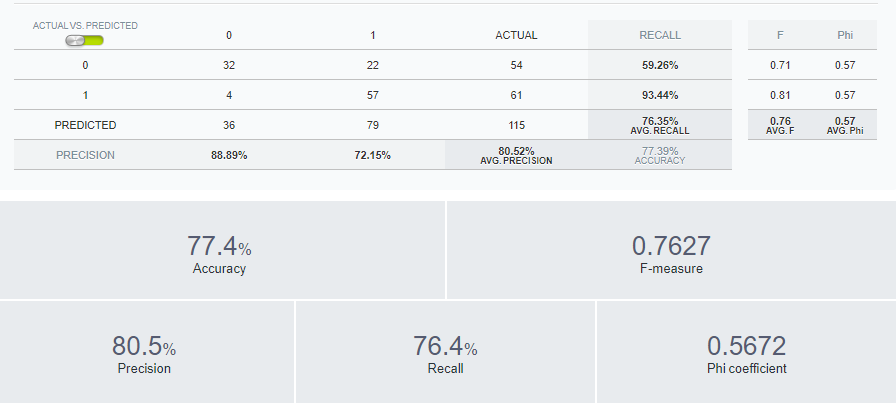
1. Random forest



1. Regresión logística

De nuevo, he vuelto a usar la herramienta OptiML para ver qué modelo con qué parámetros tiene un mejor resultado, y de nuevo ha sido la regresión logística.

Sin embargo, los resultados que da son muy parecidos a los del random forest, con tan solo UNA buena predicción más de cada clase.



#### Testing

HACER EL TESTING PARA ESTA ITERACION

#### Conclusiones de la iteración

Aunque generalmente ha dado mejores resultados, se han debido a que, en vez de tres clases, ahora tengo dos, y considero que una proporción de acierto de poco más de 2/3 es bastante mejorable. Sin embargo, como los resultados tampoco son malos del todo, he decidido darle otra oportunidad y hacer una tercera iteración, esta vez, pasando la label a las preguntas que vienen después de esta:

* *If yes, what condition(s) have you been diagnosed with?*
* *If maybe, what condition(s) do you believe you have?*

Ahora, el objetivo será predecir, si cree que tiene o no una enfermedad mental, y en el caso de que tenga, averiguar cuál.

Para ello, voy a juntar las dos columnas anteriormente mencionadas. Las respuestas de estas dos columnas dependen directamente de las dadas en la de *“Do you currently have a mental health disorder?”*: si han dicho que “No” tienen una enfermedad mental, en la columna resultado de sumar esas dos habrá un valor nulo. Si han dicho que “Sí” o “Quizás”, en la columna resultado saldrá la enfermedad que tienen/creen tener.

# Tercera iteración

## Pre-Procesado

En el script reemplazarColumnas de esta iteración he dividido las enfermedades mentales en 9 grupos:

* Anxiety disorder - 0
* Attention Deficit Hyperactivity Disorder - 1
* Autism Spectrum Disorder – 2
* Eating disorder - 3
* Mood disorder - 4
* Obsessive-Compulsive Disorder - 5
* Personality Disorder - 6
* Post-traumatic Stress Disorder - 7
* Psychotic Disorder - 8
* Otros (aquí he puesto las enfermedades mentales que solo había puesto una persona, puesto que no merecía hacer un grupo solo para esa) – 9

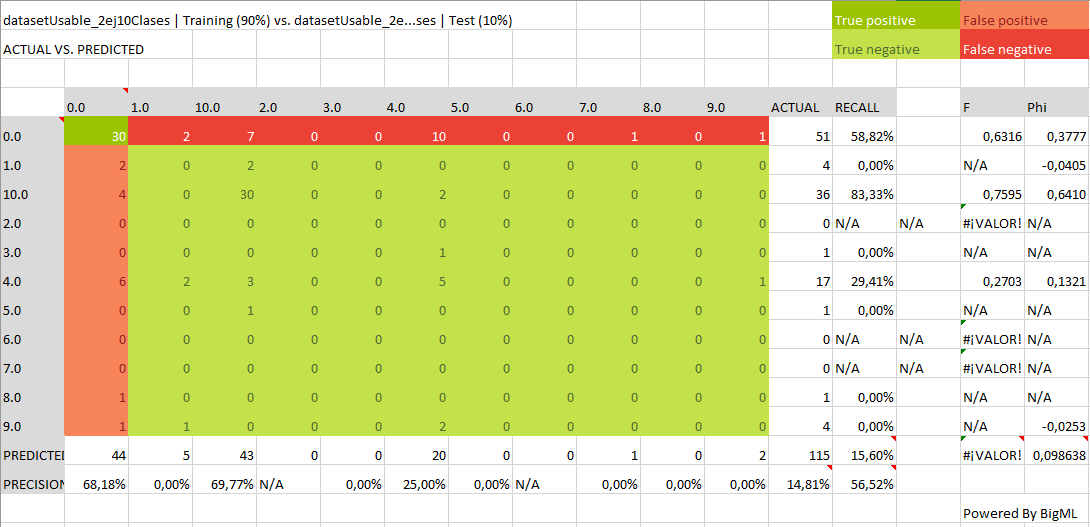
Los que no tienen ninguna los he puesto en el grupo 10.

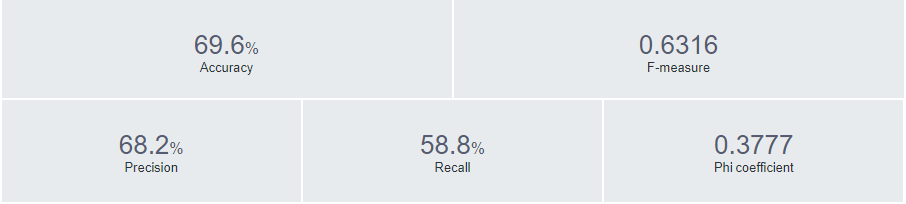
Como no soy una experta en este tema, he hecho los grupos después de una breve investigación en internet sobre las distintas enfermedades mentales. Además, como son respuestas múltiples y para reducir la complejidad de los grupos, he agrupado las respuestas que tienen varias enfermedades según la primera. Por ejemplo, “'Mood Disorder (Depression, Bipolar Disorder, etc)|Personality Disorder (Borderline, Antisocial, Paranoid, etc)',” la he puesto en “Mood disorder”. Esta aproximación me ha parecido correcta porque, en una lista, siempre se tiende a decir lo más importante primero, así que considero que esa persona tiene más parte de “Mood disorder” que de “Personality disorder”.

## Data Mining

#### Modelos y evaluaciones

1. Árbol de decisión





Este nuevo enfoque tiene una pobre precisión del 69.6%.

#### Scripts

Para hacer la clasificación de los grupos, he modificado el script reemplazarColumnas, creando a partir de él uno nuevo: reemplazarColumnas2Ej, donde he borrado la anterior columna label *(“Do you currently have a mental disorder?”)* y he renombrado las dos columnas unidas como la nueva label.

He creado una versión para la segunda ejecución de todos los scripts.

#### Conclusiones de la iteración

Nada más ver la matriz de confusión del árbol de decisión me he dado cuenta de que esta iteración no ha sido del todo buena idea. Hay demasiadas clases, y muchas de ellas apenas tienen datos. He decidido no sacar ningún modelo más, y reducir drásticamente los grupos de las enfermedades mentales a 3:

* Ansiedad/Desorden del humor
* Otra enfermedad
* Ninguna enfermedad

# Cuarta iteración

## Pre-Procesado

He agrupado las enfermedades mentales en los tres grupos mencionados en la conclusión de la iteración anterior:

* Ansiedad/Desorden del humor - 0
* Otra enfermedad – 1

Los que no tienen ninguna enfermedad los he agrupado en el número 2.

## Data Mining

#### Modelos y evaluaciones