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, así que no hay como calcular valores estadísticos. Podemos ponerlos en histogramas para verlos, pero no hay una análisis estadística sobre ellos.

1. 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 sean mejor interpretados. Una vez puesto, tenemos los valores como vectores y podemos hacer análisis sobre ellos para intentar clasificarlos entre los clases existentes: venenosos o no.

1. 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 hay como normalizarlos, pues una vez que estarán involucrados en un *encoder*, los valores serán 0 o 1.

1. 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.

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

En el conjunto de datos, es especificado que los datos tendrán un “?” cuando no estén disponibles, así que el algoritmo procura por valores iguales a “?” en la base de datos. En la especificación, esta dicho que hace falta de 2480 datos, que es exactamente lo que el algoritmo ha encontrado.

Es importante considerar algunas cosas cuanto a los datos:

* Hay valores faltando solamente en una característica llamada “*stalk-root*”;
* De las 8124 observaciones, hace falta este dato en 2480 de ellas, que es 30% de las observaciones;

Hay posibles mecanismos para actuar sobre datos indisponibles, que son: excluirlos las amuestras o la variable, crear un modelo que prediga valores faltosos o imputación.

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 classes

Como visto en la Tabla 1, la correlación no es relevante, así que la variable puede ser sacada sin problemas.

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

Si, hay muchas correlaciones. Si usáis un limite de 0.8, hay:

|  |  |  |
| --- | --- | --- |
| **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 limite de 0.8

Si usáis un limite de 0.9, tenéis:

|  |  |  |
| --- | --- | --- |
| **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 limite de 0.9

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

* 1. variables predictoras y la clase (target).

No hay correlación relevante entre las variables predictoras y la clase. La mayor correlación que hay es entre la variación ‘n’ de la variable “odor”, que tiene 0.78.

1. Detecta, si hubiera, falsos predictores.

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

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

Como las variables no están en un rango, no hay como segmentarlas. No son variables numéricas, así que o es algo, o no es, no hay mucha interpretación sobre ellas. Por otro lado, si fuera una variable numérica como edad, se podrían crear rangos de edad y segmentarlas.

1. 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 con una alta precisión si la seta es comestible o no, como por ejemplo si tiene su hábitat en hojas y si es blanco, se puede predecir con 100% de precisión el tipo de seta. Si esta agrupado y tiene su capa de color blanca, también se dice con 100% de certeza que es comestible. Hay otras reglas también, pero no identifican con 100% de certeza el tipo de la seta, como:

* si no huele a almendra, anís o no tiene olor, con 98.52% de precisión se predice que es comestible;
* si la espora es verde, con 99.41% de certeza se predice que es comestible;
* si no huele a nada, su tallo bajo en anillo es escamoso y sobre el anillo no es marrón, con 99.90% de precisión se dice que es comestible.

Así que con esa información es posible decir qué creando una variable sintética con estos datos, se puede decir la clase con más facilidad.