1. Descripción de las variables y valores estadísticos (mínimo, máximo, media, desviación, mediana, etc.). Estudia qué valores estadísticos son los convenientes según el tipo de variable y procede en consecuencia.

   Los valores no son numéricos por lo que no es posible calcular valores estadísticos. Lo que sí se puede hacer es utilizar histogramas para visualizarlos.

2. Describe y realiza modificaciones en la base datos si lo consideras necesario. Por ejemplo, qué harías con valores nominales, si los hubiera.

   Hemos puesto los valores en un *one hot encoder* para que puedan ser interpretados más claramente. Una vez creado el *one hot encoder*, los valores están en forma de vector y podemos analizarlos para su clasificación en las dos clases existentes: venenoso o no venenoso.

3. Estudia si es necesario normalizar los datos y cómo lo harías. Procede a modificar la base de datos (normalizar) si lo consideras necesario.

   Como los valores no son numéricos no es posible normalizarlos. Una vez que formen parte de un *encoder*, sus valores serán 0 o 1.

4. Detección de valores extremos (outliers) y descripción de qué harías en cada caso.

   No hay *outliers* en el conjunto de datos. Hay solamente colecciones de características posibles en cada observación, una vez que los datos son categóricos.

5. Detección de valores perdidos (missing values) y descripción de cómo actuarías para solventar el problema.

   En el conjunto de datos se especifica que los datos no disponibles tendrán un "?", por lo que el algoritmo introduce valores iguales a "?" en la base de datos. En la especificación se dice que faltan 2480 datos, que son los datos perdidos que ha encontrado el algoritmo.

   Es importante considerar algunas cosas cuanto a los datos:
   - Los valores perdidos pertenecen todos a la característica "*stalk-root*".
   - De las 8124 observaciones, falta esta característica en 2480 de ellas, que supone un 30% de las observaciones.

   Hay diferentes mecanismos para resolver el problema de los datos faltantes: excluir la muestra o variable o crear un modelo que estime el valor faltante.
   Podemos testar la correlación que hay entre la variable y las clases:

	Stalk-root_b	Stalk-root_c	Stalk-root_e	Stalk-root_r	Stalk-root_?
Class_e	-0.1783	0.2185	0.2032	0.1500	-0.3021
Class_p	0.1783	-0.2185	-0.2032	-0.1500	0.3021

Tabla 1. Correlación de la variable 'stalk-root' con las clases

Como se puede ver en la Tabla 1 la correlación no es relevante, por lo que podemos eliminar la variable.

6. Buscar correlaciones entre:
    a. las variables predictoras, lo que permitirá ver si hay variables redundantes.

    Si se utiliza un límite de 0.8 existen muchas variables correlacionadas:

Correlación entre A y B	Variable A	Variable B
0.9550972066799881	gill-attachment_a	stalk-surface-above-ring_s
0.9550972066799881	gill-attachment_a	stalk-color-above-ring_o
0.935237345539456	gill-attachment_a	stalk-color-below-ring_c
0.9550972066799881	gill-attachment_f	stalk-surface-above-ring_s
0.9550972066799881	gill-attachment_f	stalk-color-above-ring_o
0.935237345539456	gill-attachment_f	stalk-color-below-ring_c
0.8055660308028565	gill-color_b	ring-number_e
0.8508972028756072	stalk-surface-above-ring_k	stalk-shape_t
0.8508972028756072	stalk-surface-above-ring_s	stalk-shape_e
0.9550972066799881	stalk-color-above-ring_o	gill-attachment_a
0.9550972066799881	stalk-color-above-ring_o	gill-attachment_d
0.9793016123563326	stalk-color-above-ring_o	stalk-color-below-ring_c
0.9550972066799881	stalk-color-below-ring_o	gill-attachment_a
0.9550972066799881	stalk-color-below-ring_o	gill-attachment_d
0.9793016123563326	stalk-color-below-ring_o	stalk-color-below-ring_c
0.935237345539456	veil-color_w	gill-attachment_a
0.935237345539456	veil-color_w	gill-attachment_d
0.9793016123563326	veil-color_w	stalk-surface-above-ring_s
0.9793016123563326	veil-color_w	stalk-color-above-ring_o
0.9689591161987591	ring-type_o	stalk-color-below-ring_e
0.9689591161987591	ring-type_t	stalk-color-below-ring_p
0.8689269690228176	ring-number_l	veil-color_w
0.8689269690228176	spore-print-color_h	veil-type_p
0.8055660308028565	spore-print-color_w	gill-spacing_d

Tabla 2. Correlación entre variables con límite 0.8

Usando un límite de 0.9 se obtiene:

Correlación entre A y B	Variable A	Variable B
0.9550972066799881	gill-attachment_a	stalk-surface-above-ring_s
0.9550972066799881	gill-attachment_a	stalk-color-above-ring_o
0.935237345539456	gill-attachment_a	stalk-color-below-ring_c
0.9550972066799881	gill-attachment_f	stalk-surface-above-ring_s
0.9550972066799881	gill-attachment_f	stalk-color-above-ring_o
0.935237345539456	gill-attachment_f	stalk-color-below-ring_c
0.9550972066799881	stalk-color-above-ring_o	gill-attachment_a
0.9550972066799881	stalk-color-above-ring_o	gill-attachment_d
0.9793016123563326	stalk-color-above-ring_o	stalk-color-below-ring_c
0.9550972066799881	stalk-color-below-ring_o	gill-attachment_a
0.9550972066799881	stalk-color-below-ring_o	gill-attachment_d
0.9793016123563326	stalk-color-below-ring_o	stalk-color-below-ring_c
0.935237345539456	veil-color_w	gill-attachment_a
0.935237345539456	veil-color_w	gill-attachment_d
0.9793016123563326	veil-color_w	stalk-surface-above-ring_s
0.9793016123563326	veil-color_w	stalk-color-above-ring_o
0.9689591161987591	ring-type_o	stalk-color-below-ring_e
0.9689591161987591	ring-type_t	stalk-color-below-ring_p

Tabla 3. Correlación entre variables con límite de 0.9

Así que considerando una correlación alta como 0.9, podemos decir que todas las variables en la Tabla 3 están muy correlacionadas y son redundantes.

b. variables predictoras y la clase (target).
   No hay correlación relevante entre las variables predictoras y la clase. La mayor correlación que existe es entre la variación 'n' de la variable "odor", que tiene 0.78.

7. Detecta, si hubiera, falsos predictores.

   Como no hay una variable con correlación fuerte con la clase, no hay falsos predictores.

8. Estudia si fuera conveniente segmentar alguna de las variables.

   Dado que las variables no son numéricas no es posible segmentarlas.

9. Estudia si fuera conveniente crear nuevas variables sintéticas basada en las variables originales.

   De acuerdo con el dueño de los datos, hay reglas que se pueden usar para predecir si la seta es comestible o no, como por ejemplo si crece en hojas caídas y es de color blanco, se puede predecir con 100% de precisión el tipo de seta. Si esta agrupado y tiene su capa de color blanca, también puede decir con 100% de certeza que es comestible. Hay otras reglas también, pero no identifican con un 100% de certeza el tipo de la seta, como:

- si no huele a almendra, anís o no tiene olor, con 98.52% de precisión puede ser comestible.
- si la espora es verde, con 99.41% de precisión puede ser comestible.
- si no huele a nada, su tallo bajo en anillo es escamoso y sobre el anillo no es marrón, con 99.90% de certeza no es comestible.

Con esta información es posible crear variables sintéticas que ayuden a precedir la clase de una seta con mayor facilidad.