UNIVERSIDAD AUTONOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

“AVANCE I PROYECTO INTEGRADOR”

Equipo 11

Alfonso Llanos Morales 1887939

Fernando Isaí González Castillo 1819011

Daniela Monserrat Sanabria Martínez 1563836

Materia: Minería de datos.

Grupo: 002

Frecuencia: miércoles 19:00-22:00

Maestra: Mayra Cristina Berrones Reyes

28 de octubre 2020

1. **Título de la base de datos.**

   a) **Nombre de la base de datos:** Mushroom Classification

   b) **Url de la página:** https://www.kaggle.com/uciml/mushroom-classification

2. **Descripción de los datos.**

   a) **¿Qué tipos de datos son?**

   Nuestros datos están presentados en una tabla donde las columnas representan las características de los hongos y en donde cada fila es un hongo observado.

   b) **Descripción de las columnas.**

   Columnas:

1) **Clase (class):** Es la variable de respuesta de nuestra base de datos, nos dice si el hongo es venenoso (p) o comestible (e). Es una variable cualitativa binaria.

2) **Forma del sombrero (cap-shape):** Representa la parte superior del hongo puede tener forma de campana (b), cónica (c), convexa (x), plana (f), mamelonado (k) o hundido (s). Es una variable cualitativa nominal.

3) **Superficie del sombrero (cap-surface):** Variable cualitativa nominal que indica la superficie del sombrero del hongo el cual puede ser fibroso (f), asurcada o acanalada (g), escamoso (y) o liso (s).

4) **Color del sombrero (cap-color):** Variable cualitativa nominal que indica el color del sombrero del hongo, podría ser marrón (n), ante (b), canela (c), gris (g), verde (r), rosa (p), violeta (u), rojo (e), blanco (w) o amarillo (y).

5) **Magulladura (bruises):** Variable cualitativa binaria que indica si el hongo tiene magulladuras generalmente de tono azulado verdoso (t) o no las tiene (f).

6) **Olor (odor):** Representa a un aroma similar al que se percibe del hongo, este puede ser: almendra (a), anís (l), creosota o aceite (c), pescado (y), sucio (f), mohoso (m), ninguno (n), acre (p) o picante (s). Es una variable cualitativa nominal.

7) **Forma de las láminas (gill-attachment):** Variable cualitativa nominal que representa la forma de las láminas del hongo, las láminas son las estructuras laminares existentes bajo el sombrero de algunas setas. Esta variable puede tener las siguientes respuestas: adherida (a), descendente (d), libre (f) o con muescas (n).

8) **Espaciamiento de las láminas (gill-spacing):** Representa que tan separadas se encuentran las láminas del hongo, estas pueden ser distantes (d), cercanas (c) o apretadas (w), es una variable cualitativa ordinal.

9) **Tamaño de las láminas (gill-size):** Representa la dimensión de las láminas de los hongos, en donde la variable es cualitativa binaria y su posible respuesta es ancho (b) o estrecho (n).

10) **Color de las láminas (gill-color):** Representa el color de las láminas de los hongos, en donde la variable es cualitativa nominal y su posible valor es negro (k), marrón (n), ante (b), chocolate (h), gris (g), verde (r), naranja (o), rosa (p), violeta (u), rojo (e), blanco (w) o amarillo (y).

11) **Forma del tallo (stalk-shape)**: Representa si la forma del tallo se está agrandado (e) o estrechando (t), la variable es cualitativa binaria.

12) **Raíz del tallo (stalk-root):** Representa la forma de la raíz del tallo del hongo, en donde la variable es cualitativa nominal y su posible respuesta es de forma igual (e), bulboso(b), forma de garrote (c), con copa (u), con rizomorfos o cordones miceliales (z), enraizado (r) o faltante (?).

13) **Superficie del tallo sobre el anillo (stalk-surface-above-ring):** Representa como es la superficie del tallo arriba de su anillo, la variable es cualitativa nominal y esta puede ser fibrosa (f), escamosa (y), sedosa (k) o lisa (s).

14) **Superficie del tallo debajo el anillo (stalk-surface-below-ring):** Representa como es la superficie del tallo abajo de su anillo, la variable es cualitativa nominal y esta puede ser fibrosa (f), escamosa (y), sedosa (k) o lisa (s).

15) **Color del tallo sobre el anillo (stalk-color-above-ring):** Representa el color del tallo arriba de su anillo, su variable es cualitativa nominal y su posible respuesta es marrón (n), ante (b), canela (c), gris (g), naranja (o), rosa (p), rojo (e), blanco (w) o amarillo (y).

16) **Color del tallo debajo del anillo (stalk-color-below-ring):** Representa el color del tallo debajo del anillo, su variable es cualitativa y su posible valor es marrón (n), ante (b), canela (c), gris (g), naranja (o), rosa (p), rojo (e), blanco (w) o amarillo (y).

17) **Tipo de velo (veil-type):** Esta variable representa el tipo de velo del hongo, el velo es una estructura que envuelve todas o la mayor parte de las láminas del hongo cuando aún es inmaduro, cuando el hongo crece quedan rastros del velo que son identificables, está variable es cualitativa binaria y puede tener los valores de parcial (p) o universal (u).

18) **Color del velo (veil-color):** Representa el color del velo, es variable cualitativa nominal y toma los valores de marrón (n), naranja (o), blanco (w) o amarillo (y).

19) **Número de anillos (ring-number):** Esta variable cualitativa ordinal que indica el número de anillos con los que cuenta el hongo, los anillos de un hongo son los vestigios del velo que se quedan específicamente en el tallo del hongo después de que el hongo madura. Esta variable pude tener los valores de ninguno (n), uno (o) o dos (t).

20) **Tipo de anillo (ring-type):** Variable cualitativa nominal que representa la forma del anillo de los hongos, estos pueden ser en forma de telaraña o cortina (c), evanescente o que desaparece (e), llamarada (f), grande (l), ninguno (n), colgante (p), ascendente (s) o fugaz (z).

21) **Color de las esporas (spore-print-color):** Representa el color de las esporas del hongo, esta variable cualitativa nominal y puede tener los valores de negras (k), marrones (n), ante (b), chocolate (h), verdes (g), naranjas (o), violetas (u), blancas (w) o amarillas (y).

22) **Población (population):** Representa una cantidad aproximada de hongos del que se desea analizar que se encuentran en la zona, esta variable cualitativa nominal puede tener los valores de abundantes (a), agrupados (c), numerosos (n), dispersos (s), varios (v) o solitario (y).

23) **Hábitat (habitat):** Variable cualitativa nominal que representa tipo de hábitat en donde se encuentra el hongo y estos son en pastos (g), hojas (l), prados (m), caminos (p), zona urbana (u), residuos (w) o bosques (d).

3. **Justificación del uso de datos.**

   a) **¿Cuáles fueron las características que les llamó atención de los datos? ¿Qué les hizo querer trabajar con ellos?**

      La base de datos es muy descriptiva referente a las características de los hongos, contamos con columnas que nos muestran desde su forma, tamaño y color hasta el tipo de zona donde es más común encontrarlo. Al estar los datos recopilados de esta manera consideramos que sería más sencillo trabajar con ella y también nos daba una idea más clara de que tipo de técnica de minería utilizar.

   b) **¿Qué beneficios encuentran en trabajar con estos datos?**

      El tipo de datos con los que vamos a trabajar es más intuitivo con relación al tipo de técnica de minería que podemos utilizar, también nos pareció importante el formato de texto para describir las variables, que es en forma categórica, ya que todas ellas están representadas por una letra y creemos que esto nos facilitará las operaciones.

4. **Planteamiento del problema.**

   a) **Tomando en cuenta las características utilizadas en la tarea de análisis de bases de datos, elabora una problemática que te gustaría mejorar al finalizar su proyecto.**

      En la actualidad solo se han clasificado un aproximado del 10% de los 10 millones de especies de hongos que existen en el mundo, lamentablemente aún no existe un programa de clasificación estandarizado que ayude a las áreas de medicina, gastronomía y biología basado en las características físicas de los mismos y no se conoce con certeza cuales son los aspectos más relacionados a su toxicidad.

5. **Objetivo Final.**

   a) **Explica a detalle cual es el objetivo principal (y secundarios en caso de existir) para trabajar con este tipo de datos.**

      Saber cuáles son las principales características que determinan si un hongo es venenoso o comestible, así como la probabilidad de que un hongo en específico sea venenoso solamente describiendo sus atributos y detectarlo en el menor tiempo posible.

6. **Planeación de la herramienta a utilizar.**

   a) **Observando el tipo de datos que se tiene, describir cual es el tipo de técnica que se planea utilizar y dar una explicación concisa de por qué se va a trabajar con esa técnica.**

   1.- **Árboles de decisión:** Primero obtendremos las frecuencias de cada característica con respecto a la respuesta y calcularemos las probabilidades, en base a estas probabilidades construiremos el árbol de decisión donde los aspectos que influyen más en la respuesta se encontrarán en los nodos superiores e irá descendiendo conforme estás características sean menos relevantes y al final nos dará como respuesta si el hongo es venenoso o no. Trabajaremos con esta técnica ya que nos ayudará a predecir si el hongo es venenoso o no de acuerdo con el conjunto de decisiones tomadas.

   2.- **Visualización:** Esta técnica será de utilidad al momento de interpretar los resultados a lo largo del programa, ya que utilizaremos gráficas, tablas, diagramas, entre otros gráficos al momento de llegar a un resultado y al momento de presentarlos. Serán de utilidad debido a que nos brindarán un mayor entendimiento de los resultados.

   3.- **Clustering:** Esta técnica podría ser un complemento de los resultados obtenidos con el árbol de decisión o podría darnos una perspectiva diferente respecto a la relación existente entre los datos, ya que en el árbol de decisión nos enfocaríamos en la probabilidad y en el clustering veríamos que características comparten los hongos venenosos entre sí y de igual forma con los comestibles, además de identificar las características que hacen diferente a los venenosos de los comestibles.

   b) **Si ya se tiene observado algún algoritmo o herramienta dentro de esta técnica seleccionada, mencioné cual es, y porque se desea trabajar con ella.**

   - Podríamos utilizar el Gini como medida de impureza y observar que tan mezcladas están las clases en cada nodo de nuestro árbol de decisión.

   - También se podría implementar el método de bosques aleatorios y comprobar si los resultados de esta técnica son más útiles que los obtenidos con el árbol de decisión convencional.

   - Se está considerando utilizar la función *sklearn.tree.DecisionTreeClassifier()* de la librería *sklearn.tree* para la construcción del árbol.