Universidad Nacional Experimental del Táchira

Vicerectorado Académico

Decanato de Docencia

Departamento de Ingeniería en Informática

Trabajo de aplicación profesional

Proyecto especial de grado

Diseño de un clasificador de tubérculos de papa criolla para diferentes densidades de siembra según el peso fresco por calibre empleando Bayes Naive.

Autor(es): Estefany C. Salas S.

C.I.: 24820631

estefany.salas@unet.edu.ve

Tutor(es): Rossana Timaure

rttg@unet.edu.ve

San Cristóbal, Julio de 2019.

Índice general

Introducción

En Colombia, el nombre papa criolla corresponde a un morfotipo que desarrolla tubérculos con piel y pulpa de color amarillo, esta variedad ha sido clasificada como *Solanum phureja*; según Fedepapa que es la entidad gremial de carácter privado que representa a los productores de Papa en Colombia, esta variedad de papa es un alimento de alto valor nutritivo y de excelentes calidades culinarias, además de ser una de las fuentes de proteína más económica y ser considerada un producto exótico por consumidores de Europa y Estados Unidos, lo que le permite posicionarse como un producto para exportación que normalmente se comercializa en presentaciones que van desde paquetes de papas congeladas a latas de papa.

Al ser un producto de alta demanda requiere altos controles de calidad y un manejo especializado de las variables que influyen en su crecimiento, por supuesto las industrias que trabajan con este tubérculo buscan opciones que les permitan mejorar la calidad de la papa incurriendo en los menores gastos, aquí entran variables como las que han sido investigadas por el presente, que son la densidad de siembra (distancia entre plantas y distancia entre surcos) y los pesos frescos de cada calibre, en investigaciones anteriores se ha determinado que existe una relación entre estas variables, dichas investigaciones señalan que la densidad de siembra tiene un efecto sobre los tamaños y cantidad de tubérculos que la misma va a producir, esto indicaría que esta variable podría ser usada por industrias para obtener la mayor cantidad de tubérculos que se adecuen a los tamaños que ellos desean.

Esta investigación ha tenido como objetivo crear un clasificador empleando *Bayes Naive* que teniendo como variables de entrada (características) los pesos frescos de cada calibre, logre identificar con un nivel alto de aciertos la densidad de siembra que fue usada en esa cosecha, y así poder establecer la relación entre estás variables de manera más precisa. Se usó como método de clasificación *Bayes Naive* porque ha demostrado dar un gran porcentaje de aciertos al momento de clasificar y no posee limitaciones como los métodos estadísticos clásicos o la regresión espacial.

Los datos observados para realizar el modelo de clasificación fueron recolectados previamente por un estudiante de la Universidad Nacional de Colombia, a partir de estos se planteó generar un modelo empleando regresión espacial con el cual se pudieran generar datos artificiales, para luego realizar el clasificador evaluando la curva característica operativa del receptor para el caso.

La herramienta informática que ha sido empleada para desarrollar el algoritmo y realizar la clasificación es Python. Python es un lenguaje de programación reconocido por poseer una sintaxis simple y ser fácil de aprender, además es eficiente, este lenguaje permite a más personas aprender a programar de manera sencilla y así concentrarse en los problemas que buscan resolver. Actualmente se puede observar una gran tendencia al uso de Python en grandes centros de investigación como el CERN (Organización Europea para la Investigación Nuclear) y por parte de científicos en ramas como la Bioinformática, Neurofisiología, Física, Matemáticas, etc. Esto es debido a la disponibilidad de librerías de visualización, procesamiento de señales, estadísticas, álgebra, etc.; de fácil utilización y que cuentan con muy buena documentación. Los paquetes de Python denominados NumPy (Python Númerico) y SciPy (Python Científico) son pilares para la realización de trabajos científicos capaces de emular las funciones de otros lenguajes netamente científicos como Matlab.

Este documento está compuesto de cinco capítulos en los cuales se describe el desarrollo de la investigación y están estructurados de la siguiente forma:

Capitulo 1. Preliminares: En este capítulo se describe el planteamiento del problema, el objetivo general planteado y los objetivos específicos que se llevaron a cabo para cumplir con el objetivo general de la investigación.

Capítulo 2. Fundamentos teóricos: Contiene los antecedentes y la perspectiva teórica, para el entendimiento de los conceptos relacionados con esta investigación.

Capítulo 3. Fundamentos metodológicos: Describe la metodología utilizada durante la investigación, el enfoque, tipo, nivel, diseño de la misma y la metodología con la que se desarrolló el algoritmo.

Capítulo 4. Desarrollo: Detalla el desarrollo de los pasos de la metodología, la comprensión de los datos, la preparación de los mismos, el desarrollo del algoritmo y su evaluación.

Capítulo 1

Preliminares

1.1. Planteamiento y formulación del problema

Los algoritmos, la estadística, la ingeniería, la optimización y diversas ramas de la ciencia computacional son parte de las herramientas que ayudan a describir un aspecto del mundo real a través de un conjunto de datos, que en muchos casos está conformado por las muestras que son representaciones de objetos reales y las características que son la descripción de dichos objetos. (Layton, 2015)

Según VanderPlas (2017) para realizar clasificación se pueden implementar los modelos *Bayes Naive* que son un grupo de algoritmos simples y rápidos mayormente acertados para conjuntos de datos grandes, estos clasificadores están construidos sobre la base de los métodos de clasificación Bayesiana que se apoyan en el teorema de Bayes que describe la relación de las probabilidades condicionales de las cantidades estadísticas. En la clasificación Bayesiana el interés es encontrar la probabilidad de una clase L de acuerdo a las características observadas, que se puede escribir como $P\left(L|caracteristicas\right)$ y el teorema de Bayes indica como expresar esto en términos de cantidades que pueden ser computarizadas más fácilmente:

$$P\left(L|caracteristicas\right) = \frac{P\left(caracteristicas|L\right)P\left(L\right)}{P\left(caracteristicas\right)}$$

Si se intenta decidir entre dos clases (L_1 y L_2) la manera de tomar la decisión es calcular el radio de la probabilidad posterior para cada clase:

$$\frac{P\left(L_1|caracteristicas\right)}{P\left(L_1|caracteristicas\right)} = \frac{P\left(caracteristicas|L_1\right)P\left(L_1\right)}{P\left(caracteristicas|L_2\right)P\left(L_2\right)}$$

El algoritmo basado en estas ecuaciones debe seguir los siguientes pasos según Layton (2015):

- Teniendo el conjunto de datos de entrenamiento se debe calcular los valores de la probabilidad de una característica para cada clase.
- Se computa la probabilidad de que un dato de muestra pertenezca a una clase.
- Se computa la probabilidad de que un dato pertenezca a una clase.
- Se ingresan valores de prueba al modelo y se prueba la clasificación.

En investigaciones anteriores a la presente como la de Misigo y Miriti (2016) en la cual se clasificaron manzanas según su variedad, queda en evidencia que el uso de clasificadores *Bayes Naive* posee un porcentaje de exactitud mayor sobre componentes de análisis comparadas con técnicas como la lógica difusa y la red neuronal artificial perceptrón multicapa, por lo tanto la propuesta de esta investigación fue implementar un clasificador de *Bayes Naive* para entrenar un modelo que teniendo como características o entradas los pesos frescos de cada calibre de papa (el calibre es una categorización de las papas según su tamaño) bajo la hipótesis de que lo lograra hacer con un alto nivel de aciertos en cuanto a la densidad de siembra en la que fueron plantadas las plantas de papa criolla. El desarrollo de este clasificador fue en Python que según la fundación de Python (2018) es un lenguaje de programación creado en 1990 por Guido Van Rossum que se caracteriza por ser sencillo de aprender, tener estructuras de data de alto nivel que son eficientes y un simple pero efectivo acercamiento con la programación orientada a objetos. Este lenguaje actualmente es muy usado en el mundo de la ciencia porque provee herramientas que facilitan a las personas enfocarse más en el problema que buscan resolver y no en como programarlo, un ejemplo de esto son los paquetes NumPy (Python Númerico) y SciPy (Python Científico) (Challenger-Pérez *et al*, 2014).

La clasificación de las papas usualmente se realiza de manera manual, existen algunos modelos de clasificación generados por métodos estadísticos no espaciales que incurren en la violación de sus supuestos para obtener un resultado y por lo tanto su uso es poco. Al implementar un clasificador Bayes Naive estos supuestos desaparecen y se puede realizar una clasificación acertada que pueda ser considerada como un reemplazo a la clasificación manual. Según la «Norma NTC 341. Industria Alimentaria - Papa para Consumo» la clasificación de tubérculos según su diámetro se realiza bajo las clases denominadas muy grande (mayores a 90mm), grande (65-90mm), mediana (45-64mm) y

pequeña (30-44mm). Por otra parte la clasificación comercial en Colombia se realiza según grados (calibres), donde el grado 0 corresponde a tubérculos con un diámetro mayor a 90mm, el grado 1 a diámetros entre 70 y 89mm, el grado 2 entre 50 y 69mm y el grado 3 entre 35 y 49mm (Buitrago et al, 2003). Sin embargo, los datos para esta investigación siguen la clasificación de Bernal (2017) quien clasificó los tubérculos según su diámetro en las categorías menores a 2cm, (2-4]cm, (4-6]cm y mayores de 6cm, estas categorías son denominadas calibres.

Los datos que fueron empleados para la construcción, validación y prueba del clasificador fueron los datos observados en el Centro agropecuario Marengo de la Universidad Nacional de Colombia, en el departamento de Cundinamarca.

Para realizar la evaluación de los resultados obtenidos para el conjunto de datos se empleó la curva ROC, que como explica Gönen (2017) la curva ROC (*Receiver Operating Characteristic Curve*) es una herramienta estadística para evaluar la precisión de predicciones independientemente de la fuente de las mismas.

1.2. Objetivos

1.2.1. Objetivo general

Diseñar un clasificador de tubérculos de papa criolla para diferentes densidades de siembra según el peso fresco por calibre empleando *Bayes Naive*.

1.2.2. Objetivos específicos

- Diagnosticar el formato de las variables de entrada y salida para el clasificador.
- Establecer el tipo de algoritmo *Bayes Naive* a emplear y las características para el entrenamiento.
- Implementar el algoritmo de clasificación basado en *Bayes Naive*.
- Realizar las pruebas de funcionamiento y la comparación estadística.

1.3. Aportes de la investigación

Para la industria colombiana la papa criolla constituye un rubro muy importante, es un producto versátil que como lo indica Piñeros (2009) permite mediante varios procesos la obtención de papa criolla precocida y congelada, francesa precocida prefrita congelada y preformados, entre otros, para la realización de cada uno de estos productos es necesario emplear papas cuyas características se adecuen a lo que busca la industria, una característica resaltante al momento de clasificar papas es su peso. Cuando una industria cosecha papas controla que el resultado sean papas que se adecuen al tipo de producto que luego van a generar, para ello usan diferentes métodos que les permiten controlar las características incurriendo en los menores gastos posibles.

Para realizar clasificaciones de papa criolla que se basen en la relación que hay entre la densidad de siembra y los pesos frescos de los calibres de papa no existe un método o sistema que realice dicha clasificación de manera asertiva, incluso, no existen muchos métodos que empleen algoritmos que sean capaces de realizar la identificación con gran precisión empleando estadística, por eso se planteó realizar el clasificador empleando *Bayes Naive* y Python, *Bayes Naive* por ser un algoritmo de fácil implementación que ha demostrado en investigaciones pasadas tener un alto nivel de aciertos y no depende del supuesto de independencia, y se empleó Python como lenguaje de programación por ser un lenguaje que posee una sintaxis simple, un alto nivel de efectividad y librerías dedicadas al desarrollo científico, lo que permitió que el algoritmo que se diseñó sea de fácil uso para personas del campo agronómico, es decir, es una herramienta base para investigadores que quieran emplear sus propios clasificadores y evaluarlos con curvas ROC pero no tengan un alto nivel de instrucción en la parte informática.

Capítulo 2

Fundamentos teóricos

2.1. Antecedentes

Noor Amaleena Mohamad, Noorain Awang Jusoh, Zaw Zaw Htike y Shoon Lei Win , 2014 en su trabajo, *Bacteria Identification from Microscopic Morphology using Naïve Bayes*, tenían como objetivo de la investigación proponer un marco automatizado de identificación de bacterias que pudiera clasificar tres famosas clases de bacterias llamadas *Cocci*, *Bacilli* y *Vibrio* desde la morfología microscópica usando el clasificador *Bayes Naive*, para desarrollar el marco la investigación la realizaron en dos fases, la primera fue el entrenamiento del sistema empleando un conjunto de imágenes microscópicas que contenían *Cocci*, *Bacilli* y *Vibrio*, las imágenes de entrada fueron normalizadas para enfatizar el diámetro y forma de las características. En la segunda etapa emplearon el clasificador *Bayes Naive* para realizar inferencia probabilística basada en los descriptores de entrada. Para el entrenamiento utilizaron 65 imágenes de cada clase de bacteria, para las pruebas usaron 222 imágenes que poseían las tres clases de bacterias e imágenes aleatorias de humanos y aviones, durante las pruebas el sistema fue capaz de discriminar correctamente entre las tres clases de bacterias e incluso logró rechazar las imágenes que no pertenecían a ninguna de las tres clases de bacterias, como conclusión la investigación demostró como un clasificador con unas cuantas características basadas en imágenes puede proveer una alta exactitud en la identificación de bacterias según su morfología microscópica, este marco de identificación que consiste en la extracción y clasificación ha logrado un 80 % de exactitud al clasificar las tres bacterias (*Cocci*, *Bacilli* y *Vibrio*, a pesar de su naturaleza exploratoria se considera que se debe realizar más trabajo para lograr una clasificación robusta y de mayor exactitud empleando aprendizaje automático no solo para bacterias sino para cualquier otro objeto clasificable.

Misigo Ronald y Miriti Evans, 2016 en su trabajo *Classification of Selected Apple Fruit Varieties using Naïve Bayes* estudiaron la necesidad de distinguir variedades de manzanas de una manera rápida y no destructiva lo que motivó la investigación que tenía como objetivo principal investigar la aplicabilidad y el rendimiento del algoritmo de clasificación de *Bayes Naive* para distinguir las variedades de manzanas, la metodología aplicada involucró la adquisición de las imágenes, preprocesamiento y segmentación, análisis y clasificación de las variedades de manzanas. Realizaron un muestreo aleatorio y se emplearon 60 imágenes para el entrenamiento del clasificador, 30 imágenes para la validación y 60 imágenes para las pruebas, los resultados fueron positivos verdaderos, positivos falsos, negativos verdaderos y negativos falsos, luego se evaluó el rendimiento del sistema, obteniendo los valores estimados para exactitud, sensibilidad, precisión y especificidad donde se obtuvo 91 %, 77 %, 100 % y 80 % respectivamente, en conclusión, la clasificación empleando *Bayes Naive* resultó en un porcentaje mayor de exactitud que las técnicas de lógica difusa y MLP-Neural que habían sido empleadas previamente para realizar la tarea de clasificación.

Arias Victoria, Bustos Patricia y Ñústez Carlos en 1996 realizaron el trabajo *Evaluación del Rendimiento en Papa Criolla (Solanum phureja) variedad «Yema de Huevo», bajo diferentes Densidades de Siembra en la Sabana de Bogotá* donde evaluaron el rendimiento de la papa criolla, bajo diferentes densidades de siembra, utilizaron cuatro distancias entre surcos (0.70;0.80;0.90 y 1.0m), en Cundinamarca, Colombia. Las variables de rendimiento que evaluaron fueron: peso y número de tubérculos de primera, segunda y tercera clase por metro cuadrado, y también peso y número total de tubérculos por metro cuadrado. Las diferentes densidades que evaluaron no presentaron diferencias significativas para el número y peso de tubérculos de primera y segunda clase, para las distancias entre surcos menores de un metro, encontraron incrementos significativos en el peso total de tubérculos, pero se redujo el tamaño promedio de los mismos, es decir, que obtuvieron mayor número y peso de tubérculos de tercera clase.

Paraskevas Tsangaratos y Ioanna Ilia en 2016 realizaron la investigación *Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size*, teniendo como objetivo principal de la investigación comparar el rendimiento de un clasificador que implementa regresión logística con uno que implementa el algoritmo de *Bayes Naive* en evaluaciones de susceptibilidad a deslizamientos. El estudio que realizaron proporcionó una evaluación sobre la influencia de la complejidad del modelo y el tamaño del conjunto de datos de entrenamiento, mientras que identificaron el clasificador más preciso y confiable. La comparación de los dos clasificadores se basó en la evaluación de una base de datos que contiene 116 sitios ubicados en las montañas de Epiro, Grecia, donde se han

encontrado eventos graves de derrumbes. Los sitios estaban clasificados en dos categorías, áreas sin deslizamientos de tierra y derrumbes. En particular, hicieron el análisis de siete variables: unidades geológicas de ingeniería, ángulo de la pendiente, aspecto de la pendiente, promedio anual de precipitación, distancia de la red fluvial, distancia de las características tectónicas y distancia de la red de carreteras. Implementaron el análisis de multicolinealidad y la selección de características para estimar la independencia condicional entre las variables y para clasificar las variables según su importancia en la estimación de la susceptibilidad al deslizamiento.

Mediante los procesos mencionados anteriormente lograron la construcción de nueve conjuntos de datos diferentes, promover la partición les permitió crear subconjuntos de entrenamiento y validar datos de los 116 sitios originales. Cada conjunto de datos fue caracterizado por el número de variables utilizadas y el tamaño de los conjuntos de datos de entrenamiento. La comparación y validación de los resultados de cada modelo fue lograda utilizando medidas de evaluación estadística, la característica operativa de recepción y el área bajo las curvas de éxito y velocidad predictiva. Los resultados que obtuvieron indicaron que la complejidad del modelo y el tamaño del conjunto de datos de capacitación influyen en la precisión y la capacidad predictiva de los modelos concernientes a la susceptibilidad del deslizamiento. En particular, determinaron que el modelo más preciso con alto poder predictivo fue el octavo modelo (cinco variables y 92 datos de entrenamiento), con el clasificador *Bayes Naive* teniendo un rendimiento y precisión generales ligeramente más altos que el clasificador de regresión logística, 87.50 % y 82.61 % en los conjuntos de datos de validación, respectivamente. El área más alta bajo la curva fue lograda mediante el clasificador Naïve Bayes para los conjuntos de datos de entrenamiento y validación (0.875 y 0.806 respectivamente) mientras que el clasificador de regresión logística logró valores de AUC más bajos para los conjuntos de datos de capacitación y validación (0,844 y 0,711, respectivamente). Determinaron que cuando hay datos limitados disponibles, parece que se podrían obtener resultados más precisos y confiables mediante clasificadores generativos, como clasificadores *Bayes Naive*.

2.2. Bases Teóricas

2.2.1. Python

Python es un lenguaje de programación de código abierto creado por Guido van Rossum. Una de las ideas claves de van Rossum era que los programadores pasaban más tiempo leyendo código que escribiendo, entonces creo un lenguaje fácil de leer. Python es uno de los lenguajes de programación más populares y fáciles de aprender. Funciona en la mayoría de sistemas operativos

y computadoras y es usado desde la construcción de servidores web hasta crear aplicaciones de escritorio. (Althoff,2016)

2.2.2. PyCharm

PyCharm es un entorno de desarrollo integrado dedicado a Python y Django que provee un amplio rango de herramientas esenciales para programadores, que están estrechamente integrados para crear un entorno conveniente para el desarrollo productivo de Python.

2.2.3. Regresión espacial

El método de regresión espacial es un modelo estadístico para data observada en unidades geográficas como países o regiones, donde juegan un papel importante los vecinos, como indica Arbia (2014), este autor expresa que para tratar información espacial es necesario tener dos sets de información, el primero que posee los valores observados de las variables y el segundo que posee la ubicación particular donde esos valores fueron observados y las relaciones de proximidad entre todas las observaciones espaciales.

2.2.4. Probabilidad condicional

Koduvely(2015) indica que se define como la probabilidad de un evento, dado que ha ocurrido otro evento. Más formalmente, si tomamos las variables A y B, esta definición se puede reescribir de la siguiente manera:

$$P\left(A|B\right) = \frac{P\left(A,B\right)}{P\left(B\right)}$$

2.2.5. Teorema de Bayes

Hayter (2012) nos indica que para emplear las probabilidades de $P\left(A_i\right)$ y $P\left(B|A_i\right)$ a la hora de calcular $P\left(A_i|B\right)$ se debe considerar $P\left(A_1\right)$, ..., $P\left(A_n\right)$ como probabilidades previas de los eventos A_1, ...,A_n, sin embargo, la observación de los eventos B provee información adicional que permite obtener un conjunto de probabilidades posteriores, que es la probabilidad de los eventos $P\left(A_1\right)$, ..., $P\left(A_n\right)$ condicionada por B, es decir, $P\left(A_1|B\right)$, ...,$P\left(A_n|B\right)$.

Luego empleando la ley de la probabilidad total se calculan las probabilidades posteriores obteniendo el teorema de Bayes:

$$P(A_i|B) = \frac{P(A_i)\,P(B|A_i)}{P(B)} = \frac{P(A_i)\,P(B|A_i)}{\sum_{j=1}^{n} P(A_j)\,P(B|A_j)}$$

Según Koduvely(2015) Thomas Bayes empleó esta regla y formuló su famoso teorema de Bayes que puede interpretarse si $P(A)$ representa el grado inicial de creencia (o probabilidad previa) en el valor de una variable aleatoria A antes de observar B; luego, su probabilidad posterior o grado de creencia después de que se tenga en cuenta B se actualizará de acuerdo con la ecuación anterior. Entonces, la inferencia bayesiana se corresponde esencialmente con la actualización de las creencias sobre un sistema incierto después de haber hecho algunas observaciones al respecto. En el sentido, esta es también la forma en que los seres humanos aprendemos sobre el mundo. Por ejemplo, antes de visitar una nueva ciudad, tendremos cierto conocimiento previo sobre el lugar después de leer libros o en la Web. Sin embargo, poco después de llegar al lugar, esta creencia se actualizará en función de nuestra experiencia inicial del lugar. Actualizamos continuamente la creencia a medida que exploramos la ciudad nueva cada vez más.

2.2.6. Inferencia Bayesiana

Las estadísticas clásicas o frecuentistas normalmente consideran que cualquier dato físico que genere datos que contengan ruido puede modelarse mediante un modelo estocástico con valores fijos de parámetros. Los valores de los parámetros se aprenden de los datos observados a través de procedimientos tales como la estimación de máxima verosimilitud. La idea esencial es buscar en el espacio de parámetros para encontrar los valores de los parámetros que maximicen la probabilidad de observar los datos vistos hasta el momento. Ni la incertidumbre en la estimación de los parámetros del modelo a partir de los datos, ni la incertidumbre en el modelo mismo que explica los fenómenos en estudio, se tratan de una manera formal. El enfoque Bayesiano, por otro lado, trata todas las fuentes de incertidumbre usando probabilidades. Por lo tanto, ni el modelo para explicar un conjunto de datos observado ni sus parámetros son fijos, pero se tratan como variables inciertas. La inferencia bayesiana proporciona un marco para aprender la distribución completa de los parámetros del modelo, no solo los valores, que maximizan la probabilidad de observar los datos dados. El aprendizaje puede provenir tanto de la evidencia proporcionada por los datos observados como del conocimiento de dominio de los expertos. También hay un marco para seleccionar el mejor modelo entre la familia de modelos adecuados para explicar un determinado conjunto de datos. (Koduvely,2015)

2.2.7. Clasificador Bayes Naive

El clasificador *Bayes Naive* es un clasificador probabilístico basado en el teorema de Bayes, considerando la suposición de independencia ingenua. Los clasificadores Bayes Naive suponen que el efecto de un valor de variable en una clase dada es independiente de los valores de otra variable. Esta suposición se llama independencia condicional de clase. Bayes Naive a menudo puede realizar métodos de clasificación más sofisticados, es particularmente adecuado cuando la dimensionalidad de las entradas es alto. Cuando se quieren resultados más competentes, en comparación con otros métodos de salida, podemos usar la implementación de este clasificador que crea modelos con capacidades predictivas.(Misigo y Miriti, 2016)

Como indica Koduvely (2015), el nombre *Bayes Naive* proviene de la suposición básica en el modelo de que la probabilidad de una característica particular X_i es independiente de cualquier otra característica X_j dada la etiqueta de la clase C_k. Esto implica lo siguiente:

$$P\left(X_i|C_k, X_j\right) = P\left(X_i|C_k\right)$$

Usando esta suposición y la regla de Bayes, se puede mostrar que la probabilidad de clase C_k, características dadas, viene dada por:

$$P\left(C_k|\left\{X_1, X_2, ..., X_n\right\}\right) = \frac{1}{P\left(X_1, X_2, ..., X_n\right)} P\left(C_k\right) \prod P\left(X_i|C_k\right)$$

Donde, $P\left(X_1, X_2, ..., X_n\right)$ es el término de normalización obtenido al sumar el numerador en todos los valores de k. También se denomina evidencia bayesiana o función de partición Z. El clasificador selecciona una etiqueta de clase como la clase objetivo que maximiza la probabilidad de clase posterior $P\left(C_k|\left\{X_1, X_2, ..., X_n\right\}\right)$.

2.2.8. Bayes Naive Bernoulli

Como define Bonaccorso (2017) Si X es una variable aleatoria que se ajusta a la distribución de Bernoulli se puede asumir que solo tiene dos valores posibles que en este caso serán 0 y 1, y su probabilidad es:

$$P\left(X\right) = \begin{cases} p & \text{si } X = 1 \\ q & \text{si } X = 0 \end{cases}$$

$$Donde\ q = 1 - p\ y\ 0 < p < 1$$

Cuando se aplica el clasificador Bayes naive con esta representación, se asume que la aparición de un valor es independiente de la aparición del otro, y se obtiene la frecuencia de cada valor dentro de los datos. (Koller y Friedman, 2009)

2.2.9. Bayes Naive Multinomial

Según Bonaccorso (2017) una distribución multinomial es útil para modelar vectores de características donde cada valor representa, por ejemplo, el número de apariciones de un término o su frecuencia relativa. Si los vectores de características tienen n elementos y cada uno de ellos puede asumir k valores diferentes con probabilidad p_k, entonces:

$$P\left(X_1 = x_1 \cap X_2 = x_2 \cap ... \cap X_k = x_k\right) = \frac{n!}{\prod_i x_i!} \prod_i p_i^{x_i}$$

Las probabilidades condicionales $P\left(X_i|y\right)$ se calculan con un recuento de frecuencia (que corresponde a aplicar un enfoque de máxima verosimilitud), pero en este caso, es importante considerar el parámetro alfa (denominado factor de suavizado de Laplace). Su valor predeterminado es 1.0 e impide que el modelo establezca probabilidades nulas cuando la frecuencia es cero. Es posible asignar todos los valores no negativos; sin embargo, los valores más grandes asignarán mayores probabilidades a las características faltantes y esta opción podría alterar la estabilidad del modelo.

2.2.10. Bayes Naive Gaussiano

De acuerdo con Bonaccorso (2017) es útil cuando se trabaja con valores continuos cuyas probabilidades se pueden modelar usando una distribución gaussiana:

$$P\left(x\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Las probabilidades condicionales $P\left(x_i|y\right)$ también están distribuidas en Gauss; por lo tanto, es necesario estimar la media y la varianza de cada uno de ellos utilizando el enfoque de máxima verosimilitud. Esto es bastante fácil; de hecho, considerando la propiedad de un gaussiano, se obtiene:

$$L\left(\mu; \sigma^2; x_i|y\right) = \log \prod_k P\left(x_i^{(k)}|y\right) = \sum_k \log P\left(x_i^{(k)}|y\right)$$

En la formula anterior, el índice k se refiere a las muestras en el conjunto de datos y $P\left(x_i^{(k)}|y\right)$

es un gaussiano en sí mismo. Al minimizar el inverso de esta expresión se obtiene la media y la varianza para cada gaussiano asociado con $P\left(x_i^{(k)}|y\right)$ y, por lo tanto, el modelo está capacitado.

2.2.11. Curva característica operativa del receptor

Una curva característica operativa del receptor es una herramienta estadística para evaluar la precisión de predicciones, a menudo se abrevia como curva ROC o gráfico ROC, el último se utiliza con más frecuencia en la literatura de minería de datos.(Gönen,2007)

Las curvas ROC proporcionan una forma completa y visualmente atractiva de resumir la precisión de las predicciones. Son ampliamente aplicables, independientemente de la fuente de predicciones. También puede comparar la precisión de los diferentes métodos de generación de predicciones al comparar las curvas ROC de las predicciones resultantes.(Gönen,2007)

2.2.12. Glosario

Densidad de siembra : Se define como el número de plantas por unidad de área de terreno.

Hilera : Se define como una fila en la que son plantadas las semillas en distancias iguales.

Surco : Se define como hendidura que se hace en la tierra con el arado, esta hendidura se encuentra entre cada hilera de la siembra.

Tubérculo : Es un tallo subterráneo, donde se acumulan los nutrientes de reserva para la planta.

Capítulo 3

Fundamentos metodológicos

En este capítulo se detalla el enfoque, tipo, nivel, diseño de la investigación y metodología a implementar como estructura a seguir por el presente trabajo.

3.1. Enfoque de la investigación

Esta investigación posee un enfoque cuantitativo, que como lo indica Sampieri *et al* (2014) es secuencial y probatorio, cada etapa precede a la siguiente y no podemos "brincar" o eludir pasos. El orden es riguroso, aunque desde luego, se puede redefinir alguna fase. Parte de una idea que va acotándose y, una vez delimitada, se derivan objetivos y preguntas de investigación, se revisa la literatura y se construye un marco o una perspectiva teórica. De las preguntas se establecen hipótesis y determinan variables; se traza un plan para probarlas (diseño); se miden las variables en un determinado contexto; se analizan las mediciones obtenidas utilizando métodos estadísticos, y se extrae una serie de conclusiones.

3.2. Tipo o nivel de investigación

Según Pallela y Martins (2012) la investigación de campo consiste en la recolección directamente de la realidad donde ocurren los hechos, sin manipular o controlar variables, este proyecto planteó ese tipo de investigación ya que buscaba explorar los efectos de la interrelación entre los diferentes tipos de variables en lugar de los hechos.

El nivel de investigación es correlacional, porque se había planteado conocer la relación o

grado de asociación que existe entre dos o más conceptos, categorías o variables en una muestra o contexto particular, es decir, se buscó asociar variables mediante un patrón predecible para un grupo. (Sampieri *et al*, 2014).

3.3. Diseño de la investigación

El término diseño se refiere al plan o estrategia concebida para obtener la información que se desea con el fin de responder al planteamiento del problema (Sampieri *et al*, 2014), el diseño de la investigación presente es no experimental cuantitativa porque como lo indica Sampieri *et al* (2014) "es la investigación que se realiza sin manipular deliberadamente variables. Es decir, se trata de estudios en los que no hacemos variar en forma intencional las variables independientes para ver su efecto sobre otras variables. Lo que hacemos en la investigación no experimental es observar fenómenos tal como se dan en su contexto natural, para analizarlos".

De acuerdo al objetivo principal planteado por esta investigación se buscó clasificar la densidad de papa criolla a partir de sus pesos frescos, esto indica buscar la relación de estas variables, lo que conlleva a tener un diseño transeccional correlacional-causal que según Sampieri *et al* (2014) describen relaciones entre dos o más categorías, conceptos o variables en un momento determinado. A veces, únicamente en términos correlacionales, otras en función de la relación causa efecto (causales).

3.4. Metodología

El desarrollo de esta investigación en la que se había planteado la realización de un algoritmo que emplee clasificación de *Bayes Naive* comprende las siguientes fases metodológicas, según la metodología de CRISP-DM (*Cross Industry Standard Process for Data Mining*) explicada por la corporación IBM (2012):

Comprensión del problema.

Esta fase inicial se centró en la comprensión de los objetivos, requisitos y restricciones del proyecto desde una perspectiva no técnica, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. En esta fase, es muy importante la capacidad de poder convertir el conocimiento adquirido del problema en un plan preliminar cuya meta sea el alcanzar los objetivos del problema.

Las principales tareas que conformaron esta fase fueron las siguientes:

- Determinar los objetivos del problema.
- Evaluación de la situación.
- Determinación de los objetivos del proyecto propuesto.
- Producción de un plan del proyecto.

Comprensión de datos.

La segunda fase comprendió la recolección inicial de los datos, con el objetivo de establecer un primer contacto con el problema, familiarizándose con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis, si fuera el caso. Las principales tareas que se desarrollaron en esta fase del proceso fueron:

- Recolección de datos iniciales.
- Descripción de los datos.
- Exploración de datos.
- Verificación de la calidad de los datos.

Preparación de datos.

En esta fase se procedió a realizar la preparación de los datos para adaptarlos a las técnicas de clasificación, tales como técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para exploración de los datos. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato. Las principales tareas que estuvieron involucradas en esta fase fueron las siguientes:

- Selección de datos.
- Limpieza de los datos.
- Estructuración de los datos.
- Integración de los datos.
- Formato de los datos.

Modelado.

En esta fase de CRISP-DM, se seleccionó la técnica de modelado más apropiadas para el proyecto propuesto. ya que cumple con los siguientes criterios:

- Ser apropiada al problema.
- Disponer de datos adecuados.
- Cumplir los requisitos del problema.
- Tiempo adecuado para obtener un modelo.
- Conocimiento de la técnica.

Previamente al modelado de los datos se determinó un método de evaluación de los modelos que permitiera establecer el grado de bondad de ellos, para este caso la bondad de ajuste se evaluó mediante curvas ROC.

Evaluación.

En esta fase se evaluó el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerarse, además, que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. Se realizó la revisión del proceso, teniendo en cuenta los resultados obtenidos. Las tareas involucradas en esta fase del proceso fueron las siguientes:

- Evaluación de los resultados.
- Proceso de revisión.
- Determinación de futuras fases.

Capítulo 4

Desarrollo

En este capítulo se describe de manera detallada el desarrollo del clasificador de tubérculos siguiendo la metodología descrita en el capítulo 3.

4.1. Comprensión de datos

Recolección de datos iniciales.

Los datos empleados para realizar la clasificación fueron previamente recolectados por Bernal (2017). En un estudio que se realizó en el Centro agropecuario Marengo de la Universidad Nacional de Colombia, en el departamento de Cundinamarca (74°12'58.51 W; 4°40'52.92 N), el cual tiene una altitud de 2516 msnm, temperatura media de 14°C en un rango de 12°C a 18°C y precipitación media de 500 a 1000 mm, cuenta con un paisaje en planicie fluvio-lacustre y un relieve en terraza lacustre plana (que no excede al 1 %) con suelos moderadamente profundos y bien drenados. El régimen de humedad es rústico y un nivel freático a menos de 0.5m del 15 %. De acuerdo a las características de precipitación, temperatura y evapotranspiración, la zona se clasifica como Bosque Seco Montano Bajo. (Bernal, 2017)

El material vegetal utilizado corresponde al cultivo de papa criolla (*Solanum phureja*), utilizando el tubérculo como semilla con el tamaño y forma característica de la especie (tamaño mediano), ojos poco profundos, sin pudrición ni defectos en la piel. Esta variedad con un porte de planta medio y follaje verde claro, distinguida por su adaptación a días cortos, de origen y distribución en América del Sur, y con centro de diversidad genética al sur de Colombia. Con un desarrollo vegetativo que se da hasta los 35 días después de la siembra (dds), siguiendo la

oración hasta los 65 dds, fructificación hasta los 90 dds y finalmente la madurez y senescencia hasta los 120 dds. Esta variedad es precoz (120 días a 2600 msnm), su potencial de rendimiento en condiciones óptimas de cultivo es de 15 a 25 ton.ha−1, sin periodo de reposo y susceptible al virus del amarillamiento de las nervaduras de la hoja (*Potato yellow vein virus*). Se cultiva en las diferentes regiones de Colombia y en diferentes condiciones de suelo. Es la principal variedad de papa criolla cultivada en el país y hasta la presente es la variedad que se procesa para exportación como precocida congelada (Ñustez, 2011; Rodríguez y Ñustez, 2011).

A los 120 días de siembra se cosecharon los tubérculos y se contaron según su diámetro en las categorías 2 cm, (2-4] cm, (4-6] cm y > 6 cm. Para estudiar el efecto de la densidad de siembra se fijaron las distancias entre plantas de (30, 40 y 50 cm) todos con separación de 100 cm entre surcos. La siembra se realizó en surcos alineados con precisión según la densidad de siembra, utilizando tres surcos sucesivos según la geometría del lote para cada densidad, con dos repeticiones por densidad, lo que rindió un total de 18 surcos, para un total de 2841 plantas. Aunque la unidad que aportó cada dato fue la planta (tubérculos), la obvia dificultad para obtener una densidad de siembra aleatoria usando cada planta como unidad experimental, obligó a la aleatorización de las densidades de siembra, cada una con sus tres respectivos surcos (unidad experimental) dentro del lote, registrando los datos de cada planta (unidad de observación). Bajo estas condiciones, el diseño resultó ser una factorial simple en arreglo completamente al azar, tomando las distancias entre plantas como los niveles del factor. (Bernal, 2017)

Descripción de los datos.

Los datos se encuentran separados según sus características en atributos, el primero denominado planta que contiene el número que identifica cada planta, el segundo llamada densidad que posee valores de 1, 2 o 3, donde 1 implica una densidad de siembra de 30cm, 2 una densidad de 40cm y 3 una densidad de 50cm. Los siguientes cuatro atributos están identificadas como PD1, PD2, PD3 y PD4, cada uno posee valores continuos que representan el peso fresco de cada calibre en esa planta, los calibres son 4 y representan la categorización de los tubérculos según su diámetro, para el calibre 1 se toman tubérculos con diámetro menor o igual a 2cm, para el calibre 2 con diámetro mayor a 2cm y menor o igual a 4cm, en el calibre 3 mayores a 4cm y menores o igual a 6cm y finalmente el calibre 4 con tubérculos de diámetro mayor a 6cm. Los últimos dos atributos son X y Y, que indican mediante un valor entero la posición de dicha planta en la siembra. La cantidad total de plantas es 2839, para las densidades de siembra 1, 2 y 3 hay 1135, 926 y 778 plantas respectivamente, cada una con valores de PD1, PD2, PD3, PD4, X y Y.

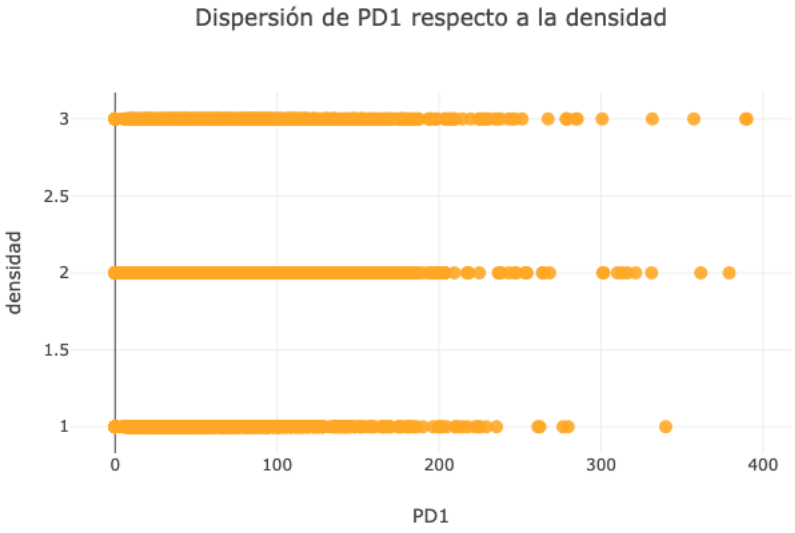
A continuación se observan los datos descriptivos de cada variable

	Densidad	PD1	PD2	PD3	PD4
Conteo	2839	2839	2839	2839	2839
Media	1.8742	74.1536gr	226.3262gr	213.3114gr	18.9698gr
Mínimo	1	0gr	0gr	0gr	0gr
Máximo	3	390.0000gr	1080.0000gr	1290.0000gr	775.9459gr
Mediana	2	62.0547gr	195.0000gr	164.7058	0.0000gr
Varianza	0.6582	2892.7971	26386.3844	36986.5868	3486.4316
Moda	1	0gr	0gr	0gr	0gr
Distribución	N/A	normal	normal	normal	normal

Cuadro 4.1: Datos Estadísticos Descriptivos.

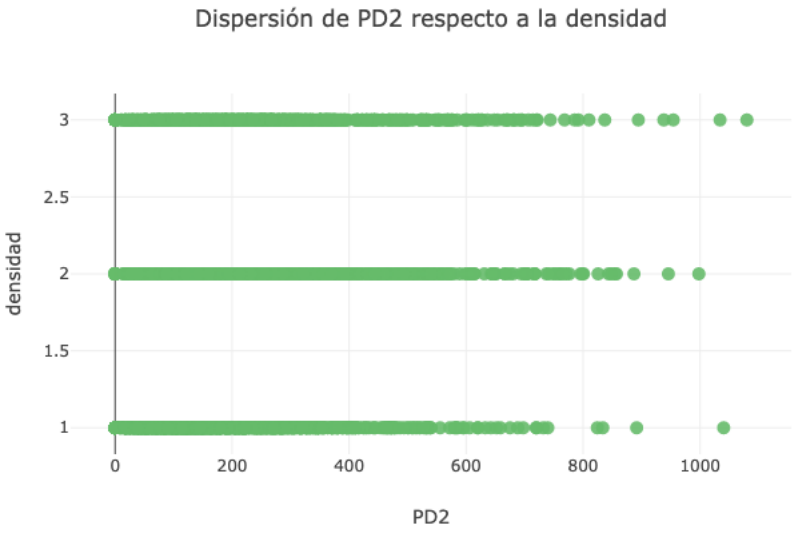
En el cuadro 4.1 se observan los datos estadísticos descriptivos de la densidad, PD1, PD2, PD3 y PD4; en cuanto a la densidad se puede establecer que la clase con mayor número de repeticiones es la 1 que representa una densidad de siembra de 30cm, las varianzas de las variables PD1, PD2, PD3 y PD4 son altas, esto indica que la data es heterogénea y se evidencia la existencia de datos extremos, en las siguientes gráficas de dispersión es pueden notar mejor estas observaciones.

Figura 4.1: Dispersión de la variable de pesos fresco del calibre 1 respecto a la densidad.



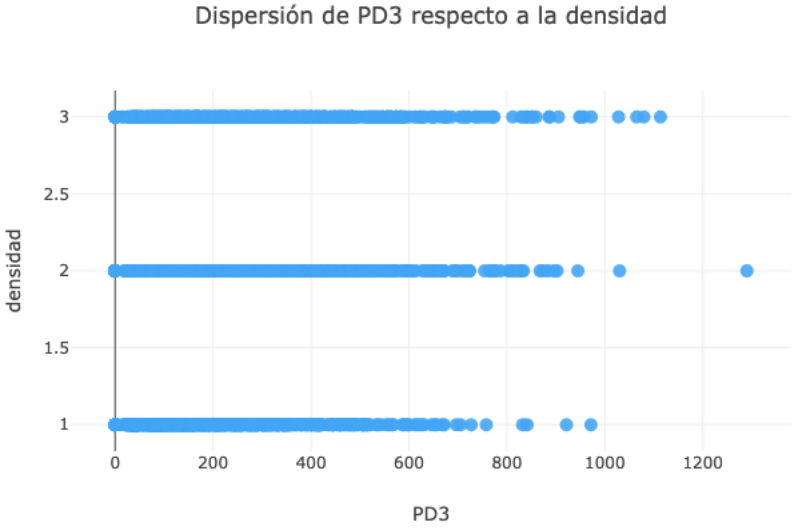
En la figura 4.1 se pueden evidenciar algunos valores extremos que incluyen al valor máximo de dicha variable que es 390gr.

Figura 4.2: Dispersión de la variable de pesos fresco del calibre 2 respecto a la densidad.



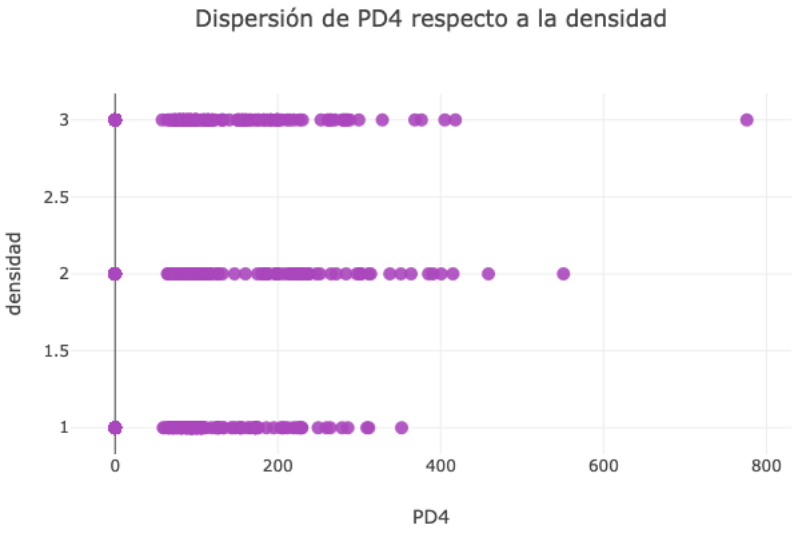
En la figura 4.2 también se pueden encontrar algunos valores extremos incluyendo al valor tope de dicha variable que es 1080gr.

Figura 4.3: Dispersión de la variable de pesos fresco del calibre 3 respecto a la densidad.



En la figura 4.3 que gráfica PD3 cuyo valor máximo es 1290gr se observan varios valores extremos.

Figura 4.4: Dispersión de la variable de pesos fresco del calibre 4 respecto a la densidad.



En la figura 4.4 se encuentran datos con valores extremos que se encuentran muy cercanos al mínimo y máximo.

Las figuras descritas anteriormente permiten determinar que se debe realizar limpieza de datos para evitar esos valores extremos o atípicos que afectan negativamente los cálculos durante la clasificación.

Exploración de datos.

Las correlaciones de las variables PD1, PD2, PD3 y PD4 para cada densidad demuestran la relación que existe entre dichas variables, en las siguientes figuras se pueden observar los valores calculados empleando el coeficiente de correlación de Spearman.

En las figuras 4.5, 4.6 y 4.7 se observan los valores de correlación entre las variables de peso para las densidades 1, 2 y 3, los valores oscilan entre 0 y 0.6 indicando que existe una correlación positiva entre los atributos PD1, PD2, PD3 y PD4.

Figura 4.5: Matriz de correlaciones para la densidad 1.

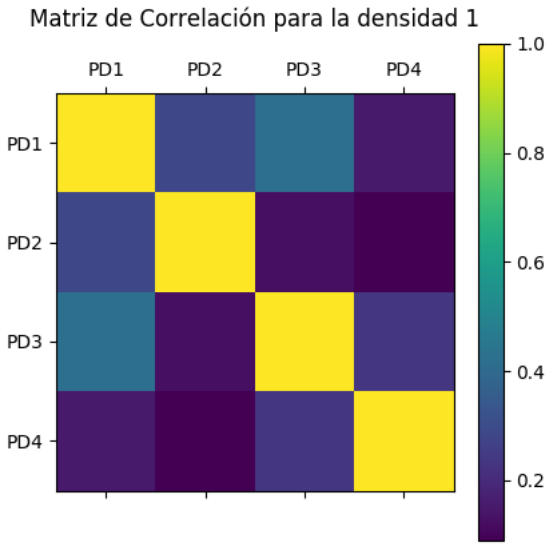


Figura 4.6: Matriz de correlaciones para la densidad 2.

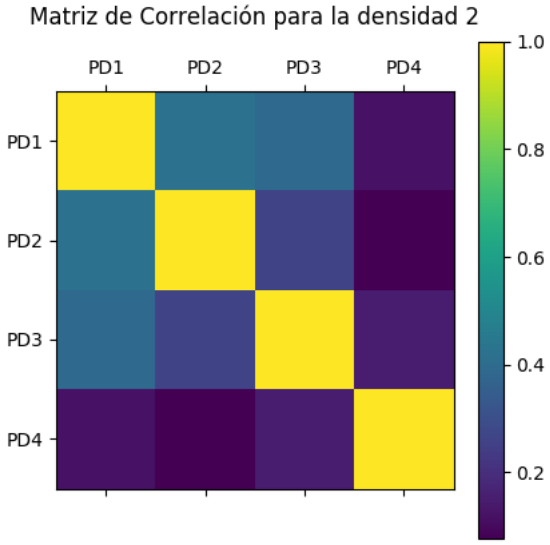
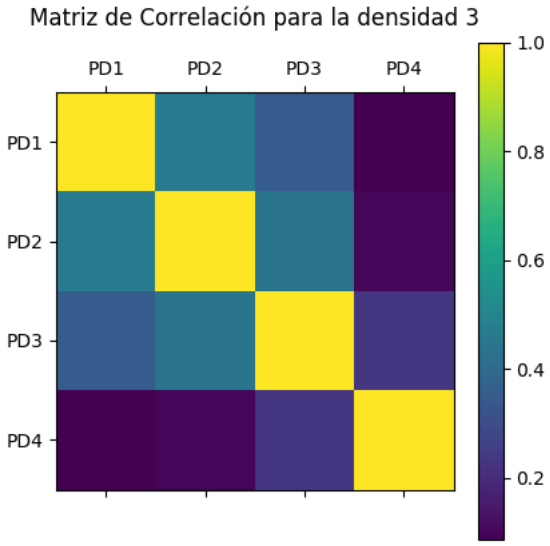


Figura 4.7: Matriz de correlaciones para la densidad 3.



Tomando en cuenta los estadísticos básicos de cada conjunto de datos se logra observar que los valores de varianza para las variables PD1, PD2, PD3 y PD4 son altos, esto podría afectar negativamente los cálculos del algoritmo de clasificación. A continuación se pueden observar los histogramas de las variables de entrada para cada densidad que permiten apreciar el patrón que siguen los datos.

Figura 4.8: Dispersión de la variable de pesos fresco del calibre 4 respecto a la densidad.

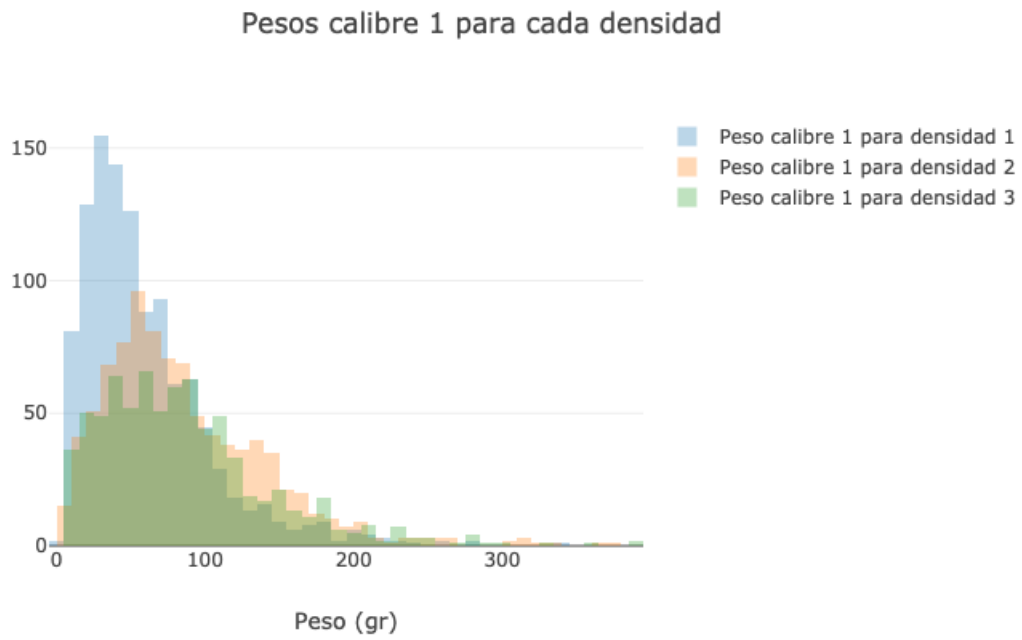


Figura 4.9: Dispersión de la variable de pesos fresco del calibre 4 respecto a la densidad.

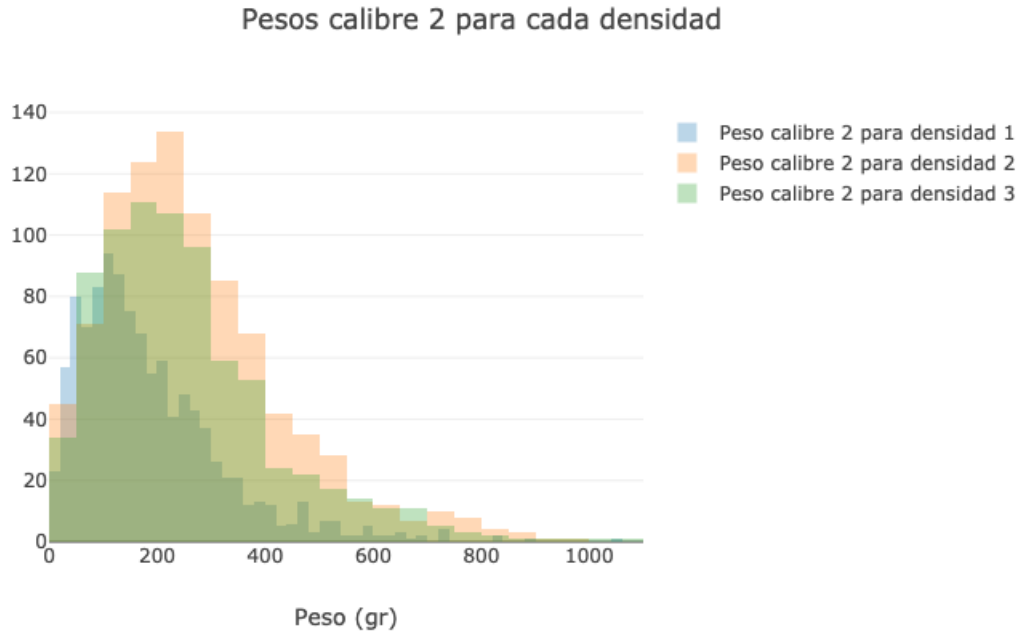


Figura 4.10: Dispersión de la variable de pesos fresco del calibre 4 respecto a la densidad.

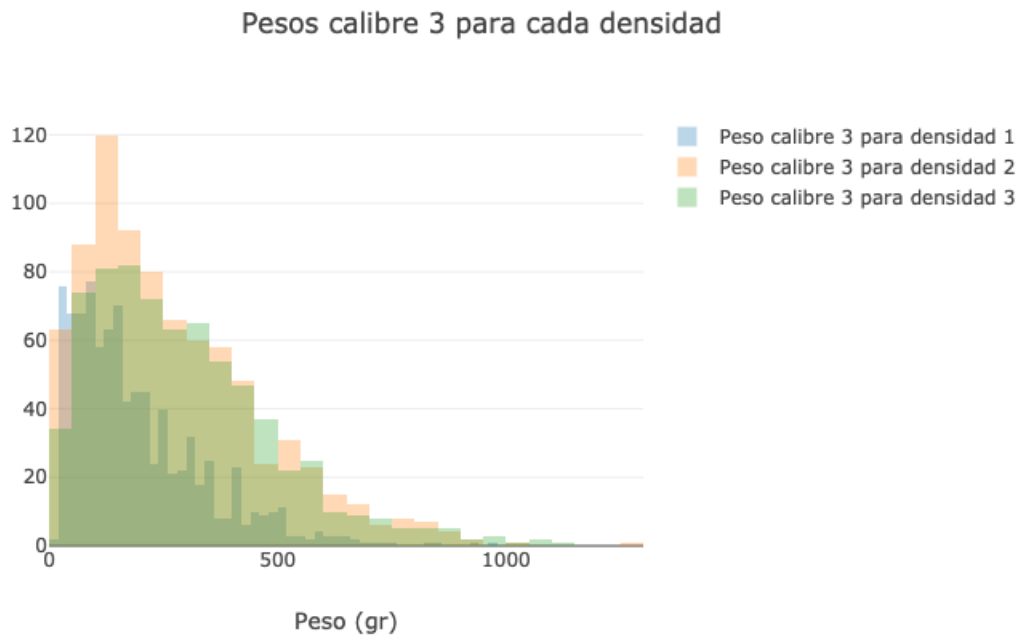
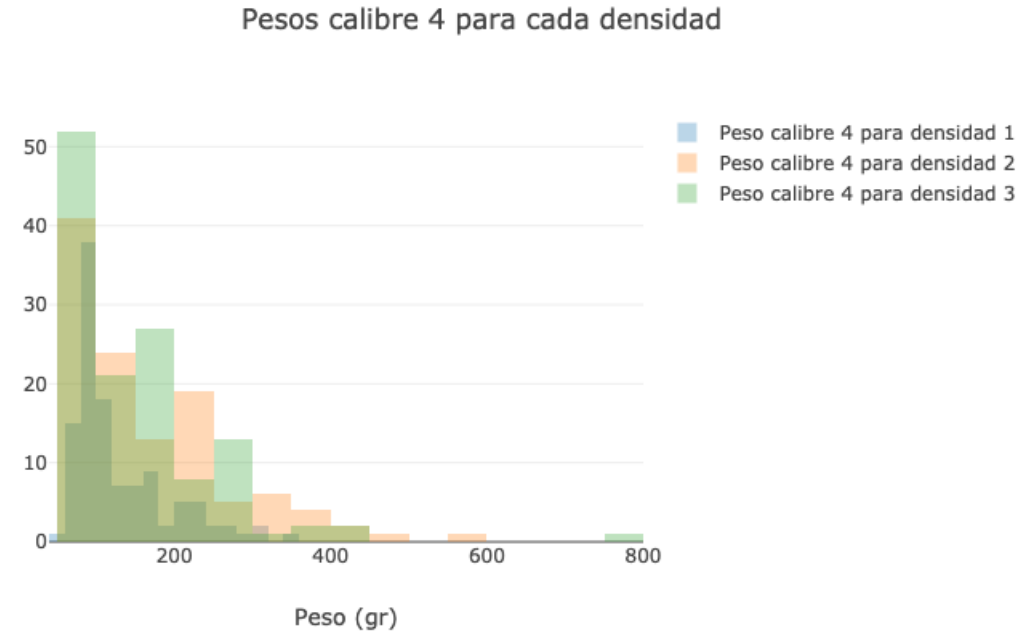


Figura 4.11: Dispersión de la variable de pesos fresco del calibre 4 respecto a la densidad.



En la figura 4.8 se encuentra el histograma de la variable PD1 para cada densidad, se puede ver el comportamiento de la variable demostrando diferentes patrones para cada densidad; en cuanto a la figura 4.9 se tiene a la variable PD2 que muestra patrones un poco mas similares para cada densidad, en la siguiente figura que es la 4.10 se evidencia la disminución de esta variable para la densidad 1 y finalmente en la figura 4.11 se hace evidente que la variable PD4 tiene valores menores respecto a las anteriores.

Los histogramas permiten determinar que la distribución de los datos de entrada se asemeja a una curva normal, lo que ayuda a inferir que se debe emplear el método gaussiano de Bayes Naive para realizar su clasificación.

Verificación de la calidad de los datos.

Luego de revisar los valores de los datos se determinó que no era necesaria la estandarización de los mismos, ya que todos se encuentran en el mismo rango en cuanto a valor y en la misma unidad de peso (gramos); además no hay valores faltantes, es decir, el set de datos está completo.

4.2. Preparación de datos

Selección de datos.

Los datos disponibles son: densidad, PD1, PD2, PD3, PD4, X y Y, sin embargo, las coordenadas de las plantas no son un dato relevante para la clasificación que se realizó porque la misma busca determinar la densidad sólo a partir de los pesos frescos por calibre, por lo tanto se excluyeron del set de datos.

Se debe tomar en cuenta que los datos de los pesos frescos por calibre fueron calculados empíricamente lo que podría implicar tener valores erróneos.

Limpieza de los datos.

No fue necesario realizar una limpieza de datos considerando la calidad de los mismos.

Estructuración de los datos.

No es necesario calcular atributos derivados para realizar la clasificación; no se debe realizar una normalización de índices porque los valores de entrada se encuentran en la misma escala y unidad, y la estimación de las varianzas indica que son homogéneos y que no existen datos atípicos o extremos.

Integración de los datos.

La media y varianza son dos datos combinados calculados a partir de los datos iniciales, estos nuevos datos son necesarios durante el cálculo probabilístico del algoritmo clasificador y se obtienen al principio de la ejecución llamando a la función **media_varianza** que retorna dos arreglos matriciales **m** y **v**, con un número de filas igual a la cantidad de clases, en este caso tres filas, y la cantidad de columnas es igual a la cantidad de características de entrada, es decir, cuatro para la data que se está empleando, cada campo de la matriz contiene la media o varianza (dependiendo de la matriz que se este revisando) para esa característica y esa clase.

Formato de los datos.

Los datos de entrada se encuentran ordenados según su clase, por lo tanto deben ser desordenados para permitir que el algoritmo de clasificación obtenga datos de cada clase para realizar los cálculos; para desordenar las filas de datos se empleó la función **random.shuffle** del paquete **numpy** para python, para que el reordenamiento se realizada de forma aleatoria.

4.3. Modelado

Técnica de modelado.

Para la clasificación de los datos se empleó un modelo probabilístico basado en el teorema de Bayes Naive para una distribución gaussiana, se seleccionó esta técnica por tener como supuesto la independencia entre las variables y su distribución espacial. Las pruebas del modelo se realizaron separando los datos en un set de entrenamiento que posee el 80 % de los datos y el otro 20 % se destino al set de pruebas, luego de obtener los resultados para el set de pruebas se realiza la

evaluación del modelo empleando la curva característica operativa del receptor.

El desarrollo del clasificador se realizó empleando Bayes Naive Gaussiano como se indica anteriormente, luego de leer los datos se llama a la función **bayes_naive_gaussiano** que recibe como entrada los siguientes parámetros:

- set_entrenamiento: es un array de numpy con las columnas número de planta, densidad, PD1, PD2, PD3, PD4, X y Y, su cantidad de filas depende de la cantidad de datos que sean destinados a formar parte del entrenamiento que en este caso es el 80 % de los datos iniciales, es decir, aproximadamente 2272 filas.
- set_test: es un array de numpy con las columnas PD1, PD2, PD3 y PD4, su cantidad de filas depende de la cantidad de datos que sean destinados a formar parte del set de pruebas que en este caso es el 20 % de los datos iniciales, es decir, aproximadamente 567 filas.
- clases_set_test: es un array de numpy y cada fila representa la clase (densidad) para los datos en el set de test en el mismo número de fila, su cantidad de filas depende de la cantidad de datos que sean destinados a formar parte del set de pruebas que en este caso es el 20 % de los datos iniciales, es decir, aproximadamente 567 filas.

Luego se ejecutan los siguientes pasos:

1. Cálculo de la media y varianza: como se indicó previamente los valores de las medias y varianzas de cada característica para cada clase son necesarios para realizar los cálculos probabilísticos, en este paso del algoritmo se llama a la función **media_varianza**, para realizar dicho procedimiento empleando para la media la fórmula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

   y para la varianza:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

2. Cálculo de probabilidades previas: ejecuta la función **pre_prob** que recibe un vector unidimensional con los valores de las clases del set de entrenamiento y aplica la fórmula

$$Prob\ Previa\,(c) = \frac{ocurrencias\ de\ la\ clase\ c}{Nro\ total\ de\ datos}$$

retornando un vector unidimensional con la probabilidad previa de cada clase.

3. Cálculo de la probabilidad posterior: es necesario calcular la probabilidad posterior del set de pruebas dada una clase **c**, para ello se llama al método **prob_caracteristica_clase** que recibe como parámetros la matriz de medias, la matriz de varianzas y el set de pruebas; tomando en cuenta la fórmula para el cálculo de probabilidades de una distribución gaussiana:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

se puede definir que la probabilidad posterior para una característica del set de prueba dada una clase **c** es:

$$P(x_i|c) = \frac{1}{\sqrt{2\pi\sigma^2_{x_i,c}}} e^{-\frac{(x_i-\bar{x}_{x_i,c})^2}{2\sigma^2_{\bar{x}_{i,c}}}}$$

donde x_i es una característica de esa fila del set de datos; el código de la función se encarga de obtener el resultado de esa fórmula para cada característica dentro de la fila del set de datos para la clase dada y luego multiplica los valores de dichas probabilidades obteniendo la probabilidad posterior para esa clase. Esta función retorna una matriz con una cantidad de filas igual a las filas del set de pruebas y por cada fila posee varios valores, cada uno de estos indica la probabilidad posterior para una clase.

4. Obtiene las probabilidades condicionales y las predicciones: finalmente para calcular la probabilidad condicional de una instancia de test se emplea la función **prob_condicional** que recibe como parámetros el set de pruebas, la cantidad de clases, el arreglo con las probabilidades posteriores, el arreglo con las probabilidades previas y las verdaderas clases del set de pruebas; empleando el teorema de Bayes obtiene la probabilidad condicional de cada clase para una fila del set de pruebas:

$$P(c_i|x) = \frac{P(c_i)\,P(x|c_i)}{\sum_{j=1}^{n} P(c_j)\,P(x|c_j)}$$

Luego se toma la probabilidad condicional mayor entre las calculadas para ese set de prueba obteniendo así la predicción.

Método de evaluación de los modelos.

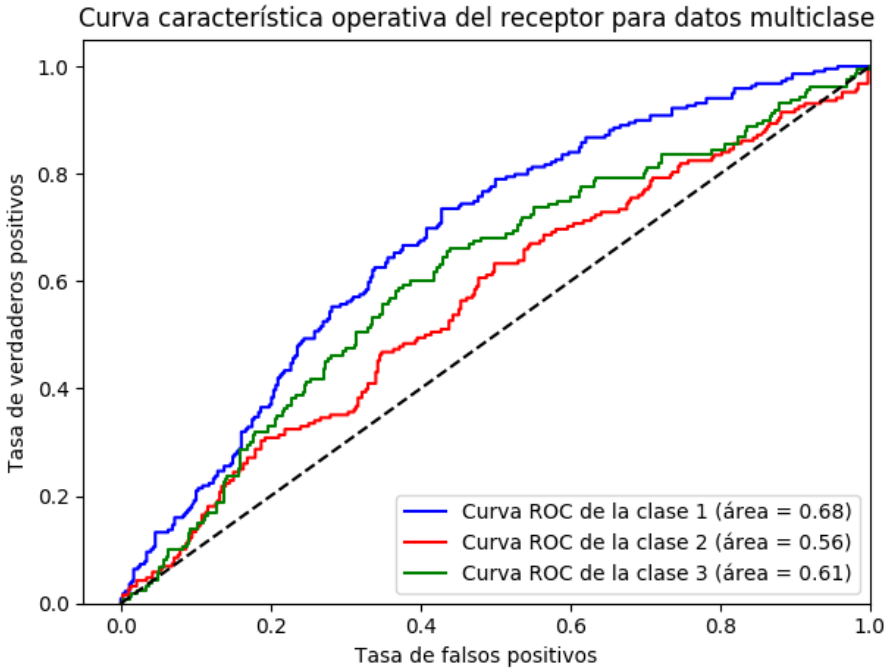
Para evaluar el modelo se emplean curvas ROC, primero se realiza el cálculo de la matriz de confusión empleando la función **confusion_matrix** del módulo **metrics** del paquete **sklearn** y posteriormente se realiza la curva característica operativa del receptor con el enfoque multiclase usando los métodos **roc_curve** y **auc** del mismo paquete, el primer método computa la curva ROC graficando las fracciones de verdaderos positivos y falsos positivos para distintos umbrales, y la función **auc** computa el área bajo la curva. Según Ato y López (1995) un área bajo la curva igual a 1 correspondería a una prueba diagnóstica perfecta y un área bajo la curva igual a 0.5 indica que se obtuvo una prueba sin precisión diagnóstica.

4.4. Evaluación

Evaluación de los resultados.

Cada vez que se ejecuta el algoritmo el mismo indica que obtuvo un porcentaje de aciertos de aproximadamente 45 % y de desaciertos de aproximadamente 55 %, en cada corrida los datos que emplea el algoritmo para entrenamiento y prueba varían por lo tanto estos valores pueden cambiar ligeramente. En cuanto a la curva característica operativa del receptor se obtiene:

Figura 4.12: Matriz de correlaciones para la densidad 3.



La curva ROC en la figura 4.12 posee trazos que indican que no es una mala clasificación pero tampoco es una clasificación óptima, si se observan los valores para el área bajo la curva de la densidad 1, 2 y 3 se obtuvieron 0.68, 0.56 y 0.61 respectivamente que según el estándar indicado anteriormente están entre una precisión regular y no tener precisión diagnóstica.

Se puede demostrar que el algoritmo está generando una clasificación empleando el método de Bayes Naive Gaussiano y está obteniendo los resultados del método de evaluación, aunque los mismos no sean los deseados.

Proceso de revisión.

Los objetivos de este proceso fueron diagnósticar el formato de las variables de entrada y salida para el clasificador, establecer el tipo de algoritmo de Bayes Naive a emplear y las características para el entrenamiento, realizar la implementación del algoritmo, realizar las pruebas de funcionamiento y la comparación estadística, todos estos objetivos fueron cubiertos durante el desarrollo de esta metodología y se logró el diseño de un clasificador de tubérculos de papa criolla para diferentes densidades de siembra a partir de los pesos frescos por calibre empleando Bayes Naive.

Determinación de futuras fases.

En cuanto al proceso actual algunas mejoras posibles podrían ser la eliminación de las características con valores iguales a cero o emplear otros métodos para clasificar los datos como las redes neuronales, se podría implementar un análisis espacial e incluso realizar una revisión más profunda a los datos, cualquiera de estás alternativas podría mejorar o empeorar el porcentaje de aciertos de la clasificación actual.

Según Rish (2001) existen características dentro de los datos que pueden afectar el rendimiento de Bayes Naive, como el impacto generado por la entropía de la distribución, la cantidad de información despreciada por asumir la independencia de las entradas, entre otros, sería bueno identificar estas características dentro del set de datos empleados para la prueba de este clasificador y así lograr determinar que está afectando la clasificación.

Capítulo 5

Conclusiones y recomendaciones

Se diseñó la solución algorítmica para el cálculo de la clasificación y predicción de la densidad de siembra de tubérculos de papa criolla a partir de sus pesos frescos por calibre empleando un clasificador Bayes Naive Gaussiano, se observó que dicho clasificador no genera predicciones óptimas para el set de datos empleado que corresponde a una cosecha realizada en el Centro agropecuario Merengo de la Universidad Nacional de Colombia; la naturaleza de los datos afecta la clasificación empleada, la mayor sospecha es la metodología empírica empleada en la clasificación de los tubérculos por calibre, es posible que las marcas de clase no sean las ideales para obtener datos que puedan ser discriminantes a la hora de realizar la clasificación.

Dentro del algoritmo se codificaron 6 funciones que se encargan de realizar los cálculos de media, varianza, probabilidades y evaluación de la clasificación, con el fin de que las mismas se adapten para distintas clasificaciones sin ser restringidos por el tamaño de los datos, la cantidad de entradas o la cantidad de clases realizando pocos ajustes.

Para la evaluación de la clasificación se empleó la curva característica operativa del receptor que es un método empleado ampliamente para la evaluación de modelos; el área bajo la curva ROC indica que los resultados del clasificador van desde no tener precisión diagnóstica a tener una precisión regular.

Para futuras pruebas de clasificación sobre el mismo set de datos se recomienda probar diferentes métodos y modelos de clasificación que por tener diferentes supuestos podrían llegar a resultados con mayor precisión, además realizar pruebas eliminando los valores iguales a 0 dentro del set de datos. En cuanto al algoritmo de clasificación, para realizar más pruebas del mismo se podrían

emplear distintos sets de datos con características estadísticas diferentes para determinar cuales son sus capacidades. Todas estás recomendaciones son procedimientos que se encontraban fuera del alcance previamente definido para este proyecto de investigación.

Referencias Bibliográficas

Althoff, C. (2016). «The Self-taught Programmer». Cory Althoff.

Arbia, G. (2014). «A Primer for Spatial Econometrics With Applications in R». Palgrave Macmillan.

Arias, V., Bustos, P., Ñustéz, C. (1996). «Evaluación del Rendimiento en papa criolla (Solanum phureja) variedad "Yema de Huevo", bajo diferentes Densidades de Siembra en la Sabana de Bogotá.» Agronomía Colombiana, Vol. XIII, No 2, pp. 152-161.

Ato, M., López, J. (1995). «IV Simposio de Metodología de las Ciencias del Comportamiento.». Universidad de Murcia.

Bernal, N., Darghan, AE., Rodríguez, LE. (2017). «Modelado del Calibre de tubérculos de papa Solanum phureja bajo diferentes densidades de siembra mediante regresión binomial negativa cero-inflada».

Bonaccorso, G. (2017). «Machine Learning Algorithms.». Packt Publishing Ltd.

Buitrago, G., López, A., Coronado, A., Osorno, F. (2004). «Determinación de las características físicas y propiedades mecánicas de papa cultivada en Colombia». Revista Brasileira de Engenharia Agrícola e Ambiental, Vol.8, No.1, pp.102-110.

Challenger, I., Díaz, Y., Becerra, R. (2014). «El lenguaje de programación Python». Ciencias Holguín, Vol. XX, No. 2, pp. 1-13.

Gönen, M. (2007). «Analyzing Receiver Operating Characteristic Curves with SAS.» SAS Institute Inc.

Hayter, A. (2012). «Probability and Statistics for Engineers and Scientists.» Estados Unidos: Brooks/Cole, Cengage Learning.

IBM. (2012). «Manual CRISP-DM de IBM SPSS Modeler». IBM Corporation.

Koduvely, H. (2015). «Learning Bayesian Models with R». Reino Unido: Packt Publishing Ltd.

Layton, R. (2015). «Learning Data Mining with Python». Packt Publishing.

Misigo, R., Miriti, E. (2016). «Classification of Selected Apple Fruit Varieties using Naive Bayes». Indian Journal of Computer Science and Engineering (IJCSE), Vol. 7, No. 1.

Mohamad, N., Jusoh, N., Htike, Z., Lei Win, S. (2014). «Bacteria Identification from Microscopic Morphology Using Naïve Bayes.» International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 4, No.2.

Piñeros, C. (2009). «Recopilación de la Investigación del Sistema Productivo Papa Criolla». Secretaria de Agricultura y Desarrollo Economico (Gobernación de Cundinamarca), Federación Colombiana de Productos de Papa.

Rich, I. (2001). «An Empirical Study of the Naïve Bayes Classifier». IJCAI 2001 Work Empir Methods Artif Intell.

Rivera, J., Herrera, A., Rodríguez, L. (2011). «Evaluación de la aptitud de procesamiento en seis genotipos de papa criolla (Solanum tuberosum Grupo Phureja)». Agronomía Colombiana, Vol. 29(1), pp. 73-81.

Sampieri, R., Fernández, C., Baptista, M. (2014). «Metodología de la Investigación». McGRAW-HILL / INTERAMERICANA EDITORES, S.A. DE C.V.

Tsangaratos, P., Ilia, I. (2016). «Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size». Catena 145, pp. 164–179.

VanderPlas, J. (2017). «Python Data Science Handbook». O'Reilly.