Universidad Nacional Experimental del Táchira

Vicerectorado Académico

Decanato de Docencia

Departamento de Ingeniería en Informática

Trabajo de aplicación profesional

Proyecto especial de grado

Paquete en lenguaje R para la clasificación de tubérculos de papa criolla para diferentes densidades de siembra empleando redes neuronales probabilísticas.

Autor(es): Jesús David Escalante Rodríguez

C.I.: 21.220.841

jesusd.escalante@unet.edu.ve

Tutor(es): Rossana Timaure

rttg@unet.edu.ve

San Cristóbal, Julio.2018

Universidad Nacional Experimental del Táchira
Vicerrectorado Académico
Decanato de Docencia
Departamento de Ingeniería Informática
Trabajo de Aplicación Profesional
Proyecto Especial de Grado

Aprobación del Tutor para presentación de la Propuesta del Proyecto Especial de Grado

Yo, Rossana Timaure titular de la cédula de identidad No. V-12.021.405. en mi carácter de Tutor(a) del Proyecto Especial de Grado titulado: , presentado por el bachiller: Jesús David Escalante Rodríguez, titular de la cédula de identidad No. 21.220.841, por medio de la presente autorizo la presentación de la Propuesta de Proyecto Especial de Grado ante los jurados designados por la comisión de Trabajo de Aplicación Profesional del Departamento de Ingeniería en Informática, en virtud de considerar que reúne los requisitos establecidos en el artículo 16 de las Normas para el Trabajo de Aplicación Profesional de la UNET.

Nombres y apellidos del tutor
C.I. V- 12.021.405
Rossana Timaure

Introducción

En los cultivos de papa criolla son muchas las variables que influyen en la cosecha de tuberculos, tales como la temperatura, altura del terreno, clima, tipo y composicion del suelo, densidad de siembra, entre otros y los encargados y personas interesadas en este proceso siempre han buscado manipularlas para no solo mejorar la calidad del producto, si no para obtenerlo con ciertas caracteristicas que mas convengan para su fin como la piel, contextura, pero muy sobre todo el tamano de los tuberculos ya que existen intereses industriales diferentes en esta ultima caracteristica.

Investigaciones anteriores como al realizada por Bernal (2017), concluyeron que la densidad de siembra es una variable con una afectacion considerable en el tamano de los tuberculos producidos por las plantas en su ciclo de cosecha, sin embargo la manipulacion de esta variable en dicha investigacion, fue de tipo lineal sin tener en cuenta, ciertas condiciones espaciales que posee como propiedades esta variable, por la misma ser distancias entre calles y surcos, y distancia entre plantas a la que se siembra o expresado de otra manera cuantas plantas son sembradas por metro cuadrado en un terreno.

La regresion espacial es una solucion que se plantea usar en el siguiente documento para el modelado de los datos recogidos y observados de la cosecha de un cultivo de tuberculos de papa criolla realizado en el Centro Agropecuario Marengo de la Universidad Nacional de Colombia, en el departamento de Cundinamarca en Colombia (74°12'58.51"W;4°40'52.92"N), el cual tiene una altitud de 2516 msnm y una temperatura media de 14 °C.

La densidad de siembra es una variable cualitativa que aunque se trata de distancias se definen en densidades predeterminadas como d1 = 50x30cm, donde d1 es una densidad definida igual a un espacio entre surcos de 50cm y distancia entre plantas de 30cm. Haciendo que el analisis de regresion espacial deba hacerse por separado para los datos de cada densidad tomados, lo que nos presenta ante la dificultad de no poder asegurar una relacion entre los resultados porque aunque

los datos tengan el mismo comportamiento al ser modelados podrian no tener una normalizacion bivariante.

Las redes neuronales clasifican basado en sus neuronas entrenadas a traves de datos conocidos con anterioridad por lo que no es de importancia el hecho de que no exista una relacion entre estas variables ya que no requiere de supuestos estadisticos a tener en cuenta para su proceso de clasificacion.

El objetivo de esta investigacion es construir un paquete en Lenguaje R que permita a traves de una entrada, propuesta como una matriz cargada con los datos observados de cultivo, clasificar los tuberculos de papa criolla cosechados en densidades de siembra segun el peso total del cultivo, peso total de tuberculos y diametro de cada tuberculo a traves de una red neuronal probabilistica entrenada con datos del cultivo anteriormente recogidos, ademas de comparar los resultados provenientes de una regresion espacial versus datos observados, para concluir cuan efectivo puede ser aplicar una regresion espacial para estos casos.

Capítulo 1

Preliminares

1.1. Planteamiento y formulación del problema

La produccion de papa criolla (*Solanum Phureja*), colombiana o yema de huevo como tambien es conocida, siempre ha tenido muchos desafios por parte de la industria, debido a los grandes intereses provenientes de los diferentes consumidores que la misma posee, siendo asi que es uno de los productos agricolas mas consumidos y con mayor importancia en el mundo, despues del arroz, maiz y trigo. Colombia es el mayor productor de papa criolla en Latinoamerica y por lo cual es un pais que posee grandes exigencias industriales de la misma (Ligarreto *et al*, 2003).

Los tuberculos de papa criolla tienen aproximadamente entre dos y ocho centimetros (2-8cm), y se pueden clasificar en tres calibres segun su diametro promedio de donde radican los intereses comerciales de la misma. Los tuberculos de entre dos y medio y cuatro centimetros son preferidos para encurtidos y pre-cocidos mientras los tuberculos promedio entre cuatro y seis y medio centimetros son preferidos para frituras en hojuelas o con mas de cinco centimetros para frituras en tiras (CORPOICA, 2009).

El crecimiento vegetal es definido por Cabezas en 2005 como «El aumento irreversible del tamaño y peso seco de las plantas (altura, área foliar, diámetro, número de células y cantidad de protoplasma) o los cambios que ocurren en una planta o población de plantas a través del tiempo, fenómeno acompañado del aumento en la complejidad estructural metabólica del organismo (diferenciación celular, número de hojas), por procesos de división y alargamiento celular, incorporación de materia y energía del ambiente (fotosíntesis, absorción de agua y de iones) y metabolización subsiguiente, la cual se traduce en multiplicación y diferenciación celular. Este proceso está ínti-

mamente relacionado con algunos factores internos como fotosíntesis, respiración, transpiración, condiciones de estrés, concentración enzimática, balance hormonal y expresión genética» (Pineros, 2009).

En la papa son dos los procesos fisiologicos asocidos directamente al rendimiento de la misma, la fotosintesis y la respiracion. Y estos dos son procesos intimamente asociados, ya que durante la fotosintesis se producen carbohidratos que son consumidos durante la respiracion, pero una gran cantidad de estos carbohidratos contribuyen al proceso de produccion incrementando el tamano de los tuberculos, el nivel y periodo de creciemiento de los tuberculos es una variable que responde directamente a la expresion de rendimiento expresada como produccion por dia, siendo asi que cerca de un noventa por ciento (90 %) del peso acumulado de los tuberculos producidos por una planta es producto de la asimilacion de dioxido de carbono, asi como otros procesos similares (Beukema, 1979). Sin embargo, estos procesos se ven afectados por condiciones externas tales como la itensidad de luz, temperatura, longitud de los dias, condiciones del suelo que son variables no controladas, o por condiciones como la cantidad de riego, fertilizantes aplicados y productos de abono para la mejora de los suelos que son variables controlables pero que requieren de costos adicionales de produccion (Pineros, 2009). Y es por ello que se sigue buscando con el manejo de variables controladas el control sobre la cantidad total producida y tamano de los tuberculos sin producir costos adicionales, siendo la densidad de siembra una de estas variables, en la cual nos enfocaremos en esta investigacion (Bernal, 2017).

Sin embargo la densidad de siembra es una variable de tipo espacial, ya que se define como la distancia entre plantas y entre linderas, siendo asi una coordenada en X y una coordenada en Y en un plano cartesiano, por lo cual un estudio estadistico de la misma para clasificar o predecir un modelo no puede ser del tipo de regresion lineal o de un analisis de varianza debido a los residuales que deben ser independientes en cada punto, no tomando en cuenta una relacion entre vecinos como si ocurre en los cultivos de papa, las plantas compiten por los recursos y por consiguiente por producir mas, haciendo asi los residuales dependientes y asi mismo no optima la aplicacion de dichos analisis (Pineros, 2009).

Se plantea el uso de regresion espacial entonces, pero el manejo de la densidad de siembra como variable es del tipo cualitativo con un valor de distancia entre linderos y plantas fijo para una cantidad de densidades definidas, implicando tener que aplicar regresion espacial para cada uno de los valores de estas densidades. Sin embargo aunque estos analisis me arrojen resultados con un comportamiento y una distribucion estadistica similar, no se puede asegurar una relacion

entre cada una de las densidades definidas basados en una normalizacion n-variante, donde n seria el numero de densidades definidas. Por ello se propone en esta investigacion el uso de una red neuronal probabilistica (PNN) donde no importan los supuestos y definiciones de un modelo lineal.

Una red neuronal artificial (ANN) es un modelo que crea una relacion entre una configuracion de senales de entrada y una senal de salida usando un modelo derivado de nuestro entendimiento de como el cerebro responde a determinados estimulos de entradas sensoriales. Asi como un cerebro usa ese conjunto de celulas interconectadas llamadas neuronas, asi las ANN usan una red de neuronal artificiales o nodos para solucionar problemas aprendidos (Lantz, 2015).

En general, las ANN son aprendices versatiles que pueden ser aplicarse a cais cualquier tarea de aprendizaje en cualquier contexto, clasificacion, prediccion, reconocimiento de patrones e incluso pudieramos hablar de tareas de supervision de procesos. Las ANN, se aplican mejor a los problemas cuyos datos de entrada y salida son bien definidos o bastante simples, sin embargo el proceso que relaciona la entrada con una determinada salida es en extremo complejo, por eso se conocen como un metodo de caja negra (Lantz, 2015).

De una manera muy arcaica, las ANN han sido usadas desde hace cincuenta anos para simular como el cerebro humano es capaz de resolver problemas. Primero surgieron las simples funciones logicas AND y OR, que ayudaron a los cientificos a entender como biologicamente el cerebro opera. Sin embargo, como el avance tecnologico y computacional ha sido incrementado y continua creciendo potencialmente durante los ultimos anos, las ANN ha incrementado su complejidad siendo usadas frecuentemente en multiples problemas como:

1. Programas de reconocimiento de voz y escritura, como los que usan los correo de voz de los servicios de transcripción y máquinas clasificadoras de correo postal.
2. La automatización de dispositivos inteligentes como los controles ambientales de un edificio de oficinas o automóviles autodirigidos y drones autoguiados.
3. Modelos sofisticados de patrones climáticos y otros fenómenos científicos, sociales o económicos.
4. La clasificacion y reconocimiento de cualquier cosa, basado en sus atributos y caracteristicas.

Existen muchos metodos y lenguajes que permiten la estructuracion y desarrollo de ANN, destacandose lenguajes como Python y softwares como Statgraphics y SPSS, sin embargo como el

objetivo de esta investigacion es trabajar con una variable de tipo espacial que amerita la aplicacion de regresion espacial antes para la normalizacion de los datos recogidos y observados, se propone el uso del lenguaje R con el paquete neuralnet para el manejo y construcciond de la red neuronal.

R es un lenguaje de programación y entornos para graficos y calculos estadisticos, es una implementacion diferente al lenguaje S (Chambers, 1998), y se desarrolla bajo un proyecto de software bajo Licencia Publica General (General Public License, GNU) de la Free Software Foundation (https://www.fsf.org/) en forma de codigo fuente. Se compila y se ejecuta en una amplia variedad de plataformas entre ellas FreeBSD, Linux, Windows y MacOS. (The R Project for Statistical Computing, 2018).

El paquete neuralnet es un paquete del lenguaje R de muy sencilla utilizacion que permite la aplicacion, visualizacion e implementacion de redes neuronales. El paquete permite configuraciones flexibles a traves de eleccion personalizada del error y función de activacion. Ademas, el calculo de pesos generalizados esta implementado (CRAN-R Project, 2016).

Esta propuesta permitira ajustar y entrenar una red neuronal artificial probabilistica que permita la clasificacion de densidades de siembra de papa yema de huevo (Solanum Phureja) teniendo como entrada las siguientes variables:

1. **Peso total del cultivo.** Variable cuantitativa expresada en gramos (gr), que indica el peso de todos los tuberculos recogidos en la siembra.

2. **Total de tuberculos recogidos.** Variable cuantitativa que indica el numero de tuberculos recogidos en la siembra.

3. **Diametro ponderado de todos los tuberculos.** Variable cuantitativa expresada en centimetros (cm), que indica el calculo del diametro ponderado (por los tuberculos ser ovoides) de cada tuberculo recogido en la siembra.

La densidad de siembra es la distancia entre plantas y leras o sencillamente cuantas plantas son sembradas por metro cuadrado de siembra.

El objetivo es comparar los resultados de clasificacion alimentando la red con datos observados y con los mismos datos normalizados espacialmente a traves de regresion espacial, para observar si es necesario el analisis de los datos recogidos antes de la clasificacion y a su vez concluir hasta que

punto la densidad de siembra afecta las variables antes mencionadas. Otra de las pruebas que se plantea usar en esta investigacion es realizar los analisis eliminando variables de entrada atisbando cuales variables permiten y generan una mejor clasificacion.

La curva ROC es una herramienta estadistica utilizada en analisis de clasificacion para determinar la capacidad discriminante de una prueba diagnostica dicotomica. Esta capacidad discriminante esta sujeta al valor umbral elegido de entre todos los posibles resultados de la variable de decision, es decir, la variable por cuyo resultado se clasifica (en nuestro caso la densidad de siembra) en un determinado grupo. La curva es el grafico resultante de representar, para cada valor umbral, las medidas de sensibilidad y especificidad de la prueba diagnostica (Benavides, SF).

Posterior a las diversas clasificaciones planteadas, se propone la construccion en lenguaje R de las curvas ROC correspondientes para determinar cuan optima es la clasificacion resultante, permitiendo asi llegar a las conclusiones deseadas.

1.2. Objetivos

1.2.1. Objetivo General

Desarrollar un paquete en lenguaje R que permita la clasificacion de tuberculos de papa criolla para diferentes densidades de siembra empleando redes neuronales probabilisticas.

1.2.2. Objetivos Específicos

- Disenar los algoritmos que permitan realizar el estudio de regresion espacial a los datos de entrada de cultivo de papa criolla.
- Disenar los algoritmos que permitan el entrenamiento de una red neuronal probabilistica que permita la clasificacion de tuberculos de papa criolla para diferentes densidades de siembra.
- Implementar y desarrollar los metodos para la clasificacion de tuberculos de papa criolla para diferentes densidades de siembra.
- Realizar la pruebas del paquete desarrollado bajo diferente escenarios que permitan concluir que tan buena es la clasificacion realizada por el mismo a taves de curvas ROC.

1.3. Justificación e Importancia

Colombia es el mayor productor, consumidor y exportador de papas diploides en el mundo; tiene una ventaja competitiva notable debido a ser centro de diversidad y poseer gran aceptacion por los consumidores debido a las caracteristicas del tuberculo. Adicionalmente, en este pais se ha desarrollado una amplia tradicion como cultivo tecnificado, con potencial de industrializacion y exportacion (Rodriguez *et al*, 2009). Ademas de contar con uno de los recursos geneticos que ofrece las mayores oportunidades de exportacion como alimento procesado exclusivo y sin competencia. Desafortunadamente, muchas de las estrategias de manejo del cultivo de la papa, han sido adaptadas a papa criolla, creando un sistema productivo poco eficiente (Piñeros, 2009). Los altos precios de los insumos agricolas y el mal manejo de los cultivos agronomicamente provoca baja productividad y amenaza la competitividad del sistema de produccion, por lo que es importante identificar factores limitantes del rendimiento y desarrollar practicas innovadoras para el cultivo, tales como una gestion nutricional integrada y equilibrada, que es una de las practicas mas eficientes para garantizar a la planta la oportunidad de expresar su potencial genetico que eventualmente se reflejara en una mejor calidad y rendimiento (Lopez *et al*, 2014).

Varias investigaciones revelan que el tamano promedio de los tuberculos, su variabilidad y su conteo, definen una unica distribucion, presentando mayor variabilidad de rangos de tamano en la etapa final de llenado, con aumento del rendimiento de los mayores tamanos por cuenta de los tamanos inferiores, lo que sugiere que un tuberculo que pasa de un calibre inferior al siguiente, no es sustituido por otro de calibre anterior, ademas, parece existir una correlacion negativa entre la variabilidad relativa del tamano del tuberculo y el numero de tuberculos por unidad de area (Struik, 1991). Es importante reconocer, que debe tenerse cuidado con la interpretacion de esta correlacion, pues hasta cierto punto, la naturaleza de los datos asociados al numero de tuberculos por planta pudiera ser similar a la de los datos composicionales, es decir, si un componente aumenta, el otro esta forzado a disminuir, lo que genera correlaciones sin conexion logica en datos verdaderamente composicionales, sin embargo, el numero de tuberculos por calibre no suma una constante, pudiera ocurrir que su maximo posea una distribucion especifica, entonces una mayor cantidad de tuberculos de un calibre podria provocar un numero menor de tuberculos en otro calibre especifico, lo que en esencia puede generar correlaciones en las que debe tenerse mucho cuidado al momento de su interpretacion. La importancia del cultivo obliga a muchos investigadores a realizar diferentes estudios para mejorarlo segun el uso que vaya a darse a los tuberculos, generando nuevas variedades resistentes a plagas y enfermedades y de facil adaptacion a diferentes climas, procurando hacer uso del espacio de siembra de forma optima, por lo que varios estudios evidencian el estudio de

la densidad de siembra para evaluar sobre todo el rendimiento. En 2017, Bernal evaluo en lugar del rendimiento, un indicador que se asocia directamente como lo es el calibre de los tuberculos, los cuales en la practica se manipulan en cuatro categorias de diametro promedio (hasta 2 cm, de 2 a 4 cm, de 4 a 6 cm y más de 6 cm). La naturaleza de esta variable imposibilita usualmente la comparacion de la respuesta (conteos de tubérculos) para cada densidad de siembra (definidas como 30cm*100cm, 40cm*100cm y 50cm*100cm) mediante analisis de varianza, pues en el caso de conteos, otras distribuciones como la Poisson y la Binomial negativa se adaptan mucho mejor al tipo de dato generado. Los modelos clasicos de regresion Poisson y binomial y negativa, en sus modalidades usuales o en la opción inflada por ceros (Cameron, 1998) en datos de conteo pertenecen a la familia de modelos lineales generalizados (Zeileis, 2008) y los desarrollos recientes permiten generar varias estadisticas que permiten su comparacion con sus contrapartes sin ceros en exceso, lo que resulta util en la eleccion del mejor modelo para relacionar predictores y respuesta en conteos.

Sin embargo, los estudios realizados hasta ahora son todos haciendo suposiciones del tipo lineal y haciendo modificaciones que permitan asi su estudio a traves de anova o regresion lineal tomando en cuenta los patrones de vecindad, mas la propuesta consiste en clasificar los datos a traves de una red neuronal que no amerita de suposiciones estadisticas o matematicas, que solo aprendiendo de lo observado es capaz de realizar una clasificacion de densidades de siembra, anadiendo aun asi para conseguir un mejor reusultado un analisis de regresion espacial previo, con el objetivo de solo con datos de siembra previos conocer a que densidad de siembra se debe cosechar para obtener un deseado calibre para fines comerciales e industriales.

El objetivo de esta investigacion es permitir una futura sustitucion de los metodos convencionales de clasificacion de los tuberculos como el tamizado y la clasificacion manual a ojo, optimizando la velocidad de cosecha y los resultados que se esperaban al momento de sembrar.

1.4. Alcance y Limitaciones

La red neuronal artificial probabilistica a implementar tiene como proposito la clasificacion de la densidad de siembra en papa criolla o yema de huevo, siendo util para incrementar las posibilidades de obtener un determinado y deseado calibre de tuberculos conociendo la densidad a la que se debe sembrar, anadiendo tambien para su optimizacion analisis de regresion espacial y un conjunto de pruebas que van desde la variacion de las variables de entrada hasta la utilizacion de los datos observados, para definir que tan buena o no es la clasificacion. Sin embargo las pruebas

se relizaran bajo solo un conjunto de datos reales que fueron tomados bajo las medidas de cuatro calibres de tuberculos y tres densidades de siembra.

Los datos disponibles del cultivo de papa se tienen en base a tres densidades de siembra, cantidad de plantas sembradas, cantidad de tuberculos recogidos, peso total del cultivo, diametro ponderado de cada tuberculo y calibre de los tuberculos y forman parte de la investigacion de Maestria del estudiante de la Universidad Nacional de Colombia, Nelson Bernal.

Los resultados de las pruebas sobre los escenarios planteados seran validados estadisticamente con ayuda del analisis de las curvas ROC generados por estos mismos resultados.

El trabajar como variable de estudio la densidad de siembra de los cultivos de tuberculos de papa criolla hace dificil el hecho de modelar los datos ya que la misma es una variable cualitativa que es definida en densidades predeterminadas, haciendo que el analisis de regresion espacial deba hacerse por separado para los datos de cada densidad tomados, lo que nos presenta ante la dificultad de no poder asegurar una relacion entre los resultados porque aunque los datos tengan el mismo comportamiento al ser modelados podrian no tener una normalizacion bivariante.

Capítulo 2

Fundamentos teóricos

2.1. Antecedentes

La clasificacion por medio de redes neuronales ha sido un hito ya marcado en el campo agronomico, muchos estudios se han realizado con el objetivo de analizar ciertos comportamientos de plantas y los beneficios que se puedan sacar de ellas. En 2011, el grupo de investigacion de sistemas de procesamiento y control de senales de la Universidad Nacional Tenaga de Malasia desarrollo un sistema de inteligencia con un enfoque novedose para la clasificacion de frutas usando tecnicas de procesamiento de imagenes digitales y redes neuronales artificiales, con el objetivo de desarrollar un metodo de clasificacion rapido con una meta del 100 % de eficiencia. El estudio se realizo con cinco frutas, manzanas, platanos, zanahorias, mangos y naranjas, extrayendo de ellas siete caracteristicas en funcion de la forma y el color. La captura de las imagenes se realizo con una camara digital convencional y las manipulaciones a los datos y construccion de la red con el software MATLAB. Los resultados obtenidos durante esta investigacion fueron de gran avance en el campo de reconocimiento de patrones en imagenes.

En 2011, Mayabiro E presento en la Universidad Nacional Experimental del Tachira, un prototipo sobre el entorno MATLAB para el calculo de la tasa de germinacion de plantulas de pimenton previamente segmentadas. El entorno desarrollado permitia establecer una clasificacion de las plantulas, hojas u objetos de la misma por medio de redes neuronales. Se realizo el entrenamiento de multiples redes neuronales multicapas con algoritmos de retropropagacion, donde aunque variaban las capas intermedias de las redes y sus funciones de transferencia fueron entrenadas con los mismos datos de entrada, validandolas con una base de datos de pruebas para seleccionar al final una con salidas similares a las deseadas.

Otro estudio realizado en 2013 por Stephen Gang Wu consistia en el estudio teorico de tecnicas de procesamiento de imagenes y datos para el reconocimiento automatico de hojas para la clasificacion de plantas. Doce caracteristicas de las plantas fueron extraidos y distribuidas en cinco variables principales que constituian el vector de entrada de una red neuronal artificial probalistica, que habis sido entrenada con 1800 hojas para clasificar 32 tipos de plantas con una precision superior al 90 %, el autor aseguraba que su metodologia de implementacion de la PNN era facil y rapida en comparacion de otras investigaciones similares.

En 2017, Bernal N realizo un estudio de campo con el cultivo de papa criolla para evaluar la influencia de la densidad de siembra asociada a distancias entre plantas de 30,40 y 50 cm y distancias entre surcos de 100 cm sobre el conteo de tuberculos de calibres inferiores a 2 cm, entre 2 y 4 cm, entre 4 y 6 cm, y de mas de 6 cm de diametro ponderado y sobre el peso fresco en gramos de los tuberculos. Los tuberculos cosechados se clasificaron y contaron mediante tamizado y se pesaron en su totalidad sin discriminar por calibre. Los modelos estadisticos empleados para modelar el comportamiento de la cosecha, evidenciaron el efecto significativo de la densidad de siembra sobre el conteo de tuberculos y calibre y se observo una razon aproximada de 40:40:20:1 desde el calibre menor al mayor. El efecto de la competicion, en todos los modelos probados resulto significativo, aumentando en la mayoria de los casos a medida que disminuia la distancia entre plantas, tanto en el patron de vecindad intrahileras como en el caso de inter e intrahileras.

2.2. Bases Teóricas

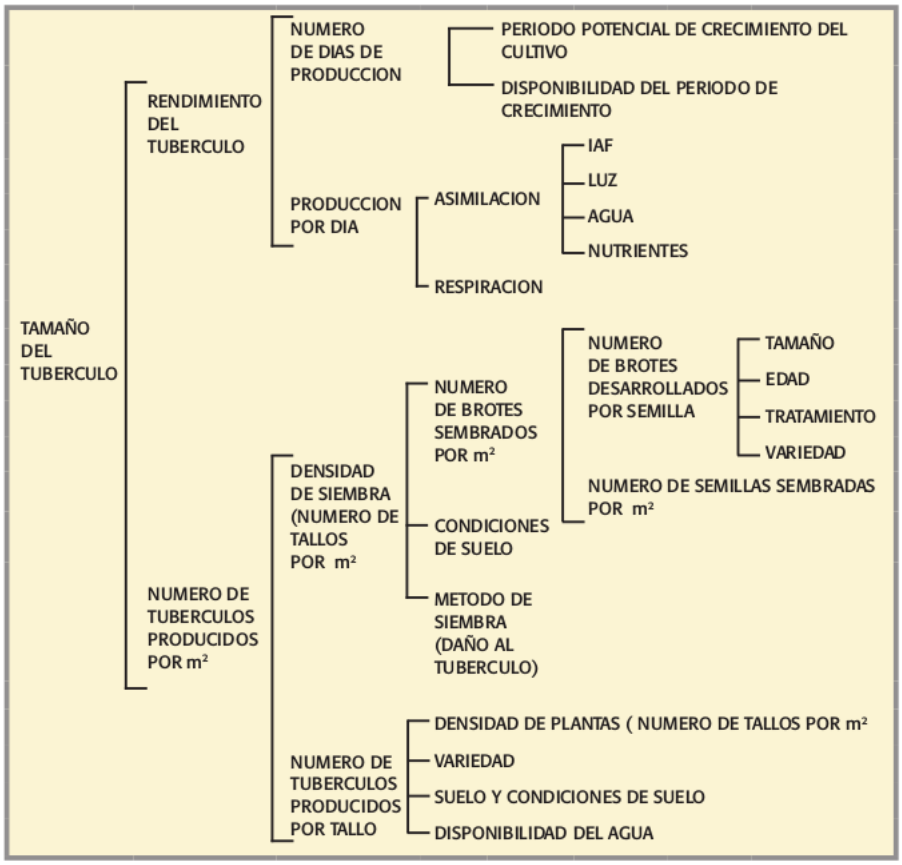
2.2.1. Papa criolla o yema de huevo (*Solanum Phureja*).

Las variables que influyen en el rendimiento de la papa pueden ser observados en la figura 2.1. Pag.13.

2.2.2. Regresión Espacial.

El análisis exploratorio de datos espaciales - AEDE, constituye una disciplina reciente que ha adquirido una especial importancia debido principalmente al avance de la tecnología en las comunicaciones y la globalización de la economía. Los sucesos que ocurren en una ubicación específica tienen repercusiones sobre sus vecinos directos e incluso sobre otros, aparentemente remotos.

Figura 2.1: Variables de influencia sobre el rendimiento de Solanum Phureja.



En el estudio de cualquier fenómeno de carácter social o económico la ubicación geográfica de los agentes constituye un aspecto importante dentro de la especificación de los modelos econométricos, ya que puede existir algún efecto espacial, que de no ser incorporado en la especificación, podría afectar la validez del modelo. Ante esta realidad y gracias al desarrollo tecnológico de los sistemas de georreferenciación de datos, surge la necesidad de contar con herramientas apropiadas para el procesamiento, descripción y análisis de la información ya que los métodos tradicionales de la estadística descriptiva no tienen en cuenta la localización geográfica de los datos.

Teniendo en cuenta que la econometría tradicional no ha incorporado el efecto de dichas circunstancias y que la estadística espacial se ocupa de otro tipo de problemas, ha surgido una disciplina a la cual se le ha dado el nombre de econometría espacial. Según Luc Anselin, uno de sus principales investigadores, "las actividades como la estimación de modelos espaciales de interacción, el análisis estadístico de la función de densidad urbana y la implementación empírica de

modelos econométricos regionales, podrían ser considerados econometría espacial" (Anselin, 1988).

Cuando se tienen observaciones georreferenciadas, se deben utilizar herramientas que permitan detectar ciertas características dentro de los datos, como son tendencia, valores atípicos, esquemas de asociación y dependencia espacial, concentración espacial o puntos calientes/fríos, entre otros. Aunque en la actualidad se tiene gran cantidad de información georreferenciada, estos datos suelen ser tratados con herramientas del análisis de series temporales (o de corte transversal, no espacial), sin usar técnicas adecuadas para el análisis estadístico espacial. Los métodos que permiten extraer dichas características de los datos georreferenciados se conocen con el nombre de análisis exploratorio de datos espaciales (AEDE) y se conciben como una disciplina dentro del análisis estadístico más general, diseñada para el tratamiento específico de los datos geográficos. El AEDE se utiliza para identificar relaciones sistemáticas entre variables, o dentro de una misma variable, cuando no existe un conocimiento claro sobre su distribución en el espacio geográfico (Chasco Yrigoyen, 2006).

2.2.3. Redes Neuronales Artificiales.

Una red neuronal artificial (ANN) es un conjunto de nodos de un programa (neuronas) interconectados entre si, simulando el proceso de pensamiento humano, se pudiera considerar como una caja negra entrenada previamente para esperar una entrada y basado en las caracteristicas o comportamiento de la misma proporcionar una determinada salida, eliminando asi la necesidad de diferentes algoritmos que deban analizar comportamientos cada uno por separado. Una red neuronal probabilistica (PNN) no es mas que una ANN que usa funciones estadisticas que escalan la variable no linealmente como una forma de campana o una distribucion normal (Stephen, 2007).

Entre las definiciones más recientes de inteligencia artificial se expresa, en forma general, la inteligencia artificial como la capacidad que tienen las máquinas para realizar tareas que en el momento son realizadas por seres humanos; otros autores como Nebendah (1988) y Delgado (1998) dan definiciones más completas y las definen como el campo de estudio que se enfoca en la explicación y emulación de la conducta inteligente en función de procesos computacionales basados en la experiencia y el conocimiento continuo del ambiente. Autores como Marr (1977), Mompin (1987), Rolston (1992), en sus definiciones involucran los términos de soluciones a problemas muy complejos.

El nacimiento de la inteligencia artificial se sitúa en los años cincuenta; en esa fecha la informáti-

ca apenas se había desarrollado, y ya se planteaba la posibilidad de diseñar máquinas inteligentes. Hoy en día se habla de vida artificial, algoritmos genéticos, computación molecular o redes neuronales. En algunas de estas ramas los resultados teóricos van muy por encima de las realizaciones prácticas.

A través de los años, se han utilizado diversas técnicas de inteligencia artificial para emular 'comportamientos inteligentes'. Al software que hace uso de dichas técnicas se le denomina, de forma genérica, 'sistema inteligente', y es cada vez más amplia la gama de aplicaciones financieras donde incide la inteligencia artificial.

Un ejemplo de esto es que al usarse una tarjeta de crédito, suelen acumularse datos sobre patrones de consumo que después se venderán a diversas empresas. Sobre la base de los pagos efectuados en dicha tarjeta de crédito, los bancos e instituciones de crédito irán elaborando un historial del usuario, el cual se utilizará para autorizar una transacción, para decidir cuándo extender el crédito y para detectar fraudes. Este tipo de procesos requiere de chequeos que suelen resultar bastante complejos, además del uso de criterios variables para poder tomar una decisión final en torno a la autorización de una cierta transacción. Claro que, al manejar enormes volúmenes de información, como los aproximadamente 16 millones de transacciones que Visa Internacional debe verificar diariamente, no resulta nada fácil poder detectar un fraude. Aunque es evidente la necesidad de automatizar procesos como éste, no es del todo obvio incorporar el comportamiento inteligente del ser humano a un programa de computadora que reemplace a un evaluador humano, ya que los sistemas de inteligencia artificial se toman como herramientas de apoyo analítico para el evaluador, mas no como una unidad autosuficiente que por sí sola pueda tomar decisiones.

Las redes neuronales artificiales son eficientes en tareas tales como el reconocimiento de patrones, problemas de optimización o clasificación, y se pueden integrar en un sistema de ayuda a la toma de decisiones, pero no son una panacea capaz de resolver todos los problemas: todo lo contrario, son modelos muy especializados que pueden aplicarse en dominios muy concretos.

Las redes neuronales emulan la estructura y el comportamiento del cerebro, utilizando los procesos de aprendizaje para buscar una solución a diferentes problemas; son un conjunto de algoritmos matemáticos que encuentran las relaciones no lineales entre conjuntos de datos; suelen ser utilizadas como herramientas para la predicción de tendencias y como clasificadoras de conjuntos de datos. Se denominan neuronales porque están basadas en el funcionamiento de una neurona biológica cuando procesa información.

2.2.4. Curvas ROC.

Una amplia gama de tests diagnósticos reportan sus resultados cuantitativamente, utilizando escalas contínuas. El análisis de curvas ROC (receiver operating characteristic curve) constituye un método estadístico para determinar la exactitud diagnóstica de estos tests, siendo utilizadas con tres propósitos específicos: determinar el punto de corte de una escala continua en el que se alcanza la sensibilidad y especificidad más alta, evaluar la capacidad discriminativa del test diagnóstico y comparar la capacidad discriminativa de dos o más tests diagnósticos que expresan sus resultados como escalas continuas.

2.2.5. Lenguaje R.

R es un conjunto integrado de *software* de código abierto para el almacenamiento, manipulación, cálculo y visualización de datos para computación y gráficación estadística, puede ser compilado y ejecutado en en Windows, Mac OS X y otras plataformas UNIX (como Linux), se distribuye usualmente en formato binario (`https://www.r-project.org/about.html`, 2018). El proyecto de *software* R fue iniciado por Robert Gentleman y Ross Ihaka. El lenguaje fue influenciado por lenguaje S desarrollado originalmente en Bell Laboratories por John Chambers y sus colegas. Desde entonces ha evolucionado para el cálculo estadístico asociado a diversas disciplinas para contextos académicos y comerciales. En R, la unidad fundamental de código compartible es el paquete, el cual agrupa código, datos, documentación y pruebas, y resulta simple de compartir con otros. Para enero del 2015 ya habían más de 6.000 paquetes disponibles en la Red Integral de Archivos de R, conocido comunmente por su acrónimo CRAN, el cual es el repositorio de paquetes . Esta gran variedad de paquetes es una de las razones por las cuales R es tan exitoso, pues es probable que algún investigador o académico ya haya resuelto un problema en su propio campo usando esta herramienta, por lo que otros usuarios simplemente podrán recurrir a ella para su uso directo o para llamarla en un nuevo código (Wickham,2015).

2.2.6. Estructura de paquetes en R/RStudio.

Requerimiento del núcleo (*core*)

1. DESCRIPTION: metadatos del package .
   La tarea del archivo Description es de gran importancia ya que es en el donde se registra la metadata, las dependencias que utiliza el paquete, la licencia y el soporte en caso de ocurrir errores con el mismo La estructura mínima para realizar un paquete en R es la siguiente:

- Package: mypackage
- Title: What The Package Does (one line, title case required)
- Version: 0.1
- Authors@R: person(”First”, ”Last”, email = ”first.last@example.com”,
- role = c(.ªut”, çre”))
- Description: What the package does (one paragraph)
- Depends: R (¿= 3.1.0)
- License: What license is it under?
- LazyData: true

2. `R/:` dirección del repositorio donde se encuentra el código del paquete (.R files).
   Se expondrán las buenas prácticas a la hora de realizar todo nuestro código en R, desde organización de las funciones, estilos de código y nombre de variables

   Organizar funciones en R: aunque puedes organizar los archivos como desees, los dos extremos son malos no colocar todas las funciones en el mismo. archivo y no crear un archivo para para función, aunque si una función es muy grande o tiene mucha documentación se puede dar el caso, los nombres de los archivos tienen que ser significado y deben de terminar en R

   - Bien
     - fit_models.R
     - utility_functions.R
   - Mal
     - foo.r
     - stuff.r

   Se puede recomendar de acuerdo al número de función utilizar prefijo

   Nombres de Objetos: Los nombres de las Variables y funciones deben de ser en minúsculas, usar el guión bajo (_) para separar palabras

   - Bien
     - day_one

    - day_1
- Mal
    - first_day_of_the_month
    - DayOne
    - dayone
    - djm1

En lo posible no usar nombres de variables existentes esto causará confusión.

Espaciado: Se recomienda colocar espacios alrededor de todos los operadores lógicos y aritméticos (=, +, -, <-, etc.). siempre coloque un espacio después de una coma, y nunca antes de ella.

- Bien
    - average <- mean(feet / 12 + inches, na.rm = TRUE)
- Mal
    - average<-mean(feet/12+inches,na.rm=TRUE)

Hay una pequeña excepción a esta regla: (:, :: y :::) no necesitan espacios alrededor de ellos.

- Bien
    - x <- 1:10
    - base::get
- Mal
    - x <- 1 : 10
    - base :: get

Dejar un espacio antes del paréntesis izquierdo, excepto en la llamada a una función

- Bien
    - if (debug) do(x)
    - plot(x,y)
- Mal
    - if(debug)do(x)

- plot(x, y)

Se Utiliza mas de un espacio es caso que de mejore a la alineación, por ejemplo:

```
list(
    total  = a + b + c,
    mean   = (a + b + c) / n
)
```

No coloque espacios alrededor del código entre paréntesis o corchetes (a menos que haya una coma)

- Bien
  - if (debug) do(x)
  - diamonds[5,]
- Mal
  - if (debug) do(x) *# no espacios alredor de debug*
  - x[1,] *# necesita un espacio despues de la coma*
  - x[1 ,] *# el espacio va despues de la coma no antes*

Llaves: Una llave de apertura nunca debe ir en su propia línea y siempre debe ir seguida de un nueva línea. Una llave siempre debe ir en su propia línea, a menos que sea seguida por otra y siempre sangría el código dentro de las llaves

- Bien
  - ```
    if (y < 0 && debug) {
      message("Y es negativo")
    }
    ```
  - ```
    if (y == 0 ) {
      log(x)
    } else {
      y ^ x
    }
    ```

- Mal
  - if (y < 0 && debug)
    message("Y es negativo")
  - if (y == 0) {
    log(x)
    }
    else {
    y ^ x
    }

Sentencias muy cortar esta bien dejarla en la misma línea.
if(y < 0 && debug) message("Y es negativo")

Longitud de Línea: cada línea debe de llevar máximo 80 carateres, si se queda sin espacio es recomendable utilzar una función separada

Sangria: Utilize sangria de 2 espacios, nunca use tablulador o multiples tabuladores o espacios. la unica expeción es cuando se define una sentencia en multiples líneas.

```
long_function_name <- function(    a = "a long argument",
                                   b = "another argument",
                                   b = "another argument",
```

Asignación: Usar el <-, y no =

- Bien
  - x <- 5
- Mal
  - x = 5

3. man/: documentación.

4. NAMESPACE: específica que objetos conforman el paquete.

Comentarios: Comente tu codigó, el comentario comienza #, los comentarios deben de explicar el porque, no el que.
use los caracteres (-) y (=) para separar lineas
Load data -
Plot data -

2.2.7. RStudio.

RStudio es un ambiente de desarrollo integrado (*Integrated Development Environment*, IDE) que ofrece herramientas de desarrollo vía consola, editor de sintaxis que apoya la ejecución de código, así como herramientas para el trazado, la depuración y la gestión del espacio de trabajo. RStudio está disponible para Windows, Mac y Linux o para navegadores conectados a RStudio Server o RStudio Server Pro (Debian / Ubuntu, RedHat / CentOS, y SUSE Linux) (https://www.rstudio.com/about/, 2018).

Capítulo 3

Fundamentos Metodológicos

A continuación se plantea la estructura a seguir por el presente trabajo, detallando el enfoque, tipo, nivel y diseño de la investigación y la metodología a implementar entre otros.

3.1. Enfoque de la investigación

La presente investigación se desarrollará siguiendo un enfoque cuantitativo, puesto que, como lo indican Pallela y Martins (2012) , "la investigación cuantitativa requiere el uso de instrumentos de medición y comparación, que proporcionan datos cuyo estudio necesita la aplicación de módelos matemáticos y estadísticos, el conocimiento está basado en hechos". Los datos a usar en el desarrollo de este trabajo investigativo fueron recolectados directamente de un cultivo y forman parte de la investigacion de Maestria del estudiante de la Universidad Nacional de Colombia, Nelson Bernal.

3.2. Tipo o nivel de investigación

Este proyecto plantea un tipo de investigación de campo, según como lo indican 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" ya que permite indagar los efectos de la interrelación entre los diferentes tipos de variable en lugar de los hechos.

En este punto se debe determinarla profundidad que abarca esta investigación, teniendo en cuenta que de acuerdo con el nivel de la investigación es definido como "grado de profundidad con que se aborda un fenómeno u objeto de estudio" (Arias, 2012).

En este sentido, se tiene que dadas las características del proyecto, se asocia con un nivel descriptivo, tal como lo indican Pallela y Martins (2012), "hace énfasis sobre conclusiones dominantes o sobre como una persona, grupo o cosa se conduce o funciona en el presente" esto debido a que se medirán los datos extraídos sin alterarlos para ser mostrados en el sistema.

Cuando se habla de un nivel descriptivo junto con una investigación de tipo de campo, en ella no se formulan hipótesis y las variables se enuncian en los objetivos de la investigación que se desarrollará.

3.3. Diseño de la investigación

Según Arias(2012), el diseño de la investigación es "la estrategia general que adopta el investigador para responder al problema planteado" (p.21) por lo que es vital establecer una correcta secuencia de pasos para elaborar el prototipo de software que dará solución a la problemática principal de la investigación.

Con este enfoque, se tiene que este trabajo seguirá un diseño no experimental, enfocado en el uso de información existente, de acuerdo con lo dicho por Pallela y Martins (2012) al definir el diseño no experimental como:

Es el que se realiza sin manipular en forma deliberada ninguna variable. El investigador no sustituye intencionalmente las variables independientes. Se observan los hechos tal y como se presentan en su contexto real y en un tiempo determinado o no, para luego analizarlos. Por lo tanto, este diseño no se construye una situación específica sino que se observan las que existen. Las variables independientes ya han ocurrido y no pueden ser manipuladas, lo que impide influir sobre ellas para modificarlas. (p.81)

Esto indica que no hay manipulación de variables. Esta investigación presenta una modalidad de proyecto especial que, como lo indican Pallela y Martins (2012), los proyectos especiales "destinados a la creación de productos que puedan solucionar deficiencias evidenciadas, se caracterizan por su valor innovador y aporte significativo" (p.92), ya que se creará un *software* aplicable al área de estudio.

3.4. Metodología

Los pasos a seguir en el desarrollo de esta investigacion seran descritos a continuacion siendo basados en los antecedentes y estudios realizados y las pautas estándar establecidas para la creación de paquetes y extensiones en R.

Creación del esqueleto del paquete.
En esta etapa se diseñaran y crearan los directorios, ficheros y objetos que conformaran el paquete.

Registrar el método para el envío y uso de funciones.

Es esta etapa del desarrollo se estableceran las dependencias sobre los paquetes de la base fuente de código R y sus métodos de conexión, considerando el manejo de versiones y los criterios de mantenimiento, además se estableceran los espacios de nombre o las estrategias para la búsqueda y utilización de las variables; unificando estos criterios a las funciones que seran diseñadas.

Diseño y codificación de las funciones.

Los métodos para el diseño de las funciones primarias en R serán los diagramas de flujo; y para su codificación se seguirán las normas de estilo para codificación en R, sugeridas por Wickham (2015) y por el creador del paquete *formatR* Xie(2017), además se establecerán la dependencia con las funciones de código base y las recomendadas para desarrollo en R.

Explorar y manipular la data.

Los datos tomados de la investigacion de Bernal, son datos reales observados de un cultivo de tuberculos de papa criolla que no han sido tratados estadisticamente o modelados, es por eso que en esta etapa se disenara un algoritmo para la carga de los mismos en estructuras de datos en R, para luego ser modelados y analizados espacialmente por medio de regresion espacial y probando de igual forma su normalizacion bivariante.

Disenar y entrenar la red neuronal probabilistica.

En esta estapa se definiran las variables de entrada y se declararan las neuronas patrones y las clases de salida que permitiran la clasificacion, para proceder con el entrenamiento de la red que

se hara a traves del metodo de jackknifing para determinar el parametro de escalamiento correcto que permita la mejor clasificacion.

Pruebas unitarias de las funciones.

Debido a que los paquetes en R están conformados, entre otros elementos por las funciones primarias, a cada una de ellas se les realizaran pruebas unitarias en dos fases, la primera con datos sintéticos que permitan comprobar cada estado del diagrama de flujo, que esquematiza la solución numérica que permite el cálculo de cada uno de los índices fisiológicos para entre otra herramientas pueden utilizarse los paquetes de R*RUnit* (Zenka, 2015) y *testthat* (Wickham, 2017), y la segunda etapa donde cada función asociada a un índice fisiológico se le realizan prueban con los conjuntos de datos de prueba que formaran parte integral del paquete y con los cuales se desarrollaran los ejemplos prácticos que conformaran la documentación que acompaã al paquete R.

Implementar y realizar pruebas de clasificacion.

El objetivo de esta etapa es definir los casos de prueba que seran usados para buscar la mejor clasificacion de densidad de siembra de tuberculos de papa criolla y realizar las pruebas con los diferentes escenarios a preparar.

Comparacion de resultados.

En este paso se procedera a la construccion de las curvas ROC asociadas a los casos de prueba planteados para asi obtener las mejores clasificaciones, permitiendo concluir las mejores relaciones entre variables, el mejor manejo de los datos y si los mismos deben ser modelados o es suficientemente buena una clasificacion con datos observados a pesar de la dependencia espacial existente.

Graficar y documentar resultados.

Finalmente se procedera a la graficacion por medio de R de los resultados, que sustenten las conclusiones realizadas y a relizar la documentacion que permita el uso de esta investigacion en futuras aplicaciones de campo.

Chequear la carga del paquete.

En esta etapa del desarrollo se utilizaran las funciones de chequear paquete que ofrece el código R; cuya finalidad es verificar cada fichero del árbol de carpetas asociadas a cada elemento de la estructura o esqueleto del paquete, que a su vez creara el archivo de documentación en LaTeX y/o HTML, compilará el código fuente y creará las librerías de enlace dinámico (*dynamic link library* DLL).

Construcción del método de distribución del paquete.

Se seleccionara la forma de distribución del paquete desde el repositorio local, creando los ficheros fuentes (en formato *tarball*) y en binario.

3.5. Aspectos administrativos

La realización de la investigación será planificada según lo establecido en el siguente diagrama:

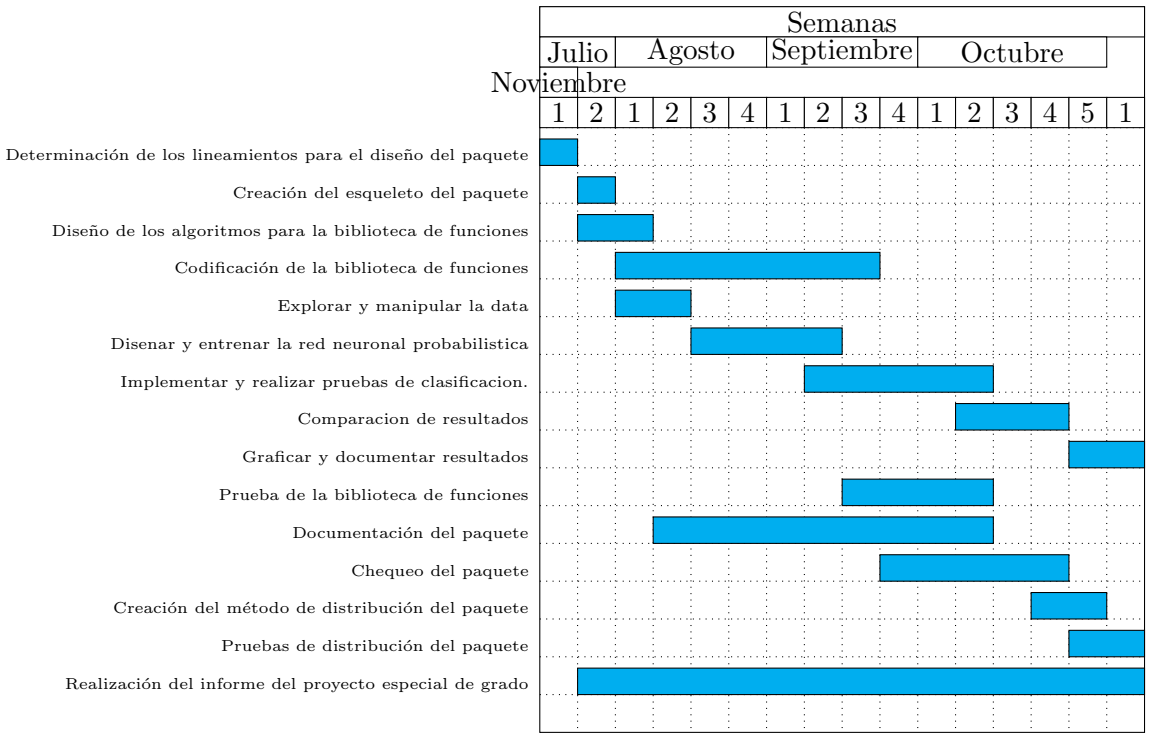


Figura 3.1: Diagrama de Gantt con la planficación del proyecto especial de grado

Referencias Bibliográficas

Available CRAN Packages By Name. Disponible en:https://cran.r-project.org/web/packages/available_packages_by_name.html. Consultada Junio 2018.

Lanz Brett, Reino Unido. Machine Learning with R. 2015.

Bohorquez Ingrid, Ceballos Ermilson. Algunos conceptos de la econometría espacial y el análisis exploratorio de datos espaciales. 2008.

Stephen Gang Wu, Forrest Sheng Bao, Eric You Xu, Yu-Xuan Wang, Yi-Fan Chang, Qiao-Liang Xiang. A Leaf Recognition Algorithm for Plant Classification Using Probabilistic Neural Network. 2007.

Nur Badariah Ahmad Mustafa, Kumutha Arumugam, Syed Khaleel Ahmed, Zainul Abidin Md Sharrif. Classification of Fruits using Probabilistic Neural Networks - Improvement using Color Features. 2011.

Bernal Nelson. Modelado del Calibre y la competición intra-específica por rendimiento de tubérculos de papa variedad Solanum phureja bajo diferentes densidades de siembra. 2017.