Universidad Nacional Experimental del Táchira

Vicerectorado Académico

Decanato de Docencia

Departamento de Ingeniería en Informática

Trabajo de aplicación profesional

Proyecto especial de grado

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 de 2018.2018

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

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 G - Suarez M, 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 íntimamente relacionado con algunos factores internos como fotosíntesis, respiración, transpiración, condiciones de estrés, concentración enzimática, balance hormonal y expresión genética» (Pineros C, 2009).

En la papa son dos los procesos fisiologicos asocidos directamente al rendimiento de la misma,

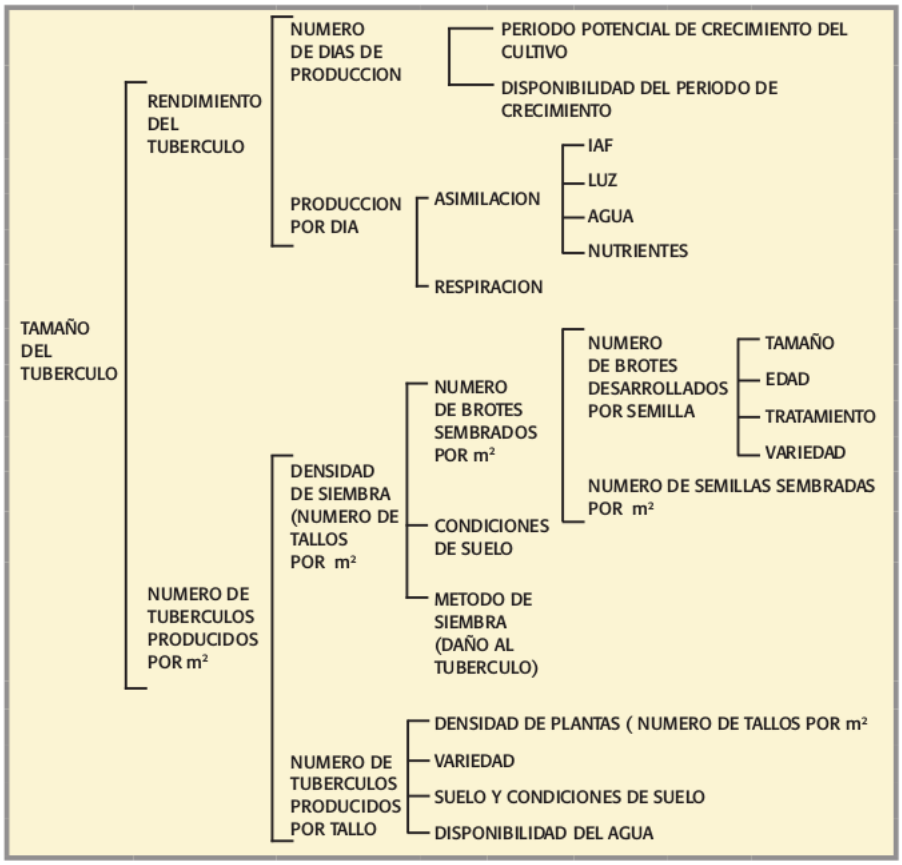
la fotosintesis y la respiracion. Y estos dos son procesos intimamente asociados, ya que durante la fotosintesis se producen carbohidratos que son consumidos durante la respiracion, pero una gran cantidad de estos carbohidratos contribuyen al proceso de produccion incrementando el tamano de los tuberculos, 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 C, 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 (Darghan E - Bernal E - Rodriguez L, 2017).

Las variables que influyen en el rendimiento de la papa pueden ser observados en la 1. Pag.3.

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 C, 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.

Figura 1: Variables de influencia sobre el rendimiento de Solanum Phureja.



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 G, 2007).

El objetivo de esta investigacion es construir un clasificador de densidades de siembra a traves de una red neuronal probabilistica, 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

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 B, 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 B, 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.

4. **Calibre de los tuberculos recogidos.** Variable cualitativa, con 4 valores posibles que definen a que calibre pertenece un tuberculo. Siendo los calibres dividivos en tuberculos menores a dos centimetros (¡2cm), entre dos y cuatro centimetros (2-4cm), entre cuatro y seis centimetros (4-6cm) y mayores a seis centimetros (6cm).

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 A, NF).

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

1.2.2. Objetivos Específicos

-

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 - Ñustez - Estrada, 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 - Gomez - Rodriguez, 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, Darghan 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. (Darghan E - Bernal E - Rodriguez L, 2017)

1.4. Alcance y Limitaciones