ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA DE INGENIERÍA EN ESTADÍSTICA INFORMÁTICA

"TÉCNICAS ESTADÍSTICAS MULTIVARIANTES PARA IDENTIFICAR Y CLASIFICAR LOS FACTORES INFLUYENTES EN LA PRODUCCIÓN DE ARROZ, ECUADOR 2017"

Trabajo de titulación

Tipo: Proyecto de investigación

Presentado para obtener el grado académico de:

INGENIERO EN ESTADÍSTICA INFORMÁTICA

AUTOR: JOSÉ LUIS CONDO LEÓN

DIRECTORA: Ing. JOHANNA ENITH AGUILAR REYES

Riobamba – Ecuador

2020

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA DE ESTADÍSTICA INFORMÁTICA

El Tribunal del trabajo de titulación certifica que: El trabajo de titulación: Tipo: Proyecto de Investigación, **"TÉCNICAS ESTADÍSTICAS MULTIVARIANTES PARA IDENTIFICAR Y CLASIFICAR LOS FACTORES INFLUYENTES EN LA PRODUCCIÓN DE ARROZ, ECUADOR 2017"**, realizado por el señor: **José Luis Condo León**, ha sido minuciosamente revisado por los Miembros del Tribunal del trabajo de titulación, el mismo que cumple con los requisitos científicos, técnicos, legales, en tal virtud el Tribunal Autoriza su presentación.

	FIRMA	**FECHA (año-mes-día)**
Dra. Jaqueline Elizabeth Balseca Castro **PRESEDENTA DEL TRIBUNAL**	JAQUELINE ELIZABETH BALSECA CASTRO Firmado digitalmente por JAQUELINE ELIZABETH BALSECA CASTRO Fecha: 2020.06.24 14:48:37 -05'00'	2020-06-24
Ing. Johanna Enith Aguilar Reyes **DIRECTORA DE TRABAJO DE TITULACIÓN**	Firmado electrónicamente por: JOHANNA ENITH AGUILAR REYES	2020-06-27
Ing. Nancy Elizabeth Chariguamán Maurisaca **MIEMBRO DE TRIBUNAL**	Firmado electrónicamente por: NANCY ELIZABETH CHARIGUAMAN MAURISACA	2020-06-24

Yo, JOSÉ LUIS CONDO LEÓN, declaro que el presente trabajo de titulación es de mi autoría y los resultados del mismo son auténticos. Los textos en el documento que provienen de otras fuentes están debidamente citados y referenciados.

Como autor asumo la responsabilidad legal y académica de los contenidos de este trabajo de titulación; el patrimonio intelectual pertenece a la Escuela Superior Politécnica de Chimborazo.

Riobamba, 18 de febrero del 2020

José Luis Condo León
060417229-6

DEDICATORIA

Dedico este trabajo de titulación a mi madre Narcisa por todo el amor y apoyo que me brinda, siendo un ejemplo de dedicación y esfuerzo para nuestra familia; a mi padre, Segundo, a mis hermanos Carlos y Edison por ser incondicionales en mi vida; a Liseth por ser mi inspiración y la luz que guía mi camino.

AGRADECIMIENTO

Agradezco a Dios por todas las bendiciones que ha colocado en mi vida, dándome fortaleza en los momentos más difíciles; a mi madre por todo el esfuerzo y dedicación que da por nuestra familia, a mi padre y hermanos por ser pilar fundamental de mi vida; a Liseth por inspirarme cada día a ser mejor; a todos mis profesores que supieron guiarme en mi formación académica; y a todos mis amigos por los buenos momentos compartidos.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE GRÁFICOS

ÍNDICE DE ANEXOS

RESUMEN

Esta investigación tuvo como objetivo comparar técnicas estadísticas multivariantes para identificar y clasificar los factores influyentes en la producción de arroz en Ecuador en el año 2017. Se uso 18 variables agronómicas, 15 cualitativas y 3 cuantitativas medidas en sembríos de arroz que participaron en la Encuesta de Superficie y Producción Agropecuaria Continua (ESPAC) 2017, la investigación no fue experimental y tuvo un alcance exploratorio relacional, se usó Análisis Exploratorio de Datos (AED), Árboles de Regresión (AR), Análisis Factorial de Datos Mixtos (AFDM) y el software R para desarrollar el estudio. El AED determinó que la producción promedio de arroz es 212.21 libras por hectárea, el 86% de los cultivos produjo entre 200 y 250 libras, se determinó que las variables superficie sembrada y superficie cosechada no tiene correlación con la variable producción de arroz. Después de la aplicación de los métodos multivariantes el modelo de AR con un valor RMSE=7.32 detectó como factores influyentes en la producción de arroz a superficie sembrada, superficie cosechada, uso de riego, uso de fertilizante orgánico y clase de semilla. Por otro lado, el AFDM con un modelo que explicó el 81% de la variabilidad de los datos generó dos agrupaciones de factores influyentes en la producción de arroz, el primer grupo está compuesto por las variables uso de fertilizantes, uso de fertilizantes químicos, uso de plaguicidas químico y uso de fitosanitarios, mientras que la segunda agrupación está compuesta por superficie sembrada y superficie cosechada. Por tal motivo se pudo determinar que los factores influyentes que tienen en común las dos técnicas son superficie sembrada y superficie cosechada, los factores más importantes para los AR son uso de riego y superficie cosechada, mientras que para el AFDM son uso de fertilizantes y superficie sembrada. Es necesario ampliar la investigación con el uso de otras técnicas de minería de datos.

Palabras Claves: <ESTADÍSTICA>**,** <MULTIVARIANTE>, <PRODUCCIÓN DE ARROZ>, <ANÁLISIS FACTORIAL DE DATOS MIXTOS>, <ÁRBOLES DE REGRESIÓN>, <FACTORES INFLUYENTES>.

SUMMARY

This research aimed to compare multivariate statistical techniques to identify and classify the influential factors in rice production in Ecuador in 2017. 18 agronomic variables, 15 qualitative and 3 quantitative measures were used in rice fields that participated in the Survey of Surface and Continuous Agricultural Production (ESPAC) 2017, the research was not experimental and had a relational exploratory scope, an Exploratory Data Analysis (AED), Regression Trees (AR), Mixed Data Factorial Analysis (AFDM) and the R software to develop the study. The AED determined that the average rice production is 212.21 pounds per hectare, 86% of the crops produced between 200 and 250 pounds, it was determined that the variables planted area and area harvested have no correlation with the rice production variable. After the application of multivariate methods, the AR model with an RMSE-7.32 value detected as influencing factors in the production of rice on sown surface, harvested area, use of irrigation, use of organic fertilizer and seed class. On the other hand, the AFDM with a model that explained 81% of the variability of the data generated two groups of factors influencing rice production, the first group is composed of the variables fertilizer use, chemical fertilizer use, use of chemical pesticides and the use of phytosanitary products, while the second group is composed of planted area and harvested area. For this reason, it was possible to determine that the influential factors that the two techniques have in common are planted area and harvested area, the most important factors for RA are use of irrigation and harvested area, while for AFDM they are use of fertilizers and surface sown. It is necessary to expand research with the use of other data mining techniques.

Keywords: <FACTORIAL ANALYSIS OF MIXED DATA>, <REGRESSION TREES>, <INFLUENCING FACTORS>, <ANALYSIS OF MAIN COMPONENTS>, <CORRESPONDENCE ANALYSIS>, <RICE PRODUCTION>, <AGRONOMIC VARIABLES>.

INTRODUCCIÓN

La agricultura es uno de los temas de investigación más importantes hoy en día, debido a la influencia que tiene sobre la economía de un país y sobre el bienestar en general la población humana. Ésta se convierte en uno de los motores básicos de desarrollo de una nación, aportando de manera significativa a la economía de las arcas de un estado y generando empleo para la sociedad.

Más que un motor de bienestar económico de una nación es la base fundamental de seguridad alimentaria, sin la generación apropiada de alimentos que cubra las necesidades básicas alimentarias de un país, éste está encaminado al fracaso y subdesarrollo.

El progreso tecnológico ha permitido que las sociedades más desarrolladas puedan satisfacer sus necesidades alimentarias, desde la revolución industrial y con el pasar de los años la brecha de hambre se ha reducido. (Klohn & Appelgren, 1999, págs. 105-126), pero ahora la sociedad se enfrenta a un nuevo problema y este es la sobrepoblación, con el pasar de los años la producción de bienes alimenticios no da abasto para cubrir las necesidad del planeta (Chonchol, 1983, págs. 189-206).

Con el fin de solucionar este problema se crea nuevas formas para la producción agropecuaria, se comienzan a tomar en cuenta factores como el clima, las tecnologías, políticas de intercambio, ayuda gubernamental, maquinaria entre otros, lo que conlleva a un desarrollo excepcional en la producción agrícola.

A nivel latinoamericano existen diversidad de investigaciones, donde mediante el estudio de los factores de producción han podido mejorar significativamente la producción agrícola de algún producto, un ejemplo de esto es Argentina en los años 90 que, con el uso de biotecnología, capacitación productiva y regulaciones económicas se planteó como un exportador agrícola (Bisang , 2003).

Existen diversos productos alimenticios de prioridad a nivel mundial, uno de los más importantes es el arroz, según un informe publicado por la (Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO), 2004, págs. 1-2) este alimento es el más importante entre 17 países Asiáticos, 9 países de América y 8 países de África, contiene muchas cualidades alimenticias que lo convierten en un producto indispensable para las personas.

Las regiones con mayor productividad de arroz en el mundo son los países asiáticos, ciertas regiones de Europa, países Latinoamericanos y Estados Unidos, esta productividad siempre se ve

incrementada según el pasar de los años, rompiendo récords anualmente (Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO) , 2018, pág. 1).

Es así que esta investigación va dirigida al estudio de la productividad del arroz y sus diferentes factores influyentes, la literatura muestra investigaciones con similitud tales como la realizada por (Muriel Osorio , 2013, págs. 43-45) en la que determina que el género de una persona está influenciado de manera significativa en la productividad del arroz, otra investigación determino que los niveles de producción de arroz con los métodos tradicionales y el sistema intensificado de cultivo de arroz(SICA) son iguales (Ochoa Herrera , 2016, págs. 43-45), en Uruguay se pudo concluir que los niveles de fósforo, arena, implantación, láminas de agua, y control de malezas están influenciando directamente sobre la productividad de la planta de arroz (Bonilla , Terra , Gutierrez , & Roel , 2015), por otro lado, una investigación en Venezuela determino que la capacidad de riego es determinante en la buena producción de esta gramínea.

Existen ciertas investigaciones en Ecuador que es conveniente resaltar, tales como el estudio realizado por (Pozo Galarraga, 2017, pág. 89) donde llego a la conclusión que el precio de los fertilizantes usados en el Ecuador es superior a los precios dispuestos en otros países, convirtiéndose en un factor determinante en la producción de esta gramínea.

Otra investigación importante se enfocó en encontrar los principales factores que influyen en la productividad de la planta de arroz, determinando que el mal manejo de plagas, mal manejo de fertilizantes, uso de mala semilla, riego inadecuado, mal manejo de suelos y tecnología nula son factores relevantes (Velásquez Burbano , 2016, págs. 120-124).

También se ha descubierto que un mayor nivel de inversión por parte de las instituciones privadas bancarias en el otorgamiento de créditos conlleva a una mayor productividad del alimento (Corporación Financiera Nacional, 2018) y la aplicación de políticas gubernamentales en la maximización de productividad del arroz no ha dado resultados favorables (Bonilla Bolaños & Singaña Tapia , 2019, pág. 70)

Debido a la mejora de calidad en producción agropecuaria de otros países es importante el mejoramiento de todos los factores agrarios del estado, mediante el uso de análisis estadístico que toma importancia pues los resultados positivos que ha aportado a esta área la ha convertido en una herramienta indispensable en el desarrollo correcto de la agricultura, siendo su fuerte la generación de modelos predictivos (Montes de Oca , Garcia Pereira, & Hernández Gómez , 2009, págs. 74-77)

El problema que se pretende solucionar con este estudio es el planteado por la Corporación

Financiera Nacional (CFN) quien en un informe resalta la disminución progresiva de la producción anual de arroz en el Ecuador, por lo que se hace indispensable encontrar los factores influyentes en la productividad de esta planta.

La presente investigación se realiza en toda el área geográfica ecuatoriana, los sujetos de investigación son todos los productores de arroz que participaron en la Encuesta de Superficie y Producción Agropecuaria (ESPAC) del año 2017.

El principal objetivo de la investigación es determinar cuáles son los factores que influyen en la producción de arroz con el uso de las técnicas estadísticas multivariantes, Árboles de decisión y Análisis Factorial Mixto, al mismo tiempo contrastar los resultados que las dos técnicas aportaron y determinar si existe alguno tipo de relación entre las variables usadas.

Antecedentes

La agricultura es una de las bases económicas principales de un país, ha permitió obtener grandes recaudaciones en el Producto Interno Bruto (PIB), además ha generado auto sustentabilidad alimentaria para el estado, la característica geográfica de los colectivos contribuye que exista mucha diversidad en la producción de alimentos; sin embargo es importante conocer que las primeras actividades agrícolas fueron la recolección de frutos silvestres a nivel del mundo, posteriormente los cultivos crecieron y se expandieron en base al regadío de agua, iniciativa que fue tomada por las primeras civilizaciones. El crecimiento constante de la producción alimentaria alrededor del mundo ha contribuido al desarrollo y progreso del hombre ya que su capacidad de poder satisfacer las necesidades alimentarias gira en torno a la diversificación de productos que forman parte de su canasta nutricional (Chonchol, 1983, págs. 189-206).

Uno de los problemas que hoy enfrenta la agricultura mundial es el rápido crecimiento de la población, la deficiente producción y la demanda alimentaria extrema, tal es la importancia de la agricultura en los pueblos que la diminución de producción de alimentos de primer orden puede transformarse en problemas de desnutrición, especialmente en los sectores sociales más vulnerables (Chonchol, 1983, págs. 189-206); sin embargo el siglo pasado, a partir de 1960 se registró grandes incrementos en la actividad agrícola al igual que la porción alimentaria de cada persona, en 1970 el crecimiento agrícola fue tal que el 50% de la población mundial consumía el alimento diario requerido, en el año 1990 la producción de alimento aumento al punto que más del 50% de la población mundial tenía un nivel alimentario satisfactorio. (Klohn & Appelgren, 1999, págs. 105-126).

A nivel latinoamericano se han implementado diversas soluciones para obtener la sustentabilidad alimentaria, un caso de la diversificación de la productividad agrícola se dió en Argentina en los años 90, donde el uso de la biotecnología junto con regulaciones económicas en favor de los productores, uso de tecnología en el cultivo y capacitación productiva, hizo que la producción alimentaria de Argentina tenga rasgos de exportación ubicando en la percha exportadora a un país que años atrás no daba abasto a las necesidades locales. (Bisang , 2003, págs. 413-442) delimitando que mientras mayor avance tecnológico y mejor tratamiento agrícola reciban las producciones, estas elevan su calidad.

En el territorio ecuatoriano la agricultura comercial de monocultivos está decayendo drásticamente en su mayoría por causas de degradación de suelo, falta de agua y sistemas agro productivos actuales (Franco W. , 2016), estas pueden ser una de las razones principales de la disminución de cosechas de productos emblemáticos en particular del arroz a pesar de que la superficie destinada para sus cosechas ha incrementado, dentro del país la producción de arroz es

una de las actividades de mayor relevancia e importancia agrícola, debido a que este producto es uno de los pilares fundamentales de consumo en los hogares ecuatorianos, teniendo beneficios y propiedades alimenticias idóneas para una correcta alimentación.

Existen diversas investigaciones que dan razón a la baja productividad agropecuaria en el Ecuador, al ser el arroz uno de los productos más importantes de la canasta básica, este es de los alimentos que se han estudiado a fondo.

El rendimiento del cultivo de arroz en el 2017 ha tenido una disminución comparándolo con el año anterior, la provincia con mayor producción de arroz fue Loja y la provincia con menor producción de este producto fue Los Ríos. Las principales causas en la reducción de la productividad de la planta de arroz son los factores que tienen que ver con el factor climático, en particular con el incremento de las lluvias. (Castro, 2017, págs. 1-8)

Un estudio importante sobre la productividad de Arroz determinó los factores que influyen en la cadena agroalimentaria en Ecuador, se realizó con información recabada entre los años 2005-2014, como resultados se halló que existe tres etapas en la estructura agroalimentaria, la producción, procesamiento y comercialización.

La producción se concentra en Guayas y Los Ríos cubriendo el 92%, las unidades productivas en su mayoría pequeños y medianos productores con el 80%, la producción ecuatoriana de arroz tan solo representa el 0,22% a nivel mundial, siendo Colombia el mercado con mayor accesibilidad, también se determinó que la producción de arroz ha disminuido en un promedio del 11% anual y que los precios están siendo afectados por la etapa de la comercialización.

Los sistemas de cultivos utilizados son el tradicional, semi-tecnificado y tecnificado con 27%, 45% y 19% respectivamente, lo que indica un grado bajo de tecnificación en la producción de arroz, los principales problemas tecnológicos que fueron detectados son desconocimiento de manejo de plagas, mal manejo de fertilizantes, uso de semilla no certificada, riego inadecuado, mal manejo de suelos, a nivel de procesamiento se pudo detectar baja infraestructura en los sistemas de almacenamiento del producto para su posterior comercialización.

Por otra parte se pudo identificar que el principal factor social que afecta la producción de arroz es la poca asociatividad entre todos los individuos que intervienen en el proceso, otro factor social influyente en la producción de arroz es el grado de educación de los productores; los factores económicos detectados fueron la poca innovación tecnológica en esta área es la falta de créditos gubernamentales y la poca efectividad de los procesos gubernamentales en la actualización de

conocimientos productivos (Velásquez Burbano , 2016, págs. 120-124).

Otras investigaciones recientes permitieron determinar que nuevas políticas gubernamentales en Ecuador enfocadas a maximizar la producción de arroz ha tenido efectos contrarios a los esperados, la utilización de insumos químicos y variedades mejoradas de la planta no ha mostrado maximizar la producción, en cambio afectó factores como la biodiversidad, asociación de producción y rol de la mujer en la producción de arroz (Bonilla Bolaños & Singaña Tapia , 2019, págs. 70-83).

Investigaciones relacionadas resaltan otros factores influyentes en la productividad del arroz, la primera fase en una industria agroalimentaria es la producción, este estudio permitió resolver que el precio de los fertilizantes usados en Ecuador son el doble en comparación con países vecinos, esto debido a políticas tributarias, gastos financieros y márgenes de comercialización, se determinó también que el arroz es parte de los productos más sensibles al cambio de precio si los precios de los fertilizantes se ven afectados, por lo que se señala que una de las problemáticas atacar son las políticas tributarias y la falta de producción de propios agroquímicos en el país. (Pozo Galarraga, 2017, pág. 89).

Es posible observar que la forma de producción de arroz en Ecuador ha sido modificada, en los últimos años la banca ha tenido una participación cada vez más elevada, siendo quienes otorgan créditos para que se realice esta producción, teniendo como consecuencia la menor participación del estado. Las agrupaciones de inversionistas también han tenido un impacto en la producción de este bien alimenticio, pues cada vez es posible observar más asociaciones para tener mayor competitividad en el mercado. (Corporación Financiera Nacional, 2018).

Planteamiento del problema

El arroz es una de los gramíneas de mayor producción a nivel nacional, elevando o disminuyendo los porcentajes del Producto interno bruto (PIB), diferentes organizaciones gubernamentales tuvieron incertidumbre respecto a los niveles de su producción en los últimos años, la Corporación Financiera Nacional (CFN) en la Ficha sectorial: Arroz de Febrero del 2018 resalta la disminución progresiva de producción de este alimento, investigaciones adyacentes mencionan que ciertos factores climáticos determinan la disminución de producción sin embrago es indispensable estudiar este fenómeno a través de los factores influyentes del cultivo y el mantenimiento de la planta durante su crecimiento.

La investigación permitirá resolver preguntas como: ¿Existen factores que influyen en los niveles de producción de arroz en el Ecuador en el año 2017?, ¿Los factores influyentes en la producción de arroz determinados por las diferentes técnicas estadísticas son los mismos?, y ¿Las investigaciones relacionadas con el tema obtuvieron resultados similares?

Justificación

El análisis estadístico es una herramienta primordial para poder recabar información sobre fenómenos y experimentos, siendo el análisis multivariante el conjunto de técnicas predominantes en el campo de análisis ya que permite generar información del fenómeno de estudio que a simple vista es difícil de observar, ayudando también a establecer posibles relaciones entre variables.

Según su conceptualización el análisis multivariante permite la visualización de relaciones entre individuos de estudio, discriminando agrupaciones caracterizadas por uno o varios factores. (Jaureguizar, 1990, págs. 191-207).

Aunque el análisis estadístico multivariante requiere de mucha más información del fenómeno de estudio que el análisis estadístico univariante, entre las diferentes ventajas o la más importante que tiene el usar el análisis estadístico multivariante se encuentra que la inferencia tendrá mayor consistencia y fiabilidad lo cual permite mayor precisión en resultados y generación de conocimiento respecto al fenómeno estudiado (Lozares Colina & López Roldán , 1991, págs. 9-29)

En el sector financiero y económico el cálculo y predicción de indicadores, así como la caracterización de información lo vuelve imprescindible, entre estas técnicas usadas se resalta de mejor forma los árboles de decisión. (Cardona Hernández, 2004, págs. 139-151)

El análisis multivariante es de las técnicas más utilizadas a nivel de agricultura para la generación de un bien de producción intervienen varios factores, como tipo de semilla, lugar de producción, condiciones, formas de producción y muchos otros factores que con el uso de técnicas multivariantes se podrá verificar su influencia en la productividad del arroz, para la determinación de posibles factores influyentes en las producción de arroz se utilizará el Análisis de Componentes Principales, Análisis Factorial Mixto y Arboles de Decisión, mismas que contribuirán a caracterizar a los individuos y clasificar los grupos de variables que contiene la base de información. El uso de estas herramientas esta respalda en muchas investigaciones donde se propone que el análisis estadístico multivariado se seleccionó como herramienta perfecta para tipificación y clasificación de fincas en este caso de sembríos (Escobar & Berdegué , 1990, págs. 34-39).

Tanto el uso del Análisis Factorial Mixto y Arboles de Decisión permiten determinar factores que influyen en la producción de la planta de arroz, siendo una de las características de estas técnicas la reducción de dimensionalidad de la información, la caracterización de los individuos y permitir la agrupación de variables, a más de determinar posibles factores influyentes se puede contrastar los resultados que cada una de ellas arrojan, haciendo que exista un aporte científico a la mejor aplicabilidad de las técnicas estadísticas multivariantes.

El arroz es el cultivo más extenso del Ecuador, ocupa más de la tercera parte de la superficie de productos transitorios del país, en términos sociales es la producción alimenticia más importante del país, a nivel nutricional es la que mayor aporte de calorías brinda a quienes lo consumen según la Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO).

Se hace relevante e importante el estudio del sector agropecuario debido a tres razones, a su elevada participación en el PIB, fuente de entrada de divisas y aporte a la constitución de políticas de soberanía alimentaria sobre las que se rige un colectivo (INEC, 2017, págs. 3-6).

Dentro de la producción agropecuaria transitoria, es posible encontrar diversos productos entre ellos el arroz, según cifras propuestas por el (INEC, 2017, págs. 3-6), es uno de los cultivos más grandes existentes en el país, se ubica en las provincias del Guayas y Los Ríos, siendo uno de los motores económicos internos del estado, generando plazas de trabajo y auto sustentabilidad al país.

Es importante el estudio de la producción agrícola de arroz, debido a que muchos productos emblemáticos como el banano y el cacao han tenido decrementos en sus niveles de producción, lo cual no es ajeno al caso del arroz, pues mediante el análisis evolutivo de área de producción y nivel de producción se ha podido observar que se está cultivando en mayor superficie y se está obteniendo niveles de producción cada vez más inferiores.

Una investigación realizada por la CFN en el año 2018 revelo que el cultivo de arroz sufrió un

decremento del 13% por año desde el 2013 hasta el 2016 (Corporación Financiera Nacional, 2018), de manera más específica en los últimos dos años se calcula que hubo decremento del 2.21% en los cultivos de arroz en el periodo 2016-2017 (INEC, 2017), por lo que es posible concluir que el arroz es uno de los productos emblemáticos ecuatorianos que está sufriendo reducción en sus cultivos y en su producción, lo que da un motivo para realiza la investigación de ésta gramínea.

OBJETIVOS

Objetivo General

Comparar técnicas estadísticas multivariantes para identificar y clasificar los factores influyentes en la producción de arroz en Ecuador año 2017.

Objetivos Específicos

- Determinar los factores influyentes en la producción de arroz.
- Analizar la dependencia entre variables cualitativas relacionadas al cultivo del arroz.
- Contrastar los factores influyentes obtenidos por cada técnica estadística multivariante.

CAPÍTULO I

1. MARCO TEÓRICO REFERENCIAL

1.1. La agricultura en el Ecuador

En la actualidad la producción agrícola ecuatoriana está dividida en ocho sectores, de los cuales la producción de pastos, cultivos permanentes, cultivos transitorios y pastos naturales son las áreas más importantes.

Es de observar que la producción de todas estas áreas está en decremento, en los últimos años se ha podido notar una disminución del nivel de productividad, según cifras del INEC obtenidas en la Encuesta de Superficie y Producción Agropecuaria Continua (ESPAC) 2017.

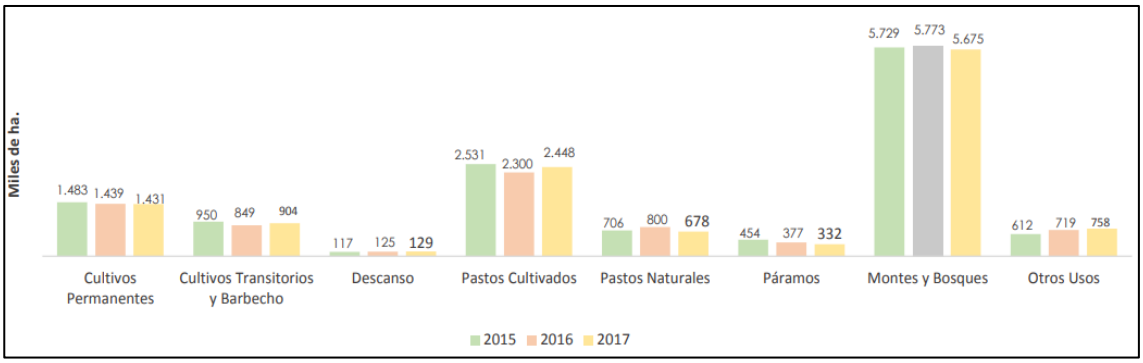


Gráfico 1-1. Superficie de producción de cada sector agropecuario en el Ecuador 2015-2017.
Realizado por: Condo León José Luis, 2019.

El sector más extenso existente en el territorio ecuatoriano son los montes y bosques seguidos por los pastos cultivados que a su vez sirven de mayor forma al área de producción de derivados animales; el tercer sector más grande es el de cultivos permanentes seguidos por los cultivos transitorios y los pastos naturales; las áreas en descanso junto con las áreas de otros usos son las de menor tamaño.

1.1.1. Pastos cultivados

Se trata de un grupo de gramíneas y leguminosas cuya principal función es generar alimento para animales, este alimento contribuye a dar proteínas, carbohidratos y generar un desarrollo positivo en el crecimiento de éste, de tal manera que puedan ser usados para obtener productos derivados, Los pastos cultivados también han tenido un decremento en su producción en los últimos tres años, la planta con mayor cultivo en esta área es la Saboya seguido por Pasto miel y Brochiaria.

1.1.2. Pastos naturales

Los pastos naturales pueden ayudar a las poblaciones dedicadas al pastoreo a adaptarse al cambio climático, debido a que el carbono captado por estos, añadido al suelo mejora la capacidad de retención del agua y con ello su capacidad para resistir las sequías, albergan una gran biodiversidad, ligeramente inferior al de los bosques, lo que disminuye la vulnerabilidad del ecosistema de la pradera natural, conformada por especies de animales, plantas y de microorganismos que residen en las tierras de pastoreo. (Riva , Valer , & Perez , 2014, págs. 7-9), estos han sufrido una disminución progresiva de su productividad en los últimos 10 años, haciendo que en la actualidad sea el menor de grupo de cultivo en el Ecuador.

1.1.3. Cultivos permanentes

También llamados cultivos perennes, entre sus principales características es posible denotar que el tiempo de su etapa de desarrollo y crecimiento es relativamente mayor a los de los cultivos transitorios y este tipo de plantas generan varias cosechas antes de terminar su vida productiva. Se ha podido observa que la producción de estos cultivos está en descenso en los últimos tres años, siendo el cacao el cultivo con mayor producción con el 37,74% seguido por la palma africana con el 20,66% del sector.

1.1.4. Cultivos transitorios

Este tipo de cultivos está caracterizado en su mayoría porque tan solo tienen una cosecha o solo un ciclo de vida productiva, los productos representativos son el maíz, la yuca, la papa, el arroz entre otros; además sufren de igual forma una disminución en su producción en los últimos 3 años, se conoce hasta el momento que la producción de maíz duro seco representa el 35% de estos cultivos, secundado por el arroz con el 33% cuyos niveles de producción anual se ha visto muy variados de un año al otro, teniendo un decaimiento en el año 2017 , la papa es el único producto transitorio que se mantiene en niveles estables de producción.

1.2. El arroz

1.2.1. Taxonomía

El arroz es una planta perteneciente a familia Fanerógama, de tipo espermatofita y subtipo angiosperma, ésta es una planta denominada como monocotiledónea, se rige bajo el tipo de plantas de orden Glumiflora y dentro esta denominación pertenece a la familia de las Gramíneas y subfamilia Panicoideas. Según su origen silvestre ésta pertenece a la Tribu Oryzae, subtribu Oryzineas del género Oryza (Tascón & García , 1985, pág. 48).

Las plantas que pertenecen a la tribu Oryzae se caracterizan por tener una sola flor, pero también es posible observarla con dos o tres flores, de todas estas tan solo una de ellas es fértil, otra de sus cualidades es que puede tener hasta 6 estambres. Dentro de esta tribu es posible encontrar la subtribu denominada como Oryzineas, que poseen características tales como la presencia de glumas con espiguillas comprimidas lateralmente, por otro lado, las que no tienen glumas poseen espiguillas sobre pedúnculos muy cortos y comprimidos (Tascón & García , 1985, pág. 50).

Un análisis a profundidad de la especie de género Oryza dictaminó que este tipo de planta posee 12 cromosomas, y dentro de este grupo se encuentra la Oryza Sativa a la que pertenece el arroz, según estas denominaciones el arroz puede tener tres tipos “Indica”, “Japónica” y “Javánica o bulo”. El arroz de tipo “Indico” es de mayor altura, hojas largas, color verde pálido, y grano largo; por otro lado el arroz de tipo “Japónico” tiene hojas rectas, color verde, granos cortos y anchos, con mayor capacidad de captar el nitrógeno lo que maximiza su producción; y el arroz de tipo “Javánica o bulo” es similar a la “Japónica” pero con las únicas diferencia de que la planta es rígida y sus hojas son más anchas (Tascón & García , 1985, págs. 52-53).

1.2.2. Anatomía

En la Figura 1-3 se muestra la estructura de la raíz de la planta de arroz, en ésta es posible notar que ésta es similar a la de las plantas acuáticas; entre sus partes más importantes se observa la Epidermis donde se puede encontrar los pelos absorbentes; la Exodermis que se caracteriza por ser la capa exterior de la corteza; la Esclerenquima que está formada por células con paredes gruesas cuya función es proteger la raíz después de la desaparición de la Epidermis y Exodermis; la Parénquima Cortical que posee la fusión metabólica de la planta, su principal tarea es generar cámaras de aire en su interior para así permitir la respiración de las raíces (Tascón & García , 1985, pág. 54).

El Cilindro Vascular de la raíz es la parte más importante, dentro de ésta se encuentra la

Endodermis que favorece el desarrollo de la planta; el Periciclo que es un conjunto de tejidos localizada en la endodermis; el Floema cuya fusión es el traslado de los alimentos; el Xilema que sirve para la conducción de agua; y la Médula que se encargar del crecimiento de la planta (Tascón & García , 1985, pág. 55).

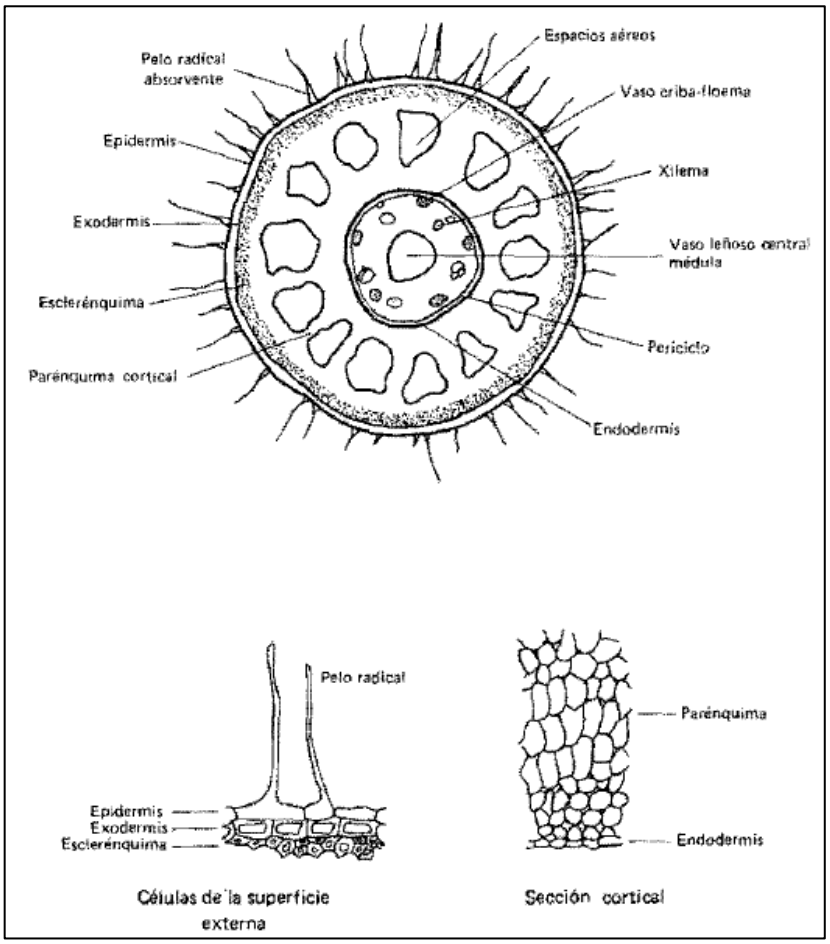


Figura 1-1. Corte transversal de una raíz de la planta de arroz
Fuente: (Tascón & García , 1985)

De la estructura de la planta de arroz el detalle de su semilla es el más relevante, en la Figura 1-1 es posible notar toda su estructura y principales partes, ésta se encuentra adherida al pericardio u ovario maduro, posee diferentes partes entre las que se desataca la Gluma, que es la capa exterior del grano de maíz; el Lemma o aristas que no siempre estas presentes; también es posible notar la presencia de Lemmas estériles que se encuentran ubicados en la base del grano; La raquilla que es la parte que une al pedicelo y la semilla; el Pericarpio que es la parte fibrosa de la semilla, compuesta por células entrecruzadas; y el tegumento y la aleurona que son la capa interior que protege a la semilla (Tascón & García , 1985, pág. 62).

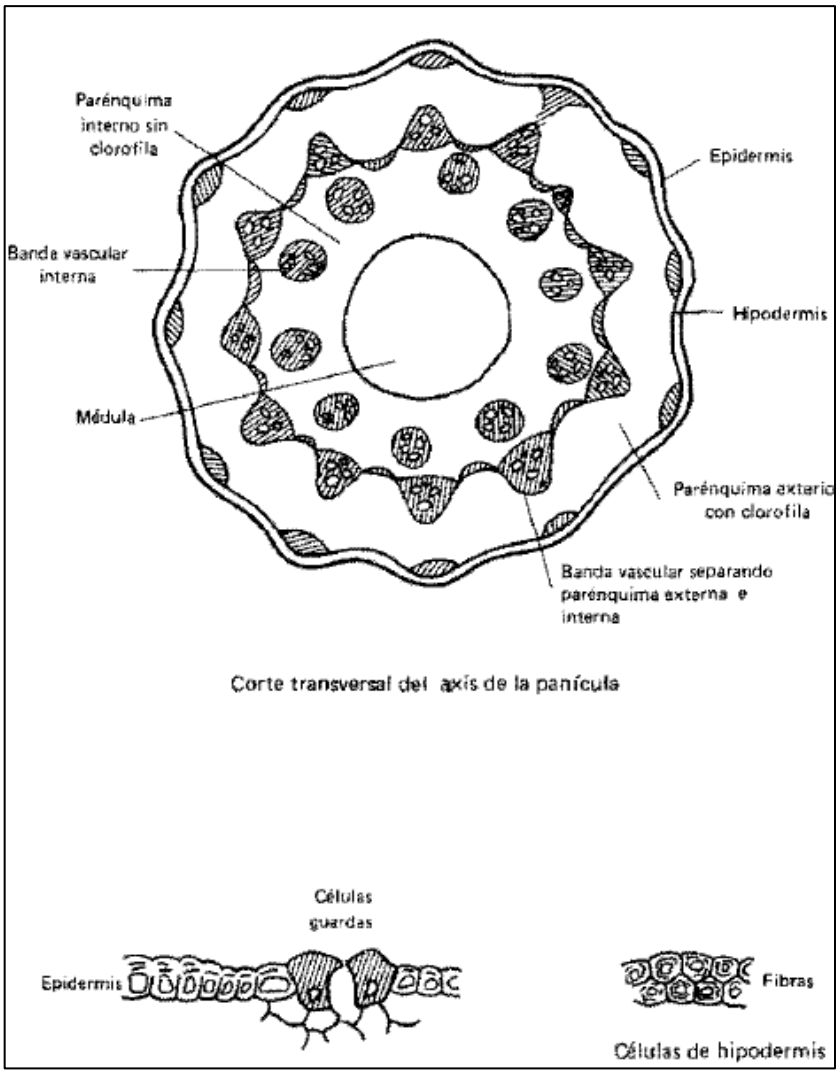


Figura 2-1. Corte transversal del axis de la panícula de la semilla de arroz

Fuente: (Tascón & García , 1985)

1.2.3. Origen

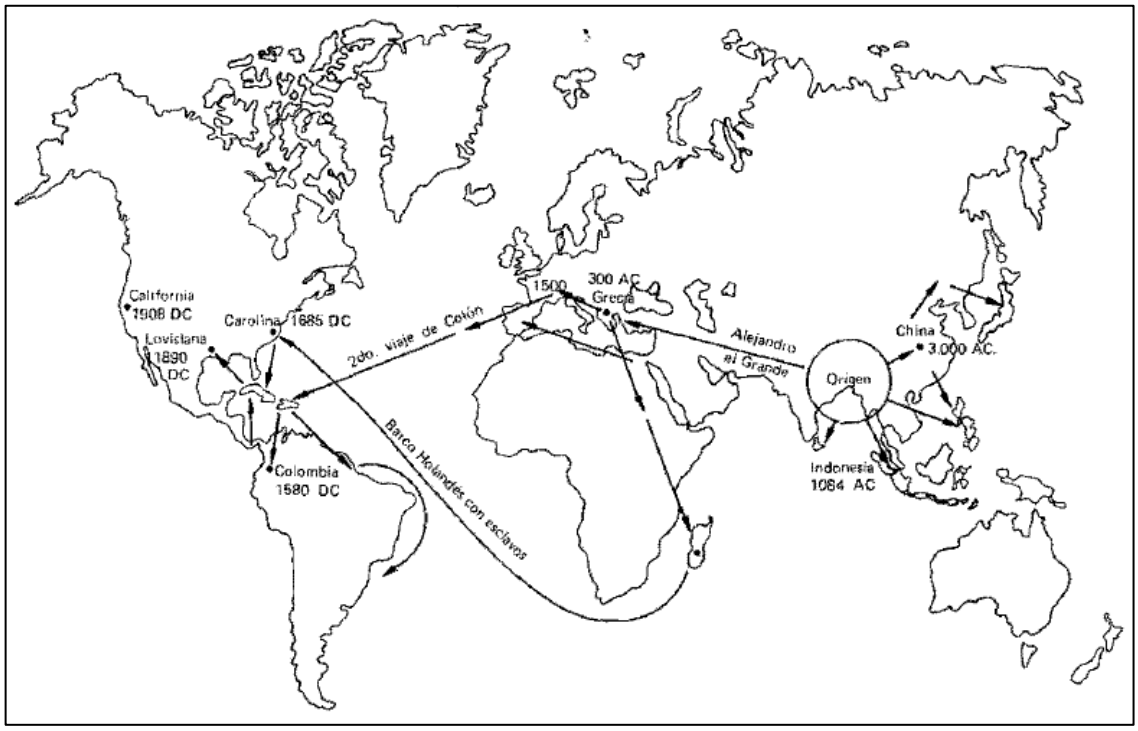


Figura 3-1. Movimiento del arroz en el mundo desde su punto de origen
Fuente: Arroz: investigación y producción: referencia de los cursos de capacitación sobre arroz dictados por el Centro Internacional de Agricultura Tropical, 1985

Establecer un origen de la planta es un reto, pero muchos autores en su literatura manifiestan y concuerdan que el origen de esta gramínea se dio en el sur de la India, donde es posible encontrar plantas silvestres del mismo tipo y las condiciones favorables para que estas se desarrollen, escritos chinos dan a notar que 3000 años antes de Cristo ya se realizaban ceremonias para la celebración de la siembra de esta planta, también se han encontrado restos fósiles de arroz en el valle de Yang-Se Kiang que datan de 3000 o 4000 años antes de Cristo. La mayor parte de los autores coinciden en que esta planta se propago desde la India hasta China, después 3000 años antes de Cristo de traslado desde China hasta Corea, para después ser trasladado desde Corea a Japón en el primer siglo antes de Cristo. En una época siguiente ya habiendo llegado este cultivo a Sri-Lanka, este fue trasladado al Mediterráneo por los persas 2000 años antes de Cristo, tiempo después y gracias a la invasión de Alejandro Magno a este territorio en el año 320 antes de Cristo este cultivo fue llevado a Grecia y a futuro se expandió por toda Europa (Tascón & García , 1985, pág. 47).

No es posible establecer la llegada de esta planta al hemisferio occidental por la falta de documentación, pero se estima que Cristóbal Colón en su segundo viaje en el año 1493 trajo consigo semillas de esta planta y los Holandeses en el siglo XVII introdujeron esta gramínea a norte América, con exactitud en Carolina (Tascón & García , 1985, pág. 48).

1.2.4. Factores medio ambientales

El arroz es una planta que puede cultivarse en lugares húmedos o con temperaturas elevadas que se encuentren a una altitud inferior a 2500 metros del nivel del mar, siendo el lugar idóneo a 45° de latitud norte y 40° al sur del Ecuador, los factores medio ambientales más influyentes para una producción idónea de la planta son temperaturas elevadas, muchas radiación solar y excesiva cantidad de agua, ésta última es un factor crítico pues se ha determinado que si no existe abundancia de agua la producción de arroz es nula (Tascón & García , 1985, pág. 19).

1.2.4.1. Temperatura

Este factor se hace importante pues afecta al crecimiento y al desarrollo de la planta, en el transcurso del crecimiento de ésta se hace necesarias diferentes temperaturas para una evolución favorable de esta gramínea, rangos de temperatura entre 0 y 4°C han resultado en la muerte de la planta, de 10 a 21°C aunque con poca investigación científica se ha determinado que tiene un crecimiento reducido y poca producción, las plantas sometidas a este rango de temperatura pueden recuperarse si ésta se eleva. En el estado de meiosis de las células madre una temperatura de 15 a 19°C causa que planta se torne estéril, el cambio de temperatura radicales en las diferentes etapas de crecimiento provoca producción deficiente, cuando la planta es expuesta a temperaturas elevadas entre 27 y 35°C también es posible notar esterilidad. Por lo que se concluye que las temperaturas desfavorables al correcto crecimiento y producción del arroz se da en rangos inferiores a 20°C y superiores a 30°C, a continuación, es posible observar rangos de temperaturas favorables para el crecimiento de la planta en sus diferentes etapas (Tascón & García , 1985, pág. 20).

Tabla 1-1: Respuesta del arroz a la variación de temperaturas

	Temperaturas Críticas (°C)		
Etapas de desarrollo	**Baja**	**Alta**	**Óptima**
Germinación	10	45	20-35
Emergencia y establecimiento de plántulas	12-13	35	25-30
Enraizamiento	16	35	25-28
Elongación de hojas	7-12	45	31
Macollamiento	9-16	33	25-31
Iniciación de panícula	15		
Diferenciación de panícula	15-20	38	
Antesis (floración)	22	35	30-33
Maduración	12-18	30	20-25

Fuente: Arroz: investigación y producción: referencia de los cursos de capacitación sobre arroz dictados por el Centro Internacional de Agricultura Tropical, 1985
Realizado por: Condo León José Luis, 2019

1.2.4.2. Radiación solar

El sol genera una longitud de onda entre 0.3 y 3.0 micrones, mientras que la tierra genera una onda entre 3 y 50 micrones, la planta de arroz usa energía solar en una onda de 0.4 a 0.7 micrones para la fotosíntesis, esta energía solar es denominada Radiación fotosintética activa y la unidad que la representa más usada es Cal/cm2/día.

El cultivo de arroz tiene varias etapas, donde cada una de ellas necesita de diferente cantidad de radiación solar para generar una planta productiva y sana, como ejemplo de esto es posible notar que poca radiación solar en la fase vegetativa afecta muy poco en el rendimiento de la planta, mientras que en la fase reproductiva hace que la producción de la planta disminuya, en la etapa de maduración del grano provoca que no se genere suficiente producción en la planta, las etapas de crecimiento de la planta de arroz y la necesidad de radiación solar en sus diferentes niveles se plasma en el Gráfico 2-1 que se presenta a continuación (Tascón & García , 1985, pág. 27).

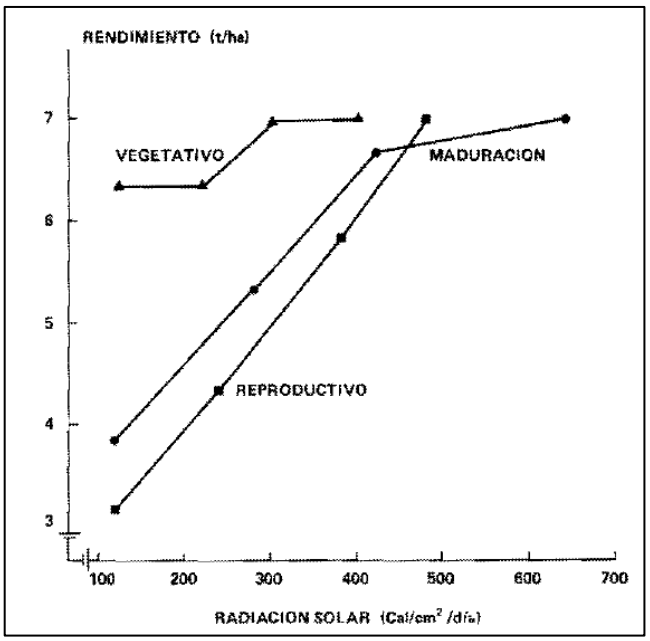


Gráfico 2-1. Efecto de la radiación solar en la planta de arroz
Realizado por: (Tascón & García , 1985)

1.2.4.3. Agua

La principal función del agua en la planta de arroz es modificar las características físicas de esta, convirtiéndolo en un factor indispensable en la supervivencia del cultivo, tan solo el 15% del agua es absorbida por la planta, lo restante es utilizada para la transpiración y fotosíntesis, debido a que en las primeras etapas de crecimiento las yemas que se responsabilizan de las hojas, macollas y penícula se encuentran debajo del agua su desarrollo está influenciado por la temperatura de esta,

en las siguientes etapas de la planta cuando las yemas se encuentran fuera del agua el responsable directo de su desarrollo es la temperatura del aire, para contrarrestar esta dependencia que tiene la planta respecto a la temperatura del aire se plantea elevar los niveles de agua para que las células de crecimiento se encuentren en el agua y no en el aire, cabe recalcar que esto puede realizarse en ciertas etapas de crecimiento de la planta. El requerimiento de agua por parte de la planta de arroz en sus diferentes etapas se detalla en la Tabla 2-1 (Tascón & García , 1985, págs. 30-31).

Tabla 2-1: Requerimientos de agua en arroz irrigado

Pérdida de agua	**mm/día**
Transpiración	1.5 - 9.8
Evaporación	1.0 - 6.2
Percolación	0.2 - 15.6
Total, pérdida diaria	5.6 - 20.4
Operación de campo	
Semilleros	40 mm
Preparación de tierras	200 mm
Irrigación de campo	1000 mm
Total	**1240 mm/cosecha**

Fuente: Arroz: investigación y producción: referencia de los cursos de capacitación sobre arroz dictados por el Centro Internacional de Agricultura Tropical, 1985
Realizado por: Condo León José Luis, 2019

1.2.4.4. Viento

Este es un factor importante ya que se ha demostrado que el viento con una velocidad lenta aumenta la productividad pues existe un mayor suministro de gas carbónico (CO_2), contrario a esto las velocidades inferiores a 0.9 metros por segundo no afecta a la planta, los vendavales o vientos secos son perjudiciales pues provoca laceraciones o resequedad (Tascón & García , 1985, pág. 32).

1.2.4.5. Suelo

Diferentes estudios demuestran que el tipo de suelo no es importante en el cultivo de la planta mientras exista niveles correctos de agua, sin embargo es posible reconocer características del suelo que permitan un desarrollo idóneo, la textura que del suelo puede ir desde arena hasta arcilla, el pH puede variar sin problemas entre 3 a 10, el contenido de materia orgánica que puede tener el suelo puede ir desde el 1 al 50%, la concentración de sal puede estar entre el 0 a 1%, y los nutrientes que este debe contener va desde la deficiencia hasta el exceso (Tascón & García , 1985, pág. 33).

1.2.5. Producción mundial

Este cultivo es uno de los más importantes a nivel internacional, debido en gran parte a su aporte nutricional, es parte incondicional de la dieta de muchas culturas, según la Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO) la producción de arroz cada año rompe récord, en el 2017 alcanzó la cifra de 759,6 millones de toneladas, la producción de arroz de los países asiáticos se vio afectado por los cambios climáticos, mientras que para los países de América latina fue mucho más favorable, por otro lado Australia tuvo buenos réditos en la producción de esta gramínea, la mayor novedad surgió en los Estados Unidos pues tuvo la menor producción de arroz en los últimos 21 años. Las predicciones de la FAO en materia de producción de arroz a nivel mundial son alarmantes, en particular para América Latina cuya producción se verá mermada en los próximos años a causas del clima (Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO) , 2018).

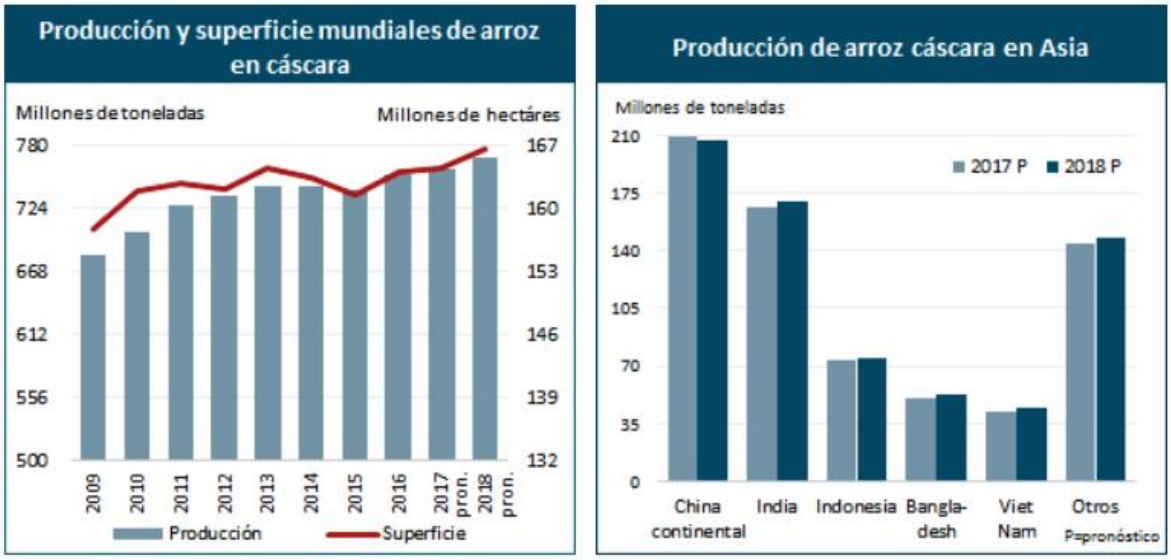


Gráfico 3-1. Producción de arroz a nivel mundial, principales productores asiáticos (2017)
Realizado por: FAO, 2018

Como es posible notar en el Gráfico 3-1 la producción de arroz se ha incrementado desde el año 2009, teniendo tan solo una baja en su producción en el año 2015, los mayores productores de este cultivo a nivel asiático son China, India e Indonesia (Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO) , 2018)

1.2.6. Producción en el Ecuador

El arroz es el cultivo más extenso del Ecuador, ocupa más de la tercera parte de la superficie de productos transitorios del país, en términos sociales es la producción alimenticia más importante del país, a nivel nutricional es la que mayor aporte de calorías brinda a quienes lo consumen según

la (FAO).

Según un pronóstico la FAO, para el 2017 se prevé un crecimiento en la producción internacional de arroz de 0.9% anual. Esto está influenciado principalmente por el continente asiático, donde se pronostica que se registre la mayor parte del aumento de la producción mundial. En cuanto a América Latina y el Caribe, las limitaciones impuestas por los elevados costos de producción y los precios poco atractivos, han impedido la recuperación de considerables superficies de cosecha (Castro, 2017, págs. 1-8).

Los sistemas de manejo de la producción arrocera dependen de la estación climática, zona de cultivo, disponibilidad de infraestructura de riego, ciclo vegetativo, tipo y clase de suelo niveles de explotación y grados de tecnificación.

De acuerdo con los datos del Ministerio de Agricultura, Ganadería, Acuicultura y Pesca del Ecuador (MAGAP) en el año 2009 la superficie sembrada de arroz era de aproximadamente 371 mil hectáreas. (Delgado , 2011, págs. 7-10).

En Ecuador, el rendimiento nacional de producción arrocera para el primer ciclo del 2016 fue de 4.16 t/ha. La provincia de mayor rendimiento fue Loja con 8.46 t/ha; mientras que la de menor rendimiento fue Los Ríos con 3.46 t/ha.

Los problemas fitosanitarios como el vaenamiento y manchado de grano fueron las principales causas que impactaron en la productividad. El rendimiento promedio de los productores que participaron en el plan semilla fue 0.64 t/ha superior al rendimiento de los productores que no participaron. La propagación del cultivo en su mayoría se realizó por medio de semilla, cuya principal variedad fue INIAP 2014.

La superficie promedio sembrada por agricultor fue de 4.28 hectáreas, el rendimiento promedio nacional de arroz fue de 3.92 t/ha. La provincia de Loja registró el mayor rendimiento, siendo 9.54 t/ha; mientras que Los Ríos presentó el rendimiento más bajo con 3.05 t/ha. Comparando con el mismo ciclo del año 2016, se evidencia una reducción en la producción nacional de 6%. Entre las principales variables que influenciaron en la reducción del rendimiento se encuentran los factores climáticos, especialmente el incremento del volumen en las precipitaciones que soportó el litoral ecuatoriano, lo que impactó en la severidad e incidencia de enfermedades como el manchado y vaenamiento de grano.

A continuación, en el Gráfico 4-1 se presenta la producción nacional de arroz medido en toneladas en el periodo 2002-2017, donde es posible observar que siempre existen variaciones en los niveles de este, pero en el año 2017 se presentó la cifra más pequeña de producción en los últimos 16 años con tan solo 1’066.614 toneladas, por lo que es posible asumir una disminución progresiva

de producción nacional en este producto en los últimos 3 años, según cifras del Sistema de Información Pública Agropecuaria (SIPA)

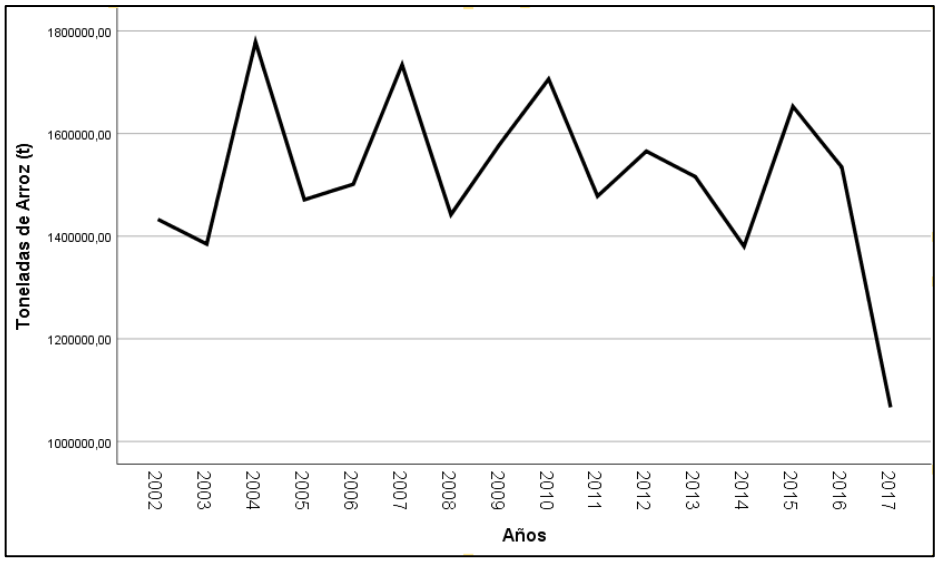


Gráfico 4-1. Producción de arroz en el Ecuador 2002 - 2017
Realizado por: Condo León José Luis, 2019

1.3. Teoría Estadística

Debido al desarrollo exponencial que ha tenido el área de la informática, ciencias como la matemática y la estadística tienen un crecimiento elevado en sus diferentes teorías, es posible notar que ahora el estudio de estas ciencias comienza a partir de muy temprana edad llegando hasta puntos muy elevados de la investigación científica; la importante utilidad de los métodos que posee esta ciencia la han convertido en materia obligatoria en licenciaturas como Medicina, Economía e Ingenierías.

En la actualidad concretamente en la estadística es posible obtener desarrollos y avances científicos importantes, paquetes accesibles y herramientas que faciliten realizar cálculos que hechos de forma manual serían casi imposibles lograr; esta puede tener varias etapas al realizar investigación científica entre las que resalta las siguientes:

1.3.1. Análisis exploratorio de datos (AED)

El AED es obligatorio para conocer y visualizar la estructura espacial a nivel univariante y multivariante de la información, siendo un grupo de herramientas estadísticas que recopilan, analizan e interpretan a los datos, extrayendo toda la información posible de estos, aportando a generar nuevas hipótesis sobre ellos (Batanero, Godino , & Estepa , 1991, págs. 25-31); además permiten conocer el ajuste de distribución, la presencia de datos faltantes y atípicos de las variables utilizadas (Universidad Autónoma de Madrid, 2003, págs. 3-6); a la misma vez que ayudan a determinar

las posibles relaciones existentes entre las variables y los individuos de estudio (Salvador Figueras & Gargallo , 2018)

Una de las características importantes del AED es que trata de extraer toda información posible de un conjunto de datos y a su vez permite generar conjeturas sobre ellos, también recaba nuevos conocimientos sobre el fenómeno estudiado; otra característica relevante es que marca la pauta o el camino del cómo proceder en las siguientes etapas de análisis pues aclara el panorama general del fenómeno y la estructura que posee los datos.

Básicamente las técnicas empleadas están regidas según un parámetro y este es el tipo de variable con la que se va a trabajar en la investigación, pueden ser variables cualitativas y cuantitativas; las variables cualitativas generalmente contienen información relacionada a cualidades de las unidades de investigación y pueden dividirse en nominales y ordinales, por otro lado las variables cuantitativas tienen la características de poseer números que son mediciones de alguna característica de las unidades de investigación y pueden dividirse en discretas y continuas. Entre los métodos gráficos más importantes para representar variables de tipo cualitativas es posible encontrar la tabla de frecuencias, gráfico de barras, gráfico de pastel; algunos de los métodos gráficos para representar información de variables numéricas pueden ser el gráfico de tallos y hojas, histogramas, diagramas de caja entre otros.

Particularmente para las variables de tipo numérica es posible obtener mucha más información con el cálculo de indicadores, los cuales pertenecen a dos grandes grupos que son los indicadores de posición y los de dispersión. Las medidas de posición pretenden describir un valor aproximado alrededor del cual se encuentran las observaciones de la variable, entre los distintos indicadores de este grupo de medidas es posible mencionar a la media muestral, media de datos agrupados, mediana entre muchas más. Por otro lado, las medidas de dispersión permiten al investigador medir de cierta manera la variabilidad existente en un conjunto de datos y sus principales indicadores son el rango muestral, varianza muestral, distancia intercuartil, desviación estándar.

Dentro de este análisis lo más esencial es la comprobación de supuestos sobre los parámetros y constante búsqueda de ajuste de las variables estadísticas de una distribución normal.

1.3.2. Imputación de información faltante

Otro paso en el análisis inicial de los datos es la imputación de datos faltante, la falta de información o datos faltantes en las investigaciones es algo común para un investigador y esto se debe en su gran mayoría a la falta de respuesta que tiene una pregunta realizada a la unidad de investigación o a la no posible medición de cierta característica en el individuo de estudio, es por

esto que tener una base de información completa es ideal pues la imputación o relleno de información faltante es innecesaria y el error en el análisis se reduce significativamente.

Tal como lo resalta (Useche & Mesa, 2006, págs. 127-152) unas de las consecuencias causadas por los datos faltantes pueden ser resultados incompletos o inválidos, distorsión de las frecuencias marginales o conjuntas de las variables, sesgos en los estimadores, reducción del tamaño de la muestra, incremento del error de muestreo, representación de estratos mínima, disminución de relación entre variables y estimadores muy difíciles de obtener. Por tal motivo la imputación de datos faltantes se hace necesario, esto en razón a que es permanente la presencia de este problema, se realiza este procedimiento en función del tipo de variables que se usa en la investigación (Medina & Galván , 2007, págs. 19-20). Este problema es tan común que incluso en ambientes controlados como en diseño de experimentos se produce el fenómeno de los valores perdidos y esto puede tener consecuencias en algunos casos como perder la validez del proceso (Franco & Melo , 2006, págs. 35-56).

Las ventajas y desventajas de la imputación de información pueden ser relativas, puesto que la aplicación correcta de estos métodos ayudaría a disminuir el sesgo si es que existiere, pero el mal uso de estas puede llegar afectar las distribuciones de las variables generando sesgo y poca confiabilidad en los resultados obtenidos (Useche & Mesa, 2006, pág. 140). No existe método de imputación correcto o estándar, todo dependerá de la naturaleza de la investigación y el criterio del investigador según su conocimiento del estudio. Existen infinidad de métodos de imputación, pero todas están marcadas por el número de variables y caracterización del problema, según varios autores este conjunto de métodos puede clasificarse como:

Imputación de datos simples en el cuál la obtención de la estimación es llevada a cabo con el uso de la información de la misma variable o de las variables correlacionadas (Otero García B. , 2011, pág. 20). Entre los métodos representativos de este tipo de imputación se tiene la Imputación con media que consiste principalmente en la sustitución de los valores perdidos con el uso de la media aritmética; Imputación con media condicionada para datos agrupados que se caracteriza por crear agrupaciones de individuos según las covariables relacionadas con la variable imputada; Imputación con variables ficticias, se trata de crear una variable indicadora y así determinar las observaciones con información faltante (Medina & Galván , 2007, pág. 27);Imputación usando una distribución desconocida, procede imputando la información faltante de receptores con el uso de información completa de otros individuos llamados donantes; Imputación con regresión, estima valores faltantes de la variable Y a partir de las variables $(X_1, X_2, \dots, X_p)$ correlacionadas; Imputación por máxima verosimilitud(MV): estima un parámetro con los datos completos, sustituye valores faltantes usando el parámetro y repite el algoritmo hasta que exista convergencia y los parámetros no varíe entre dos repeticiones sucesivas del algoritmo (Medina & Galván , 2007, pág.

29).

El otro tipo de imputación llamado Imputación de datos múltiples se caracteriza por asignar para cada dato faltante un número finito de posibles valores imputables, con el fin de combinar este número finito de posibilidades y dar un solo valor que llenara el dato faltante (Otero García B. , 2011, pág. 20). La utilización de este métodos tiene una base teórica y supuesto que necesitan ser cumplidos, entre los que se puede denotar que el patrón de todos los datos faltantes debe ser aleatorios, se requiere una correlación elevada entre las variables imputadas y las variables del modelo predictor, es necesario que el modelo de análisis tenga relación significativa con el modelo que se usó para la imputación (Rubin D. , 1996, págs. 473-489).

Diversos autores en sus estudios plasmaron que existen varias clasificaciones que se les puede dar a los métodos de imputación de información, investigaciones como (Puerta Goicoechea, 2002, págs. 17-22), (Platek, 1986, págs. 48-53), (Useche & Mesa, 2006, págs. 127-152) muestran una visión más detallada de la clasificación que a continuación se muestra.

Técnicas fundamentadas en información externa: este tipo de técnicas se utilizan cuando las variables usadas pertenecen a una encuesta anterior o con diferentes cualidades; entre las técnicas usadas en esta clasificación se tiene le método deductivo y Tablas Look-up; Técnicas determinísticas: está basada en la repetición de la encuesta bajo las mismas condiciones lo que producirá respuestas iguales; los principales métodos de esta clasificación son imputación de la media, imputación de la media por clases, imputación mediante regresión, emparejamiento de media, imputación usando el vecino más cercano, iteración con máxima verosimilitud, redes neuronales, modelos de series de tiempo; Técnicas estocásticas: son técnicas que al repetir el método de imputación produce resultados diferentes y se toma el valor más frecuente o el aproximado para imputar la información faltante; entre las que se tiene a imputación aleatoria de un caso particular, imputación aleatoria de un caso perteneciente a una clase, imputación Hot-Deck, imputación jerárquica Hot-Deck, imputación por regresión aleatoria, imputación por regresión logística.

1.3.2.1. Imputación de variables cuantitativas con Regresión Lineal

La finalidad de este método es imputar información faltante de una variable dependiente (Y) usando modelos de regresión lineal generados a partir de variables independientes $(X_1, X_2, \dots, X_p)$ (Otero García D. , 2011, págs. 21-24). Asumiendo que, Y sigue una distribución Normal, se presentan dos variantes del modelo de regresión:

El primero es la imputación por regresión determinística, cuyo modelo es

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_p X_{ip} \quad (1.1)$$

Donde $(\hat{\beta}_o, \hat{\beta}_1, \dots, \hat{\beta}_p)$ son los coeficientes de la regresión, $(X_{i1}, X_{i2}, \dots, X_{ip})$ son los valores del individuo i en las variables independientes que están condicionadas a tener observaciones no faltantes.

En segundo lugar, está la imputación mediante regresión estocástica que se rige por el modelo

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_p X_{ip} + \varepsilon \quad (2.1)$$

Donde se incorpora el valor del residuo de estimación $\varepsilon \sim N(0, \hat{\sigma}_{1,\dots,p})$

1.3.2.2. Imputación de variables cualitativas con Bosques aleatorios

Este método de imputación trabaja con variables de tipo cualitativas, cuantitativas o de manera conjunta, basa su funcionalidad en la creación de m número de árboles de decisión generados a partir de muestras diferentes de los datos, obtenidos los m arboles se obtiene un promedio de imputación en el caso de las variables cuantitativas, en cambio para las variables cualitativas se imputará con el valor más frecuente en las predicciones. (Planchuelo Gómez, 2017, pág. 27)

El proceso de imputación de este método puede resumirse en los siguientes pasos:

En una primera parte, dadas $(X_1, X_2, \dots, X_p)$ variables con n números de individuos se elige la variable X_s con valores perdidos en los individuos $i_{mis}^{(s)} \epsilon \{1, \dots, n\}$,a continuación se divide la base de datos en 4 partes:

1. Los valores observados de X_s, denotados por $b_{obs}^{(s)}$.
2. Los valores perdidos de X_s, denotados por $y_{mis}^{(s)}$.
3. Las variables diferentes a X_s con observaciones en $i_{obs}^{(s)} \epsilon \{1, \dots, n\} \setminus i_{mis}^{(s)}$, denotados por $x_{obs}^{(s)}$.
4. Las variables diferentes a X_s con observaciones perdidas $i_{mis}^{(s)}$ denotadas por $x_{mis}^{(s)}$.

Los índices $i_{obs}^{(s)}, i_{mis}^{(s)}$ denotan valores observados y valores perdidos, condición que no se cumple completamente para ninguno de los dos grupos de observaciones.

En la segunda parte del algoritmo

1. Se genera una estimación inicial de los valores perdidos usando la media o cualquier otro método de imputación.
2. Se ordena todas las variables $X_s, s = 1, \dots, p$ de acuerdo al número de valores perdidos,

iniciando con la cantidad más baja.

3. Para cada variable X_s se ajusta un modelo de bosques aleatorios con respuesta $y_{obs}^{(s)}$ y predictores $s_{obs}^{(s)}$. Después de esto se predice los valores perdidos $y_{mis}^{(s)}$ por la aplicación de los bosques aleatorios generados en el grupo de observaciones $x_{mis}^{(s)}$.
4. Este procedimiento se repite hasta que se cumple un criterio de detención del algoritmo, el criterio γ se cumple tan pronto la diferencia entre la nueva matriz imputada y la matriz previa a esta aumenta por primera vez con respecto a ambos tipos de variables (continua y categórica), si es el caso y se presentan.

Esta diferencia generada por el conjunto de N variables continuas está definida como:

$$\Delta_N = \frac{\sum_{j=1}^{N}\left(X_{new}^{imp} - X_{old}^{imp}\right)^2}{\sum_{j=1}^{N}\left(X_{new}^{imp}\right)^2} \tag{3.1}$$

Donde $X_{new}^{imp}\ y\ X_{old}^{imp}$ denota la nueva y antigua base de datos imputada.

Por otro lado, la diferencia generada por las F variables categóricas está definida como:

$$\Delta_F = \frac{\sum_{j=1}^{F}\sum_{i=1}^{n} I_{X_{new}^{imp} \neq X_{old}^{imp}}}{\#\,NA} \tag{4.1}$$

Donde $\#NA$ es el número de valores perdidos en las variables categóricas.

La efectividad del método puede ser evaluado con el uso del Error cuadrático medio Normalizado (NRMSE) definido como:

$$NRMSE = \sqrt{\frac{mean((X^{true} - X^{imp})^2)}{var(X^{true})}} \tag{5.1}$$

Donde X^{true} es el conjunto de datos completos, X^{imp} es el conjunto de datos imputados, $var(X^{true})$ es la varianza empírica calculada sobre el conjunto de información inicial que contiene valores faltantes.

Estudios demuestran que el método de imputación por arboles aleatorios es superior en muchos sentidos a otros métodos de imputación (Misztal , 2013, págs. 169-173).

1.3.3. Normalidad

El análisis clásico estadístico basa su uso en varios supuestos que deben cumplir el conjunto de datos, el más importante del análisis multivariante es que las variables medidas en la investigación sean normales multivariadas, el cumplimiento de este supuesto da validez a los resultados

obtenidos, si se cumple la normalidad multivariada implicaría normalidad univariada lo que no sucede, al contrario.

Para tener una idea global de lo que se trata la normalidad de un conjunto de datos es necesario primero definir la normalidad univariante.

1.3.3.1. Distribución normal univariada

Una variable X se dice que es normal si función de densidad es:

$$f(x,\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\frac{1}{2}(x-\mu)^2}{\sigma^2}} \tag{6.1}$$

La variable se distribuye con $N(\mu,\sigma^2)$, dicho de otra manera, con media μ y varianza σ^2 (Cuadras, 2007 , pág. 24)

1.3.3.2. Distribución normal multivariada

La distribución normal multivariada se deriva de la distribución normal univariada, siendo una generalización de ésta.

Según (Cuadras, 2007 , pág. 25) dado un conjunto de variables $X = \left(X_1, X_2, \dots, X_p\right)'$ cumplen que son normales multivariadas si su densidad es:

$$f(x,\mu,\sum) = \frac{|\sum|^{-\frac{1}{2}}}{\left(\sqrt{2\pi}\right)^p} e^{\frac{1}{2}(x-\mu)'\sum^{-1}(x-\mu)} \tag{7.1}$$

Propiedades:

- $E(X) = \mu$
- $E\left((X-\mu)(X-\mu)'\right) = AI_pA' = \sum$
- Cada una de las variables marginales X_i es normal univariante
$$X_i \sim N(\mu_i, \sigma_{ii}), \quad i = 1, \dots, p$$
- Toda combinación lineal de las variables marginales X_i son normales univariantes
$$Z = b_0 + b_1X_1 + \cdots + b_pX_p \sim N(0,1)$$
- La función de densidad conjunta de un grupo de variables $\left(X_1, X_2, \dots, X_p\right)$ con $\sum = diag\ (\sigma_{11}, \sigma_{22}, \dots, \sigma_{pp})$ y estocásticamente independientes es el producto de las funciones de densidad marginales
$$f\left(x_1, \dots, x_p;\ \mu;\ \sum\right) = f(x_1;\ \mu_1; \sigma_{11}) * \dots * f\left(x_p;\ \mu_p; \sigma_{pp}\right)$$
- $U = (x-\mu)\sum^{-1}(x-\mu)' \sim X_p^2,\ U = YY' =$

$\sum_{i=1}^{p} Y_i^2, donde\ Y \sim N(0,1)\ y\ son\ independientes$ (Cuadras, 2007 , pág. 26)

Existen diversas pruebas que pueden contrastar este supuesto, una investigación realizada por (Porras Cerron , 2016, págs. 141-146) comparo diversas técnicas como Mardia, Shapiro-Wilk, HZ y Royston, donde en un ambiente controlado de simulación se creó diversas muestras pseudoaleatorias provenientes de una distribución normal multivariada, generando diferentes escenarios al variar factores como tamaño de muestra, número de variables y variabilidad de los datos, los resultados determinaron que no existen diferencias significativas en la potencia que tiene cada una de estas pruebas, sin embargo es necesario rescatar que el contraste de Shapiro-Wilk tiene una mayor potencia en comparación al resto cuando la muestra y el número de variables son grandes, investigaciones similares de (Pedrosa , Juarros Basterretxea , Robles Fernandez, Basteiro, & Garcia Cueto, 2015, págs. 245-254) determinaron que esta prueba es mucho más sensible en sus resultados en comparación al resto.

1.3.3.3. Prueba de Shapiro-Wilk generalizada

Dado un conjunto de vectores $X_1, \dots, X_p$ idénticamente distribuidos $N_p(\mu, \sum)$ su estandarización es:

$$Z = \frac{(X - \mu)}{\sum^{\frac{1}{2}}} \sim N_p(0, I_d) \tag{8.1}$$

Partiendo de este principio las componentes generadas con la estandarización $Z_1, \dots, Z_p$ están incorrelacionadas, por lo que Z es normal multivariada bajo la condición de que estas componentes sean normales.

La estandarización de cada una de las componentes se define por:

$$Z_i = \frac{\left(X_i - \tilde{X}\right)}{S^{\frac{1}{2}}} \tag{9.1}$$

Donde $\tilde{X}$ es la media muestral y S es la matriz de varianzas y covarianzas muestral.

Por otro lado, el estadístico de Shapiro-Wilk se define como:

$$W = \sum_{i=1}^{\frac{n}{2}} a_{i,n}(Z_{(n-i+1):n} - Z_{i:n}) \tag{10.1}$$

Donde $Z_{1:n} < \cdots < Z_{n:n}$ son los datos estandarizados ordenados, el estadístico tiene como objetivo calcular la distancia existente entre las muestras estandarizadas ordenadas y la mediana, para luego comparar estas distancias con las distancias que habría en una muestra que se ajuste a

una distribución normal.

Teniendo estos conceptos claros se propone el siguiente estadístico de Shapiro-Wilk generalizado

$$W^* = \frac{1}{p}\sum_{i=1}^{p} W_{Z_i} \tag{11.1}$$

donde W_{Z_i} es el estadístico de Shapiro-Wilk analizado en la i-ésima componente estandarizada $(Z_{i1}, \dots, Z_{in})$, con $i = 1, \dots, p$. Con el uso del estadístico W^* se rechaza H_0 en una prueba de tamaño α si $W^* < c_{\propto;n,p}$ donde $c_{\propto;n,p}$ satisface la siguiente ecuación (Porras Cerron , 2016, pág. 144).

$$\propto = P\{W^* < c_{\propto;n,p} \,/\, H_0 \;\; es\; verdadero\} \tag{12.1}$$

1.3.3.4. Contraste de hipótesis

1.- Hipótesis

$\boldsymbol{H_0}$: El conjunto de variables estadísticas se ajusta a una distribución normal multivariante.

$\boldsymbol{H_1}$: El conjunto de variables estadísticas no se ajusta a una distribución normal multivariante.

2.- Nivel de significancia

$$\alpha = 5\%$$

3.- Estadístico

$$W^* = \frac{1}{p}\sum_{i=1}^{p} W_{Z_i} \tag{13.1}$$

4.- Región de rechazo

Es posible utilizar dos criterios para rechazar la hipótesis nula, el primero está determinado por la comparación del estadístico $W^* < c_{\propto;n,p}$, $Rechazar\ H_o$, el otro criterio usa la comparación del valor p, $Si\ p < \propto\ Rechazar\ H_o$.

1.3.4. Modelo de regresión

El modelo de regresión es una función matemática que relaciona la variable dependiente Y con una o más variables independientes $(X_1, X_2, \dots, X_p)$, esta función matemática permite predecir valores de la variable dependiente a partir de una o varias variables independientes, existe dos tipos de regresión importantes regresión lineal simple o regresión lineal múltiple.

La generación de un modelo que cumpla con los supuestos básicos permitirá concluir que las variables explicativas se están relacionando de buena manera con la variable respuesta, lo que implica que se podrá utilizar dichas variables en cualquier análisis multivariado, sabiendo que se obtendrá buenos resultados en predicción y bondades de modelos. Si esto se no cumple la estandarización de las variables cuantitativas resuelve el problema, haciendo que propiedades de distribuciones los supuestos de regresión se cumplan a cabalidad.

1.3.4.1. Regresión lineal simple

Se caracteriza por tener una variable dependiente Y y una variable independiente X, la función que relaciona estas variables es una línea recta (Walpole & Myers, págs. 389-391).

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{14.1}$$

Donde β_0 es la ordenada en el origen o el punto donde la recta generada pasa por el origen, $\beta_1$es la pendiente o inclinación de la recta y ε es el error aleatorio. La estimación de los coeficientes de la regresión se realiza de tal manera que

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{15.1}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \tag{16.1}$$

$$\text{Donde } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

También es necesario conocer los siguientes estimadores que permitirán la evaluación del modelo.

- Suma de cuadrados totales $SCT = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- Suma de cuadrados de la regresión $SCR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- Suma de cuadrados de los residuos $SCRes = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Coeficiente de determinación $r^2 = \frac{SCR}{SCT}$ (Walpole & Myers, 2012, pág. 396)

1.3.4.2. Regresión lineal múltiple

De manera análoga a la regresión simple se caracteriza por tener una variable dependiente Y y p variables independientes $(X_1, X_2, \dots, X_p)$ relacionándose según la siguiente función lineal (Walpole & Myers, 2012, pág. 443).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \quad (17.1)$$

Donde β_0 es el valor de la variable dependiente cuando las $X_1, X_2, \ldots, X_p$ toman valores de 0, β_1 estima el cambio que ha tenido Y por cada variación de la variable X_1, manteniendo constantes los valores de las variables $(X_2, X_3, \ldots, X_p)$ y de manera análoga para los coeficientes restantes de la regresión, ε es el error de estimación causado por variables que no se controlan.

La estimación de los coeficientes de regresión se realiza de tal manera que

$$\hat{\beta} = (X^t X)^{-1} X^t Y \quad (18.1)$$

Y la estimación de la varianza del error es:

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (19.1)$$

- Para la validación del modelo múltiple es necesario conocer los siguientes estimadores: Variabilidad total $VT = \sum_{i=1}^{n} (y_i - \bar{y})^2$
- Variabilidad explicada $VE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
- Variabilidad no explicada $VNE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
- Descomposición de la variabilidad $VT = VE + VNE$
- Coeficiente de determinación $R^2 = \frac{VE}{VT} = 1 - \frac{VNE}{VT}$
- Coeficiente de determinación ajustado $R^2\ ajustado = 1 - \frac{\frac{VNE}{(n-(p+1))}}{\frac{VT}{(n-1)}}$ (Walpole & Myers, 2012, pág. 445)

Tanto para la regresión lineal simple como para la múltiple se generan los residuos o errores, mismo que permiten la comprobación de la idoneidad del modelo generado, estos se calculan como:

$$e_i = y_i - \hat{y}_i \sim N(0, \sigma^2) \quad (20.1)$$

Los residuos son la diferencia existente entre la variable dependiente y su estimación generada por el modelo de regresión, también considerados errores de estimación.

En la evaluación del modelo de regresión múltiple, existen supuestos fundamentales que el modelo debe cumplir para poder ser considerado bueno y un predictor idóneo de una variable dependiente, entre ellos se tiene la independencia de los residuos, homocedasticidad y normalidad de los residuos.

1.3.4.3. Independencia

Básicamente este supuesto comprueba que los residuos no tengan correlación, si esto no se cumple se tiene problemas de auto correlación entre los individuos de un grupo de observaciones. Una manera de conocer si existe independencia de los errores es graficar los residuos en el eje vertical y los valores pronosticados en el eje horizontal, si existe un patrón en los puntos no existe independencia de los residuos, y si estos puntos están ubicados de manera aleatoria dentro del plano existe independencia de los residuos. El test de Durbin-Watson brinda una prueba estadística para verificar la independencia de los residuos (Bouza Herra , 2018, pág. 83).

Contraste de la prueba

1.- Hipótesis

$\boldsymbol{H_0}$: Existe auto correlación de los residuos

$\boldsymbol{H_1}$: No existe auto correlación de los residuos

2.- Estadístico

$$D = \sum_{i=2}^{n} \frac{(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\hat{\varepsilon}^2{}_t} \in [0,4] \qquad (21.1)$$

Donde $\hat{\varepsilon}_t, \hat{\varepsilon}_{t-1}$ son los residuos o errores de estimación en el tiempo $t\ y\ t-1$

3.- Región de rechazo

Tabla 3-1: Regiones de rechazo de la prueba de Durbin-Watson

Estadístico	Decisión
Si D ≅ 2	No existe correlación
Si 1,5<D<2,5	No existe correlación
Si D <1,5	Existe Correlación
En otro caso	Indecisión

Fuente: Modelo de regresión y sus aplicaciones, 2018
Realizado por: Bouza Herrera Carlos, 2018

Usando los valores de la Tabla 3-1 de Durbin-Watson

Tabla 4-1: Regiones de rechazo de la prueba de Durbin-Watson según valores de distribución

Estadístico	Decisión
Si D > d_U	No existe correlación
Si D <d_L	Existe correlación
Si $d_L < D < d_U$	Indecisión

Fuente: Modelo de regresión y sus aplicaciones, 2018
Realizado por: Bouza Herrera Carlos, 2018

1.3.4.4. Homocedasticidad

Este supuesto verifica que la variable dependiente $\boldsymbol{Y}$ tenga variabilidad parecida en todo el conjunto de observaciones de una variable independiente X_{ij}, la variabilidad se evalúa en cada combinación (Y, X_{ij}). El no cumplimiento de este supuesto genera que la estimación por mínimo-cuadrados sea ineficiente. La manera visual de verificar este supuesto es graficar los valores pronosticados versus los residuos absolutos, si en alguna parte de la serie existe una expansión de los puntos se puede deducir que no existe homocedasticidad, si no sucede este fenómeno se puede asumir homocedasticidad. El test de Breusch-Pagan permite el contraste de la homocedasticidad.

Contraste de la prueba

1.- Hipótesis

$\boldsymbol{H_0}$: Existe heterocedasticidad en función de las variables independientes

$\boldsymbol{H_1}$: No existe heterocedasticidad en función de las variables independientes

2.- Nivel de significancia

$$\alpha = 5\%$$

3.- Estadístico

$$BP = R^2 p = \frac{VE}{2} \sim \chi_p^2 \tag{22.1}$$

4.- Región de rechazo

$$Si\ BP > \chi_{p,\propto}^2\ o\ p < \alpha, \text{Rechazo } \boldsymbol{H_0}$$

1.3.4.5. Normalidad de los residuos

Determina si los residuos del modelo siguen una distribución normal, este supuesto es el de mayor importancia pues si no se cumple la estimadores de mínimos-cuadrados no será eficiente y los intervalos de confianza y las pruebas de hipótesis pierden validez. Para poder verificar este supuesto de manera gráfica el QQ-Plot e histograma de los residuos ayuda a determinar normalidad, la prueba de Shapiro-Wilk permite determinar este supuesto de manera analítica.

Contraste de la prueba

1.- Hipótesis

$\boldsymbol{H_0}$: Los residuos se ajustan a una distribución normal

$\boldsymbol{H_1}$: Los residuos no se ajustan a una distribución normal

2.- Nivel de significancia

$$\alpha = 5\%$$

3.- Estadístico

$$W^* = \frac{1}{p}\sum_{i=1}^{p} W_{Z_i} \tag{23.1}$$

4.- Región de rechazo

$$W^* < c_{\propto;n,p}\ ,\quad Rechazar\ H_o$$

1.3.5. Análisis de datos atípicos

El análisis de datos atípicos tiene su razón de ser debido a gran medida a obtener resultados correctos cercanos a la realidad del fenómeno estudiado, la buena calidad de la información es fundamental para una correcta investigación pues puede conducir a conclusiones certeras en el estudio, por otro lado, la mala calidad de esta puede causar problemas como pérdida de tiempo y dinero. Una de las principales causas de la mala calidad de información son los denominados valores atípicos, muchos investigadores han logrado definir este término como:

Un dato atípico es una observación que se desvía mucho de otras observaciones y despierta sospechas de ser generada por un mecanismo diferente" (Hawkins, 1980, pág. 1), también pueden denominarse como aquellas observaciones con cualidades distintas a las demás (Ocaña Peinado, 2018), las consecuencias de este tipo de información es que deteriora la calidad de las decisiones de una investigación (Amón Uribe & Jiménez Ramírez , 2015, págs. 186-190), distorsiona el cálculo de los parámetros de una población o las estimaciones de los estadísticos de una muestra.

El conjunto de datos atípicos, aunque diferente al resto de la muestra pueden ser importantes porque están representando a un estrato importante de individuos de estudio y el hecho de omitirlos generaría pérdida fundamental de información, poseen cualidades tales como el enmascaramiento y el empantanamiento (Trivez , 1994, págs. 26-27).

Se explica el enmascaramiento como el efecto que tienen un dato atípico sobre otro, se debe a que la ocurrencia del uno hace que la ocurrencia del otro pase desapercibida y tan solo después de la eliminación del primero sea posible la detección del segundo; por otra parte, el empantanamiento hace referencia a que una observación es considerada atípica solo si está en presencia de otro dato considerado atípico, con la eliminación de uno el otro deja de poseer la característica de atípico (Muñoz Garcia & Amón Uribe , 2013, pág. 14).

Entre los métodos más utilizados para la detección de datos atípicos multivariado se pueden citar a la distancia de Mahalanobis, Regresión Lineal, Componentes principales, búsqueda de proyecciones, un método adaptable entre muchos más. La técnica más conveniente para utilizar por las características de la muestra es la distancia de Mahalanobis.

1.3.5.1. Distancia de Mahalanobis

Este método estima parámetros de la distribución multivariada de los datos, permitiendo generar un centro de masa y determinando la distancia existente entre cada individuo de la investigación con respecto a este centro (Manoj & Senthamarai Kannan , 2015, pág. 2319), el criterio de decisión es muy simple, toda aquella distancia de Mahalanobis que se encuentre alejado de manera considerable del centro de masa será especificada como un dato atípico, una manera visual de poder determinar los datos atípicos es generar la gráfica de las distancias de Mahalanobis versus los valores teóricos de una distribución Chi-cuadrado (Filzmoser , págs. 2-3).

La definición teórica de este método es:

$$MSD_i = \sqrt{(x_i - \tilde{x})^T - S_n^{-1}(x_i - \tilde{x})} \tag{24.1}$$

Representa la distancia entre el individuo i y el vector de medias de los datos, donde S_n es la matriz de covarianzas muestral y se define como:

$$S_n = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \tilde{x})(x_i - \tilde{x})^T \tag{25.1}$$

Según la teoría, si los datos poseen una distribución normal multivariada, las distancias generadas tienen una distribución chi-cuadrado con p grados de libertad, por lo que es posible concluir que aquellas observaciones con distancias de Mahalanobis grandes son considerados datos atípicos (Muñoz Garcia & Amón Uribe , 2013, págs. 14-16).

Le regla general de detección es que una observación x_i es denominada como un atípico si $MSD_i \geq C_{p,1-\propto} \sim X_p^2$ (Nordhausen & Ruiz Gazen, 2017, pág. 18), por otro lado (Zambrano , 2019, pág. 4) manifiesta que cada MSD_i debe ser comparado con un valor crítico de la tabla de la distribución $F_{(1-\propto,p,n-p-1)}$, donde p es el número de variables, n es el número de observaciones y $\propto = 1 - (1 - 0.05)^p$, si MSD_ise encuentra por encima de éste valor crítico será denominado como un dato atípico.

En este método el efecto de enmascaramiento puede jugar un papel relevante pues generalmente disminuye el valor de la distancia de Mahalanobis de un dato atípico quitándole la denominación que poseía, de manera opuesta el empantanamiento puede producir que las distancias se incrementen en individuos que no tienen características de observaciones denominadas atípicas. Otro problema es que al momento de estimar la matriz de varianzas y covarianzas para después calcular las distancias de Mahalanobis los datos atípicos ya están influenciando en éstas estimación por lo que el problema puede ser solucionados con la estimación robusta de la matriz de varianzas y el vector de medias (Leys , Klein , & Dominicy, 2017, págs. 2-5).

Se debe considerar que si el porcentaje de datos atípicos es superior a $\frac{1}{1+p}$ la mayoría de los estimadores serán erróneos y cualquier proceso estadístico usando los mismo serán incorrectos (Cousineau & Chartier, 2010, págs. 63-65).

1.3.6. Árboles de clasificación y regresión

En los últimos años el número de artículos científicos que usan análisis estadísticos multivariado ha aumentado, demostrando su correcto funcionamiento para afrontar y resolver problemas de manera mucho más práctica y sencilla en comparación con los métodos clásicos, los beneficios de este tipo de análisis van desde la exploración de datos hasta la modelación matemática, cabe resaltar que es importante la correcta elección de la técnica para resolver problemas de manera confiable (Zuluaga Dominguez , 2011, págs. 143-157).

Es así que las ciencias de la computación tienen varios campos donde se desempeñan, una de las ramificaciones científicas de esta es la minería de datos, cuya función más importante es descubrir conocimiento, buscando patrones de información escondidos en grandes cantidades de datos. (Valero Orea , Salvador Vargas , & Garcia Alonso, 2010, págs. 33-39).

La minería de datos está en la capacidad de exponer diferentes técnicas ya sea para clasificación o predicción, con el uso de una tipología de datos cuantitativos, cualitativos o mixtos. Entre las técnicas multivariantes de minería de datos más representativas es posibles destacar los Árboles de decisión, Arboles de regresión, Sistema de Máquina de Vectores, Redes Neuronales, Bosques Aleatorios, Aprendizaje profundo entre otros.

Una de las técnicas más utilizadas en la investigación científica son los Árboles de decisión, debido mayoritariamente a que su aplicación no está condicionada a que los datos cumplan supuestos estadísticos, diversas investigaciones han demostrado que un modelo de regresión lineal es superior a los árboles de regresión, siempre y cuando la regresión se construya de manera adecuada y respetando todos los supuestos, lo que en la mayoría de los casos no sucede, por lo que los árboles de decisión y regresión son una alternativa viable (Díaz Sepúlveda, 2012, págs. 2-7).

Los árboles de decisión es un algoritmo simple que usa el criterio de mayor porcentaje de ganancia de información, de manera recursiva elige los factores que mejor caracteriza los datos (Valero Orea , Salvador Vargas , & Garcia Alonso, 2010, págs. 33-39), permite la segmentación, estratificación, predicción, reducción de datos, filtrado de variables, identificación de interacciones, fusión de categorías y la discretización de variables continuas (Berlanga Silvente , Rubio Hurtado, & Vila Baños , 2013, págs. 65-79).

La presentación de resultados es mediante la generación de la gráfica en forma de árbol que

permite organizar información dadas en varias etapas, cada una de las ramas generadas en este árbol está creada en base a probabilidades de ocurrencia de ciertos patrones que contiene la base de datos. (Lind , Marchal , & Wathen , 2012, págs. 764-765).

1.3.6.1. Ventajas

- La forma gráfica de los resultados agilita la visualización de los resultados obtenidos.
- Ayuda a la rápida toma de decisiones.
- Como toda técnica de minería de datos reduce la influencia innecesaria de variables independientes (Pérez, Tecnicas de segmentación, conceptos, herramientas y aplicaciones, 2011, págs. 65-79).
- No está sujeta al cumplimiento de supuesto que condicionan su uso.

1.3.6.2. Desventajas

- Para que esta técnica genere resultados confiables el número de datos a analizar debe ser considerablemente grande.

1.3.6.3. Elementos de un Árbol

Nodo raíz: este nodo contiene un grupo inicial de sujetos desde donde se desprenderán otros nodos, también llamado muestra de aprendizaje.

Nodos intermedios: también llamados nodos hijos los cuales generan más segmentaciones en la muestra a partir de ellos.

Nodo terminal: es el punto de segmentación del conjunto de datos donde ya no es posible realizar más divisiones a la muestra puesto que los individuos pertenecen a una sola categoría o solo existe un individuo.

Subárbol: es el producto de la poda del árbol de decisión (Márquez Pérez, Useche , Mesa , & Idés Chacón, 2017, págs. 9-40).

1.3.6.4. Algoritmo

Dada la variable dependiente Y y las variables independientes $(X_1, X_2, \dots, X_p)$, se pretende establecer una relación entre la variable dependiente y las independientes, de tal manera que sea posible la inferencia de Y usando las $(X_1, X_2, \dots, X_p)$ variables independientes. La variable Y determina que metodología y algoritmo se usará en la construcción del árbol, puesto que si esta

es de tipo cualitativa se genera un árbol de decisión y si es de tipo cuantitativa se crea un árbol de regresión.

Visto desde el punto de vista matemático, se pretende estimar la probabilidad condicionada de la variable aleatoria Y respecto a $(X_1, X_2, \dots, X_p)$ y se define como:

$$P[Y = y|\, x_1, x_2, \dots, x_p] \tag{26.1}$$

El procedimiento como primer punto crea un nodo raíz, después de manera iterativa genera nodos hasta obtener nodos terminales, el objetivo fundamental de la técnica es particionar los nodos según las variables usadas y finalizar obteniendo nodos terminales homogéneos (Díaz Sepúlveda, 2012, pág. 8), la noción de impureza o nivel de homogeneidad de los individuos del nodo se la puede definir como:

$$Impureza\ de\ un\ nodo = \frac{Número\ de\ individuos\ que\ cumplen\ la\ característica\ en\ el\ nodo}{Número\ total\ de\ individuos\ en\ el\ nodo} \tag{27.1}$$

Por lo que mientras más homogéneo sea el nodo el valor del cociente se aproximara a 0 o a 1

Generados estos indicadores, cada vez que se produzca un nodo hijo se busca la variable y el punto de corte ideal que permitan dos próximos nodos menos impuros, este procedimiento es repetitivo hasta obtener nodos terminales caracterizados por contener a muy pocos individuos (Bacallao Guerra & Bacallao Gallestey , 2010, págs. 133-139)

1.3.6.5. Generación de nodos intermedios

Tras generar el nodo raíz, este se debe dividir en nodos homogéneos, la manera de lograrlo es obtener la fracción de ocurrencia de cada categoría contenidas en las variables independientes, para decidir qué variable está siendo significativa en el nodo analizado solo es necesario deducir la fracción que más cerca este a 0 o a 1. Antes de llegar a tomar una decisión sobre cómo se construirá el nodo es necesario determinar la bondad de división (impureza del nodo).

Para poder tener un ejemplo de la funcionalidad de la bondad de división se toma a X_1 como la variable predictora y a el valor de c como punto de corte o argumento de elección de grupo, de esta manera se genera la Tabla 5-1.

Tabla 5-1: Generación de nodo derecho e izquierdo en los Árboles de decisión

	$Y = 0$	$Y = 1$	
Nodo Izquierdo (τ_L) $x_1 \leq c$	n_{11}	n_{12}	n_{f1}
Nodo Derecho (τ_R) $x_1 > c$	n_{21}	n_{22}	n_{f2}
	n_{c1}	n_{c2}	

Fuente: Comparación entre Árboles de Regresión CART y Regresión Lineal, 2012
Realizado por: Diaz Sepúlveda Juan Felipe, 2012

Dado que la teoría de los árboles de decisión condiciona a que la variable respuesta Y solo puedan tomar valores entre 0 o 1, entonces la $P[Y = 1|\tau_L] = \frac{n_{12}}{n_{f1}}$ y $P[Y = 1|\tau_R] = \frac{n_{22}}{n_{f2}}$.

Es necesario también conocer el concepto de entropía, para el lado izquierdo del nodo que se define como:

$$i(\tau_L) = -\frac{n_{11}}{n_{f1}}\log\left(\frac{n_{11}}{n_{f1}}\right) - \frac{n_{12}}{n_{f1}}\log\left(\frac{n_{12}}{n_{f1}}\right) \tag{28.1}$$

Para el lado derecho del nodo se establece como:

$$i(\tau_R) = -\frac{n_{21}}{n_{f2}}\log\left(\frac{n_{21}}{n_{f2}}\right) - \frac{n_{22}}{n_{f2}}\log\left(\frac{n_{22}}{n_{f2}}\right) \tag{29.1}$$

Finalmente, la bondad de división (s) mide el nivel de impureza que se obtiene cuando se pasa de un nodo padre a los nodos hijos, se define como:

$$\Delta I(s, \tau) = i(\tau) - P[\tau_L]i(\tau_L) - P[\tau_R]i(\tau_R) \tag{30.1}$$

Donde τ es el nodo padre de τ_L y τ_R, $P[\tau_L] = \frac{n_{f1}}{n_{f1}+ n_{f2}}$ y $P[\tau_R] = \frac{n_{f2}}{n_{f1}+ n_{f2}}$ se definen como la probabilidad de que un individuo sea asignado dentro del lado izquierdo (τ_L) o derecho (τ_R) del nodo (Díaz Sepúlveda, 2012, pág. 9).

1.3.6.6. Generación de nodos terminales

El algoritmo usado para la generación del árbol establece que se debe dividir los nodos hasta que ya no sea posible, por lo que se puede definir al nodo terminal como el nodo que no se pueda dividir, generalmente esto sucede cuando solo queda un individuo en el nodo. Cabe tener en cuenta que al realizar todas las divisiones posibles a un árbol este se va a "sobre-ajustar" lo que quiere decir que el modelo va a poder clasificar o inducir valores casi a la perfección, pero solo para la muestra con la fue generado, por otro lado, va a tener grandes errores en la inducción con datos externos a esta. Por este motivo debe introducirse el término de "poda", cuya funcionalidad

es determinar cuáles son los nodos más relevantes sobre la muestra para luego utilizarlo en la predicción (Díaz Sepúlveda, 2012, pág. 10).

1.3.6.7. Poda del Árbol

El término poda en este contexto implica encontrar un subárbol que sea el mejor predictor de los datos, para así no utilizar el árbol sobre ajustado, bajo el criterio de costo-complejidad la función que permite la poda se denota por:

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \propto |T| \tag{31.1}$$

Donde T es el árbol d decisión, $|T|$ es el número de nodos terminales y $\propto$ es la medida del equilibrio entre el árbol más largo (todo aquel nodo que contenga un solo integrante) y el más chico (aquel de un solo nodo) (Díaz Sepúlveda, 2012, pág. 10).

Cabe resaltar que en la creación del modelo de árboles no necesariamente todas las variables independientes serán usadas, por otro lado, la variable dependiente puede ser usada una o varias veces como la variable que separa los nodos.

1.3.6.8. Árboles de clasificación

Los árboles de clasificación se denominan de tal manera debido a que su variable respuesta es de tipo cualitativa.

Impureza del nodo en los Árboles de clasificación

Dada una variable cualitativa Y dicotómica que puede tomar valores de 0 o 1 y la obtención de un árbol de clasificación saturado, donde para el nodo menos impuro la impureza es de 0 y se debe cumplir que $P[Y = 1|\tau] = 0\ o\ P[Y = 1|\tau] = 1$, el nodo más impuro se caracteriza por tener una impureza igual a 1 con $P[Y = 1|\tau] = \frac{1}{2}$ (Serna Pineda, 2009, pág. 11), por lo que se puede definir a la impureza de un nodo como:

$$i(\tau) = \emptyset(\{Y = 1|\tau\}) \tag{32.1}$$

Esta función de impureza cumple con características como:

- $\emptyset \geq 0$
- $\forall\ p\ \in (0,1), \quad \emptyset(p) = \emptyset(1-p) \quad y \quad \emptyset(0) = \emptyset(1) < \emptyset(p)$

Las medidas de impureza más comunes entre los árboles de decisión son:

- El error mínimo o error de Bayes $\emptyset(p) = min(p, 1-p)$
- Índice de Gini $\emptyset(p) = p(1-p)$
- Entropía $\emptyset(p) = -p\log(p) - (1-p)\log(1-p)$

Donde $0\log(0) \coloneqq 0$

Para la comprobación del nivel de impureza en los nodos cuando se trabaja con variables cualitativas utiliza el índice de impureza de Gini y el índice Binario, por otro lado, cuando opera con variables cuantitativas usa la prueba de Homogeneidad de varianzas (Mendoza Vega , 2018, pág. 3).

Generación de nodos terminales en árboles de clasificación

Para lograrlo se debe tener una medida de la calidad de un árbol de clasificación que verifique la homogeneidad de los nodos terminales, por lo que es posible concluir que la calidad de un árbol de clasificación se mide verificando la calidad de los nodos terminales.

Generado un árbol *T*, se define

$$R(T) = \sum_{\tau \in \hat{T}} P[\tau] r(\tau) \qquad (33.1)$$

Donde $\hat{T}$ es el conjunto de nodos terminales de T y $r(\tau)$ es la medida de calidad del nodo τ (Díaz Sepúlveda, 2012, pág. 12)

Poda del árbol de clasificación

La finalidad de la poda es obtener un subárbol T^* de un árbol T de tal medida que $R(T)$ sea el mínimo, un método para obtener la medida de impureza del nodo ($r(\tau)$) es determinar el costo de mala clasificación.

Costo de mala clasificación

Dada la variable dependiente Y que solo puede tomar valores de 0 o 1, se establece a $c(i|j)$ como el costo de mala clasificación de que un individuo de la clase j sea clasificado como si fuera de la clase i, también si $c(i|i) = 0$ el costo de clasificación errónea sería cero (Díaz Sepúlveda, 2012, pág. 12), generalizando esta lógica hacia el nodo se tiene que τ es determinado como un nodo de la clase j si

$$\sum_i \{c(j|i)P[Y=i|\tau]\} \leq \sum_i \{c(1-j|i)P[Y=i|\tau]\} \quad (34.1)$$

Partiendo de la anterior ecuación es posible determinar que el costo de mala clasificación del nodo τ es $R(\tau) = P[\tau]r(\tau)$, por lo que análogamente se puede demostrar que el costo de la mala clasificación del árbol T es:

$$R(T) = \sum_{\tau \epsilon \hat{\tilde{T}}} R(\tau) \quad (35.1)$$

1.3.6.9. Árboles de regresión

La diferencia que más peso tiene con los árboles de clasificación es que la variable respuesta para los árboles de regresión es de tipo cuantitativa, también se debe generar una medida de impureza como propiedad del nodo, un criterio de división de los nodos para generar el árbol en su totalidad y un indicador de costo-complejidad para podar este árbol (Díaz Sepúlveda, 2012, págs. 16-17). El método de generación del árbol es el mismo que cuando se genera un árbol de clasificación, la diferencia radica en el cálculo de los criterios, es así como la impureza de un nodo τ es la varianza de la respuesta dentro del nodo y se define como:

$$i(\tau) = \sum_{sujeto_{i \epsilon \tau}} \left(Y_i - \hat{Y}(\tau)\right)^2 \quad (36.1)$$

Donde $\hat{Y}(\tau)$ se define como el promedio de los $Y_i s$ contenidos dentro del nodo τ, la división de un nodo en dos nodos hijos $\tau_L \; y \; \tau_R \; se$ e realiza mediante la función:

$$\emptyset(s,\tau) = i(\tau) - i(\tau_L) - i(\tau_R) \quad (37.1)$$

Además, el costo del árbol se define bajo la función:

$$R(T) = \sum_{\tau \epsilon \hat{\tilde{T}}} i(\tau) \quad (38.1)$$

1.3.7. Análisis factorial de datos mixtos (AFDM)

El análisis factorial de datos mixtos es un análisis que permite el análisis simultáneo de un grupo o varios grupos de variables que fueron medidas sobre un conjunto de individuos, lo que lo diferencia del análisis factorial común es que integra el análisis de agrupaciones y genera agrupaciones y factores latentes en base a variables con distinta naturaleza de medida. Este análisis en primer lugar realiza un análisis de componentes principales obteniendo los autovalores

de cada variable lo que a su vez permite generar una matriz que corresponde a los coeficientes para el análisis factorial, después se vuelve a realizar el análisis de componentes principales usando la matriz anteriormente calculada y obteniendo los factores latentes o factores de caracterización de los individuos de la muestra. (Gonzáles, 2009, págs. 1-15)

Entre los muchos beneficios que aporta esta técnica podemos encontrar que:

Estudia el parecido entre los individuos de la muestra, permite observar la existencia de relación entre las variables usadas y la cercanía existente entre los diferentes grupos que forman las variables latentes.

Otra de las bondades del AFDM es que ayuda a equilibrar la influencia de los diferentes factores cualitativos en el análisis simultaneo de las variables, generando indicadores o factores latentes. La principal función de este análisis es generar indicadores de las inercias entre los individuos. A parte de poseer una amplia gama de visualización gráfica de resultados. (Abascal Fernandez & Landaluce Calvo, 2002, págs. 109-122)

Esta técnica también permite describir a varios grupos de individuos, no está limitada a obtener solo las tipologías de los individuos si no a medir la relación entre los grupos de variables y los grupos de individuos, conociendo cuales son las variables latentes o influyentes sobre un colectivo. (Fernandez Aguirre, 2013, págs. 305-322).

La definición literaria plantea que "Es un método multivariante que pretende expresar p variables observables como una combinación lineal de m variables hipotéticas o latentes, denominadas factores" según (Cuadras, 2007).

Es una de las herramientas del análisis Multivariante que analiza datos mixtos (Categóricos/cuantitativos) que tienen una aproximación analítica similar al Análisis de Componentes principales y al Análisis de Correspondencias Múltiples. La mayor ventaja del AFDM es que permite estudiar datos respetando su naturaleza, siendo el estudio de la similaridad de los individuos y la relación entre variables su fuerte. (Zubcoff, 2017).

Es posible notar varias diferencias del AFDM con respecto al Análisis de Componentes Principales (ACP), entre las más destacables se muestra que en el ACP la tabla de correlaciones existe, las componentes principales son funcionales, en cambio sí en el AFDM las correlaciones no existieran el modelo factorial debería ser validado mediante algún test estadístico, por otro lado la diferencia más significativa es que el Análisis Factorial se limita a trabajar con variables latentes de tipo cuantitativas y el AFDM funciona con el uso tanto de variables cualitativas y cuantitativas.

1.3.7.1. Análisis factorial univariado

Para poder entender el AFDM primero es necesario conocer cuáles son las teorías fundamentales en las que basa su funcionamiento, de tal manera es necesario definir el modelo unifactorial, éste considera p variables medidas sobre un colectivo de individuos y el uso de un solo factor común F (Cuadras, 2007 , pág. 86), es así que el modelo se define como:

$$X_i = a_i F + d_i U_i \,. \quad i = 1, \dots, p \tag{39.1}$$

Suposiciones del modelo

- Las variables y los factores están estandarizados
- Los p+1 factores están incorrelacionados

Características del modelo unifactorial

- El único factor F contiene la variabilidad explicada de las p variables
- El factor único U_i es la parte de la variabilidad que no está siendo explicada por el factor común
- a_i es la saturación de la variable X_i en el factor F

$$cor(X_i, X_j) = a_i a_j, \quad \text{i} \neq \text{j} \tag{40.1}$$

- a_i es la saturación al mismo tiempo es el coeficiente de correlación entre la variable X_i en el factor F
- La proporción de la variabilidad que explica el único factor y la correlación entre variables depende únicamente de las saturación u correlaciones entre ellas

1.3.7.2. Análisis factorial multivariado

Este modelo considera m factores comunes que son de las que depende las variables observables, contiene a la misma vez p factores únicos que explican la variabilidad que no está siendo explicada por los factores comunes (Cuadras, 2007 , pág. 88), a continuación, se presenta el modelo multifactorial de manera matricial

$$\mathrm{X} = \mathrm{AF} + \mathrm{DU} \tag{41.1}$$

Donde, X es el vector columna de las variables $\mathrm{X} = (X_1, \dots, X_p)'$, A se denomina como la matriz factorial, D es la matriz diagonal que contiene las saturaciones entre las variables y los factores únicos y se define como $D = diag(d_1, \dots, d_p)$, finalmente $U = (U_1, \dots, U_p)$ son los factores únicos.

El objetivo primordial de este análisis es encontrar e interpretar la matriz factorial A, pero el uso de este modelo está sujeto a diversos supuestos, entre los cuales se puede resaltar que:

- Los factores comunes y los factores únicos están incorrelados entre si

$$cor(F_i, F_j) = 0, \quad i \neq j = 1, \dots, m \tag{42.1}$$

$$cor(U_i, U_j) = 0, \quad i \neq j = 1, \dots, p \tag{43.1}$$

- Los factores comunes están incorrelados con los factores únicos

$$cor(F_i, U_j) = 0, \quad i = 1, \dots, m\,, \quad j = 1, \dots, p \tag{44.1}$$

- Los factores comunes y los factores únicos son variables estandarizadas con media 0 y varianza 1

Los coeficientes a_{ij} son las saturaciones entre cada variable X_i y el Factor F_j, de tal manera que la matriz factorial contiene los coeficientes de saturaciones y se define como:

$$A = \begin{pmatrix} a_{11} & \dots & a_{1m} \\ a_{21} & \dots & a_{2m} \\ \vdots & \dots & \vdots \\ a_{p1} & \dots & a_{pm} \end{pmatrix}$$

Por otro lado, un indicador importante del análisis es el cálculo de la comunalidad que es la parte de la variabilidad de las variables explicada por los factores comunes definida como $h_i^2 = a_{i1}^2 + \cdots + a_{im}^2$, la cual de manera análoga se puede explicar diciendo que la $Variabilidad = comunalidad + unicidad$, suponiendo variables observables estandarizadas tenemos que $1 = h_i^2 + d_i^2$.

Para complementar el análisis, y obtener la matriz de correlaciones reducidas es necesario sustituir los unos de la diagonal de la matriz R por las comunalidades, es así que:

$$R^* = \begin{pmatrix} h_1^2 & r_{12} & \cdots & r_{1p} \\ r_{21} & h_2^2 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & h_p^2 \end{pmatrix}$$

De esta manera se verifica que la matriz de correlaciones se construye como:

$$R = R^* + D^2 \tag{45.1}$$

1.3.7.3. Generación del método de AFDM

Determinada ya las bases conceptuales del análisis Factorial univariado y multivariado es necesario definir el significado, funcionalidad y propiedades del AFDM; muchos autores lo definen como una técnica que trabaja simultáneamente con variables cuantitativas y cualitativas, conocido también como datos mixtos (Pagés, 2015, pág. 67), la metodología actual de cómo resolver el análisis de datos mixtos se basa en la transformación de variables cuantitativas a cualitativas, quebrando la variación de los intervalos dentro de las clases y resolviendo este problema con Análisis de correspondencias múltiple (ACM), este procedimiento es muy fácil de realizar pero los resultados no son estables, debido a la gran pérdida de información que la transformación de las variables conlleva.

Con el fin de resolver dicho problema Brigitte Escofier sugiere la introducción de variables de tipo cuantitativo en el ACM en el año 1979, y en 1990 Gilbert Saporta sugiere la introducción de variables de tipo cualitativo en el Análisis de Componentes Principales (ACP), las ideas planteadas por los dos autores presentaban los mismos resultados y es así como nace el AFDM, generando buenos resultados y mucho potencial en resolución de problemas de este tipo en comparación con otras técnicas.

Para poder tener mejor comprensión del tema es necesario denotar varios elementos y características importantes como lo son:

Dentro de la muestra se tiene I individuos, cada uno de estos tiene un peso igual a p_i, por lo que todos estos tienen el mismo peso dentro de la muestra a menos que sea especificado lo contrario, de tal manera se verifica que $p_i = \frac{1}{I}\ \forall\ i$, y se comprueba que $\sum_i p_i = 1$.

Como requisito para el uso de esta técnica todas las K_1variables cuantitativas $\{k = 1, K_1\}$ deben estar estandarizadas y reducidas, por otro lado, dentro de las Q variables cualitativas $\{q = 1, Q\}$ debe existir K_q categorías $\{k_q = 1, K_q\}$, lo que determina que las suma de las variables cuantitativas y cualitativas se defina como $K = K_1 + K_2$.

1.3.7.4. Representación de variables

Se establece que R^I será el espacio generado en función de I, este espacio es la matriz diagonal que contienen todos los pesos de los individuos, denotado como D:

$$D(i,j) = \begin{cases} 0 & if\ j \neq i \\ p_1 & if\ j = i \end{cases} \qquad (46.1)$$

Por lo general los individuos tienen los mismos pesos, su construcción matricial esta denotada

por:

$$D = \left(\frac{1}{I}\right) I_d \tag{47.1}$$

Donde, I es el número de individuos y I_d es la matriz identidad.

Al igual que en el ACP las variables cuantitativas son representadas por un vector de longitud 1, y análogamente al método de ACM las q variables cualitativas están representadas por una nube N_q de sus indicadores centrados del conjunto de variables K_q. La nube genera un subespacio E_q de dimensión $K_q - 1$, donde E_q es el conjunto de funciones centradas constantes en las clases de la partición definidas por q.

Para que N_q posea las mismas propiedades de inercia que en el ACM, se debe estandarizar como en el ACP, lo que provoca que el indicador k_q deba ser dividido a p_{k_q} y a cada una de estas divisiones atribuirles un peso.

Para la obtención de la inercia propia de ACM es necesario que los pesos p_{k_q}sean divididos para J promedios de acuerdo con el número de variables usadas, lo que es indeseable pues cuando las variables cualitativas se enfrentan con las variables cuantitativas las inercias generadas no se promedian.

Por lo tanto, al proceder en esta dirección, se usa una propiedad fundamental del ACM, donde para obtener la inercia proyectada de $N_q$en una variable centrada y es necesario calcular la raíz cuadrada del radio de correlación entre q y y, denotado como $\eta^2(q, y)$. Cuando se busca la dirección de v del espacio R^I que maximiza la inercia proyectada de la nube N_K es posible establecer el criterio:

$$\sum_{k \epsilon K_1} r^2(k, v) + \sum_{q \epsilon Q} \eta^2(q, v) \tag{48.1}$$

Es así como se establece el criterio expresado por Gilbert Saporta, geométricamente como las k variables se encuentran estandarizadas, las coordenadas de proyección de la variable k en v es igual a $cos(\theta_{kv}) = r(k, v)$, donde θ_{kv} es el ángulo entre los vectores k y v, de manera similar como v esta centrada $\eta^2(q, y) = \cos^2\left(\theta_{qv}\right)$, donde θ_{qv} es el ángulo entre v y está proyectada en E_q, este criterio esta expresado como:

$$\sum_{k \epsilon K_1} cos^2(\theta_{kv}) + \sum_{q \epsilon Q} cos^2(\theta_{qv}) \tag{49.1}$$

1.3.7.5. Representación de individuos

Dado el espacio R^Kcompuesto por K_1variables cuantitativas y K_2 variables cualitativas, los valores euclidianos de la diagonal son iguales a los pesos, 1 para las variables cuantitativas y p_{k_q}para las categorías de cada variable cualitativa, por lo que se establece la distancia entre el individuo i y l como:

$$d^2(i,l) = \sum_{k \epsilon K_1} (x_{ik} - x_{lk})^2 + \sum_{q \epsilon Q} \sum_{k \epsilon K_q} p_{k_q} \left(\frac{y_{ik_q}}{p_{k_q}} - \frac{y_{lk_q}}{p_{k_q}} \right)^2 \quad (50.1)$$

Las variables de tipo cuantitativo contribuyen a esta distancia de la misma manera que lo hacen en ACP y las variables de tipo cualitativo lo hacen igual que en el ACM, un caso importante de las distancias es la que se genera entre un individuo y el centro de masa o el origen O, claramente se asume un centrado de las variables cuantitativas y se denota como:

$$d^2(i,O) = \sum_{k \epsilon K_1} x_{ik}^2 + \sum_{q \epsilon Q} \sum_{k_q \epsilon K_q} p_{k_q} \left(\frac{y_{ik_q}}{p_{k_q}} - 1 \right)^2 = \sum_{k \epsilon K_1} x_{ik}^2 + \sum_{q \epsilon Q} \frac{1 - p_{q(i)}}{p_{q(i)}} \quad (51.1)$$

Donde $q(i)$ son las i categorías de la variable q y $p_{q(i)}$es la proporción asociadas a cada $q(i)$.

De esta manera se asegura el balance de la contribución que tienen los dos tipos de variables en el análisis, y es posible medir la influencia de cada variable por su contribución de inercia a todos los individuos.

Todas las consideraciones establecidas en R^I se transponen en R^K, particularmente en el subespacio R^K generado por K_q categorías de la variable q la proyección de la nube de individuos tiene una inercia de $K_q - 1$ distribuida isotrópicamente en todas las direcciones del subespacio de dimensión $K_q - 1$.

Se denotará a $F_s(i)$ como la proyección del individuo i en el eje del rango s que en conjunto forman la proyección de los individuos en el eje de la inercia; F_s es el coeficiente de correlación entre las variables cuantitativas con los factores; y $F_s(k_q)$ es la coordenada de proyección en el eje de los rangos del centro de gravedad de los individuos que poseen la categoría k de la variable q.

1.3.7.6. Relaciones de transición de $\boldsymbol{R^K}$ hacia $\boldsymbol{R^I}$

Se usarán las fórmulas de ACP para determinar la relación existente entre el espacio R^K hacia el espacio R^I, de esta manera se define a $G_s(k)$ como la coordenada de la columna k en el eje del

rango s y se denota como:

En el caso de una variable cuantitativa:

$$G_s(k) = \frac{1}{\sqrt{\lambda_s}} \sum_i p_i x_{ik} F_s(i) = r(k, F_s) \tag{52.1}$$

En el caso de una variable cualitativa, dado k_q categorías de una variable q con frecuencia relativa de p_{k_q}:

$$G_s(k) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{\sqrt{p_{k_q}}} \sum_i p_i y_{ikq} F_s(i) = \frac{1}{\sqrt{\lambda_s}} F_s(k_q) \tag{53.1}$$

Donde $F_s(k_q)$ es la coordenada a lo largo del eje del rango s, del centro de gravedad del individuo que posee la categoría (k_q); al igual que en el ACM, $\frac{1}{\sqrt{\lambda_s}}$ es el coeficiente de la coordenada de una categoría como un indicador (esto dentro R^I), dicho de otra manera, es el baricentro de los individuos que poseen esto (dentro R^K).

1.3.7.7. Relaciones de transición de $\boldsymbol{R^I}$ hacia $\boldsymbol{R^K}$

Esta relación es indispensable en el ACM para conocer la posición de un individuo de acuerdo con las categorías que posee, esto raramente se explica en el ACP, en el AFDM éste se expresa como:

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{k \epsilon K_1} x_{ik} G_s(k) + \frac{1}{\sqrt{\lambda_s}} \sum_{k_q \epsilon K_2} p_{k_q} \left(\frac{y_{ik_q}}{p_{k_q}} - 1 \right) G_s(k_q) \tag{54.1}$$

El primer miembro es del ACP, éste expresa que un individuo está del lado de las variables para los cuales tiene un promedio superior, y es opuesto para las variables para las cuales tiene un valor por debajo del promedio. El segundo miembro es del ACM, sobre los $\frac{1}{Q}$ coeficientes, este puede ser expresado de acuerdo con $F_s(k_q)$, gracias a la ecuación anterior relacionando $G_s(k_q)$ con $F_s(k_q)$:

$$\frac{1}{\sqrt{\lambda_s}} \sum_{k_q \epsilon K_2} \left(y_{ik_q} - p_{k_q} \right) F_s(k_q) = \frac{1}{\sqrt{\lambda_s}} \sum_{k_q \epsilon K_2} \left(y_{ik_q} F_s(k_q) \right) \tag{55.1}$$

La ecuación anterior expresa que un individuo esta sobre el coeficiente λ_s , en el baricentro de las categorías que éste posee, cabe recalcar que en la relación de transiciones se expresa las coordenadas de un individuo acordando que estas son las categorías.

El coeficiente es $\sqrt{\lambda_s}$ si las categorías están representadas por la proyección de los indicadores

(dentro de R^I), es λ_s si las categorías están representadas por los centros de gravedad de los individuos que poseen las mismas categorías (dentro de R^K).

Se puede determinar que un individuo se encuentra del lado de las variables cuantitativa para las que tienen un valor y del lado de las categorías que posee.

1.3.7.8. Implementación

La solución más apropiada para la implementación de este método es con el uso del paquete estadístico FactorMineR contenido en el software libre R, para poder aplicarlo de manera correcta las variables cuantitativas deben estar centradas y reducidas. Otra manera de hacerlo es con el uso de softwares que permita el ACP, para obtener una representación de las categorías se necesita los centros de gravedad de las variables cualitativas y deben ser introducidas como suplementarias (Pagés, 2015, pág. 72).

CAPÍTULO II

2. MARCO METODOLÓGICO

2.1. Hipótesis de investigación

"Existen factores que influyen en la producción arroz según la base de datos ESPAC 2017"

2.2. Identificación de variables

Para realizar la investigación se usó la base de datos ESPAC 2017, específicamente la sección de cultivos transitorios y dentro de ésta a todos los individuos productores de arroz que participaron en la encuesta, se tomaron en cuenta solo las variables con información completa en los individuos, lo que generó que un total de 15 variables cualitativas y 3 variables cuantitativas.

2.3. Población y muestra

Desde el año 2002 el INEC emplea la metodología de muestreo de marcos múltiples (MMM) para realizar la ESPAC (INEC, 2016, pág. 9), el (MMM) es una combinación entre muestreo de marco de áreas (MMA) y muestreo con el marco de lista (MML); el marco de áreas (MA) realiza la segmentación de la superficie total del estado por estratos en base a la intensidad productiva de cada área del país; por otro lado, el marco de lista (ML) sirve para constatar las principales explotaciones dedicadas a determinados cultivos, los cuales son tomados en cuentan para de tal forma mejorar la calidad de estimaciones estadísticas (Guamán Daquilema & Mullo Guaminga , 2018, pág. 44) .

Por tal motivo es posible delimitar que la población de investigación de la ESPAC 2017 son todos los terrenos con explotación agropecuaria dentro de la superficie continental ecuatoriana (INEC, 2016, pág. 31); y la muestra son todos los terrenos que participaron en dicha encuesta, siendo un total de 22698 cultivos, que comprende todo tipo de sembríos. Esta investigación secciona la información tomando como muestra a todos los cultivos que produjeron arroz en el año 2017 alcanzando 2805 casos de investigación.

El 68.49% de los cultivos que participaron en la encuesta pertenecen a la provincia del Guayas,

el 26.17% a Los Ríos, el 2.94% a Manabí, 1.44% a El Oro, el 0.38% a Sucumbíos y el 0.58% a otras provincias, determinando como puntos principales de recolección a Guayas y Los Ríos (INEC, 2017, pág. 16).

2.4. Tipo de investigación

Según la literatura mostrada en (Hernández Sampieri & Baptista Lucio, 2014, págs. 88-100) existen diversas características que una investigación posee, entre las más relevantes se puede citar que este estudio es de tipo retrospectivo, debido a que los datos utilizados son históricos recogidos en el año 2017; en función de la naturaleza de sus fuentes la mayor parte del estudio fue basado en fuentes bibliográficas; y está enfocando el análisis de un grupo de individuos; también se define a esta investigación de tipo no experimental, ya que se trató de estudiar un fenómeno natural donde no se tuvo control en absoluto de las variables medidas.

Otra característica del estudio es que tiene un alcance exploratorio relacional, en razón a que pretendió detectar información que a simple vista no es visible mediante la categorización, caracterización y relacionamiento de los individuos y variables; también generará nuevo conocimiento en el campo de la Estadística.

2.5. Técnicas y métodos

La investigación aquí presentada tiene carácter exploratorio de información, cuyo objetivo principal fue la búsqueda de factores que influyeron en la producción de arroz; con el fin de concluir con este objetivo y en base a los antecedentes de este estudio, se utilizó los datos de la ESPAC 2017, dentro de ésta la sección de cultivos transitorios; seleccionando a todos los terrenos cuyo cultivo sea el arroz.

El análisis en su totalidad se realizó con el uso del software libre R, como primer paso en la investigación se generó el AED que expuso las primeras características de la base de datos, para esto se usó código base del software; a continuación, se procedió con la imputación de información faltante, se usó Regresión Lineal para imputar las variables de tipo cuantitativas y la librería missForest para imputar variables cualitativas; el uso de Regresión Lineal está sujeto a varios supuestos que deben ser comprobados, para esto se utilizó la librería (lmtest) que determina la bondad del modelo obtenido, la Normalidad de los residuos se comprueba con función (shapiro.test()), la independencia con la función (dwtest) y homocedasticidad con la función

(bptest()).

Tras el análisis inicial de verificación de bondad y ajuste de la información, se procedió a la búsqueda de los factores que influyen en la producción de arroz, esto se logró de dos formas, con el uso de Árboles de decisión y AFDM, para obtener los árboles de decisión se usó la librería (rpart) y (rpart.plot) para la visualización de resultados; por otro lado el AFDM en R se lo realizó con el uso de las librerías (FactoMineR) y (factoextra).

2.6. Operacionalización de variables

Tabla 1-2: Operacionalización de variables

N°	Código	Variable	Descripción	Tipo de variable	Escala de medida	Categorías
1	ct_k504	Condición económica de cultivo	Forma de asociación económica para producción	Variable Cualitativa	Nominal	• Solo • Asociado • Invernadero
2	ct_rotacion	Rotación de cultivos	Cambio de tipo de sembrío en un mismo lugar	Variable Cualitativa	Nominal	• Si • No
3	ct_k506	Semilla de más uso	Clase de semilla según su preparación anterior al sembrado	Variable Cualitativa	Nominal	• Común • Mejorada • Hibrida nacional • Hibrida internacional
4	ct_k510h	Superficie sembrada en hectáreas	Cantidad de hectárea donde se realiza el sembrío	Cuantitativa	Razón	
5	ct_k511h	Superficie cosechada en hectáreas	Cantidad de hectáreas, donde se realizó la cosecha del producto	Cuantitativa	Razón	
6	ct_k513	Uso de riego	Utilización de cualquier fuente de agua para regar el sembrío	Variable Cualitativa	Nominal	• Si • No
7	ct_k514	Uso de fertilizantes	Utilización de cualquier tipo de fertilizantes en el sembrío	Variable Cualitativa	Nominal	• Si • No
8	ct_k515	Uso de fitosanitarios	Utilización de cualquier tipo de fitosanitarios en el sembrío	Variable Cualitativa	Nominal	• Si • No
9	ct_k518	Producción en libras por hectárea	Cantidad de libras de producción del sembrío	Cuantitativa	Razón	
10	ct_afecta_prod	Problemas del sembrío	Tipo de factor(problemas) que afecto a la producción del sembrío	Variable Cualitativa	Nominal	• Sequía/Heladas • Plagas/enfermedades • Inundación/ exceso de agua

						• Semilla • Prácticas inadecuadas/Falta de prácticas • Edad de la plantación • Ninguna
11	ct_prepara_suelo	Preparación de suelo	Realización de algún método de preparación del suelo antes de sembrar	Variable Cualitativa	Nominal	• Si • No
12	ct_deshierbe	Deshierbe	Realización de extracción de hierbas no deseadas para la siembra	Variable Cualitativa	Nominal	• Si • No
13	ct_aporque	Aporque	Realización de acumulación de tierra en el tallo de la planta	Variable Cualitativa	Nominal	• Si • No
14	ct_tutoreo	Tutoreo	Mantenimiento de la planta en todo su ciclo de vida hasta la producción de esta	Variable Cualitativa	Nominal	• Si • No
15	ct_forg	Uso fertilizante orgánico	Utilización de cualquier tipo de fertilizante orgánico en el sembrío	Variable Cualitativa	Nominal	• Si • No
16	ct_fqui	Uso fertilizante químico	Utilización de cualquier tipo de fertilizante químico en el sembrío	Variable Cualitativa	Nominal	• Si • No
17	ct_porg	Uso plaguicida orgánico	Utilización de cualquier tipo de plaguicida orgánico en el sembrío	Variable Cualitativa	Nominal	• Si • No
18	ct_pqui	Uso plaguicida químico	Utilización de cualquier tipo de plaguicida químico en el sembrío	Variable Cualitativa	Nominal	• Si • No

Fuente: INEC**,** 2017. (Encuesta de superficie y producción agraria continua).
Realizado por: Condo León José Luis, 2019

CAPÍTULO III

3. MARCO DE RESULTADOS Y DISCUSIÓN DE RESULTADOS

3.1. Análisis exploratorio de datos

Variable Cuantitativa: Superficie sembrada

Tabla 1-3: Resumen estadístico de la variable superficie sembrada

Estadístico	Valor
Mínimo	0,02
Máximo	874,94
Media	22,57
Mediana	4
Moda	20.99
Desviación estándar	59,13
Varianza	3497,17
Cuartil 2	1
Cuartil 3	21,16
Coeficiente de asimetría	6,90
Coeficiente de Curtosis	62,10
Longitud de variable	2805
Datos perdidos	0

Realizado por: Condo León José Luis, 2019

La Tabla 1-3 muestra que en promedio la superficie sembrada es de 22,57 hectáreas, la mínima es de 0,02 hectáreas y la de mayor tamaño es de 874,94 hectáreas, el tamaño más común de este tipo de sembrío es de 20.99 hectáreas, en contraste con la media es posible notar que la desviación estándar está muy alejada. Analizando el coeficiente de asimetría se puede determinar que la variable sigue una distribución asimétrica positiva y leptocúrtica, dicho de otra manera, los datos de esta variable se agrupan en mayor cantidad a la derecha de su media y más apuntada de lo normal.

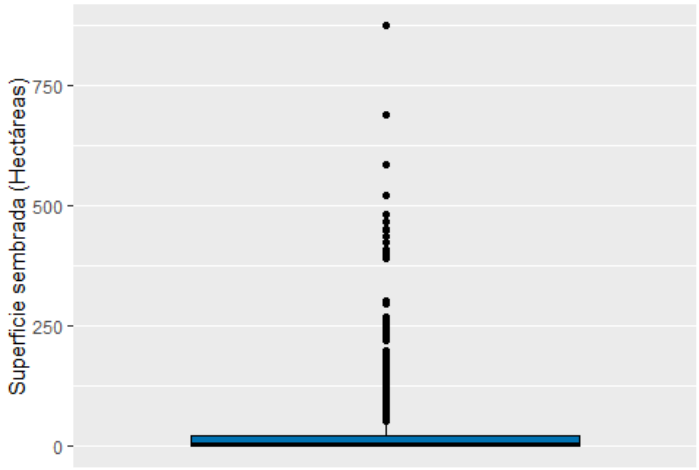


Gráfico 1-3. Diagrama de caja de la variable superficie sembrada

Realizado por: Condo León José Luis, 2019

En el Gráfico 1-3 es posible notar que existen varias mediciones distantes en relación con la media y los cuartiles, el 25% ≤ 1h , el 50% ≤ 4 h y el 75% ≤ 21.16 h.

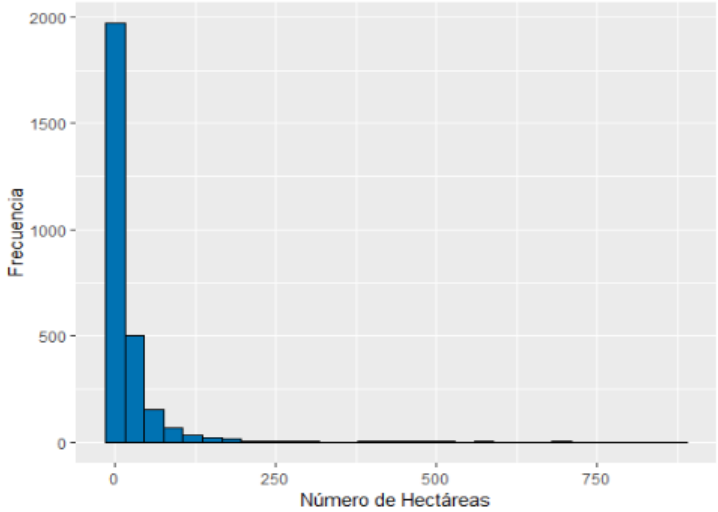


Gráfico 2-3. Histograma de frecuencia de la variable superficie sembrada

Realizado por: Condo León José Luis, 2019

Tabla 2-3: Distribución estadística de frecuencia de la variable superficie sembrada

Clase	Mínimo	Máximo	Marca de clase	Frecuencia Absoluta (f_i)	Frecuencia Relativa (h_i)	Frecuencia Absoluta Acumulada (F_i))	Frecuencia Relativa Acumulada (H_i)
1	0	200	100	2763	98.5	2763	98.5
2	200	400	300	21	0.7	2784	99.3
3	400	600	500	16	0.6	2800	99.8
4	600	800	700	4	0.1	2804	100.0
5	800	1000	900	1	0.0	2805	100.0

Realizado por: Condo León José Luis, 2019

La Tabla 2-3 muestra que el 98,5% de las superficies sembradas está comprendida en un intervalo entre 0 a 200 hectáreas con un total de 2763 sembríos, tan solo 42 cultivos de arroz superan las 200 hectáreas de tamaño y solo 1 sembrío está comprendido en un tamaño entre 800 a 1000 hectáreas.

Variable cuantitativa: Superficie cosechada

Tabla 3-3: Resumen estadístico de la variable superficie cosechada

Estadístico	Valor
Mínimo	0,02
Máximo	874,94
Media	22,19
Mediana	4
Moda	20,99
Desviación estándar	58,89
Varianza	3467,44
Cuartil 2	1
Cuartil 3	21,17
Coeficiente de asimetría	6,98
Coeficiente de Curtosis	66,25
Longitud de variable	2805
Datos perdidos	15

Realizado por: Condo León José Luis, 2019

El tamaño de las superficies cosechadas en promedio es de 22,19 hectáreas, la de menor tamaño es de 0,02 hectáreas y la mayor es de 874,94 hectáreas, la superficie más común que se cosecha es de 2099 hectáreas, la deviación estándar se encuentra muy alejada de la media. El coeficiente de asimetría deduce que la distribución de la variable es asimétrica positiva y el coeficiente de curtosis determina que esta es leptocúrtica. La variable muestra un 0,005% de valores perdidos.

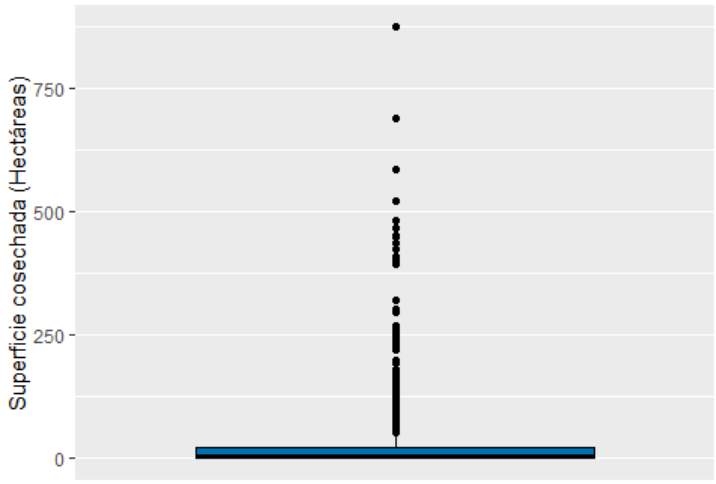


Gráfico 3-3. Diagrama de caja de la variable superficie cosechada

Realizado por: Condo León José Luis, 2019

El diagrama de caja correspondiente a la variable Superficie cosechada determina la existencia de varios datos muy alejados respecto a los cuartiles. Es posible notas que cerca del 25% de las superficies cosechadas es menor a 1 hectárea, el 50% es menor igual a 4 hectáreas y el 75% de estas es menor a 21,17 hectáreas.

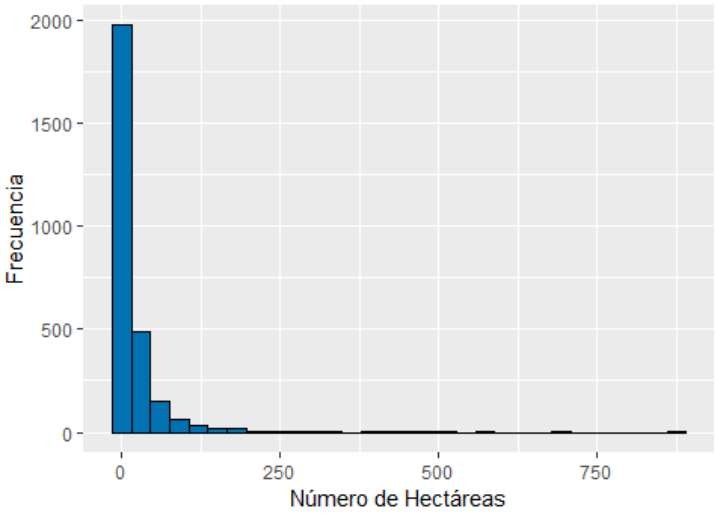


Gráfico 4-3. Histograma de frecuencia de la variable superficie cosechada

Realizado por: Condo León José Luis, 2019

Tabla 4-**3:** Distribución estadística de frecuencia de la variable superficie cosechada

Clase	Mínimo	Máximo	Marca de clase	Frecuencia Absoluta (f_i)	Frecuencia Relativa (h_i)	Frecuencia Absoluta Acumulada (F_i))	Frecuencia Relativa Acumulada (H_i)
1	0	200	100	2749	98.5	2749	98.5
2	200	400	300	20	0.7	2769	99.2
3	400	600	500	16	0.6	2785	99.8
4	600	800	700	4	0.1	2789	100.0
5	800	1000	900	1	0.0	2790	100.0

Realizado por Condo León José Luis, 2019

La d.e.f de la variable muestra que el 98,5 % de las superficies cosechadas tienen un tamaño menor a 200 hectáreas con un total de 2749 terrenos con esta característica, tan solo el 1,5% de estos sembríos tiene una superficie mayor a 200 hectáreas, cabe destacar que existe un sembrío que está comprendido en el intervalo entre 800 a 1000 hectáreas cosechadas.

Variable Cuantitativa: Producción

Tabla 5-3: Resumen estadístico de la variable producción por hectárea

Estadístico	Valor
Mínimo	25
Máximo	300
Media	212,21
Mediana	215
Moda	130
Desviación estándar	26,71
Varianza	713,1685
Cuartil 2	210
Cuartil 3	220
Coeficiente de asimetría	-2,99
Coeficiente de Curtosis	15,60
Longitud de variable	2805
Datos perdidos	15

Realizado por: Condo León José Luis, 2019

La Tabla 5-3 presenta un resumen estadístico de la variable Producción de arroz por hectárea que esta medida en libras; en promedio la producción de los sembríos alcanza 212,21 libras, el cultivo que menor producción generó es de 25 libras, y el de mayor tamaño es de 300 libras, siendo la cantidad de 130 libras la producción más común en estos sembríos, la desviación estándar no se encuentra tan alejada respecto a la media. La mayor parte de las mediciones se encuentra acumulada en la parte izquierda de la media, por lo que se puede decir que la variable tiene una distribución asimétrica negativa y posee características de una variable leptocúrtica. Es posible notar la existencia de 15 valores perdidos en esta variable.

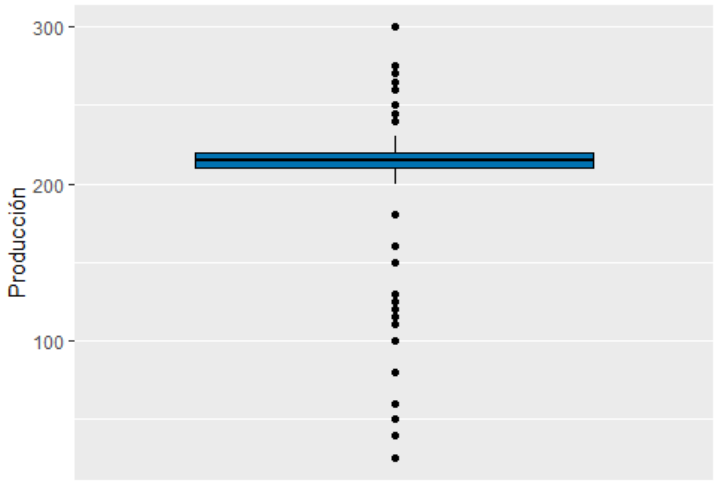


Gráfico 5-3. Diagrama de caja de la variable producción
Realizado por: Condo León José Luis, 2019

La información mostrada por el diagrama de caja muestra la existencia de valores extraños en la variable Producción pues se encuentran muy distanciados respecto a los cuartiles, el 25% de los cultivos produjeron cantidades menores o iguales a 21 libras, el 50% de estos produjeron cantidades menores o iguales a 215 libras y el 75% produjo cantidades menores o iguales a 220 libras.

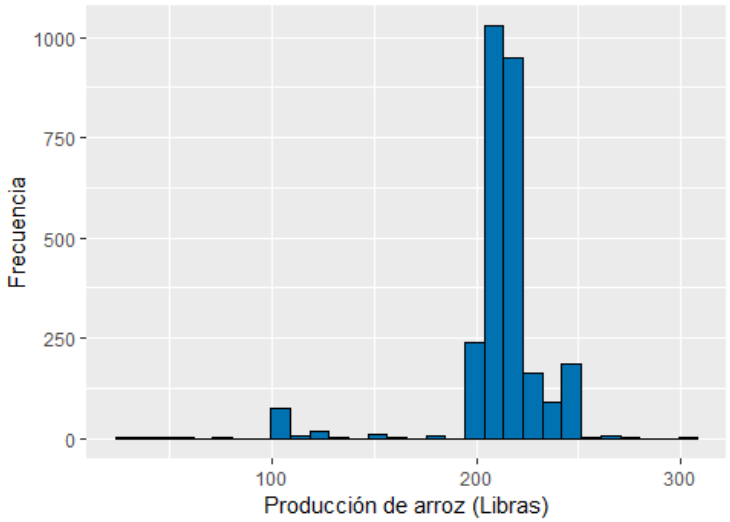


Gráfico 6-3. Histograma de frecuencia de la variable producción
Realizado por: Condo León José Luis, 2019

Tabla 6-3: Distribución estadística de frecuencia de la variable producción por hectárea

Clase	Mínimo	Máximo	Marca de clase	Frecuencia Absoluta (f_i)	Frecuencia Relativa (h_i)	Frecuencia Absoluta Acumulada (F_i))	Frecuencia Relativa Acumulada (H_i)
1	0	50	25	6	0.2	6	0.2
2	50	100	75	80	2.9	86	3.1
3	100	150	125	33	1.2	119	4.3
4	150	200	175	245	8.8	364	13.0
5	200	250	225	2414	86.5	2778	99.6
6	250	300	275	12	0.4	2790	100.0

Realizado por: Condo León José Luis, 2019

El 86,5% de los datos aproximadamente 2414 cultivos tuvo una producción comprendida entre 200 a 250 libras, 6 culticos tuvieron una producción comprendida entre 0 a 50 libras, tan solo 12 de estos tienen una producción comprendida entre 250 a 300 libras.

Variable cualitativa: Condición económica

Tabla 7-3: Resumen estadístico de la variable condición económica

Clase	Frecuencia	Porcentaje
Solo	2805	100%
Asociado	0	0
Invernadero	0	0

Realizado por: Condo León José Luis, 2019

Tal como se muestra en la Tabla 7-3, El 100% de los cultivos fueron financiados por un solo productor, haciendo que una asociación o invernadero no sea característica disponible en ninguno de los sembríos.

Variable cualitativa: Rotación de cultivo

Tabla 8-3: Resumen estadístico de la variable rotación de cultivo

Clase	Frecuencia	Porcentaje
Si	374	13%
No	2431	87%

Realizado por: Condo León José Luis, 2019

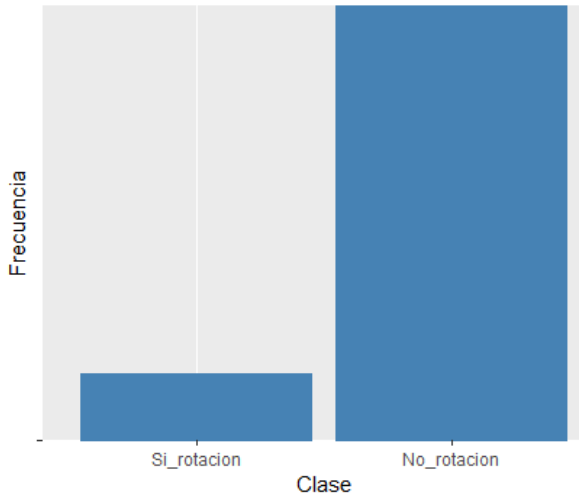


Gráfico 7-3. Diagrama de barras de la variable rotación de cultivo

Realizado por: Condo León José Luis, 2019

Por lo que se muestra en la Tabla 8-3, el 87% de los sembríos no realizó la rotación del cultivo en ninguna etapa del crecimiento de planta de arroz, con un total de 2431 terrenos con esta característica, por otro lado 374 de los sembríos si realizo la rotación del cultivo.

Variable cualitativa: Clase semilla

Tabla 9-3: Resumen estadístico de la variable clase semilla

Clase	Frecuencia	Porcentaje
Común	1502	54%
Mejorada	741	26%
Hibrida nacional	561	20%
Hibrida internacional	0	0%
Datos perdidos	0	0%

Realizado por: Condo León José Luis, 2019

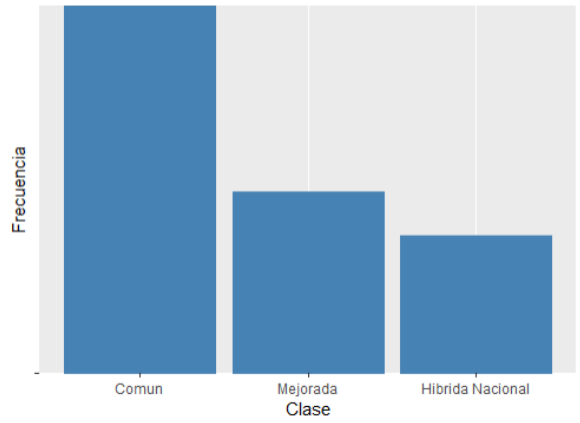


Gráfico 8-3. Diagrama de barras de la variable clase semilla

Realizado por: Condo León José Luis, 2019

En el 54% de los cultivos utilizaron semilla de tipo común para sembrar, este tipo de semilla fue

utilizada 1502 veces, seguida por la semilla de tipo mejorada con un 26% y la semilla de tipo hibrida nacional con el 20%, no se cultivó ningún sembrío con semillas de tipo hibrida internacional; la variable no presenta datos perdidos.

Variable cualitativa: Uso de riego

Tabla 10-3: Resumen estadístico de la variable uso de riego

Clase	Frecuencia	Porcentaje
Si	1705	61%
No	1100	39%

Realizado por: Condo León José Luis, 2019

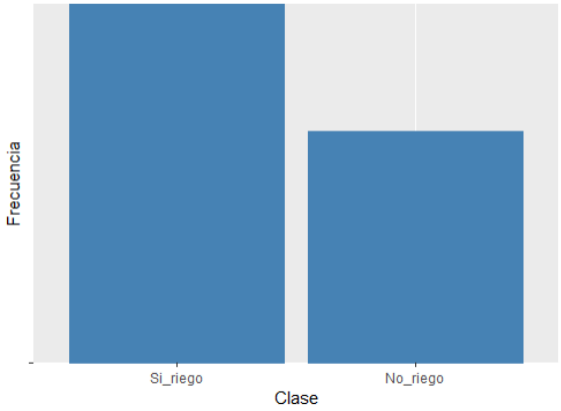


Gráfico 9-3. Diagrama de barras de la variable uso de riego

Realizado por: Condo León José Luis, 2019

En el 61% de los sembríos de arroz se realizó el uso de riego, a diferencia de los 1100 sembríos que lo usaron.

Variable cualitativa: Uso de fertilizantes

Tabla 11-3: Resumen estadístico de la variable uso de fertilizantes

Clase	Frecuencia	Porcentaje
Si	2757	98%
No	48	2%

Realizado por: José Condo, 2019

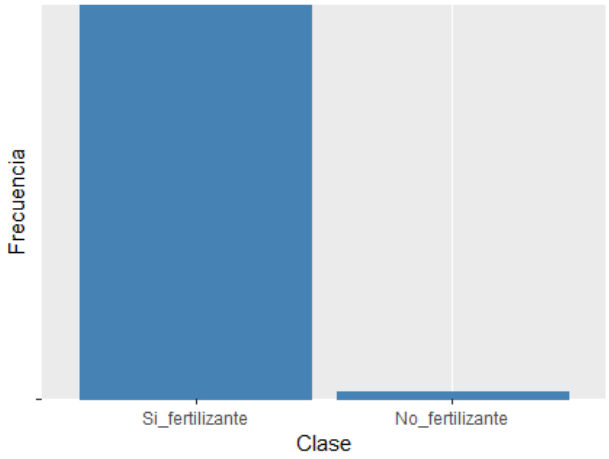


Gráfico 10-3. Diagrama de barras de la variable uso de fertilizantes

Realizado por: Condo León José Luis, 2019

El 98% de los sembríos utilizo fertilizantes para el tratamiento de la planta de arroz, tan solo el 2% de los cultivos no usaron este tipo de tratamiento en el crecimiento de las plantas.

Variable cualitativa: Uso de fitosanitarios

Tabla 12-3: Resumen estadístico de la variable uso de fitosanitarios

Clase	Frecuencia	Porcentaje
Si	2738	97%
No	67	3%

Realizado por: Condo León José Luis, 2019

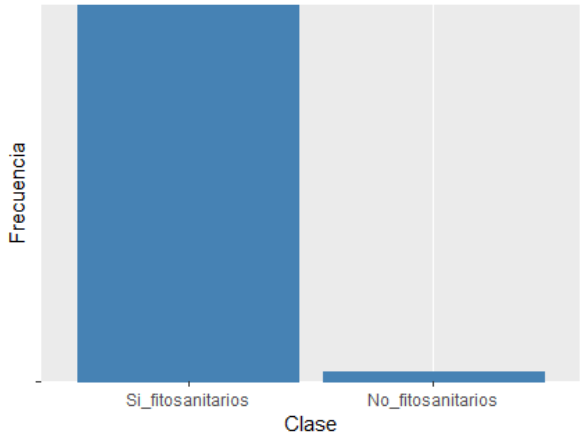


Gráfico 11-3. Diagrama de barras de la variable uso de fitosanitarios

Realizado por: Condo León José Luis, 2019

El 97% de los cultivos realizo el uso de fitosanitarios en los sembríos, tan solo el 3% o visto de otra manera 67 cultivos no realizo el uso de los fitosanitarios en las plantas.

Variable cualitativa: Problemas de sembrío

Tabla 13-3: Resumen estadístico de la variable problemas de sembrío

Clase	Frecuencia	Porcentaje
Sequia/Heladas	51	1%
Plagas/Enfermedades	1673	60%
Inundación/Exceso de agua	122	5%
Semilla	100	4%
Practicas inadecuadas/Falta de practicas	60	2%
Edad de la plantación	3	0%
Ninguna	796	28%

Realizado por: Condo León José Luis, 2019

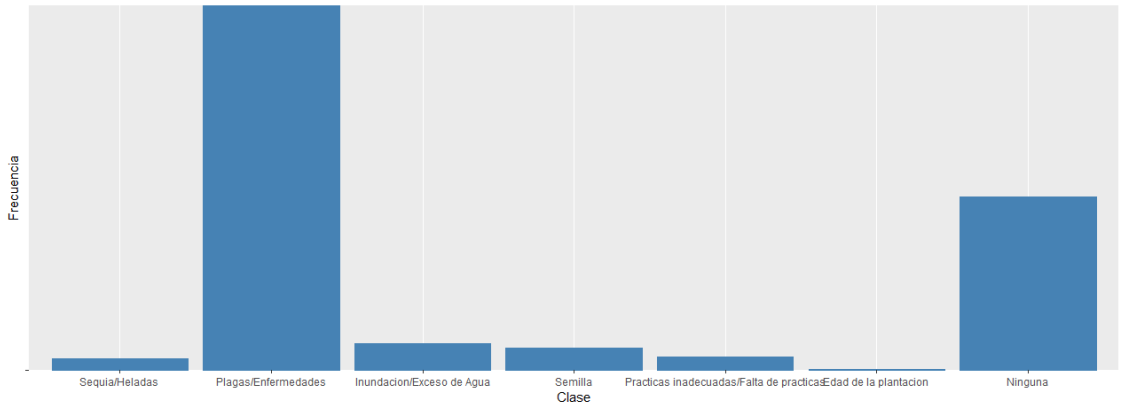


Gráfico 12-3. Diagrama de barras de la variable problemas de sembrío
Realizado por: Condo León José Luis, 2019

La Tabla 12-3 muestra un resumen estadístico de la variable Problemas de sembrío, misma que es corroborada por el Gráfico 12-3, donde se muestra que el mayor problema que existe en los cultivos de arroz son las Plagas/enfermedades, este es el mayor problema con el 60% de la información, seguido por Inundaciones/excesos de agua con un 5%; el 28% de los cultivos no tuvieron ningún problema en sus plantas.

Variable cualitativa: Preparación de suelo

Tabla 14-3: Resumen estadístico de la variable preparación de suelo

Clase	Frecuencia	Porcentaje
Si	2661	95%
No	144	5%

Realizado por: Condo León José Luis, 2019

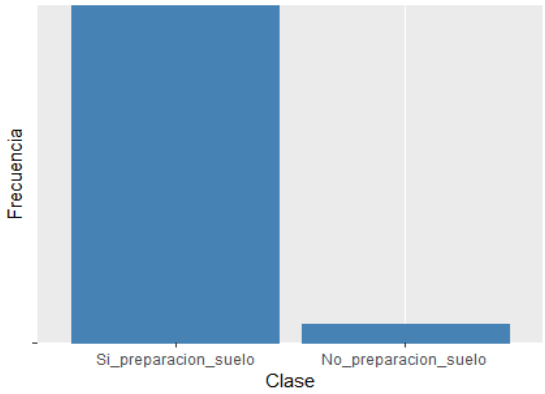


Gráfico 13-3. Diagrama de barras de la variable preparación de suelo

Realizado por: Condo León José Luis, 2019

Casi todos los cultivos analizados realizaron la preparación del suelo con un total de 2661 sembríos, siendo el 95%, tan solo el 5% de estos no realizo este proceso en los cultivos.

Variable cualitativa: Deshierbe

Tabla 15-3: Resumen estadístico de la variable deshierbe

Clase	Frecuencia	Porcentaje
Si	2189	78%
No	616	22%

Realizado por: Condo León José Luis, 2019

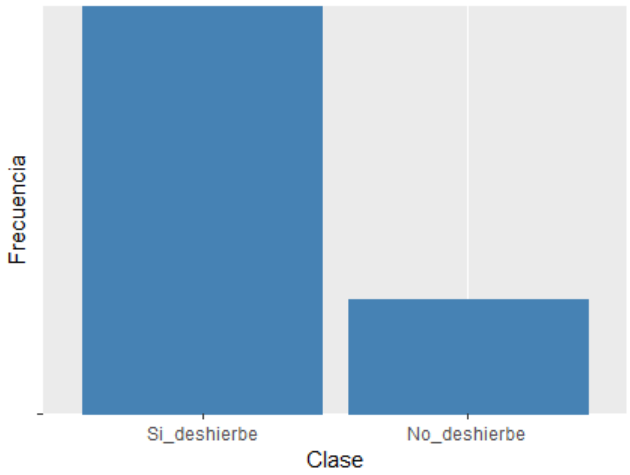


Gráfico 14-3. Diagrama de barras de la variable deshierbe

Realizado por: Condo León José Luis, 2019

El 78% de los terrenos cultivados con arroz tuvieron un proceso de deshierbe en algún punto de su crecimiento, por otro lado, el 22% de los cultivos no tuvo ningún tipo de tratamiento de limpieza del suelo.

Variable cualitativa: Aporque

Tabla 16-3: Resumen estadístico de la variable aporque

Clase	Frecuencia	Porcentaje
Si	15	1%
No	2790	99%

Realizado por: Condo León José Luis, 2019

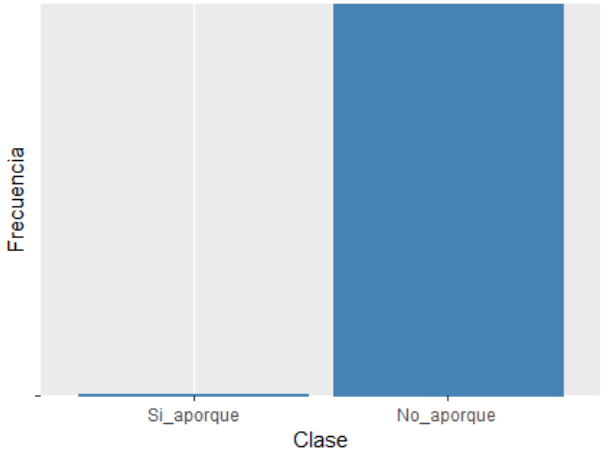


Gráfico 15-3. Diagrama de barras de la variable aporque

Realizado por: Condo León José Luis, 2019

El 99% de los cultivos no realizo el aporque del terreno en ninguna etapa de crecimiento de los sembríos, tan solo en 15 sembríos cumplieron con este procedimiento.

Variable cualitativa: Tutoreo

Tabla 17-3: Resumen estadístico de la variable tutoreo

Clase	Frecuencia	Porcentaje
Si	15	1%
No	2790	99%

Realizado por: Condo León José Luis, 2019

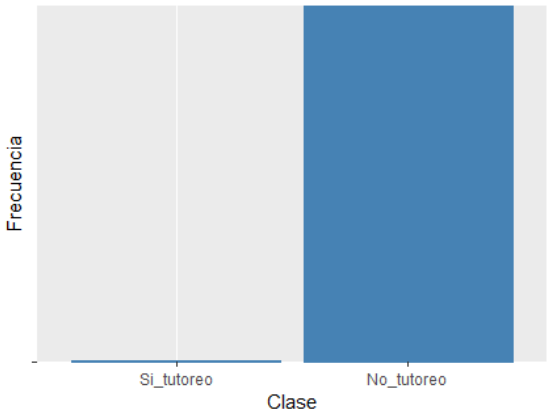


Gráfico 16-3. Diagrama de barras de la variable tutoreo

Realizado por: Condo León José Luis, 2019

Al igual que la variable Aporque, el 99% de los cultivos no tuvieron asesoramiento o Tutoreo profesional, tan solo 15 de los cultivos cumplieron con esta característica.

Variable cualitativa: Uso fertilizante orgánico

Tabla 18-3: Resumen estadístico de la variable uso fertilizante orgánico

Clase	Frecuencia	Porcentaje
Si	207	7%
No	2598	93%

Realizado por: Condo León José Luis, 2019

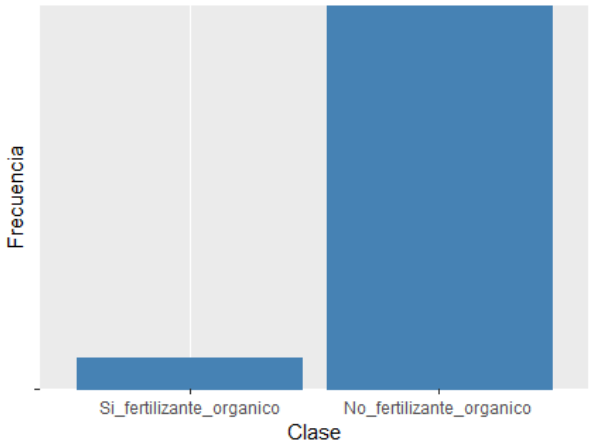


Gráfico 17-3. Diagrama de barras de la variable uso fertilizante orgánico

Realizado por: Condo León José Luis, 2019

El uso del fertilizante orgánico tan solo fue aplicado en el 7% de los sembríos, siendo el 93% restante de cultivos los que no utilizaron este tipo de tratamiento en los sembríos.

Variable cualitativa: Uso fertilizante químico

Tabla 19-3: Resumen estadístico de la variable uso fertilizante químico

Clase	Frecuencia	Porcentaje
Si	2756	99%
No	49	1%

Realizado por: Condo León José Luis, 2019

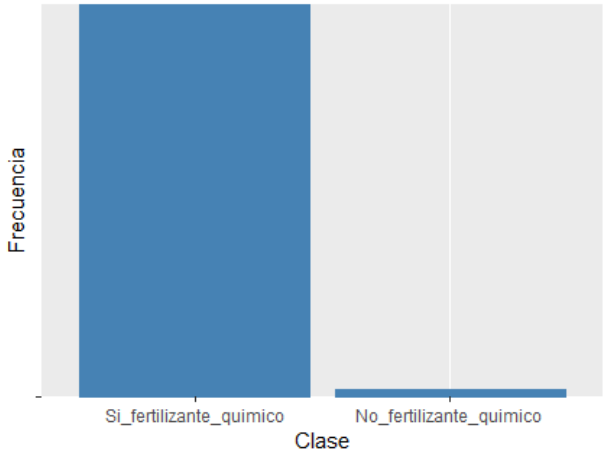


Gráfico 18-3. Diagrama de barras de la variable uso fertilizante químico

Realizado por: Condo León José Luis, 2019

Casi la totalidad de los cultivos analizados han sido tratados con fertilizantes químicos en alguna etapa antes de la producción de la planta, siendo el 99% de los cultivos los que tienen esta característica.

Variable cualitativa: Uso plaguicida orgánico

Tabla 20-3: Resumen estadístico de la variable uso plaguicida orgánico

Clase	Frecuencia	Porcentaje
Si	24	1%
No	2781	99%

Realizado por: Condo León José Luis, 2019

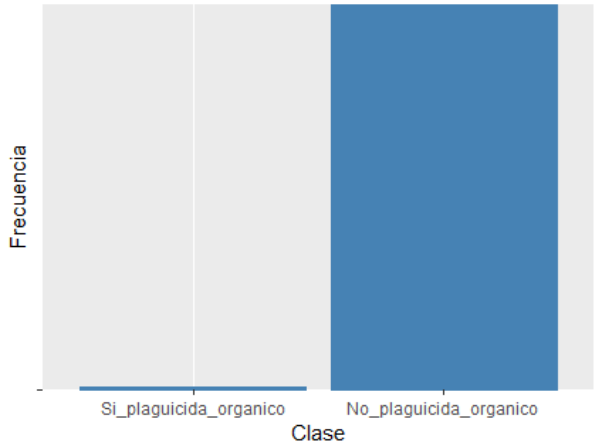


Gráfico 19-3. Diagrama de barras de la variable uso plaguicida orgánico

Realizado por: Condo León José Luis, 2019

El 99% de los cultivos de arroz no tuvo un tratamiento de plaguicida orgánico, tan solo 24 de los 2805 si poseen esta característica.

Variable cualitativa: Uso plaguicida química

Tabla 21-3: Resumen estadístico de la variable uso plaguicida químico

Clase	Frecuencia	Porcentaje
Si	2736	98%
No	69	2%

Realizado por: Condo León José Luis, 2019

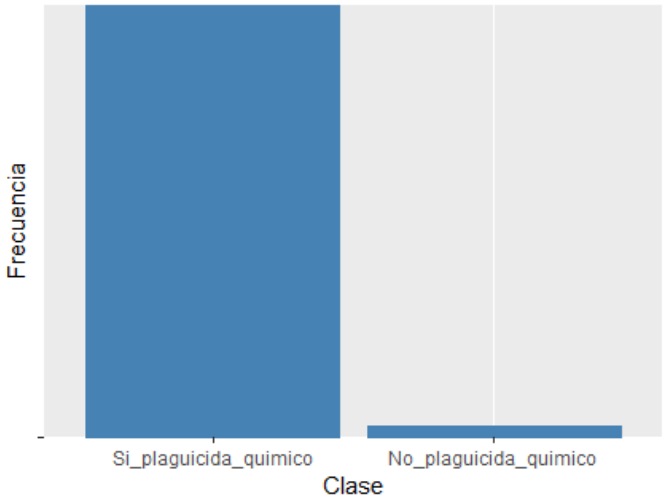


Gráfico 20-3. Diagrama de barras de las variables uso plaguicida químico

Realizado por: Condo León José Luis, 2019

El 98% de los cultivos hicieron uso de los plaguicidas químicos en alguna etapa del crecimiento de la planta.

3.2. Análisis bivariado de datos

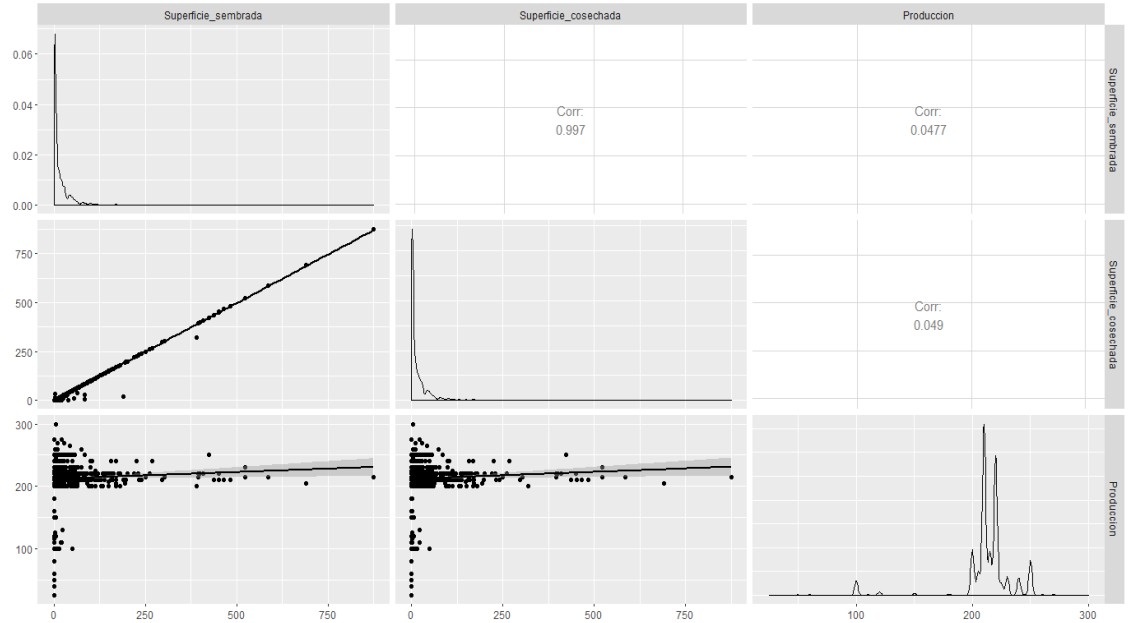


Gráfico 21-3. Análisis de distribución y correlación de Pearson de las variables cuantitativas

Realizado por: Condo León José Luis, 2019

El análisis de correlación de Pearson mostrado en el Gráfico 21-3 determina que las variables Superficie sembrada y Superficie cosechada tiene una relación lineal positiva casi perfecta con un coeficiente de 0.99, dicho de otra manera, este par de variables tienen una marcada dependencia, a medida que una aumenta la otra también lo hace. Las variables Producción y Superficie sembrada tienen una relación lineal positiva muy débil con un coeficiente de 0.047, sucede lo mismo entre las variables Producción y Superficie cosechada, existe una relación lineal muy débil casi imperceptible con un coeficiente de Pearson de 0.049.

La densidad mostrada por las variables superficie sembrada y superficie cosechada se asemeja a la de una variable con distribución Gamma, la forma de densidad de la variable Producción es posible que se ajuste a una distribución normal.

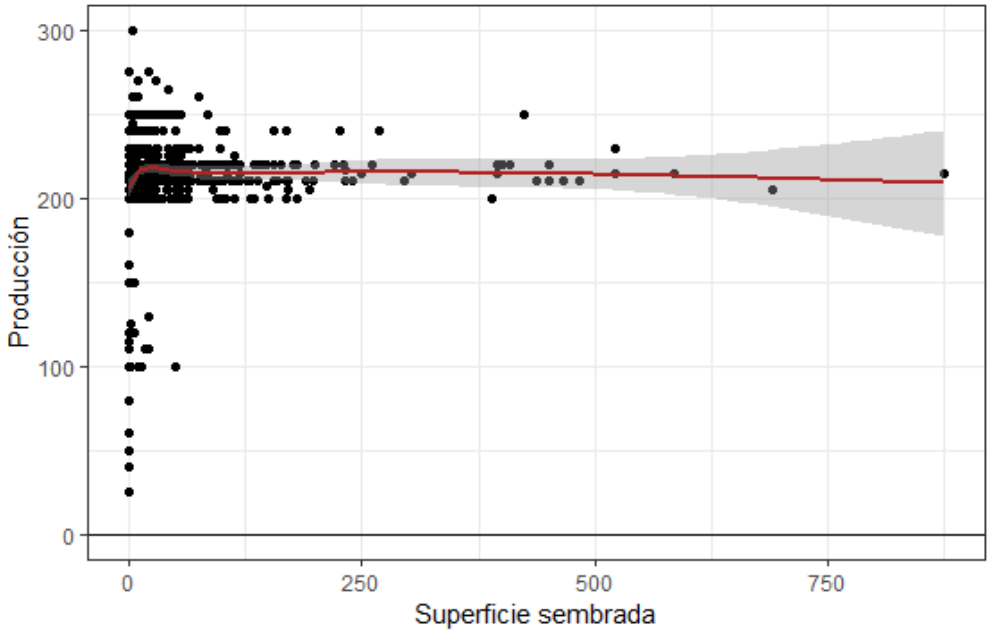


Gráfico 22-3. Diagrama de correlación de variables producción y superficie sembrada
Realizado por: Condo León José Luis, 2019

Con un coeficiente de correlación de 0.047 sumado a la información proporcionada por la Gráfica 22-3, es posible notar la falta de relación entre las variables Superficie sembrada y la variable Producción, salvo contados casos, la producción de los sembríos no se eleva cuando la superficie sembrada es más grande.

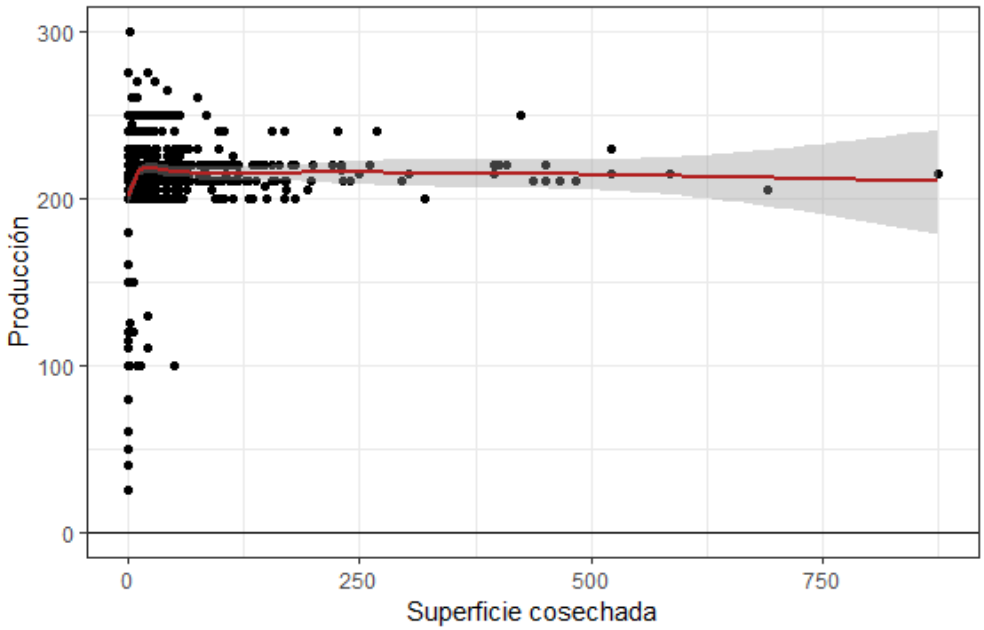


Gráfico 23-3. Diagrama de correlación de variables producción y superficie cosechada

Realizado por: Condo León José Luis, 2019

Como se puede observar en el Gráfico 23-3 las variables Producción y Superficie cosechada no poseen una relación marcada, los niveles de producción no se elevan cuando la superficie cosechada es más grande, cabe recalcar que existen varios casos particulares cuya extensión cosechada es muy grande por lo que su producción va a ser más grande que los demás.

3.3. Técnicas de interdependencia

3.3.1. Imputación de datos faltantes

Se detectaron 15 datos faltantes por cada una de las variables cuantitativas Superficie cosechada y Producción, el relleno de información faltante se realizó con un modelo de regresión lineal simple denotado como $y_i = \beta_0 + \beta_1 x_i$, donde y_i es el individuo i con información faltante en la variable y, β_0 y β_1 son los coeficientes de regresión y x_i es el individuo i con información completa en la variable x, tras la imputación los principales estadísticos de los datos se presentan en la Tabla 22-3

Tabla 22-3: Estadísticos de variables imputadas

Estadístico	Superficie sembrada	Superficie cosechada	Producción
Mínimo	0,03	0,03	25
Máximo	874,94	874,94	300
Media	22,56	22,27	212,20
Mediana	4	4	212,60
Datos perdidos	0	0	0

Realizado por: Condo León José Luis, 2019

Al comparar los estadísticos de las variables con y sin imputación es posible deducir que no han cambiado en gran medida, es decir la imputación no afecto en la estimación de los diferentes estadísticos y debido a la poca cantidad de datos faltantes en comparación con el total de individuos se puede asumir una imputación exitosa.

3.3.2. Contraste de normalidad multivariante

1.- Hipótesis

$\boldsymbol{H_0}$: El conjunto de variables estadísticas (X_1: Superficie sembrada en hectáreas, X_2: Superficie cosechada en hectáreas, X_3: Producción en libras) se ajustan a una distribución normal multivariante.

$\boldsymbol{H_1}$: El conjunto de variables estadísticas (X_1: Superficie sembrada en hectáreas, X_2: Superficie cosechada en hectáreas, X_3: Producción en libras) no se ajustan a una distribución normal multivariante.

2.- Nivel de significancia

$\alpha = 5\%$

3.- Estadístico

K_p^2 =0,42055, p =0,00

4.- Región de rechazo

$Si\ p < \propto\ Rechazar\ H_o$

Si 0,0 < 0,5 $Rechazar\ H_o$

5.- Decisión

Dado que se obtuvo un valor *p* menor que el nivel de significancia se rechaza la hipótesis nula, por tal razón se deduce que el conjunto de variables estadísticas (X_1: Superficie sembrada en hectáreas, X_2: Superficie cosechada en hectáreas, X_3: Producción en libras) no se ajustan a una distribución normal multivariante.

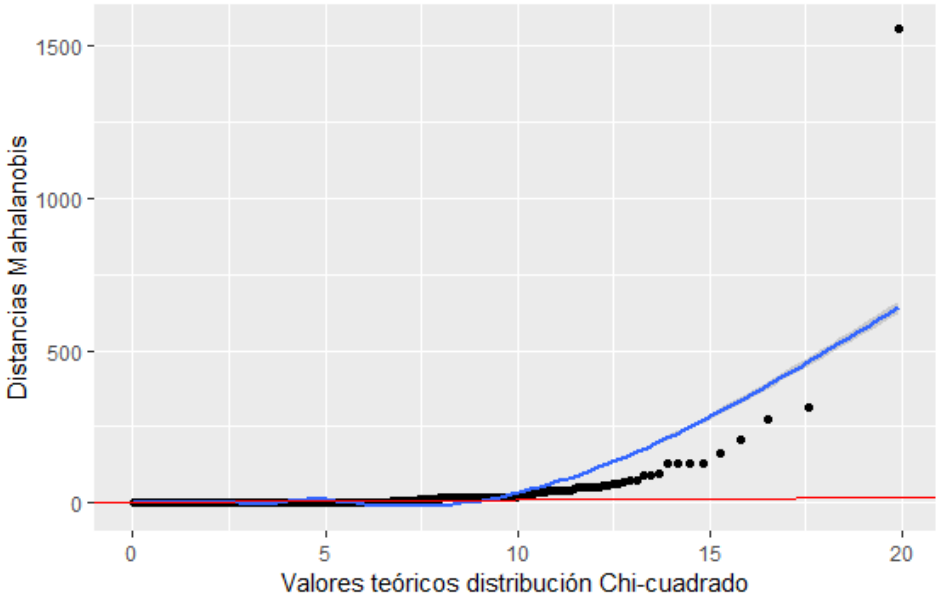


Gráfico 24-3. Prueba gráfica de normalidad multivariante de Johnson y Wichern

Realizado por: Condo León José Luis, 2019

El Gráfico 24-3 demuestra que los datos no se ajustan a una distribución normal, mucha de la información se encuentra muy alejada de la línea de normalidad, corroborando la información proporcionada por el contraste de normalidad multivariante de Shapiro-Wilk.

3.3.3. Detección de datos atípicos con distancia de Mahalanobis

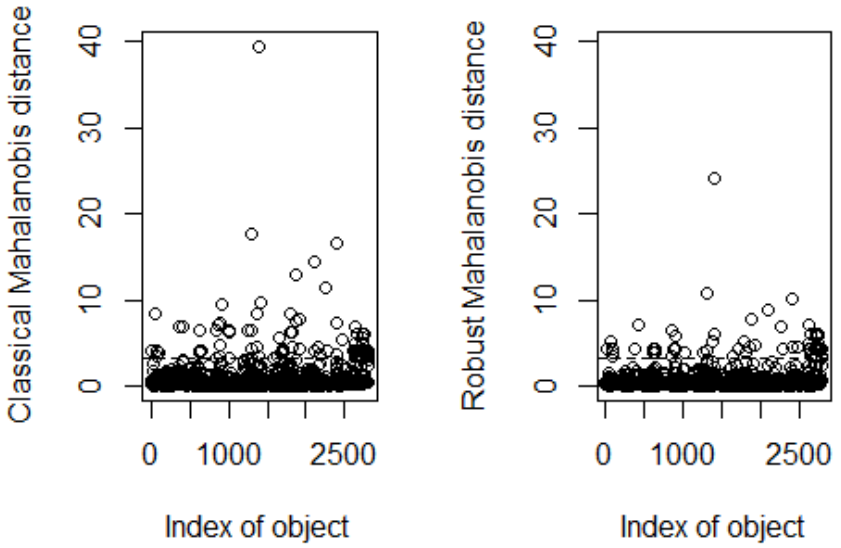


Gráfico 25-3. Datos atípicos determinados por distancias de Mahalanobis

Realizado por: Condo León José Luis, 2019

El análisis de las distancias de Mahalanobis clásico y robusto determino la existencia de 144 datos atípicos, esto debido a que los individuos tuvieron distancias por sobre el punto de corte, determinado por $F_{(1-\propto,p,n-p-1)} = F_{(1-0.05,\ 3,\ 2805-3-1)} = 3.36$, ésta información es posible

notarla en la Gráfica 25-3donde existen varias distancias que se encuentran extremadamente alejada del punto de corte marcada por una línea horizontal.

Para una mejor comprensión en el contraste de los estadísticos que a continuación se usará se decidió denominar a X_1: ***Superficie Sembrada***, X_2: ***Superficie cosechada***, X_3: ***Producción en libras.***

Tabla 23-3: Contraste de estadísticos de datos con y sin información atípica

	Información completa			**Información sin datos atípicos**			**Datos atípicos**		
Estadístico	X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3
Mínimo	0,02	0,02	25	0,02	0,02	130	0,13	0,07	25
Máximo	874,94	874,94	300	303,40	303,40	300	874,94	874,94	250
Media	22,57	22,27	212,2	18,40	18,26	216,8	99,51	96,41	127
Mediana	4	4	212,6	4,50	4	215	1	1	100
Varianza	3497,18	3470	709,36	1137,60	1139,45	188,23	41128,69	41011,25	2709,93
Desviación estándar	59,14	58,91	26,63	33,72	33,75	13,71	202,80	202,51	52,05
Coeficiente de asimetría	6,90	6,95	-3,01	3,74	3,74	0,66	1,87	1,91	0,84
Coeficiente de curtosis	64,99	65,82	15,68	21,13	21,11	7,05	5,24	5,37	2,51
Número de Elementos	2805	2805	2805	2661	2661	2661	144	144	144

Realizado por: Condo León José Luis, 2019

Es posible observar que los cambios más importantes obtenidos al retirar los datos atípicos fueron en la desviación estándar, varianza, coeficiente de asimetría y curtosis, lo que demuestra que esta información extraña estaba modificando las distribuciones originales de las variables, lo que a su vez afecta de manera directa en el cálculo de los estadísticos y estimaciones.

Dado que el objetivo principal es determinar las características y factores influyentes de los sembríos con mayor y menor producción, la extracción de los valores extraños sería el paso correcto para dar, pues ayudaría a que las características del grupo de datos se reflejen en los resultados y no tener la influencia incorrecta que producen los valores atípicos.

3.3.4. Validación del modelo de regresión múltiple

Con el fin de determinar si las variables cuantitativas Superficie sembrada y Superficie cosechadas sirven de manera idóneo como variables predictoras para la variable Producción en libras se requiere la comprobación de todos los supuestos que la regresión lineal exige, de esta manera se verifica que existe una relación significativa entre las variables respuesta y las

predicciones, a continuación, se resuelve lo idóneo que sería la utilización de estas variables.

3.3.4.1. Independencia de los residuos

1.- Hipótesis

$\boldsymbol{H_0}$: Existe auto correlación de los residuos

$\boldsymbol{H_1}$: No existe auto correlación de los residuos

2.- Estadístico

$D = 0{,}61, p = 0.00$

3.- Decisión

Según la información proporcionada por la Tabla 2-3 y el estadístico $D = 0.61$, se determina la existencia de auto correlación en los residuos, dicho de otra manera, los residuos de las predicciones tienen un patrón determinado en ciertas posiciones, por lo que no se rechaza la $\boldsymbol{H_0}$.

3.3.4.2. Normalidad de los residuos

1.- Hipótesis

$\boldsymbol{H_0}$: Los residuos de la regresión lineal múltiple se ajustan a una distribución normal univariada

$\boldsymbol{H_1}$: Los residuos de la regresión lineal múltiple no se ajustan a una distribución normal univariada

2.- Nivel de significancia

$\alpha = 5\%$

3.- Estadístico

$W^* = 0{,}63816, \quad p = 0{,}00$

4.- Decisión

Según el contraste de Shapiro-Wilk con un estadístico $W^* = 0{,}63$ y un valor $p = 0{,}00$ se determina que los residuos producidos por la regresión lineal no se ajustan a una distribución normal univariante, por lo que se rechaza la hipótesis nula.

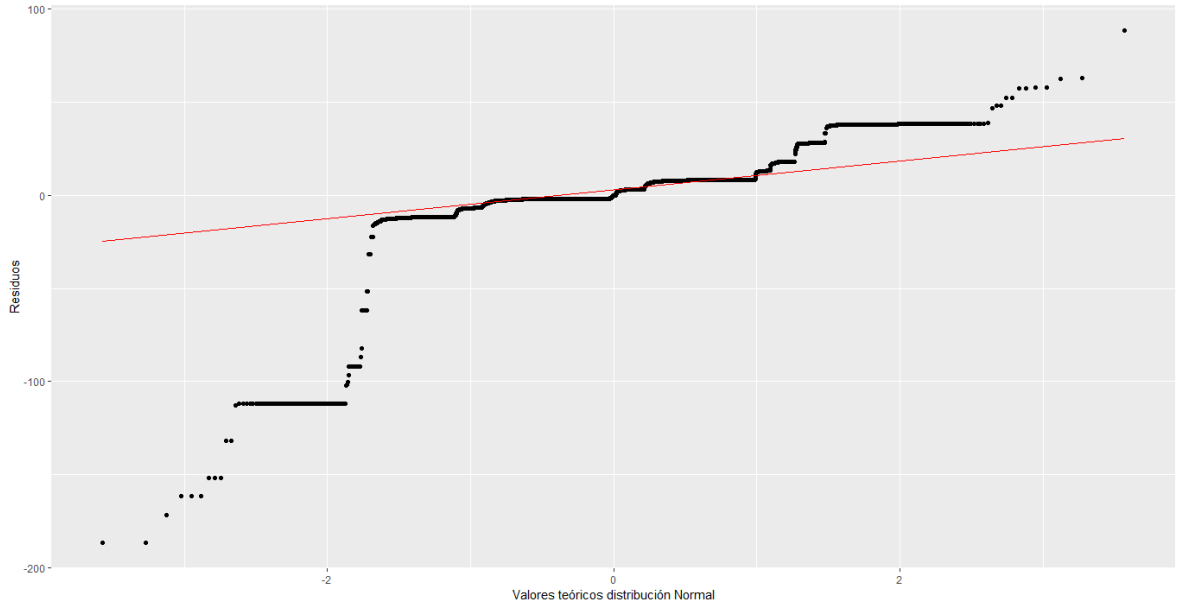


Gráfico 26-3. Prueba gráfica de normalidad univariada de los residuos de regresión
Realizado por: Condo León José Luis, 2019

El análisis gráfico muestra que los residuos producidos por el modelo de regresión lineal múltiple no se ajustan a una distribución normal, esto en razón a que estos valores se encuentran muy distanciados de la línea de normalidad, lo que corrobora los resultados producidos por el contraste de normalidad univariante de Shapiro-Wilk.

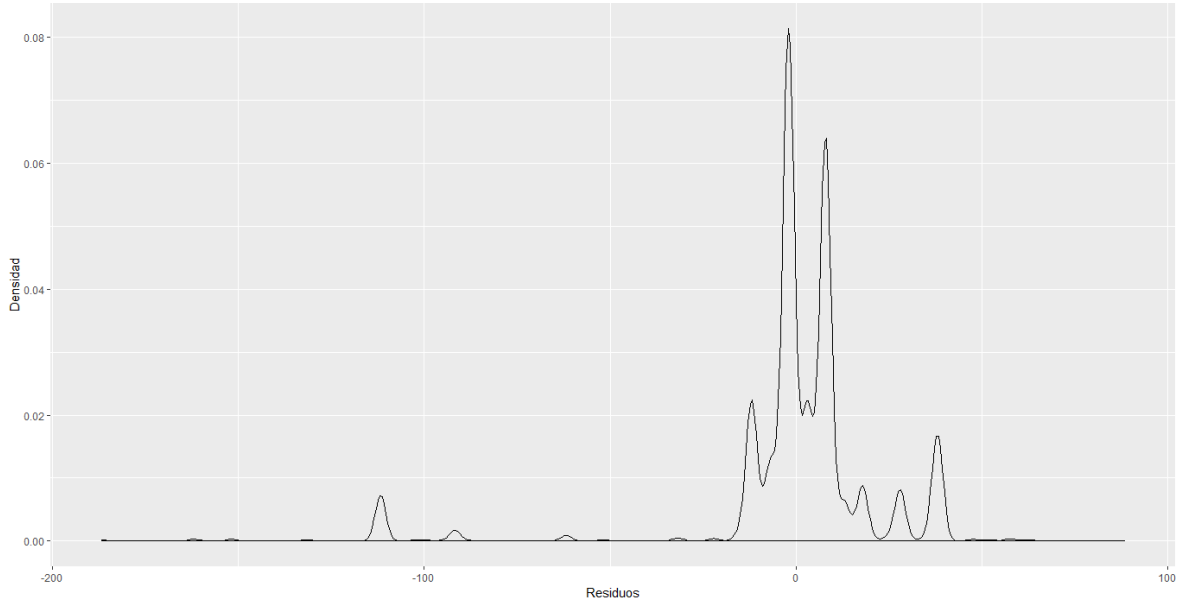


Gráfico 27-3. Función de densidad de los residuos de la regresión
Realizado por: Condo León José Luis, 2019

El Gráfico 27-3 muestra que la densidad producida por los residuos no se asemeja a la de una distribución normal, por lo que esta es otra evidencia de la falta de normalidad de los residuos.

3.3.4.3. Homocedasticidad

1.- Hipótesis

$\boldsymbol{H_0}$: Existe homocedasticidad en función de las variables independientes

$\boldsymbol{H_1}$: No existe homocedasticidad en función de las variables independientes

2.- Nivel de significancia

$\alpha = 5\%$

3.- Estadístico

$BP = 14{,}126\,, \qquad p = 0{,}0008563$

4.- Decisión

El contraste de hipótesis sobre homocedasticidad con un estadístico $BP = 14.12$ y un valor $p = 0.00$ determina que no existe homocedasticidad en función de las variables independientes, por lo que se rechaza la hipótesis nula.

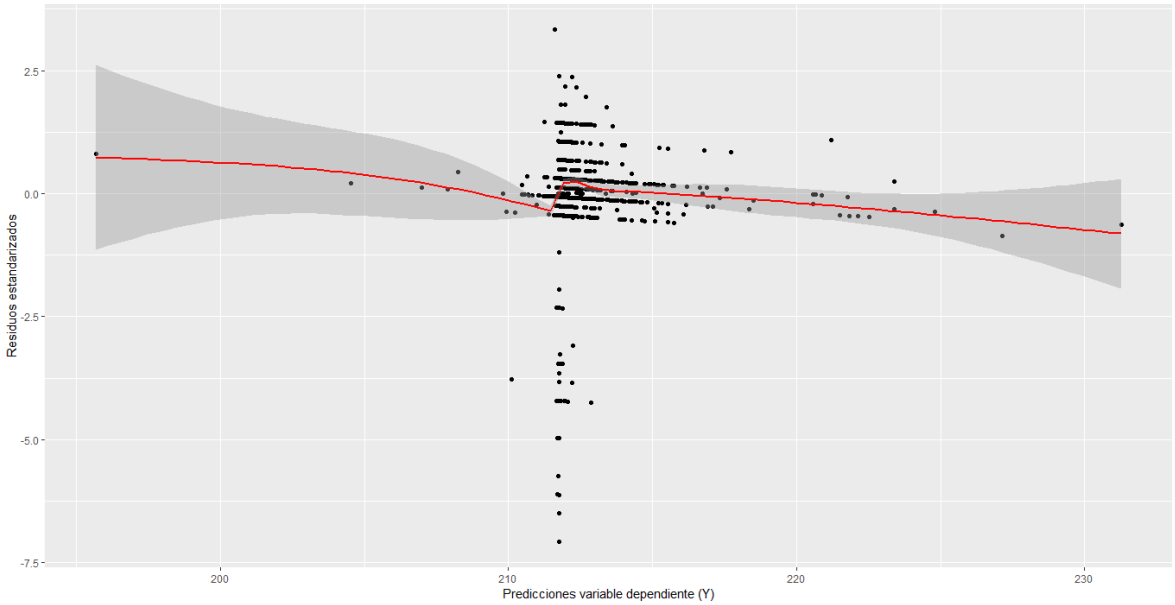


Gráfico 28-3. Prueba de homocedasticidad en función de las variables independientes
Realizado por: Condo León José Luis, 2019

El Gráfico 28-3 corrobora los resultados presentados por el contraste de hipótesis, no existe homocedasticidad, pues existe niveles de expansión crecientes y decrecientes de los residuos estandarizados.

Tras realizar las pruebas que determinan el cumplimiento de los supuestos es posible resolver que las variables Superficie sembrada y Superficie Cosechada no serán buenas predictoras de la variable Producción en libras, esto en razón a que no se cumplió con ninguno de los supuestos que un modelo correcto exige, el uso de AFDM exige variables explicativas relacionadas con la variable respuesta, por tal razón se realizó la estandarización de los datos, lo que lleva a cumplir con todos estos supuestos.

3.4. Árboles de regresión

3.4.1. Generación del modelo base de árboles de regresión

Como primer paso se generó un árbol de regresión con el uso de la base de datos sin mediciones atípicas, a partir de esto se creó el árbol base de la forma:

$$Árbol_{base} = Variable_{Respuesta} \sim Variables_{Predictoras} \tag{1.3}$$

$$\begin{aligned} Árbol_{base} = {} & \text{Produccion} \sim \text{Superficie\_sembrada} + \text{Superficie\_cosechada} \\ & + \text{Condicion\_economica} + \text{Rotacion\_cultivo} + \text{Clase\_semilla} \\ & + \text{Uso\_riego} + \text{Uso\_fertilizantes} + \text{Uso\_fitosanitarios} \\ & + \text{Problemas\_sembrio} + \text{Preparacion\_suelo} + \text{Deshierbe} \\ & + \text{Aporque} + \text{Tutoreo} + \text{Uso\_fertilizante\_organico} \\ & + \text{Uso\_fertilizante\_quimico} + \text{Uso\_plaguicida\_organico} \\ & + \text{Uso\_plaguicida\_quimico} \end{aligned} \tag{2.3}$$

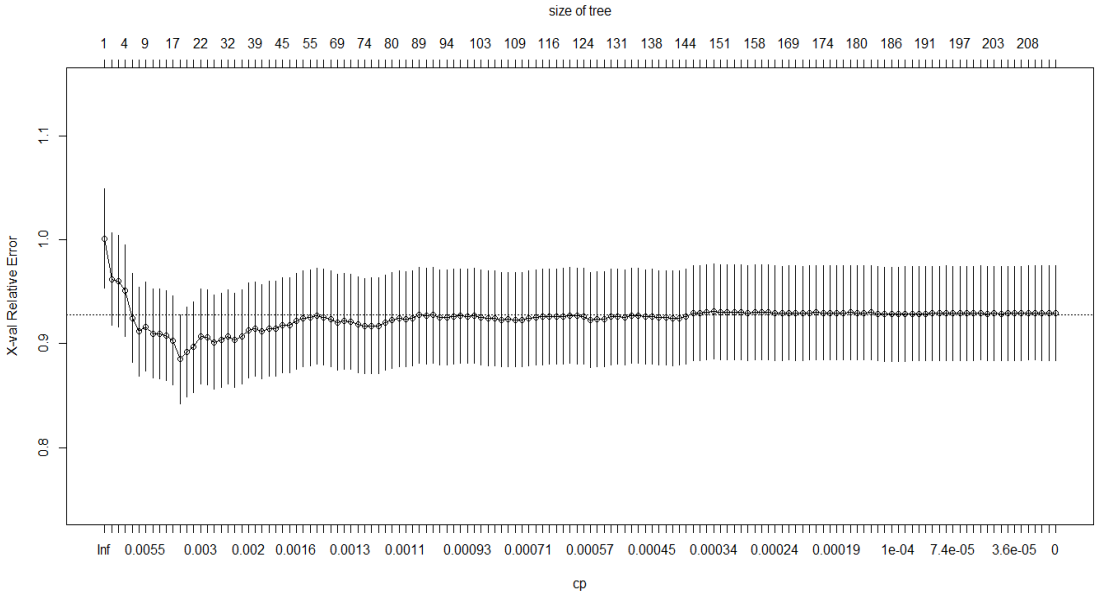


Gráfico 29-3. Generación de nodos del árbol base
Realizado por: Condo León José Luis, 2019

Como es posible notar en el Gráfico 29-3, la probabilidad de clasificación correcta del $Árbol_{base}$ cae drásticamente a partir del nodo 8, un análisis a profundidad demuestra que el árbol base necesitó generar 211 nodos o divisiones para lograr clasificar a todos los individuos del conjunto de datos; por otra parte, también se determinó que las variables influyentes que intervienen en la construcción del árbol base son: Clase de semilla, Deshierbe, Preparación de suelo, Problemas de sembrío, Rotación de cultivo, Superficie cosechada, Superficie sembrada, Uso de fertilizante orgánico, Uso de plaguicida orgánico y Uso de riego; Es importante destacar que las variables

que mayor peso tuvieron en la generación de nodos son Superficie sembrada, Superficie cosechada, Problemas de sembrío, Uso de riego y Clase de semilla.

La forma más común para determinar la validez del Árbol es con el uso del Error Cuadrático Medio (MSE), pues podrá determinar cuan fiables son las predicciones, de esta manera se define como:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(real_i - estimado_i)^2 \tag{3.3}$$

Donde n es el número de datos, $real_i$ es la medición del individuo i en la variable y $estimado_i$ es el valor estimado por el modelo para el individuo i.

Una breve validación del modelo del $Árbol_{base}$, usando el 90% de los datos como muestra de entrenamiento y un 10% como muestra de prueba determino el $MSE = 79{,}56$ y $RMSE = 8{,}92$, por lo que se puede deducir que con el uso de este modelo base del árbol se tendrá un error de estimación promedio de 9.92 libras de producción de arroz. Debido a que el $Árbol_{base}$ es sobre ajustado es necesario podarlo y determinar el mejor modelo con el uso de las variables influyentes correctas, para poder visualizar las verdaderas características de la base de datos.

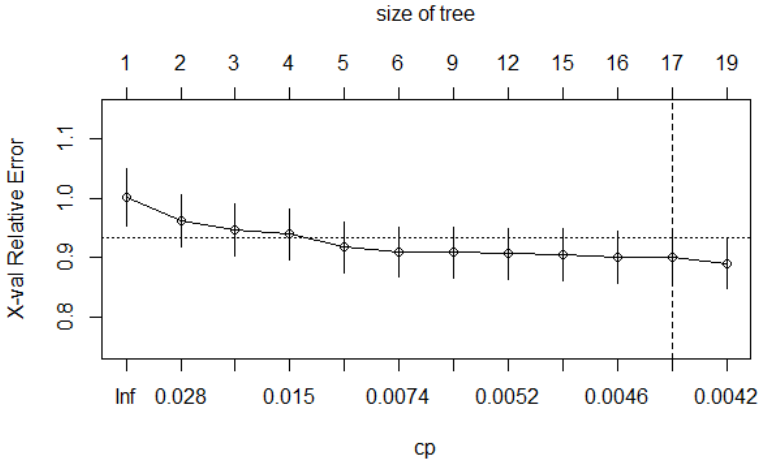


Gráfico 30-3. Generación de nodos del árbol podado
Realizado por: Condo León José Luis, 2019

3.4.2. Generación del modelo de árboles de regresión podado

Tras podar el árbol con parámetros $cp = 0.004$ y 17 nodos este se construye como:

$$\begin{aligned} Árbol_{podado} = \text{Produccion} \sim \text{Clase\_semilla} + \text{Deshierbe} + \text{Preparacion\_suelo} \\ + \text{Problemas\_sembrio} + \text{Superficie\_sembrada} \\ + \text{Superficie\_cosechada} + \text{Uso\_fertilizante\_organico} + \text{Uso\_riego} \end{aligned} \tag{4.3}$$

Las variables que más aporte tuvieron en la generación de los nodos en el árbol podado son:

Superficie sembrada, Superficie cosechada, Problemas de sembrío, Uso de riego, Uso de fertilizante orgánico y Clase de semilla; en el proceso de validación del modelo $Árbol_{podado}$ se obtuvo un $MSE = 77.79$ y $RMSE = 8.82$, lo que quiere decir que la poda no afecto en gran medida a la estimación del modelo original.

En búsqueda del modelo idóneo que represente de mejor manera la información se generó el modelo $Árbol_{defecto}$ que permite el software R usando con el método ANOVA y la construcción de éste se realizó como:

$$Árbol_{defecto} = \text{Produccion} \sim \text{Superficie\_sembrada} + \text{Superficie\_cosechada} + \text{Uso\_fertilizante\_organico} + \text{Uso\_riego} \quad (5.3)$$

Como es posible observar en el modelo $Árbol_{defecto}$ solo se vieron involucradas 4 variables predictoras para su construcción y la generación de los nodos está regida en función de las variables Uso de riego, Superficie cosechada, Superficie sembrada y Uso de fertilizante orgánico; en su validación, el valor de los indicadores $MSE = 73{,}78$ y $RMSE = 8.59$ lo que implica que el error de estimación promedio ha disminuido con respecto a los anteriores modelos.

3.4.3. Generación del modelo de árboles de regresión óptimo

Habiéndose generado 3 modelos de árboles y observando los resultados, fue posible determinar que existen 6 variables que están siendo determinantes tanto en la construcción como en la generación de nodos; por otro lado, existieron variables que no estaban aportando de manera significativa a la construcción de los modelos. Se utilizaron las 6 variables más influyentes en todos los modelos para construir un árbol óptimo y parsimonioso, y se presenta a continuación:

$$Árbol_{óptimo} = \text{Produccion} \sim \text{Superficie\_sembrada} + \text{Superficie\_cosechada} + \text{Uso\_fertilizante\_organico} + \text{Uso\_riego} + \text{Clase\_semilla} \quad (6.3)$$

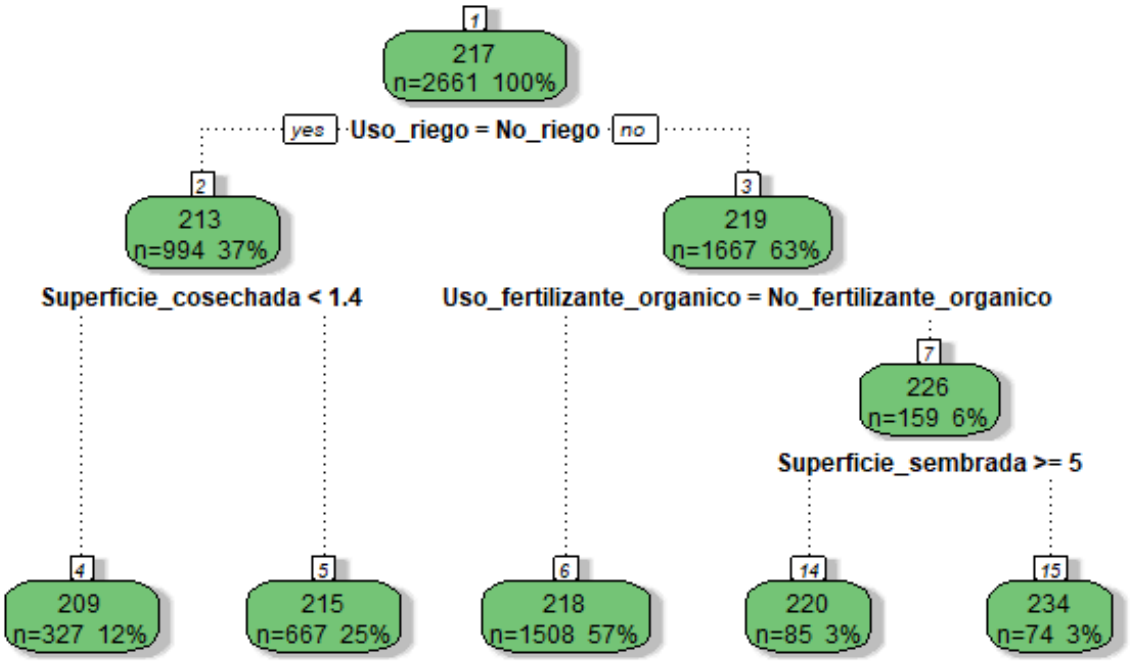


Gráfico 31-3. Árbol de regresión óptimo
Realizado por: Condo León José Luis, 2019

Tras generar el nuevo árbol con un valor MSE=53.58 y con un variación del RMSE=7.32 libras por hectárea se denomina a este como el mejor modelo pues se ajusta de mejor manera a la variabilidad de los datos, tal como se muestra en el Gráfico 31-3, las observaciones de la variable Producción de arroz en libras están determinadas principalmente por las variables Uso de riego, Superficie cosechada, Uso de fertilizante orgánico y Superficie sembrada, el modelo $Árbol_{óptimo}$ generó 4 nodos padres y 5 nodos terminales, definiendo Uso de riego como la variable que mejor separa al conjunto de información, pudiendo clasificar a los individuos en un 100%, lo que la convierte en el factor más influyente en la Producción de arroz; La Superficie cosechada y el Uso de fertilizante orgánico pueden considerarse como factores influyentes de segundo orden, ya que son ramificaciones del factor principal, la variación de sus valores y categorías puede determinar una diferente productividad de arroz. La superficie sembrada y la clase de semilla son factores influyentes tan solo para el grupo de sembríos que uso riego y fertilizantes orgánicos.

Según el $Árbol_{óptimo}$ se generó 4 principales reglas, con las que se puede determinar los niveles de producción de arroz, estas son:

La regla 1 se generó a partir de la variable Uso de riego, determinó que al no realizar el uso de riego el promedio de producción es de 213 libras de arroz, característica que posee el 37% de los individuos de la base de datos o 994 sembríos de arroz; por otro lado, los sembríos que realizaron el uso de riego obtuvieron una productividad de 219 libras de arroz en promedio para los 1667 casos que tienen ésta características, pudiendo establecer que con el uso de riego en los sembríos de arroz se puede aumentar en promedio 6 libras de arroz por cada hectárea de cultivo.

La segunda regla se generó a partir de la variable Superficie cosechada, la cual decreta que si en el cultivo no se realizó el Uso de riego y la Superficie cosechada es menor a 1.4 hectáreas, la producción en promedio será de 209 libras de arroz, el 12% de los cultivos analizados posee esta característica; mientras que por otro lado el 25% de los individuos tienen en promedio una producción de 215 libras de arroz, esto en razón a que no se realizó el uso de riego y la superficie sembrada es mayor o igual a 1.4 hectáreas de terreno; por tal razón se deduce que si la superficie cosechada es mayor a 1.4 hectáreas la producción se eleva 6 libras por cada hectárea.

Otra regla que generó el modelo es que cuando se hizo uso del riego y no se utilizó el fertilizante orgánico la productividad en promedio fue de 218 libras de arroz, esta característica se observa en 1508 sembríos, deduciendo que el 57% de los productores ecuatorianos que participo en la ESPAC 2017 hizo uso de riego, pero no aplico ningún fertilizante orgánico.

Por otro lado, todos aquellos productores que hicieron uso de riego y aplicaron algún tipo de fertilizante orgánico obtuvieron una producción en promedio de 226 libras de arroz, si a esto se añade que la superficie sembrada es mayor o igual a 5 hectáreas, la productividad de arroz se eleva en 8 libras por cada hectárea; si la superficie sembrada es menor a 5 hectáreas y tiene las características mencionadas anteriormente la productividad fue de 220 libras de arroz.

Con el uso de los árboles de regresión se determina que el factor más influyente en los niveles de producción de los sembríos de arroz es el uso de riego; como factores secundarios es posible encontrar la superficie de terreno cosechada y el uso de fertilizantes orgánicos que influyen de manera significativa; como es lógico una superficie sembrada amplia genera mayor productividad, lo que lo clasifica como un factor influyente importante.

3.4.4. Generación del modelo de bosques aleatorios

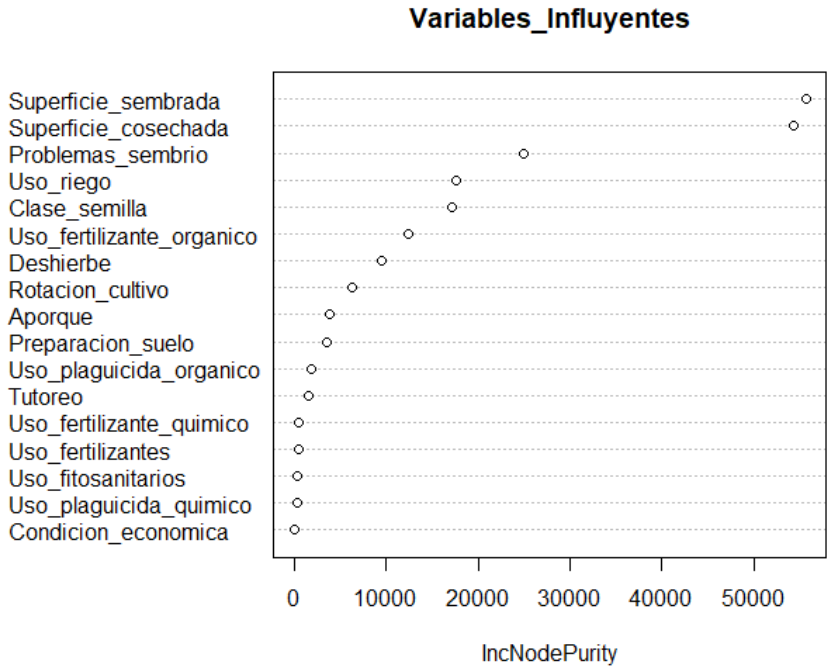


Gráfico 32-3. Variables influyentes en la producción de arroz, según Bosques Aleatorios
Realizado por: Condo León José Luis, 2019

Unas de las falencias en la aplicabilidad de los árboles de decisión o regresión es que los modelos pueden estar sobre ajustados y que arboles generados a partir de la misma base de datos resulten tener diferentes factores influyentes, con el fin de solucionar dicho inconveniente nace la metodología de Bosques Aleatorios cuyo principal objetivo es generar una cantidad n de árboles con la misma base de datos y determina que los factores que más se repiten en todos los árboles son los de mayor influencia.

De esta manera con el uso de la librería (randomForest), el comando varImpPlot() y la generación de 300 árboles aleatorios, se determina que las variables más influyentes en la productividad del arroz son Superficie sembrada, Superficie cosechada, Problemas de sembrío, Uso de riego, Clase de semilla y Uso de fertilizantes orgánico, tal como se muestra en el Gráfico 2-4; corroborando los resultados obtenidos con el uso de los árboles aleatorios.

3.5. Análisis factorial de datos mixtos

3.5.1. Generación del modelo de AFDM

El AFDM inicio con la creación del modelo usando las 18 variables, anterior a esto se preparó las variables de tipo cuantitativas realizando una estandarización sobre ellas, el primer indicador que se debe conocer es el total de la variabilidad explicada por el modelo, en el Gráfico 33-3 se detalla

que la primera componente tan solo explica el 13% de la variabilidad de los datos y la segunda componente explica el 9.1%.

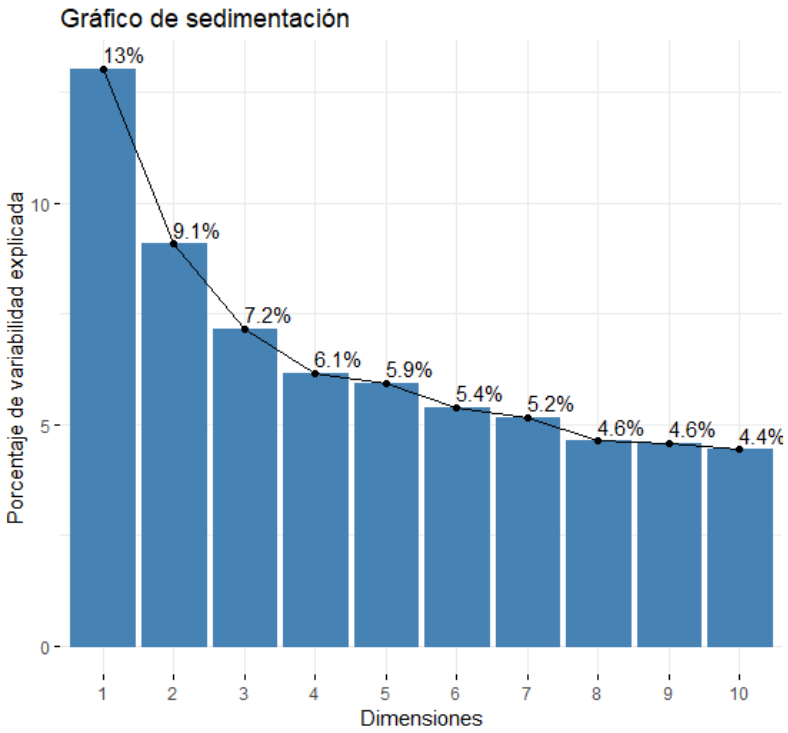


Gráfico 33-3. Gráfico de sedimentación de modelo base de AFDM
Realizado por: Condo León José Luis, 2019

Es posible notar que ninguno de las componentes explica de buena manera la variabilidad de la información, las 10 componentes en conjunto explican tan solo el 65.49% de la información, por lo que a simple vista se lo puede considerar un modelo no explicativo.

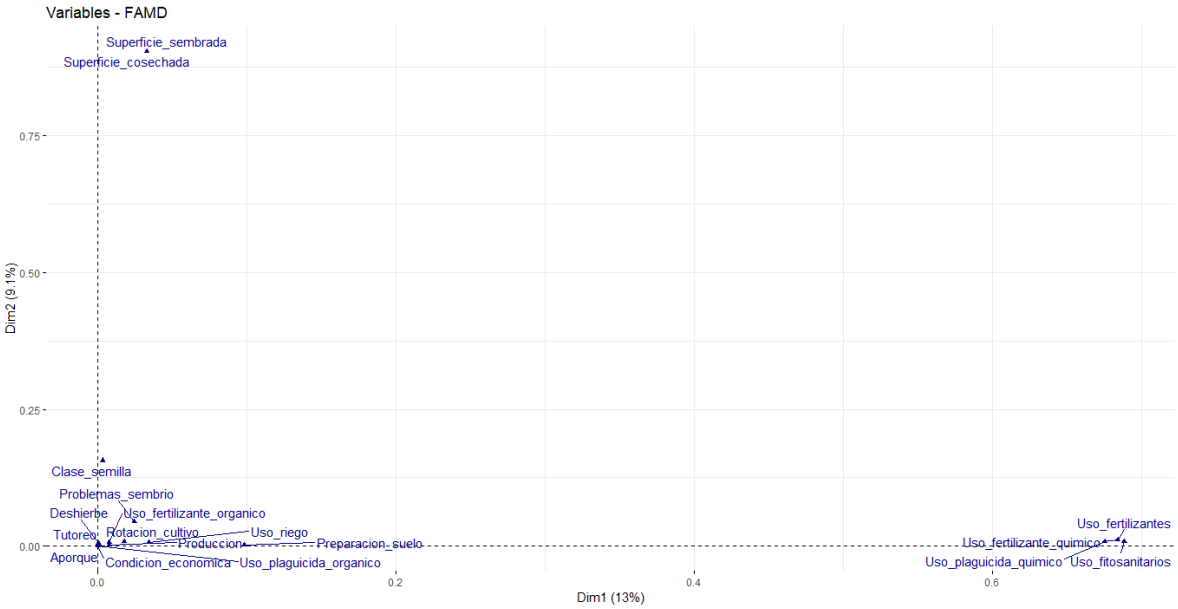


Gráfico 34-3. Correlación variables vs componentes de modelo base de AFDM
Realizado por: Condo León José Luis, 2019

Tal como se muestra en el Gráfico 34-3, las variables uso de fertilizantes, uso de fertilizante químico, uso de plaguicida químico y uso de fitosanitarios tienen una correlación muy alta con la primera componente; por otro lado, las variables superficie sembrada y superficie cosechada tienen una correlación elevada con la componente dos; el resto de las variables no están puntuando como significativas en la formación de las dos primeras componentes. Esta información es posible comprobarla en el Gráfico 35-3.

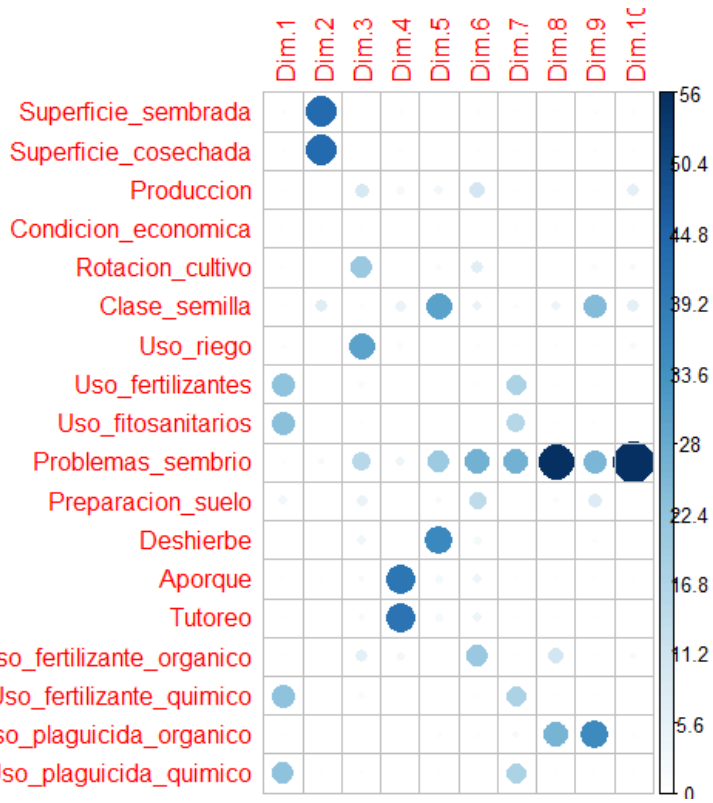


Gráfico 35-3. Niveles de correlación variables vs componentes de modelo base de AFDM

Realizado por: Condo León José Luis, 2019

El análisis de correlación expone de manera clara que la componente uno se relaciona de manera significativa con las variables uso de fertilizantes, uso de fitosanitarios, uso de fertilizante químico y uso de plaguicida químico; la componente 2 está altamente relacionada con las variables superficie sembrada y superficie cosechadas.

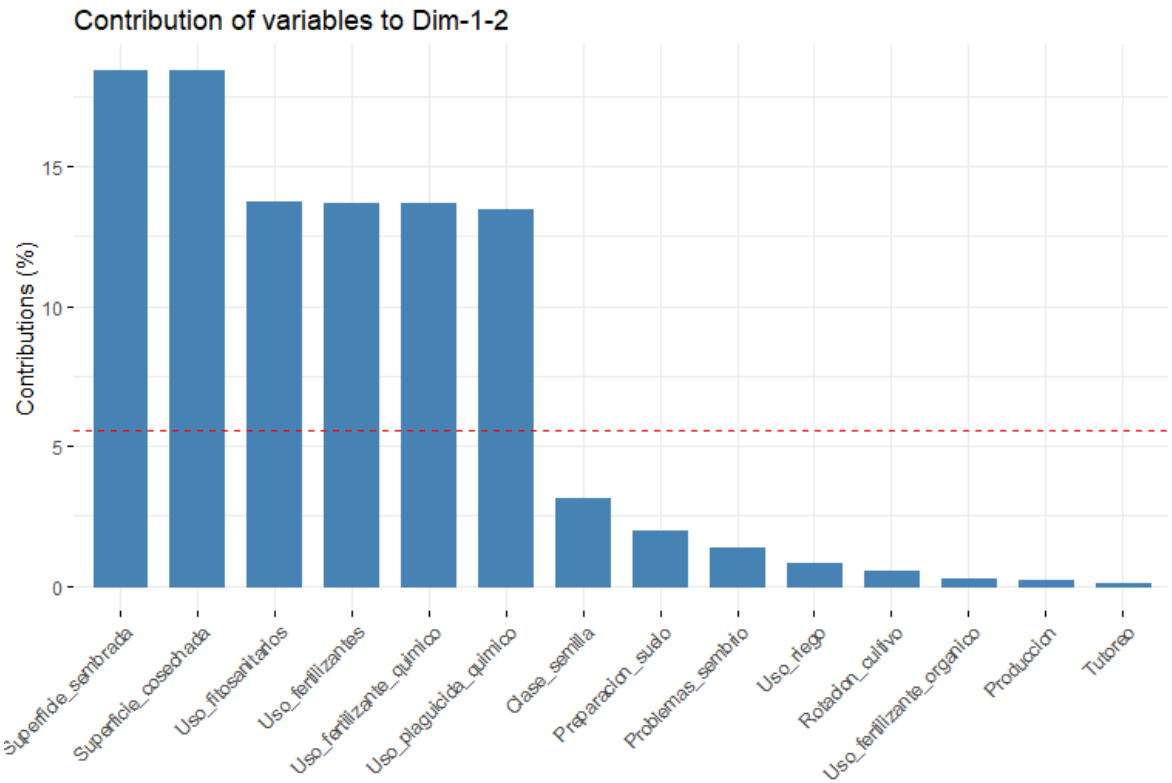


Gráfico 36-3. Contribución de variables a componentes de modelo base de AFDM
Realizado por: Condo León José Luis, 2019

Las variables superficie sembrada y superficie cosechada contribuyeron con un 19% cada una a la construcción de la componente 2, mientras que las variables uso de fitosanitarios, uso de fertilizantes, uso de fertilizante químico y uso de plaguicida químico contribuyeron con un 13% cada una a la creación de la componente 1.

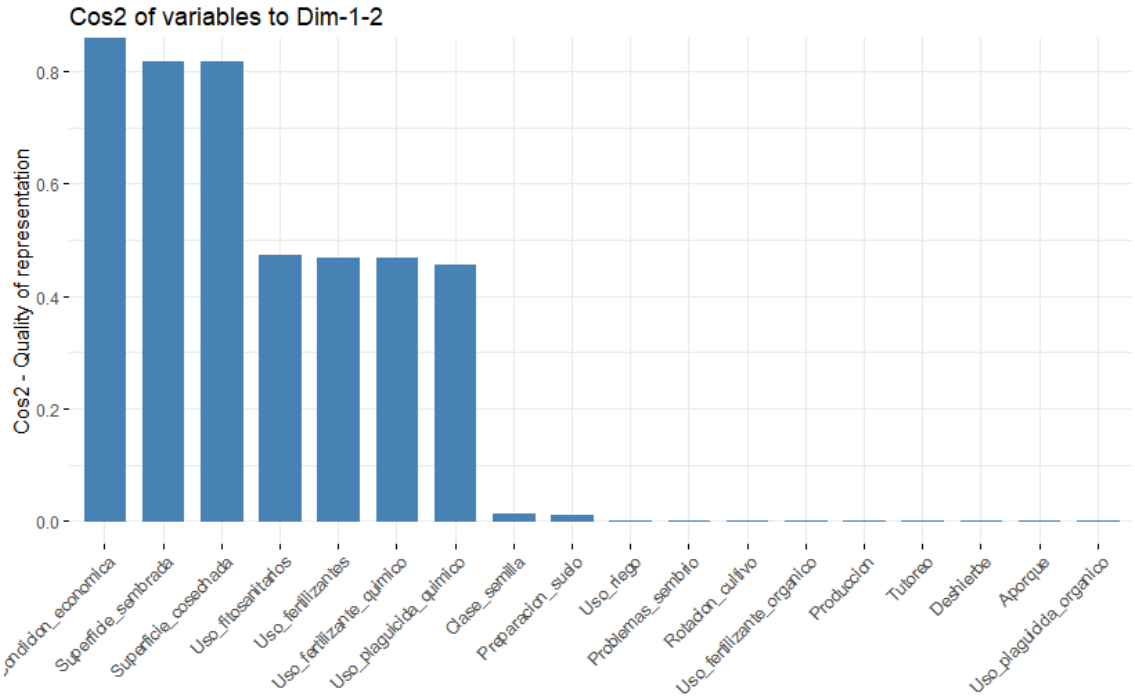


Gráfico 37-3. Representación de la información de variables en modelo base de AFDM
Realizado por: Condo León José Luis, 2019

Con valores cercanos a 1, la información de las variables condición económica del cultivo, superficie sembrada y superficie cosechada está bien representada dentro del modelo factorial, su influencia y variabilidad se ve reflejadas dentro de las componentes; caso contrario es lo que sucede con las variables uso de fitosanitarios, uso de fertilizantes, uso de fertilizante químico y uso de plaguicida químico, tan solo el 50% de su variabilidad se ver reflejada dentro del modelo factorial.

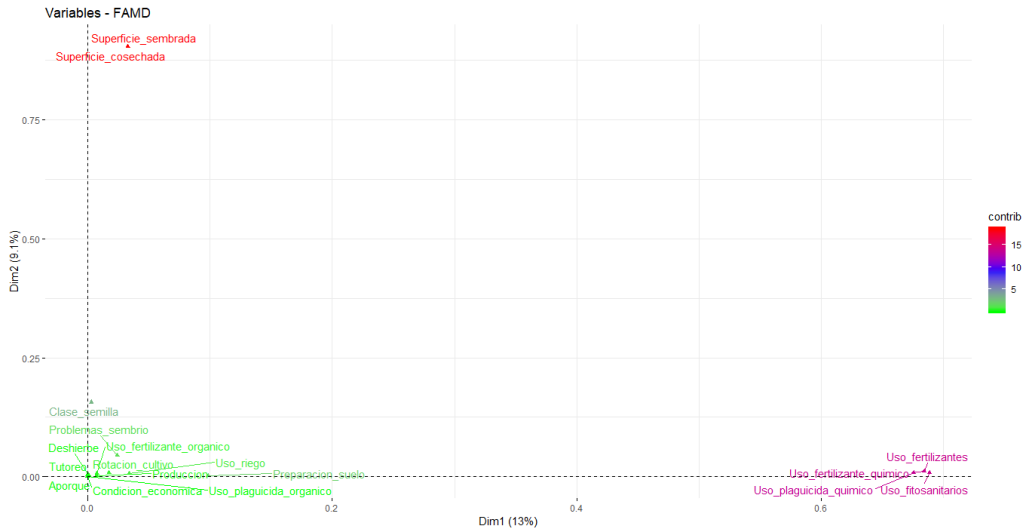


Gráfico 38-3. Correlación de variables vs componentes 1 y 2 de modelo base de AFDM
Realizado por: Condo León José Luis, 2019

Como se muestra en el Gráfico 38-3 es notoria la generación de 3 grupo de variables, la primera agrupación guarda estrecha relación con la componente 1 y está compuesta por la variable uso de fertilizante, uso de fertilizante químico, uso de plaguicida químico y uso de fitosanitarios; el segundo grupo de variables tiene una correlación alta con la componente 2 y está conformado por las variables superficie sembradas y superficie cosechada; el tercer grupo de variables no guarda relación con ninguna componente y no está aportando de manera significativa a la creación de estas.

Con el fin de conseguir un modelo que explique gran parte de la variabilidad de la información se decidió eliminar del modelo de AFDM a todas aquellas variables que no aportan a la generación de las dos componentes principales.

3.5.2. Generación del modelo óptimo de AFDM

El nuevo modelo esta generado con las variables uso de fertilizantes, uso de fertilizantes qguímicos, uso de plaguicida químico y uso de fitosanitarios como variables representativas de la componente 1; superficie sembrada y superficie cosechada denominadas como variables influyentes para la componente 2.

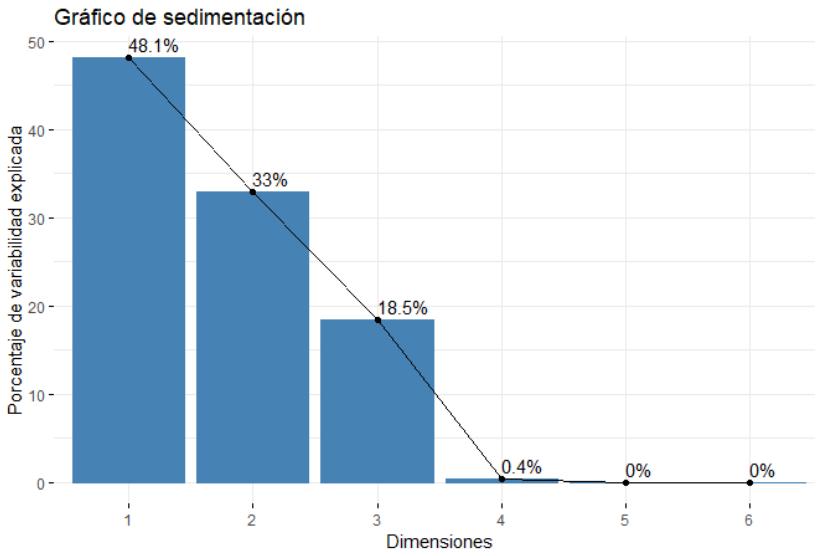


Gráfico 39-3. Gráfico de sedimentación de modelo óptimo de AFDM
Realizado por: Condo León José Luis, 2019

La exclusión de las variables no significativas permitió que la variabilidad de los datos explicada por el modelo óptimo mejore significativamente, con el uso de dos componentes este modelo logra explicar 81% de la información, tal como es posible observar en el Gráfico 39-3.

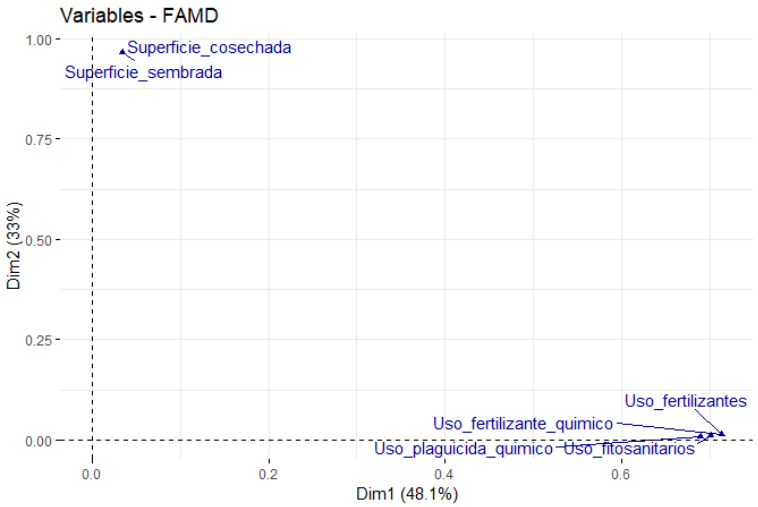


Gráfico 40-3. Correlación variables vs componentes 1 y 2 de modelo óptimo de AFDM

Realizado por: Condo León José Luis, 2019

Tal como era de esperar, el Gráfico 40-3 muestra que la componente 1 tiene alta correlación con las variables uso de fertilizantes, uso de fertilizantes químicos, uso de plaguicidas químicos y uso de fitosanitarios, por otra parte, la componente 2 tiene una relación marcada casi perfecta con las variables cuantitativas superficie sembrada y superficie cosechada; de tal manera es posible notar la creación de dos agrupaciones de variables influyentes en el conjunto de datos de producción de arroz.

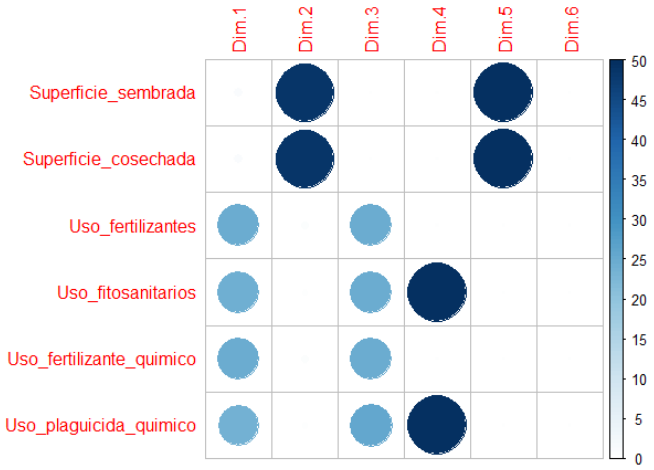
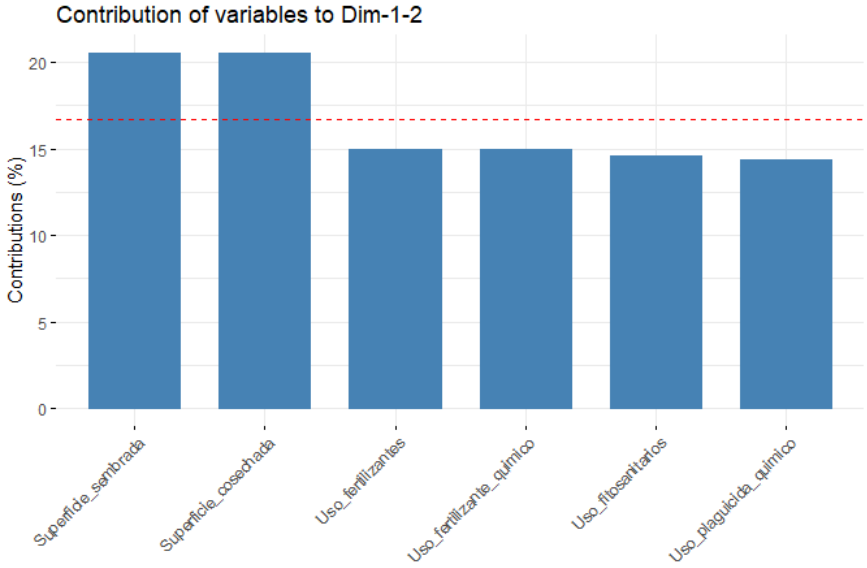


Gráfico 41-3. Correlación variables vs todas las componentes de modelo óptimo de AFDM

Realizado por: Condo León José Luis, 2019

Según muestra el Gráfico 40-3 y 41-3, la componente 2 tiene una relación muy marcada con las variables superficie sembrada y superficie cosechada, con una correlación igual a 0.96, por otro lado es notorio que las variables uso de fertilizantes, uso de fitosanitarios, uso de fertilizantes químicos y uso de plaguicidas químicos tienen un relación con la componentes uno pero no tan fuerte; respectivamente sus correlaciones son 0.71, 0.70, 0.71 y 0.68, cabe destacar que el modelo

óptimo explica el 100% de la variabilidad de la información con el uso de 6 componentes, la componente 3 se relaciona con el mismo grupo de que la componente 2, la componente 4 se relaciona con las variables uso de fitosanitarios y uso de plaguicida químico y la componente 5 se relaciona con las mismas variables que la componente 1.



Gráfica 42-3. Contribución de variables a componentes 1 y 2 de modelo óptimo de AFDM
Realizado por: Condo León José Luis, 2019

Las variables superficie sembrada y superficie cosechada se establecen como las variables más importantes en la formación de la componente uno, esto en razón a que cada una contribuye con el 20% a la generación de esta dimensión; la componente 2 determina su formación según las variables uso de fertilizantes, uso de fertilizantes químicos, uso de fitosanitarios y plaguicida químico, las primeras dos contribuyendo con el 15% y el resto contribuyendo con 14% a la formación de la componente.

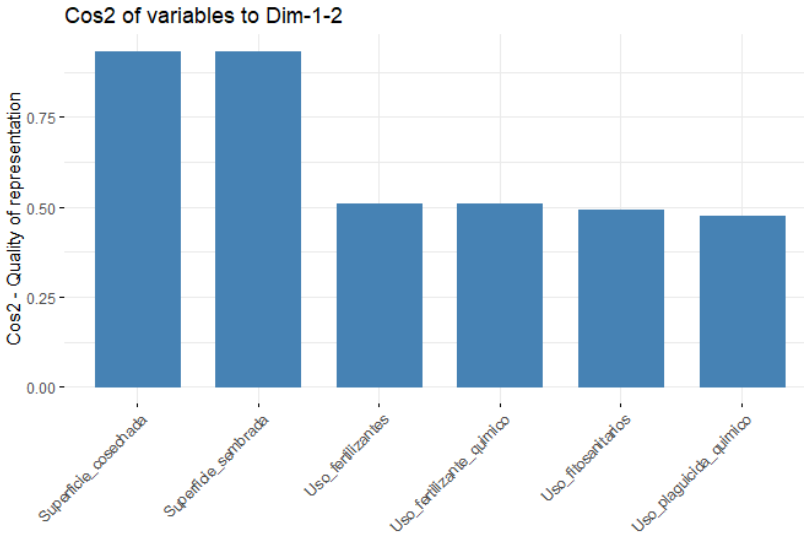


Gráfico 43-3. Representación de la información de variables en modelo óptimo de AFDM

Realizado por: Condo León José Luis, 2019

Como es posible observar en el Gráfico 43-3, el 93% de la información de las variables superficie cosechada y superficie sembrada está representada en el modelo óptimo de AFDM, en cambio, tan solo 50 % de la variabilidad de uso de fertilizantes, el 50% de la variabilidad de uso de fertilizantes químicos, el 49% de la variabilidad de uso de fitosanitarios y el 47% de la variabilidad de uso de plaguicidas químicos están representadas en el modelo factorial.

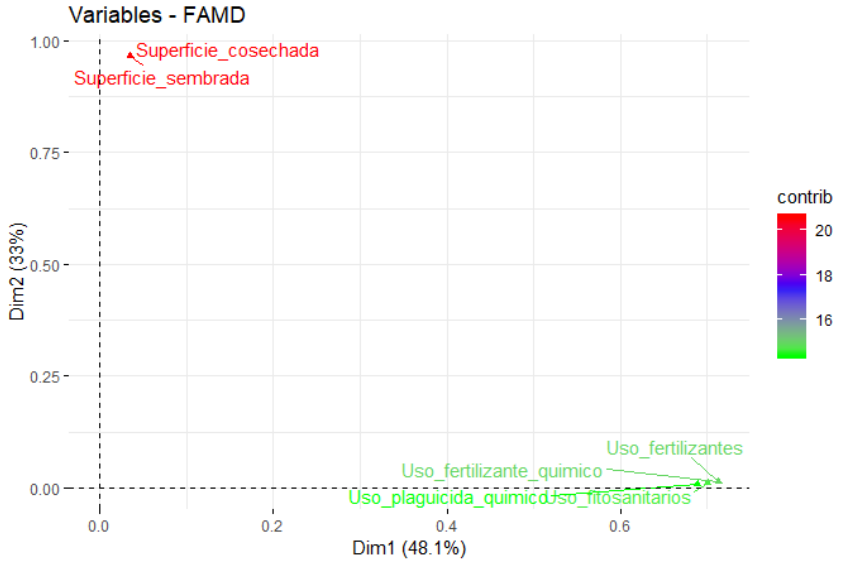


Gráfico 44-3. Nivel de relación variables vs componente 1 y 2 de modelo óptimo de AFDM

Realizado por: Condo León José Luis, 2019

Se muestra que la agrupación de variables influyentes en la componente uno tiene una relación relativamente fuerte y su contribución es alta, a su vez que las variables influyentes en la componente dos están estrechamente relacionadas con esta, siendo su contribución un tanto menor en comparación con las variables influyentes en la componente uno.

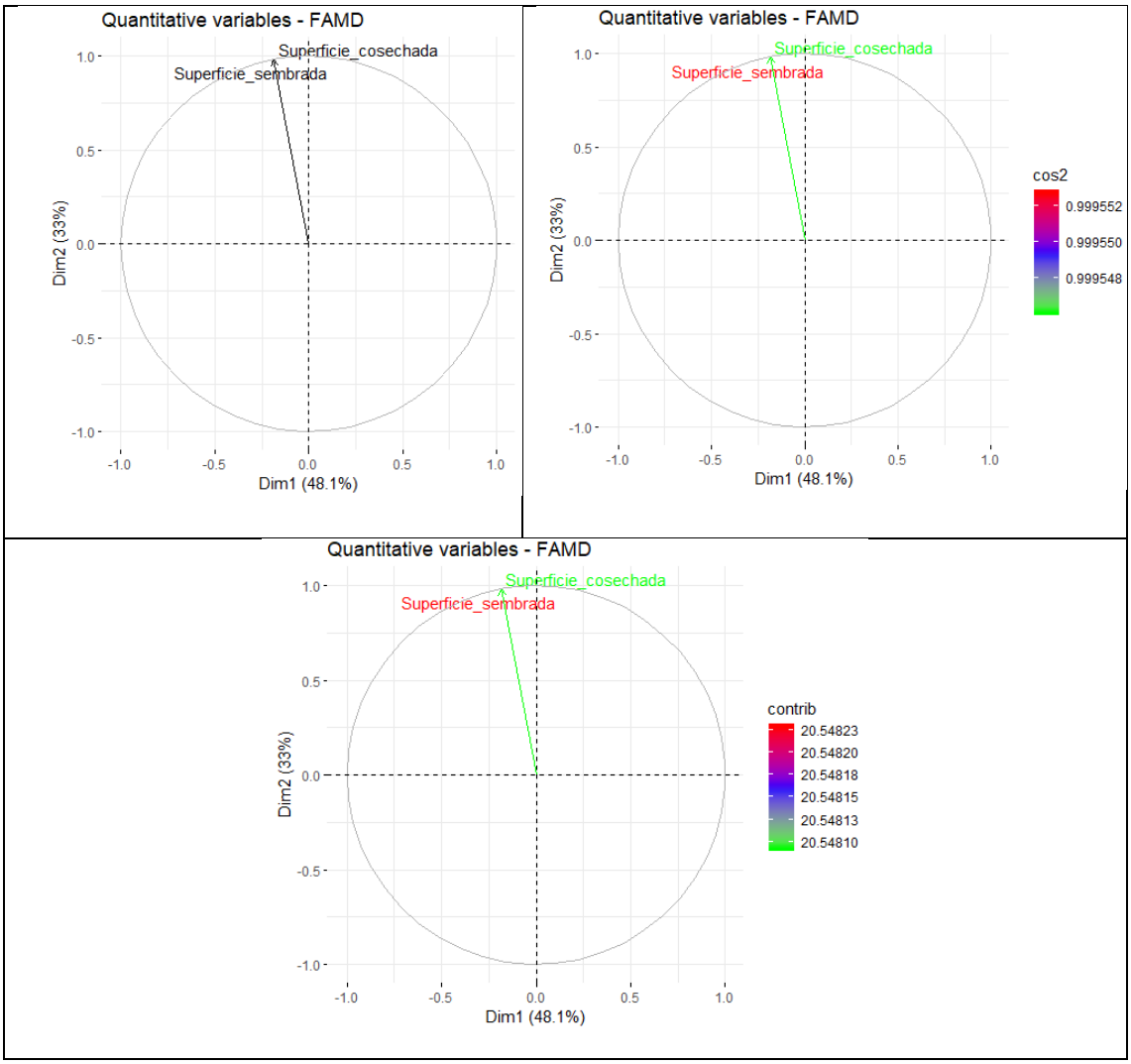


Gráfico 45-3. Representación de las variables cuantitativas en el plano factorial
Realizado por: Condo León José Luis, 2019

Las variables superficie sembrada y superficie cosechada puntúan muy alto en su relación con la componente 1, las dos tienen la misma dirección, por lo que se intuye que las mediciones de las dos variables cuantitativas son similares; cabe decir que la variables superficie sembrada tienen una representación ligeramente mayor por el modelo, lo que la hace que contribuya de mejor manera la formación de la componente uno; dentro de la información obtenida por la ESPAC 2017 se determina que las superficies sembradas con arroz no son iguales a las superficies cosechadas.

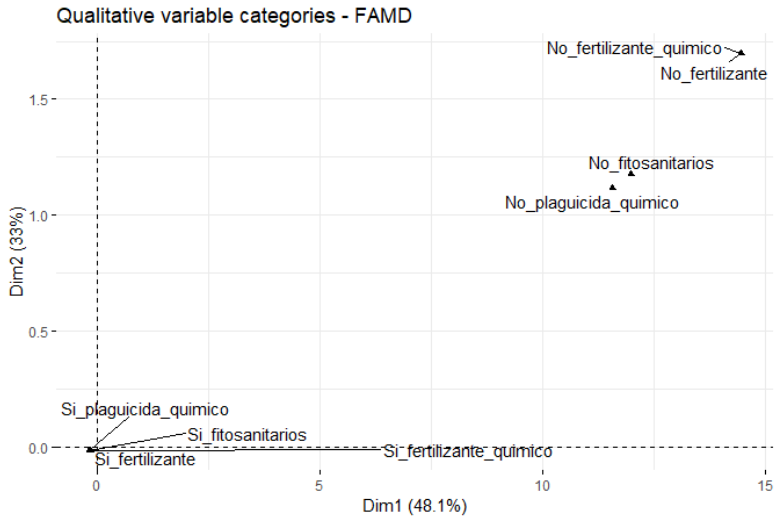


Gráfico 46-3. Representación de variables cualitativas en modelo óptimo de AFDM

Realizado por: Condo León José Luis, 2019

Es notoria la puntuación alta de ciertas categorías de las variables cualitativas en las dos componentes, se deduce según el Gráfico 46-3 que las mediciones de las variables superficie sembrada y superficie cosechada serán influenciadas al no usar fertilizantes, fertilizante químico, fitosanitarios o plaguicida químico.

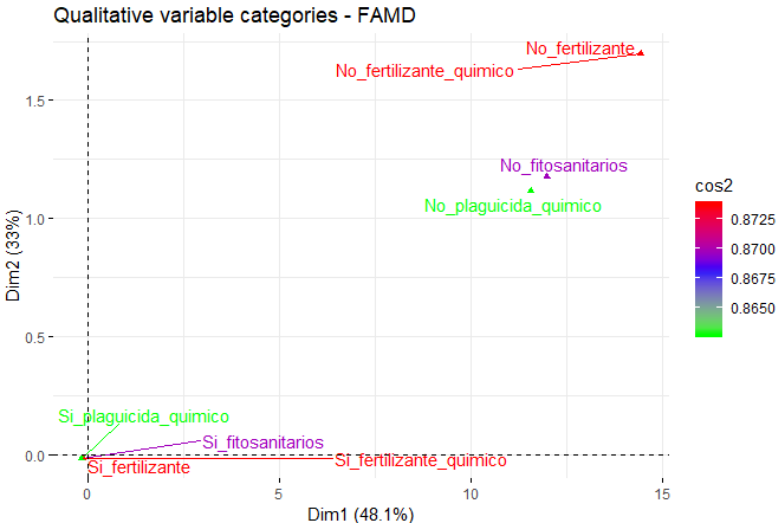


Gráfico 47-3. Niveles de representación de variables cualitativas modelo óptimo de AFDM

Realizado por: Condo León José Luis, 2019

Las variables uso de fertilizante, uso de fertilizante químico y sus categorías están teniendo más representatividad en el AFDM en comparación a las otras variables cualitativas; por otro lado, la que menos representatividad tiene en comparación con el resto es la variable uso de plaguicida químico, por lo que el orden de importancia de estas variables según su influencia en los datos es uso de fertilizante, uso de fertilizante químico, uso de fitosanitarios y uso de plaguicida químico.

El 85% de la variabilidad de la categoría no uso de fertilizante químico, el 85% de la variabilidad

no uso de fertilizante, el 59% de la variabilidad de la categoría no uso de fitosanitarios y el 56% de la variabilidad de la categoría no plaguicida químico se encuentra representadas en el plano factorial.

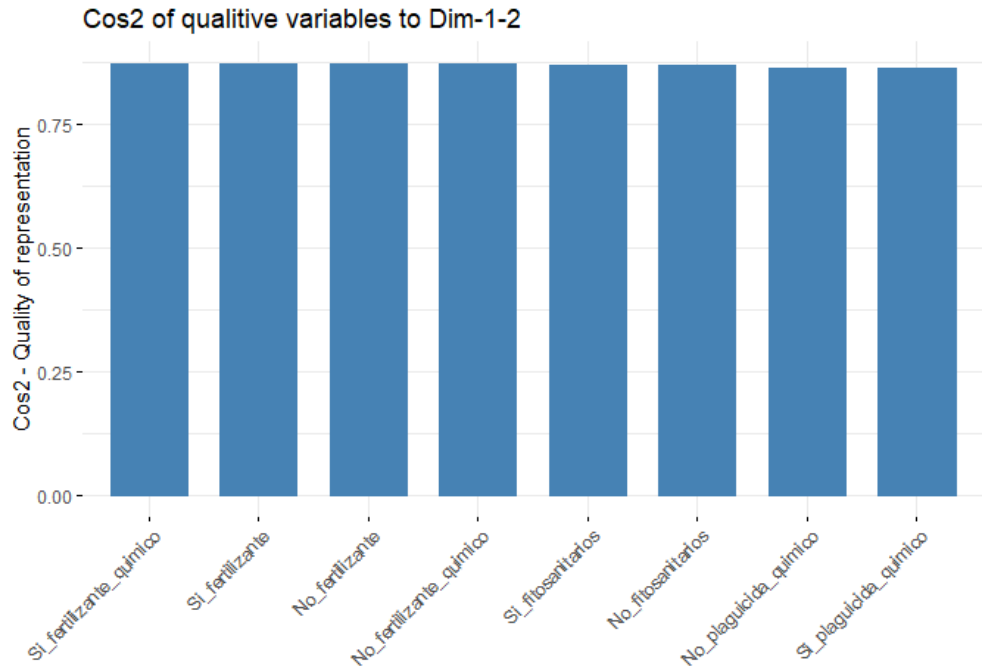


Gráfico 48-3. Representación de categorías de variables cualitativas modelo óptimo AFDM
Realizado por: Condo León José Luis, 2019

La información de las variables cualitativas se encuentra bien representada en el modelo, debido a que los valores de cos2 son iguales a 92%, pudiendo decir que la componente uno contiene la variabilidad de los individuos con categorías si uso de fertilizante químico, si uso de fertilizante, no uso de fertilizante y no uso de fertilizante químico; en cambio; la componente dos está recibiendo la variabilidad de los individuos con categorías si uso de fitosanitarios, no uso de fitosanitarios, no uso de plaguicida químico y si uso de plaguicida químico.

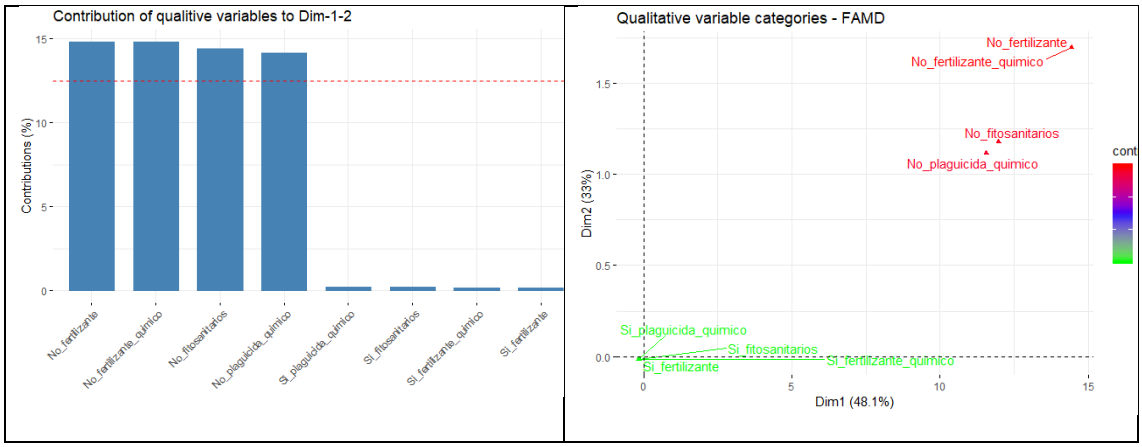


Gráfico 49-3. Contribución de categorías de variables cualitativas modelo óptimo AFDM
Realizado por: Condo León José Luis, 2019

Realizado el análisis en el Gráfico 49-3, es posible corroborar esa información en el Gráfico 48-3, esto en razón a que las categoría no uso de fertilizante, no uso de fertilizantes químicos, no uso de fitosanitarios y no uso de plaguicidas químicos contribuye con porcentajes menores a 15% cada una a la construcción de la componente 1; las categorías si uso de plaguicida químico, si uso de fitosanitarios, si uso de fertilizante químico y si uso de fertilizantes no contribuyen en gran medida a la formación de la componente 2.

3.6. Discusión de resultados

Como consecuencia de la industrialización y la mejora de las formas de producción de alimento se tienen diversas investigaciones que determinar que el factor más importante en una agricultura productiva y fértil es el riego, una infraestructura para el riego permite al agricultor obtener seguridad en el momento de la cosecha. (Klohn & Appelgren, 1999, págs. 105-126), factor que en la presente investigación también fue influyente con el uso de árboles de regresión.

En diversas investigaciones se ha demostrado que la manipulación de un factor en un proceso puede resultar en la obtención de distintos resultados en cierto factor, es así que una investigación realizada en Venezuela donde se utilizó semillas de arroz de un espécimen llamado centauro modificando los niveles de nitrógeno da como resultado que esta semilla junto con la modificación controlada de los niveles de nitrógeno en el crecimiento de esta planta permite mayores niveles de producción de arroz, también se pudo demostrar que el uso de componentes principales y en general el uso de análisis estadístico multivariado es de mucha importancia, pues permite representar la interacción de ambos factores en sus diferentes niveles (Acevedo Barona , y otros, 2011, pág. 192). Otra información importante de un estudio latinoamericano cuyo objetivo fue determinar la base genética de los cultivos de arroz, permitió determinar parentesco entre variedades de arroz comerciales liberadas desde el 2003 a 2014 y variedades ancestrales, en el transcurso de los años la semilla de este producto ha ido mutando teniendo a Venezuela como el

país con mayor parentesco entre estas dos variedades de semilla, y a Ecuador como el país con menor parentesco entre las semillas comerciales y semillas ancestrales, se debe tener en cuenta esta información pues puede ser un determinante en la producción de este sembrío. (Berrio Orozco , 2016), de acuerdo con los resultados obtenidos, la clase de semilla está siendo determinante para la producción de arroz en menor medida.

A nivel latinoamericano en concreto en Uruguay se trató con la teoría de que el potasio es un determinante de la baja producción de la planta de arroz, para determinar si esto es real o no se realizó sembríos a los cuales se les aplico cierta cantidad de potasio en la etapa de cultivo, lo que dio como resultado que el potasio no es un factor que disminuye la producción de arroz, al contrario, lo incrementa significativamente. (Ferrando , 2017, págs. 59-64), lo que concuerda pues el uso de fertilizantes es un factor importante en la producción de arroz.

El diseño experimental con el uso de varios factores en la producción arrocera es muy frecuente, es así que en la búsqueda del mejoramiento genético de los niveles de producción de arroz se evaluó características agronómicas, morfológicas y niveles de producción de la planta con técnicas estadísticas multivariadas, dando como resultados que la correcta aplicación e interpretación de las técnicas estadísticas permitirá el objetivo de mejorar la genética productiva de la planta de arroz, el uso de la técnica de componentes principales fue inútil pues no está representando de buena forma la variabilidad de la información que se posee, siendo el Análisis factorial discriminante una herramienta recomendable para este tipo de investigación pues permite el uso de muchas variables sin sesgo en los resultados (Morejón , Diaz , & Pérez , 2001, págs. 43-48)

Visto desde un punto de vista multivariante de la forma de producción del arroz y la cantidad producida por diferentes técnicas de agricultura es posible denotar que cada factor o forma distinta de tratamiento de esta planta influye en la cantidad que esta puede producir. Diversos estudios respaldan esta idea, tales como (Morejón & Díaz , 2017, págs. 115-121) quienes mediante un análisis de conglomerados lograron determinar las diferencias entre los niveles de producción de varias líneas genéticas de arroz, llegando a resultados como que las características físicas de la planta producida están muy correlacionadas entre sí y esta se ven directamente influidas por los factores de cultivo en especial los que guardan relación con los niveles de lluvia, también se determinó que la aplicación de técnicas estadísticas multivariadas aporta de manera significativa al mejoramiento genético de la planta de arroz.

El análisis de componentes principales y el análisis discriminante técnicas estadísticas multivariadas permite de cierta manera la clasificación de factores influyentes de un fenómeno,

usando estas técnicas se realizó una investigación en el año 1994 sobre la clasificación de los sistemas de producción agrícola en Venezuela, tomando como caso de estudio la producción de arroz, se estudiaron 41 variables de las cuales se obtuvieron 23 componentes que explican el 80% de la variabilidad de los datos, la primera componente se encuentra compuesta por variables relacionadas al tamaño, escala, dimensión o superficies de la finca, la segunda agrupa a variables referidas a fuentes de fertilización, la tercera componente agrupa a variables relacionadas a dosis de fertilización, la cuarta está compuesta por factores relacionadas a tipos de riego y limitaciones por malezas, la quinta está conformada por tipo de mano de obra y tipo de aplicación de agroquímicos, la sexta se conforma por enfermedades, las demás componentes se encuentran conformadas por variables que no están siendo significativas para el modelo, otros resultados confirman que la capacidad de riego es determinante en la producción de arroz, fue posible también discriminar cuatro grupos como bueno, mediano, regular y mal manejo en el cultivo de arroz y los factores de mayor incidencia en la producción de arroz son la dimensión, tamaño o escala de la finca, fuentes y dosis de fertilización y tipo de riego (Demey, Adams, & Freites, 1994, págs. 475-497)

Otra investigación realizada en Uruguay donde hacen uso de del análisis de conglomerados enfocándose en los factores de manejo agronómico y la distribución espacial de las propiedades del suelo donde se cultiva el arroz arrojó que existen dos conglomerados importantes respecto a los tipos de propiedades de cultivo de arroz, el primer grupo resalta la diferencia existente entre los niveles de arena y fósforo, el otro grupo está determinado por la implantación, la lámina de agua y el control de malezas, se demuestra una vez la importancia de la aplicabilidad del análisis estadístico multivariante, porque según el análisis realizado es posible tomar medidas sobre cierto grupo de propiedades para mejorar la producción de arroz (Bonilla , Terra , Gutierrez , & Roel , 2015, págs. 112-121)

Un análisis estadístico multivariado de la producción arrocera en el Ecuador en el año 2017 mediante componentes principales arrojo resultados como que los niveles de producción arrocera dependen de la superficie sembrada y cosechada en determinado periodo, otro importante factor influyente en los niveles de producción es el clima o los niveles de humedad siendo el invierno la época de mayor producción, también se generaron 3 componentes principales que contenían agrupaciones de variables en razón de procesos de producción, industrialización y comercialización y clima, otro resultado obtenido fue que las exportaciones de arroz tenderán a disminuir en los próximos 20 años (Montalvo Roca & Zurita Herrera , 2017, págs. 1-6)

Un reciente estudio en la provincia de Tungurahua en comunidades rurales determino sobre cultivos de papas que los productores aun consideran el conocimiento ancestral como guía para

la producción de este producto, demostrando que las costumbres de antaño aún están presentes en la producción de bienes de consumo humano, cabe recalcar que estas poblaciones ya utilizan ciertas características de producción actual como semillas mejoradas y este tipo de factores claramente afectan los niveles de producción de cualquier cultivo. (Pomboza Temaquiza, 2017, págs. 157-163)

Un gran ejemplo de la aplicabilidad del análisis estadístico en la productividad del arroz se ve reflejada en la investigación realizada en la provincia de Guayas, donde mediante el uso de diseño de experimentos estadísticos y ANOVA fue posible demostrar la igualdad en los niveles de producción de arroz cuando se usa los métodos tradicionales de agricultura y el Sistema Intensificado de Cultivo de Arroz(SICA), estos resultados contradicen investigaciones de otros países donde llegaron a la conclusión de que hubo incremento de los niveles de producción con el uso del SICA, por lo tanto es posible considerar que con el uso de técnicas estadísticas queda demostrado que la forma de cultivo tradicional del arroz y SICA arrojan mimo niveles de producción en la ciudad del Guayas (Ochoa Herrera , 2016, págs. 43-45)

Otra investigación realizada en la ciudad de Santiago de Cali en el año 2013 demostró que el género de la persona que cultiva el arroz está influyendo de manera directa en la alta o baja producción de esta planta, se muestra que las parcelas manejadas por hombres tienen mayor producción que las manejadas por mujeres, se dedujo que esto podría estar causado por posibles factores como el accesos a recursos de producción, a mercados y servicios de extensión, el uso intensivo de insumos y de mano de obra y la posesión de tierras y dimensión de las áreas cultivadas, la investigación concluyó que la producción llevada por mujeres tiene bajos niveles de uso de tecnología en gran medida a las pocas oportunidades con las que cuenta una mujer en esta área (Muriel Osorio , 2013, págs. 43-45)

Ciertas investigaciones enfocadas en el análisis de producción agrícola revelan resultados que debería tenerse en cuenta pues guardan relación directa con el tipo de cultivo aquí estudiado, tanto el plátano como el arroz son de los productos de mayor producción en Ecuador por lo que de cierta manera guardan cierta relación agrícola, es así que se realizó un análisis de los niveles de producción del plátano en el Ecuador usando el Análisis Factorial de Datos Mixtos(AFDM) obteniendo como resultados que los factores que influyen en los niveles de producción del plátano son el factor superficie y el factor denominado como factor uso y cuidado misma que no fue significativa en un modelo de regresión (Guamán Daquilema & Mullo Guaminga , 2018, págs. 59-60).

CONCLUSIONES

- Previo al análisis multivariante de datos, el análisis exploratorio de datos establece que en promedio la producción de arroz por hectárea es de 212.21 libras, siendo 130 libras la cantidad de producción más común entre los sembríos, el nivel de producción más bajo y alto son 25 y 300 libras respectivamente, también se determina que el 86.5% de la producción fluctúa entre 200 y 250 libras. Cabe destacar que en promedio la superficie sembrada es de 22.57 hectáreas y la superficie cosechada es de 22.19 hectáreas y que el 98.5% de las superficies sembradas y cosechadas es menor a 200 hectáreas. Otras características importantes en los sembríos de arroz son que el 61% de ellos si hacen uso de riego, el 93% no hace uso de fertilizante orgánico, el problema más común abarcando el 60% de los sembríos son las plagas y enfermedades, predomina el uso de la semilla de tipo común con 54%, el 98% de los sembríos usa cualquier tipo de fertilizante, el 97% de ellos hacen uso de fitosanitarios y el 99% hacen uso de fertilizantes y plaguicidas químicos. Por otro lado, se resalta que las variables superficie sembrada y superficie cosechada siguen una distribución asimétrica positiva por lo que se asume mayor acumulación de datos a la derecha de sus promedios, mientras que la variable producción de arroz sigue una distribución asimétrica negativa, también se puede denominar a las tres variables anteriores como leptocúrticas por tal razón sus datos presentan alto grado de concentración alrededor de la media.
- Los árboles de regresión con un modelo cuyo RMSE=7.32 libras por hectárea determina que el 57% de los cultivos hace uso de riego, pero no utiliza fertilizante orgánico, por otro lado 25% de los cultivos no hacen uso de riego y la superficie cosechada es superior a 1.4 hectáreas. Ésta técnica establece que la superficie sembrada, superficie cosechada, uso de riego, uso de fertilizante orgánico y clase de semilla son factores que influyen en gran medida sobre los niveles de producción, se denomina como principal factor el uso de riego del sembrío ya que al emplearlo la producción de arroz en promedio se eleva 6 libras por hectárea, cuando la superficie cosechada es superior o igual a 1.4 hectáreas la producción incrementa 6 libras por hectárea, si se hace uso de fertilizantes orgánicos la producción se eleva 8 libras por hectárea y si la superficie sembrada es superior a 5 hectáreas la producción subirá 14 libras por hectárea. Si se realiza el uso de riego, uso de fertilizante orgánico y si la superficie sembrada es superior a 5 hectáreas la producción se incrementa 25 libras por hectárea en comparación cuando no se usan.
- El análisis factorial de datos mixtos con un modelo que explica el 81% de la variabilidad de los datos usando dos componentes genera dos agrupaciones de factores influyentes, la primera componente está conformada por las variables: uso de fertilizantes, uso de

fertilizantes químicos, uso de plaguicidas químicos y uso de fitosanitarios, por otro lado, la segunda componente está conformada por las variables: superficie cosechada y superficie sembrada.

- Según el AFDM las variables uso de fertilizantes, uso de fertilizantes químicos, uso de plaguicidas químicos y uso de fitosanitarios tienen una dependencia alta entre ellas, debido a cada una de estas explica el 16% de la componente uno y los valores del cos2 para cada variable es igual a 87.25, 87, 86, y 86% respectivamente. Cabe destacar que las variables se relacionan mucho más cuando los individuos de estudio no usan fertilizantes, no usan fertilizantes químicos, no usan fitosanitarios y no usan plaguicidas químicos pues los niveles de contribución de estas categorías al modelo factorial son del 14%. Por otro lado, los árboles de regresión determinan que el uso de riego y el uso de fertilizantes orgánicos están estrechamente relacionados, ya que la primera variable explica el 100% de la variabilidad de los datos y la segunda nace en consecuencia de la primera explicando el 63% de la variabilidad de la información.
- Tanto el AFDM y los árboles de regresión coinciden en que las variables cuantitativas superficie sembrada y superficie cosechada son factores influyentes en los niveles de producción de arroz. Por otro lado, solo la variable uso de fertilizante orgánico coincide como un factor importante en las dos técnicas, la variable clase de semilla se refleja como un factor, pero no contribuye en gran medida a ninguno de los modelos analizados.

RECOMENDACIONES

- Es importante mejorar los sistemas de regadío usados en la producción de arroz pues es el factor más relevante en la productividad de ésta gramínea.
- Es necesario mejorar las políticas de adquisición y distribución de fitosanitarios a nivel nacional para que así los agricultores tengan las herramientas necesarias para obtener una buna productividad en sus sembríos.
- Ampliar esta investigación con el uso de Bosques aleatorios y Redes neuronales para así eliminar las desventajas que presenta el uso de los Árboles de regresión.
- Socializar los resultados con los productores de arroz en el Ecuador para que puedan tomar medidas sobre sus sembríos en base a esta investigación.
- Debido a la gran utilidad que tienen las técnicas de minería de datos en la resolución de problemas y presentando ventajas frente a las técnicas estadísticas clásicas es recomendable la enseñanza de éstas en la carrera de Ingeniería Estadística en la ESPOCH.
- Es necesario enfocar nuevas medidas de control de calidad para establecer una mejor recolección de información dentro del INEC, para así mejorar la calidad de la información de la ESPAC en futuras ediciones.

GLOSARIO

ACP:	Análisis de componentes principales
AED:	Análisis exploratorio de datos
AFDM:	Análisis factorial de datos mixtos
CFN:	Corporación Financiera Nacional
ESPAC:	Encuesta de superficie y producción agropecuaria
FAO:	Organización de las Naciones Unidas para la Agricultura y la Alimentación
INEC:	Instituto Nacional de Estadísticas y Censo
MAGAP:	Ministerio de Agricultura, Ganadería, Acuicultura y Pesca del Ecuador
PIB:	Producto interno bruto
SICA:	Sistema intensificado de cultivo de arroz
SIPA:	Sistema de Información Pública Agropecuaria

BIBLIOGRAFÍA

ABASCAL FERNANDEZ, & LANDALUCE CALVO. "Análisis factorial múltiple como técnicas de estudio de la estabilidad de los resultados de un análisis de componentes principales". *Questiio,* 26, 2002, 109-122, Consulta:5-05-2019, Disponible en: https://upcommons.upc.edu/handle/2099/4175

ABELLANA SANGRA, R., & FARRAN CODINA, A. "Identificación, impacto y tratamiento de datos perdidos y atípicos en epidemiología nutricional", *Revista Española de Nutrición Comunitaria*, 21, 2015, 188-194, Consulta:10-06-2019, ISSN 1135-3074, Disponible en: http://www.renc.es/imagenes/auxiliar/files/RENC2015supl1MISSING.pdf.

ACEVEDO BARONA , M., SALAZAR , M., CASTRILLO FUENTES , W., TORRES ANGARITA, O., REYES RAMONE , E., NAVAS , M., . . . TORRES TORO , E. "Efecto de la densidad de siembra y fertilización nitrogenada sobre el rendimiento de granos de arroz del cultivar centauro en Venezuela", *Agronomía tropical*, 2011, 61, 30-35, Consulta: 12-10-2019, Disponible en: http://www.scielo.org.ve/scielo.php?script=sci_arttext&pid=S0002-192X2011000100002

AMÓN URIBE , I., & JIMÉNEZ RAMÍREZ, C. "Hacia una metodología para la selección de técnicas de depuración de datos", *Avances en Sistemas e Informática*, 2015, 6, 186-190, Consulta: 05-05-2019, Disponible en: http://www.redalyc.org/html/1331/133112608019/

BACALLAO GUERRA , J., & BACALLAO GALLESTEY , J. "Imputación múltiple en variables categóricas usando Data Argumentation y Árboles de Clasificación", *Revista Investigación Operacional,* 2010, Cuba, 31, 133-139, Consulta: 05-05-2019, Disponible en: https://biblat.unam.mx/es/revista/investigacion-operacional/articulo/imputacion-multiple-en-variables-categoricas-usando-data-augmentation-y-arboles-de-clasificacion

BATANERO, C., GODINO , J., & ESTEPA , A. "Análisis exploratorio de datos: Sus posibilidades en la enseñanza Secundaria", *Suma*, Enero de 1991, 9, 25-31, Consulta: 05-05-2019, Disponible en: http://www.ugr.es/~batanero/pages/ARTICULOS/anaexplora.pdf

BERLANGA SILVENTE , V., RUBIO HURTADO, M., & VILA BAÑOS , R. "Cómo aplicar árboles de decisión en SPSS", *REIRE*, (2013), 6(1), 65-79, Consulta: 03-05-2019, Disponible en: http://diposit.ub.edu/dspace/bitstream/2445/43762/1/618361.pdf

BERRIO OROZCO , L. "Diversidad genética de las variedades de arroz FLAR libradas entre 2003 - 2014". *AGron Mesoam* , (2016), 217-231., Consulta: 01-09-2019, Disponible en:

http://www.scielo.sa.cr/pdf/am/v27n2/1021-7444-am-27-02-00217.pdf

BISANG , R. "Apertura económica, innovación y estructura productiva: la aplicación de biotecnología**",** *Desarrollo Económico* , (2003), 43(171), 413-442, Consulta: 01-04-2019, Disponible en: www.jstor.org/stable/3455892

BONILLA , C., TERRA , J., GUTIERREZ , L., & ROEL , Á. "Cosechando los beneficios de la agricultura de precisión en un cultivo de arroz en Uruguay", *Agrociencia Uruguay,* (6 de Febrero de 2015), 112-121, Consulta: 16-04-2019, Disponible en: http://www.scielo.edu.uy/scielo.php?script=sci_arttext&pid=S2301-15482015000100013

BONILLA BOLAÑOS, A. G., & SINGAÑA TAPIA , D. A. "La productividad agrícola más allá del rendimiento por hectárea: Análisis de los cultivos de arroz y maíz duro en Ecuador, *La granja*", (2019), Ecuador, 70-83, Consulta 5-12-2019, Disponible en: http://scielo.senescyt.gob.ec/pdf/lgr/v29n1/1390-3799-lgr-29-01-00070.pdf

BOUZA HERRA C. N. , *Modelos de regresión y sus* aplicaciones, Libro, Cuba, (2018), Consulta: 17-6-2019, Disponible en: https://www.researchgate.net/profile/Carlos_Bouza/publication/323227561_MODELOS_DE_REGRESION_Y_SUS_APLICACIONES/links/5a871265a6fdcc6b1a3abe40/MODELOS-DE-REGRESION-Y-SUS-APLICACIONES.pdf

CARDONA HERNÁNDEZ, P. "Aplicación de árboles de decision en modelos de riesgo crediticio", *Revista Colombiana de Estadística* , (2004), Colombia, 27(2), 139-151, Consulta: 2-2-2019, Disponible en: http://emis.ams.org/journals/RCE/ingles/V27/V27_2_139Cardona.pdf

CASTRO, M. (2017). "Rendimiento de arroz con cáscara, primer cuatrimestre 2017". Quito: Ministerio de agricultura y ganaderia, Consulta:5-05-2019, Recuperado de: http://sipa.agricultura.gob.ec/descargas/estudios/rendimientos/arroz/rendimiento_arroz_primer_cuatrimestre_2017.pdf

CHACÓN, P. (2010). "Cultivo de pastos", Consulta:5-05-2019, Recuperado de: https://www.swisscontact.org/fileadmin/user_upload/COUNTRIES/Peru/Documents/Publications/MANUAL_PASTOS_CULTIVADOS.pdf

CHONCHOL, J. "Agricultura, Alimentación y Eenergía" , *Fondo de Cultura Económica* , (1983), 189-206, Consulta: 3-9-2019, Disponible en: www.jstor.org/stable/23395635

CORPORACIÓN FINANCIERA NACIONAL. (2018). "Cultivo de Arroz, Molienda o Pilado de arroz", Consulta:5-05-2019, Recuperado de: https://www.cfn.fin.ec/wp-content/uploads/2018/04/Ficha-Sectorial-Arroz.pdf

CORREA, J. C., & SALAZAR , J. C. "¿Qué es Projection Pursuit? ", *Revista Estadística de Colombia,* Consulta:5-05-2019, (1997).

COUSINEAU, D., & CHARTIER, S. "Outliers detection and treatment: a review". *International Journal of Psychological Research*, (2010). , 58-67.

CUADRAS C. "Nuevos métodos de análisis multivariante*"*. 30 ed. . Barcelona: CMC Editions.

DELGADO, F. "El arroz. Quito: Departamento de arroz Ecuaquimica", [Consulta: 12 octubre 2019]. Disponible en https://www.ecuaquimica.com.ec/info_tecnica_arroz.pdf

DEMEY, R., ADAMS, M., & FREITES, H. "Uso del método de análisis de componentes principales para la caracterización de fincas agropecuarias". *Agronomía tropical*, 475-497. [Consulta: 12 octubre 2019]. Disponible en: http://sian.inia.gob.ve/revistas_ci/Agronomia%20Tropical/at4403/Arti/demey_j.htm

SEPÚLVEDA, J. "Comparación entre Árboles de regresión CART y Regresión lineal". (2012) Universidad de Colombia, 2-7.

ESCOBAR, G., & BERDEGUÉ, J. "Tipificación de sistemas de producción agricola". (1990) Santiago de Chile. [Consulta: 12 octubre 2019]. Disponible en: https://idl-bnc-idrc.dspacedirect.org/bitstream/handle/10625/3969/49675.pdf?sequence=1#page=11

FERNANDEZ AGUIRRE, K. "Nuevo procedimiento metologógico para el análisis exploratorio de una tabla estructurada en diversos conjuntos de individuos". *Estadística española*, (2013) pp. 305-322. [Consulta: 12 octubre 2019]. Disponible en: file:///C:/Users/Jose/Downloads/182-3.pdf

FERRANDO, M. "Respuesta del arroz a la fertilización potásica en el sistema uruguayo de manejo inundado". *Agrociencia* (2017) Uruguay, 59-64. [Consulta: 12 octubre 2019], Disponible en http://www.scielo.edu.uy/scielo.php?script=sci_arttext&pid=S2301-15482017000200059&lang=pt

Filzmoser, P. A "Multivariate Outlier Detection Method", Department of Statistics and Probability Theory.

Franco, D. C., & Melo, O. O. "Estimación de información faltante, imputación y estadísticos de prueba en modelos mixtos a dos vías de clasificación". (Junio de 2006)., *Revista Colombiana de Estadística*, 29(1), 35-56.

FRANCO , W. "Biodiversidad productiva y asociada en el Valle Interandino Norte del Ecuador", *Bioagro*, (2016)., 182-192. [Consulta: 12 octubre 2019] Disponible en http://www.scielo.org.ve/pdf/ba/v28n3/art05.pdf

FRIEDMAN , J. H., & STUETZLE, W., "Projection Pursuit for Data Analysis", *Modern Data Analys,* (1982), 123-147.

Gonzáles, V. M., "Análisis de los espacion de representación del STTATIS y del AFM en el estudio de movilidad biográfica en Bogotá 1993", *Revista colombiana de estadística*, (2009), 32(1), 1-15. [Consulta: 12 octubre 2019]

GUAMÁN DAQUILEMA , S. E., & MULLO GUAMINGA , H. S. "Análisis estadístico multivariante para el estudio de los factores que influyen en la producción del plátano en el Ecuador, Periodo 2014 -2016". Escuela Superior Politécnica de Chimborazo , (2018). Riobamba:. Disponible en http://dspace.espoch.edu.ec/bitstream/123456789/8969/1/226T0043.pdf

HARALD CRAMÉR , C. , (Noviembe de 2018). Universidad de la república de Uruguay. Reuperado de http://eva.fcea.edu.uy/pluginfile.php/109030/mod_resource/content/0/stcap13_atipicosPena.pdf

HAWKINS, D. M. CHAPMAN & HALL, "Identification of outliers" (1980). *Londo, 2017* . Disponible en https://www.springer.com/la/book/9789401539968

HERNÁNDEZ SAMPIERI, R., & BAPTISTA LUCIO, P. (2014). "Metodología de la investigación". (McGRAW-HILL, Ed.) México, México, México: Mc Graw HILL. Recuperado de http://observatorio.epacartagena.gov.co/wp-content/uploads/2017/08/metodologia-de-la-investigacion-sexta-edicion.compressed.pdf

INEC. (2016). "Encuesta de Superficie y Producción Agropecuaria Continua": Metodología. Recuperado de https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_agropecuarias/espac/espac-2018/Metodologia%20de%20la%20operacion%20estadistica%20ESPAC%202018.pdf

INEC. (2017). "Encuesta de Superficie y Producción Agropecuaria Continua 2017". Instituto Nacional de Estadístias y Censos, Quito. Disponible en http://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_agropecuarias/espac/espac_2017/Informe_Ejecutivo_ESPAC_2017.pdf

JAUREGUIZAR, E. O. "Evolución de mamiferos cenozoicos sudamericanos: Un estudio basado en técnicas de análisis multivariado". *Comisión de investigación científica de buenos aires*, (1990), 191-207. Disponible en https://www.researchgate.net/profile/Edgardo_Ortiz-Jaureguizar/publication/230600193_Evolucion_de_las_comunidades_de_mamiferos_cenozoicos_sudamericanos_un_estudio_basado_en_tecnicas_de_analisis_multivariado/links/0912f501c928b677cb000000/Evolucion-de-las-

KLOHN, W., & APPELGREN, B. "Agua y Agricultura". *CIDOB d'Afers Internacionals*, (1999). 105-126. Disponible en http://www.jstor.org/stable/40586152

LEYS , C., KLEIN , O., & DOMINICY, Y.. "Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance", *Journal of Experimental Social Psychology* , (2017), Belgium,

LIND , D., MARCHAL , W., & WATHEN , S. "Estadística aplicada a los negocios y la economia", Mc *Graw Hill* , 2012, México, (15 ed.), Disponible en https://www.academia.edu/16035082/Estadistica_aplicada_a_los_negocios_y_la_economia_15_edicion?auto=download

LOZARES COLINA , C., & LÓPEZ ROLDÁN , P. "El análisis Multivariado: Definición, Criterios y Clasificación". *Revista de Sociología*, (1991). 9-29. Disponible en https://ddd.uab.cat/pub/papers/02102862n37/02102862n37p9.pdf

MANOJ , K., & SENTHAMARAI KANNAN , K. "Outlier Detection in Multivariate Data". *Applied Mathematical Sciences*, (2015). , 9(47), 2317-2324. Disponible en http://dx.doi.org/10.12988/ams.2015.53213

MÁRQUEZ PÉREZ, V., USECHE , L., MESA , D., & IDÉS CHACÓN, A.." Estrategia de imputación con la media bajo el uso de árboles de regresión". *Comunicaciones en Estadística*, (Junio de 2017) 10 (1), 9-40.

MARTINEZ FLORES , L. "AFitomejoramiento y racionalidad social: los efectos no intencionales de la liberacion de una semilla lupino en Ecuador. Paralelos", (2016). 71-91. Disponible en http://www.scielo.org.co/pdf/antpo/n26/n26a04.pdf

MEDINA, F., & GALVÁN , M. "Imputación de datos faltantes", *CEPAL*, (2007). Santiago de Chile:.

Mendoza Vega , J. B. (23 de Abril de 2018). Rpubs. Disponible en https://rpubs.com/jboscomendoza/arboles_decision_clasificacion

MISZTAL , M. (2013). "Some remarks on the Data Imputation using "MissForest" Method". Lodz: University of Lodz

MONTALVO ROCA , D., & ZURITA HERRERA , G. "Análisis estadístico de la producción arrocera en Ecuador". (2017). Guayaquil: ESPOL. Disponible en http://www.dspace.espol.edu.ec/xmlui/bitstream/handle/123456789/41208/D-71681.pdf?sequence=-1&isAllowed=y

MONTES DE OCA , L. R., GARCIA PEREIRA, A., & HERNÁNDEZ GÓMEZ , A. "Uso de técnicas de análisis multivariable aplicadas en la obtención de modelos de predicción de propiedades realcionadas con los sistemas agrícolas". *Ciencias técnicas agropecuarias*, (2009), 74-77. Disponible en http://www.redalyc.org/pdf/932/93215937014.pdf

MOREJÓN , R., & DÍAZ , S. "Aplicación web SISDAM y técnicas estadística multivariadas en la selección de líeas de arroz en los Palacios". *Cultivos Tropicales,* . (Enero - Marzo de 2017). 115-121. Disponible en http://scielo.sld.cu/pdf/ctr/v38n1/ctr15117.pdf

MOREJÓN , R., DIAZ , S., & PÉREZ , N. "Aplicación de técnicas multivariadas a la clasificación morfoagronómica de genotipos de arroz obtenidos en la estación experimental "Los palacios"". *Cultivos Tropicales*, (2001). 43-48. Disponible en http://www.redalyc.org/html/1932/193218206008/

MUÑOZ GARCIA , J. A., & AMÓN URIBE , I.. "Técnicas para detección de outliers multivariantes". *Revista en Telecomunicaciones e Informática*, (Enero de 2013), 11-25.

MURIEL OSORIO , J. "Diferencias en el rendimiento de la producción de arroz en el norte de Perú bajo la variable género". *Universidad del Valle*, (2013). 43-45. Disponible en http://bibliotecadigital.univalle.edu.co/bitstream/10893/5663/1/0461886-p.pdf

NORDHAUSEN , K., & RUIZ GAZEN, A. (April de 2017). University of Toulouse. Disponible en University of Turku.

OCAÑA PEINADO, F. M. .Universidad de Granada. (Noviembe de 2018). Disponible en https://www.ugr.es/~fmocan/MATERIALES%20DOCTORADO/Tratamiento%20de%20outliers%20y%20missing.pdf

OCHOA HERRERA , E. A. "Evaluación de un sistema de intensificación del cultivo de arroz(SICA) bajo condiciones ambientales de Churute", (2016). Universidad de Cuenca, Guayas, Ecuador. Cuenca:. Disponible en http://dspace.ucuenca.edu.ec/bitstream/123456789/23951/1/tesis.pdf

ORGANIZACIÓN DE LAS NACIONES UNIDAS PARA LA AGRICULTURA Y LA ALIMENTACIÓN (FAO) . (Junio de 2018). http://www.fao.org/. Disponible en http://www.fao.org/economic/est/publicaciones/publicaciones-sobre-el-arroz/seguimiento-del-mercado-del-arroz-sma/es/

ORGANIZACIÓN DE LAS NACIONES UNIDAS PARA LA AGRICULTURA Y LA ALIMENTACIÓN (FAO). (2004). "El arroz y la nutrición humana". Recuperado de: http://www.fao.org/rice2004/es/f-sheet/hoja3.pdf

OTERO GARCÍA , B. “Imputación de datos faltantes en un sistemas de información sobre conductas de riesgo”. 2011, Universidad de Santiago de Compostela. Coruña:. Disponible en http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_616.pdf

OTERO GARCÍA , D. ”Imputación de datos faltantes en un sistema de Información sobre Conductas de Riesgo”. (2011) Coruña. Disponible en http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_616.pdf

PAGÉS, J. (2015). “Multiple Factor Analysis by Example Using R. Boca Raton: CRC Press”. Disponible en https://www.oreilly.com/library/view/multiple-factor-analysis/9781498786690/

PEDROSA , I., JUARROS BASTERRETXEA , J., ROBLES FERNANDEZ, A., BASTEIRO, J., & GARCIA CUETO, E.. “Pruebas de bondad de ajuste en distribuciones simétricas, ¿qué estadístico utilizar?”, (Marzo de 2015), Universidad Psicológica, 245-254.

PÉREZ , C. “Tecnicas de segmentación, Conceptos, herramientas y aplicaciones”. *REIRE*, . (2011). 65-79. Disponible en http://diposit.ub.edu/dspace/bitstream/2445/43762/1/618361.pdf

Pérez , C., & Santín , D.. “Minería de datos: Técnicas y herramientas”. (2007), : Ediciones Paranino S.A , Madrid. Disponible en http://diposit.ub.edu/dspace/bitstream/2445/43762/1/618361.pdf

PLANCHUELO GÓMEZ, Á. (2017)., “Comparativa de Análisis de imputación de datos faltantes con análisis de casos completos en pruebas diagnósticas”. Madrid : Universidad Complutense de Madrid, . Disponible en https://eprints.ucm.es/43961/1/TFM_PlanchueloGomez.pdf

PLATEK, R.. “Metodología y tratamiento de la no respuesta. Seminario internacional de estadística en EUSKADI EUSTAT”. (1986), (págs. 48-53), Disponible en http://www.eustat.eus/prodserv/datos/vol0010.pdf

POMBOZA TEMAQUIZA, P. P, “Prácticas ancestrales en cultivos de papa”. *Acta Agronómica*, . (2017). , 157-163. Disponible en http://www.scielo.org.co/pdf/acag/v66n2/0120-2812-acag-66-02-00157.pdf

PORRAS CERRON , J. C.. “Comparación de pruebas de normalidad multivariada”. *Anales Científicos*, (2016), 141-146, Disponible en http://dx.doi.org/10.21704/ac.v77i2.483

POZO GALARRAGA, C. E. (2017). “Incidencia de la variación de los precios financieros y de eficiencia de los fertilizantes químicos en la estructura de costos de producción y en la rentabilidad de los cultivos de arroz, maiz duro, quinua, banano y caña de azúcar. Periodo 2013-2016. Quito:

Pontificia Universidad Católica del Ecuador". Disponible en http://repositorio.puce.edu.ec/handle/22000/14242

PUERTA GOICOECHEA, A.. "Imputación basada en árboles de clasificación". *Eustat*. (2002), Disponible en http://www.eustat.eus/document/datos/ct_04_c.pdf

RAO , C. R.. "The use and interpretation of principal components analysis in applied research". *Sankhya*, (Diciembre de 1964), 26, 329-358. Disponible en https://www.jstor.org/stable/25049339?seq=1#page_scan_tab_contents

RIVA , J., VALER , F., & PEREZ , J. (2014). "Manejo de pastos naturales", Lima. Disponible en http://www.paccperu.org.pe/publicaciones/pdf/147.pdf

RODRÍGUEZ JAUME , M., & MORA CATALÁ , R. (Noviembre de 2001). Repositorio Institucional de la Universidad de Alicante. Disponible en https://rua.ua.es/dspace/bitstream/10045/8145/1/EXPLORATORIO.pdf

RUBIN , D. "Multiple Imputation After 18+ Years". *Journal of the American Statistical Association*, (June de 1996), 473-489. Disponible en https://www.jstor.org/stable/2291635?seq=1#page_scan_tab_contents

RUBIN , D., & LITTLE, R.. "Statistical analysis with missing data". *Wiley*, (2002), 140. Disponible en https://www.wiley.com/en-us/Statistical+Analysis+with+Missing+Data%2C+2nd+Edition-p-9780471183860

RUIZ MUÑOZ , D. (s.f.). Manual de Estadística.

SALVADOR FIGUERAS, M., & GARGALLO , P.." Introducción a las finanzas". (1 de Noviembre de 2018) Disponible en https://ciberconta.unizar.es/leccion/aed/ead.pdf

SERNA PINEDA, S. C. (2009). "Comparación de Árboles de regresión y Clasificacion y Regresión Logística. Medellin": Universidad Nacional de Colombia.

TASCÓN , E., & GARCÍA , E. "Arroz: investigación y producción: referencia de los cursos de capacitación sobre arroz dictados por el Centro Internacional de Agricultura Tropical", (1985) Disponible en https://cgspace.cgiar.org/handle/10568/54403

TRIVEZ , J.. "Efectos de los distintos tipos de outliers en las predicciones de los modelos ARIMA". *Estadística Española*, (1994), 36(135), 21-58.

UNIVERSIDAD AUTÓNOMA DE MADRID. "Métodos Gráficos del Análisis Exploratorio de Datos Espaciales". *www.asepelt.org*, (2003). 3-6. Disponible en

http://www.asepelt.org/ficheros/File/Anales/2003%20-%20Almeria/asepeltPDF/93.PDF

USECHE, L., & MESA, D.. "Una introducción a la imputación de valores perdidos". Terra, (2006), 22(31), 127-152. Disponible en https://www.redalyc.org/pdf/721/72103106.pdf

VALERO OREA , S., SALVADOR VARGAS , A., & GARCIA ALONSO, M., "Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decision y el algoritmo de los k vecinos más cercanos". *Recursos Digitales para la Educación y la Cultura*, (2010)., KAAMBAL, 33-39. Disponible en http://www.itsmotul.edu.mx/ccita2011/documentos/Recursos_digitales.pdf#page=34

VELÁSQUEZ BURBANO , V. A. (2016)**.** "Análisis económico, social, político de la cadena agroalimnetaria del arroz en el Ecuador, periodo 2005-2014". Pontificia Universidad Católica del Ecuador, 120-124. Disponible en http://repositorio.puce.edu.ec/handle/22000/12428

WALPOLE, R., & MYERS, R. (2012). "Probabilidad y estadística para ingenieros". México : PEARSON EDUCACIÓN.

ZAMBRANO , A. (14 de Octubre de 2019). Rpubs. Recuperado de: https://rpubs.com/alexjzc/graf

ZUBCOFF, J. J. (2017). "FAMD(Factor analysis of mixed data)". Alicante: Universidad de Alicante. Disponible en https://rua.ua.es/dspace/bitstream/10045/72567/3/Analisis-multivariante-FAMD.pdf

ZULUAGA DOMINGUEZ , C. M.. "Análisis Estadístico Multivariado: una herramienta estratégica para el control de procesos y calidad en la insdustria alimentaria". *Publicaciones e Investigaciones*, (2011), 143-157. Disponible en http://oaji.net/articles/2017/5082-1501177330.pdf

ANEXOS

ANEXO D: ACCESO A INFORMACIÓN INVESTIGADA

Acceso a la información

Para poder obtener acceso a los datos usad, diríjase al siguiente enlace:
https://www.ecuadorencifras.gob.ec/informacion-de-anos-anteriores-espac/

ANEXO E: CÓDIGO DE PROGRAMACIÓN USADO

Lectura de información

```
setwd("Dirección dentro del computador")
library(readr)
datos <-read_table2("datos.txt", na ="NA", col_names = T)
dcl<-datos[ ,c(1,2,3,6,7,8,10,12,13,14,15,16,19,22,23 )]
dct<-datos[,c(4,5,9)]
```

Preprocesamiento de la información

Categorización de variables cualitativas

```
dclf1<-as.data.frame(dcl)
dclf1$Condicion_economica<-factor(dclf1$Condicion_economica, levels=c(1,2,3),
labels=c("Solo", "Asociado", "Invernadero"))
dclf1$Rotacion_cultivo<-factor(dclf1$Rotacion_cultivo, levels=c(1,2), labels=c("Si_rotacion",
"No_rotacion"))
dclf1$Clase_semilla<-factor(dclf1$Clase_semilla, levels=c(1,2,3,4,5), labels=c("Comun",
"Mejorada", "Hibrida Nacional", "Hibrida Internacional", "Comun"))
dclf1$Uso_riego<-factor(dclf1$Uso_riego, levels=c(1,2), labels=c("Si_riego", "No_riego"))
dclf1$Uso_fertilizantes<-factor(dclf1$Uso_fertilizantes, levels=c(1,2),
labels=c("Si_fertilizante", "No_fertilizante"))
dclf1$Uso_fitosanitarios<-factor(dclf1$Uso_fitosanitarios, levels=c(1,2),
labels=c("Si_fitosanitarios", "No_fitosanitarios"))
dclf1$Problemas_sembrio<-factor(dclf1$Problemas_sembrio, levels=c(1,2,3,4,5,6,7),
labels=c("Sequia/Heladas", "Plagas/Enfermedades", "Inundacion/Exceso de Agua", "Semilla",
"Practicas inadecuadas/Falta de practicas", "Edad de la plantacion", "Ninguna"))
dclf1$Preparacion_suelo<-factor(dclf1$Preparacion_suelo, levels=c(1,2),
labels=c("Si_preparacion_suelo", "No_preparacion_suelo"))
dclf1$Deshierbe<-factor(dclf1$Deshierbe, levels=c(1,2), labels=c("Si_deshierbe",
"No_deshierbe"))
dclf1$Aporque<-factor(dclf1$Aporque, levels=c(1,2), labels=c("Si_aporque", "No_aporque"))
dclf1$Tutoreo<-factor(dclf1$Tutoreo, levels=c(1,2), labels=c("Si_tutoreo", "No_tutoreo"))
dclf1$Uso_fertilizante_organico<-factor(dclf1$Uso_fertilizante_organico, levels=c(1,2),
labels=c("Si_fertilizante_organico", "No_fertilizante_organico"))
dclf1$Uso_fertilizante_quimico<-factor(dclf1$Uso_fertilizante_quimico, levels=c(1,2),
labels=c("Si_fertilizante_quimico", "No_fertilizante_quimico"))
dclf1$Uso_plaguicida_organico<-factor(dclf1$Uso_plaguicida_organico, levels=c(1,2),
labels=c("Si_plaguicida_organico", "No_plaguicida_organico"))
dclf1$Uso_plaguicida_quimico<-factor(dclf1$Uso_plaguicida_quimico, levels=c(1,2),
labels=c("Si_plaguicida_quimico", "No_plaguicida_quimico"))
```

Análisis descriptivo de variables cuantitativas

Variables Producción

```
min(dct$Produccion, na.rm = T)
max(dct$Produccion, na.rm = T)
mean(dct$Produccion, na.rm = T)
median(dct$Produccion, na.rm = T)
which.max(dct$Produccion) # moda
sd(dct$Produccion, na.rm = T)
var(dct$Produccion, na.rm = T)
quantile(dct$Produccion, 0.25, na.rm = T)
quantile(dct$Produccion, 0.75, na.rm = T)
skewness(dct$Produccion, na.rm = T) #- Asimetria
kurtosis(dct$Produccion, na.rm = T) #- Curtosis
length(dct$Produccion)
summary(dct$Produccion, na.rm = T)
bp <- ggplot(dct, aes(factor(0), Produccion))
bp + geom_boxplot(fill="#0072B2", color="#000000") + xlab("") +ylab("Producción")+
scale_x_discrete(breaks = NULL)
bh<-ggplot(data=dct, aes(dct$Produccion)) +  geom_histogram(fill="#0072B2",
color="#000000")+ # dibujamos el histograma
  labs(x="Producción de arroz (Libras)", y="Frecuencia")
h3<- hist(x = dct$Produccion)
th3<-table.freq(h3)
h3<- hist(x = dct$Produccion, breaks = 4)
th3<-table.freq(h3)
```

Análisis descriptivo de variables cualitativas

```
ggplot(data=dclf1, aes(x=dclf1$Condicion_economica, y="")) +   geom_bar(stat="identity",
position="stack")+  labs(x="Clase", y="Frecuencia")
```

Análisis bivariado de variables

```
library(GGally)
ggpairs(dct,lower = list(continuous = "smooth"))
# Superficie sembrada vs Producción
ggplot(data = dct, aes(x = dct$Superficie_sembrada  , y = dct$Produccion)) +
  geom_point()+
  geom_smooth(color = "firebrick") +
  geom_hline(yintercept = 0) +
  theme_bw()+
  labs(x="Superficie sembrada", y="Producción")
# Superficie cosechada vs Producción
ggplot(data = dct, aes(x = dct$Superficie_cosechada  , y = dct$Produccion)) +
  geom_point()+
  geom_smooth(color = "firebrick") +
  geom_hline(yintercept = 0) +
  theme_bw()+
  labs(x="Superficie cosechada", y="Producción")
```

Imputación datos cuantitativos

```
f1<- function (x)
{
  mp<-as.matrix(x)
```

```
  cm <- cor(mp,use = "complete.obs")
  nc<-ncol(cm)
  nf<-nrow(mp)
  for (i in 1:nc){for (j in 1:nc){if(i==j){cm[i,j]<-0}}}
  for (i in 1:nc){    v1<- i    v2<-ifelse(max(cm[ ,v1])>=(-1*(min(cm[ ,v1]))), r1<-
which.max(cm[ ,v1]), r1<-which.min(cm[ ,v1]))
    r<-coef(lm(mp[ ,v1]~mp[ ,v2]))
    cr1<-r[1]
    cr2<-r[2]
    for (j in 1:nf){ifelse(is.na(mp[j,i]), mp[j,i]<-(cr1+(cr2*mp[j, v2])), mp[j,i]<-mp[j,i])}}
  res<-as.data.frame(mp)
  return(res)
}
dctf1<-f1(dct)
dctf2<-dctf1
```

Detección de datos atípicos

```
library(chemometrics)
df.1<-dctf2
md.1 <- Moutlier(df.1, quantile = 0.99, plot = T)
md.1$cutoff #punto de corte
summary(md.1$rd)

# Q-Plot
df <- data.frame(md.1$rd)
p <- ggplot(df, aes(sample = dctf2$Produccion))
p + stat_qq() + stat_qq_line(colour="red")+
  labs(x="Distancias de Mahalanobis", y="Producción")
# Histograma
aux<-data.frame(md.1$rd)
# Histogramas de Distancias de Mahalanobis
ggplot(data=aux, aes(aux$md.1.rd )) +
  geom_histogram()+ # dibujamos el histograma
  labs(x="Distancias de Mahalanobis", y="Frecuencias")
# Diagrama de densidad
ggplot(data=aux, aes(aux$md.1.rd)) +   geom_density(alpha=0.7)+   labs(x="Distancias de
Mahalanobis", y="Densidad")
# Identificación de los datos atípicos
atipicosdmhr<-which(md.1$rd > md.1$cutoff)
Modelo de regresión
model<-lm( formula = Produccion ~ Superficie_sembrada + Superficie_cosechada, data=dctf2)
summary(model)
# 1. Multicolinealidad
library(car)
vif(model)
# 2. Supuesto Independencia
library(lmtest)
dwtest(model)
# 3. Normalidad de los residuos
```

```
## Cálculo
model_norm<- rstudent( model )
shapiro.test( model_norm )
library(nortest)
ad.test(model_norm)
# Gráfico
# QQplot - Residuos
library(ggplot2)
df <- data.frame(residuals(model))
p <- ggplot(df, aes(sample = residuals.model.))
p + stat_qq() + stat_qq_line(colour="red")+
  labs(x="Valores teóricos distribución Normal", y="Residuos")
# Histogramda de residuos
dh1<-qplot(model$residuals,
      geom="histogram",
      binwidth = 0.5,
      xlab = "Residuos",
      fill=I("red"),
      col=I("black"),
      alpha=I(.2))
dh1
# Densidad de los residuos
dh2<-ggplot(data=aux, aes(model$residuals)) +
  geom_density(alpha=0.7)+ # dibujamos el diagrama de densidad
  labs(x="Residuos", y="Densidad")
dh2
# 4. Homocedasticidad
library(lmtest)
bptest( model )
# Gráfico
val_ajust <- fitted( model) # valores ajustados # valores ajustados #Valores ajustados: valores
ajustados (valores de la variable respuesta) para las observaciones originales de la predictora.
val_resid <- residuals( model) # residuos # residuos #Residuos: diferencia entre valor observado
de la respuesta y valor ajustado por el modelo.
val_resid_estanda <- rstudent( model) # residuos estudentizados # residuos estudentizados
#Estadísticos: residuos estudentizados del modelo ajustado.
ggplot(data = dctf2, aes(x =  fitted(model), y = rstudent(model)) ) + geom_point()+
stat_smooth(colour="red")+
  labs(x="Predicciones variable dependiente (Y)", y="Residuos estandarizados")
# 5. Linealidad
# Residuos vs independientes
ggplot(data = dctf2, aes(x = Superficie_sembrada  , y = model$residuals)) +  geom_point()+
geom_smooth(color = "firebrick") +  geom_hline(yintercept = 0) +
theme_bw()+labs(x="Superficie sembrada", y="Residuos")
ggplot(data = dctf2, aes(x = Superficie_cosechada , y = model$residuals)) +  geom_point()+
geom_smooth(color = "firebrick") +  geom_hline(yintercept = 0) +  theme_bw()+
labs(x="Superficie cosechada", y="Residuos")
ggplot(data = dctf2, aes(x = Produccion , y = model$residuals)) +  geom_point()+
```

```
geom_smooth(color = "firebrick") +  geom_hline(yintercept = 0) +  theme_bw()+
labs(x="Producción", y="Residuos")
```

Árboles de regresión

```
library(DMwR2)
library(rpart)
library(rpart.plot)
library(rattle)
library(caret)
library(AmesHousing)
library(rsample)
library(tidyverse)
# Eliminación de datos atípicos
mat<-cbind(dctf2, dclf1)
mat<-mat[-atipicosdmhr,]
# Generación del modelo idóneo
mat1<-mat[,c(1,2,3, 6, 10,7,15)]
tbase3 <- rpart(Produccion ~ ., data =mat1, method = "anova")
printcp(tbase3)
plotcp(tbase3)
rpart.plot(tbase3)
summary(tbase3)
fancyRpartPlot(tbase3)
library(party)
modelo1 <- ctree(Produccion ~ ., data = mat1)
plot(modelo1)
summary(modelo1)
modelo1

#Validación del modelo
set.seed(12345) # muestra aleatoria fija
nr<-trunc(dim(mat)[1]*0.9)
rndSample <- sample(1:nrow(mat), nr) # Generación de la muestra de entrenamiento
tr <- mat[rndSample, ] # Obtención de la muestra de entrenamiento
ts <- mat[-rndSample, ] # Obtención de la muestra de prueba
ps <- predict(modelo1, ts) # Probabilidades de pertenecer a un grupo(muestra prueba)
ps1<- cbind(ps, ts$Produccion) # Probabilidad de pertenecer a un grupo junto con grupo
head(ps1)
head(ps1)
tetn1<-mean(abs(ps1 - ts$Produccion)) # Media de error
tetn2<-mean(abs((ps1 - ts$Produccion)^2)) # Media de error cuadrático
tetn1
tetn2
sqrt(tetn2) # RMSE
```

Análisis factorial de datos mixtos

```
#Libraries
library("FactoMineR")
library("factoextra")
library("corrplot")
```

```
library("dplyr")
# Estandarización de variables cuantitativas
datct<-dctf2[-atipicosdmhr,]
datcl<-dclf1[-atipicosdmhr,]
de<- as.data.frame(scale(datct))
# Matriz estandarizada sin atípicos
mat<-cbind(de, datcl)
# Preparación de la matriz para AFDM
bd.r<-mat
str(bd.r)
tail(bd.r)
# Generación de un modelo optimo
res.famd<-FAMD(bd.r,graph=F,ncp=10)
print(res.famd)
# Extraer los autovalores
eig.val<-get_eigenvalue(res.famd)
round(eig.val,2)
# variabilidad explicada por cada dimensión, y variabilidad acumulada explicada
# Gráfico de sedimentación (los porcentajes de inercia explicados por cada dimensión del
AFDM)
fviz_screeplot(res.famd,addlabels=TRUE,main="Gráfico de sedimentación",
ylab="Porcentaje de variabilidad explicada", xlab="Dimensiones")
# Resultados del AFDM para las variables
var<-get_famd_var(res.famd)
var$coord
var$contrib
var$cos2

# 1 "$ coord" "Coordenadas para las variables"
# 2 "$ cos2" "Calidad en el mapa factorial"
# 3 "$ contrib" "Contribuciones a los componentes principales"

# Gráfico de variables - Correlación de variables con los factores
round(head(var$coord),2)
fviz_famd_var(res.famd,"var",          repel=TRUE, # Evitar superposición de texto (lento)
ggtheme=theme_minimal())
fviz_famd_var(res.famd,"var",col.var="blue4",
repel=TRUE,ggtheme=theme_minimal())
# Contribuciones de variables a las dimensiones del AFDM
round(head(var$contrib,4),3)
corrplot(var$contrib,is.corr=FALSE)
## Contribuciones de variables para la dim 1 y dim2
## 1/longitud(variables) = 1/10 = 10%
fviz_contrib(res.famd,choice="var",axes=1:2,top=14)
# Diagrama de barras del cos2 de variables en las dos primeras dimensiones
fviz_cos2(res.famd,choice="var",axes=1:2)
## Las variables más importantes (o contribuyentes)
## se pueden resaltar en la gráfica de correlación
```

```
fviz_famd_var(res.famd,"var",col.var="contrib",  gradient.cols=c("green","blue","red"),
repel=T)
# Descripción de la dimensión
## Variables de acuerdo con sus contribuciones a las dimensiones del AFDM
res.desc<-dimdesc(res.famd,axes=c(1,2),proba=0.05)

## Descripción de la dimensión 1
res.desc$Dim.1
## Descripción de la dimensión 2
res.desc$Dim.2
# Resultados del AFDM para los individuos
ind<-get_famd_ind(res.famd)
# 1 "$ coord" "Coordenadas para los individuos"
# 2 "$ cos2" "Calidad para los individuos"
# 3 "$ contrib" "Contribuciones de los individuos"
# Gráfico de individuos
fviz_famd_ind(res.famd)

## Variables cuantitativas (ACP)
quanti.var<-get_famd_var(res.famd,"quanti.var")
# representar las variables cuantitativas (círculo de correlaciones)
fviz_famd_var(res.famd,"quanti.var",repel=T,  col.var="black")

# Calidad de representación en el plano factorial
round(head(quanti.var$cos2,4),3)

#cos2 de variables en todas las dimensiones
corrplot(quanti.var$cos2,is.corr=FALSE)
## Total cos2 de variables en Dim.1 y Dim.2
## diagrama de barras de las variables cos2
fviz_cos2(res.famd,choice="quanti.var",axes=1:2)

# Color por valores de cos2: calidad en el mapa de factores
fviz_famd_var(res.famd,"quanti.var",col.var="cos2",
gradient.cols=c("green","blue","red"),
        repel=T)
# Contribuciones de variables cuantitativas a las dimensiones del ADFM
round(head(quanti.var$contrib,4),3)
corrplot(quanti.var$contrib,is.corr=FALSE)
## Las variables más importantes (o contribuyentes), se pueden resaltar en la gráfica de
correlación
fviz_famd_var(res.famd,"quanti.var",col.var="contrib",
gradient.cols=c("green","blue","red"),  repel=T)
```

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

DIRECCIÓN DE BIBLIOTECAS Y RECURSOS PARA EL APRENDIZAJE Y LA INVESTIGACIÓN

UNIDAD DE PROCESOS TÉCNICOS

REVISIÓN DE NORMAS TÉCNICAS, RESUMEN Y BIBLIOGRAFÍA

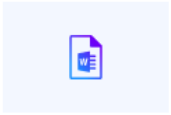
Fecha de entrega: 18 / 06 / 2020

INFORMACIÓN DEL AUTOR/A (S)
Nombres – Apellidos: José Luis Condo León
INFORMACIÓN INSTITUCIONAL
Facultad: Ciencias
Carrera: Ingeniería en Estadística Informática
Título a optar: Ingeniero en Estadística Informática
f. Analista de Biblioteca responsable: 0074-DBRAI-UPT-2020  

Jaime Alberto Tapia Salinas <jaime.tapia@espoch.edu.ec>
Para: hola_josecondo@yahoo.es

mié., 4 mar. a las 8:29

Jose, buen dia sirvase adjunto encontrar el abstract de su trabajo de titulacion
Exitos en su vida profesional
Porfessor Jaime Tapia
Docente Centro de Idiomas



Jose Condo.docx
14.3kB

ABSTRACT

This research aimed to compare multivariate statistical techniques to identify and classify the influential factors in rice production in Ecuador in 2017. 18 agronomic variables, 15 qualitative and 3 quantitative measures were used in rice fields that participated in the Survey of Surface and Continuous Agricultural Production (ESPAC) 2017, the research was not experimental and had a relational exploratory scope, an Exploratory Data Analysis (AED), Regression Trees (AR), Mixed Data Factorial Analysis (AFDM) and the R software to develop the study. The AED determined that the average rice production is 212.21 pounds per hectare, 86% of the crops produced between 200 and 250 pounds, it was determined that the variables planted area and area harvested have no correlation with the rice production variable. After the application of multivariate methods, the AR model with an RMSE-7.32 value detected as influencing factors in the production of rice on sown surface, harvested area, use of irrigation, use of organic fertilizer and seed class. On the other hand, the AFDM with a model that explained 81% of the variability of the data generated two groups of factors influencing rice production, the first group is composed of the variables fertilizer use, chemical fertilizer use, use of chemical pesticides and the use of phytosanitary products, while the second group is composed of planted area and harvested area. For this reason, it was possible to determine that the influential factors that the two techniques have in common are planted area and harvested area, the most important factors for RA are use of irrigation and harvested area, while for AFDM they are use of fertilizers and surface sown. It is necessary to expand research with the use of other data mining techniques.

Keywords: <FACTORIAL ANALYSIS OF MIXED DATA>, <REGRESSION TREES>, <INFLUENCING FACTORS>, <ANALYSIS OF MAIN COMPONENTS>, <CORRESPONDENCE ANALYSIS>, <RICE PRODUCTION>, <AGRONOMIC VARIABLES>.