UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA INFORMÁTICA

Laboratorio 1 - Análisis Estadístico

Integrantes:	Maximiliano Arévalo
	Benjamín Muñoz
Curso:	Análisis de Datos
Sección	A-1
Profesor:	Max Chacón Pacheco

25 de Mayo de 2020

Tabla de contenidos

1. Introducción

Este trabajo tiene como objetivo conocer y comprender la base de datos escogida, ya que como se utilizará a lo largo del semestre, es importante comprender que significa cada elemento de esta y la relación que tienen entre sí.

Se utiliza la base de datos 'Soybean', en la cual se buscará un primer conocimiento de los datos y como estos se relacionan mediante un análisis estadístico. Como objetivos específicos se encuentra la representación de los datos mediante gráficos y distintas pruebas de hipótesis que se pueden realizar para comprobar la independencia entre las variables.

El presente informe inicialmente da a conocer la explicación del problema, explicando la base de datos junto con sus clases y variables. Luego se presenta el análisis estadístico, en el cual se explican los procedimientos realizados comenzando con la presentación de los gráficos obtenidos a partir de los datos junto a un breve análisis de estos. También se muestran las pruebas estadísticas que permitirán dar el análisis estadístico como tal. Finalmente se presentará la conclusión, en la cual se hará una retrospectiva del trabajo, analizando el cumplimiento de los objetivos y resumiendo los logros del trabajo.

Es importante señalar que la base de datos a utilizar corresponde a 'Soybean (Large)' la cual se obtuvo del 'UCI Machine Learning Repository' Michalski and Chilausky (1980), y que para el análisis estadístico se utilizará el lenguaje de programación R.

2. Descripción del problema

El cultivo de la semilla de soya es importante para la agronomía, especialmente en países como Estados Unidos, Brasil, Argentina y China Intagri (2016) debido a su alto valor de proteína y aceites, y también por su versatilidad al momento de hacer alimentos. A su vez, la producción de soya ha crecido bastante junto a su importación los últimos años, por lo que su producción también apoya a la economía de sus productores.

Esta semilla sin embargo, puede presentar distintas enfermedades al momento de plantarla, cada una con distintas características. Por lo mismo, en la década de los 80 se construyó una base de datos que contiene las distintas enfermedades que presenta esta semilla (y la planta en general) con las distintas variables que caracteriza a cada una, con el objetivo de investigar dichos problemas.

El problema principal en este caso sería con el set de datos disponible, poder estudiar las distintas enfermedades y ver si con las variables que se presentan se logre realizar un análisis estadístico con el fin de entender mejor el conjunto de datos que se presenta, descubrir las distintas relaciones entre las variables y poder hacer estudios futuros.

2.1. Descripción de la base de datos

La base de datos utilizada corresponde a Soybean, específicamente Soybean (Large), la cual contiene información de diferentes enfermedades y características que afectan a la soya y su planta. Dentro de sus clases se encuentran un total de 19 enfermedades, aunque en los diferentes estudios y papers que utilizan esta base de datos consideran 15 de ellas (esto debido a que las últimas cuatro poseen pocas instancias). También cuenta con 35 atributos que indican diferentes características de cada observación, como por ejemplo: germinación, area dañada, tamaño de semilla, etc. Es importante señalar que la base de datos posee en total 307 instancias y que los atributos son del tipo categórico, contando además con la presencia de valores desconocidos (representados por ¿") y que los valores correspondientes a los atributos son representados como números pero estos están asignados a diferentes posibles valores categóricos.

2.2. Descripción de clases y variables

A continuación se procederá con la descripción de las clases y variables pertenecientes al dataset utilizado. Se consideran las 15 enfermedades con mayor cantidad de datos dentro del dataset.

2.2.1. Clases

- *diaporthe-stem-canker*: Enfermedad bastante común, pero no tan fácil de reconocer. Es similar a la pudrición por *phytophthora*, puede matar plantas enteras o partes de ella.
- *charcoal-rot*: Enfermedad en la raíz de la soya que se transmite por el suelo, principalmente producida por el calor y la sequía. Las plantas afectadas pueden morir prematuramente y marchitarse o atrofiarse con frecuencia debido a la putrefacción.
- *rhizoctonia-root-rot*: Enfermedad común de la raíz de la soya que causa el mayor daño a las plántulas, aunque también afecta plantas viejas. Puede matar y atrofiar plantas o simplemente producir lesiones superficiales.
- *phytophthora-rot*: Enfermedad que puede matar y dañar a las plántulas y plantas pudriendo el tallo, puede afectar desde la temporada de crecimiento de la semilla hasta casi la cosecha de esta. Se favorece por las condiciones del suelo húmedo y cálido.
- *brown-stem-rot*: Enfermedad en la cual un patógeno infecta el tallo interno, los síntomas pueden ser o no visibles, por lo que es posible que la planta se vea afectada sin identificar una división en los tallos de las plantas afectadas.
- *powdery-mildew*: Enfermedad esporádica que tiene lugar al final de la temporada durante los períodos de temperaturas frías, puede causar defoliación y pérdida de rendimiento. Una de sus características es que las hojas presentan un recubrimiento blanco y polvoriento en las hojas afectadas, similar a la harina.
- *downy-mildew*: Enfermedad típicamente superficial que generalmente no causa pérdida de rendimiento pero puede producir defoliación de las plantas. Ocurre en períodos de

alta humedad y temperaturas moderadas, una de sus principales características es la presencia de hongos color tostado en el envés de las hojas infectadas.

- *brown-spot*: Enfermedad foliar común en la soya, tiene una incidencia alta que rara vez se desarrolla y causa una pérdida significativa de rendimiento. Causa pequeñas manchas oscuras en las hojas. También es conocida como "mancha marrón Septoria"

- *bacterial-blight*: Enfermedad de la soya común durante climas fríos y húmedos. Puede confundirse con la "mancha marrón Septoria", pero la diferencia es que en "bacterial blight"las manchas presentan un halo, generalmente afecta hojas jóvenes.

- *bacterial-pustule*: Enfermedad común y frecuente en climas cálidos y húmedos, una de sus características es la existencia de pequeñas protuberancias de color canela en el envés de las hojas, además no tiene abertura en las pústulas.

- *purple-seed-stain*: Enfermedad que puede causar una defolación severa de las plantas y reduce el rendimiento, una de sus características es que las plantas afectadas pueden desarrollar una planta púrpura que generalmente se hace evidente durante la etapa de formación de las semillas.

- *anthracnose*: Enfermedad del tallo que ocurre en condiciones húmedas y cálidas, los síntomas no son visibles generalmente hasta que la planta alcanza la madurez. Puede reducir el rendimiento, rodales y la calidad de la semilla. Desarrolla manchas marrones irregulares en tallos y vainas, las áreas infectadas presentan pequeñas espinas negras.

- *phyllosticta-leaf-spot*: Enfermedad que produce manchas bien definidas, redondas e irregulares con un centro de color claro que puede tener pequeñas manchas negras dentro de este.

- *alternarialeaf-spot*: Enfermedad que afecta las hojas formando manchas causadas por hongos. Las que presentan anillos concéntricos, tienden a aparecer primero en hojas inferiores, las cuales se abren y se caen.

- *frog-eye-leaf-spot*: Enfermedad común que puede causar una defoliación severa de las hojas en ambientes cálidos y húmedos. Las manchas poseen un anillo marrón rojizo o

púrpura que rodea manchitas redondas.

2.2.2. Variables

Observación: 'lt' y 'gt' significan 'leather than' y 'greater than' respectivamente, el que aparezca un '?' como posible valor significa que es desconocido. Mientras que 'dna' significa que no aplica.

- *date*: Corresponde a la fecha de aparición de la enfermedad, los posibles valores son: april, may, june, july, august, september, october, ?.
- *plant-stand*: Corresponde al estado del soporte de la planta, los posibles valores son: normal, lt-normal, ?.
- *precip*: Corresponde a la cantidad de precipitación recibida por la planta, es decir: lluvia, llovizna, nieve, aguanieve o granizo. Los posibles valores son: lt-norm, norm, gt-norm, ?.
- *temp*: Corresponde a la temperatura ambiente a la que estan expuestas las plantas, los posibles valores son: lt-norm, norm, gt-norm, ?.
- *hail*: Corresponde si las plantas fueron expuestas al granizo, los posibles valores son: yes, no, ?.
- *crop-hist*: Corresponde al historial de cultivo, comparándolo con años anteriores. Los posibles valores son: diff-lst-year, same-lst-yr, same-lst-two-yrs, same-lst-sev-yrs, ?. Es decir: diferente al año pasado, igual al año pasado, igual a los dos últimos años, igual a los últimos siete años.
- *area-damaged*: Corresponde al área dañada de la planta, los posibles valores son: scattered, low-areas, upper-areas, whole-field, ?. Es decir: dispersas, áreas bajas, áreas superiores o todo el campo.
- *severity*: Corresponde a la severidad de los daños, los posibles valores son: minor, pot-severe, severe, ?. Para menor, potencialmente grave y grave.

- *seed-tmt*: Corresponde a la presencia de una enzima llamada Tocopherol Methyltransferase en la semilla, los posibles valores son: none, fungicide, other, ?.
- *germination*: Corresponde a la germinación de la semilla, los posibles valores son: 90-100 %, 80-89 %, lt-80 %, ?.
- *plant-growth*: Corresponde al crecimiento de la planta, los posibles valores son: norm, abnorm, ?. (normal o anormal).
- *leaves*: Corresponde a la condición de las hojas, los posibles valores son: norm, abnorm.
- *leafspots-halo*: Corresponde al halo de las manchas de las hojas, los posibles valores son: absent, yellow-halos, no-yellow-halos, ?.
- *leafspots-marg*: Corresponde al margen de las manchas de las hojas, los posibles valores son: w-s-marg, no-w-s-marg, dna, ?.
- *leafspot-size*: Corresponde al tamaño de las manchas de las hojas, los posibles valores son: lt-1/8, gt-1/8, dna, ?.
- *leaf-shread*: Corresponde a si las hojas están trituradas o destrozadas, los posibles valores son: absent, present, ?.
- *leaf-malf*: Corresponde a si las hojas presentan malformaciones, los posibles valores son: absent, present, ?.
- *leaf-mild*: Corresponde al crecimiento de moho en la hoja, los posibles valores son: absent, upper-surf, lower-surf, ?.
- *stem*: Corresponde a la condición del tallo, los posibles valores son: norm, abnorm, ?.
- *lodging*: Corresponde a una condición de la soya que afecta su potencial de rendimiento y produce pérdidas en la cosecha, los posibles valores son: yes, no, ?.
- *stem-cankers*: Corresponde a una llaga en el tallo, los posibles valores son: absent, below-soil, above-soil, above-sec-nde,?. (ausente, bajo el suelo,).

- *canker-lesion*: Corresponde a una lesión en el tallo, los posibles valores son: dna, brown, dk-brown-blk, tan, ?. (no aplica y algunos tipos de lesiones)
- *fruiting-bodies*: Corresponde a la presencia de cuerpo frutal en la planta, los posibles valores son: absent, present, ?.
- *external decay*: Corresponde al deterioro externo, los posibles valores son: absent, firm-and-dry, watery, ?.
- *mycelium*: Corresponde a la presencia de micelio, los posibles valores son: absent, present, ?. El micelio corresponde al talo de los hongos, conformado por filamentos ramificados que permiten la nutrición de estos.
- *int-discolor*: Corresponde a la decoloración interna, los posibles valores son: none, brown, black, ?.
- *sclerotia*: Corresponde a identificar si esta presente el esclerocio en la planta (masa compacta de micelio que contiene reservas alimenticias), los posibles valores son: absent,present,?.
- *fruit-pods*: Corresponde al estado de la vaina de la fruta, los posibles valores son: norm, diseased, few-present, dna, ?.
- *fruit spots*: Corresponde a identificar si existen manchas en la fruta, los posibles valores son: absent, colored, brown-w/blk-specks, distort, dna, ?.
- *seed*: Corresponde a la condición de la semilla, los posibles valores son: norm, abnorm, ?.
- *mold-growth*: Corresponde al crecimiento de moho en la semilla, los posibles valores son: absent, present, ?.
- *seed-discolor*: Corresponde a la decoloración de la semilla, los posibles valores son: absent, present, ?.
- *seed-size*: Corresponde al tamaño de la semilla, los posibles valores son: norm, lt-norm, ?.

- *shriveling*: Corresponde a si la planta marchitó, los posibles valores son: absent, present, ?.
- *roots*: Corresponde a la condición de la raíz, los posibles valores son: norm, rotted, galls-cysts, ?.

3. Análisis Estadístico

La base de datos utilizada para esta experiencia corresponde a "Soybean", la cual contiene 19 clases que se ubican en la primera columna correspondientes a enfermedades que afectan tanto a la soya como a la planta en general, es decir, su raíz, tallo, hojas, etc. Mientras que las variables contienen cualidades con respecto a las condiciones a las que fue expuesta cada observación como por ejemplo: la fecha en la cual se identificó la enfermedad, la cantidad de agua que recibió producto de precipitaciones, daños en diferentes partes de la planta, etc.

Para la primera parte del análisis, se consideraron todas las entradas del dataset utilizado (ya que cada uno representa una observación de cierta enfermedad), para obtener las frecuencias de aparición de ciertos elementos y así representarlos con gráficos. En primer lugar, se revisan las frecuencias de cada clase para saber cuales enfermedades son las que tienen más instancias. Para ello se hace el siguiente gráfico de barras. Se debe tener en cuenta también que de las 19 clases, solo las primeras 15 se usan para el estudio, ya las últimas 4 tienen muy pocas instancias, estas corresponden a: 2-4-d-injury, herbicide-injury, cyst-nematode, diaporthe-pod-&-stem-blight. La figura muestra que las 4 enfermedades más comunes de la semilla son: alternarialeaf-spot, brown-spot, anthracnose, y brown-stem-rot.

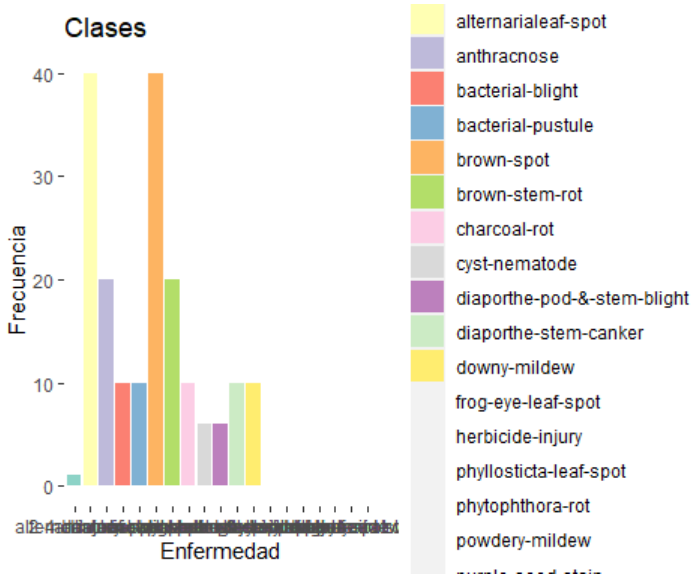

Figura 1: Gráfico de barra de las enfermedades

Entre estas, según Mantecón (2008) una de las enfermedad más común en Argentina es el brown-spot, la cual depende del historial de cultivo que en la base de datos se representa con el atributo crop-hist. Para este se hace una prueba de dependencia con chi cuadrado, entre la enfermedad y dicho atributo, la cual da un p-valor de 1.344 e-06, y como este es menor a la significación de 0.05, demuestra que el brown spot depende del historial de cultivo. En el siguiente grafico se muestra las frecuencias de crop hist en la enfermedad, y la que tiene un mayor número de instancias es cuando el historial de cultivo es el mismo en los 7 años, seguida de un historial igual en los ultimos dos años.

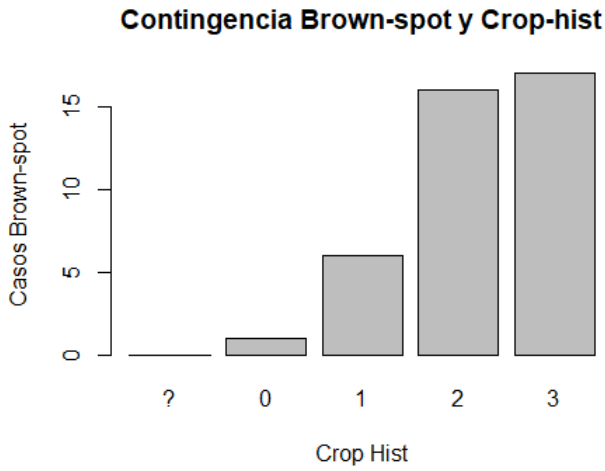

Figura 2: Contingencia de crop-hist y brown-spot

Una de las características importantes dentro del concepto de las enfermedades de la soya corresponde al área dañada de la planta, ya que además es una característica visual muy sencilla de identificar debido a la superficie anormal que aparece en ella. Se consideran todas las observaciones del dataset con su respectivo valor de área dañada para realizar un gráfico de barras y así representar la información existente. En el siguiente gráfico se puede apreciar las frecuencias de los distintos niveles de daños para todos los datos, donde 'scattered' se representa con un 0, 'low-areas' con un 1, 'upper-areas' con un 2, 'whole-field' con un 3 y '?' indica que la entrada no contenía el dato del área dañada.

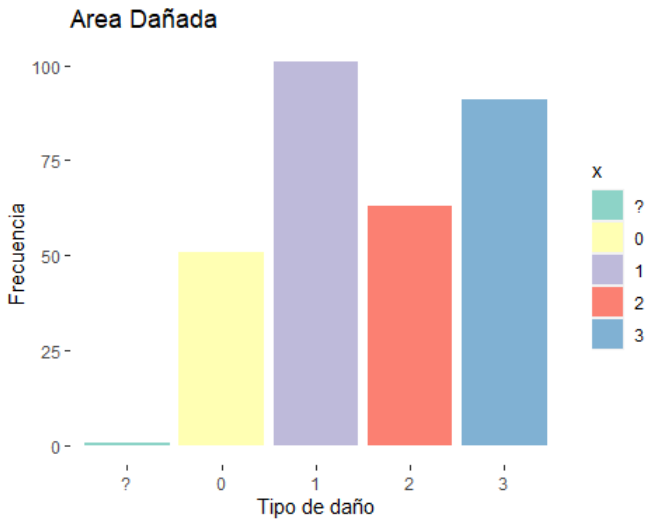

Figura 3: Gráfico de barra del área dañada

Se puede apreciar que las plantas poseen mayormente daños en las áreas bajas, luego le sigue en su campo entero, disminuye la frecuencia a una menor cantidad en las áreas superiores. Disminuyendo un poco más para daños en áreas dispersas y finalmente las plantas que no poseen áreas dañadas son casi cero. Además se evalúa la dependencia entre las enfermedades y el area dañada, esta entrega un valor p de 0.00049, y dada la significación de 0.05, se ve que el area dañada depende de una enfermedad.

Otra característica relevante corresponde al estado de las hojas de la planta, ya permite identificar anormalidades en su superficie. Lo cual es atribuible a alguna enfermedad, por lo que el siguiente gráfico muestra las frecuencias de las condiciones de las hojas de todas las observaciones del dataset, es decir, permite identificar cual condición de las hojas es más frecuente cuando se posee una enfermedad. La condición 'norm' se representa con 0 y 'abnorm' se representa con 1. Además para estos estudios se omiten las otras condiciones de la hoja, ya que las otras cualidades aparecen solo si la hoja es anormal (Michalski et al. (1982)), por eso el análisis se enfocará más en si la hoja es normal o no.

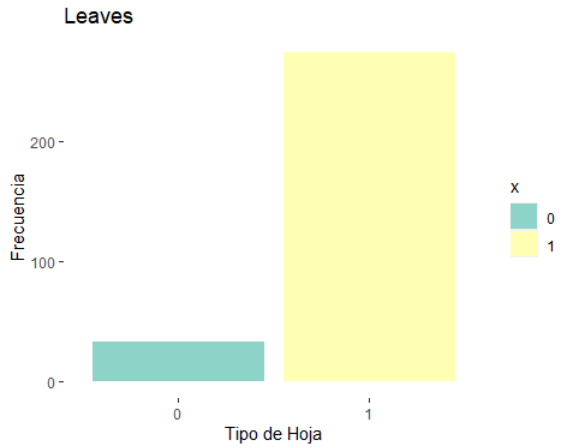

Figura 4: Gráfico de barra de la condición de las hojas

Como se aprecia en el gráfico anterior, los datos contienen las características de las observaciones que posee cierta enfermedad. Por lo que tiene sentido que en su mayoría se tengan hojas anormales, ya que están siendo afectadas por diversas enfermedades.

La temperatura a la cual es expuesta la planta tiene mucha relación con su capacidad de producción y rendimiento, tanto el calor como el frío pueden afectar la planta y así mismo beneficiar el avance de alguna enfermedad. Es por esto que esta variable también es considerada para el análisis de frecuencia, la cual se muestra en el siguiente gráfico:

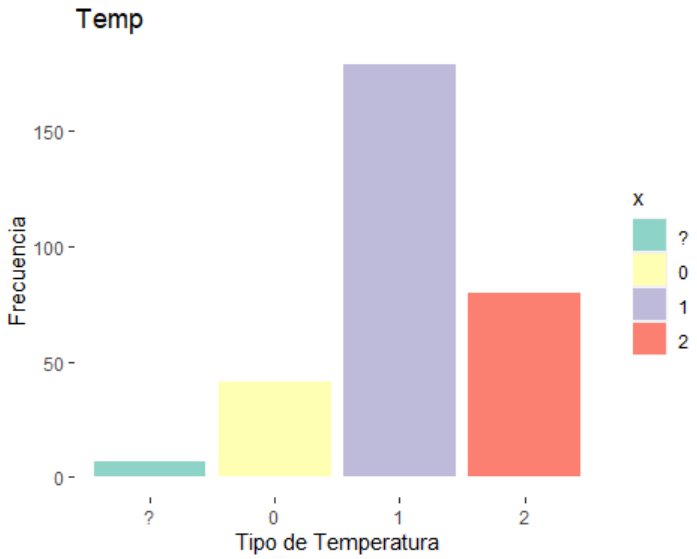

Figura 5: Gráfico de barras de la temperatura

Donde 'lt-norm' corresponde a 0, 'norm' corresponde a 1 y 'gt-norm' corresponde a 2. Se puede identificar que en la mayoría de las observaciones la temperatura a la cual fueron expuestas las plantas fue normal, pero es importante considerar que mayor a lo normal y menor a lo normal también tienen una cantidad no baja de apariciones. Además puede presentarse el caso de que las enfermedades sean producto de la temperatura, por lo que se hace un test de chi cuadrado que resulta con p-valor de 2.2 e -16 , lo cual es menor a la significación de 0.05 entregada, por lo que existe dependencia entre la temperatura y la enfermedad. El siguiente gráfico muestra como es la temperatura en la distintas enfermedades, donde más 100 de ellas se presentan por presentar alta o baja temperatura en la semilla.

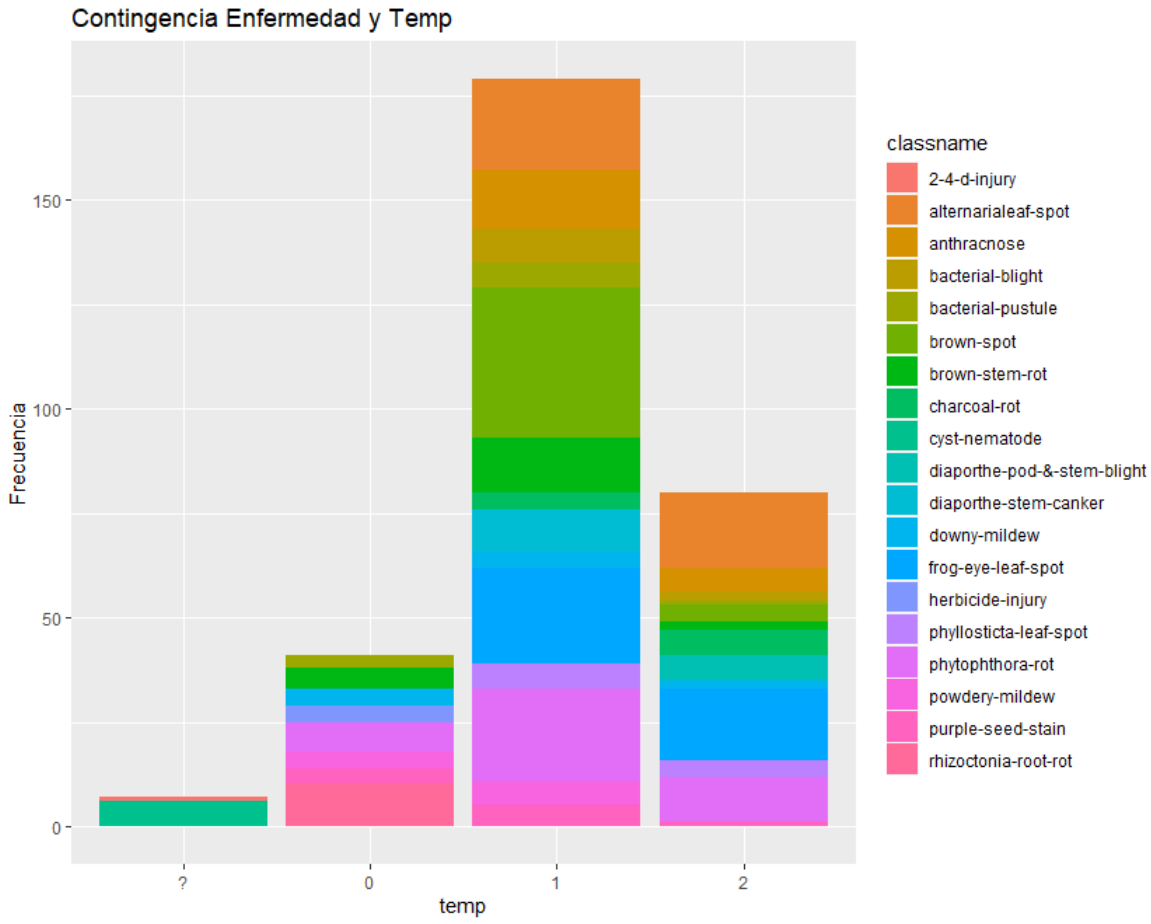

Figura 6: Relación entre las enfermedades y la temperatura

Por otro lado, también se considera el tamaño de las semillas para identificar si tiene que ver con la aparición de las enfermedades, ya que Li et al. (2019) afirma que es una característica importante al momento de sembrar la semilla de soya. El siguiente gráfico muestra los tipos de tamaño de las semillas de la base de datos:

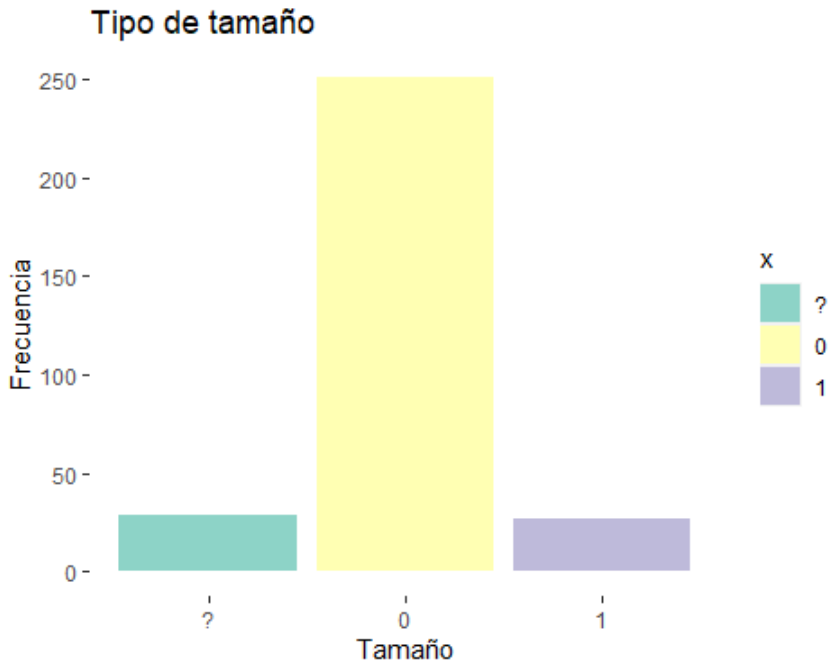

Figura 7: Gráfico de barras del tamaño de la semilla

Como se puede apreciar, en su mayoría las semillas corresponden a un tamaño normal (el 0 corresponde a 'norm') mientras que para menor que normal hay pocas observaciones en comparación a las de tamaño normal. Por lo que se puede identificar que las enfermedades afectan de igual manera las semillas de tamaño normal.

Otra característica importante a mencionar es la precipitación, ya que Schuster et al. (2019) comenta que la enfermedad "phytophthora-rot"tiene causa directa por el tipo de cultivo que recibe la semilla. Para comprobar esto se hace la prueba chi-cuadrado correspondiente con una significación de 0.05, obteniendo un p-valor de 7.438e-14, el cual nos reafirmaría que existe dependencia entre dicha enfermedad y el tipo de cultivo de la semilla. En el siguiente gráfico se muestra la relación entre el historial de riego y dicha enfermedad

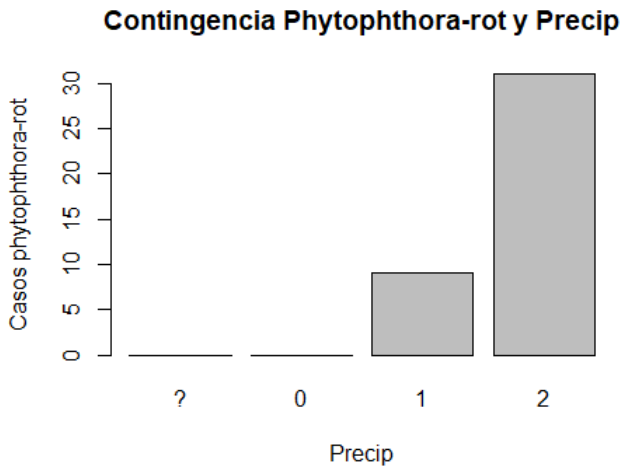

Figura 8: Contingencia phytophthora-rot y precipitation

Finalmente se quiere ver si hay una relación entre las enfermedades y su mes de ocurrencia, se busca saber si el mes puede influir en la aparición de alguno de estos padecimientos. La siguiente figura muestra los meses por frecuencia para todas las instancias.

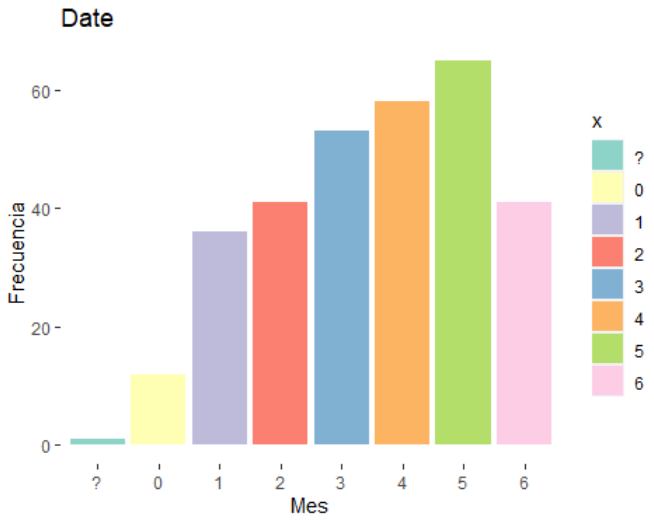

Figura 9: Gráfico de barras de los meses

Se observa que los meses con mayor frecuencia son 3, 4, y 5 que corresponden a Julio, Agosto y Septiembre. Como el mes 5 tiene una alta frecuencia se sospecha que la enfermedad puede depender del mes, para lo cual se realiza una prueba más de chi cuadrado, que da un p-valor de 2.2 e-16, el cual es menor a la significancia entregada de 0.05, por lo que efectivamente la enfermedad puede depender del mes de ocurrencia.

Además esto se puede ver también en el siguiente gráfico que representa a la tabla de contingencia de los meses y las enfermedades, donde se observa que hay grupos de las enfermedades alternarialeaf-spot, brown-spot y downy-mildew con varias instancias en los meses 3, 4, y 5, por lo que se concluye que en este periodo son más propensas a aparecer.

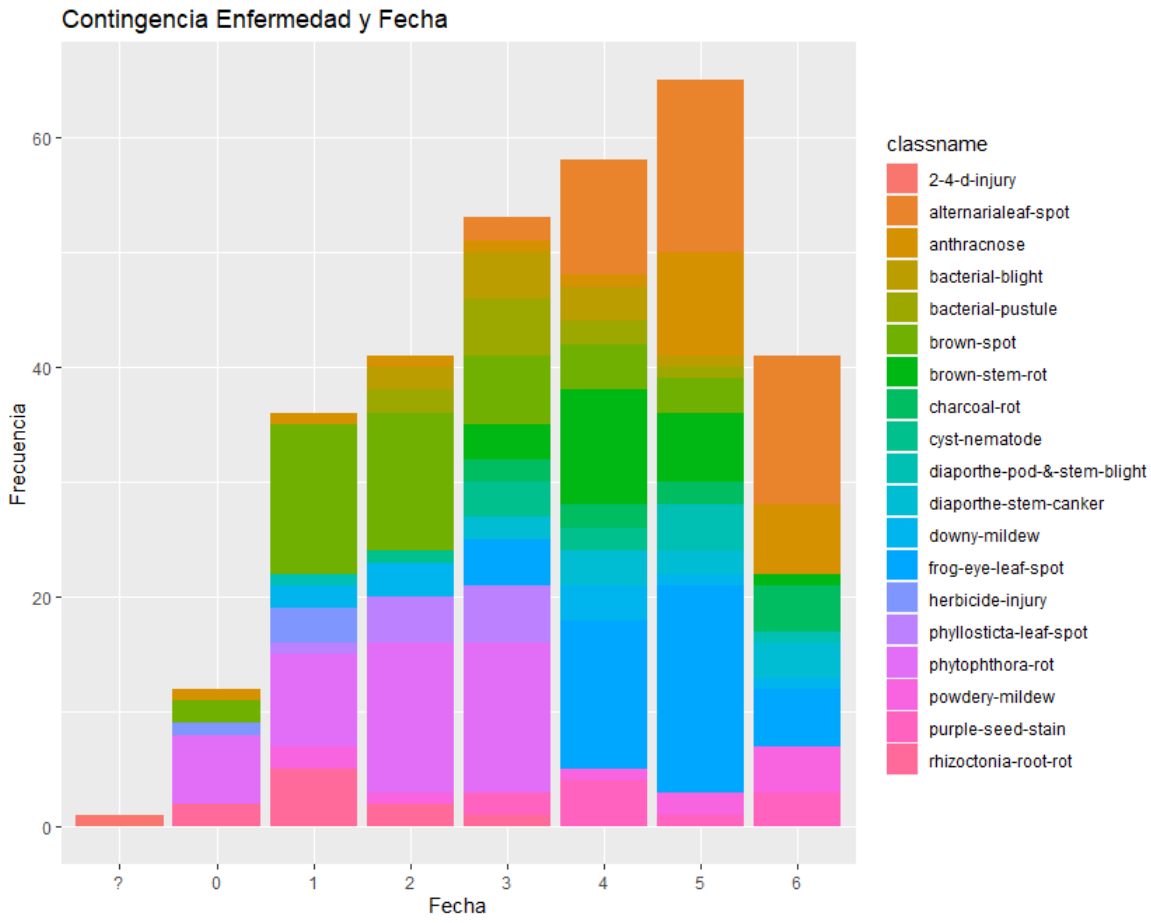

Figura 10: Contingencia entre enfermedades y meses de aparición

4. Conclusiones

La base de datos de Soybean contiene un total de 19 clases que representan distintas enfermedades que afecta a la soya y su planta en general, cada una de las columnas representa una cualidad determinada encontrada para cada observación perteneciente al dataset, entre ellas se encuentra la fecha de ocurrencia de la enfermedad, si existen áreas dañadas de la planta, características de las hojas, etc. Esta información es sumamente relevante para el ámbito de la cosecha y producción de soya, porque el hecho de que los datos sean reales permite trabajar con ellos para tratar de encontrar relaciones o patrones dentro de las diferentes enfermedades y las respectivas características.

Como se identificó en la sección del análisis estadístico, al graficar las frecuencias de ciertas características como el área dañada, los meses de ocurrencia de las enfermedades o la temperatura a la cual fueron expuestas las plantas, es posible identificar como se comportan los datos dentro de la dataset, permitiendo reconocer cuales son los valores de cada característica que más o menos se repiten para comprender si existen relaciones entre determinadas características y las enfermedades correspondientes.

Para determinar con mayor exactitud si existe dependencia entre las variables se aplicó el test de chi cuadrado, por ejemplo para el caso de la temperatura se desea conocer si esta es causante de las enfermedades. En el gráfico de la figura 6 se pudo apreciar que para las distintas enfermedades, casi 100 de ellas eran producto de una alta o baja temperatura, por lo que gracias a la prueba de chi cuadrado realizada anteriormente se identificó que efectivamente existe dependencia entre la temperatura y las enfermedades.

Finalmente, cabe mencionar que R es un lenguaje de programación muy útil ya que fue desarrollado para resolver problemas con fines estadísticos, y para este caso en particular se tenían variables categoricas, de las cuales se obtuvo bastante información sobre comportamiento y dependencia entre dichas variables y las clases de enfermedades dentro del dataset. Al representar la información mediante gráficos, es posible realizar una conclusión previa de manera más sencilla antes de realizar las pruebas estadísticas como tal, por lo que fue importante también para este trabajo haber mostrado los datos con gráficos de barras.

Bibliografía

Intagri (2016). Soya: Importancia nacional e internacional. https://www.intagri.com/articulos/noticias/soya-importancia-nacional-e-internacional.

Li, J., Jia, L., Zhang, A., Mateen Khattak, S., Sun, W., Gao, M., and Wang (2019). Soybean seed counting based on pod image using two-column convolution neural network. https://ieeexplore.ieee.org/document/8716704.

Mantecón, J. D. (2008). Efficacy of chemical and biological strategies for controlling the soybean brown spot (septoria glycines). https://scielo.conicyt.cl/scielo.php?pid=S0718-16202008000200011script=sci$_a rttexttlng = e$.

Michalski, J., Davis, V., Bisht, J., and Sinclair (1982). Plant/ds: an expert consulting system for the diagnosis of soybean diseases. http://ebot.gmu.edu/bitstream/handle/1920/1565/82-04.pdf?sequence=1isAllowed=y.

Michalski, R. S. and Chilausky, R. L. (1980). Soybean (large) data set. http://archive.ics.uci.edu/ml/datasets/Soybean+(Large).

Schuster, I., Lopes da Silva, F., Dalla Nora, T., de Almeida, B., Borém de Oliveira, A., and Volpato, L. (2019). Snp markers associated with soybean partial resistance to phytophthora sojae.