Año: 2022 Volumen: 1
Número 8, ISSN: 2448-623X

Algoritmos de clasificación en la cosecha del brócoli

R. Juárez-Del-Toro[1], J. Castrejón-Lozano[2], F. Salas-Pérez[3]

***Resumen*—El aprendizaje automático se ha convertido en una herramienta de análisis de datos de mayor uso en nuestros tiempos y en casi todas las áreas del conocimiento. Esto se debe en gran medida a la creciente disponibilidad de información a la que podemos tener acceso y que son generados por sensores, encuestas, fotografías, dispositivos móviles, etcétera. Los algoritmos de clasificación son parte de los algoritmos de aprendizaje automático supervisado y permiten la clasificación de los datos nuevos, en categorías bien definidas con datos de entrenamiento. La agricultura es una fuente natural de información muy amplia y permite a los productores determinar por ejemplo el tiempo adecuado de cosecha, los volúmenes y rendimientos de producción o bien el análisis riesgos de manera anticipada. En una empresa agrícola regional que siembra brócoli es indispensable determinar el tiempo preciso de madurez en el que debe cosecharse el producto. Esta información le permite al agricultor planear la contratación de personal, el uso de máquinas, la disponibilidad de espacios físicos, entre otros. En este trabajo se hace una comparación de algoritmos de clasificación del aprendizaje automático supervisado, escritos en Pyhton, que determinan el momento correcto de cosecha. El artículo también repasa las principales ventajas y desventajas del uso de estos algoritmos de clasificación en este tipo de análisis en la agricultura.**

Palabras claves—aprendizaje automático, algoritmo de clasificación, agricultura, IA, día de cosecha

***Abstract*—Machine Learning has become the most widely used data analysis tool in our world in all areas of knowledge. The information available every day which is generated by sensors, questionnaires, the internet, mobile devices, etc. is mind-blogging. Classification is a supervised learning algorithm to identify a category of new observations based on training data. Agriculture generates priceless information to forecast the harvest dates, crop yield, or the risk analysis. A local farm where vegetables are the main product wish to determine the precise maturity time in which the product must be harvested. The data analysis allows for planning the employee recruiting, for using the machinery and equipment, for the availability of physical spaces, among others. In this work, a number of classification algorithms are performed: Linear models for example Logistic regression and Support Vector Machine; and nonlinear models for example K-nearest Neighbours, Naïves Bayes, Multilayer Perceptron, Stochastic Gradient Descent,Decision Tree and Forest Random. decision tree, random forest; all written in Python. The classifiers learned how to identify the correct moment (features values) when the plant is ready to be harvested. The article also includes the missing data and the regression analysis.**

Keywords— agriculture, AI, classification algorithm, harvesting date, ML

I. Introducción

El objetivo de este trabajo es comprobar la efectividad de los algoritmos clasificadores que existen en el aprendizaje automático supervisado para identificar cuando el ÁREA DEL FLORETE de la planta de brócoli ha alcanzado un tamaño adecuando para que pueda cosecharse. Las variables denominadas características son las variables de mayor significancia e influencia en la identificación del área del florete de la planta. En este trabajo, se analizaron sólo variables biométricas del brócoli. Las variables biométricas son el conjunto de características que describen a una planta, como sus dimensiones, peso, diámetro, número de hojas, áreas, etcétera. La idea es identificar aquellas características biométricas que permitan indicar cuando el brócoli tiene el tamaño correcto, o el área del florete correcto para ser cosechado. Esta información facilitaría el proceso de planeación de la cosecha, por ejemplo, la contratación anticipada del personal y maquinaria de cosecha, la disponibilidad de espacios físicos para el almacenamiento, y hasta la prevención de riesgos en la cosecha como plagas o condiciones climáticas inesperadas, ayudando a optimizar los ingresos para el rancho. Esta es la forma en que los cosechadores expertos miden el tiempo exacto de cosecha en el campo. Las herramientas de clasificación por aprendizaje automático brindan una valiosa herramienta a partir de la cual es posible estimar la fecha correcta de cosecha. La predicción de cosecha de los cultivos es una tarea relativamente difícil debido a la complejidad de la relación entre el proceso de crecimiento de la planta y los factores ambientales como el clima.

El brócoli es un cultivo que ha recibido una atención considerable con respecto a la predicción de la fecha de cosecha Marshall y Thompson (1987), Pearson et al. (1993), Fujime y Okuda (1994), Default (1997). Wurr et al. (1992)

[1] *Tecnológico Nacional de México. Instituto Tecnológico Superior de Lerdo. División de posgrado. Av. Tecnológico 1555, Colonia Periférico. C.P. 35150, Ciudad Lerdo, Durango. México. Raymundo.jt@itslerdo.edu.mx*
[2] *Universidad Politécnica de Gómez Palacio. Ingeniería en Tecnologías de Manufactura. Carretera El Vergel-La Torreña Km 820, Colonia El Vergel. C.P. 35120, Gómez palacio, Durango. México**gjcastrejon@upgop.edu.mx.
[3] *Universidad Autónoma de Coahuila. Unidad Torreón. Facultad de Contaduría y Administración. Blvrd. Revolución 151 oriente, Colonia centro. C.P. 27000, Torreón, Coahuila. México. francisco.salas@uadec.edu.mx*

Año: 2022 Volumen: 1
Número 8, ISSN: 2448-623X

estableció una relación cuadrática entre el logaritmo natural del diámetro de la cabeza y los grados diarios efectivos acumulados desde el inicio de la cabeza, que podría usarse para predecir cuándo las cabezas alcanzan un tamaño específico. Se desarrolló un modelo informático para predecir las fechas de cosecha después de tomar muestras de la cabeza en una etapa joven. Grevsen (1998) describió una relación cuadrática similar entre el logaritmo del diámetro de la cabeza de brócoli y los grados-día acumulados. Se obtuvo una marcada mejora en la bondad del ajuste al incluir la densidad de plantas y el cultivar en la relación. La relación se puede usar para pronosticar cuándo las cabezas alcanzan cierto tamaño. Las cabezas nunca se iniciaron antes de que se formaran al menos siete hojas visibles y se sugirió que el número de hojas se puede usar para indicar el momento más temprano posible para el muestreo del diámetro de la cabeza.

La presente contribución tiene la siguiente estructura, la sección 1 presenta la introducción y el estado del arte sobre el trabajo. En la sección 2 se presenta la descripción de variables, el análisis de correlación y la selección de variables significativas y características. En la sección 3 presentamos los resultados de la aplicación de los algoritmos de clasificación, tomando en cuenta el área del florete de la planta. La sección 4 presenta las conclusiones del trabajo y finalmente se presentan los agradecimientos y las referencias más importantes.

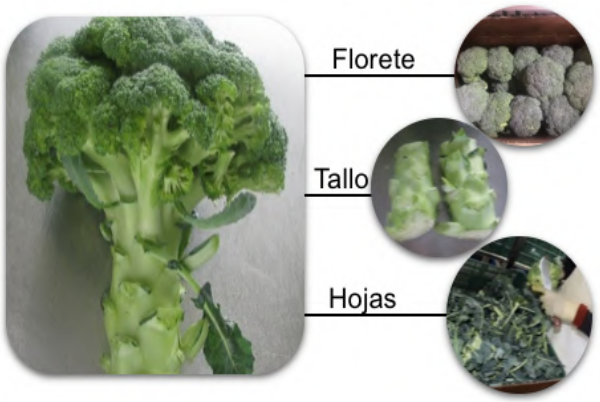

Figura 1.1. Partes de la planta del Brócoli.

II. Parte Técnica del Artículo

Rancho Medio Kilo (RMK), en Aguascalientes México, es una finca familiar, social y ecológicamente responsable, que produce, procesa y distribuye vegetales como brócoli, coliflor, zanahoria, calabaza, acelga, espinaca, cebolla, lechuga, tomate, chile, etc. El brócoli convencional representa el 39% de la producción de RMK para todos sus productos y se utiliza únicamente en dos variedades etiquetadas como V106 y V194. La diferencia principal entre ellas es la capacidad de soportar el clima frío de invierno (V194) pero no es el centro de esta contribución. En tiempo de cosecha, RMK tiene que contratar el personal necesario, tener disponible la maquinaria y todos los recursos adicionales para el proceso de cosecha. En el RMK, la superficie de producción se divide en parcelas o *Tablas* que se dividen a su vez en secciones. Hay cuatro Tablas etiquetadas como Ch127, Ch143 y Ch108. Cada Tabla de alrededor de cuatro hectáreas tiene cuatro secciones de aproximadamente una hectárea. Las variedades de brócoli V106 y V194 se cultivan al azar en mayo en Ch127. Las variedades de brócoli V106 y V194 también se cultivan al azar en julio en Ch143. Solo el brócoli V194 se cultiva en septiembre en Ch108. Entonces la variedad de brócoli V194 es la variedad resistente al frío. En este trabajo solo se aplican diferentes clasificadores para determinar el tiempo preciso de cosecha de brócoli, con el objetivo de tomar decisiones en cuanto a la planificación estacional del cultivo, y para equilibrar la cantidad de medios de producción utilizados, como mano de obra humana, maquinaria, anticipación de recursos, y también para estar preparado. para reaccionar contra plagas o condiciones climáticas inesperadas. El rancho desearía anticiparse respecto a un período de treinta días antes del inicio de la cosecha.

La Figura 2.1 ilustra el ciclo de cultivo del brócoli en un gráfico en una granja local, desde la siembra hasta la cosecha. Hay tres períodos de siembra en el rancho. Cada período de siembra comienza alrededor de cuatro o cinco días después del anterior. El eje horizontal indica el día exacto del calendario y la semana desde el comienzo del primer período de siembra. El ciclo de cultivo del brócoli oscila alrededor de 14 semanas desde la primera etapa de siembra hasta el último día de cosecha. Ver Figura 2.2. Hay un período de proyección de 40 días a partir del inicio de cada etapa de siembra en una Tabla. El período de cosecha comienza alrededor de los 70 días u 11 semanas después de cada inicio de siembra. En el período de cosecha, el rendimiento del cultivo aumenta hacia el máximo de 8 a 10 días después del inicio. Después de eso, el rendimiento de los cultivos disminuye. La forma del rendimiento del cultivo en cada etapa de siembra en el gráfico tiene una forma gaussiana.

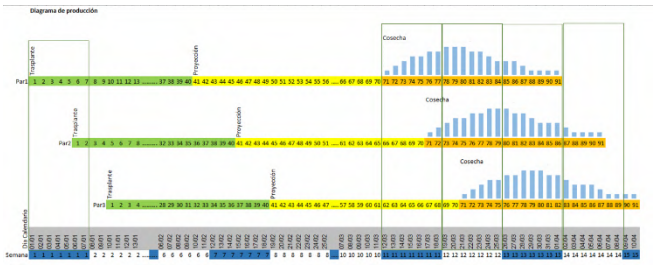

Figura 2.1. Ciclo del cultivo del Brócoli en la empresa de horatlizas.

El ciclo vegetativo del brócoli oscila entre 58 y 120 días

dependiendo de las características genéticas de las variedades, el manejo agronómico y las condiciones climáticas al momento de la siembra, cultivo y cosecha. Jaramillo et al. (2006) y Maroto Borrego y Baixauli Soria (2007) dicen que desde la germinación de la semilla hasta la plántula hay aproximadamente 30 días. La plántula se caracteriza por la formación de hojas y raíces, tres o cuatro hojas bien formadas con 10-12cm de altura y está lista para trasplantar a campo. Desde que las plántulas han sido trasplantadas a los 40 días, el primordio floral ha subido y se visualiza perfectamente. La planta de brócoli ahora tiene 70 días. La altura, diámetro del tallo, biomasa, número de hojas y área foliar muestran un incremento logarítmico existiendo también una proliferación de hojas. Según Jaramillo et al. (2006), el cierre del dosel ocurre alrededor de los 35 días después del trasplante y muestra un desarrollo acelerado de las hojas para la captación de radiación. La cabeza floral aparece de 40 a 45 días después del trasplante y cuando la planta tiene de 18 a 20 hojas. A partir de este momento comienza un crecimiento lineal de la planta y en especial de la cabeza floral. Se confirma por Jaramillo et al. (2006) que la tasa de emisión foliar, la tasa de evolución de la superficie foliar y la tasa de crecimiento del tallo, disminuyen entonces. Maroto (1989), estudió la influencia de las condiciones físicas del suelo y la humedad, en la emergencia floral. La inflorescencia se presenta cuando las flores aún no están abiertas y dura de 20 a 25 días. La inflorescencia presenta un crecimiento exponencial en diámetro y biomasa. Alrededor de los 55 días después del trasplante comienza un período de crecimiento lento. Luego, de 60 a 65 días después del trasplante, aparece un período más rápido hasta la cosecha. En esta etapa se produce la translocación de fotoasimilados. Luego aumenta el diámetro del tallo y la altura presenta un segundo pico de crecimiento debido al aumento del tamaño de la cabeza. El ciclo del cultivo depende de las variedades, las condiciones edafoclimáticas y hídricas, e incluso de las prácticas culturales y de fertilización.

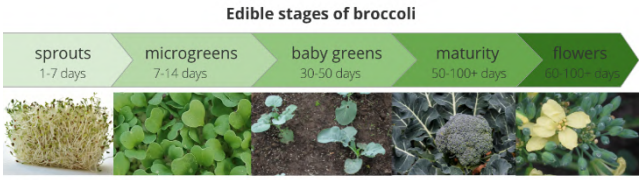

Figura 2.2. Ciclo de crecimiento del Brócoli.

La base de datos y todas las variables se agruparon en tres grupos principales de variables: Biometría e identificación (ID), que describen la edad de la planta y la ubicación real en el gráfico. El grupo de biometría se describe más adelante en un subtema. El grupo de identificación describe la variedad y ubicación de cada planta en el campo, e incluye la observación, número de planta, variedad, cuadro, sección, surco y estación. Debido a la importancia económica las variables de pronóstico en esta contribución se desea determinar el tamaño correcto del florete de la planta para determinar el día de cosecha. A pesar de que el segundo y tercer período de cosecha representan los 80% de la producción total, el primer periodo de cosecha es el más importante. En esta contribución se toman en cuenta todas las observaciones posibles en los datos: de mayo a diciembre de 2019; aun teniendo claro que existen tres períodos de cosecha diferentes en un año (2019): de mayo a agosto para Ch127; de julio a octubre para Ch143; y de septiembre a diciembre por Ch108. El brócoli convencional es el cultivo objetivo en esta contribución. La fecha de cosecha del brócoli se refiere a los períodos durante los cuales realmente ocurre la cosecha del cultivo. No se extienden al período subsiguiente en el que algunos productos se almacenan en el campo después de la cosecha. Las fechas que se muestran indican los períodos en los que se plantan y cosechan los cultivos en la mayoría de los años. No tienen en cuenta fechas excepcionalmente tempranas o tardías de siembras y cosechas dispersas, ni temporadas anormales causadas por condiciones climáticas o económicas.

A. Variables de biometría vegetal

Los datos de biometría de crecimiento de brócoli se recopilaron en RMK de cada uno de los tres períodos de cultivo en 2019. Uno para cada tabla de cultivo en la tierra de RMK: Ch127, Ch143 y Ch108, respectivamente. Algunas variables biométricas del crecimiento de las plantas se obtuvieron utilizando métodos de recolección destructivos. Tales variables incluyen peso fresco y seco de hojas, tallo, ramas laterales, cogollo y la planta completa. El resto de las variables biométricas de las plantas se obtuvieron mediante métodos no destructivos. Tales variables incluyen el número de hojas y hojas dobladas, la longitud y el diámetro del tallo desde la base de la planta, y la superficie foliar y de la flor. En la Figura 1.1 se ilustran las partes principales del brócoli. Como se ha mencionado, la altura, el diámetro del tallo, la biomasa, el número de hojas y el área foliar muestran primero un aumento logarítmico con una proliferación de hojas. Luego, cuando aparece el capítulo floral y la planta tiene de 18 a 20 hojas, comienza un crecimiento lineal para la planta y especialmente el capítulo floral. Finalmente, la inflorescencia presenta un crecimiento exponencial en diámetro y biomasa. Luego aumenta el diámetro del tallo y la altura presenta un segundo pico de crecimiento debido al aumento del tamaño de la cabeza. Debido a este hecho, es más difícil pronosticar el rendimiento de un cultivo o la fecha de cosecha a intervalos periódicos durante la temporada de crecimiento. Sin embargo, se debe seleccionar cuidadosamente un conjunto apropiado de tales características para incluirlo en el modelo de pronóstico y

como un indicador útil del resultado final de la fecha de cosecha y el rendimiento del cultivo, que se miden sin mucho error y multicolinealidad. Finalmente, un método de recopilación destructiva permanente de observaciones sobre las características biométricas no es económicamente factible.

B. Análisis de correlación

El análisis de correlación es el análisis estadístico para realizar la selección de predictores o en este caso de clasificadores. El factor de correlación denominado coeficiente de Pearson, cuyo valor está en el intervalo $[-1,1]$, determina el orden de importancia respecto al pronóstico. El análisis de correlación nos permite seleccionar los principales predictores significativos. Una correlación perfecta también es un efecto indeseable en las relaciones variables y debe evitarse. Se realiza un análisis de correlación del $99\% de significancia, entre la variable de interés y el grupo de biometría, y las observaciones en los tres periodos de cosecha en un año (2019): De mayo a agosto para Ch127; de julio a octubre para Ch143; y de septiembre a diciembre por Ch108. La Tabla I presenta predictores más significativos para la variable ÁREA DEL FLORETE. Son 10 variables significativas de un total de 20 variables biométricas, es decir sólo el 50% de las variables fueron significativas. El PESO FRESCO TOTAL, TALLO PESO FRESCO, CABEZA PESO FRESCO, TALLO PESO SECO, CABEZA PESO SECO y PESO SECO TOTAL, son todas variables destructivas y se refieren al peso de cada uno en gramos. Las variables DIAMETRO TALLO BASE, DIAMETRO TALLO 5 CM, DIAMETRO TALLO 15 CM y DIAMETRO TALLO 20 CM, son todas variables NO destructivas. Es deseable que exista significancia sólo de variables NO destructivas sin embargo esto no es así en este caso. Incluso, el nivel de significancia es mayor de las variables destructivas.

TABLA I. ANÁLISIS DE CORRELACIÓN DE CARIABLES CLASIFICADORAS RESPECTO AL ÁREA DEL FLORETE. NIVEL DE SIGNIFICANCIA AL 99%. N=172 DATOS FALTANTES

Variables clasificadoras	Área del Florete. Coeficiente de Pearson
PESO FRESCO TOTAL	0.627
TALLO PESO FRESCO	0.360
CABEZA PESO FRESCO	0.889
DIAMETRO TALLO BASE	0.490
DIAMETRO TALLO 5 CM	0.321
DIAMETRO TALLO 15 CM	0.210
DIAMETRO TALLO 20 CM	0.490
TALLO PESO SECO	0.255
CABEZA PESO SECO	0.920
PESO SECO TOTAL	0.554

El resto de las variables como: DIAS A INFLORESCENCIA, NUMERO DE HOJAS, NUMERO DE HOJAS DOBLADAS, AREA FOLIAR, LONGITUD DEL TALLO, HOJAS PESO FRESCO, HOJAS PESO SECO, DIAMETRO TALLO 10CM y el PESO DE LOS BROTES LATERALES no fueron significativas para el grupo. La gráfica de la Figura 2.3 muestra el comportamiento de las variables significativas de la Tabla I. Es importante notar en este punto que las variables significativas presentan datos faltantes que limitan su uso en los algoritmos de regresión y de clasificación. La variable CABEZA PESO SECO no se incluye en la gráfica de la Figura 2.3 por esta razón. Entonces es necesario el análisis de datos faltantes para evitar sus efectos en los análisis posteriores.

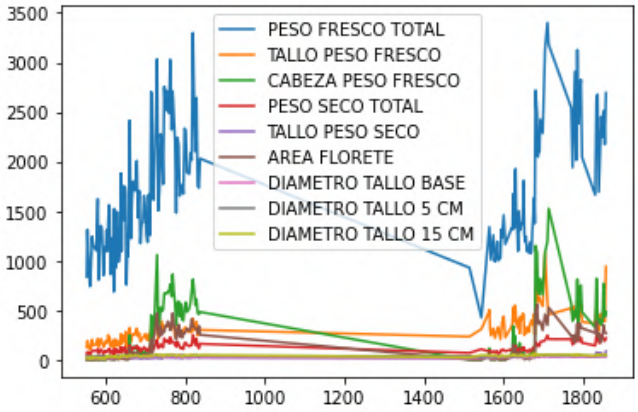

Figura 2.3. Variables significativas para el ÁREA DEL FLORETE.

C. Análisis de datos faltantes

Las variables significativas obtenidas del análisis anterior reducen el conjunto de variables original en el 50% sin embargo no pueden utilizarse en el análisis de clasificación sin antes hacer un análisis de datos faltantes. Este análisis es parte fundamental del llamado pre-procesamiento de datos. Hay varias opciones para evitar un efecto indeseable de este tipo de datos en los algoritmos de Aprendizaje Automático. Entre estas opciones de tratamiento de este tipo de datos se encuentra su eliminación y varios tipos de interpolación o extrapolación numérica. En este trabajo haremos una eliminación tanto de las observaciones faltantes como de las variables con pocos datos disponibles. Las variables con pocos datos disponibles fueron: CABEZA PESO SECO y DIAMETRO A 20 CM, y se eliminaron de la base de datos.

D. Selección de clasificadores

El grupo de variables de estudio en esta parte del trabajo se reduce al considerar un análisis de clasificación inicial por Árbol de Decisión, que permite identificar las variables precisas que son características principales del algoritmo de clasificación. Los rasgos o características son las variables de

un experimento científico, son características de un fenómeno bajo observación que pueden ser medidas u observadas. Cuando estas características alimentan un algoritmo de aprendizaje automático, el algoritmo intenta descubrir patrones entre ellas. Estos patrones se usan para generar las salidas de estas redes de aprendizaje. Las salidas de la red se denominan etiquetas. Cuando la salida obtenida de ciertas características recibe cierta etiqueta por la red, entonces se dice que cae en cierta categoría existente. Un clasificador por árbol de decisión reduce el conjunto de datos en subgrupos cada vez más pequeños sobre diferentes criterios. Una vez que el árbol divide los datos en un solo dato, este se ubica en una de las posibles clases existentes. La Figura 2.4 muestra el árbol de decisión y los clasificadores más importantes en el proceso: CABEZA PESO FRESCO, TALLO PESO FRESCO, DIAMETRO TALLO BASE y DIAMETRO TALLO 15 CM. Estas variables serán las características para los demás clasificadores.

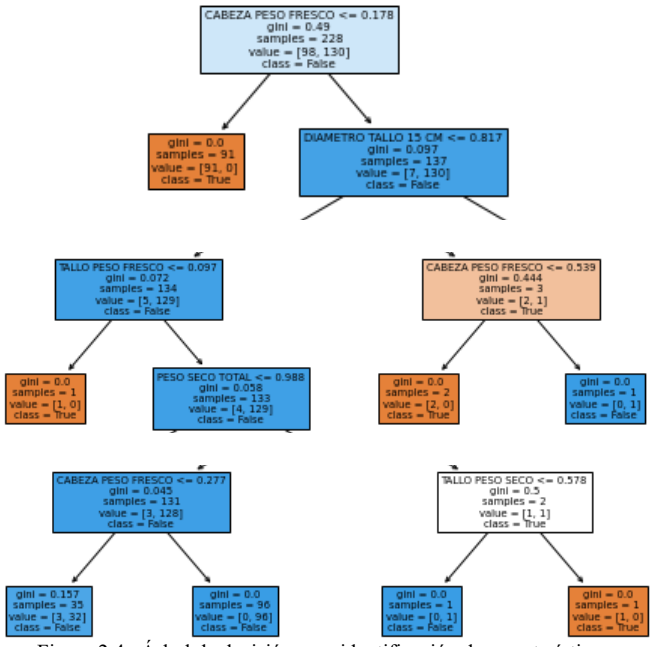

Figura 2.4. Árbol de decisión para identificación de características.

Los siete algoritmos de clasificación mostrados en la Tabla II son los más comunes y los que aplicaremos en este trabajo. Estos algoritmos fueron seleccionados como los de mayor uso y eficiencia en el aprendizaje automático supervisado.

TABLA II. LISTA DE ALGORITMOS DE CLASIFICACIÓN DEL APRENDIZAJE AUTOMÁTICO SUPERVISADO MÁS COMUNES

Algoritmos Clasificadores	Tipo
Multi Layer Prceptron MLP	Lineal
Logistic Regressión	Lineal
Naive Bayes	No Lineal
K-Neighbors	No Lineal
Support Vector SVC	Lineal
Gaussian SGD	Lineal
Aleatorian Decisión Tree	No Lineal

El clasificador por Perceptrón Multicapa MLP implementa el uso de las redes neuronales para entrenarse por backpropagation y mapea o clasifica los datos de entrada en conjuntos apropiados de salida. EL MLP se caracteriza por varias capas de neuronas de entrada, conectados por flechas dirigidas entre la entrada y la salida. La clasificación por Regresión Logística genera predicciones sobre los datos de prueba hacia una escala binaria de 0 y 1. Si el valor de un dato es 0.5 o mayor, entonces se clasifica automáticamente como de la clase 1, mientras que si su valor está por debajo de 0.5, entonces se clasifica perteneciente a la clase 0. Cada una de las características también se etiquetan con ceros y unos. La Regresión Logística es un clasificador lineal y se usa cuando existe cierta relación lineal entre los datos. El clasificador por Naive-Bayes determina la probabilidad de que una observación pertenezca a una clase, calculando la probabilidad de que un evento ocurra dado que algunos eventos de entrada han ocurrido. Cuando esto sucede, se asume que todos los predictores de una clase tienen el mismo efecto sobre la salida, ya que los predictores son independientes. El clasificador por K-Vecinos opera revisando la distancia desde una observación particular hacia los valores de entrenamiento. El grupo de datos que arrojan la menor distancia entre estos datos de entrenamiento y los de prueba sirve para determinar la clase a la que pertenecen. La clasificación por Vectores de Soporte trabaja dibujando líneas que separan los diferentes grupos de datos y agrupan a las posibles clases que existen. Los datos de un lado de la línea pertenecen a una clase y los datos al otro lado son de otra clase diferente. Este clasificador maximiza la distancia entre la línea y los puntos a cada lado de ella para incrementar la visualización de las clases. Cuando se grafican los puntos de prueba es muy claro identificar a que clase pertenece cada dato. El clasificador lineal por Gradiente Descendente (SGD) optimiza cualquier clasificador lineal (SVM, Regresión logística, etc.). Es decir, el Gradiente Descendente no es un clasificador en sí, sino una herramienta de optimización que minimiza o maximiza la función de perdida de algún clasificador lineal. Cada algoritmo de clasificación está complementado por la Matriz de Confusión, que es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. Uno de los beneficios de las matrices de

confusión es que facilitan ver si el sistema está confundiendo dos clases. La matriz de confusión del árbol de decisión anterior se muestra en la Figura 2.5.

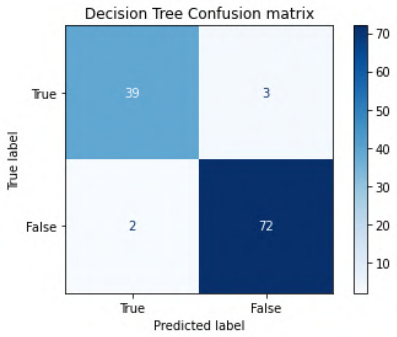

Figura 2.5. Matriz de confusión para el árbol de decisión.

En esta matriz puede notarse que ambas categorías quedan muy bien identificadas y definidas, descritas por el mayor número de observaciones o conteos sobre la diagonal principal. El caso de los falsos positivos está escasamente descrito por los conteos en la diagonal secundaria de la matriz.

E. Preprocesamiento de los datos

El procesamiento de los datos en el aprendizaje profundo es fundamental para implementar cada herramienta de la Inteligencia Artificial. El preprocesamiento de los datos incluye al análisis de datos faltantes; la conversión de los datos categóricos a numéricos, ya que el aprendizaje automático sólo trabaja con este tipo de datos; el escalamiento de características; y la separación de datos de entrenamiento y de prueba. El análisis de datos faltantes ya se abordó en la sección anterior y en este caso no se cuenta con datos categóricos. Antes del escalamiento de las características y la separación de los datos de entrenamiento y los de prueba, se presenta un análisis gráfico de las variables predictoras o clasificadoras obtenidas. Este análisis permitirá confirmar porque estas variables significativas son las características adecuadas para los algoritmos de clasificación. El análisis gráfico incluye la matriz de correlación que confirma la significancia entre las características obtenidas. El diagrama a pares muestra la distribución de la muestra en pares de variables y da una idea clara de que tan definidas o claras están distribuidas las clases bajo análisis. El diagrama a pares tiene el mismo propósito que la gráfica de coordenadas paralelas donde cada fila de las variables características se representa por una línea que atraviesa un conjunto de ejes paralelos, un eje por cada dimensión o número de variable. El análisis de correlación de las variables características comprueba la significancia e influencia de este grupo de variables. Los valores de Pearson obtenidos se muestran en la matriz de correlación de la Figura 2.6.

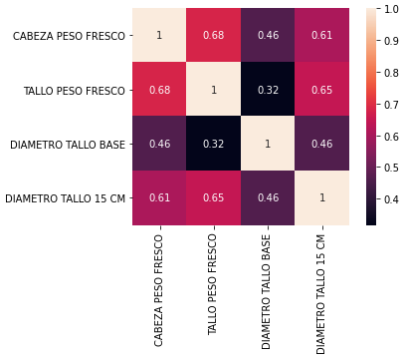

Figura 2.6. Matriz de correlación entre variables catracterísticas

El diagrama a pares de la Figura 2.7 muestra la relación a pares de las variables características. Puede notarse una clara separación de ambas clases en los datos. En algunas combinaciones la separación de ambas no es tan clara pero este efecto es aceptable para la implementación de los algoritmos de clasificación.

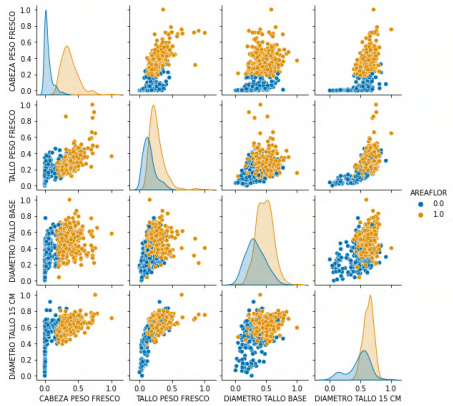

Figura 2.7. Diagrama a pares de las variables características.

La Figura 2.8 muestra la gráfica de coordenadas paralelas de las variables significativas. El color de cada línea permite identificar a las clases en los datos. En este ejemplo en particular, las líneas de color rojo se identifican como una clase con una relativa separación de las líneas de color azul, que representan la segunda clase en los datos. Es decir, es posible hacer una separación de dos clases bien definidas en los datos de las variables características.

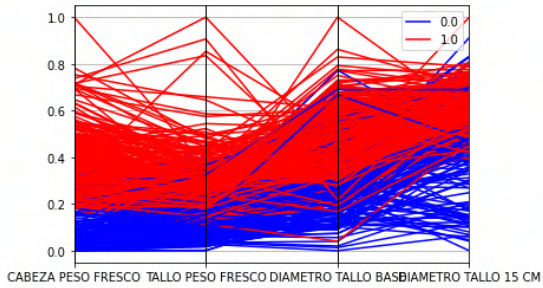

Figura 2.8. Curva de magnetización

El escalamiento de las variables características es la forma de homologar los datos para ser utilizados sin hacer distinciones numéricas entre ellos. Existen dos tipos de escalamiento de las variables características, la Normalización y la Estandarización de los datos. El primero se refiere a transformar los valores en un rango de [0, 1], mientras que la estandarización se encarga de imponer media cero y varianza uno a los datos normalizados. La separación de los datos se realiza comúnmente para evitar el efecto de sobre-ajuste u overfitting y generalmente se separa en dos conjuntos de datos, de entrenamiento y de prueba. Dos tercios de estos datos se usan para formar el primer subconjunto y un tercio para formar el subconjunto de prueba. Tanto el procesamiento como la separación de los datos se realiza a través de las siguiente instrucción en Python:

```
#Línea de código en Python para el escalamiento de
los datos y la separación en entrenamiento y prueba.

#Escalamiento de los datos
scaler = MinMaxScaler()
data4update[cols2]                                  =
scaler.fit_transform(data4update[cols2])

# Separación en datos de entrenamiento y de prueba
X_train,     X_test,     y_train,     y_test     =
train_test_split(feature,target,shuffle=True,test_
size=0.2,random_state=1)
```

III. Resultados

Después de haber realizado el análisis de datos faltantes y la identificación de variables significativas respecto al ÁREA DEL FLORETE, se realiza un suavizado de los datos que permite la aplicación de los algoritmos de clasificación. En el análisis de datos faltantes se eliminaron todas las observaciones faltantes y aquellas variables cuyo número de datos faltantes es mayor al 50% del número total de observaciones totales. El grupo de variables reducido contiene sólo las variables que el árbol de decisión identificó como de mayor influencia sobre el ÁREA DEL FLORETE.

El clasificador basado en MLP tiene una capa oculta con 4 neronas y un solver de la familia del método quasi-Newton. La función de activación es la función RELU y conecta a las neuronas en la capa oculta. La matriz de confusión para este clasificador se muestra en la Figura 3.1. En esta matriz de confusión puede observarse como la mayoría de observaciones pueden ubicarse en una de las dos clases que existen y sólo un mínimo número de observaciones caen en la categoría de falsos positivos, es decir que no puede definirse a que clase pertenecen. Esto sugiere una eficiencia alta de este clasificador para el ÁREA DEL FLORETE.

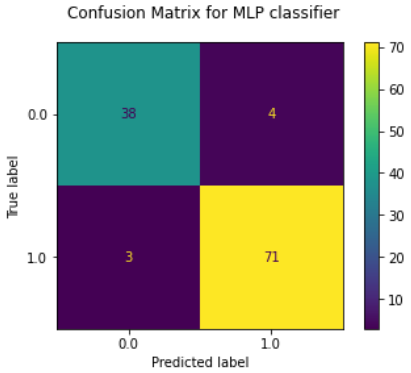

Figura 3.1. Matriz de confusión para el clasificador por MLP

Un comportamiento similar se presenta en el resto de clasificadores seleccionados. La mayoría es capaz de ubicar las observaciones en una de las dos categorías respecto al ÁREA DEL FLORETE. Sólo unas cuantas observaciones no pueden identificarse claramente y corresponden a los falsos positivos en la matriz. Es el caso del clasificador por Regresión Logística, K-vecinos, SVC, SGD y por árbol de decisión. Sólo el clasificador de Naive-Bayes no es capaz de clasificar adecuadamente las observaciones. La primera de las categorías la ubica completamente como falsos positivos. Las gráficas de las Figuras 3.2-3.6 muestran las matrices de confusión de los métodos seleccionados. La Figura 3.3 muestra la Matriz de confusión del clasificador menos efectivo de los que se utilizaron. En esta gráfica puede verse claramente que los datos de la primera clase no pueden ser ubicados de manera correcta y se toman como falsos positivos.

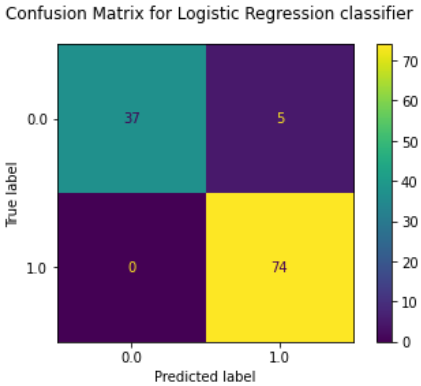

Figura 3.2. Matriz de confusión para el clasificador por Regresión Logística

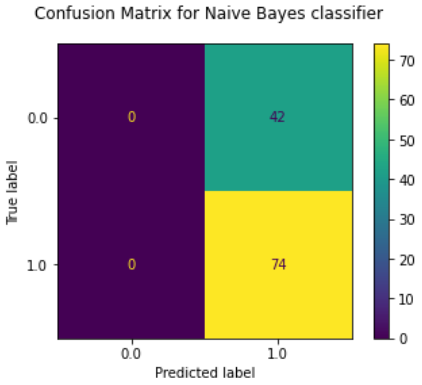

Figura 3.3. Matriz de confusión para el clasificador Naive-Bayes

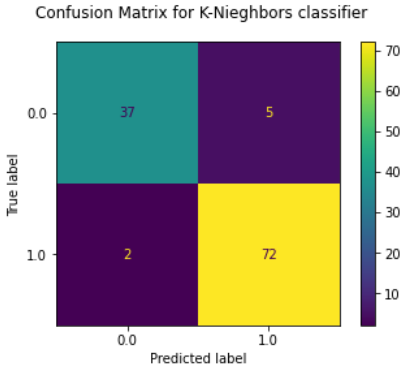

Figura 3.1. Matriz de confusión para el clasificador por K-vecinos

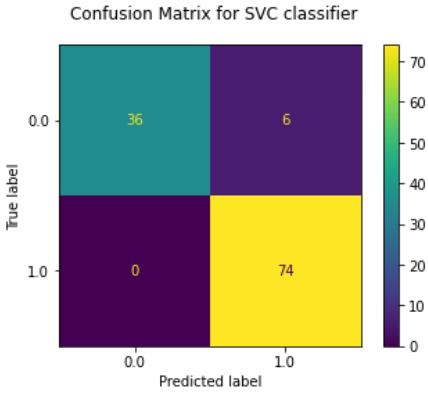

Figura 3.4. Matriz de confusión para el clasificador por SVC

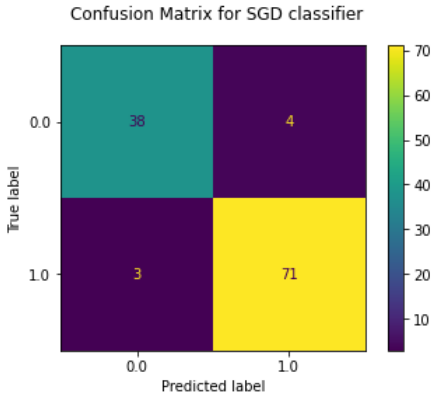

Figura 3.5. Matriz de confusión para el clasificador por SGD

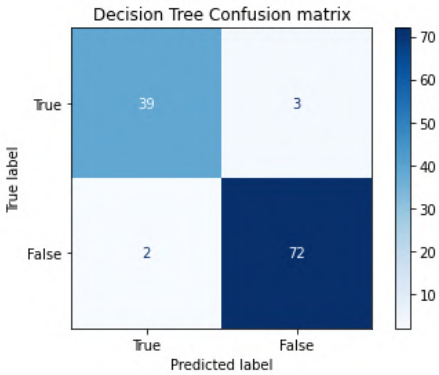

Figura 3.6. Matriz de confusión para el clasificador por árbol de decisión

Una forma más precisa y cuantitativa de medir la eficiencia de cada método consiste en graficar los valores de eficiencia de cada método y comparar la diferencia entre los métodos. La Figura 3.7 muestra la eficiencia obtenida en cada clasificador para la misma base de datos. Nótese que se ha excluido el valor de la eficiencia del clasificador por Naive-Bayes para una observación más precisa en la escala de valores. La Tabla III muestra los valores de eficiencia que se usaron para ilustrar la gráfica de la Figura 3.7.

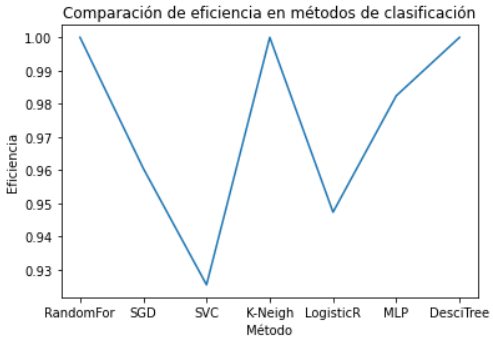

Figura 3.7. Eficiencia de cada algoritmo de clasificación.

TABLA III. VALORES DE EFICIENCIA DE LOS ALGORTIMOS DE CLASIFICACIÓN

Algoritmos Clasificadores	Eficiencia
Multi Layer Prceptron MLP	0.98
Logistic Regressión	0.94
Naive Bayes	0.77
K-Neighbors	0.99
Support Vector SVC	0.92
Gaussian SGD	0.96
Aleatorian Decisión Tree	0.99

IV. DISCUSIÓN, CONCLUSIÓN Y RECOMENDACIONES

En este artículo exploramos la aplicación de los algoritmos de clasificación del aprendizaje automático supervisado, que son más comunes en la literatura. Un algoritmo de clasificación identifica de manera automática la clase a la que pertenecen los datos. La importancia del problema de clasificación radica en identificar cuando una planta de brócoli está lista para cosecharse. El criterio que se usa en este trabajo para establecer si una planta de brócoli puede cosecharse es el valor de su ÁREA DEL FLORETE. Las variables más significativas a esta variable fueron CABEZA PESO FRESCO, TALLO PESO FRESCO, DIAMETRO TALLO BASE y DIAMETRO TALLO 15 CM. La aplicación de siete clasificadores: MLP, Regresión Logística, Naive-Bayes, K-Vecinos, SVC, SGD y Árbol de DecisióN, demostró la utilidad de la mayoría de estos algoritmos del Aprendizaje Automático Supervisado. Sólo el clasificador de Naive-Bayes obtuvo una eficiencia baja del 77%. El resto de los clasificadores pueden considerarse como muy buenos para resolver el problema de clasificación planteado. Los mejores de ellos fue el clasificador por K-Vecinos, y ambos clasificadores de decisión, tanto el clásico como el bosque aleatorio con una eficiencia perfecta de 100%. Finalmente puede decirse que los clasificadores por K-Vecinos y los de Árbol de Decisión, pueden utilizarse para ubicar con seguridad, y en base a las variables: CABEZA PESO FRESCO, TALLO PESO FRESCO, DIAMETRO TALLO BASE y DIAMETRO TALLO 15 CM; cuando la planta del brócoli está lista para cosecharse.

V. AGRADECIMIENTOS

Se agradece el uso de los datos a la empresa de hortalizas LA HUERTA que proporcionó los datos de la cosecha del brócoli durante el año 2019. Actualmente se siguen desarrollando proyectos con esta empresa para la predicción del rendimiento de la producción. Este proyecto es un proyecto social y replicable a otras empresas agrícolas de la Comarca Lagunera.

VI. REFERENCIAS

[1] Dufault, R.J., 1997. Determining heat unit requirements for broccoli harvestin coastal south carolina. Journal of the American Society for Horticultural Science 122, 169–174.
[2] Fujime, Y., Okuda, N., 1994. The physiology of flowering in brassicas, especially about cauliflower and broccoli, in: ISHS Brassica Symposium-IX Crucifer Genetics Workshop 407, pp. 247–254.
[3] Grevsen, K., 1998. Effects of temperature on head growth of broccoli (brassica oleracea l. var. italica): Parameter estimates for a predictive model. The Journal of Horticultural Science and Biotechnology 73, 235–244.
[4] Hara, P., Piekutowska, M., Niedbala, G., 2021. Selection of independent variables for crop yield prediction using artificial neural network models with remote sensing data. Land 10, 609.
[5] Jaramillo, N., D´ıaz, D., et al., 2006. El cultivo de las cruciferas: Brocoli, coliflor, repollo, col china. Technical Report. Corporacion Colombiana de Investigacion Agropecuaria.
[6] Maroto, J., 1989. Horticultura herb´acea. Ediciones Mundiprensa Madrid–Espa˜na. Pág. 45–53.
[7] Maroto Borrego, J.V., Baixauli Soria, C., 2017. Cultivos hort´ıcolas al aire libre. Cajamar Caja Rural
[8] Marshall, B., Thompson, R., 1987. Applications of a model to predict the time to maturity of calabrese brassica oleracea. Annals of Botany 60, 521–529.
[9] Pearson, S., Hadley, P., Wheldon, A., 1993. A reanalysis of the effects of temperature and irradiance on time to flowering in chrysanthemum (den-dranthema grandiflora). Journal of horticultural science 68, 89–97.
[10] Wurr, D., Fellows, J.R., Hambidge, A.J., 1992. The effect of plant density on calabrese head growth and its use in a predictive model. Journal ofHorticultural Science 67, 77–85. 57

VII. BIOGRAFÍA

Juárez Del Toro Raymundo. Docente de asignatura en el Instituto Tecnológico Superior de Lerdo TECNM. Pertenece al Sistema Nacional de Investigadores SNI del CONACYT. Asesor de tesis para estudiantes del posgrado en Ingeniería Mecatrónica. Es Físico Matemático del Instituto Politécnico Nacional IPN, en la ciudad de México. Obtuvo una maestría y un doctorado en ciencias en Teoría de Control Automático en el Centro de Investigación y de Estudios Avanzados, CINVESTAV-IPN, en la ciudad de México. Sus líneas de investigación son en su mayoría multidisciplinarias. Desarrolla proyectos multidisciplinarios a través del uso de herramientas matemáticas e ingenieriles. Algunas de estas herramientas son: control óptimo, ciencia de decisiones, inteligencia artificial, control robusto, métodos matemáticos, métodos numéricos, gestión eficiente de la energía, industria 4.0, robótica, visión artificial, entre otros.

Salas Francisco G. Es ingeniero electrónico por el Instituto Tecnológico de La Laguna, Torreón, México, y recibió los grados de Maestro y Doctor en Ciencias en Ingeniería Eléctrica, de la misma institución, en 2007 y 2013, respectivamente.
Desde 2017 es profesor investigador en la Facultad de Contaduría y Administración de la Universidad Autónoma de Coahuila, Torreón, México. Pertenece al Sistema Nacional de Investigadores (SNI) del CONACyT. Es líder de un cuerpo académico en su institución de adscripción. Ha dirigido varios proyectos de investigación y desarrollo tecnológico con la participación de instituciones educativas y empresas privadas. Sus intereses en investigación incluyen

Año: 2022 Volumen: 1
Número 8, ISSN: 2448-623X

inteligencia computacional aplicada a la toma de decisiones, control difuso y modelado y control de robots.

Castrejón Lozano Juan Gerardo es originario de Torreón, Coahuila. Terminó sus estudios de ingeniería en electrónica en el 2002. Obtuvo el grado de Maestro en Ciencias en 2005 y de Doctor en Ciencias en el 2008, en el Instituto Tecnológico de la Laguna. Actualmente trabaja como Profesor-Investigador en la Universidad Politécnica de Gómez Palacio. Entre sus temas de interés se encuentran la programación de algoritmos numéricos de alto desempeño para la estimación y el aprendizaje automático, así como la ingeniería basada en conocimiento.